Author:
Clayton, Gemma

Title:
Incorporating external evidence syntheses in the design and analysis of trials

Author:
Clayton, Gemma

Title:
Incorporating external evidence syntheses in the design and analysis of trials

INCORPORATING EXTERNAL EVIDENCE SYNTHESES IN THE DESIGN AND ANALYSIS OF TRIALS

GEMMA LOUISE CLAYTON

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF BRISTOL IN ACCORDANCE WITH THE REQUIREMENTS FOR AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN THE FACULTY OF HEALTH SCIENCES

BRISTOL MEDICAL SCHOOL MAY, 2019

$\approx$ 71,000 words

Summary

When a trialist is designing a trial, related work will often be used to inform several aspects. This information is often used informally, such as using a systematic review to indicate whether a gap in the current evidence base justifies a new trial. External evidence can be used more formally, by explicitly incorporating it in a Bayesian framework through a prior distribution.

Funders often highlight the importance of taking into account existing evidence when planning a new trial. Researchers and trialists acknowledge that existing evidence should be used to inform new research to reduce research waste. However, the prevalence of explicitly using external evidence through informative prior distributions is low and there is still much controversy around its use in all stages of a trial.

In this thesis, we explore whether and how trialists could use a synthesis of external evidence in the design and analysis of a clinical trial, through a Bayesian analysis. We begin with a survey and qualitative study to capture the current use of evidence synthesis by trialists and reasons why it might not be used in practice. In the remainder of the thesis, we assess and extend methods in areas where external evidence could have the most benefit to a trialist. We focus on the following three case studies:

- External evidence on likely *bias* in a trial, based on information from meta-analyses within meta-epidemiological studies.

- External evidence on the likely *effect size* in a trial, based on information from similar trials within meta-analyses.

- External evidence on likely *outcomes* in a trial, based on information from similar patients within trials.

Acknowledgements

First, I would like to thank my supervisors, Hayley Jones and Julian Higgins, for their guidance and support throughout the last three years. I feel extremely grateful and lucky to have worked with them both.

Also, thank you to Daisy Elliott, Nicky Welton and Jonathan Sterne who provided helpful direction and feedback in some important stages of this thesis.

I would like to thank the MRC ConDuCT-II Hub for Trials Methodology Research for funding my PhD and providing an invaluable experience.

Throughout this thesis I have had the opportunity to collaborate with many people and would therefore like to thank the MRC Evidence Synthesis Working Group, Asbjørn Hróbjartsson, Helene Moustgaard, and the wider MetaBLIND team. I would also like to thank Novartis for the opportunity of a three-month internship in Basel. Thanks, in particular to Laurence Colin, Baldur Magnusson, Yue Li and Asher Schachter.

I would also like to thank my family and friends. In particular, my friend Adam Trickey, for his continued positivity and enthusiasm throughout. My parents, David and Louise Clayton as well as my two brothers, Tom and James, and sister-in-law, Becky.

Last but not least, this thesis is dedicated to Lily Clayton and Chris Newell.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .... DATE:..

Preface

Below is a list of publications that have arisen from research undertaken throughout this PhD along with the Contributions section from each publication.

G. L. Clayton, I. L. Smith, J. P. T. Higgins, B. Mihaylova, B. Thorpe, R. Cicero, K. Lokuge, J. R. Forman, J. F. Tierney, I. R. White, L. D. Sharples, and H. E. Jones. The invest project: investigating the use of evidence synthesis in the design and analysis of clinical trials. Trials, 18, 2017. ISSN 1745-6215. doi: 10.1186/s13063-017-1955-y.

"The project was initially conceived by LS, as a collaboration between members of the MRC HTMR's Evidence Synthesis Working Group. All authors were involved in the design of the survey and interpretation of the results. IS implemented the online version of the survey, with support from BT and RC. GC, IS, BT, RC, KL, JF, JT, IW and LS promoted the survey during the ICMTC. RC and BT wrote code to extract the data, which were analysed by GC. GC and HEJ wrote the paper, with IS, JH, BM, KL, JT, LS and IW contributing to critical revisions. All authors read and approved the final manuscript."

G. L. Clayton, A. D. Schachter, B. Magnusson, Y. Li, and L. Colin. How often do safety signals occur by chance in first-in-human trials? Clinical and Translational Science, 11(5):471–476, 2018. ISSN 1752-8054. doi: 10.1111/cts.12558.

"L.C., G.C., and B.M. wrote the article; A.S. and L.C. designed the research; L.C., G.C., B.M., and Y.L. performed the research; G.C.analyzed the data."

Contents

List of Figures 11

List of Tables 16

1 Introduction 19

1.1 Rationale of thesis . . . 19

1.2 Clinical trials . . . 21

1.2.1 Phases of a clinical trial . . . 21

1.2.2 Approaches to clinical trials . . . 23

1.2.3 Analysis of a clinical trial . . . 24

1.3 Evidence synthesis methods . . . 28

1.3.1 Introduction to evidence synthesis methods . . . 28

1.3.2 Meta-analyses . . . 28

1.3.3 Other types of evidence synthesis methods . . . 33

1.4 Current uses of external evidence in trials . . . 34

1.4.1 Aims of the thesis . . . 42

1.4.2 Structure for the thesis . . . 42

2 Current practice and attitudes: the INVEST survey 46

2.1 Introduction and aims . . . 46

2.1.1 Aims . . . 47

2.2 Methods . . . 49

2.2.1 All participants . . . 49

2.2.2 Use of evidence synthesis to inform trial design . . . 50

2.2.3 Use of evidence synthesis to inform trial analysis . . . 51

2.2.4 Analysis of survey responses . . . 51

2.3 Results . . . 52

2.3.1 Use of evidence synthesis to inform trial design 55
2.3.2 Use of evidence synthesis to inform trial analysis 60
2.3.3 Barriers to the use of evidence synthesis methods 65
2.4 Discussion 65
2.4.1 Overview of key findings 65
2.4.2 Limitations 67
2.4.3 Comparison to recent reviews of evidence synthesis methods in trial design 71
2.5 Conclusions and rationale for remainder of thesis 73

3 Eliciting current views and exploring barriers: a qualitative study 75

3.1 Introduction 75
3.1.1 Context and objective 75
3.1.2 Background to qualitative research 76
3.2 Methods 80
3.2.1 Recruitment and sampling (eligibility criteria) 80
3.2.2 Data collection 81
3.2.3 Data analysis 83
3.3 Results 84
3.3.1 Participants 84
3.3.2 Key findings 87
3.4 Discussion 107
3.4.1 Overview of key findings 107
3.4.2 Reflexivity 109
3.4.3 Methodological strengths 109
3.4.4 Limitations 110
3.4.5 Relevance of this study in relation to other studies 111
3.4.6 Implications for future research 112

3.5 Conclusions and implications for remainder of thesis 115

4 A meta-epidemiological investigation of the impact of blinding on estimated treatment effects in randomised clinical trials: the MetaBLIND study 117

4.1 Context and overview 117

4.2 Introduction and aims 118

4.3 Background to meta-epidemiological methods 120

4.3.1 Welton *et al* model 122

4.4 Methods 125

4.4.1 Screening and data extraction 125

4.4.2 Classifications 126

4.4.3 Data analysis 128

4.5 Results 134

4.5.1 Comparison between binary and continuous outcomes 150

4.6 Discussion 153

4.6.1 Overview of key findings 153

4.6.2 Limitations 155

4.6.3 Comparison to recent meta-epidemiological studies 156

4.6.4 Conclusions and rationale for remainder of thesis 162

5 Statistical considerations in meta-epidemiological evidence of non-dichotomous study characteristics 164

5.1 Context and overview 164

5.2 Meta-epidemiological considerations for categorical study characteristics with an application to the role of unclear risk of bias trials 166

5.2.1 Description of models 166

5.2.2 Case study application 174

5.2.3 Results 175

5.2.4 Using the outputs of meta-epidemiological models to inform a new trial 182

5.2.5 Summary 184

5.3 Meta-epidemiological considerations for continuous study characteristics: examining the association of sample size on estimated treatment effects . . 185

5.3.1 Introduction and aims 185

5.3.2 Description of the models 186

5.3.3 Case study application: sample size 187

5.3.4 Results 191

5.3.5 Summary 196

5.3.6 Comparison to other studies 196

5.4 Conclusions 197

6 Comparing methods for incorporating external evidence on the effect size in sample size calculations 199

6.1 Introduction and aims 199

6.2 Background 200

6.2.1 Classical power - (1) 203

6.3 Case study: The CRASH trial 207

6.4 A comparison of methods for power calculations 210

6.4.1 Bayesian analysis of the new trial with an informative prior distribution - (2) 211

6.4.2 Updated meta-analysis - (3) 216

6.4.3 Updated cost-effectiveness analysis, EVSI - (4) 221

6.5 Results 225

6.5.1 Bayesian analysis of the new trial with an informative prior distribution - (2) 225

6.5.2 Updated meta-analysis - (3) . . . 226

6.5.3 Updated cost-effectiveness analysis, EVSI - (4) . . . 226

6.6 Discussion . . . 230

6.6.1 Overview of key findings . . . 230

6.6.2 Are these methods implementable in routine trial design? . . . 232

6.6.3 Towards recommendations . . . 233

6.6.4 Areas for further research and consideration . . . 234

6.6.5 Conclusions . . . 236

7 Using external evidence on adverse outcomes with an application in first in human studies 238

7.1 Context and overview . . . 238

7.2 Introduction and aims . . . 239

7.3 Summary of the available empirical data . . . 241

7.4 Methods . . . 246

7.4.1 Outcome . . . 246

7.4.2 Description of the model . . . 250

7.4.3 Model covariates . . . 251

7.4.4 Model selection and fit . . . 251

7.4.5 Predicting from the model . . . 252

7.5 Results . . . 254

7.5.1 Univariable analyses . . . 254

7.5.2 Model diagnostics . . . 256

7.5.3 Model estimates and interpretation . . . 256

7.5.4 Using the pooled placebo data to inform a new trial . . . 261

7.6 Discussion . . . 265

7.6.1 Overview of key findings . . . 265

7.6.2 Limitations . . . 266

7.6.3 Previous studies looking at safety data in early phase trials 267
7.6.4 Implications for routine safety data in practice and further research recommendations 268
7.6.5 Conclusions 270

8 Discussion **272**
8.1 Key findings 272
8.2 Implications/recommendations 278
8.3 External evidence in a *routine* clinical trial setting: does it have a place? . . . 281
8.4 Future research 283
8.5 Final conclusions 284

References **286**

Appendices **315**

A Documents and Tables **315**

B WinBUGS code **346**

List of Figures

1.1 Prior and trial data regarding a treatment effect in a clinical trial combined to make a posterior distribution in two scenarios. 27
2.1 **Views of respondents on whether evidence synthesis methods should be used to inform trial design.** The type of evidence synthesis method is summarised across five aspects of trial design: whether a trial is needed (n=104), choice of population (n=103), choice of interventions (n=103), choice of outcomes and follow-up time (n=101), sample size (n=103). 56

2.2 **Comparisons between desirable and current practice in use of evidence synthesis methods in trial design.** This is summarised by type of evidence synthesis method, among survey respondents involved in trial design to inform five aspects of trial design: whether a trial is needed (n=84), choice of population (n=82), choice of interventions (n=82), choice of outcomes and follow-up time (n=81), sample size (n=81). Numbers displayed are percentages. 59

2.3 **Views of respondents on whether evidence synthesis should be used to inform trial analysis.** This is summarised across three aspects of trial analysis: the treatment effect, potential biases arising from trial conduct and other quantities (of n=100 people who answered this question). 61

2.4 **Comparisons between desirable and current practice in use of evidence synthesis methods in trial analysis.** This is summarised among survey respondents involved in trial analysis to inform three aspects of trial analysis: the treatment effect (n=68), potential biases arising from trial conduct (n=69) and other quantities (n=68). Numbers displayed as percentages. . . . 63

2.5 Comparisons between ideals and current practice of evidence synthesis methods *amongst statisticians only* in use of evidence synthesis methods to inform three different aspects of trial analysis: the treatment effect (n=54), potential biases arising from trial conduct (n=54) and other quantities (n=54). 64

2.6 **Barriers to the use of evidence synthesis (higher scores indicate greatest perceived barriers).** 3 points were assigned to the greatest barrier, 2 points to the second and 1 to the third. For example, 38 respondents ranked time constraints as the greatest barrier (3 x 38 = 114 points), 21 ranked it second (2 x 21 = 42) and 11 third (1 x 11=11). 66

2.7 Key findings from the INVEST survey 68

3.1 Implications for further research based on key findings 113

4.1 Study flow diagram. 129

4.2 Estimated RORs and effects on heterogeneity associated with blinding status of patients, healthcare providers and outcome assessors. 144

4.3 RORs from individual meta-analyses and from analyses combined across all meta-analyses. Results for individual meta-analyses are frequentist estimates with confidence intervals, based on comparing the summary odds ratio from studies with the study characteristic of interest with the summary odds ratio from studies without the characteristic. The overall estimates of RORs are results based on the Bayesian hierarchical model described in the main text. CD numbers are identifiers of individual Cochrane reviews, from the Cochrane Database of Systematic Reviews. 151

4.3 (cont.) 151

4.3 (cont.) 152

4.3 (cont.) 152

4.3 (cont.) 153

4.4 Plot of the predictive distributions of the effect of bias from MetaBLIND with the mean effects of bias from other meta-epidemiological studies . . . 160

5.1 Plot of the individual probabilities of unclear RoB trials being high risk trials against their estimated treatment effects and 95% CrIs. 179

5.2 Summary of sample size for each RoB category. 180

5.3 Plot of the individual probabilities of unclear RoB trials being high risk trials, when predicted by sample size against their estimated treatment effects and 95% CrIs. 181

5.4 Plot of the mean posterior distribution of B against study size, for all unclear RoB trials, for both the probability model and the probability model when using sample size as a predictor. Shown for smaller study sizes. 181

5.5 Plot of the individual probabilities of unclear RoB trials being high risk trials, when predicted by sample size against their study size. 182

5.6 Plot of the association between sample size and the predicted treatment effect when $\frac{1}{n}$ is used as the function to represent the relationship which sample size has on treatment effects. 194

5.7 Plot of the association between sample size and the predicted treatment effect when $\frac{1}{\sqrt{n}}$ is used as the function to represent the relationship which sample size has on treatment effects. 195

6.1 The standard normal distribution 204

6.2 Results from a random effects meta-analysis looking at the effect of cortisterioids versus placebo after traumatic brain injury. The black diamonds represent the odds ratios of the individual studies, and the horizontal lines their 95% confidence intervals. The grey squares represent the weight which each study contributes to the overall pooled result. The results of the 16 trials have been pooled in an inverse-variance (I-V) weighted random effects meta-analysis to give an overall weighted average of the treatment effect. . 208

6.3 Power curves for the two-sided test $H_0 : \delta = 0$ vs. the alternative hypothesis $H_1 : \delta \neq 0$. Methods 2(i)) and 3(i) require a prior (to average across) and uses the predictive distribution. Method 3(ii) doesn't require a prior and instead updates the meta-analysis. Method 2(ii) uses the predictive distribution as a prior. We assume $\alpha = 0.05$ and a 'typical' standard deviation σ of 4.52 (estimated as the median across studies). Additionally, for methods (1), (2(ii)) and (3(ii)) a hypothetical effect is set to detect an OR=0.85. 227

6.4 EVSI values according to total sample size of a new trial informing probability of the event and treatment effect in both groups (in terms of quality adjusted life years (QALYs) , ENBS is not worked out here and is therefore not given in terms of cost 228

6.5 Power curves for the two-sided test $H_0 : \delta = 0$ vs. the alternative hypothesis $H_1 : \delta \neq 0$. Methods 2(i)) and 3(i) require a prior (to average across) and uses the predictive distribution. Method 3(ii) doesn't require a prior and instead updates the fixed effect meta-analysis. Method 2(ii) uses the predictive distribution as a prior. We assume $\alpha = 0.05$ and a 'typical' standard deviation σ of 4.52 (estimated as the median across studies). Additionally, for methods (1), 2(ii) and 3(ii) a hypothetical effect is set to detect an OR=0.85. When $\tau = 0$, methods 3(i) and 3(ii) reduce to 2(i) and 2(ii) respectively and therefore curves 3(i)= 2(i) and similarly 3(ii)=2(ii). 229

6.6 Key findings from a case study analysis 232

7.1 Event rate by study. 247

7.2 Distribution of log(ALT/ULN). 249

7.3 Distribution of observed log(ALT/ULN) values vs. the predicted log(ALT/ULN) values. 249

7.4 Simulated mean proportion of events from the posterior predictive distribution (light blue) and the observed proportion of events (dark blue). 256

7.5 Model-based predicted probabilities of a subject experiencing an event of ALT > ULN 258

7.6 Predicted probability of an ALT>ULN for a subject, assuming other covariates are fixed at the mean population values. 25% and 75% quartiles are also displayed. The median probability is assumed. 259

7.7 Probability of observing a given number of subjects with an event (ALT > ULN) by chance in the hypothetical cohort described by Table 7.9. 262

7.8 Probability of observing a given number of subjects with an event (ALT > ULN) by chance using the population average probability 263

7.9 Key findings 265

8.1 Key future research needs. 284

A.1 The INVEST survey. 316
A.1 The INVEST survey. (cont.) 317
A.2 Consent form. 318
A.3 Participant information sheet. 319
A.3 Participant information sheet (cont.) 320
A.4 Example of topic guide 321
A.4 Example of topic guide (cont.) 322
A.5 RORs from individual meta-analyses and from analyses combined across all meta-analyses. Results for individual meta-analyses are frequentist estimates with confidence intervals, based on comparing the summary odds ratio from studies with the study characteristic of interest with the summary odds ratio from studies without the characteristic. The overall estimates of RORs are results based on the Bayesian hierarchical model described in Section 4.3.1. CD numbers are identifiers of individual Cochrane reviews, from the Cochrane Database of Systematic Reviews. 336
A.5 (cont.) 337
A.5 (cont.) 338
A.5 (cont.) 339
A.5 (cont.) 340

List of Tables

1.1 Summary of opportunities for evidence synthesis to inform design, conduct and analysis of a clinical trial. 38
2.1 Respondent characteristics 53
2.2 How evidence synthesis ‘should’ be used in trial design 57

2.3 Actual uses of evidence synthesis in trial design during the last 10 years, compared with uses considered desirable 58

2.4 Types of evidence synthesis used in trial design 59

2.5 How evidence synthesis methods 'should' be used to inform aspects of trial analysis 61

2.6 Actual uses of evidence synthesis in trial analysis during the last 10 years, compared with uses considered desirable 62

3.1 Participant and job-related characteristics 86

3.2 Table summarising themes (and subthemes), supported by a quote 89

4.1 Characteristics of included Meta-analyses and Trials. Overall dataset and main analyses (Ia, Ib, IIa, IIb, III). 136

4.2 Adjusted analyses. 147

4.3 Secondary analyses. 148

4.4 Predictive Distributions of the Effect of Bias 159

4.5 Comparison with other studies 161

5.1 Posterior summaries from each of the models described, examining the influence of not blinding participants. 178

5.2 Probability model using sample size, results in more detail 180

5.3 Posterior summaries from each of the models described, examining the influence of sample size on estimated treatment effects 193

6.1 Summary of the closed form equations for methods (1) – (3). Method (4) does not give a closed form solution because the posterior distribution is updated using MCMC sampling and is therefore not included in this table. 224

7.1 Demographics of the placebo database. 243

7.2 Raw incidence (unadjusted for study effect) of safety signals in pooled database of placebo-treated healthy volunteers. 244

7.3 Hypothetical situations in FIH studies and the corresponding probability of observing the same events under placebo. 245

7.4 Results of univariable analyses for each variable fit in a random effects logistic regression model with a random effect on study. Results are reported as odds ratios and the between study variance on the log odds scale. 254

7.5 LOO deviance table when deciding on the final model. 255

7.6 Final model coefficient estimates for the logistic model of the probability of a subject experiencing an event of ALT > ULN. 257

7.7 Mean and standard deviations of the variables used in the model 257

7.8 Characteristics of subjects who had an elevated ALT>ULN (N=77) compared to those who did not (N=1157) 260

7.9 Hypothetical cohort of active-treated subjects and individual model-based predictions for each individual's placebo twin 261

7.10 2 by 2 analysis 264

A.1 Differences between the topic guide for statisticians and clinicians. 315

A.2 Log of amendments to topic guide for statisticians. 323

A.3 COREQ checklist. 324

A.4 Characteristics of included trials. 332

A.5 Associations between reported study characteristics 341

1 Introduction

1.1 Rationale of thesis

> "I believe that the more you know about the past, the better you are prepared for the future."
>
> *Theodore Roosevelt*

Randomised controlled trials (RCTs) are considered the gold standard method to compare the relative efficacy of competing interventions and represent a significant investment of clinical research resources [1, 2]. Funders often highlight the importance of taking into account existing evidence when planning a new trial [3, 4, 5] in order to reduce research waste [6]. It is important that RCTs are designed and conducted in a way that allows the research question to be addressed in the most efficient and cost-effective manner, given what is already known from previous research [7]. It is fully ingrained into researchers, and in particular trialists, that existing evidence should be used in some way to inform new research [1, 8].

Existing evidence can be used to inform all stages of a clinical trial: planning of a trial before it begins, monitoring of a trial in progress, and analysis and reporting of the results of a new trial alongside other relevant research [7]. However, the amount to which existing evidence can be used to inform the design and analysis of trials can vary greatly [8, 9, 10].

There are 'traditional' uses of previous evidence, for example, in sample size calculations, to inform parameters about which we need to make assumptions. Similarly, in the early design stage, existing evidence can be collated and summarised in a systematic review, and subsequently used to inform whether a gap in the current evidence base justifies a new trial [11]. In 2007, Clarke *et al* [12] made a plea for reports of clinical trials to begin and end with an updated systematic review. However, it is still unclear precisely how trialists are using external evidence in the design and analysis of trials. Furthermore, the formal explicit use of external information, such as from other trials, is largely omitted from methodological guidelines.

A formal way of incorporating previous evidence is by use of informative prior distributions, in a Bayesian framework [13]. A Bayesian approach combines the new trial data with existing data, as represented by the prior distribution [14]. One advantage of this approach is it allows the analyst to make use of all relevant data and put their work in the context of existing research, and there have been numerous calls to do this [6, 12]. In a Bayesian framework, informative prior distributions can, in theory, be assigned to any parameter during the design or analysis stage of a trial when there is existing data available. However, there is still much controversy about the use of informative priors.

In this thesis, we focus on exploring the use of informative priors in trials. I use Bayesian methodology to synthesise external trial data and subsequently assess how these data can be used to inform parameters during the design and analysis stage of a trial, such as the effect size or bias. Therefore, the aim of this thesis is to extend, develop and assess the feasibility of methods to use external evidence, based on an existing body of data (such as a collection of studies), to inform the design and analysis of an individual trial. This thesis does not look at elicitation methods to form priors from expert opinion. Instead, we will include consideration of how best to form suitable priors based on existing evidence and how a trialist could use these priors in a clinical trial setting. An underlying theme throughout the thesis is whether trialists would be happy to implement these methods in practice, and a survey and extensive qualitative study is undertaken to evaluate this issue.

In this chapter we provide background about what a clinical trial is and how the results of a clinical trial can be used to inform medical practice and policy. Section 1.2 explains the different phases of a trial and how the majority of trials are analysed in a frequentist framework. We then describe the idea of analysing a trial in a Bayesian framework, to allow the incorporation of external evidence to be combined with trial data. Section 1.3 introduces different types of evidence syntheses, including definitions, and focusses on the role evidence synthesis has to inform policy. In Section 1.4, we return to the possible opportunities of using external evidence in trials, highlighting the areas on which this thesis focusses. Finally, a chapter synopsis outlining the structure of the thesis is given.

1.2 Clinical trials

In medical research, there are often competing interventions and the aim is to try and understand which intervention is better. An intervention can be a new drug, surgical technique, device, or treatment plan. There is usually a hypothesised ′superior′ intervention which we want to compare to the standard intervention (or placebo). We use the term "intervention" to refer to any intervention, placebo or other control. The best way to determine which intervention is better is to do a clinical trial [2]. In a standard parallel group trial these are allocated to groups, which we refer to as experimental and control groups.

1.2.1 Phases of a clinical trial

The earliest phase of a clinical trial of a pharmaceutical product begins by trying to identify pathways of how a molecule (which will potentially be the new drug) interacts, i.e. biological processes within cells [15]. This is very much at the heart of the transition from basic science to translational medicine. These pathways are explored by scientists in a laboratory who are looking for interactions between molecules and cells, that could have some clinical relevance. When there is an interaction, this translates to potential clinical relevance for a particular group of patients and generates a scientific hypothesis. In order to test that hypothesis, a trial is designed which requires patients to be assigned to different treatment or control arms, to separate the impact of a treatment or therapy. Notably, only a tiny number of these hypotheses go forward to be evaluated in an early phase trial. This phase is known as first-in-human (FIH) or phase I trials and usually occurs in a small number of healthy individuals. In some cases, this phase occurs in patients with the target disease, often oncology trials.

The main aim of phase I trials is to check tolerability and establish safe doses of a new drug, in order to move to phase II. Phase II trials often involve eligible patients with the disease of interest, where the main aim is to find the correct dose and to establish dose frequencies, administration routes, and endpoints, before moving to a phase III trial. A phase III trial will again include even more patients, who are assigned to different treatment or control arms, to enable the comparison of treatments and determine efficacy. If efficacy is confirmed, the trial will then progress to the final stage, known as phase IV. In phase IV, the intervention is rolled out to a broader population and any adverse events are monitored over a longer period of time. In this thesis, we will consider case studies of potential uses of external evidence syntheses to inform FIH studies and phase III trials.

Pilot and feasibility studies are often used to inform aspects of a phase III trial design, such as predicting recruitment rates [16]. They are also used to see if it is feasible for the main trial to go ahead [17], for example, whether the intervention can be delivered and implemented as planned [18]. The primary aim of feasibility studies is to establish if a new trial can be done: they do not analyse the outcome of interest. Instead, they are used to estimate parameters which are needed in the design of the main study. In contrast, pilot studies are closely related to what the main study plans to be but are smaller in size. The primary aim of pilot studies is to see how all the components of the main study work together. Pilot studies therefore emulate the main trial, but on a much smaller scale [19].

1.2.2 Approaches to clinical trials

One of the simplest designs of a phase III clinical trial, compares the group of participants who receive the experimental intervention to the other group of participants, who receive either the control intervention or a placebo. Each group of participants is then followed over (ideally) the same period of time. Information on a relevant outcome(s) is collected at either one or multiple times. An ideal study would be designed such that at the end of the study we could attribute any difference in the outcomes between the two groups, to the difference in interventions. For this to be the case (unbiased estimate of the intervention effect) we require the two groups to be similar in terms of all other characteristics (e.g. similar age groups, similar proportions of males and females etc). After receiving the intervention, we want the patients to behave in exactly the same way (except for any effect of the interventions) until we measure the difference between them. This is achievable through the following steps:

- Randomisation to ensure each group has participants with the same characteristics. If one group is systematically different to the other this can cause confounding.

- Adequate concealment of which intervention is next in the randomisation sequence. A failure to adequately conceal which treatment will be received by the participant may introduce confounding or selection bias [20]. It is important to conceal allocation from the people recruiting patients into the trial. This is called allocation concealment.

- The participant, the person(s) providing or delivering the intervention and the person(s) assessing the outcome should be blinded to the intervention delivered where possible. Sometimes an individual has more than one of these roles: for example, if the outcome is a questionnaire filled out by the patient, they will be both the participant and the person assessing the outcome. Failure to blind those delivering the intervention (i.e. healthcare providers) can result in performance bias and failure to blind the outcome assessor can result in detection bias [21].

- It is important that each participant has their outcome measured. Failure to do so could cause incomplete outcome data and therefore attrition bias.
- If only some outcomes were reported (relative to those presented in the protocol and planned to be reported) and selection depends on the results, this is known as reporting bias.

If any of these steps either cannot be implemented or implementation fails, then each can cause the results of a trial to become biased.

1.2.3 Analysis of a clinical trial

In statistical analysis, and thus when analysing a trial, there are two different underlying philosophies: frequentist or Bayesian. In general, either approach can be used for any data [22]. For example, if one type of analysis is in a frequentist framework, there will usually be an equivalent within the Bayesian framework [23]. Bayesian approaches will often identify themselves as such by use of the word Bayesian, whereas a frequentist approach is not usually explicitly stated, as it is considered most traditional and more widely accepted as the default framework.

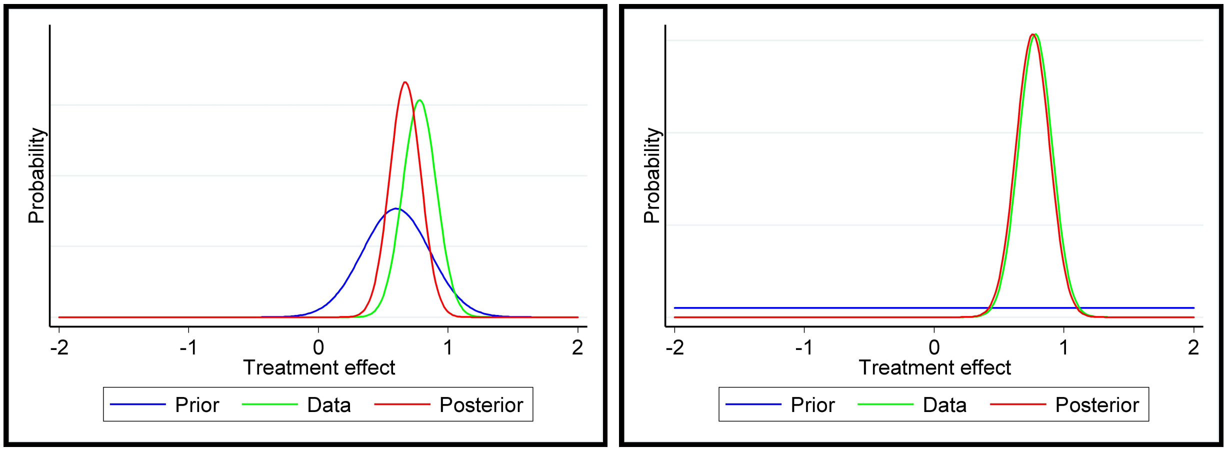
Suppose in a clinical trial we have a simple null hypothesis, that there is no difference between the two interventions and the alternative hypothesis is that there is a difference. The theory behind frequentist and Bayesian approaches to examining this hypothesis is fundamentally different and therefore determines how trial results are interpreted; mainly because of the way probability is viewed for each. A frequentist framework quantifies if a clinical trial was repeated over and over again, how often would these results have been observed if the null hypothesis was false. In contrast, a Bayesian approach asks a direct probability question: given the results observed in these data, what is the probability the null hypothesis is false [24]? The focus of a Bayesian analysis is on how the trial should change our belief about the treatment effect [14]. This forces us to state a reasonable belief concerning the plausibility of different values of the treatment effect before a trial has started as a ′prior distribution′. This has historically been considered controversial [14, 25]. Combined with the data for the trial, a revised belief about the treatment effect is produced (known as the posterior distribution).

Seemingly, the most discussed area in Bayesian analysis is the selection of an appropriate prior distribution [25, 26]. Priors can either be uninformative (intended to represent an absence of any prior knowledge about the value of the treatment effect) or informative. Informative priors may be further divided into: (i) those based on previous evidence in a specific clinical area; (ii) those based on expert opinion from knowledge of the clinical area; and (iii) default ‘sceptical' or ‘enthusiastic' priors, representing the idea that the treatment effect in general is relatively unlikely or likely [13]. In this thesis, we focus on the formulation and use of those prior distributions based on previous evidence in a specific clinical area.

Many authors have shown there are no issues or there should at least be no controversy if priors are uninformative [25, 27]. A pivotal study in the late 1980s looked at the practical use of adopting Bayesian methods in the pharmaceutical industry [28], compared to the classical or more widely known frequentist approach [29]. This study received much attention and dismissed many negative theories of using non-informative priors in Bayesian analysis of medical data. However, there is still much controversy about the use of informative priors [27]. Partly because of this, clinical trial data are usually analysed in isolation, without explicit reference to the wider evidence base and are therefore conducted in a frequentist framework [13]. However, there is often relevant external evidence available [6, 30]. For example, a meta-analysis of results of a set of similar previous trials could provide information on the likely size of the intervention effect [6]. A meta-analysis is a statistical method which combines intervention effects (or any other parameter) across multiple studies to summarise all available evidence about the intervention effect [31] (and is described in more detail in Section 1.3.2).

A meta-analysis which analytically summarises the existing studies can be used to formulate a sensible prior distribution of likely treatment effects. The process is represented by Figure 1.1a: the prior information (from the meta-analysis) and the data from the new trial are combined to form the posterior distribution. This is summarised by statistics such as the mean and 95% intervals. The less precise the external information, the more the combined (posterior) results will be driven by the new trial. For example, in Figure 1.1b the horizontal line represents a lack of any relevant prior information (i.e. an ′uninformative′ or ′vague′ prior), such that any value of the intervention effect is felt to be equally likely. In this case, the posterior distribution is driven entirely by the data. The results obtained are then the same as if we had followed a classical (non-Bayesian) approach.

The computation of the posterior distribution in Bayesian analyses are often carried out using Markov chain Monte Carlo (MCMC) simulation methods. One of the most common ways to do this is to use Gibbs Sampling, which generates samples from the posterior distribution. In this thesis, Bayesian software WinBUGS, version 1.4.3 (MRC Biostatistics Unit, Cambridge, UK) [32] and Stan (via RStan) are used. To fit these models, the user has to specify the statistical distribution (likelihood) for each piece of data, the prior distribution for the parameters, the function of parameter that each data source provides an estimate of, and how these relate to each other, i.e. the model. MCMC is equivalent to integrating over one or more density functions and is achieved by simulating random variables from known statistical distributions. A Markov Chain is a random process which has the specific property that the future depends only on the current state of the process



(a) Prior information (from the meta-analysis) and the data from the new trial are combined to form what is known as the 'posterior distribution'.

(b) Prior information (representing any value of the intervention effect being equally likely) and the data from the new trial are combined to form what is known as the 'posterior distribution'.

Figure 1.1: Prior and trial data regarding a treatment effect in a clinical trial combined to make a posterior distribution in two scenarios.

and not the past. Thus, initial values for the chain need to be specified. Once the chain has converged the posterior distribution of the parameter of interest is summarised using general summary statistics. The median and 2.5th and 97.5th percentiles are often used, especially when the posterior distribution is skewed. The 95% credible interval (CrI) is defined as the interval between the 2.5th and 97.5th percentiles.

Rather than incorporating prior information on the treatment effect, Bayesian approaches can also incorporate external evidence on other parameters, examples of these are given in Section 1.4.

1.3 Evidence synthesis methods

1.3.1 Introduction to evidence synthesis methods

The results of clinical trials offer one of the best sources of evidence to help determine whether an intervention is effective or not, clinically, economically, or both. However, the result of one clinical trial often has little impact. Instead, it often takes the synthesis of multiple trials to determine if an intervention should be adopted in practice [30].

Evidence synthesis is the collation and combination of multiple sources of evidence in order to answer a specific research question. This includes conducting a literature review (systematic or otherwise) and providing a narrative summary of the results. More formally, a systematic review may be undertaken to collate all evidence that fits pre-specified eligibility criteria in order to address a specific research question [33]. To provide a numerical summary of a body of evidence, a meta-analysis is often performed. This statistically combines results from two or more studies addressing the same research question. In the following section, the background and explicit formulae of a meta-analysis are given, as we make use of these in Chapters 4 to 6.

1.3.2 Meta-analyses

Most meta-analyses use either a fixed effect or a random effects statistical model which are based on differing assumptions [34, 35, 36]. They can also be performed in a frequentist or Bayesian framework. Since we refer to both in this thesis, both frameworks are described below.

Frequentist approaches

Fixed effect meta-analysis

A fixed effect meta-analysis is usually interpreted as assuming all studies are estimating the same (fixed) underlying intervention effect [37], i.e. differences amongst studies are due to chance or random error. The type of effect estimate will be related to the outcome variable of interest. For binary outcomes, usually odds ratios or risk ratios including its standard error) are synthesised. For these ratio measures, meta-analysis is performed on the log scale. For continuous outcomes, we usually seek a mean difference and its standard error from each study or the mean, standard deviation, and number of participants in each group [38]. These effect estimates are usually extracted from articles [39].

Suppose we have studies $i = 1, 2, ..., n$. We define Y_i to be the effect estimate in study i and s_i^2 its variance (assumed known). For each of the studies, we apply a weight w_i to the estimate Y_i in our meta-analysis, calculated by:

$$\mathrm{w}_i = \frac{1}{s_i^2} \tag{1.1}$$

One of the most common methods in the frequentist framework is the inverse-variance method. The pooled estimate from the inverse-variance (fixed-effect) method can be found using the formula for the maximum likelihood estimate for θ:

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

The variance of the pooled effect size, $\hat{\theta}$, is given by:

$$V(\hat{\theta}) = \frac{1}{\sum_{i=1}^{\mathrm{n}} s_i^2} \tag{1.3}$$

The maximum likelihood estimate is used because it is asymptotically unbiased, efficient, and normally distributed. If we assume that $\hat{\theta}$ is approximately normally distributed, then an approximate 95% confidence interval for the pooled estimate can be obtained from the formula $\hat{\theta} \pm (1.96 \times \sqrt{V(\hat{\theta})})$.

Random Effects Meta-Analysis

There is debate about combining studies in to a single summary estimate through a fixed effect meta-analysis when some studies could be very context specific [40]. There is a strong assumption in a fixed effect meta-analysis that particular characteristics across studies will not influence the size of the treatment effect (such as the conduct of the trial) [41, 42, 43]. All meta-analyses are susceptible to heterogeneity. This is likely due to differences in trial conduct, which vary across studies and as such will likely impact upon the underlying treatment effect across studies [44]. When there is more variation in the treatment effect estimates observed across studies than is expected by chance alone, this is defined as statistical heterogeneity [45]. A random effects meta-analysis allows for this heterogeneity. The I^2 statistic can be used to describe the percentage of the total variation across studies that is estimated to be due to between-study heterogeneity [41, 44].

A random effects meta-analysis assumes the underlying treatment effect varies across studies with mean θ and between-study variance, denoted by τ^2. We now account for both the within-study variation *and* between-study variation when estimating the new pooled effect $\hat{\theta}$. To compute $\hat{\theta}$ we must first estimate the between-study variance $\hat{\tau}^2$. This is often estimated by the DerSimonian and Laird (1986) method of moments [31] where the weights are from the fixed effect meta-analysis. If $\hat{\tau}^2$ is > 0 this indicates that heterogeneity exists. We now calculate $\hat{\theta}$, using updated weights:

$$\hat{\theta} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{Y}_i \mathrm{w}_i^*}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{w}_i^*} \tag{1.4}$$

where

$$\mathrm{w}_i^* = \frac{1}{s_i^2 + \hat{\tau}^2} \tag{1.5}$$

The variance of the pooled effect size, $\hat{\theta}$, is given by:

$$V(\hat{\theta}) = \frac{1}{\sum_{i=1}^{\mathrm{n}} s_i^2 + \hat{\tau}^2} \tag{1.6}$$

In contrast to a fixed effect meta-analysis, a random effects meta-analysis assumes a distribution of intervention effects across studies [41]. The pooled estimate is then interpreted as the average intervention effect across studies. Therefore, in a random effects meta-analysis individual studies could have an intervention effect that varies considerably away from the average value.

In addition to summarising the average intervention effect of the studies in a random effects meta-analysis, it is also possible to derive a predictive distribution [42]. The predictive distribution summarises the true intervention effect in a new study that is like those already in the meta-analysis, allowing for the additional variability between the studies [46]. An approximate 95% prediction interval for the underlying effect is $\hat{\theta} \pm (1.96\sqrt{\hat{\tau}^2 + V(\hat{\theta})})$.

Bayesian approaches

The majority of this thesis is conducted within a Bayesian framework and in Chapters 4 and 5, we make specific reference to Bayesian meta-analysis models for binary data. In Chapter 6, we also use the Bayesian model for normal data. These are briefly introduced below.

Bayesian random effects meta-analysis model for normal data

The estimates Y_i come from the following random effects distributions:

$$\mathrm{Y}_i|\theta_i \sim N(\delta_i, {s_i}^2) \tag{1.7}$$

The unknown random effects δ_i are assumed to come from a common distribution:

$$\delta_i|\theta, \tau \sim N(\theta, \tau^2) \tag{1.8}$$

This model is in fact the basis for the frequentist random effects inverse variance approach (equation (1.4)). In a fixed effect model, δ_i would be assumed to be the same for all studies and therefore equal to θ. In a Bayesian framework, the parameters, θ and τ which are estimated by the model and assumed unknown, need to be given prior distributions. These can be either non-informative or informative. The random effects mean, θ, is usually given a very vague prior such as $\theta \sim N(0, 10^5)$. One of several options for a vague prior for the between study SD is $\tau \sim Unif(0, 5)$. However, informative empirical priors for the between study variance have been suggested by Turner *et al* [47].

Most meta-analyses are based on normal approximations to the distribution of Y_i. However, in a Bayesian framework it is straightforward to model data exactly rather than using approximations.

Bayesian random effects meta-analysis model for binary data

For binary data Smith *et al* [48] suggest the following model:

$$r_{t,i} \sim Binomial(p_{t,i}, n_{t,i}) \tag{1.9}$$

$$r_{c,i} \sim Binomial(p_{c,i}, n_{c,i}) \tag{1.10}$$

$$logit(p_{c,i}) = \mu_i \tag{1.11}$$

$$logit(p_{t,i}) = \mu_i + \delta_i \tag{1.12}$$

$$\delta_i \sim N(\theta, \tau^2) \tag{1.13}$$

where $r_{t,i}$, $r_{c,i}$, $n_{t,i}$, $n_{c,i}$ are the number of events and patients in the treatment and control arm, respectively. In the ith study in the treatment arm, the probability of an event is $p_{t,i}$ and in the control arm is $p_{c,i}$. The $logit$ of the event probability in the control arm is μ_i. As with the normal random effects model, the treatment effect in study i is δ_i, the mean value of all treatment effects across all studies is θ, and τ^2 is the heterogeneity (variance) between the study estimates. The additional parameters μ_i are usually given $N(0, 10^5)$ prior distributions.

1.3.3 Other types of evidence synthesis methods

An extension of *any* of the above meta-analysis models in the context of healthcare interventions is a network meta-analysis (NMA), which allows the simultaneous comparison of the effectiveness of multiple interventions through the use of direct and indirect evidence [49, 50]. An economic decision model can be used to evaluate intervention effects formally in the context of other factors such as costs and potential harms and make decisions on use of interventions in practice [51]. Sometimes, a value of information (VoI) analysis is used to assess whether there is value in conducting a new study, and to identify the optimal design for such a study within an analytical modelling framework (based on a decision model) [52, 53]. Expected value of sample information (EVSI) is a specific type of VoI analysis used to assess the ability of a new trial to inform a cost-effectiveness assessment of the intervention to determine the optimal sample size [54].

Large-scale empirical ("meta-epidemiological") studies of randomised trials allow us to learn about biases, by comparing results of trials suffering from specific limitations with results of trials that are free of these limitations. Bayesian hierarchical models for meta-epidemiological research developed by Welton *et al* [55] estimate the average amount of bias expected in high risk of bias (RoB) studies, the average variability in this bias within meta-analyses, and variability in average bias between meta-analyses. In future chapters, we analyse and develop meta-epidemiological models.

1.4 Current uses of external evidence in trials

Evidence synthesis methods play an important role in identifying and summarising information in the current evidence base. There are different levels of external data which may be synthesised, for example, individual participant data (IPD) from trials, whilst summary level data may be combined in a meta-analysis. In this thesis, we explore how a trialist can use information from such syntheses to inform the design and analysis of their clinical trial, with particular reference to the impact on the current evidence base. Building upon the overview of how external evidence is used to inform trials in Section 1.1, we now describe this in more detail. A summary of the application of evidence synthesis methods, using different levels of evidence, is provided in Table 1.1, with consideration of each stage of a trial.

Before design

Before the design stage of a trial, funders, such as the National Institute for Health Research (NIHR), often request a systematic review and, where possible, a meta-analysis, *to justify a gap in the evidence base* [5, 56]. If there are no relevant previous trials, a search strategy might be requested by funders to support this. Relevant systematic reviews might include existing clinical trials, early phase trials, non-randomised comparisons, animal studies, or qualitative research studies [57]. If a meta-analysis exists, this will provide information on effect sizes from 'similar' trials. Burke *et al* show how an existing meta-analysis of similar phase II trials can be used to inform whether a phase III trial should be conducted [58]. This is conducted within a Bayesian framework and the authors strongly suggest documenting sensitivity to the choice of meta-analysis model.

Design

In the design stage, a systematic review or meta-analysis may highlight the *population* and particular subgroups that warrant further investigation [59]. For example, the population in the Lung ART trial [59] was defined by subgroup results from an IPD meta-analysis of postoperative radiotherapy vs none. Similarly, results from evidence syntheses, including network meta-analyses, decision models, and VoI analyses, can be used to choose which *interventions and comparators* to trial [60] and characteristics of these, e.g. dose or duration of treatment [11, 59]. For example, an IPD meta-analysis comparing prophylactic cranial irradiation vs. none in patients with small-cell lung cancer informed the choice of comparators in the trial [59].

A systematic review can help inform the *choice of outcomes* [11, 57] in a new trial and how they should be defined and, if relevant, the *duration of follow-up* [59]. For example, a systematic review may highlight adverse events that should be monitored, in particular events that are expected and related. A systematic review and/or meta-analysis may also provide information on the parameters needed for *sample size calculations* [11, 59], such as the standard deviation, control group outcome rates, plausible effect sizes, loss to follow-up, and correlation coefficients [57]. For example, an IPD meta-analysis comparing chemotherapy vs standard radiotherapy informed the control group survival rate used as the basis for the sample size calculation in the OUTBACK trial [59]. Alternatively, EVSI calculations can be used to assess the ability of a new trial to inform cost-effectiveness assessment of the intervention to determine sample size and reduce decision uncertainty

[54].

Information from meta-analyses can be used to inform parameters within a trial. External evidence might be used to improve the estimation of *secondary or 'nuisance' parameters* involved in trial analysis which are often poorly estimated. An example of a secondary or nuisance parameter is the intra-class correlation coefficient (ICC) in cluster randomised trials [61] and between-centre variability in multi-centre trials. When calculating the sample size for cluster randomised trials, an estimate of the variability between clusters, the ICC, is required. It was common practice to look back at previous studies to informally estimate the ICC value, however, sample size calculations can be sensitive to the choice of this value. Turner *et al* suggest a Bayesian approach to allow for the uncertainty around pooled ICC values from multiple studies. The authors describe how priors for the ICC, based on a meta-analysis of existing studies, can be used to inform sample size calculations for cluster randomised trials.

Monitoring (conduct)

During the monitoring stage of the trial, external evidence on *recruitment rates* can be used to inform rates in new hospitals or centres. Observed *adverse event rates* can be compared with predictions from a synthesis of historic data to see if they are higher than expected by chance [62]. For example, the incidence of cancer was calculated in an interim analysis in patients taking ezetimibe and this was then compared to the expected incidence from two larger ongoing trials [62]. The incidence was not higher than expected and so the trial continued. Emerging trial results considered in the context of results from previous studies might be used to make the *decision to stop a trial early* [59]. For example, subgroup results from a meta-analysis of IPD (of postoperative platinum-based chemotherapy vs. none for non-small-cell lung carcinoma (NSCLC)) were used to stop the trial and cancel the planned phase III component due to clear discrepancies [59].

Analysis and reporting

Factors such as measures of effects (event rates, mean difference etc.) from previous trials might influence the *choice of statistical model*. Prognostic or predictive factors identified through evidence synthesis may be used to stratify or adjust trial analyses [59]. For example, an IPD meta-analysis validated predictive risk factors which were then used to inform the basic statistical analysis of the new trial by adjusting for them in the model [59]. Choice of the most important covariates to be recorded for imputation modelling might be informed by patterns of missing data in previous trials.

A well designed RCT is considered the highest level of evidence and often referred to as the gold standard study design to compare treatments [2]. However, RCTs are susceptible to bias throughout [21, 63]. External evidence about *typical biases* associated with undesirable study characteristics, e.g. inadequate blinding, might come from 'meta-epidemiological' studies [55, 64, 65] or from opinions elicited from experts [66].

Individual patient data from previous trials can inform patient level parameters. In an individual trial setting, Pocock *et al* make use of previous information on historical control outcomes, not only in the design stage to randomise fewer patients but also in the analysis stage [67]. An informative prior is used on the mean outcome in the control arm based on a weighted average of the means of the new and historical controls. This can be advantageous as it more precisely estimates the treatment effect comparisons based on both types of control.

An *updated* systematic review [12] or meta-analysis including the new trial results [46] should be reported to put the results in the context of the wider evidence base [59]. For example, EORTC 62931 trial report uses results of meta-analysis and subsequent trials to place trial results in the context of other results from similar studies [59]. An existing meta-analysis might be used to form a prior distribution for the *treatment effect* in a new study.

This can then be updated using the trial data in a Bayesian statistical analysis.

Table 1.1: Summary of opportunities for evidence synthesis to inform design, conduct and analysis of a clinical trial.

Stages of a clinical trial	**Opportunities in which previous evidence might be used**
Before design	**To justify the need for a new trial in light of the existing evidence base.** A systematic literature review and, where appropriate, quantitative synthesis, could be used to assess the need for the new trial [5, 9, 56].
Design	**Choice of population.** A systematic review may highlight the population and particular subgroups that warrant further investigation [59]. **Choice of interventions and comparators.** Results from evidence syntheses, including NMA, decision models, and VoI analyses can be used to choose which interventions and comparators to trial [60] and characteristics of these, e.g. dose or duration of treatment [11, 59]. **Choice of outcomes and length of follow up.** A systematic review may help inform the choice of outcomes [11, 57] in a new trial and how they should be defined and, if relevant, the duration of follow-up [59]. **Sample size calculations.** A systematic review and/or meta-analysis may provide information on the parameters needed for sample size calculations [11, 57, 59]. Alternatively, EVSI calculations can be used to assess the ability of a new trial to inform cost-effectiveness assessment of the intervention and reduce decision uncertainty [54]. **To inform secondary parameters.**

Continued on next page

Table 1.1 – continued from previous page

Stages of a clinical trial	**Opportunities in which previous evidence might be used**
	External evidence might be used to improve the estimation of 'nuisance' parameters involved in trial analysis which are often poorly estimated, such as the intra-class correlation coefficient in cluster randomised trials [61] and between-centre variability in multi-centre trials.
Monitoring (conduct)	**Recruitment and consent.** For example, good or poor recruitment rates in previous relevant trials can inform site selection in a new multi-centre trial [11]. **To deal with adverse events.** Observed adverse event rates can be compared with predictions from a synthesis of historic data to see if they are higher than expected by chance [62]. **To decide whether to stop an ongoing trial.** Emerging trial results considered in the context of results from previous studies might be used to make the decision to stop a trial early [59].
Analysis	**To inform the statistical analysis plan.** Factors such as measures of effects (event rates, mean difference etc.) from previous trials might influence the choice of statistical model. Prognostic or predictive factors identified though evidence synthesis may be used to stratify or adjust trial analyses [59]. **To adjust for potential biases.** External evidence about typical biases associated with undesirable study characteristics, e.g. inadequate blinding, might come from 'meta-epidemiological' studies [64], allowing the analyst to assess the sensitivity of the findings to alternative model assumptions. **To assess the trial treatment effect in the context of existing evidence.**

Continued on next page

Table 1.1 – continued from previous page

Stages of a clinical trial	Opportunities in which previous evidence might be used
	An existing meta-analysis might be used to form a prior distribution for the treatment effect in a new study, which can then be updated using the trial data in a Bayesian statistical analysis.
Reporting	**To report the new trial results in the context of the wider evidence base.** An updated systematic review [12] or meta-analysis including the new trial results [46] should be reported to put the results in the context of the wider evidence base [59].

Most of these uses do not involve informative priors, which is what we focus on in this thesis. Despite these suggestions and examples in the methodological literature, informative priors are not routinely used within the clinical trials community and instead only applied in specific areas [14, 68, 69, 70]. Spiegelhalter [13] describe how informative priors are rarely used even within Bayesian analyses. Possible barriers to use of Bayesian methods across clinical trials include concerns about misspecification and subjectivity of the prior distribution [24] and the perspective of wanting to look at data of the clinical trial in isolation [71]. Deaton *et al* allude to this and continue the "lumpers and splitters" debate about meta-analysis, illustrating lots of issues about combining studies and firmly define everything as heterogeneous and context specific [40]. Davey-Smith and Egger discuss this issue at great length and construct arguments from both sides [72], arguing a potential reason why results of clinical trials and meta-analyses, and thus the proper account of previous evidence, are under used in clinical trials. However, it is less clear why informative priors are not used to (i) improve the estimation of key parameters and/or (ii) put trial results in the context of the existing evidence base.

In this thesis we consider three case studies of potential uses of external evidence syntheses in the design or analysis of a clinical trial, each relating to a different level of a hierarchy of sources of external evidence.

- **(i) External evidence on *bias* from meta-analyses within meta-epidemiological studies.** First, when there are potential unavoidable limitations in the methodology of the new trial, the analyst could attempt to account for these via incorporation of an informative prior distribution for the likely amount of bias. An example of an unavoidable limitation is the allocated treatment being unblinded to the patient or personnel, which can cause bias in the treatment effect estimate. The Bayesian analysis would then provide a treatment effect estimate from the new study that has been adjusted for the likely bias, allowing the analyst to assess the sensitivity of the findings.

- **(ii) External evidence on *effect sizes* from trials within meta-analyses.** Second, there are methods in the literature [13, 73, 74] which enable trialists to use information from an existing meta-analysis to inform sample size calculations, but it is unclear which of these (if any) to use in practice. We compare methods, when the focus of inference is the new trial or the updated random effects mean to see exactly what size a trial needs to be in order to have an impact on the current evidence base. This can potentially reduce the sample size needed for a new trial and give an indication to trialists of whether their new trial will have any impact to update an existing meta-analysis and therefore to change future policy. We also compare our results when using EVSI, based on the ability of the new trial to change the decision based on a cost-effectiveness model.

- **(iii) External evidence on *outcomes* from patients within trials.** Third, a synthesis of ′similar′ studies might be used to inform other parameters in the analysis which may be poorly estimated, such as when an analyst compares adverse events between two interventions. Therefore, external evidence synthesis could be used to inform the control group event rate when the event is rare and adverse outcomes are underpowered.

Qualitative work can explore people's views, attitudes, and previous experiences to allow an in-depth understanding on topics which may be multi-faceted [75, 76, 77]. In this thesis, we perform a questionnaire and qualitative study to explore current attitudes and experiences of trialists using informative priors, in the form of real trial data, to inform the design and analysis stage of a trial.

1.4.1 Aims of the thesis

There are several key aims to this thesis:

- To explore the current use of evidence synthesis in trials and the potential barriers to such use through a questionnaire and qualitative study.
- To extend and develop Bayesian methodology to synthesise relevant external data for each of the three case studies.
- To assess how this data can be used to inform the design and analysis stages of a trial. This will include considering how best to form suitable priors based on existing evidence and how a trialist would use these priors in a clinical trial setting.

1.4.2 Structure for the thesis

In Chapter 2 we explore the current use of and opinions about use of evidence synthesis methods in trial design and analysis. A survey was undertaken at the International Clinical Trials Methodology Conference (ICTMC) 2015 [78]. A subset of questions focused on the use of previous evidence in the analysis stage of a trial and specifically identifies the views on use of informative prior distributions in a Bayesian statistical framework, for each of: (i) the treatment effect, (ii) potential biases and (iii) secondary parameters. We also explore the potential barriers to the use of external evidence. The findings of the 'INVEST' (INVestigating the use of Evidence Synthesis in the design of clinical Trials) survey inform the rationale for the remainder of the thesis.

Chapter 3 describes a qualitative study that builds upon the findings of the INVEST survey. The aim of the study is to explore in more detail trialists' perspectives and experiences of analysing clinical trials in the context of external evidence, conducted through in-depth semi-structured interviews. The study targets key people within the clinical trials community, such as methods leads, those writing grant applications, and chief investigators. In the interviews we elicited views on the three case studies. We also explore the barriers to the use of informative priors across these settings and contrast the results to our INVEST survey.

In Chapters 4 and 5 we consider our first case study of (i), using external evidence on the likely amount of *bias* from meta-analyses within meta-epidemiological studies. Previous meta-epidemiological studies of blinding have generally compared intervention effect estimates from trials described as "double-blind" with those from trials not described as "double-blind". However, "double blind" is an ambiguous term as it is unclear precisely which parties were blinded [79]. In Chapter 4 we describe the first meta-epidemiological study to disentangle the impact of different types of blinding to enable the clear separation of the two main types of blinding-related bias: "performance bias" (systematic differences in the care provided to the participants in the comparison groups other than the intervention under investigation, which can be addressed by blinding the participants and care providers), and "detection bias" (systematic differences between comparison groups in how outcomes are ascertained, addressed by blinding outcome assessors), first described in 1.2.2. We examine, separately, the impact of blinding participants, care providers, and outcome assessors, and explore how this might vary by type of outcome. This chapter also extends current methodology to develop models which: combine binary and continuous outcomes; and stratify the average bias.

Although it seems reasonable to adjust for trials at a high RoB in meta-analyses, the results of such meta-epidemiological studies are still rarely used in this way [80]. Furthermore, meta-epidemiological studies have never, to our knowledge, been used, to inform bias adjustment within an individual trial setting. Instead, limitations of the trial, such as non-blinding, are often written descriptively in an end of study report or publication. There is usually no quantification about the likely amount of bias or how it may affect the interpretability of the findings. Therefore, in this thesis, we address some of the limitations of previous meta-epidemiological models [55] and extend them to try and improve their potential for implementation in practice. Chapter 5 explores statistical considerations of modelling meta-epidemiological data [55] with an application of these methods using published data from a recent meta-epidemiological study [81]. We extend current methodology to model associations between treatment effect estimates and categorical and numerical study characteristics. This includes jointly modelling unclear and high RoB trials in relation to the common reference group of low RoB trials; and modelling continuous bias predictors. Finally, we estimate the probability an unclear RoB trial is at a high RoB [82]. We also discuss the use of meta-epidemiological evidence to adjust for a potentially biased treatment effect in an individual trial.

It can be assumed any previous trials of the same intervention should inform the decision as to whether to conduct a further trial [3, 10]. However, it is less clear how the results of such a synthesis should be used by trialists, when designing their trial and by funders, when deciding whether to invest in the proposed new trial. Chapter 6 explores the use of existing evidence to inform the design stage of a trial, more precisely, the use of an existing meta-analysis to inform sample size calculations [73, 74]. An example looking at the effect of cortisterioids after traumatic brain injury is used for illustration [83, 84]. This chapter compares methods for explicitly incorporating information on the intervention effect from a previous meta-analysis for use in sample size calculation. We describe and compare the following methods: (1) standard power calculations; (2) calculations based on the power of a Bayesian analysis of the new trial with an informative prior distribution based on a meta-analysis [13]; (3) calculations based on the power of an updated meta-analysis [73, 74] and (4) EVSI calculations, based on the ability of the new trial to change the decision in a cost-effectiveness model. We discuss the advantages and disadvantages of each for their use in practice.

Chapter 7 presents an application of Bayesian mixed modelling to a collection of placebo arm data in FIH studies to estimate the incidence of safety events. In this chapter we investigate how to use this existing data to make predictions and inform the analysis of FIH studies [85]. The results are used to assess whether a safety event observed of the investigational drug is likely to be due to chance or caused by the compound under investigation. An example of placebo data from seventy-seven FIH studies is used for illustration.

Finally, Chapter 8 summarises the key findings from the thesis and discusses areas for further research.

2 Current practice and attitudes: the INVEST survey

The work presented in this chapter was performed in collaboration with the MRC Evidence Synthesis Working Group which included Isabelle Smith, Julian Higgins, Borislava Mihaylova, Benjamin Thorpe, Robert Cicero, Kusal Lokuge, Julia Forman, Jayne Tierney, Ian White, Linda Sharples, and Hayley Jones.

2.1 Introduction and aims

One of the overarching aims of this thesis is to develop methodological approaches that use existing evidence syntheses to inform the design and analysis stage of an individual trial. Despite the promotion of reducing research waste [6], it is still currently unclear *whether* and *how* trialists are using existing evidence syntheses to inform the design and analysis of a trial in practice.

In a survey of 24 investigators whose trials were included in an update of a Cochrane review, only 8 (33%) indicated that a previous review had influenced trial design, and only 2 (8%) had used the previous Cochrane review [10]. More recently, reviews of trials funded by the NIHR Health Technology Assessment (HTA) programme found that the majority (77% of those funded between 2006 and 2008 [11], and 100% of those funded in 2013 [57]) referenced a systematic review in the funding application. When a systematic review was not referenced, there were valid reasons for this, such as there being no relevant systematic review addressing the proposed research question [57]. Arguably of more interest is whether and how a cited review was used to inform trial design. Jones *et al* [11] found only 54% (20/37) of trials that referenced a systematic review used the review in some way. The recent review of Bhurke *et al* [57] found that 94% (32/34) of the trials examined used the referenced systematic review to justify the treatment comparison in the new trial, but that other uses were relatively infrequent. The other most common uses were in selection of a definition or outcome (16%), to inform the standard deviation (9%) or to inform duration of follow up (6%). Tierney *et al* describe examples of how meta-analyses of IPD have informed trial design, conduct and analysis in practice [59].

To our knowledge, there are no recent studies investigating the extent of use of evidence synthesis in the design of trials funded through streams other than the NIHR HTA programme or in trial *analyses*. Therefore, the aim of this chapter is to explore the current use of and opinions about use of evidence synthesis methods in trial design and analysis. To do this, a survey 'INVEST' was undertaken at the ICTMC, 2015. The purpose of this survey was to (i) summarise current evidence synthesis use in trial design and analysis across clinical trials teams, (ii) capture current opinions of trialists and methodologists on such use, and (iii) understand any barriers to use in practice.

The survey will therefore provide a snapshot of the current views of trialists of how and what evidence synthesis methods are being used to inform both trial design and analysis, and also help determine appropriate areas for the consideration of such methods in the remaining chapters of this thesis.

2.1.1 Aims

The following aims are:

Aim 1: Evidence synthesis methods in trial design

- To summarise respondent's views and opinions on the use of evidence synthesis methods to inform trial *design*.

This will identify if respondents are using evidence synthesis methods to justify or inform (i) whether a trial is needed; (ii) the choice of population; (iii) the choice of interventions;

(iv) the choice of outcomes and follow-up time; (v) sample size calculations. We also explore how these methods are implemented, whether they are conducted by the trials team or based on previously published evidence syntheses, such as an existing systematic review and/or meta-analysis.

Aim 2: Evidence synthesis methods in trial analysis

- To summarise respondent's views and opinions on the use of evidence synthesis methods to inform trial *analysis* in the following three areas:
    - Treatment effect, such as using information from a previous meta-analysis;
    - Potential biases arising from trial conduct, for example, evidence on the likely amount of bias from methodological limitations;
    - Other quantities that need to be estimated in the analysis such as correlations and baseline event rates.

Aim 3: Barriers to the use of evidence synthesis methods in trial design and analysis

- To understand the barriers to the use of evidence synthesis methods in *both* trial design and analysis.

This will identify the most important barriers to such use and help identify areas for future research.

A supplementary aim is to characterise the sampling frame:

- To summarise the characteristics of delegates who attended the conference and those who responded;
- To infer potential differences between respondents, whether there are differences between groups of trialists and their background or if previous experience can explain

this.

2.2 Methods

The sampling frame consisted of all delegates at the two-day ICTMC on 16-17th November 2015. The conference was open to both those involved and those who have an interest in clinical trials methodology. Approximately 638 people registered to attend the conference across a range of disciplines including trialists, clinicians, statisticians, health economists, information specialists and qualitative researchers. 95% of the registered delegates were from the UK and the Republic of Ireland, with the remaining 5% from Australia, Canada, Denmark, France, Germany, Holland and the United States. The main UK research centres represented were Aberdeen, Birmingham, Bristol, Cambridge, Cardiff, Coventry, Glasgow, Leeds, London, Liverpool, Manchester, Oxford and Southampton. Conference delegates were first invited to take part in the survey during the opening plenary session, then by researchers from the INVEST team during breaks. The survey could be completed either on paper or online, with a closing date of 18th December 2015. The survey is available in full in the Appendix, Figure A.1.

The survey will be presented in two parts: first, descriptively summarising all participants in the survey and second, describing only the subset of participants involved in trial design and trial analysis. All participants will be characterised using their job role, setting and which aspects of clinical trials they are involved in.

2.2.1 All participants

Following details about their job role, job setting and the length of time they had spent working in clinical trials, respondents who indicated that they had been involved in trial design (and/or analysis) were further asked questions about whether and how they have used evidence synthesis (i.e. in practice). All respondents were then asked about their views on the use of evidence synthesis in trial design and analysis. They were also asked to rank what they considered to be the three greatest barriers to such use. There were eight potential barriers listed, and an 'other' category allowing free text. The subsets of respondents who indicated that they had been involved in trial design (and/or analysis) were used to contrast views on whether evidence synthesis methods should be used versus current use in practice.

2.2.2 Use of evidence synthesis to inform trial design

Respondents who indicated they had personally been involved in trial design were asked to consider any trials in which they had been involved over the last 10 years and to specify, if applicable, how evidence synthesis had been used in practice. A matrix style layout was chosen to allow multiple responses, with rows for each area of trial design and columns for types of evidence synthesis method. In addition to (i) a description of previous evidence, (ii) a systematic review and (iii) a meta-analysis, we listed three evidence synthesis methods that extend meta-analysis: (iv) NMA; (v) an economic decision model; (vi) a VoI analysis [86]. A final option of 'none of these methods' was included. Respondents were provided with a brief definition of these evidence synthesis methods to reduce ambiguity. The areas of trial design listed were: (i) whether a trial is needed; (ii) the choice of population; (iii) the choice of interventions; (iv) the choice of outcomes and follow-up time; (v) sample size calculations. Respondents were also asked to indicate whether any evidence synthesis used had been performed by the trial team or previously published by others.

We also asked all respondents which of the listed evidence synthesis methods they thought *should* be used to inform aspects of trial design. This question was formatted to match the earlier question about how those involved in trial design were using evidence synthesis methods, facilitating comparison between ideal and current practices.

2.2.3 Use of evidence synthesis to inform trial analysis

Respondents who indicated they had personally been involved in trial analysis were asked which (if any) of three types of external evidence they had used in practice, during the last 10 years: (i) external information about the treatment effect (including a meta-analysis); (ii) evidence around the likely size of potential biases arising from trial conduct (e.g. blinding infeasible); and (iii) other quantities involved in the analysis (e.g. correlations or baseline event rates).

We asked all survey respondents whether each of these three types of external evidence *should* be used to inform trial analysis. For each of these, the options were 'yes', 'no', and 'don't know'. An additional 'don't understand' response was also included since we anticipated that some of these uses of evidence synthesis might be new concepts to some respondents [87, 88].

2.2.4 Analysis of survey responses

Our main analysis is descriptive, as sample sizes were not sufficient for a robust assessment of associations or subgroup comparisons. Missing responses were excluded from denominators and are indicated in footnotes in the tables that follow.

For the subsets of respondents involved in trial design or analysis, we compared their responses for *desirability* versus *actual use* of evidence synthesis. For each of the five aspects of trial design, we categorised each respondent who indicated they had been involved in trial design into one of the following: 'used and think desirable', 'used but don't think desirable', 'not used and don't think desirable' and 'not used but think desirable'. For each of the three aspects of trial analysis, we added three categories to these options: 'used and don't know whether desirable', 'not used and don't know whether desirable', and 'don't understand'.

To summarise responses about the three greatest barriers to use of evidence synthesis, we assigned three points to the first (greatest perceived) barrier, two to the second and one to the third for each respondent. If a respondent had ticked three barriers but not indicated a ranking, each was assigned two points. No points were allocated for respondents who did not answer the question. For each potential barrier, the scores were then summed across respondents, so that higher overall scores indicated greater perceived barriers.

Although highly exploratory in nature because of small numbers, we examined answers to specific questions for two subgroups: the perceived barriers to use of evidence synthesis in practice by statisticians specifically, statisticians' use versus perceived desirability of using evidence synthesis in trial analysis, and the views of health economists on VoI analyses.

2.3 Results

There were 106 respondents, of whom 54 (51%) were statisticians, 8 (8%) were health economists and 18 (17%) worked in trial management. These are overlapping categories, i.e. respondents were asked to select all roles that applied to them. All respondents had spent some time working in the area of trials: 86 (81%) for at least 3 years and 32 (30%) for more than 10 years. 96 (91%) respondents indicated that they had been involved in the design, setting up or running of trials (77 (80%) in a clinical trials unit, and 9 (9%) in industry). 85 (80%) indicated that they had been involved in trial design, 71 (67%) in trial conduct, 73 (69%) in statistical analysis, and 52 (49%) had been involved in undertaking a systematic review of trials. Only 3 (3%) respondents indicated that they had not been involved in any of these. Full details are shown in Table 2.1.

Table 2.1: Respondent characteristics

N=106	**n**	**%**
Job/role[1]		
Clinician	12	11.3
Clinical co-ordinator	1	0.9
Data management	5	4.7
Epidemiologist	5	4.7
Health economist	8	7.5
Information specialist	2	1.9
Programmer	1	0.9
Qualitative researcher	10	9.4
Statistician	54	50.9
Student[2]:	10	9.4
Clinician	1/10	
Epidemiologist	1/10	
Health services	1/10	
Statistician	4/10	
Statistician and data manager	1/10	
Unspecified	2/10	
Trial management	18	17.0
Other:	15	14.2
Academic researcher	4/15	
Chief/principal investigators	4/15	
Director of trials unit/CEO	2/15	

Continued on next page

[1] This question was 'tick all that apply' so respondents could have selected more than 1 and therefore the percentages do not add up to 100%

[2] Student disciplines will have already been counted if they ticked one of the available options

Table 2.1 – continued from previous page

N=106	**n**	**%**
Network coordinator	1/15	
Research Funder	1/15	
Systematic reviewer	1/15	
Trial methodologist	3/15	
Involved in design, setting up or running trials in your job/role		
None at all	10	9.4
Clinical trials unit only	59	55.7
Industry only	1	0.9
Clinical trials unit and Industry	6	5.7
Clinical trials unit and other setting:	12	11.3
CTU & Academia	10/12	
CTU & NHS/Hospital	2/12	
Industry and other setting:	2	1.9
MRC unit	2/2	
Other[3]:	16	15.1
Academia	9/16	
NHS/Hospital	4/16	
Academia/Hospital	3/16	
Involved in a trials unit at some point	77	72.6
Time spent working in the area of clinical trials		
Not at all	0	0
0-2 years	20	18.9
3-5 years	30	28.3
6-10 years	14	13.2
11-20 years	26	24.5

Continued on next page

[3] Could also have ticked any of the available options

Table 2.1 – continued from previous page

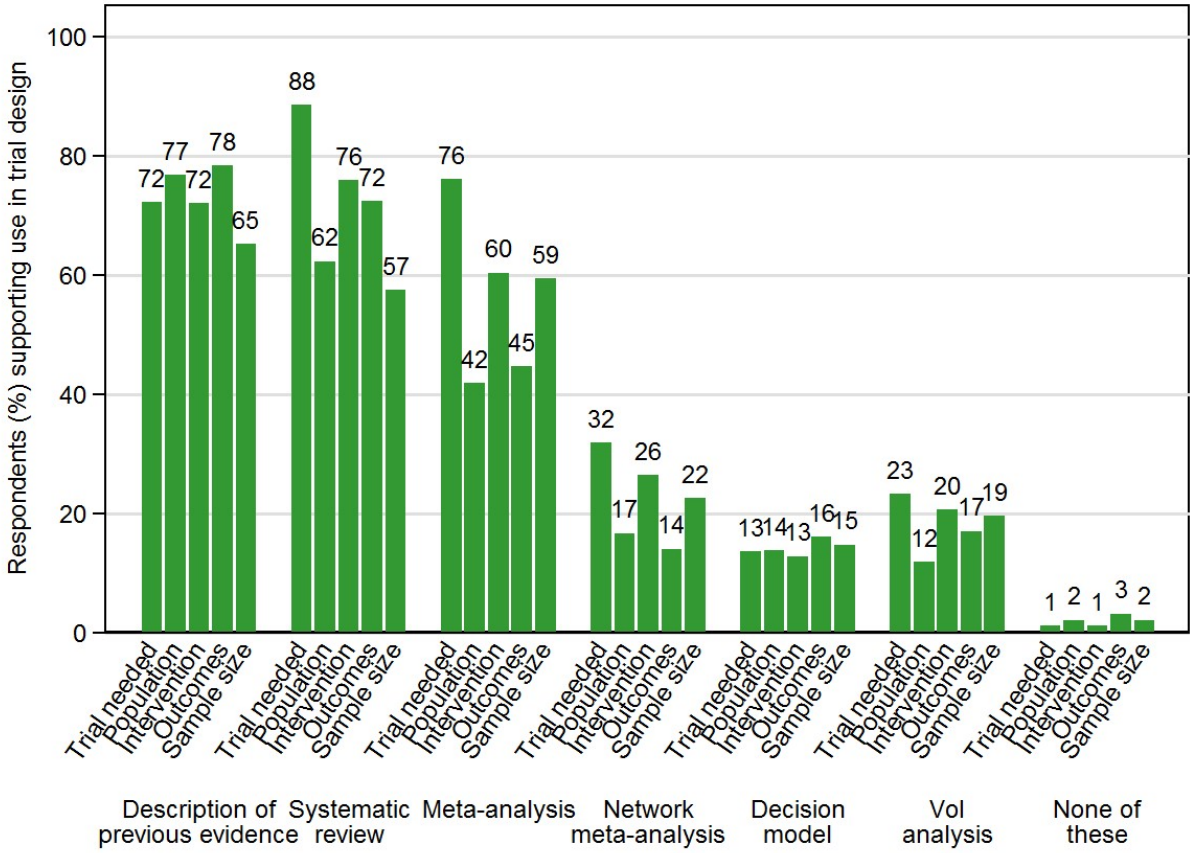
N=106	**n**	**%**
Over 20 years	16	15.1
Aspects of clinical trials have you been involved in		
Trial design	85	80.2
Trial conduct	71	67.0
Statistical analysis	73	68.9
Undertaking a systematic review of trials	52	49.1
None of these	3	2.8

2.3.1 Use of evidence synthesis to inform trial design

Figure 2.1 summarises the views of respondents on the *desirability* of using evidence synthesis in trial design. Support for using a description of previous evidence or a systematic review to inform each aspect listed was high. For most aspects of design, support was slightly higher for a simple description of previous evidence than a systematic review. In contrast, there was slightly more support for a systematic review to inform whether a trial is needed (92/104 or 89% systematic review versus 75/104 or 72% description of previous evidence) and the choice of interventions (78/103, 76% versus 74/103, 72% respectively). Over 50% of respondents also felt that a meta-analysis should be used to inform whether a trial is needed, the choice of interventions and the sample size.

Fewer respondents indicated support for use of more complex analyses (NMA, decision models and VoI analyses). For example, only 19% (20/101 respondents) indicated that VoI analyses should be used to inform sample size calculations. Of these respondents, 55% (11/20 respondents) were statisticians and 20% (4/20 respondents) were health economists, including one person who identified themselves in both roles. However, six of the eight health economists (75%) supported such use of VoI calculations across at least

Figure 2.1: **Views of respondents on whether evidence synthesis methods should be used to inform trial design.** The type of evidence synthesis method is summarised across five aspects of trial design: whether a trial is needed (n=104), choice of population (n=103), choice of interventions (n=103), choice of outcomes and follow-up time (n=101), sample size (n=103).



one aspect of design.

All respondents indicated support for using some form of evidence synthesis in at least three of the five aspects of trial design that were listed. Seven respondents, all of whom had experience in trial design, suggested that no form of evidence synthesis was required for one or two specific aspects, most commonly ‘choice of outcomes and follow-up time’ (3/101 or 3% of respondents). Full results are shown in Table 2.2.

Of the 85 respondents who indicated involvement in trial design, Figure 2.2 contrasts their

Table 2.2: How evidence synthesis 'should' be used in trial design

N=106	Description of previous evidence		Systematic review		Meta-analysis		NMA		Decision model		VoI analysis		None of these	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Whether a trial is needed[1]	75	72.1	92	88.5	79	76	33	31.7	14	13.5	24	23.1	1	1
Choice of population[2]	79	76.7	64	62.1	43	41.7	17	16.5	14	13.6	12	11.7	2	1.9
Choice of interventions[3]	74	71.8	78	75.7	62	60.2	27	26.2	13	12.6	21	20.4	1	1
Choice of outcomes and follow-up time[4]	79	78.2	73	72.3	45	44.6	14	13.9	16	15.8	17	16.8	3	3
Sample size[5]	67	65	59	57.3	61	59.2	23	22.3	15	14.6	20	19.4	2	1.9
Any design aspect[6]	88	87.1	93	92.1	87	86.1	38	37.6	28	27.7	35	34.7	7	6.9

[1] 2 respondents with missing data. [2] 3 respondents with missing data.
[3] 3 respondents with missing data. [4] 5 respondents with missing data.
[5] 3 respondents with missing data. [6] 5 respondents with missing data.

views on how evidence synthesis methods should be used versus their own use during the last 10 years. Full results are shown in Table 2.3.

Slightly more respondents indicated they had *used* a description of previous evidence to inform aspects of trial design than had indicated that such use was *desirable*. For example, 82% (69/84) had used a description of previous evidence to decide whether a trial is needed, compared with 71% (60/84) indicating support for such use. Of the 69 respondents who had used a description of previous evidence in this way, 14 (20%) did not indicate that such use was desirable. In contrast, our results suggested that trial design practitioners would like to be using each of the other five types of evidence synthesis more than they currently do in practice. This pattern was consistent across all aspects of trial design. For example, only 50% (42/84) of respondents had used a meta-analysis to inform whether a trial is needed, whereas 74% (62/84) thought it was desirable. 93% (39/42) of those who had used a meta-analysis to inform whether a trial is needed felt that such

Table 2.3: Actual uses of evidence synthesis in trial design during the last 10 years, compared with uses considered desirable

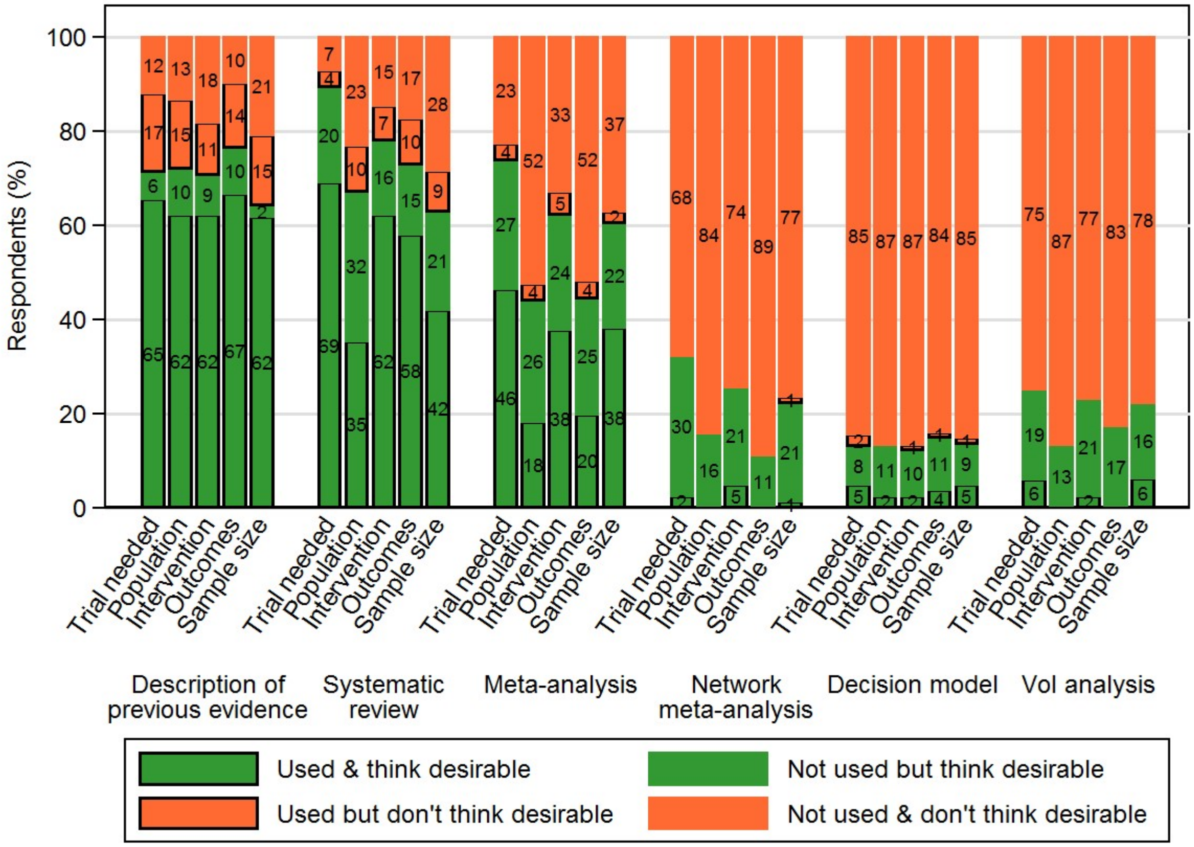
N=85	Yes used, yes desirable		Not used, yes desirable		Not used, not desirable		Yes used, not desirable		Total used		Total should be used	
Whether a trial is needed[1]												
Description of previous evidence	55	65.5	5	6	10	11.9	14	16.7	69	82.2	60	71.4
Systematic review	58	69	17	20.2	6	7.1	3	3.6	61	72.6	75	89.2
Meta-analysis	39	46.4	23	27.4	19	22.6	3	3.6	42	50	62	73.8
NMA	2	2.4	25	29.8	57	67.9	0	0	2	2.4	27	32.2
Decision model	4	4.8	7	8.3	71	84.5	2	2.4	6	7.2	11	13.1
VoI analysis	5	6	16	19	63	75	0	0	5	6	21	25
Choice of population[2]												
Description of previous evidence	51	62.2	8	9.8	11	13.4	12	14.6	63	76.8	59	72
Systematic review	29	35.4	26	31.7	19	23.2	8	9.8	37	45.2	55	67.1
Meta-analysis	15	18.3	21	25.6	43	52.4	3	3.7	18	22	36	43.9
NMA	0	0	13	15.9	69	84.1	0	0	0	0	13	15.9
Decision model	2	2.4	9	11	71	86.6	0	0	2	2.4	11	13.4
VoI analysis	0	0	11	13.4	71	86.6	0	0	0	0	11	13.4
Choice of intervention[3]												
Description of previous evidence	51	62.2	7	8.5	15	18.3	9	11	60	73.2	58	70.7
Systematic review	51	62.2	13	15.9	12	14.6	6	7.3	57	69.5	64	78.1
Meta-analysis	31	37.8	20	24.4	27	32.9	4	4.9	35	42.7	51	62.2
NMA	4	4.9	17	20.7	61	74.4	0	0	4	4.9	21	25.6
Decision model	2	2.4	8	9.8	71	86.6	1	1.2	3	3.6	10	12.2
VoI analysis	2	2.4	17	20.7	63	76.8	0	0	2	2.4	19	23.1
Choice of outcomes and follow-up time[4]												
Description of previous evidence	54	66.7	8	9.9	8	9.9	11	13.6	65	80.3	62	76.6
Systematic review	47	58	12	14.8	14	17.3	8	9.9	55	67.9	59	72.8
Meta-analysis	16	19.8	20	24.7	42	51.9	3	3.7	19	23.5	36	44.5
NMA	0	0	9	11.1	72	88.9	0	0	0	0	9	11.1
Decision model	3	3.7	9	11.1	68	84	1	1.2	4	4.9	12	14.8
VoI analysis	0	0	14	17.3	67	82.7	0	0	0	0	14	17.3
Sample size[5]												
Description of previous evidence	50	61.7	2	2.5	17	21	12	14.8	62	76.5	52	64.2
Systematic review	34	42	17	21	23	28.4	7	8.6	41	50.6	51	63
Meta-analysis	31	38.3	18	22.2	30	37	2	2.5	33	40.8	49	60.5
NMA	1	1.2	17	21	62	76.5	1	1.2	2	2.4	18	22.2
Decision model	4	4.9	7	8.6	69	85.2	1	1.2	5	6.1	11	13.5
VoI analysis	5	6.2	13	16	63	77.8	0	0	5	6.2	18	22.2

[1] 1 respondent with missing data. [2] 3 respondents with missing data. [3] 3 respondents with missing data. [4] 4 respondents with missing data. [5] 4 respondents with missing data.

Table 2.4: Types of evidence synthesis used in trial design

N=85	Description of previous evidence		Systematic review		Meta-analysis		NMA		Decision model		VoI analysis	
	n	%	n	%	n	%	n	%	n	%	n	%
Previously published evidence syntheses (such as a systematic review or meta-analysis)	62	72.9	59	69.4	46	54.1	5	5.9	7	8.2	1	1.2
Conducted by the clinical trial team	59	69.4	55	64.7	31	36.5	2	2.4	7	8.2	6	7.1

Figure 2.2: **Comparisons between desirable and current practice in use of evidence synthesis methods in trial design.** This is summarised by type of evidence synthesis method, among survey respondents involved in trial design to inform five aspects of trial design: whether a trial is needed (n=84), choice of population (n=82), choice of interventions (n=82), choice of outcomes and follow-up time (n=81), sample size (n=81). Numbers displayed are percentages.



use was desirable.

Some 96% (78/81) of respondents claimed to have used some form of evidence synthesis to inform sample size calculations in the last 10 years, close to the 99% (80/81) who indicated support for such use (data not shown). Making the same comparison but excluding the less formal ‘description of previous evidence’, we found a larger discrepancy: 62% (50/81) had used evidence synthesis methods to inform sample size calculations, compared with 84% (68/81) indicating that this is desirable (data not shown). Only 6% (5/81) of respondents had used a VoI analysis to inform sample size calculations, compared with 22% (18/81) indicating that VoI analysis should be used for this. All five respondents who had used VoI in this way were in support of its use. For all types of evidence synthesis methods except VoI analyses, which was mostly conducted by the clinical trials team, use of previously published evidence syntheses was most common, although only marginally (See Table 2.4).

2.3.2 Use of evidence synthesis to inform trial analysis

Of the 106 participants who were asked all questions on the survey, only 100 (94%) answered questions on trial analysis. 79% (79/100) of respondents indicated that external information about the treatment effect should be used to inform aspects of the analysis (See Figure 2.3; Table 2.5). Similarly, 69% (69/100) expressed support for using external information related to potential biases in trial analysis, and 67% (67/100) for use of external evidence on other quantities which are usually poorly estimated. While only a few respondents (5% or less) indicated that external evidence should not be used in these ways, between 15 and 30% selected the ‘don’t know’ or ‘don’t understand’ options.

73/106 (69%) respondents indicated they were involved in trial analysis. Figure 2.4 contrasts the views of this subsample on how evidence synthesis methods should be used to

Figure 2.3: **Views of respondents on whether evidence synthesis should be used to inform trial analysis.** This is summarised across three aspects of trial analysis: the treatment effect, potential biases arising from trial conduct and other quantities (of n=100 people who answered this question).

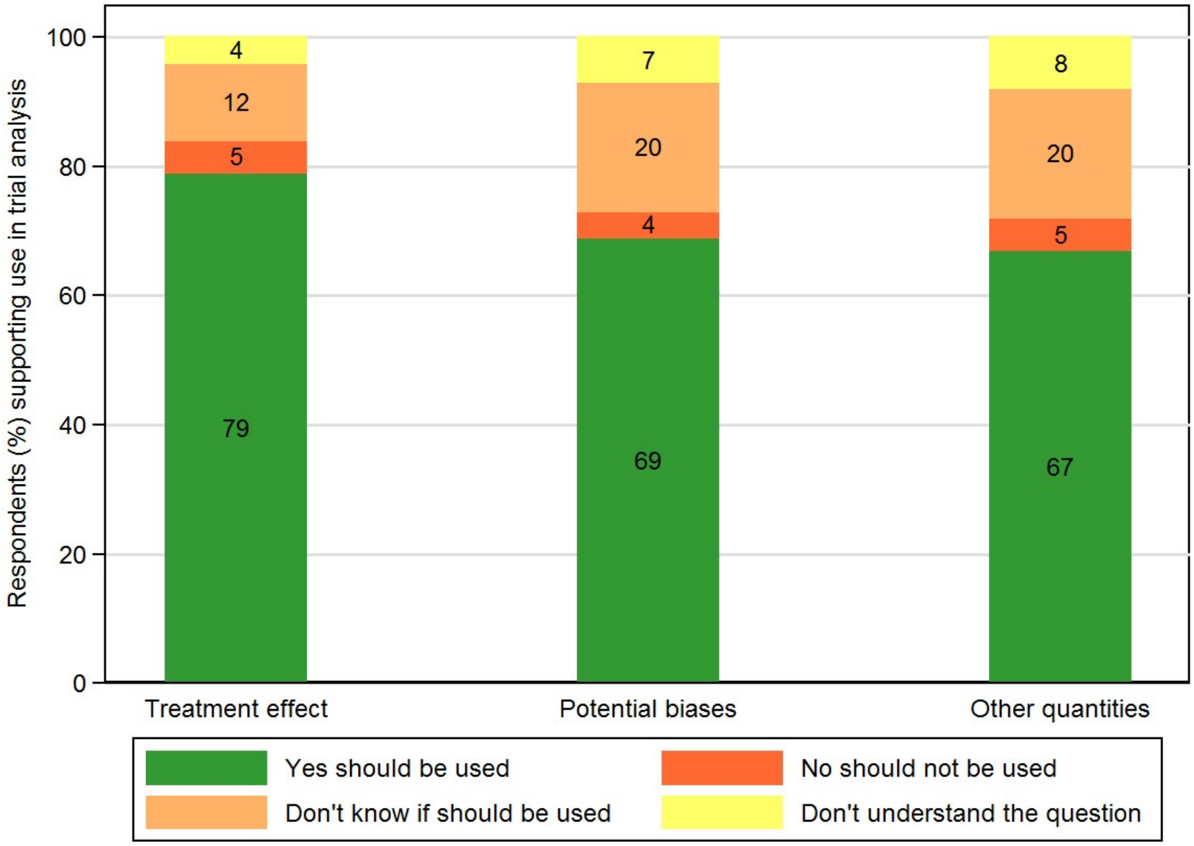


Table 2.5: How evidence synthesis methods ‘should’ be used to inform aspects of trial analysis

N=100	Yes		No		Don’t know		I don’t understand	
	n	%	n	%	n	%	n	%
External information about the treatment effect (including a meta-analysis)	79	79	5	5	12	12	4	4
External information related to potential biases arising from trial conduct (e.g. blinding infeasible)	69	69	4	4	20	20	7	7
External information about other quantities involved in the analysis (e.g. correlations or baseline event rates)	67	67	5	5	20	20	8	8

inform aspects of analysis versus their own use in practice. 52% (35/68) indicated that, during the past ten years, they had used external information about the treatment effect to inform trial analysis, compared with 79% (54/68) indicating support for such use. 97% (34/35) of those who had used external information in this way felt that such use was desirable. Whilst 63% (20/32) of respondents who had not used external information about the treatment effect in trial analysis also felt such use was desirable, 22% (7/32) were not sure. Similar patterns were seen for using external evidence on potential biases and other quantities. Full results are shown in Table 2.6. A sensitivity analysis including only statisticians suggested slightly less use of external evidence in each of the three areas (See Figure 2.5).

Table 2.6: Actual uses of evidence synthesis in trial analysis during the last 10 years, compared with uses considered desirable

N=73	**Yes used, yes desirable**		**Not used, yes desirable**		**Not used, not desirable**		**Yes used, not sure**		**Yes used, not desirable**		**Not used, not sure**		**Don't understand**	
Treatment effect[1]	34	50	20	29.4	5	7.4	1	1.5	0	0	7	10.3	1	1.5
Potential biases[2]	28	40.6	19	27.5	4	5.8	2	2.9	0	0	12	17.4	4	5.8
Other quantities[3]	33	48.5	16	23.5	5	7.4	2	2.9	0	0	10	14.7	2	2.9

[1] 5 respondents with missing data. [2] 4 respondents with missing data. [3] 5 respondents with missing data.

Figure 2.4: **Comparisons between desirable and current practice in use of evidence synthesis methods in trial analysis.** This is summarised among survey respondents involved in trial analysis to inform three aspects of trial analysis: the treatment effect (n=68), potential biases arising from trial conduct (n=69) and other quantities (n=68). Numbers displayed as percentages.

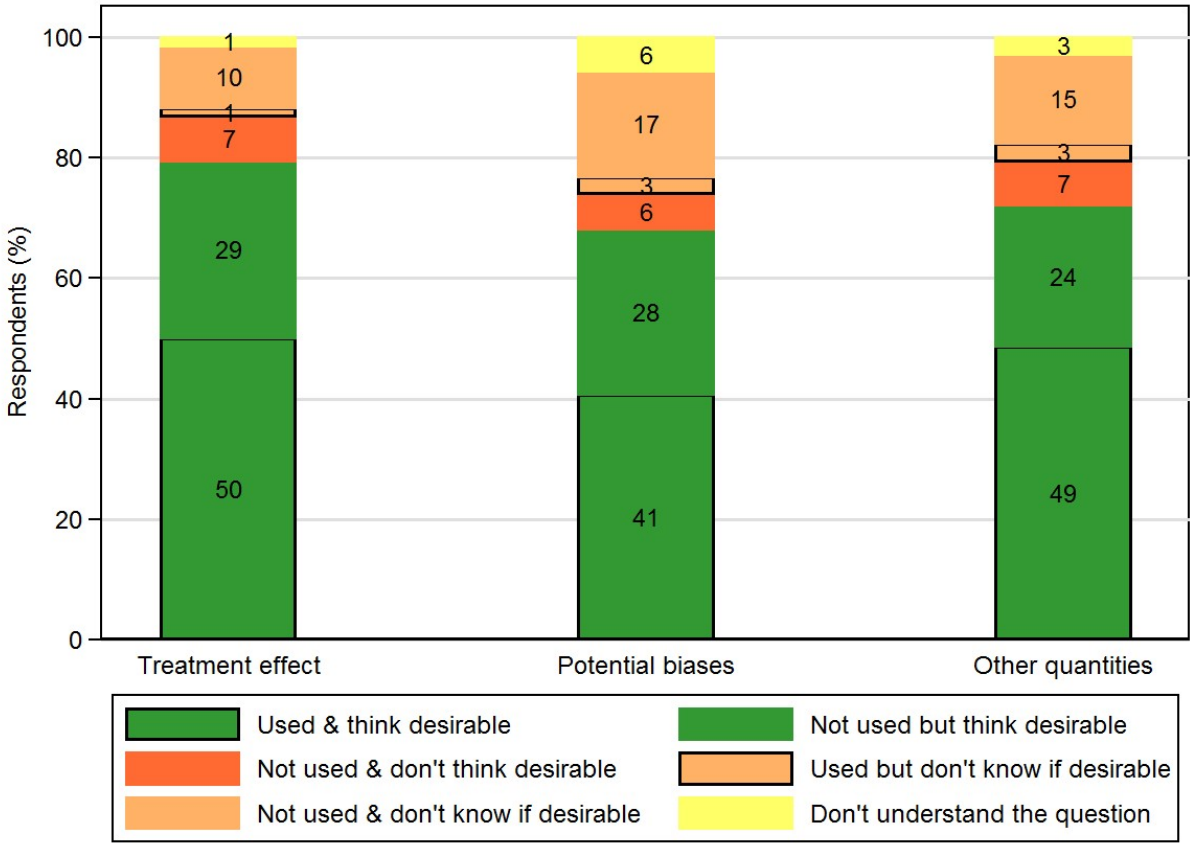
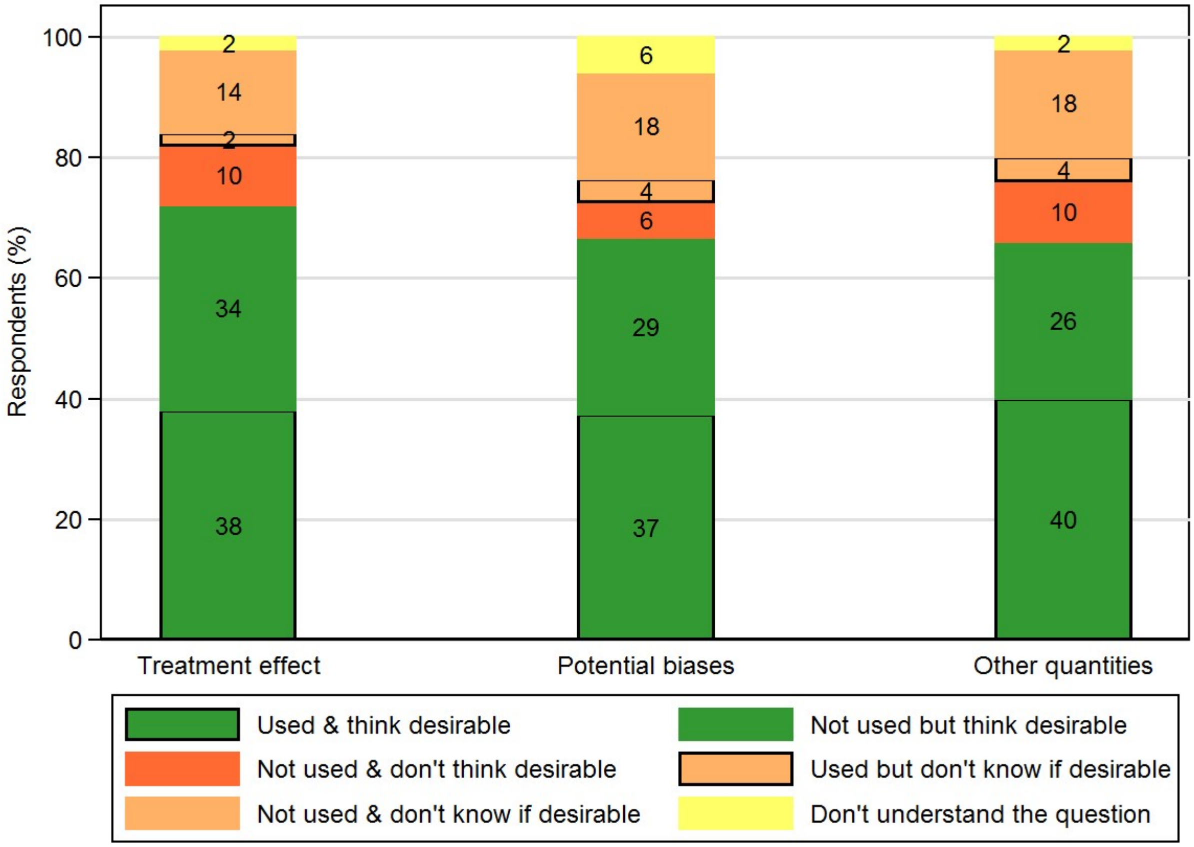


Figure 2.5: Comparisons between ideals and current practice of evidence synthesis methods *amongst statisticians only* in use of evidence synthesis methods to inform three different aspects of trial analysis: the treatment effect (n=54), potential biases arising from trial conduct (n=54) and other quantities (n=54).



2.3.3 Barriers to the use of evidence synthesis methods

Figure 2.6 shows the barriers to using evidence synthesis, ordered by their perceived importance. The bars show the total number of points awarded to each barrier, split by the number of points it acquired by being ranked the first, second and third greatest barrier. Of the 106 participants who were asked all questions on the survey, only 103 (97%) answered questions on barriers to the use of evidence synthesis methods. 87% (90/103) of respondents answered this question fully. By far the greatest perceived barrier was time constraints. This was followed by a belief the trial was the first in the area, and a belief that previous trials were different from the current trial. Of those selecting 'Other', reasons included complexity of the trials and the "Chief Investigator had more evidence than previously published information." 'Objections to using evidence syntheses (from you or colleagues)' was the lowest scoring barrier of those listed. The conclusions remained unchanged when the analysis was restricted to statisticians only (data not shown).

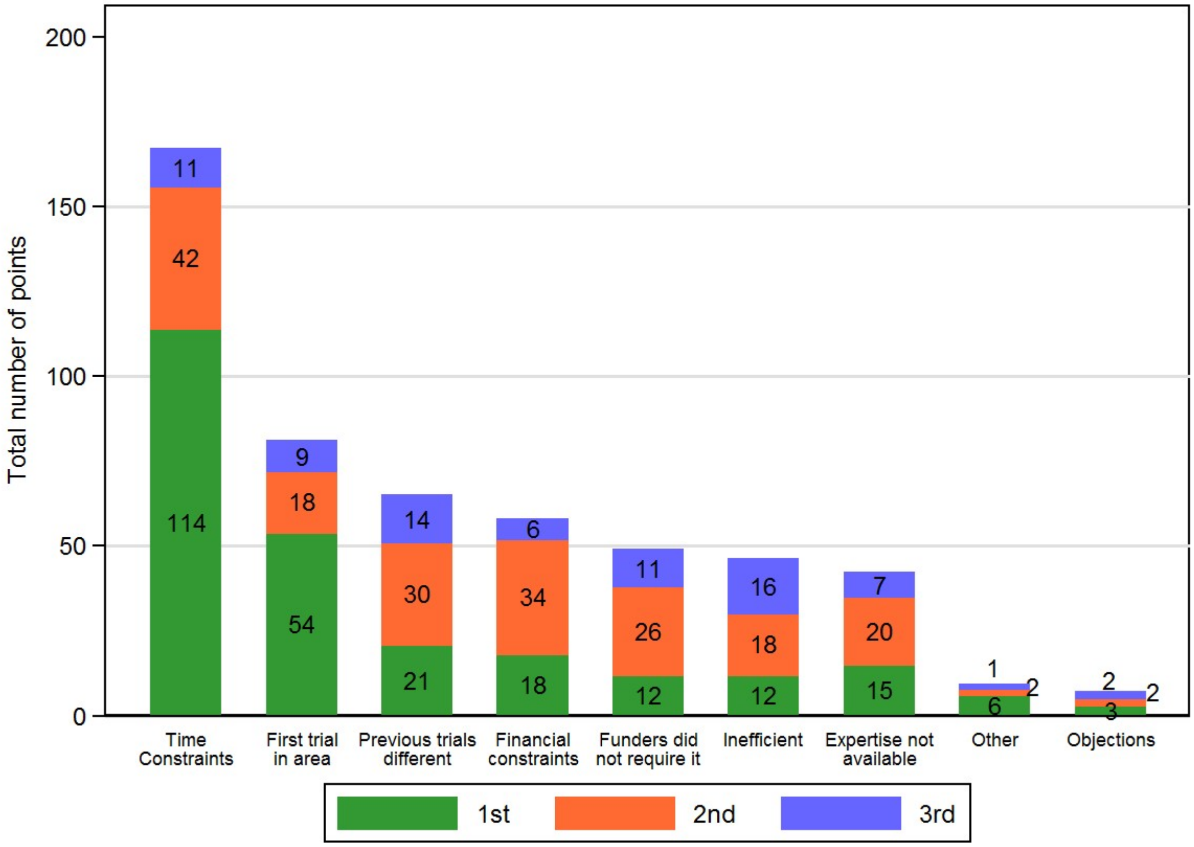
2.4 Discussion

This survey had three aims: first, to summarise and contrast the use of evidence synthesis methods to inform aspects of trial design; second, to summarise and contrast the use of evidence synthesis methods to inform aspects of trial analysis and third, to understand potential barriers to such use of evidence synthesis in both design and analysis. The key findings and limitations are displayed in Figure 2.7 and now discussed.

2.4.1 Overview of key findings

Our INVEST survey indicates a high level of support for use of evidence synthesis to inform aspects of trial design and analysis. Support was generally high for using a de-

Figure 2.6: **Barriers to the use of evidence synthesis (higher scores indicate greatest perceived barriers).** 3 points were assigned to the greatest barrier, 2 points to the second and 1 to the third. For example, 38 respondents ranked time constraints as the greatest barrier (3 x 38 = 114 points), 21 ranked it second (2 x 21 = 42) and 11 third (1 x 11=11).



scription of previous evidence, a systematic review or a meta-analysis when designing a trial. Fewer respondents indicated support for use of NMA, decision models and VoI analyses. Only a few respondents (approximately 5%) felt that external evidence about particular parameters should not be used in the analysis of a trial, however many (up to 20%) did not know if such evidence should be used in practice. Our results indicate some discrepancies between the evidence synthesis methods people think should be used and what they are using in current practice. In particular, respondents did not appear to be using systematic reviews, meta-analyses, NMAs, decision models and VoI analyses as much as they wanted across all aspects of trial design.

The greatest perceived barrier to using evidence synthesis methods in trial design or analysis was time constraints, followed by a belief the new trial was the first in the area. The second biggest barrier of, 'the first trial in this area' seems a reasonable one, although, it could still be possible to use evidence synthesis for some unknowns that go into design considerations. One of the barriers to evidence synthesis methods in practice was "Chief Investigator had more evidence than previously published information". Arguably, this may be true in some settings, in that pooling over a disparate set of studies may obscure the key uncertainties.

2.4.2 Limitations

Of approximately 638 attendees of the conference, 106 (17%) completed the survey, half of whom were statisticians. Some 95% of our sampling frame were from the UK and Republic of Ireland, so the results may not be generalisable to the international clinical trials community. Although the support of evidence synthesis methods is more promising than we may have initially thought it is likely that we sampled more methodologists than trialists. Our sampling frame consisted of conference delegates closely involved in trial design and analysis, who are likely to have a strong interest in promoting good practice. As such, we

Figure 2.7: Key findings from the INVEST survey

Characteristics of the sampling frame

- Of approximately 638 attendees of the conference, 106 (17%) completed the survey, half of whom were statisticians.
- It is likely we sampled more methodologists than applied trialists, given the focus of the conference on trial methodology. Methodologists are likely to be more supportive of more advanced methods, but their views could be quite different from other trialists.

Aim 1: Evidence synthesis methods in trial design

- Support was generally high for using a description of previous evidence, a systematic review or a meta-analysis in trial design.
- Generally, respondents did not seem to be using evidence syntheses as often as they felt they should. For example, only 50% (42/84 of relevant respondents) had used a meta-analysis to inform whether a trial is needed compared with 74% (62/84) indicating that this is desirable.
- Only 6% (5/81 relevant respondents) had used a VoI analysis to inform sample size calculations versus 22% (18/81) indicating support for this.

Aim 2: Evidence synthesis methods in trial analysis

- Surprisingly large numbers of participants indicated support for and previous use of evidence syntheses in trial analysis. For example, 79% (79/100) respondents indicated that external information about the treatment effect should be used to inform aspects of the analysis.

Aim 3: Barriers to the use of evidence synthesis methods in trial design and analysis

- The greatest perceived barrier to using evidence synthesis methods in trial design or analysis was time constraints, followed by a belief the new trial was the first in the area.
- Evidence syntheses can be resource-intensive, but their use in informing the design, conduct and analysis of clinical trials is widely considered desirable.
- We advocate additional research, training and investment in resources dedicated to ways in which evidence syntheses can be undertaken more efficiently, offering the potential for cost savings in the long term.

might expect our sample to answer some of the questions more favourably than the wider population of people involved in clinical trials. In particular, half of respondents were statisticians (51%), who may be expected to be more open to advanced statistical methods (such as using evidence syntheses to improve precision in estimates of some parameters) compared with other contributors to the design, conduct, or delivery of trials. Statisticians are also influential members of the multi-disciplinary teams that are involved in trial design and may be useful advocates for increased use of available evidence in trial design. Although it would have been interesting to explore differences across research centres and countries, we chose not to collect such geographical data to protect anonymity and minimise the burden of survey completion.

The large proportions of respondents who indicated that they had either used evidence synthesis to inform trial analysis or that they believed evidence synthesis should be used in this way were surprising. Even more surprisingly, a sensitivity analysis including only statisticians provided slightly lower estimates of these proportions, although the small sample size precludes strong assertions. It is unlikely that these relatively advanced methods are being used so frequently in practice, we feel that it is likely the questions related to trial analysis were misunderstood to mean the general use of previous evidence whilst the intention had instead been to elicit views on the use of informative prior distributions in a Bayesian statistical framework. This explanation appears to be supported by the result that fewer statisticians than non-statisticians claim to be using external evidence in this way, i.e. it is likely that confusion about these questions was higher among non-statisticians although there is no direct evidence of this. For example, respondents might have interpreted the incorporation of 'external information about the treatment effect (including a meta-analysis)' in trial analysis as meaning including the new trial results in an updated meta-analysis. In order to gain feedback on the questions, the early versions of the survey were piloted to colleagues known to the study team who had some involvement in clinical trials. Although the survey was piloted, it is still possible there was ambiguity about how the questions regarding the use of external evidence to inform the analysis stage of a trial were interpreted. Instead, examples of such uses may have been more useful and helped to clarify that it was the more formal use of evidence synthesis that we were interested in.

Since this was a paper and online survey, we ensured paper and online versions were identical. In doing so, we allowed for missing data in the online survey to exactly match the paper survey, rather than including missing data queries. It was also anonymous in order to get participants to answer truthfully. As such, it was not possible to go back to ask respondents how they had interpreted some questions. In the question regarding the use of existing evidence to inform parameters in trial analysis, it was also not stated whether the intention was as a primary or secondary analysis.

The question regarding the use of a systematic review or meta-analysis to inform a sample size calculation could instead indicate a variety of uses. For example, if a respondent had ticked yes, it is not known how external evidence is being used. Informally, this may include, justification for the choice of the effect size and standard deviation estimates, or more, formally, translated external evidence into a prior distribution and used either conditional or simulated power.

The use of decision models was low. A possible reason for this includes the way in which the question was phrased. Decision models can only be used in trial design through VoI analyses and have no other use. Therefore, having this as a separate option for decision model and VoI analyses could have confused respondents.

To summarise the barriers to use of evidence synthesis, we assigned scores based on an arbitrary assumption of linearity, i.e. such that an individual's highest ranked barrier is 3 times as important as his/her third barrier. These scores, although helpful for summarising data, might not reflect respondents' true views. We intended all listed barriers to be interpreted as reasons why a trial team might not seek or carry out evidence synthesis. However, it is possible that some respondents who chose ‘Believed to be the first trial in the area’ could have been thinking of the situation where a literature search or systematic review reveals no previous trials. The extent of this barrier would then be over-estimated.

2.4.3 Comparison to recent reviews of evidence synthesis methods in trial design

The INVEST survey provides generally higher estimates of use of systematic reviews in trial design than the recent review of Bhurke *et al* [57], with the exception of ‘justification of the trial’ (Bhurke *et al* 94%, versus INVEST 73%). For example, 68% of our respondents indicated that they had used a systematic review to inform choice of outcomes and follow up time, whereas only 16% and 6% of trials reviewed by Bhurke *et al* had used a review to inform these two aspects respectively. Similarly, 51% of our respondents said they had used a systematic review to inform sample size calculations, seemingly in contrast to Bhurke *et al's* findings that only 9% of trials had used a review to inform the standard deviation and 3% to ‘estimate the difference to detect or margin of equivalence’. It is possible that other trials in the Bhurke *et al* review relied on pilot trials to inform these parameters [16, 17], while the INVEST results seem to suggest that relevant information will often be available from evidence syntheses. However, the results are not directly comparable since we asked respondents to consider all trials they had been involved in during the last 10 years whereas Bhurke *et al* investigated whether evidence synthesis had been used in specific individual trials. On the other hand, Bhurke *et al* reviewed only publicly funded (NIHR HTA) trials, while trialists attending ICTMC are likely to also participate in company funded trials, for which less justification is required and there is possibly a stronger expectation for independently clear results.

In agreement with Bhurke *et al* we found that important barriers to the use of evidence synthesis in practice include a new trial being the first in its area or being different from trials included in a previous review. However, by directly asking trialists instead of relying on documentation, we were able to see that the greatest barrier is time constraints. In attempt to overcome the issue of time constraints when synthesising evidence, many methods for rapid reviews have been proposed over recent years [89, 90]. Khangura *et al* [90] developed their own eight step approach of conducting a rapid review having reviewed the current literature. Implementation of their approach in HTA trials has been successful and can be applied to other types of trials [91]. However, more training on approximate methods and rapid reviews is needed to support their wider use in practice. Investment in adequate resources and training at this stage could lead to cost savings in the longer term, by reducing waste in research.

We found less support for the use of NMAs, decision models and VoI analyses in trial design, which may be because they are more complex to conduct and require a higher investment of time and expertise. These methods could further help inform decisions but also require additional assumptions and ’a priori’ parameter estimates such as the cost-effectiveness threshold and parameters related to structural uncertainties in the case of VoI, which may not be available. A policy framework on when and how to perform such analyses and how they are used could be a useful next step [53].

We also note that most individual trials investigate a specific research question for one particular treatment: for example, in 2014, 80% of trials were still two arm trials [92]. In contrast, NMAs, decision models and VoI analyses are commonly used to make decisions and inform policy when there is a choice between a number of concurrent treatment options [93]. These methods could be considered less relevant in the design and analysis of an individual two-armed trial. VoI analyses, in particular, are usually commissioned in high value trials, often in situations with many treatments and uncertainty as to which is best. However, a NMA could be more relevant to inform the interventions of a two arm trial if used at the earlier part of the design process [49].

Trial-based economic analyses are sometimes secondary to the clinical aspect rather than being fully integrated within a trial design [54], meaning that the use of decision models and VoI analyses to inform trial design can be limited. Only 6% (5/84) of our respondents had used a VoI analysis to inform whether a trial is needed, although all of those who had used a VoI analysis were in favour of its use more generally. Models in health economic analyses are a strongly simplified representation of disease history and treatment effects and are framed around a particular decision setting (e.g. UK) using setting-specific values for healthcare use, costs and health benefits. These values may change over time and are likely to be different in other settings. Streamlining of decision modelling and VoI analyses would therefore be particularly challenging. Despite the assumptions and limitations of a VoI analysis, its potential to guide the need for and the design of new studies warrant its wider consideration and further development [94]. We explore a specific case of this in Chapter 6 on the use of EVSI for sample size calculations.

We did not explore the views of funders or reviewers specifically, but this could be another valuable avenue for future research, given the critical role they could play in minimising research wastage.

2.5 Conclusions and rationale for remainder of thesis

The results of the survey highlighted that trial teams responding to the INVEST survey at the ICTMC generally reported they are using evidence synthesis in trial design and analysis more than we might have expected, but less than they might like to. This motivates the need for methods development and guidance. Time constraints was identified as the greatest barrier to more widespread use, so we discuss the practical implications of implementing such methods throughout the remainder of this thesis.

In trial design, for both whether a trial was needed and for choosing an intervention, more respondents said that a systematic review, rather than a less formal description of previous evidence, should be used. It therefore seems that respondents felt the need for a thorough, systematic approach in order to show convincingly whether there is a gap in the evidence base that merits a new trial. For the other aspects of trial design, there may not be sufficient available evidence to warrant a systematic review, so that a less formal description of previous evidence might be felt to be adequate.

It is likely that we sampled more methodologists than trialists, given the focus of the conference. Methodologists are likely to be more supportive of such methods but unrepresentative of trialists in general. Furthermore, it is likely the questions related to trial analysis were not fully understood by respondents from this brief written survey. To address this, a qualitative study is conducted in the next chapter to investigate more thoroughly how trialists are currently using evidence synthesis to inform both design and analysis, and the potential barriers to an increased amount of such use.

The survey also highlighted specific areas where trialists would like to be using more evidence synthesis methods such as sample size calculations. This motivates Chapter 6 to compare current methods of using the results of a meta-analysis to inform sample size calculations. Together with further work in the following chapters, we will explore the consideration of all relevant prior information, including the statistical and clinical relevance.

3 Eliciting current views and exploring barriers: a qualitative study

This study was a collaboration with Daisy Elliott (DE), a qualitative researcher, and my supervisors Hayley Jones (HJ) and Julian Higgins (JH). I was the lead investigator of the study, forming the idea with HJ and JH and conducting all interviews and analyses.

3.1 Introduction

3.1.1 Context and objective

The results of the survey in the previous chapter indicated that approximately 50% of those involved in trial analysis *reported* they had used evidence synthesis to inform at least one area of the analysis [78]. However, we suspect the true proportion may be much lower. Respondents may have either (i) answered the question more favourably because most were methodologists rather than trialists or (ii) misinterpreted the question as referring to the variety of different uses of evidence synthesis in a more informal manner (rather than the use of informative priors, of which, the prevalence is low (see Section 1.4)), or both. For example, respondents may have interpreted this as including the new trial results in an updated meta-analysis rather than the more formal use of synthesising existing trials in a Bayesian analysis.

Our aim in the survey had been to elicit views on how a synthesis of existing evidence could be incorporated, for each of: (i) the treatment effect, (ii) potential biases, and (iii) nuisance parameters. Interestingly, only about 5% of responders (n=5) felt that external evidence should not be used in the analysis of a trial, however about 20% did not know if such evidence should be used. In this chapter, we therefore undertake a qualitative study to explore the attitudes and experiences of trialists towards incorporating external evidence, through the Bayesian design or analysis of a trial. Since there is less in the literature exploring the use of informative priors to inform the analysis stage [13], we focus *more* on this than the design stage. This study builds upon our initial findings in the INVEST survey, exploring the barriers to the use of informative priors, through semi-structured interviews. Qualitative work can explore individuals' views, attitudes and previous experiences to allow an in-depth understanding on multi-faceted topics [75, 95, 96, 97, 98], which may not be picked up in quantitative study designs, such as surveys.

We explore *how* external evidence could be useful to trialists; *which* types of external evidence might be considered most relevant and *what* level of such use might be acceptable in practice. A secondary aim of our study is to improve our understanding of how current methodology is chosen; specifically, what kind of information informs this choice, and by whom. We want to find the areas where external information could have the most benefit over a frequentist analysis to minimise future research waste. We hope this will enable recommendations to be made regarding the use of external evidence syntheses during the design and analysis of clinical trials. We also hope to identify the research and/or training needs to facilitate the incorporation of external evidence.

3.1.2 Background to qualitative research

In this section, a short overview of qualitative research is given. Two of the most common approaches to qualitative research are grounded theory (Glaser & Strauss [99]) and thematic analysis [76]. A grounded theory approach is used when you have no preconceptions about the data and a literature review is conducted *after* the data analysis. Conversely, in a thematic analysis a literature review is often conducted *before* any data collection.

Within qualitative research, there are two different approaches which can be broadly categorised into an ′inductive′ and ′deductive′ approach. In an inductive approach patterns are developed from the content of the data, whereas in a deductive approach patterns from the data are driven by existing ideas. An inductive approach is usually preferred as it shows findings have emerged analytically, thus reducing potential bias [76]. An inductive approach is inherent in a grounded theory framework whilst a thematic analysis can use either approach. Since we had some idea of the literature regarding the use of evidence synthesis methods (Chapter 2) and the use of informative priors (or lack of) in trial design and analysis (Section 1.4), a thematic analysis with an inductive approach was conducted in the study and is described in more detail later on.

Prior to the analysis, eligible participants are sampled. Sampling in qualitative research can be viewed as more flexible than in a quantitative framework. Some of the most common sampling methods in qualitative research are purposeful and snowball sampling. Purposeful sampling is a sampling strategy which intends to recruit participants based on particular characteristics of the eligibility criteria, in order to achieve variation between respondents [100, 101]. Snowball sampling is a type of convenience sampling; often used to improve recruitment by asking participants to suggest other people in similar job roles [102]. It can therefore aid comparisons of similar or disparate views between such groups of participants.

There are several data collection methods in qualitative research such as interviews, focus groups, observations etc. Each have their strengths and weaknesses and very much dependent on the research question. For example, focus groups can be particularly useful for gathering general opinions or beliefs and generating discussion amongst people, whilst a disadvantage can be if there are particularly dominant or controversial views others in the group could follow their opinions [103]. On the other hand, interviews are advantageous when looking at the experiences and views of particular respondents. The interview process is usually ‘semi-structured’ meaning the interviewer has a list of topics which they would like to be discussed with the participant [104]. These topics are usually phrased as open-ended questions to allow a broad discussion of such topics and facilitate and inductive approach. Conducting a semi-structured interview allows each participant to be asked similar questions and aid comparison between individuals, whilst allowing sufficient flexibility for discussion. A ’topic guide’ describing these questions/topics is often used. It is common for the topic guide to be modified, with some questions added or replaced as the study progresses. This may occur when relevant topics not originally listed in the topic guide are raised in interviews and the researcher may want to raise the topic in future interviews.

Braun and Clarke [76] describe a six-step approach to a thematic analysis. The first step involves getting to know the data (i.e. interview transcripts) by reading and re-reading it to get familiar with all aspects. Second, the researcher labels parts of the data which are interesting and form part of an answer to the research question. These are known as codes. Codes are generally short with an analytical meaning. As the analysis is often an iterative process, these steps may taken for the first two to three interviews, and then the topic guide modified if applicable and so on. Within this stage, specific codes can be refined to support the same meaning, i.e. labelling relevant codes the same if the meaning is the same across interviews. This can involve a process of ’constant comparison’ whereby the researcher explores similarities and differences between interviewees [105].

The third step involves putting all the codes together from multiple interviews to see if there are any emerging patterns, i.e. potential themes. Deviant cases can also emerge here, whereby most interviewees or a subset of interviewees are broadly emerging to create a theme whilst a minority of views seem to diverge from this.

The fourth step is a process in which potential themes are checked against the raw data. This is an important step in which themes are often modified, merged or separated to tell the true story of the data [104]. The fifth step builds upon step four by defining and naming themes, ensuring each theme tells a story. This is generally where potential subthemes emerge; subthemes can be more specific to some groups (or more general) but help to tell the story of the overall theme. When no new key findings are emerging from new interviews, this is known as 'data saturation' and determines the end of further data collection [106]. Although this may be viewed as a subjective decision, data saturation is usually determined when in the final two to three interviews no new concepts are added to the key findings [107]. The final stage involves telling the overall story of the findings and putting the findings in the context of other research.

Due to the subjective nature of qualitative research, it is likely the researcher brings their own views and implicitly informs the research [77]. A reflexivity stance is often required when discussing results to highlight the implications and impact the researcher's views can have on the results.

These qualitative approaches have previously been used to explore issues in evidence synthesis methods. For example, a qualitative study by Lorenc *et al* explored the process of how researchers dealt with issues relating to complex data in systematic reviews and meta-analyses [108]. These authors used purposeful sampling to ensure a range of participants' views were explored amongst different fields. In general, the results consistently highlighted a lack of consensus on which methods researchers felt they should be using to deal with heterogeneity. Interestingly, many researchers were using their own judgement to deal with this. Methods were therefore chosen based on their own experience, rather than any methodological guidance. As such, the justification of these methods was often not explicitly written in the end of study report.

3.2 Methods

3.2.1 Recruitment and sampling (eligibility criteria)

We aimed to sample and interview a number of individuals working in clinical trials units in England. Individuals were eligible for inclusion in this study if they were at least a certain *position* but also had at least certain *responsibilities*. We aimed to sample a range of individuals, across multiple locations, from the following positions: methods leads, lead trial statisticians, trialists writing grant applications and leads of NIHR funded trials. Both clinicians and statisticians were eligible to take part. We also selected participants whose responsibilities broadly fell into one or more of the following categories: (i) conducting the analysis of trials in practice (ii) planning such analysis and/or (iii) responsible for people doing these things.

Individuals involved in developing evidence synthesis methods were excluded, for example members of the evidence synthesis working group with the Medical Research Council. Since we were recruiting individuals across various locations including some academic ones, ethical approval was obtained by the University of Bristol on 27/04/2017 (Reference number 48101). The consent form is shown in Figure A.2.

An initial list of potential participants was drawn up, based on connections members of the study team had with colleagues working in trials. The lead researcher (GC) contacted potential participants, via email, to explain the purpose of the study and ask whether they were willing to take part in an interview. A participant information sheet was also included to provide more detail about the study. This is shown in Figure A.3. Following the initial list of participants drawn up, a combination of snowball [102] and purposeful sampling [95] was used. Purposeful sampling was used to achieve maximum variation between respondents, in which participants were sampled based on their position and responsibilities. Whilst snowball sampling was used to help aid recruitment. We asked participants at the end of their interview if they could suggest anyone who they thought may be suitable to take part in the study. The sampling strategy intended to recruit individuals who were at different stages of their career and also had differing roles within a trials unit.

We did not specifically refer to Bayesian analysis in the participant information sheet, or state that participants should have experience with Bayesian analysis, thus encouraging a more diverse range of views from participants with different experiences and views. Instead, the participant information sheet (Figure A.3) talked about exploring 'trialists' views and experiences of analysing trials in the context of the wider evidence base.' Recruitment was driven by data saturation (whereby data collection continues until no new themes emerged).

3.2.2 Data collection

Interviews were semi-structured. Topic guides were used to ensure similar areas were covered in each interview, with sufficient flexibility to allow new issues of importance to emerge [109]. The topic guide was initially developed with suggestions from all members of the study team. An example topic guide is given in Figure A.4. There were two versions of the topic guide; one for clinicians and the other for statisticians. These topic guides were very similar, with some questions rephrased for clinicians to focus more on the conceptual ideas of using previous evidence in different scenarios. For these differences, see Table A.1.

The topic guide was iteratively modified in light of previous interviews. This included adding and rephrasing questions. For example, access to data and the issues surrounding this became an emerging theme when discussing how these methods could be implemented in practice. The following question was therefore added to both versions of the topic guide: “How would you access data needed [prompt: would such data need to be collated by you or another colleague?]”. Each time the topic guide was modified, it was saved as a new version. A log was kept of all amendments (Table A.2).

To meet our aim of understanding how analysis methods were chosen, we began the interviews by trying to elicit which methods trialists were currently using and have used throughout their career to analyse trials. We then explored the process of how previous evidence was considered, when designing and analysing a trial. This included how external evidence was collected and synthesised. We then explored participants’ views and experiences on the formal incorporation of existing data via a Bayesian analysis. As we did not require participants to have any knowledge of Bayesian analysis, the essentials of a Bayesian approach were explained to participants who did not have such understanding. For example, I explained how a meta-analysis of results of a set of similar previous trials could provide information on the likely size of the intervention effect in a new trial and be explicitly incorporated with the data from the new trial. We also wanted to know if participants knew of any colleagues using Bayesian analyses.

The latter part of the interview was used to examine hypothetical scenarios of how external evidence syntheses could be incorporated into a trial. This included our three case study scenarios, external evidence on: *bias* from meta-analyses within meta-epidemiological studies; *effect sizes* from trials within meta-analyses in sample size calculations and *outcomes* from patients within trials to help inform adverse events. We also explored potential barriers to implementing these methods in practice. At the end of the interview information was collected regarding the participant’s demography: years in profession, type of trials involved in, type of unit, affiliation, and university and their highest degree. Interviews were recorded using an audio recorder (encrypted) and were expected to last up to 60 minutes.

3.2.3 Data analysis

Interview recordings were transcribed verbatim whole to conduct a comprehensive analysis. Transcripts were analysed thematically and inductively by GC, using techniques of constant comparison. The first three transcripts were analysed separately by GC to avoid making any assumptions about potentially emerging themes. Following the first three interviews, GC met with an experienced qualitative researcher (DE) and consolidated several codes which captured the same idea into one.

Codes within transcripts were analytically summarised such that each code could be interpreted on its own [76]. This involved detailed coding. For example, emerging themes were compared with other codes across the dataset, to see if there were any shared or disparate views amongst particular subgroups [110]. These subgroups included methods leads, lead statisticians, trialists writing grant applications and NIHR leads and/or dependent on their role responsibilities. As some trialists came from the same trials unit, similarities and differences were also compared within trials units. If applicable, emerging themes were further sub-divided by type of characteristics [111]. Where applicable, deviant cases were described at the end of the main theme [96]. The coding was conducted using the qualitative data analysis software, NVivo (Version 11).

The initial coding was cross-checked by the qualitative researcher (DE) and discussed with GC, with inconsistencies resolved by discussion. GC met with another member of the study team (HJ) who is an expert in her field of evidence synthesis and trials, to double code the first transcript. This involved HJ highlighting the key parts of the transcript, making notes of any interesting points and summarising her key impressions of the data, with four to six points of the key findings. Although it was expected that different labels/codes would be assigned to parts of the transcript by GC and HJ, the overall meaning of the code should be the same, i.e. this is a check of the reliability of coding. The overall meaning and interpretation of codes were similar, and any minor disparity was discussed until there was consensus. For example, HJ analytically interpreted 'Strong impression that it would be hard to change practice – particularly to persuade people that using informative priors at the analysis stage would be beneficial." And GC had described this as "Would have to convince a lot people to use previous information, formulated as a prior to inform the current trial." Although the meaning is the same, we discussed how best to articulate this to ensure it was interpretable to someone with no previous knowledge in the field. The final code was "Felt it would be hard to change practice and show it was beneficial to use informative priors". In that time, we were happy that consistency in the interpretation of the coding was reached.

The iterative process of identifying emerging themes was continued until saturation was reached, i.e. until no new themes were emerging so that maximum variation within the sample had been achieved. The first 13 interviews were coded in full; with each interview analysed soon after it had taken place. After analysing the first 13 interviews, we hypothesised that maximum variation had been reached [110]. Therefore, the last three interviews were conducted to check that no new codes emerged which directly related to the key findings. We report our study according to the consolidated criteria for reporting qualitative research (COREQ) [112] and a summary table can be found in Table A.3.

3.3 Results

3.3.1 Participants

Among the 16 interviewees, three had a clinical background (two of which were Chief Investigators) and 13 had a statistics background. 25% (4/16) had greater than 10 years' experience working in trials. All had experience of working on RCTs. Most participants had experience of working on observational studies, whilst only some had experience of working on feasibility studies (including pilot studies). Interviews lasted an average of 54 minutes (range = 37 - 79 minutes). Table 3.1 provides participant and job-related characteristics.

Table 3.1: Participant and job-related characteristics

ID	Job title	Role description	Years in career	Types of trials	Highest qualification
P1	Trial statistician	Conducting analysis of trials in practice	3.5	RCTs, feasibility studies, primary & secondary care	MSc Statistics
P2	Senior statistician - Applied Health Research	Lead trial statistician & writing grant applications	6	RCTs, observational studies, surgery	MSc Statistics
P3	Professor of Haematology	Responsible for people planning/conducting the analysis of a clinical trial & lead of NIHR funded trials	>10	RCTs, FIH, translational studies	PhD, Clinical
P4	Associate Professor (Reader) of Population and Public Health Sciences	Responsible for people planning/conducting the analysis of a clinical trial & lead of NIHR funded trials	>10	RCTs, feasibility studies	PhD, Statistics
P5	Professor of Cardiac Surgery	Responsible for people planning/conducting the analysis of a clinical trial & lead of NIHR funded trials	>10	RCTs, FIH, translational studies	MD, Clinical
P6	Clinical PhD student	Conducting analysis of trials in practice	5	RCTs	PhD, Clinical
P7	Trial statistician	Conducting analysis of trials in practice	5	RCTs, feasibility studies	MSc Statistics
P8	Trial statistician	Planning & conducting the analysis of a clinical trial	3.5	RCTs, feasibility studies	PhD, Statistics
P9	Trial statistician	Conducting analysis of trials in practice	3.5	RCTs, feasibility studies, primary & secondary care	MSc Statistics
P10	Principle statistician	Responsible for people planning/conducting the analysis of a clinical trial	>10	RCTs, feasibility studies	MSc Statistics
P11	Senior statistician	Planning & conducting the analysis of a clinical trial	7	RCTs, feasibility studies	MSc Statistics
P12	Senior statistician	Conducting analysis of trials in practice	4	RCTs, feasibility studies	MSc Statistics
P13	NIHR Research Fellow	Conducting analysis of trials in practice	0.5	RCTs, feasibility studies, surgery	MSc Statistics
P14	Senior statistician	Planning & conducting the analysis of a clinical trial & writing grant applications	6.5	RCTs, feasibility studies, observational data/cohorts	MSc Statistics
P15	Head of Statistics	Planning & conducting the analysis of a clinical trial & writing grant applications	5	RCTs, observational data/cohorts	MSc Statistics
P16	Senior statistician, PhD student	Conducting analysis of trials in practice	5	RCTs, feasibility studies	MSc Statistics

RCT=Randomised Controlled Trial; NIHR=National Institute for Health Research; FIH=First-in-human. Locations were omitted to ensure anonymity.

3.3.2 Key findings

We first describe the overall current practice of evidence synthesis methods in trials. We then report the findings in four themes, displayed in Table 3.2. Each of the themes have their own subthemes, supported by quotations.

Current practice of evidence synthesis methods

Consistent with the findings of our INVEST survey, we found participants, across all trials units, were using existing evidence synthesis informally in a number of ways to inform the design of a new trial. Uses of evidence included the justification of the new trial, the choice of outcomes and parameters in sample size calculations.

The sourcing of previous evidence was often instigated by the clinician, who may forward or share an existing systematic review with the study team. The evidence, usually a systematic review, was typically being used to support the need for the trial and to demonstrate equipoise in order to get funding for a trial, rather than to inform the design.

> Ch inv, P5: "Obviously, there's the evidence of equipoise. There's the evidence of the knowledge gap."

> Ch inv, P3: "And unless you can convince a funder or ethics committee of equipoise to two treatment arms then you won't get, you can't do the study. There's got to be some sort of uncertainty."

Many participants followed previous statistical analysis plans (SAPs) within their unit, which were usually based on standard operating procedures (SOPs).

> Senior stat, P12: "We all tend to work quite collaboratively so I think we quite informally just run ideas past other people. I'm writing the SAP for [Study

> name omitted] which I know I should have my final version done already, but completely I was looking at other people's analysis"

We commonly found the statistical section in the protocol was written vaguely by the senior statistician or methods lead, rather than the trial statistician or chief investigator. The chief investigators' then had some input into the analysis plan, often with the trial statistician and the senior statistician, articulating the specifics of the research question.

> Ch inv, P3: "And I've already had input into the statistical analysis plan. . .to the extent that I've sat in front of the [senior statistician] and [trial statistician] I've tried to articulate what the research question is. And the [senior statistician] has translated into statistical language to explain to the [trial statistician]."

However, there was disparity between the two chief investigators' in our sample in terms of where responsibility ultimately lies, regarding which methods were used in the analysis. Specifically, in relation to the statistics section in the protocol and the SAP.

> Ch inv, P5: ". . . but I think it has to be set up by a statistician in terms of what would be my qualifications to sign it off as it were."

> Ch inv, P3: "The ultimate responsibility has to be with the chief investigator for all aspects of the study, that includes the analysis but obviously I'd be mad to argue with the statistics team!"

Themes

We now describe our findings in four themes.

Table 3.2: Table summarising themes (and subthemes), supported by a quote

Themes/Subthemes	Supporting quotes
Theme 1: Personal feelings	
1a: Favour simplicity and standard statistical methods	Princ stat, P10: “It’s generally in a way just the simplest technique that will get the job done and not overcomplicating it. Calculations get confused enough as it is!“
1b: Lack of confidence in Bayesian methods	Trial stat, P9: “It was a black-box moment of it went into the system and came out and I didn’t really know what had gone on in between [laughs]. Very bad statistician!”
1c: Relevance of prior data	Trial stat, P9:”If you’ve got a few bad pennies in there that’s going to make everything a bit skewed.”
1d: Lack of concern regarding bias adjustment	Methods lead, P4: “If you’ve got a significant result the last thing you want to do is talk it away!”
Theme 2: Perceived practical challenges of use	
2a: Access to data	Trial Stat, P8: “I suppose if there was, if there was consistency in the way the studies were reported and there was a way, a simple way of collecting all of the high-quality evidence together very quickly, then that would obviously be a big help but yes, I suppose that’s a bit of a pipe dream really.”
2b: Lack of roles/expertise within the team to identify and conduct a systematic review	Ch inv, P5: “One of the problems is there’s probably a shortage of systematic review capacity. So, finding systematic reviewers is really tough actually.”
2c: Aversion to software	Methods lead, P4: “I think a lot of it is accessibility of the software because Stata it’s just very straightforward, WinBUGS it’s not. So, I think that’s a massive hurdle. If you could do it in Stata people would probably do it.”
2d: Time and financial constraints	Ch inv, P3: “I’m trying to get funding for a study now to do this comparison, I can’t easily spend loads of money having a statistician spending ages trying to make a brilliantly efficient trial design...”
2e: No methodological guidelines	Trial stat, P8: “So, I think that would be a helpful if there was, I mean certainly if there was some sort of guidance that had been produced elsewhere.”
Theme 3: Concerns regarding acceptance of Bayesian methods in practice	
3a: Concerned trials team would not understand	Senior stat, P11: “I think actually clinicians and things are more familiar with the frequentist approach rather than Bayesian and actually it can be more difficult when you say, ‘I’ve used Bayesian methods’ and they think ‘Oh, what have you done?’”
3b: Harder to publish	Methods lead, P4: “Reviewers could be like ‘what on earth have you done? I’ve worked in trials all my life and I’ve never done this.”
Theme 4: Perceived impact of making use of existing evidence	
4a: Safety signals could be picked up faster	Ch inv, P3: “So, we do you make use of it but obviously in a suboptimal way and I can imagine that doing this kind of approach for adverse events for example would offer greater safety and allow safety signal to become obvious in my study earlier maybe so therefore better.”
4b: Powering a trial based on a meta-analysis could be more efficient	Senior stat, P1: “You’re not gonna sort of waste time and money showing an effect size in a single trial when you might be able to do it in a combination with existing studies. I think that’s quite sensible, but I guess it’s a case by case basis.”

Theme 1: Personal feelings

This theme summarises how trialists chose which methods to analyse their trials with, including their perceptions on using external evidence syntheses to inform these choices. We found trialists were using standard statistical models (such as linear, logistic regression), which they have always used. We also found they favoured simpler methods, rather than methods they perceived to be more complicated. Some felt they were happy with how they analysed trials in general. As such, participants felt that, if they were to bring in external information, they would need to know it was worthwhile. Many statisticians did not feel confident in using Bayesian methods. External evidence, by definition, belongs to someone else; a common finding with all participants was it can be difficult to trust such evidence, with additional concerns it could bias their own trial results. Surprisingly, trialists did not feel that their own trials were likely to be biased, even if some elements of the trial design could not be implemented (such as blinding the outcome assessor). Instead, they were happy to describe such trial design features as a limitation in the discussion, rather than using any formal bias adjustment.

Subtheme 1a: Favour simplicity and standard statistical methods

We found relatively standard statistical models were used to analyse trials. For example, logistic regression was used for binary outcomes, linear regression for continuous outcomes and Cox regression for survival outcomes. When trials had repeated outcomes or were part of a multicentre trial, many statisticians used mixed effect models to account for the hierarchical nature of the data.

> Senior stat, P2: “So, like logistic regression, linear regression. Erm mixed modelling, survival analysis, usually just Cox regression.”

Participants felt they wanted to analyse their trials using the simplest methods, and additionally using methods they had always used.

> Princ stat, P10: “It’s generally in a way just the simplest technique that will get the job done and not overcomplicating it. Calculations get confused enough as it is! “

Some participants also believed there was nothing wrong with current methods and would need to see a big enough motivating example to suggest otherwise.

> Clinical PhD student, P6: “I guess you’ve got to start off with what’s the problem with the current methods and is there a problem with the current statistical methods we’re using and what is the problem and how big is that problem and do we need to change our methods because of that problem. The current methods, analysis methods, seem to be pretty robust.”

Subtheme 1b: Lack of confidence in Bayesian methods

We found that statisticians (across all years of experience) had a negative perception of Bayesian methods. We found this was commonly a combination of generally not liking Bayesian methods and not feeling comfortable conducting such analyses. The latter of which could be attributed to a lack of confidence but also a lack of expertise if they have not been taught Bayesian methods.

> Senior stat, P11: “I personally don’t go anywhere near them. I think I did do a course in Bayesian stuff, but I just don’t think I work that way and I don’t feel comfortable using Bayesian methodology, so I personally would shy away from it.”

Possible reasons for not feeling confident in Bayesian methods could be not having the understanding or skill set through a lack of education or experience.

> Trial stat, P9: “It was a black-box moment of it went into the system and came out and I didn’t really know what had gone on in between [laughs]. Very bad statistician!”

Participants later in their career and more senior tended to have stronger opinions on the use of Bayesian methods, which were mainly negative when initially asked.

> Methods lead, P4: “I guess it’s not the standard thing to do and not what I’ve been taught - or still teach my students to do [laughter] – and generally don’t . . . unless I’m doing a mixed-treatment comparison I try, and steer clear of Bayesian methods [laughs] at all costs!”

> BRI007, Senior stat, P12: “I don’t know anything about Bayesian methods.”

Clinicians were equally rather hesitant to the use such statistics.

> Ch Inv, P5: “Are you creating some sort of Bayesian statistic? ... Yeah, so I don’t understand about that.”

> Ch Inv, P3: “I have to be honest with you. . . virtually zero.”

We therefore found trials were not analysed in a Bayesian framework and subsequently external information was not incorporated through informative priors. Most participants were not aware of colleagues using Bayesian methods either.

In relation to one of our three case studies: many participants did not want to do any formal bias adjustment (described in Subtheme 1d). A potential reason for this stemmed from a lack of confidence and understanding in meta-epidemiological methods.

> Trial stat, P8: “I’d need to have a much deeper understanding I think of how

> those adjustments work in order to do it.”

However, one statistician (a deviant case) was in favour of using bias adjustment, providing they were able to understand the method more. We found a lack of understanding in meta-epidemiological methods for bias adjustment amongst those who were more in favour of bias adjustment, and interpreting trial results in light of potential limitations.

> Trial stat, P9: “If I was confident in the methods... as you’re saying it right now, my head is blown by the idea, but if I knew the methods well enough, definitely, because it sounds like something that would... I’d somehow be able to take into account that impact.”

Subtheme 1c: Relevance of prior data

A lot of the concerns about formal use of existing evidence in a Bayesian framework surrounded trust in the data; from how different the population of the external data was to the trial population and whether such evidence would introduce bias into their trial. Many mentioned they were not sure how relevant external data are to their trial population.

> Senior stat, P11: “I am always a bit uncertain with meta-analysis about how you can group together different trials because they are different trials. They don’t use the same patient groups and there are different intricacies in there.”

> Trial stat, P9: ”If you’ve got a few bad pennies in there that’s going to make everything a bit skewed.”

Others felt this would also mean making extra assumptions above those made by the standard methods, particularly around how external data was collated, assessed for bias and its applicability to their trial.

> Methods lead, P4: “I think it opens a big can of worms that isn’t open if you just do a frequentist analysis like you’ve always done.”

Some participants also felt that a new trial should be analysed in isolation rather than being combined or mixed with existing data.

> Senior stat, P12: “I think I’m not by nature a Bayesian, so I would tend to be more on the no, let’s just go in fresh and see, especially if you’re doing a big trial and you can kind of say but yeah, all sorts of things have changed.”

There were also major concerns about how relevant the existing data in meta-epidemiological studies, or any study, is to tell them how biased their result could be. There were specific concerns expressed about how use of such studies would affect the interpretation of the trial results.

> Trial stat, P8: “I don’t know but whether it would plague the interpretation of the results making it more difficult for people reading it to understand what was actually going on. Yes, I’d be a bit cautious I think about that”

Subtheme 1d: Lack of concern regarding bias adjustment

We found that trials being at RoB from lack of blinding or inadequate allocation concealment was not a concern to most trialists. In general, most statisticians did not worry about whether a trial was biased.

> GC: “Yeah, so do you think if the patients are aware of the treatment that they’re receiving, do you think that could potentially bias the treatment effect?”

> Senior stat, P1: “Yes, I suppose it can do but I’ve never been particularly wor-

> ried about it."
>
> Senior stat, P12: "I think nobody in a trial is trying to be unblinded or is trying to cheat. I think you have to work from that but there's unconscious bias."

If a trialist thought their trial was at a potentially high RoB or there were potential design limitations, they often reported this as a limitation in the discussion with no quantification of how this may affect the results.

> Ch Inv, P3: "So it's absolutely critical to report that as a potential source of bias. And as far as I know all you can say is, is this potential source of bias means...well sometimes you can predict the potential direction of bias I guess but not always."
>
> Senior stat, P1: "I haven't done it, not thought about it or even something we discuss as being potentially biased, just see it as a limitation."

Many participants (including both chief investigators) believed they would not get funding for a trial unless they either ran a trial with all aspects of methodological components fulfilled, such as adequate allocation concealment etc, or they had a very good reason why it was not possible.

> Ch Inv, P3: "You have to have a compelling case and if it's pretty obvious that in your design when someone reviews your funding application that implicit in your trial there's going to be a significant chance of bias then they'll just say, 'sorry there's got to be a better way of doing it'."
>
> Senior stat, P2: "I guess in more recent, I guess quite often in trials I feel like they might not actually get accepted by funding bodies if they aren't able to blind people or they don't have an objective outcome and aren't blinded."

The aim of ‘meta-epidemiology’ is to learn about average bias across many trials and the amount of variability in that bias. However, this was something many participants had not heard of before, and many did not think it that bias adjustment of results was something they would do in practice. Even in a scenario where a trialist has got a significant result in their analysis which was subject to methodological limitations, bias adjustment is not a consideration.

> Methods lead, P4: “If you’ve got a significant result the last thing you want to do is talk it away!”

Theme 2: Perceived practical challenges of uses

Following on from theme 1, which looked more at an individual’s perception, theme 2 summarises the practical challenges participants felt they (and/or the wider trials unit) would face if they wanted to formally incorporate external evidence, in practice. One of the most common issues surrounded the logistics of accessing external data and the corresponding consideration of anonymisation. A systematic review was seen as one the most obvious ways to access multiple data sources. However, almost all of the trials units in this study did not have direct access to a systematic review team. Many participants viewed systematic reviewers as having a different skill set to most trialists. Many felt even if they had access to the data, they would then have to learn Bayesian software and therefore worried about the extra time and financial pressures this could have. It was frequently brought up that, in order for these methods to be used in practice, there would need to be guidelines and/or requirements to use such methods by funders.

Subtheme 2a: Access to data

The majority of participants mentioned there was no single repository whereby individual patient data is accessible. However, there is a shift in terms of trying to get individual patient data shared across institutions [113, 114].

> Trial stat, P8: “I suppose if there was, if there was consistency in the way the studies were reported and there was a way, a simple way of collecting all of the high-quality evidence together very quickly, then that would obviously be a big help but yes, I suppose that’s a bit of a pipe dream really.”

> Senior stat, P11: “I think I’ve heard it talked about the trial in a certain area you always collect certain variables and then those variables could be uploaded to a dataset and then it actually creates a big one. Everyone’s trial data gets compiled together and then you do have a big database that you could then use to inform sample size calculations and other things like that.”

Furthermore, more senior members recognised that access to a pooled repository of multiple data sources can bring its own challenges such as making sure data are anonymised.

> Princ stat, P10: “If it’s publicly funded you need to make the data available and that seems reasonable but there’s still always an administrative exercise in getting through approvals and getting that and for somebody to create a dataset that can be shared without risking identifiable data and stuff.”

Subtheme 2b: Lack of roles/expertise within the team to identify and conduct a systematic review

One of the main sources of existing evidence is a systematic review. However, the majority of trials units were not integrated with a systematic reviews team. As such, there was typically no individual responsible for identifying or performing systematic reviews within the clinical trials team.

> Ch Inv, P5: "One of the problems is there's probably a shortage of systematic review capacity. So, finding systematic reviewers is really tough actually."

In situations where systematic reviews were identified, they often came from the chief investigator to support the case for funding. Less senior members of trials units (such as trial statisticians) often relied on the clinician having used previous evidence to inform the design stage, rather than look for evidence themselves.

> Trial Stat, P8: "Yes so this is kind of... my only experience of using sort of raw data to inform the design of the study and on one case... in one case it was – there was a lot of quite, you know, high quality, detailed data that was sourced from one centre so it happened to be the one that the clinician was... where they were based and so then that was used in order to try and inform the parameters of the sample size calculations that we were doing and that was an example."

Subtheme 2c: Aversion to software

One of the most common issues that prevented trialists from analysing their trial in a Bayesian framework was Bayesian software. Although there are some Bayesian packages in Stata and SAS, most participants seemed unaware of this.

> Methods lead, P4: “I think a lot of it is accessibility of the software because Stata it’s just very straightforward, WinBUGS it’s not. So, I think that’s a massive hurdle. If you could do it in Stata people would probably do it.”

> Princ stat, P10: "It’s also about software and knowing how to implement it even if you wanted to.”

Some participants who had experience of using a wholly Bayesian software package, WinBUGS, were unsatisfied with its interface.

> Research fellow, P13: “Not a fan and not a fan of their messages. So many pitfalls because you could have those issues of non-convergence. . . WinBUGS, first of all it’s been developed by academics so they’re not really looking for profit, so you don’t have the same amount of time. You can see in the result of Bugs; the user interface is really poor.”

Subtheme 2d: Time and financial constraints

Several participants discussed how external pressures can impact the time they have to spend assessing the quality of external evidence and its relevance to their trial. Even when not explicitly incorporating external data, the majority did not find it easy to access it, summarise and justify its use in current practice.

> Methods lead, P4: “If we’re designing a trial that’s using something else, I can spend a long time thinking about how we justify and looking for papers to try and justify the sample size.”

More senior trial members, such as the chief investigator and methods leads, identified concerns regarding how much extra time would be needed to implement Bayesian methods, and the implicit costs associated with this.

> Ch Inv, P3: “I’m trying to get funding for a study now to do this comparison, I can’t easily spend loads of money having a statistician spending ages trying to make a brilliantly efficient trial design. That will obviously make my application better but it’s having some resource to do that first. Finding a trial where it works and then one that does – oh yes, we’ll use that – already spent tens of thousands of pounds on salaries and time of a statistician”

Trial statisticians also felt the pressures of not having enough time to think about using new methods and implementing them.

> Senior stat, P12: “I think the truth is as well I’m so busy at work and under such time pressures there’s times when I’ve not felt like what I’ve done has been wrong, but I’ve thought, do you know what, I haven’t got the time. I don’t feel I’ve got the time and it’s usually you’re forced to do it because somebody’s kicked up a fuss.”

Subtheme 2e: Lack of methodological guidelines

Traditionally, trialists use the Consolidated Standards of Reporting Trials (CONSORT) guidelines to determine what should and should not be reported. As more assumptions are made in a Bayesian analysis about the external evidence and model assumptions, some felt they would not know how to report such an analysis.

> Methods lead, P4: “Yeah, and I guess when you’re putting a trial together, you’d follow the CONSORT guidelines whereas there might be additional stuff that really should be in the paper that maybe you wouldn’t know to put. I don’t know. I wouldn’t be as clear what to include and what not to include”

To increase the uptake of these methods in practice, (in particular using external evidence to inform adverse event rates and sample size calculations) trialists felt methodological guidelines were needed with clear case studies.

> Trial stat, P8: “So, I think that would be a helpful if there was, I mean certainly if there was some sort of guidance that had been produced elsewhere.”
>
> Research fellow, P13: “So one thing I can say about Bayesian is you do find some material, but you don’t find a lot of ways to apply it really hands on”.

Theme 3: Concerns regarding acceptance of Bayesian methods in practice

The use of Bayesian methods, with its inherent summarising of prior information and combining this information with the results of the current trial, has long been debated in the literature. Theme 3 summarises concerns about such methods being accepted in practice. Many trialists think chief investigators, who ultimately sign off the analysis plan, would not understand Bayesian methods, and moreover, would not encourage their use. Since Bayesian analyses of trials are rarely seen in journals, some also thought publishing a Bayesian analysis could be problematic.

Subtheme 3a: Concern that trials team would not understand

Most statisticians were concerned about whether the trial team, and in particular the chief investigators and clinicians, would understand these methods.

> Senior stat, P11: “I think actually clinicians and things are more familiar with the frequentist approach rather than Bayesian and actually it can be more difficult when you say, ‘I’ve used Bayesian methods’ and they think ‘Oh, what have you done?’”

> Senior stat, P1: “They don’t understand it [chief investigator] I think we’re getting to the point where people generally understand a treatment effect and a confidence interval and that’s fine. If we start adding all these other things in, they’re going to be like okay, too much variability, why is there so much variability?”

When the clinicians were asked about Bayesian methods, their initial response did match the statisticians’ perceptions.

> Ch Inv, P3: “This is kind of universes apart from what most clinicians would understand. It’s totally different and I’ve got colleagues who do RCTs and compare treatment A vs B and don’t give this a second thought.”

> Ch Inv, P5: “Are you creating some sort of Bayesian statistic? ... Yeah, so I don’t understand about that.”

Subtheme 3b: Harder to publish

Some senior statisticians (and in particular methods leads) were concerned that if Bayesian methods were used, it would make it much harder to publish their trials.

> Methods lead, P4: “Whether it would be accepted by decision makers. I guess if you’re doing anything that’s not the norm, you’d just be a bit scared, even getting it published. Reviewers could be like ‘what on earth have you done? I’ve worked in trials all my life and I’ve never done this.”

Some also thought there could be an issue with ethics, which could be a potential barrier to these methods being used in practice.

> Trial stat, P9: “I think it’s an interesting idea. I don’t know how you’d get the, I

> guess the approval for doing such a thing, like ethics, I can imagine you might have an issue with."

Theme 4: Perceived impact of making use of existing evidence

We have seen the potential barriers in themes 1-3, but there were also lots of enthusiasm about making more use of existing data from many trialists. Many participants felt they could make much more use of existing data, in particular aspects of trials. These aspects included making more use of existing safety data, so that rare events could be picked up faster, and powering a trial to update the existing evidence base to have a bigger impact on changes to practice/policy.

Other participants, who were initially hesitant, could also appreciate the value that a Bayesian framework allows for the incorporation of previous evidence. They also saw a possible ethical impact of evidence by potentially influencing the initiation of new trials.

> Senior stat, P12: "Well it's the whole ethical thing isn't it of doing further trials when you already know the answer or it's just putting it in the wider context isn't it so yeah, it's sort of downgrading or upgrading whatever you've found in the wider context."

Furthermore, many participants, acknowledged they were already using existing evidence, but not in a Bayesian framework.

> Senior stat, P12: "I suppose it's something that people do informally but not in a structured Bayesian [way]. I think that's what's true."

Many thought that making more use of existing data was advantageous, as a lot of time and money is invested in trials, for the information collected not to be used again.

> Senior stat, P1: "We don't want to do a trial that wastes a) time and b) money so if we had existing evidence which would cut down time and money then I think we should do it to start with."

Subtheme 4a: Safety signals could be picked up faster

Participants from all trials units remarked that adverse events were typically summarised descriptively because the control arm did not have an adequate response rate i.e. events were rare. This meant, due to low event rates, the response rate in the experimental arm could not be statistically compared to the response rate in the control arm. Most trials units were therefore led by the clinician or chief investigator to list the adverse events and then rely on their experience to identify a potential safety concern. There was also no mention of using specific expected rates from external evidence to inform trialists of the population adverse event rate if there have been lots of other studies looking at the same control intervention.

This generally involved listing the expected and unexpected events, based on a chief investigator's personal experience, with no mention to specific expected rates based on any external evidence, such as similar studies.

> Senior stat, P2: "So, I think you always have the rule that if there was less than 10 events, then we just present them as counts, so present how many occurred in each group and not do any formal analysis."

> Trial stat, P9: "Me personally, I only get given a list... I hope that when I create the table of expected events and then I say, 'Oh, five people experienced this,' that the clinicians within the trial team would then say, 'Ooh, that's a bit high.'"

Most participants acknowledged that describing adverse events and relying on the clinician or data monitoring committee (DMC) to identify anything unexpected, may not be the best way of doing it.

> Princ stat, P10: "I rely on the DMC quite a lot basically. I don't think we've got good methods for looking at adverse event rates really. It's often just listings or tabulations."

When asked about their views on using existing data in a similar population to predict what the expected adverse event rate would be in the control arm, many thought this was a good idea. Clinicians and senior statisticians thought this would allow safety signals to be picked up faster.

> Ch Inv, P3: "So, we do make use of it but obviously in a suboptimal way and I can imagine that doing this kind of approach for adverse events for example would offer greater safety would allow safety signal to become obvious in my study earlier maybe so therefore better."

> Trial stat, P9: "No, I never have, but now you've said it, it seems like such an obvious tool, it should exist [laughs]."

> Senior stat, P12: "I think if I'm totally honest I don't think it had occurred to me to think about doing it and now I can completely see the motivation for doing it. It's finding the time and prioritising as well"

Subtheme 4b: Powering a trial based on a meta-analysis could be more efficient

The concept of using an existing meta-analysis to power a new trial, based on its ability to impact an existing meta-analysis, was unfamiliar to all participants. Traditionally, the sample size for a new trial is calculated without explicitly incorporating existing evidence. In current practice existing evidence from the literature might instead be used to provide information on the likely event rate in the control group or the between-person standard deviation in outcome measurements. The key assumption here is they are not based on the idea that the trial will be analysed or interpreted together with others. However, the evidence base in its entirety is commonly what impacts on policy and/or clinical practice. Therefore, the meta-analysis rather than the trial might be considered to be what is of most interest. Multiple researchers, primarily methodologists [6, 73, 74] have suggested it may be beneficial to conduct a meta-analysis of previous evidence and work out what a new trial would need to show in order to demonstrate an intervention is effective. Having briefly explained to participants that it is possible to power a new trial based on an existing meta-analysis, the majority saw this as an advantage. Many thought it was a very attractive idea and could make the trial more efficient. Many also recognised that they did have some idea about the potential effect from a previous meta-analysis.

> Senior stat, P1: "You're not gonna sort of waste time and money showing an effect size in a single trial when you might be able to do it in a combination with existing studies. I think that's quite sensible, but I guess it's a case by case basis."

> Methods lead, P4: "The body of evidence [meta-analysis] is going to change practice. I guess then I can see that taking into account prior evidence might be a good thing to do."

There were, however, reservations about powering trials based on an existing meta-analysis. Some participants felt they would only do this as part of a sensitivity analysis to see how much the sample size would differ.

> Methods lead, P4: "Equally I'm not sure. Still I think I'd prefer just to do a meta-analysis afterwards I think."

> Princ stat, P10: "As I say, I think I'd be a bit reluctant that that was my main

> analysis or that that was the only analysis, but I think as part of a sensitivity analysis or something that's probably a good way of looking at it."

Others thought an integral part of the trial results was to publish them in isolation, without any form of Bayesian analysis or statistical incorporation of existing evidence.

> Ch Inv, P6: "I think it will still, at least for the participants and all the staff involved, we still deserve to have the results published in isolation as that individual trial".

Others thought there could be potential to recruit fewer patients, which would also make the trial cheaper.

> Ch Inv, P3: "...because that means you can deliver a much bigger study or a more cheaper study of more patients faster and that's better for everyone."

> Trial stat, P8: "I think it's quite sensible probably to power it based on you know making a change to that if it means that you, you know, you're gonna recruit less participants."

3.4 Discussion

3.4.1 Overview of key findings

A key finding of this qualitative study was that Bayesian methods, and therefore informative priors, were not used in practice by the interviewees. Most were not aware of colleagues using them either. Many barriers to explicitly using external data in trials were practical, but there was also an important concern that these methods would not be accepted by researchers in the field, or ethics boards.

One of the key practical concerns was the lack of infrastructure most trials units have to access external sources of data. In particular, many of the trials units in this study did not have an integrated systematic review team. Participants also raised the practical issue that it is not easy to use Bayesian software. Inherently, there could be a steep learning curve in implementing these methods in practice. This may include the extra work it would take to synthesise external prior data and subsequently model it; which intrinsically has time and financial burdens.

The concerns about the lack of acceptance of Bayesian methods in trials arose from its negative perceptions amongst researchers and the wider trials community such as editors and ethics boards. There was an additional fear that it would be harder to publish these methods within the clinical trials community. Because of these perceived issues, we found trialists had an internal discomfort about formally incorporating external data with their own trial data. This included a lack of confidence to implement Bayesian methods, concerns over relevance of the prior data and favouring simpler, standard methods.

Although there were many concerns, participants still thought they could be making more use of existing data. Participants felt there could be positive clinical implications for informing safety decisions. Rather than relying on what the clinicians says (which is often based on their personal experience), the adverse event rate would be informed by previous studies and ideally an evidence based synthesis. Although explicitly using an existing meta-analysis to inform a new trial by actually powering the trial based on its ability to impact on the meta-analysis was unfamiliar to all participants, many thought it was a good idea. A potential advantage was that it could be possible to recruit fewer patients, making the trial cheaper. For some, this raised concerns about ethics, but others thought the possibility to recruit fewer people was an advantage. Many also said it was multiple trials, rather than an individual trial, that changes policies. If these methods were to be used routinely in practice, there were calls for methodological guidelines [115].

3.4.2 Reflexivity

In qualitative research, the results are inherently and indirectly influenced by the views of the researcher who conducts the study [96]. In this study, part of the sample we targeted were trial statisticians. Having worked as a trial statistician for two years before starting my PhD, it is possible my own views could have influenced the results. In the first two interviews of statisticians, I felt they did not elaborate on some statistical methods, potentially knowing that I realised what they meant. For this reason, I tried to keep the questions as open ended as possible and used probing questions to ask what they meant by certain things to gain a deeper understanding. When meeting participants, one of the first questions they would ask was about my PhD (closely related to the aims of the study) which could force my views onto them. I therefore decided not to tell participants about my PhD at the start, so that their views (positive or negative) were completely their own.

Interviewing people for the first time was strange in the beginning as I wanted to agree with what they were saying and/or give my opinion. I had tried to prepare myself for this, but it was a lot more difficult in reality. After the first few interviews, I tried to speak less and appear that I was listening without giving any encouragement. Although, initially I found silences daunting, I learnt these periods meant participants were thinking and they often then gave more elaborate responses. Feeling more comfortable with silences and asking statisticians to elaborate on methods, allowed the data to be richer and for participants to elaborate on points without prompts.

3.4.3 Methodological strengths

Interviews were inductive rather than deductive; this was achieved by keeping the questions open ended. The topic guide was adapted when new themes were emerging in order to explore potential new areas which we had not thought of. Therefore, the flexibility in the topic guide allowed new themes to develop.

We did not use the term 'Bayesian' in the information sheet, instead referring to 'existing evidence' (see Section 3.2 for more detail). By not using 'Bayesian' at any point prior to the interview, this meant participants did not have any preconceived ideas before the interview. It also helped ensure a range of participants were in the sample, making it as diverse as possible.

3.4.4 Limitations

In any qualitative study, it is possible that maximum variation within the sample was not reached [76]. In our sample, only three clinicians were included. This was in part due to the individuals who we wanted to recruit, as methods leads and NIHR leads were more likely to have a statistical background rather than a clinical background.

Snowball sampling was used to identify new participants after initial key contacts in each group were sampled from colleagues known to the study team. This has the potential to bias the final sample by affecting how diverse the participants are in general and across potential subgroups. However, following snowball and purposeful sampling, we sampled from each group of our intended population. There was also only one person who said they would like to take part but then could not find time. Furthermore, no one dropped out, which reduces potential bias in the final sample of participants.

As the term 'Bayesian' was not used in the participant information sheet, it is possible that we missed people who had a potentially strong view on such methods. However, this term was purposely avoided as we were trying to get a range of participants without any preconceived views. It is also a finding in itself that there were numerous trialists, in multiple locations, who were not using Bayesian methods in trials units. Furthermore, the potential barriers to these methods were explored in further detail and, again, form a major part of the findings.

3.4.5 Relevance of this study in relation to other studies

Unsurprisingly, one of the key findings of the study related to the ‘relevance of prior information’ which is one of the most discussed areas of Bayesian analysis [116]. This key finding is also consistent with our INVEST survey [78] which found two of the top three barriers were ‘First trial in area’ and ‘Previous trials different’. In the INVEST survey, we also found ‘time constraints’ was the biggest barrier to the use of existing evidence across both the design and analysis of trials. Although we did not identify time constraints as an overarching theme in this study, this issue was implied as one of the practical challenges to using external evidence in trials in Theme 2. A more detailed exploration in our study revealed the extra time needed to conduct a systematic review (given that systematic review teams are often not integrated into clinical trial teams) was a concern. We also found trialists found it difficult to access and collate other external data, either aggregate or IPD. Our study found there was an additional time aspect for statisticians to learn new methods and software, such as WinBUGS.

Lilford [117] argues the assumption of equipoise in RCTs is misleading to the patients being invited to participate in a new trial. This is particularly so because, more often than not, some evidence exists before an RCT either on similar treatments in the same disease area or the same treatments in other disease areas. Lilford therefore contends that the wording of equipoise in patient leaflets, as suggested by Donovan *et al* [118], is misleading, because it is quite likely that we do know *something* about different subgroups from previous ‘similar’ studies. The finding of our qualitative study is consistent with this: trialists recognised that they did have some idea about the potential effect from a previous meta-analysis and it is the accumulation of evidence that is likely to change practice:

> Methods lead, P4: “The body of evidence [meta-analysis] is going to change practice. I guess then I can see that taking into account prior evidence might be a good thing to do.”

However, they did not explicitly incorporate this information into the trial design or analysis.

Brocklehurst *et al* [30] describe their experience of stopping a large trial because of emerging external evidence. The authors concluded that it remains unclear how trial investigators should consider external evidence during a trial and subsequently make decisions regarding whether future recruitment should continue, stop, or be reduced. Our study appears consistent with Brocklehurst *et al* [30] in the overarching finding that it remains unclear to trialists (including investigators) the process by which external evidence should be considered, and at precisely which stages of a trial. For example, when given the hypothetical scenario of using external evidence to inform adverse event rates, it was unclear to our participants where the data would come from, how it would be synthesised, how to determine its relevance to their trial, and how to incorporate it statistically.

3.4.6 Implications for future research

The main implications based on the key findings are displayed in Figure 3.1.

Our study showed that trialists favour simplicity in their analyses. We therefore need real world examples or case studies to convince trialists these methods are worth the extra time to invest in learning and implementing them in potentially new software. Many issues were related to the fact that WinBUGS is the main tool for Bayesian analysis whereas most trials units use Stata or SAS. This is a problem, but more Bayesian techniques are being implemented in Stata. Furthermore, most funders and the Medicines and Healthcare products Regulatory Agency require software to be validated, which calls for recommen-

Figure 3.1: Implications for further research based on key findings

Implication 1: Real world examples or case studies are needed to motivate scenarios in which external data might be used in trial analyses: specifically, for (i) safety signals and (ii) powering a trial based on an existing meta-analysis.

Implication 2: Analyses that incorporate external evidence need it to be implemented in software that is commonly used and validated.

Implication 3: Methodological guidelines are needed in order to guide statisticians as to the necessary steps one should take and consider in incorporating external evidence into trial analyses.

Implication 4: There is a need to improve access to existing data that might inform trial analyses; and for further encouragement to trials units to make their data available in databases or data platforms.

dations to build upon existing Bayesian software in Stata. Ultimately, real word examples and case studies showing potential gains over current methods in a more user-friendly software should be developed.

The use of informative priors in sample size calculations and using evidence synthesis methods for rare events was perceived as attractive. It was these aspects of a trial where most trialists thought improvements could be made to current methods. One of the potential reasons external evidence has not been used to inform adverse event rates routinely is that there is less of a concern about safety in phase III trials, as opposed to early phase trials. In cases where event rates are low in the control arm is inadequate, it may be useful to use evidence-based models rather than relying on clinicians for these rates.

In relation to sample size calculations, it may not always be appropriate or possible to power a trial based on an existing meta-analysis, but the option could be explored during the design stage. If it is not possible, for example because the intervention is sufficiently different, or the research question is different, then this should be made clear. This may call for transparency on *how* evidence synthesis has been used.

We found that statisticians across all levels do not feel confident applying Bayesian methods; and in particular the extra assumptions that have to be made regarding the relevance of prior information. This implies that specific methodological guidelines are needed to guide statisticians as to the necessary steps one should take and consider [100]. The CONSORT guidelines provide guidance on what and how aspects of trials should be reported in a trial report [119, 120]. Some participants suggested they would be more likely to include external evidence in trials if they had methodological and reporting guidelines to follow, or if funders asked for it. Many participants were also concerned as to whether Bayesian methods would be accepted in practice by funders and decision makers, as well as other colleagues and clinicians. We believe by creating guidelines, trialists and decision makers would feel more confident using informative priors, particularly in situations where existing evidence could be advantageous over current practice.

Another major concern is that the infrastructure is not in place. Many trials units are not integrated with a systematic review team. Second, there is not an easily accessible repository of relevant data, ideally individual patient data. Although there have been many calls in the UK for a platform of all individual trial data, it is probably still a while off [121]. As such, we still need to first improve access to existing data so that it can be incorporated. A possible solution to bridge this gap is for each trials unit to synthesise each of their own trials to learn about adverse event rates and predict the probability of a patient with particular characteristics having an event. This could then be used to look at potential sample size parameter assumptions and to predict adverse events. Furthermore, most trials from the same trials unit have a greater chance of being more similar in terms of population and setting. For example, a trials unit that specialises in cardiovascular trials is more likely to run trials in similar populations.

3.5 Conclusions and implications for remainder of thesis

The results of the qualitative study highlighted that trialists felt they could be making more use of existing data to inform the design and analysis of a clinical trial in particular scenarios. We found that trialists are using existing evidence in a lot of ways; ranging from using it to justify that there is a gap in the evidence base (such as using a systematic review to show there is an unanswered clinical question) or using it to inform parameters in the sample size calculations (such as the expected control group event rate). However, trialists rarely (never in our sample) explicitly combine this previous evidence statistically with their actual trial through a Bayesian analysis. This was consistent with the results from our INVEST survey. We found the main reasons a Bayesian analysis was not conducted could be categorised into three main areas (i) personal feelings of a trialist (lack of confidence in Bayesian methods and relevance of the data); (ii) perceived practical challenges of use (hard to access data, anonymisation issues, Bayesian software) and (iii) concerns about lack of acceptance in the field (negative perceptions of Bayesian methods).

We found trialists do not think about bias adjustment, partly because they were not aware of meta-epidemiological methods. In Chapters 4 and 5, we extend current meta- epidemiological methodology, looking specifically at how a trialist might use such evidence for bias adjustment in their trial. Second, although trialists were unaware of explicitly incorporating information from an existing meta-analysis into sample size calculations, many thought it was a good idea. In Chapter 6, a comparison of methods is undertaken to discuss the advantages and limitations of incorporating external evidence in such calculations in practice. We also found participants felt current methods for determining how likely adverse events were a safety concern could be improved. In particular, participants felt they could use existing data to inform adverse event rates when the event rate in the control arm is not adequately powered. In Chapter 7, we develop methods which use a synthesis of control data to inform the likely event rate in the control arm of a new study.

In Chapters 5, 6, and 7, specific attention will be drawn to how existing data could be pooled and summarised and used by a trialist in practice.

4 A meta-epidemiological investigation of the impact of blinding on estimated treatment effects in randomised clinical trials: the MetaBLIND study

The MetaBLIND study was a collaboration with the following researchers: Helene Moustgaard (HM), Asbjørn Hróbjartsson (AH), Hayley Jones (HJ), Julian Higgins (JH) Jelena Savović (JS), Jonathan Sterne (JS) and the wider MetaBLIND team Phillippe Ravaud (PR), Isabelle Boutron (IB), Lars Jørgensen (LJ), David Laursen (DL), Mette Frahm Olsen (MFO) and Asger Paludan-Müller (APM). My role has included management of the dataset, performing all statistical analyses, including extending current methods and drafting some of the written outputs.

4.1 Context and overview

We know RCTs offer one of the best sources of evidence to answer a specific research question [2]. However, RCTs may still have their limitations, despite best efforts by the trial team [21, 63] (see Section 1.2.2 for examples of the different types of biases). Trial analyses can adjust for biases in an RCT context, in the same way this has been proposed for the meta-analysis context [55]. Meta-epidemiological studies provide empirical evidence for such adjustments. We can then use this information about such biases to adjust the treatment effect estimate in a new study, allowing the analyst to assess the sensitivity of their findings. In an example where it was not possible to blind a particular party, this would enable the trial team to answer, "what result would we have seen if we were able to blind the study". An opportunity arose for me to get involved in a major new meta-epidemiological study, MetaBLIND. I, therefore, describe the study and some novel methodological developments I made. This is the first of two chapters on using meta-epidemiological studies to adjust for bias in an RCT.

4.2 Introduction and aims

Meta-epidemiology is a recognised method to quantify the difference, on average, in treatment effects between trials which differ by a characteristic or set of characteristics. The first meta-epidemiological study [20] analysed a set of meta-analyses that included similar trials, to investigate the impact of several methodological factors that could potentially bias estimated treatment effects. Schulz *et al* reported that concealment of allocation was very important, with treatment effects exaggerated by 41% (95% CI 27% to 51%) on average among trials without relative to trials with adequate allocation concealment, whilst double-blinding was less important, with treatment effects exaggerated by 17% (95% CI 4% to 29%) on average. Since that pivotal study a number of similar analyses have been conducted, with inconsistent results [122, 123].

In 2012, a combined re-analysis of seven meta-epidemiological studies reported that, on average, odds ratios were exaggerated by 13% (95% CI 4% to 21%) in trials without double-blinding and this was greater in trials with subjectively measured outcomes, 22% (95% CI 8% to 35%) [64]. However, in both types of outcomes there was evidence of heterogeneity in the average bias across meta-analyses. A key limitation of these previous meta-epidemiological studies arises from the ambiguity of "double blind": it is unclear from this term precisely which parties were blinded [79, 124]. Previous studies have relied predominantly on the labelling of trials as "double blind" in published trial reports and not attempted to access information on actual trial conduct. Importantly, the comparison between "double-blind" trials and not "double-blind" trials has not enabled a separation of the impact of lack of blinding of outcomes assessors and the impact of lack of blinding of patients and/or healthcare providers. Previous studies have also not taken into account whether the outcomes were assessed and reported by patients or by other observers. Overall, these conceptual and methodological limitations may, at least partially, explain the inconsistent results.

We therefore describe the first meta-epidemiological study aimed at disentangling the impact of different types of blinding to enable the clear separation of the two main types of blinding-related bias: ′performance bias′ and ′detection bias′ (described in Section 1.2.2). We examine separately the impact of blinding participants, healthcare providers and outcome assessors, and examine how this might vary by type of outcome. We analysed the data using a hierarchical bias model proposed by Welton *et al* [55], which has been previously applied in meta-epidemiological research [64, 81] and is described in the next section. Our study also introduces the following methodological novelties. Previous studies have modelled binary and continuous data separately. We therefore model continuous and binary data simultaneously in a single model, assuming a mixture of normal and binomial likelihoods but modelling the underlying bias on the same scale. We also extend this model to incorporate covariates, in particular, to explore the association between degree of subjectivity of the outcome and the average magnitude of bias.

The five main analyses investigate:

- (Ia) the effect of blinding patients to the treatment they are receiving, in outcomes which are assessed by the patient (patient reported outcomes) [performance bias and detection bias];
- (Ib) the effect of blinding patients where the person assessing the outcome is blinded to the treatment the patient has received [performance bias];
- (IIa) the effect of blinding those providing care or those making a decision on a treatment for patient, in outcomes which are assessed by the same person (i.e. those which gave them a particular treatment, defined as healthcare provider decision out-

comes) [performance bias and detection bias];

- (IIb) the effect of blinding those providing care or those making a decision on a treatment for patients, in outcomes where the person assessing the outcome is blinded to the treatment the patient has received [performance bias];
- (III) the effect of blinding the outcome assessor, in outcomes which are subjectively assessed (by the same person) [detection bias].

The chapter is structured as follows. Section 4.3 give the background to meta-epidemiology and current methods. Section 4.4 details the methods applied to MetaBLIND including the extension of current statistical methodology. The chapter concludes with a discussion of the findings and the implications for using the results of meta-epidemiological studies to adjust for biases in a new trial.

4.3 Background to meta-epidemiological methods

To assess the RoB of methodological components in a trial, such as allocation concealment and blinding, it is common to use a RoB tool. This assessment is often conducted as part of a systematic review and, when applicable, to quantitatively synthesise effect estimates from studies, in a meta-analysis. The Cochrane Risk of Bias tool is one way of carrying this out. The tool has individual methodological components set within several domains. These domains include 'selection', 'performance', 'detection', 'attrition', 'reporting' and any other bias which is not covered in those domains. Example of how these biases can occur in trials are given in 1.2.2. Each individual methodological component is classified as being either at a 'high RoB', 'low RoB' or 'unclear RoB'.

Since 2012, the Cochrane RoB tool has included separate components for different types of blinding: 'Double-blinding' has been replaced with 'Blinding (participants and personnel)′ and ′Blinding (outcome assessment)′. These now fall under the ′performance′ and ′detection′ bias domains. We therefore use this information to inform the assessment of blinding status for each trial. A ′low RoB′ trial is defined as a trial with the methodological component, i.e. with blinding of the outcome assessor. A ′high RoB′ trial is a trial without the methodological component or one in which it is unclear.

There are different methods used in meta-epidemiology to estimate the average amount of bias attributed to a specific methodological limitation. One of the first studies was by Schulz *et al* [20] who wanted to see if methodological limitations were associated with evidence of bias in estimating treatment effects. He used logistic regression to model the association of treatment effects between studies at a high RoB and studies at a low RoB, for a specific methodological flaw. Thus, assuming the underlying association was fixed or constant across studies.

Sterne *et al* extended this method by allowing the average bias to vary across meta-analyses in an intuitive two stage approach. In the first stage, the amount of bias attributed to a specific methodological limitation is estimated within each meta-analysis. This is quantified using ratio of odds ratios (RORs), $\mathrm{ROR} = \frac{\mathrm{OR}_{\text{high-risk}}}{\mathrm{OR}_{\text{low-risk}}}$, based on comparing the summary odds ratio from studies without the study characteristic of interest with the summary odds ratio from studies with the characteristic. In the second stage, the amount of bias is averaged across meta-analyses to get the average ROR. The between study variability in the average bias is denoted by φ^2.

Welton *et al* hypothesised that it is likely trials will have been influenced differently by the effects of blinding or any flawed study characteristic [125]. Therefore, they developed a model that allows the bias between trials and additionally the bias within meta-analyses to vary and implemented this in a one stage Bayesian framework. We first outline this hierarchical model, as described by Welton *et al* model. We build upon this model to combine binary and continuous outcomes and stratify the average magnitude of bias and degree of subjectivity in observer reported outcomes.

4.3.1 Welton *et al* model

The Welton *et al* model allows for the treatment effect to vary across studies, the average amount of bias across meta-analyses to vary and additionally the study specific bias across trials to vary in a one stage approach. It therefore assumes biases are broadly similar (exchangeability assumption) within a meta-analysis, and assumes the average bias is broadly similar (exchangeability assumption) across meta-analyses.

The outcome $r_{a,i,m}$ for arm a of trial i in meta-analysis m is assumed to have a binomial likelihood (for given denominator $n_{a,i,m}$):

$$r_{a,i,m} \sim \text{binomial}\left(p_{a,i,m}, n_{a,i,m}\right) \tag{4.1}$$

The probability of success, $p_{a,i,m}$, is modelled by a logistic regression:

$$logit(p_{a,i,m}) = \begin{cases} \mu_{i,m} & \text{control arm} \\ \mu_{i,m} + \delta_{i,m} + \beta_{i,m}C_{i,m} & \text{treatment arm} \end{cases} \tag{4.2}$$

$\mu_{i,m}$ is the log-odds of success in the control arm. $\delta_{i,m}$ is the treatment effect, on the log odds ratio scale, in each study. We assume each $\delta_{i,m}$ is normally distributed with mean treatment effect d_m and variance τ_m^2.

$$\delta_{i,m} \sim N\left(d_m, \tau_m^2\right) \tag{4.3}$$

$$C_{i,m} = \begin{cases} 1 & \text{if study } i \text{ at a high or unclear RoB} \\ 0 & \text{if study } i \text{ not at RoB} \end{cases} \tag{4.4}$$

$C_{i,m}$ indicates whether the methodological limitation is present, for example, if the outcome assessor is blinded ($C_{i,m} = 1$, non-blinded; $C_{i,m} = 0$, blinded). $\beta_{i,m}$ is the bias in treatment effect in study i of meta-analysis m. A hierarchical model is put on the study-specific biases, $\beta_{i,m}$, that capture the nature of the empirical evidence that is available to inform these parameters:

$$\beta_{i,m} \sim N\left(b_m, \kappa^2\right) \tag{4.5}$$

$$b_m \sim N\left(b_0, \varphi^2\right) \tag{4.6}$$

The ROR is given by $exp(b_0)$. κ^2 is the average increase in between-trial heterogeneity among studies with, relative to those without, the characteristic. φ^2 is the between meta-analysis variability in mean bias, that is, the variation in average bias across meta-analyses. We note that between study heterogeneity, τ_m, is estimated separately for each meta-analysis, whereas κ is shared across all meta-analyses.

To study the impact of potential confounding factors, such as whether or not patients were blinded, on the difference in intervention effects, we include binary covariates at the trial level. Using the same notation, we add the term $b_1 x_{i,m}$ to equation (4.2):

$$logit(p_{a,i,m}) = \begin{cases} \mu_{i,m} & \text{control arm} \\ \mu_{i,m} + \delta_{i,m} + \beta_{i,m} C_{i,m} + b_1 x_{i,m} & \text{treatment arm} \end{cases} \tag{4.7}$$

$x_{i,m}$ is the value of the covariate in the data for each trial within each mth meta-analysis and b_1 is the regression coefficient of the trial level covariate.

All of the models presented are implemented and carried out using WinBUGS version 1.4.3 (MRC Biostatistics Unit, Cambridge, UK [126]). For location parameters (overall mean bias, baseline risk, treatment effects), Normal (0, 1000) priors were assumed. Vague priors were assumed with a modified Inverse Gamma (0.001, 0.001) prior on all variance components to allow increased weight on small values. This was chosen from the earlier

BRANDO analysis by Savovic *et al* [127] who found this prior to perform the best (with the lowest average mean squared error) having conducted a simulation study. It is well known with this type of modelling that variance components can be sensitive to the prior distributions [128].

Meta-analyses with only one high risk or only one low risk trial are prevented from contributing to the estimation of κ. This is implemented using the 'cut' function in WinBUGS, originally applied in the BRANDO analysis [64]. In situations where there is only one high or low RoB trial, it is not possible to estimate κ. When there is only one high RoB trial, it does not make sense to estimate the variability in bias between the high RoB trials. When there is only one low RoB trial, it is not possible to estimate both between study heterogeneity and κ. We apply this to all of the following models with a binary study characteristic.

For meta-analyses with few studies, the between study heterogeneity, τ_m, can be imprecisely estimated. We therefore conducted sensitivity analyses with a hierarchical model on this parameter for each of the five main analyses. We took the log of the between study standard deviation from each meta-analysis, which we assumed was normally distributed, thereby allowing us to borrow strength from the meta-analyses which estimated the between study heterogeneity more precisely. Although we may expect the estimates for τ_m to change, we checked our parameters of interest (b_0, κ, φ) did not change. For all analyses, 2 parallel chains were run, with a burn-in of 250,000 iterations followed by at least a further 1,000,000 iterations, with a thinning of 5. We assessed convergence by checking the agreement of each of the chains in history plots and density plots. 95% CrIs are provided with each parameter estimate.

4.4 Methods

4.4.1 Screening and data extraction

1042 Cochrane reviews (all reviews published or updated between 01/02/2013 (Cochrane Library Issue 2 2013) and 18/02/2014) were screened for informative meta-analyses. We define a meta-analysis as ′informative′ if it includes at least one contrast of the type of blinding status. For example, a meta-analysis containing at least one trial with blinded patients and one trial with non-blinded patients would be ′informative′ for patient blinding. The meta-analysis had to be informative for at least one of the blinding contrasts (patient, healthcare provider or outcome assessor blinding) to get through the screening stage. Since a Cochrane review often contains multiple meta-analyses, the first meta-analysis was checked for informativeness, based on the criteria above. If it was not informative, the second meta-analysis (or analysis) in the review was checked and so on, until an informative meta-analysis was identified. Once an informative meta-analysis was identified, subsequent meta-analyses were not checked. The Cochrane RoB tool was used initially as an assessment of blinding status to check informativeness.

Outcome measures were classified as observer-reported, patient-reported (via interviewer or directly reported), healthcare provider decision outcomes or as mixed (in cases where the outcome was a mixture of more than one category, e.g. both patient and observer-reported elements). The screening process identified 395 potentially informative meta-analyses, of which 226 were potentially informative in relation to blinding of outcome assessors and 169 in relation to blinding of patients or healthcare providers. For pragmatic reasons, we continued with a random subsample of 120 of the 226 meta-analyses that were potentially informative in relation to blinding of outcome assessors and all 169 meta-analyses in relation to blinding of patients or healthcare providers, i.e. 289 meta-analyses in total. Figure 4.1 shows the flow of data through the study, from contributing studies to final datasets.

Trial characteristics and blinding information were extracted manually from trial publications. Trial results were extracted automatically from the Cochrane Database of Systematic Reviews via the Archie database interface: number of patients in intervention and control groups, for binary outcomes the number of events, and for measurement scale outcomes the means and standard deviations. Extraction of the name of the Cochrane review group, and review authors' RoB assessments for the domains "Allocation concealment" and "Incomplete outcome data" was automated. In cases where trial publications could not be retrieved, they were requested from the review authors. If the blinding status of trial participants was unclear and the trial was published after 1999, authors were contacted via e-mail, asking for information on the blinding status of all groups within the trial.

4.4.2 Classifications

The blinding status of patients, healthcare providers and outcome assessors was assessed using a modified algorithm derived from that of Akl *et al* [129]. "Blinded" is defined as being unaware of the intervention status of individual patients or the patient's own blinding throughout the trial (as opposed to, for example, being unaware of the hypothesis of the trial). Healthcare providers were coded as "blind" if all staff involved in patient treatment and care were described as blinded, and as "non-blind" if all or a subgroup were described as non-blinded. Staff responsible for the determination of any healthcare provider decision outcomes was thus also covered by the coding of blinding status of "healthcare providers". The assessment of blinding status was done by two observers independently (APM, DL, LJ, MFO, HM or AH), and differences were resolved by discussion. The team differentiated between what we took to be definitive information on blinding status ("definitely yes"/"definitely no") based on explicit description or on contact with trial authors, and assessments based on other information in publications ("probably yes"/"probably no").

Classification of interventions as 'experimental' and 'control' was based on descriptions in the trial publications, except when some treatment was clearly labelled as "placebo", "control", "standard care" or "treatment as usual". In any of the latter cases we followed the labelling used by the review authors and classified these interventions as 'controls'. To ensure consistent comparisons of estimated bias across meta-analyses, we excluded meta-analyses in which intervention classifications were unclear. We excluded meta-analyses in which trials did not all have the same type of outcome (e.g. patient-reported), unless there was an informative subset of trials with the same type of outcome. We also excluded trials included in more than one meta-analysis with the same outcome, if the meta-analyses were due to be included in the same meta-epidemiological analysis. Such trials were removed at random until the trial only occurred within one meta-analysis. After removal of individual trials, some meta-analyses were no longer informative. Meta-analyses were classified according to whether the underlying hypothesis was of benefit (e.g. reduced mortality or increased live birth rate in the experimental intervention group); or of harm (when an intervention is assessed based on a hypothesis that some unwanted occurrence will increase, or some wanted occurrence decrease in the intervention group, typically "adverse events").

Observer-reported outcomes were subdivided into "objective: all-cause mortality", "objective: other than total mortality" (e.g. automatized non-repeatable laboratory tests), "subjective: pure observation" (e.g. assessment of radiographs) and "subjective: interactive" (e.g. assessment of clinical status). Subjective observer-reported outcomes were scored 1-3 according to degree of subjectivity (i.e. the extent to which determination of the outcome depended upon the judgement of the observer, 1 meant low degree of subjectivity). For example, the outcome FEV1 increase was assigned a score of 1 and improvement in depression a score of 3. The scoring of subjectivity was done by two observers (HM and

MFO) independently and masked to any results of trials or meta-analyses, with differences resolved by discussion. The classification of outcome type and experimental and comparison interventions were conducted to facilitate comparisons with an earlier meta-epidemiological study [10]. We further categorized experimental interventions as alternative/complementary or conventional medicine.

We excluded trials with binary outcomes in which no or all participants had the outcome event and trials with continuous outcomes where the required information for calculating the standardized mean difference (SMD) was missing.

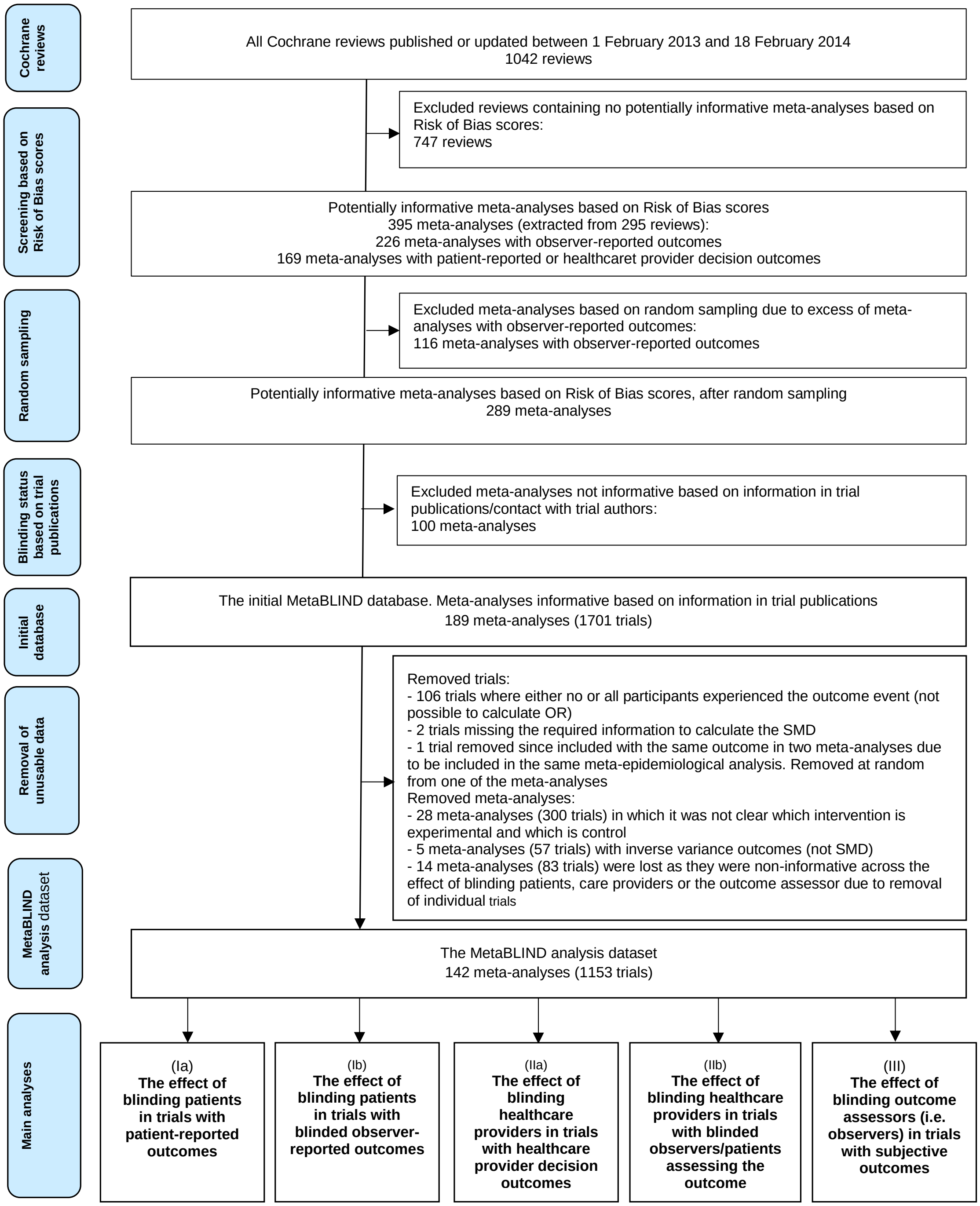
4.4.3 Data analysis

Intervention effects of binary outcomes were modelled as log odds ratios (ORs) and coded such that an OR<1 meant a beneficial intervention effect. For continuous outcomes the SMD and corresponding standard error were used and coded such that SMD<0 meant a beneficial intervention effect.

Bayesian hierarchical models for meta-epidemiological research developed by Welton *et al* [55] were used to estimate the average amount of bias associated with lack of each type of blinding, the average variability in this bias within a meta-analysis (quantified by the standard deviation [SD] increase in between trial heterogeneity, κ), and variability in average bias between meta-analyses (quantified by the SD of mean bias across meta-analyses, φ) . This model is explicitly described in Section 4.3.1.

To maximize the number of studies included in the analysis, I extended the Welton model to include both binary and continuous outcomes. As this is one of two methodological novelties in this chapter, we explain the rationale of why this is potentially appropriate in more detail in the following section.

Figure 4.1: Study flow diagram.


Cochrane reviews
All Cochrane reviews published or updated between 1 February 2013 and 18 February 2014
1042 reviews
Excluded reviews containing no potentially informative meta-analyses based on Risk of Bias scores:
747 reviews
Screening based on Risk of Bias scores
Potentially informative meta-analyses based on Risk of Bias scores
395 meta-analyses (extracted from 295 reviews):
226 meta-analyses with observer-reported outcomes
169 meta-analyses with patient-reported or healthcaret provider decision outcomes
Random sampling
Excluded meta-analyses based on random sampling due to excess of meta-analyses with observer-reported outcomes:
116 meta-analyses with observer-reported outcomes
Potentially informative meta-analyses based on Risk of Bias scores, after random sampling
289 meta-analyses
Blinding status based on trial publications
Excluded meta-analyses not informative based on information in trial publications/contact with trial authors:
100 meta-analyses
Initial database
The initial MetaBLIND database. Meta-analyses informative based on information in trial publications
189 meta-analyses (1701 trials)
Removal of unusable data
Removed trials:
- 106 trials where either no or all participants experienced the outcome event (not possible to calculate OR)
- 2 trials missing the required information to calculate the SMD
- 1 trial removed since included with the same outcome in two meta-analyses due to be included in the same meta-epidemiological analysis. Removed at random from one of the meta-analyses
Removed meta-analyses:
- 28 meta-analyses (300 trials) in which it was not clear which intervention is experimental and which is control
- 5 meta-analyses (57 trials) with inverse variance outcomes (not SMD)
- 14 meta-analyses (83 trials) were lost as they were non-informative across the effect of blinding patients, care providers or the outcome assessor due to removal of individual trials
MetaBLIND analysis dataset
The MetaBLIND analysis dataset
142 meta-analyses (1153 trials)
Main analyses
(Ia) The effect of blinding patients in trials with patient-reported outcomes
(Ib) The effect of blinding patients in trials with blinded observer-reported outcomes
(IIa) The effect of blinding healthcare providers in trials with healthcare provider decision outcomes
(IIb) The effect of blinding healthcare providers in trials with blinded observers/patients assessing the outcome
(III) The effect of blinding outcome assessors (i.e. observers) in trials with subjective outcomes

	(Ia)	(Ib)	(IIa)	(IIb)	(III)
*All**	21 (155)	14 (95)	35 (226)	14 (96)	51 (413)
*Hypothesis of benefit**	**18 (132)**	**14 (95)**	**29 (173)**	**13 (91)**	**46 (397)**
*Hypothesis of harm**	3 (23)	0	6 (53)	1 (5)	5 (16)

*Meta-analyses contributing with trials whose outcome measures were categorized as “mixed” (i.e. not possible to classify as either patient-reported, healthcare provider decision or observer-reported since they contained elements from more than one of these types) not counted here. “Mixed” outcome trials did not contribute to the main analyses.

Transforming continuous outcomes to binary outcomes on the log odds ratio scale

Since meta-analyses on the SMD scale can be difficult to interpret, Anzures-Cabrera *et al* provide an explanation of how it is possible to convert such continuous outcomes on the SMD scale to binary outcomes on the log odds ratio scale using a multiplicative factor of 1.81 [130]. We first show how 1.81 is derived and second interpret the assumptions we make in our meta-epidemiological study.

Suppose we have a continuous outcome with a true underlying mean and standard deviation for a single arm of a two arm study, that is, $X_1 \sim N(\mu_1, \nu_1^2)$. As always, we have the summary statistics of the sample mean and sample standard deviation. Now suppose there exists an underlying (because we only have summary level data) cut point C where the observations less than or equal to C determine the number of events in that arm, say m_1^C. We can then estimate these probabilities or risks as $r_1^C = \frac{m_1^C}{n_1}$. Now let $P_1^C = P(X_1 \leq C)$ be the true probabilities of the events in arm or group 1. If μ_1 increases, then $P_1^C = P(X_1 \leq C)$ decreases. This dependency can be expressed as $P_1^C(\mu_1)$. We can also take the logit of this, $logit(P_1^C(\mu_1))$. Anzures-Cabrera *et al* looked at whether there were instances when the relationship between μ_1 and $logit(P_1^C(\mu_1))$ was linear across values of μ_1, i.e. $\mu_1 = a + b[logit(P_1^C(\mu_1))]$ or equivalently $\frac{\mu_1}{b} = \frac{a}{b} + logit(P_1^C(\mu_1))$. Setting $K = \frac{\nu_1}{b}$ gives

$$K\frac{\mu_1}{\nu_1} = \frac{a}{b} + logit(P_1^C(\mu_1)) \tag{4.8}$$

For two groups with a single cut point, C, and SD (assumed the same for both groups) then:

$$K\frac{\mu_1 - \mu_2}{\nu} = logit(P_1^C(\mu_1)) - logit(P_2^C(\mu_2)) = log(OR)^C \tag{4.9}$$

Now suppose the continuous data (or again one arm of a study) have a logistic distribution with mean μ_1 and standard deviation ν_1. The cumulative distribution function is given by:

$$P_1^C(\mu_1) = \frac{1}{1 + e^{\frac{\pi(C-\mu_1)}{\nu_1\sqrt{3}}}} \tag{4.10}$$

It can be shown that:

$$logit(P_1^C(\mu_1)) = \frac{\pi(C-\mu_1)}{\nu_1\sqrt{3}} \tag{4.11}$$

which is a linear relationship between $logit(P_1^C(\mu_1))$ and μ_1 for a given C and ν. The following property of a logistic distribution can be derived for the difference between two logits for a common cut point, C, for two groups with means μ_1 and μ_2 :

$$log(OR)^C = \frac{\pi(C-\mu_1)}{\nu\sqrt{3}} - \frac{\pi(C-\mu_2)}{\nu\sqrt{3}} = -\frac{\pi}{\sqrt{3}}\frac{\mu_1-\mu_2}{\nu} \tag{4.12}$$

that holds for *any* cut point with $K = -\frac{\pi}{\sqrt{3}}$. Thus,

$$log(OR)^C = -\frac{\pi}{\sqrt{3}} * SMD \approx -1.81 * SMD \tag{4.13}$$

In our study, we therefore modelled continuous and binary data simultaneously in a single model, assuming a mixture of normal and binomial likelihoods but modelling the underlying bias on the same scale. For the continuous outcomes, this required re-expressing the SMDs as log odds ratios by multiplying the SMD and standard error by 1.81 [130, 131]. Although there is a minus sign in the derivation by Anzures-Cabrera et al, we multiply the SMD by positive 1.81 as both types of outcome were already coded such that an OR<1 and SMD<0 meant a beneficial treatment effect. By transforming the SMD to the log odds ratio scale, an underlying logistic distribution is assumed for continuous variables. We assume approximate an normal distribution for the effect estimate and therefore assumed normal likelihood for each trial i:

$$SMD * 1.81 = lnor \sim Normal\left(\theta_{i,m}, (1.81 * s_{i,m})^2\right) \tag{4.14}$$

$$\theta_{i,m} = \delta_{i,m} + \beta_{i,m} \tag{4.15}$$

$\beta_{i,m}$ are assumed to have same model structure as above (equations (4.5) and (4.6)) so that b_0, κ and φ are informed by both types of data together. We assume a binomial likelihood for the binary data with a logit link function (as described by equations (4.1) and (4.2)) and a normal likelihood for the continuous data with an identity link function (as described by equations (4.14) and (4.15)). To check the underlying assumption holds, we present the RORs separated by outcome type.

For the primary analyses we categorised trials as a "high risk" of bias if the blinding status of relevant parties were "definitely no" or "probably no" or in which it was unclear and "low risk" of bias if the blinding status were "definitely yes" or "probably yes". We conducted univariable analyses for each type of contrast in blinding status (i.e. of patients, healthcare providers and outcome assessors) using all informative meta-analyses for that characteristic.

In assessing the effect of blinding outcome assessors, we also studied the impact of our subjectivity scores on the difference in intervention effect. To do this, I extended the Welton *et al* model [55] to incorporate a three-level categorical covariate (low versus moderate versus high) at the meta-analysis level in the second methodological novelty in this chapter. The reference group was chosen to be the low level of subjectivity and indicator variables for high and moderate were added. Thus, allowing a different average amount of bias for each meta-analysis, to be estimated, by the level of subjectivity [55, 132]. Using the same notation as previously, we add the terms $b_1 x_{1m} + b_2 x_{2m}$ to equation (4.5):

$$\beta_{i,m} \sim N\left(b_m + b_1 x_{1m} + b_2 x_{2m}, \kappa^2\right) \tag{4.16}$$

b_1 is the regression coefficient of moderate subjectivity compared to low subjectivity. Similarly, b_2 is the regression coefficient of high subjectivity compared to low subjectivity. x_{1m}

is the indicator variable for moderate subjectivity in the mth meta-analysis. Similarly, x_{2m} is the value of the indicator variable for high subjectivity in the mth meta-analysis. As described previously, equation (4.6) remains the same and vague prior distributions are assumed for b_1 and b_2. The association between degree of subjectivity of the outcome and the average magnitude of bias is interpreted for each level of bias in comparison to the reference group. We denote $exp(b_0)$ as the magnitude of bias for outcomes with a low degree subjectivity score. $exp(b_1)$ is the magnitude of bias for outcomes with a moderate degree compared to a low degree subjectivity score. Similarly, $exp(b_2)$ is the magnitude of bias for outcomes with a high degree compared to a low degree subjectivity score. The code for both models can be found in Appendix B.

The risk of confounding by other limitations in trial design was assessed in multivariable analyses by re-running the main analyses with adjustment in the model for: concealment of the allocation sequence, RoB due to incomplete outcome data (attrition bias) , trial size and blinding status of patients (the latter only in the analysis of outcome assessor blinding [III]). The model is described in equation (4.7). We adjusted for each of these characteristics in turn in separate analyses: we did not adjust for combinations of the covariates. The multivariable model allows the covariate (potentially confounding variable) to act on the mean bias but not to explain additional variability in bias; the models therefore included the same variance components as the main univariable analyses. In sensitivity analyses we excluded trials in which it was unclear whether the relevant parties had been blinded: (i) excluding only those with a classification of “unclear”, (ii) excluding trials with an “unclear”, “probably yes” or “probably no” classification, such that only “definitely yes” and “definitely no” remained.

In order to facilitate comparison of our study with previous meta-epidemiological studies we also compared trials described as “double blind” or “triple blind” having both patients, care providers and outcome assessors described explicitly as blind in the publication with those trials for which neither was the case. Secondary analyses were stratified by outcome type (for example, objective outcomes (and subtypes), outcomes classified as mixed).

4.5 Results

Table 4.1 shows characteristics of the 142 meta-analyses and 1,153 trials included in the analysis dataset. The median year of publication of the included trials was 2003, and the median sample size was 768 patients for meta-analyses and 106 for trials. Twenty-two meta-analyses (16%) assessed outcomes related to adverse effects of the treatment , followed by resource use (19 meta-analyses [13%]) and clinician assessed outcomes (12 meta-analyses [9%]). There were 68 meta-analyses with observer-reported outcomes, of which subjectively assessed outcomes were reported most often, in 53 meta-analyses (78%), followed by all-cause mortality (11 meta-analyses [16%]). Of 1153 trials included in the analysis dataset 1112 trials (96%) had a parallel trial design and 753 (65%) were drug trials. Trial authors were contacted in 5% of trials (54/1153). The authors response rate was 52% (28/54). This reduced the fraction of trials with "unclear" blinding status from 8% (95/1153) to 6% (67/1153). Full details can be found in Appendix A, Table A.4 .

Various methodological characteristics were highly correlated with each other across trials. There were very high correlations amongst types of blinding whilst there were lower correlations between blinding and other (non-blinding) RoB judgements. For example, trials with blinded patients were more likely to have blinded outcome assessors (OR, 75.0 [95% CI, 38.6 to 145.8]), compared to trials with unblinded patients. Trials with blinded outcome assessors were more likely to have adequate allocation concealment (OR, 3.0 [95% CI, 2.2 to 4.0]) and complete outcome data (OR, 2.0 [95% CI, 1.5 to 2.8]). Compared to trials with unblinded patients, trials with blinded patients were more likely to have blinded outcome assessors (OR, 75.0 [95% CI, 38.6 to 145.8]). Full details on these aspects can be found in Appendix A, Table A.5. Figure 4.2 summarises results for each of the five main analyses (Ia, Ib, IIa, IIb, III). For illustration, forest plots of results from individual meta-analyses are presented for each of the main analyses in Appendix A, Figure A.5.

Table 4.1: Characteristics of included Meta-analyses and Trials. Overall dataset and main analyses (Ia, Ib, IIa, IIb, III).

	Overall dataset		**Ia**		**Ib**		**IIa**		**IIb**		**III**	
	Meta-analyses	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**
	142	**1153**	**18**	**132**	**14**	**95**	**29**	**173**	**13**	**91**	**46**	**397**
Outcome measures according to clinical area												
Adverse events (as adverse effects of the treatment)	22 (15.5)	129 (11.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
All-cause mortality	7 (4.9)	143 (12.4)	0 (0.0)	0 (0.0)	2 (14.3)	27 (28.4)	0 (0.0)	0 (0.0)	2 (15.4)	27 (29.7)	0 (0.0)	0 (0.0)
Cause-specific mortality	1 (0.7)	11 (1.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (2.2)	11 (2.8)
Clinician-assessed outcomes	12 (8.5)	95 (8.2)	0 (0.0)	0 (0.0)	1 (7.1)	11 (11.6)	1 (3.4)	3 (1.7)	0 (0.0)	0 (0.0)	11 (23.9)	92 (23.2)
Composite end point inc. mortality or major morbidity	2 (1.4)	16 (1.4)	0 (0.0)	0 (0.0)	2 (14.3)	12 (12.6)	0 (0.0)	0 (0.0)	1 (7.7)	7 (7.7)	1 (2.2)	9 (2.3)

Continued on next page

Table 4.1 – continued from previous page

	Overall dataset		Ia		Ib		IIa		IIb		III	
	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials
	142	**1153**	**18**	**132**	**14**	**95**	**29**	**173**	**13**	**91**	**46**	**397**
Global improvement	3 (2.1)	14 (1.2)	0 (0.0)	0 (0.0)	2 (14.3)	5 (5.3)	0 (0.0)	0 (0.0)	2 (15.4)	5 (5.5)	2 (4.3)	12 (3.0)
Laboratory-reported outcomes	5 (3.5)	45 (3.9)	0 (0.0)	0 (0.0)	1 (7.1)	2 (2.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (2.2)	4 (1.0)
Lifestyle outcomes	5 (3.5)	100 (8.7)	1 (5.6)	2 (1.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (6.5)	63 (15.9)
Major morbidity event	5 (3.5)	44 (3.8)	0 (0.0)	0 (0.0)	3 (21.4)	24 (25.3)	0 (0.0)	0 (0.0)	3 (23.1)	24 (26.4)	5 (10.9)	44 (11.1)
Mental health outcomes	7 (4.9)	61 (5.3)	2 (11.1)	9 (6.8)	1 (7.1)	4 (4.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	5 (10.9)	52 (13.1)
Other outcomes (not classified elsewhere)	15 (10.6)	145 (12.6)	5 (27.8)	79 (59.8)	1 (7.1)	2 (2.1)	4 (13.8)	16 (9.2)	2 (15.4)	4 (4.4)	5 (10.9)	48 (12.1)

Continued on next page

Table 4.1 – continued from previous page

	Overall dataset		Ia		Ib		IIa		IIb		III	
	Meta-analyses	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**
	142	**1153**	**18**	**132**	**14**	**95**	**29**	**173**	**13**	**91**	**46**	**397**
Pain	5 (3.5)	17 (1.5)	3 (16.7)	8 (6.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (7.7)	7 (7.7)	1 (2.2)	2 (0.5)
Perinatal outcomes	5 (3.5)	34 (2.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (3.4)	2 (1.2)	0 (0.0)	0 (0.0)	1 (2.2)	9 (2.3)
Pregnancy outcomes	8 (5.6)	28 (2.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (3.4)	3 (1.7)	0 (0.0)	0 (0.0)	6 (13.0)	23 (5.8)
Quality of life	3 (2.1)	19 (1.6)	2 (11.1)	6 (4.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (2.2)	13 (3.3)
Radiological outcomes	2 (1.4)	11 (1.0)	0 (0.0)	0 (0.0)	1 (7.1)	8 (8.4)	0 (0.0)	0 (0.0)	1 (7.7)	8 (8.8)	2 (4.3)	11 (2.8)
Resource use	19 (13.4)	133 (11.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	19 (65.5)	133 (76.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

Continued on next page

Table 4.1 – continued from previous page

	Overall dataset		Ia		Ib		IIa		IIb		III	
	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials
	142	**1153**	**18**	**132**	**14**	**95**	**29**	**173**	**13**	**91**	**46**	**397**
Surgical and device-related outcomes	4 (2.8)	20 (1.7)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (10.3)	16 (9.2)	0 (0.0)	0 (0.0)	1 (2.2)	4 (1.0)
Symptoms or signs of illness or condition	6 (4.2)	35 (3.0)	5 (27.8)	28 (21.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (7.7)	9 (9.9)	0 (0.0)	0 (0.0)
Withdrawals/ dropouts/ compliance	6 (4.2)	53 (4.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Type of experimental intervention												
Pharmacologic	95 (66.9)	728 (63.1)	12 (66.7)	48 (36.4)	10 (71.4)	74 (77.9)	19 (65.5)	121 (69.9)	10 (76.9)	78 (85.7)	25 (54.3)	195 (49.1)
Surgical	3 (2.1)	12 (1.0)	1 (5.6)	4 (3.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (2.2)	4 (1.0)

Continued on next page

Table 4.1 – continued from previous page

	Overall dataset		Ia		Ib		IIa		IIb		III	
	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials
	142	**1153**	**18**	**132**	**14**	**95**	**29**	**173**	**13**	**91**	**46**	**397**
Psychosocial, behavioural or educational	17 (12.0)	204 (17.7)	1 (5.6)	42 (31.8)	3 (21.4)	17 (17.9)	3 (10.3)	10 (5.8)	1 (7.7)	2 (2.2)	9 (19.6)	101 (25.4)
Other	27 (19.0)	209 (18.1)	4 (22.2)	38 (28.8)	1 (7.1)	4 (4.2)	7 (24.1)	42 (24.3)	2 (15.4)	11 (12.1)	11 (23.9)	97 (24.4)
Field of experimental intervention												
Conventional medicine	137 (96.5)	1100 (95.4)	17 (94.4)	127 (96.2)	14 (100)	95 (100)	29 (100)	173 (100)	12 (92.3)	84 (92.3)	44 (95.7)	368 (92.7)
Alternative/ complementary medicine	5 (3.5)	53 (4.6)	1 (5.6)	5 (3.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (7.7)	7 (7.7)	2 (4.3)	29 (7.3)
Type of comparison intervention												
Placebo or no treatment	57 (40.1)	442 (38.3)	8 (44.4)	36 (27.3)	1 (7.1)	11 (11.6)	12 (41.4)	47 (27.2)	2 (15.4)	16 (17.6)	17 (37.0)	160 (40.3)

Continued on next page

Table 4.1 – continued from previous page

	Overall dataset		**Ia**		**Ib**		**IIa**		**IIb**		**III**	
	Meta-analyses	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**
	142	**1153**	**18**	**132**	**14**	**95**	**29**	**173**	**13**	**91**	**46**	**397**
Other inactive (Standard care)	38 (26.8)	452 (39.2)	4 (22.2)	76 (57.6)	7 (50.0)	55 (57.9)	9 (31.0)	84 (48.6)	5 (38.5)	46 (50.5)	17 (37.0)	176 (44.3)
Active comparison	47 (33.1)	259 (22.5)	6 (33.3)	20 (15.2)	6 (42.9)	29 (30.5)	8 (27.6)	42 (24.3)	6 (46.2)	29 (31.9)	12 (26.1)	61 (15.4)
Hypothesis of benefit	114 (80.3)	971 (84.2)	18 (100)	132 (100)	14 (100)	95 (100)	29 (100)	173 (100)	13 (100)	91 (100)	46 (100)	397 (100)
Observer-reported outcome*	68 (47.9)	640 (55.5)	0 (0.0)	0 (0.0)	14 (100)	95 (100)	0 (0.0)	0 (0.0)	10 (76.9)	73 (80.2)	46 (100)	397 (100)
All-cause mortality	11 (16.2)	170 (26.6)	0 (0.0)	0 (0.0)	2 (14.3)	27 (28.4)	0 (0.0)	0 (0.0)	2 (20.0)	27 (37.0)	0 (0.0)	0 (0.0)
Other objective	4 (5.9)	39 (6.1)	0 (0.0)	0 (0.0)	1 (7.1)	2 (2.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

Continued on next page

Table 4.1 – continued from previous page

	Overall dataset		**Ia**		**Ib**		**IIa**		**IIb**		**III**	
	Meta-analyses	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**	**Meta-analyses**	**Trials**
	142	**1153**	**18**	**132**	**14**	**95**	**29**	**173**	**13**	**91**	**46**	**397**
Subjective	53 (77.9)	431 (67.3)	0 (0.0)	0 (0.0)	11 (78.6)	66 (69.5)	0 (0.0)	0 (0.0)	8 (80.0)	46 (63.0)	0 (0.0)	0 (0.0)
Binary or measurement scale outcome												
Binary	110 (77.5)	885 (76.8)	9 (50.0)	42 (31.8)	11 (78.6)	78 (82.1)	25 (86.2)	151 (87.3)	11 (84.6)	82 (90.1)	32 (69.6)	289 (72.8)
Continuous	31 (21.8)	265 (23.0)	8 (44.4)	87 (65.9)	3 (21.4)	17 (17.9)	4 (13.8)	22 (12.7)	2 (15.4)	9 (9.9)	14 (30.4)	108 (27.2)
Inverse variance	1 (0.7)	3 (0.3)	1 (5.6)	3 (2.3)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

(Ia) The effect of blinding patients in trials with patient-reported outcomes

Information was available on 18 informative meta-analyses with a hypothesis of benefit, containing 132 trials. Patient blinding was assessed as “probably yes” or “definitely yes” in 33 of these trials (25%) and “probably no”, “definitely no” or “unclear” in 99 (75%). The overall ROR was 0.91 (95% CrI, 0.61 to 1.34) comparing trials with lack of or unclear blinding of patients with those with blinded patients. The average increase in between-trial heterogeneity among the trials with non-blind patients was estimated to be an SD of 0.22 (95% CrI, 0.02 to 0.60) and the between–meta-analysis variation in average bias was estimated to be SD, 0.20 (95% CrI, 0.01 to 0.74).

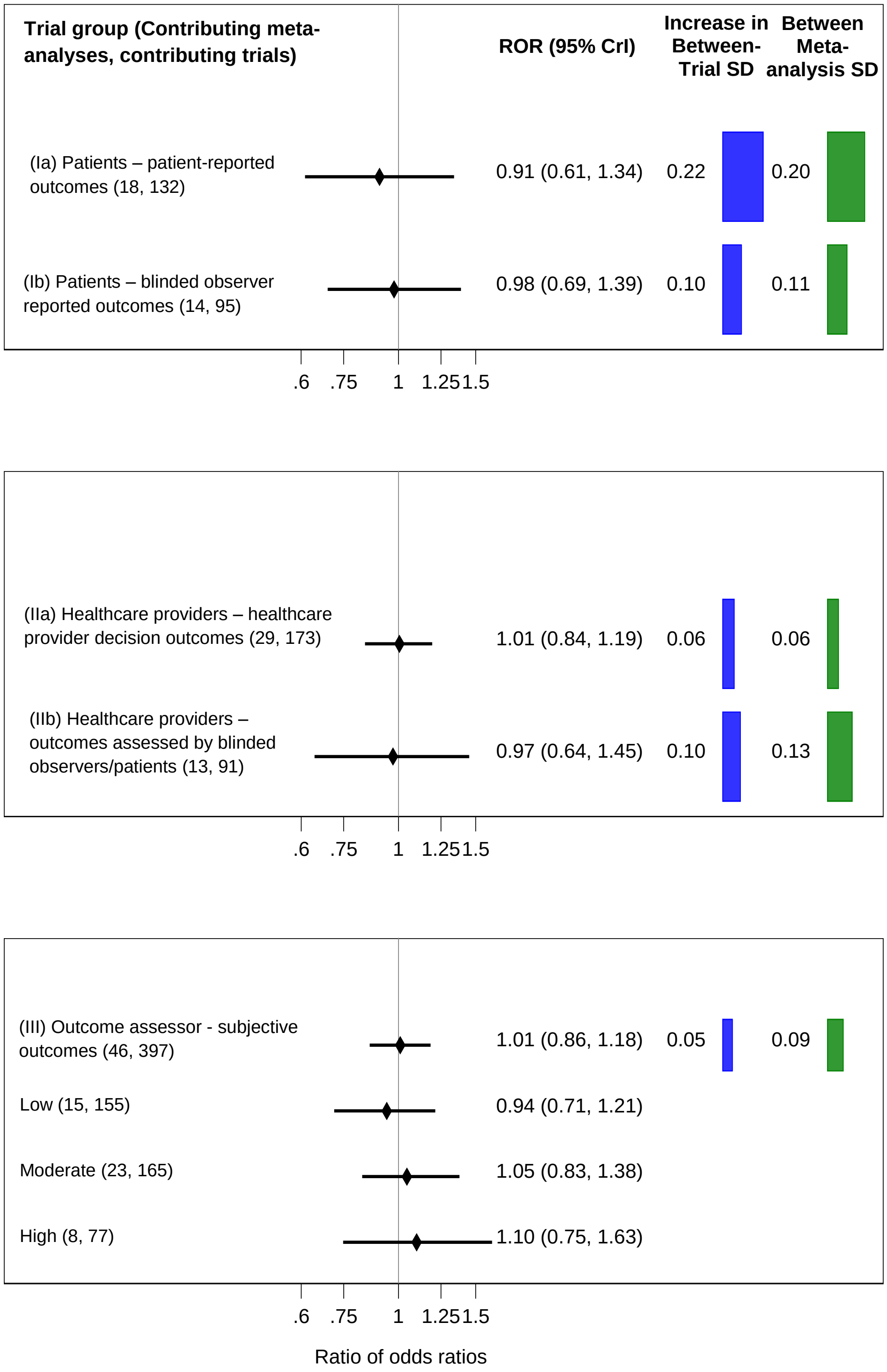
(Ib) The effect of blinding patients in trials with blinded observer-reported outcomes

Information was available on 14 informative meta-analyses with a hypothesis of benefit, containing 95 trials. Patient blinding was assessed as “probably yes” or “definitely yes” in 57 (60%) of these and “probably no”, “definitely no” or “unclear” in 38 (40%). The ROR was 0.98 (95% CrI, 0.69 to 1.39) when comparing effect estimates from trials with lack of or unclear blinding of patients with effect estimates from trials with blinded patients. The increase in between-trial heterogeneity among the trials with non-blinded patients was SD = 0.10 (95% CrI, 0.01 to 0.60).

(IIa) The effect of blinding healthcare providers in trials with healthcare provider decision outcomes

Information was available on 29 informative meta-analyses with a hypothesis of benefit, containing 173 trials. Healthcare provider blinding was assessed as “probably yes” or “definitely yes” in 93 of these trials (54%) and “probably no”, “definitely no” or “un-

Figure 4.2: Estimated RORs and effects on heterogeneity associated with blinding status of patients, healthcare providers and outcome assessors.



clear" in 80 (46%). The overall ROR was 1.01 (95% CrI, 0.84 to 1.19) when comparing effect estimates from trials with lack of or unclear blinding of healthcare providers with effect estimates from trials with blinded healthcare providers. The estimated increase in between-trial heterogeneity among the trials with non-blinded healthcare providers was SD = 0.06 (95% CrI, 0.01 to 0.30).

(IIb) The effect of blinding healthcare providers in trials with blinded observers/patients assessing the outcome

Information was available on 13 informative meta-analyses with a hypothesis of benefit, containing 91 trials. Healthcare provider blinding was assessed as "probably yes" or "definitely yes" in 61 trials (67%) and "probably no", "definitely no" or "unclear" in 30 (33%) of these trials. The overall ROR was 0.97 (95% CrI, 0.64 to 1.45) when comparing effect estimates from trials with lack of or unclear blinding of healthcare providers with effect estimates from trials with blinded healthcare providers. The increase in between-trial heterogeneity among the trials with non-blinded healthcare providers was SD = 0.10 (95% CrI, 0.01 to 0.59).

(III) The effect of blinding outcome assessors (i.e. observers) in trials with subjective outcomes

Information was available on 46 informative meta-analyses with a hypothesis of benefit containing 397 trials. Outcome assessor blinding was assessed as "probably yes" or "definitely yes" in 199 of these trials (50%) and "probably no", "definitely no" or "unclear" in 198 (50%). The overall ROR was 1.01 (95% CrI, 0.86 to 1.18) when comparing effect estimates from trials with lack of or unclear blinding of outcome assessors with effect estimates from trials with blinded outcome assessors, in meta-analyses with subjectively assessed outcomes. The increase in between-trial heterogeneity among the trials with non-blinded outcome assessors was SD = 0.05 (95% CrI, 0.01 to 0.22). Investigating the impact of the degree of subjectivity, using our extension to the Welton model, described in Section 5.4.3, equation (1.10), gave RORs of 0.94 (95% CrI, 0.71 to 1.21), 1.05 (95% CrI, 0.83 to 1.38) and 1.10 (95% CrI, 0.75 to 1.63) for the meta-analyses with the lowest, middle and highest score for subjectivity, respectively.

For each of the main analyses, excluding (Ia), there appeared to be only limited between–meta analysis heterogeneity in mean bias, φ , although the 95% CrIs were wide (between meta-analyses SD (Ia) 0.20 (0.01, 0.74), (Ib) 0.11 (0.01, 0.55), (IIa) 0.06 (0.01, 0.26), (IIb) 0.13 (0.01, 0.82) and (III) 0.09 (0.01, 0.31) for the five main analyses).

For each of the five main analyses, separate adjustment for concealment of the allocation sequence, attrition and trial size did not change the result (Table 4.2). Estimated increases in between trial heterogeneity and estimates of between–meta-analysis variability in average bias were also little changed, compared with the unadjusted main analyses. Excluding trials with "unclear" blinding status of the relevant group from the unadjusted main analyses did not change the results substantially.

Secondary analyses

Secondary analyses looking separately at the effect of blinding patients, healthcare providers or outcome assessors across different types of outcomes can be found in the Table 4.3. For example, an analysis based on observer-reported outcomes classified as objective also showed no effect of outcome assessor blinding status (ROR 0.94 [95% CrI 0.61 to 1.26]) (meta-analyses with a hypothesis of benefit only). Analyses comparing trials described as "double blind" (or "triple blind") with those not so described or "unclear" did not show any effect when meta-analyses with any type of outcome were included (ROR 0.99 [95%

Table 4.2: Adjusted analyses.

	Median posterior estimates (95% CrI)	Allocation concealment	Incomplete outcome data	Trial size	Excluding unclears [1]
(Ia)	ROR	0.91 (0.61, 1.35)	0.91 (0.63, 1.31)	0.89 (0.59, 1.29) [2]	1.10 (0.72, 1.69)
	φ	0.20 (0.02, 0.74)	0.17 (0.01, 0.70)	0.18 (0.02, 0.74)	0.19 (0.02, 0.76)
	κ	0.21 (0.01, 0.61)	0.18 (0.01, 0.60)	0.18 (0.01, 0.60)	0.23 (0.02, 0.61)
(Ib)	ROR	1.07 (0.74, 1.56)	1.08 (0.63, 1.31)	0.99 (0.69, 1.39)	1.00 (0.70, 1.44)
	φ	0.11 (0.01, 0.57)	0.10 (0.01, 0.52)	0.10 (0.01, 0.54)	0.11 (0.01, 0.58)
	κ	0.10 (0.01, 0.57)	0.13 (0.01, 0.72)	0.10 (0.01, 0.57)	0.10 (0.01, 0.60)
(IIa)	ROR	1.03 (0.84, 1.23)	0.98 (0.72, 1.58)	1.00 (0.83, 1.19)	0.97 (0.77, 1.18)
	φ	0.07 (0.01, 0.29)	0.06 (0.01, 0.28)	0.06 (0.01, 0.27)	0.08 (0.01, 0.36)
	κ	0.06 (0.01, 0.28)	0.07 (0.01, 0.30)	0.06 (0.01, 0.29)	0.07 (0.01, 0.39)
(IIb)	ROR	1.03 (0.67, 1.54)	1.07 (0.80, 1.17)	0.98 (0.63, 1.44)	0.96 (0.64, 1.45)
	φ	0.13 (0.01, 0.80)	0.12 (0.01, 0.77)	0.13 (0.01, 0.82)	0.14 (0.01, 0.82)
	κ	0.10 (0.01, 0.60)	0.09 (0.01, 0.60)	0.09 (0.01, 0.58)	0.10 (0.01, 0.68)
(III)[3]	ROR	1.04 (0.89, 1.23)	1.02 (0.87, 1.19)	1.03 (0.88, 1.21)	1.01 (0.85, 1.20)
	φ	0.10 (0.01, 0.36)	0.08 (0.01, 0.33)	0.10 (0.01, 0.34)	0.11 (0.01, 0.35)
	κ	0.05 (0.01, 0.21)	0.05 (0.01, 0.19)	0.06 (0.01, 0.25)	0.06 (0.01, 0.25)

[1] Number of meta-analyses, trials: (Ia)=(16, 116); (Ib)=(14, 94); (IIa)=(28, 160); (IIb)=(13, 90); (III)=(43, 365).

[2] One meta-analysis (3 trials) were removed which did not specify the size of the trial due to the format given in the review.

[3] Adjusted for patient blinding ROR=1.03 (95% CrI: 0.87 to 1.23), φ=0.10 (95% CrI: 0.01 to 0.32), κ=0.06 (95% CrI: 0.01 to 0.22).

CrI, 0.86 to 1.09]), nor when only meta-analyses with subjective observer-reported outcomes and a hypothesis of benefit were included (ROR 1.11 [95% CrI, 0.86 to 1.44]) (Table 4.3).

Table 4.3: Secondary analyses.

	N	**ROR**	φ	κ
	(MA, trial)	**(95% CrI)**	**(95% CrI)**	**(95% CrI)**
Lack of double blinding or unclear double blinding (vs double blind):				
All outcomes	(94, 722)	0.99 (0.86, 1.09)	0.07 (0.01, 0.29)	0.06 (0.01, 0.18)
All outcomes – benefit	(74, 583)	1,02 (0.90, 1.13)	0.06 (0.01, 0.27)	0.07 (0.01, 0.19)
All outcomes – harms	(20, 139)	0.64 (0.38, 1.04)	0.15 (0.01, 0.89)	0.13 (0.01, 1.23)
Observer-reported outcomes	(36, 37)	1.04 (0.84, 1.25)	0.14 (0.01, 0.57)	0.08 (0.01, 0.23)
Subjectively assessed observer-reported outcomes *Same outcomes as analysis (III)*	(27, 221)	1.11 (0.86, 1.44)	0.13 (0.01, 0.61)	0.09 (0.01, 0.42)
Mortality within observer-reported outcomes	(6, 124)	0.87 (0.45, 1.32)	0.35 (0.02, 1.25)	0.08 (0.01, 0.26)
Patient-reported outcomes *Same outcomes as analysis (Ia)*	(13, 53)	0.89 (0.57, 1.40)	0.15 (0.01, 0.83)	0.12 (0.01, 0.88)
Healthcare provider outcomes *Same outcomes as analysis (IIa)*	(24, 147)	0.98 (0.79, 1.19)	0.07 (0.01, 0.31)	0.07 (0.01, 0.36)
The effect of blinding patients in trials with the following outcomes:				
Private patient-reported outcomes *Subset of analysis (Ia)*	(14, 120)	1.06 (0.67, 1.69)	0.22 (0.02, 0.85)	0.32 (0.02, 0.63)

Continued on next page

Table 4.3 – continued from previous page

	N	ROR	φ	κ
	(MA, trial)	(95% CrI)	(95% CrI)	(95% CrI)
Patient and observer-reported outcomes (blinded) with mixed outcomes *Analysis (Ia) and (Ib) with mixed outcomes*	(34, 277)	0.94 (0.74, 1.19)	0.11 (0.01, 0.48)	0.12 (0.01, 0.52)
Patient and observer-reported outcomes (blinded) without mixed outcomes *Analysis (Ia) and (Ib) without mixed outcomes*	(32, 267)	0.95 (0.76, 1.21)	0.11 (0.01, 0.44)	0.13 (0.01, 0.52)
The effect of blinding healthcare providers in trials with the following outcomes:				
Observer-reported outcomes assessed by blind observers	(11, 78)	1.05 (0.56, 1.58)	0.11 (0.01, 0.67)	0.11 (0.01, 0.61)
All outcomes jointly including mixed	(42, 250)	1.01 (0.86, 1.19)	0.06 (0.01, 0.26)	0.06 (0.01, 0.26)
The effect of blinding outcome assessors in trials with the following outcomes:				
Any objective outcomes	(15, 207)	0.94 (0.61, 1.26)	0.23 (0.02, 0.82)	0.13 (0.02, 0.39)
All-cause mortality	(11, 168)	0.91 (0.51, 1.31)	0.29 (0.02, 1.15)	0.1 (0.02, 0.32)
Subjective interactive outcomes *Analysis (III) excluding subjective pure observation outcomes*	(15, 145)	1.22 (0.94, 1.58)	0.08 (0.01, 0.39)	0.16 (0.01, 0.53)
Subjective pure observation outcomes *Analysis (III) excluding subjective interactive outcomes*	(31, 252)	0.92 (0.76, 1.12)	0.1 (0.01, 0.39)	0.05 (0.01, 0.20)

Continued on next page

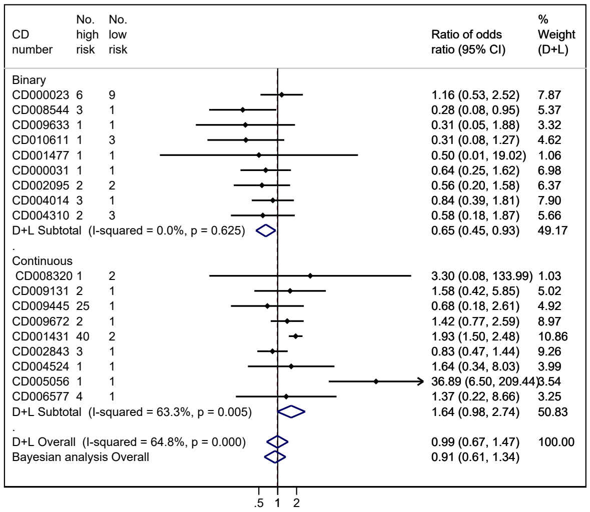
Table 4.3 – continued from previous page

	N	ROR	φ	κ
	(MA, trial)	(95% CrI)	(95% CrI)	(95% CrI)
Observer-reported outcomes without mixed outcomes	(61, 604)	1.01 (0.88, 1.14)	0.1 (0.01, 0.33)	0.08 (0,01, 0.22)
Observer-reported outcomes including mixed outcomes	(65, 624)	1.01 (0.89, 1.14)	0.09 (0.01, 0.30)	0.08 (0.01, 0.21)

4.5.1 Comparison between binary and continuous outcomes

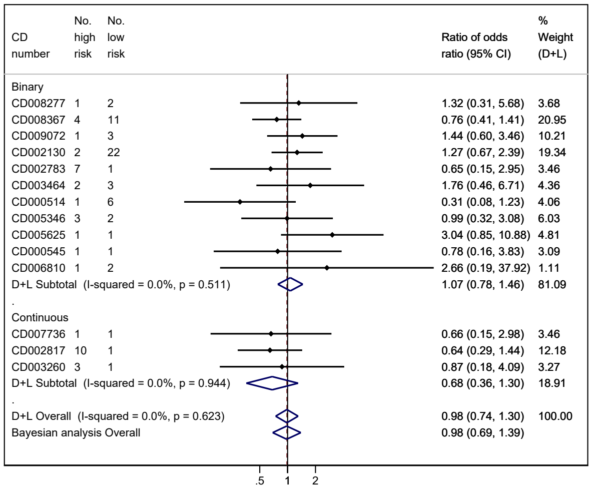
Figure 4.3 shows each of the five main analyses split by binary and continuous outcomes. We see that for all but analysis (Ia) the results from binary and continuous outcomes are comparable so there is no particular reason to doubt the assumptions made. In analysis (Ia) the meta-analysis with Cochrane number CD005056 appears to be an outlier. Having looked at the original data from this Cochrane study it is correct with only two studies, each showing the opposite effect, making the difference in the treatment effects extreme. As this pattern was not seen in the other four analyses, we do not see this as evidence against the model assumptions. As our models are based on the Bayesian hierarchical model conducted in WinBUGS, the studies with only one low RoB study would be given relatively little weight, driven by its prior distribution.

Figure 4.3: RORs from individual meta-analyses and from analyses combined across all meta-analyses. Results for individual meta-analyses are frequentist estimates with confidence intervals, based on comparing the summary odds ratio from studies with the study characteristic of interest with the summary odds ratio from studies without the characteristic. The overall estimates of RORs are results based on the Bayesian hierarchical model described in the main text. CD numbers are identifiers of individual Cochrane reviews, from the Cochrane Database of Systematic Reviews.



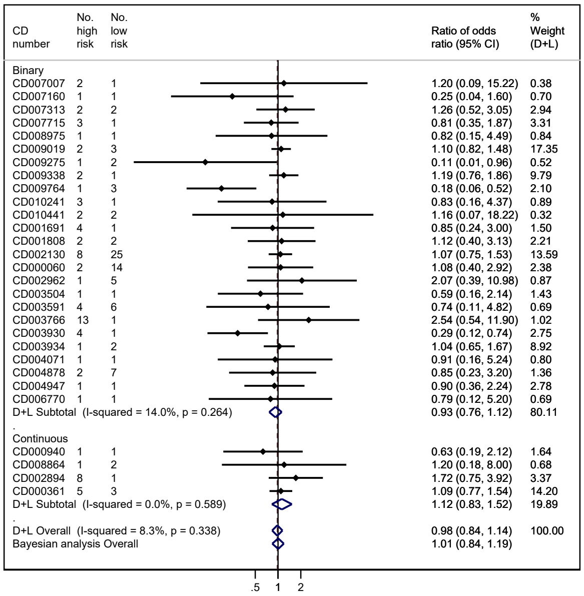
(a) (Ia) The effect of blinding patients in trials with patient-reported outcomes

Figure 4.3: (cont.)



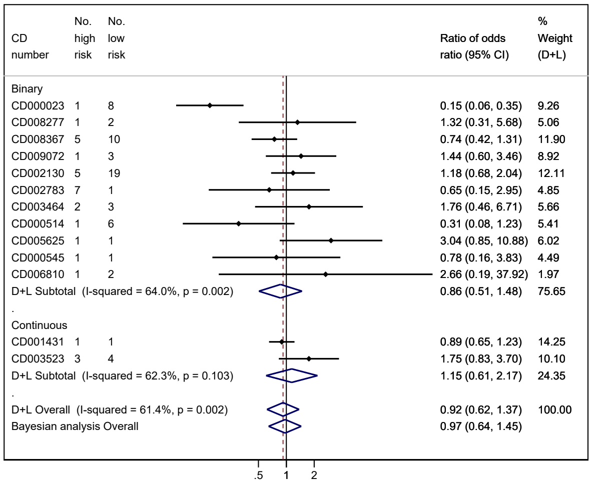
(b) (Ib) The effect of blinding patients in trials with blinded observer-reported outcomes

Figure 4.3: (cont.)



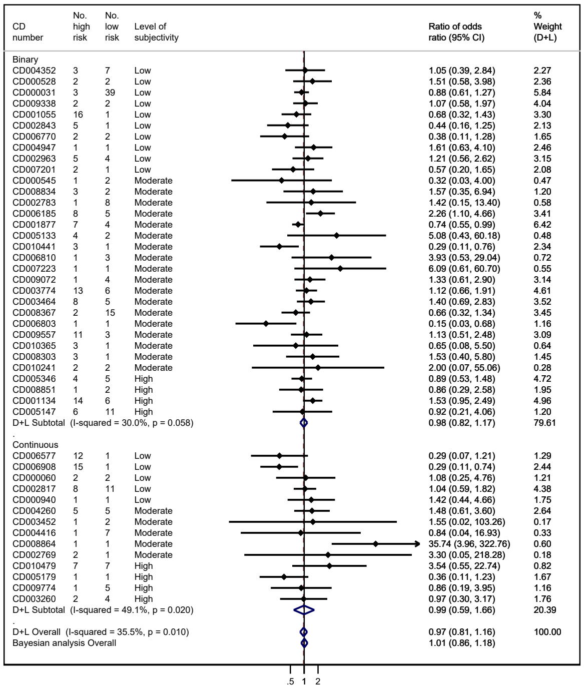
(c) (IIa) The effect of blinding healthcare providers in trials with healthcare provider decision outcomes

Figure 4.3: (cont.)



(d) (IIb) The effect of blinding healthcare providers in trials with blinded observers/patients assessing the outcome

Figure 4.3: (cont.)



(e) (III) The effect of blinding outcome assessors (i.e. observers) in trials with subjective outcomes

4.6 Discussion

4.6.1 Overview of key findings

In this chapter, the first meta-epidemiological study was conducted to look separately at the effect of blinding patients, those providing care or deciding on other aspects of a patient's care and outcome assessors. We estimated the influence of study design characteristics on average intervention effect estimates and on heterogeneity within a meta-analysis. To our surprise, we found no evidence for bias, on average, resulting from a lack of blinding of patients, those providing care or those or deciding on other aspects of a patient's care or outcome assessors in randomised clinical trials. This was the case even among trials with the most subjectively measured outcomes. For all the results, CrIs were wide showing considerable uncertainty in our estimates which encompassed both considerable difference and no difference. Our main findings were unaffected by adjustment for possible confounders and in sensitivity analyses. Estimates of the difference in between-trial heterogeneity associated with lack of blinding provided only weak evidence of an increase, except in the analysis concerning patient blinding and patient-reported outcomes. The same pattern was found when comparing "double-blind" with not "double blind" trials.

Often, meta-epidemiological studies are published with either all binary outcomes or all continuous outcomes. However, since some meta-epidemiological studies can be small, due to the pooling of data being labour intensive, we combined binary and continuous outcomes in a single model. We have shown it can be helpful to combine meta-analyses with binary and meta-analyses with continuous outcomes, by re-expressing standardised mean differences from continuous outcomes, as log odds ratios. One advantage of this is it allows the underlying bias to be modelled on the same scale, whilst still using a binomial likelihood for the binary outcomes. The incorporation of more studies can improve precision of the estimate in some settings. We also extended the Welton *et al* model to include meta-analysis level covariates rather than conducting separate analyses. This can be advantageous when the separate analyses have small sample sizes and therefore lower power. This can also benefit the analyst by producing stratified analyses to enable the average bias to be compared across levels of subjectivity or any other relevant categories, whilst making use of all the trials in the dataset.

This result is based on a sample of meta-analyses from Cochrane reviews with no restriction on topic, and thus comprising a broad sample of clinical areas, interventions and outcomes. We based our classification of trials as blinded or non-blinded on information on actual trial conduct, rather than relying on the ambiguous label of "double blind". This was achieved by studying explicit descriptions in the trial publications and, if necessary, contacting trial authors. To our knowledge, contact to trial authors or information on trial conduct has been attempted in no other meta-epidemiological study. Accessing information on actual trial conduct allows classification of the blinding status of each group of participants in a trial separately: patients, healthcare providers and outcome assessors. This in turn enables disentangling of the effects of blinding outcomes assessors and patients. Individual trial outcomes were carefully classified, based on information in trial publications, according to who was involved in their determination, so as to separately look at patient reported outcomes, healthcare provider and observer reported outcomes.

4.6.2 Limitations

This study attempts to measure the effect of blinding of patients, those providing care or deciding upon the treatment and of the process of outcome assessment through an observational design: trials with inadequate blinding are compared with trials in which blinding was adequate, within meta-analyses. As with any observation design, there will be a higher risk of confounding. We have compared trials within meta-analyses, which should reduce some confounding since these sets of trials should be comparable regarding interventions, patient populations etc. However, trials within a meta-analysis may still differ in a variety of other ways, including other methodological aspects and therefore residual confounding is possible. The optimal design for a study attempting to measure precisely the impact of lack of blinding of outcome assessors would be experimental, as opposed to our observational design. The study would be randomised in design (half the patients randomised to be blinded, half not blinded).

Figure 4.3(a) shows the effect of bias from the continuous and binary data are in opposite directions. Although this is surprising, the remaining four analyses do not show this, suggesting that we do not see this as evidence against the model assumptions. However, further work could look at this in more detail through a simulation study. A simulation study would determine at which point the model assumptions break down and provide guidelines on when it may be appropriate/not appropriate to pool binary and continuous outcomes in a meta-epidemiological model.

Our study had potentially lower power than we anticipated, as we had to disregard lots of meta-analyses which were non-informative.

When trial publications explicitly and clearly described blinding status, we did not contact the authors for further information. Further, only about half of the trial authors that were contacted by the team (HM) answered and the remaining unclear RoB trials could therefore lead to some misclassification of blinding status. However, our main result persisted after exclusion of all trials where blinding status was not based on direct explicit description or information gained through contact with authors.

4.6.3 Comparison to recent meta-epidemiological studies

The absence of an apparent effect of lack of blinding in our study contrasts the effects of lack of double blinding detected in earlier meta-epidemiological studies. As is common in epidemiology, more recent studies have generally estimated less of an effect than early studies. Earlier meta-epidemiological studies on blinding have had somewhat conflicting results but have generally found evidence of some effect of lack of blinding [20, 122, 123]. The 2012 combined re-analysis of seven earlier studies found an average of 13% exaggeration of intervention effects (odds ratios) and 22% in subjective outcomes with lack of or unclear double blinding compared to double blinding [64]. When our data were analysed based on the labelling as "double blind" vs. non-"double blind" (or unclear) for comparison with previous studies we did not find evidence of overestimation. Either for all outcomes combined (ROR 0.99 [95% CrI, 0.86 to 1.09]), or for subjective observer-reported outcomes specifically (ROR 1.11 [95% CrI, 0.86 to 1.44]. However, the CrIs were wide and overlap with the confidence intervals from the previous studies comparing "double blind" and non-“double blind” trials.

A 2012 systematic review included all randomised clinical trials, in any clinical area, directly comparing blinded and non-blinded assessment of the same outcome within the same trial within the same patients [133]. This does sidestep a lot of the important issues of confounding involved in the (indirect) meta-epidemiological approach, which we have taken. This is, however, still observational, as it does not include randomising groups to blinded and non-blinded assessments. The systematic review found evidence of substantial bias, concluding that non-blinded outcome assessors of subjective binary outcomes exaggerated odds ratios by 36% on average, in stark contrast to our results [133]. This discrepancy could be due to confounding in meta-epidemiology (as this is based on comparing trials within meta-analyses rather than within-trial comparisons). Alternatively, it is possible we included less subjective outcome types, or a set of studies that are different in some other way. However, it may be also be that the cohort of trials studied by Hróbjartsson *et al* was not representative of medical trials in general. This is because, the particular trial design using both blinded and non-blinded outcome assessment may have been chosen by the trialists, precisely due to suspicions of a particular susceptibility of the outcome measures to observer bias. Further studies measuring these effects, for trials with different types of outcome measures, would be warranted and could serve to qualify in important ways the assessment of RoB in randomised clinical trials. Such studies would require careful consideration of the outcome types involved, by whom they were assessed, and the specific groups blinded.

The contrast with other results of existing studies is less clear when comparing with studies that have estimated separately the effect of blinding of patients, healthcare providers and outcome assessors. A 2016 systematic review of meta-epidemiological studies found a limited number of smaller studies making such separate comparisons [65]. To explore this further, in our MetaBLIND study, we calculated the predicted mean bias expected in a new meta-analysis without the study characteristic to see if the average bias effect from previous studies lies within this interval. This predictive interval takes into account the between meta-analysis variability. This is shown in Table 4.4. We found all previous studies were consistent with our study except the 2012 Hróbjartsson *et al* study (as described earlier), which compared the blinded and non-blinded assessments within the same trial. This is shown in Table 4.5 and Figure 4.4. Four studies estimated the effect of blinding patients: Balk 2002 (ROR 0.95 [95% CI, 0.70 to1,13]) [134], Bialy 2014 (ROR 0.96 [95% CI, 0.78 to 1.19]) [135], Chaimani 2013 (ROR 0.85 [95% CI, 0.31 to 2.13]) [136] and Nuesch 2009 (difference in standardized mean difference -0.15 [95% CI, -0.39 to0.09]) [137]. Two studies estimated the effect of blinding personnel: Balk 2002 (ROR 0.98 [0.75, 1.20]) and Bialy 2014 (ROR 1.01 [0.82, 1.23]). Four studies estimated the effect of blinding outcome assessors: Balk 2002 (ROR 1.02 [0.82, 1.22]), Bialy 2014 (ROR 1.08 [0.85, 1.33]), Chaimani 2013 (ROR 0.87 [0.63, 1.20]) and Hartling 2014 (ROR 1.00 [0.82, 1.22]) [138]. A 2017 meta-epidemiological study looking at physiotherapy trials and analysing separately the effect of blinding patients and outcome assessors found estimates of the difference in standardized mean difference (dSMD) indicating underestimation of treatment effect associated with lack of blinding, but with wide confidence intervals: lack of blinding of patients dSMD 0.12 (-0.06, 0.30) and lack of blinding of outcome assessors dSMD 0.07 (-0.08, 0.22) [139].

Table 4.4: Predictive Distributions of the Effect of Bias

Study design characteristic	Average bias (95% CrI)	Increase in between-trial heterogeneity (95% CrI)	Variation in average bias (95% CrI)	Predicted mean bias in a new meta-analysis	Predicted bias in a new Trial
Lack of/unclear blinding of participants (versus blinding)					
Patient reported outcomes (Ia)	0.91 (0.61, 1.34)	0.22 (0.02, 0.60)	0.20 (0.01, 0.74)	0.91 (0.41, 1.95)	0.91 (0.32, 2.62)
Patient reported outcomes where the outcome assessor is blinded (Ib)	0.98 (0.69, 1.39)	0.10 (0.01, 0.60)	0.11 (0.01, 0.55)	0.98 (0.56, 1.73)	0.98 (0.46, 2.05)
Lack of/unclear blinding of outcome assessor (versus blinding)					
Observer reported outcomes	1.01 (0.89, 1.14)	0.09 (0.01, 0.30)	0.08 (0.01, 1.21)	1.01 (0.73, 1.36)	1.01 (0.69, 1.45)
Subjective observer reported outcomes (III)	1.01 (0.86, 1.18)	0.05 (0.01, 0.22)	0.09 (0.01, 0.31)	1.01 (0.73, 1.41)	1.01 (0.68, 1.43)
Lack of/unclear double blinding (versus double blinding)					
Outcomes (excluding a harms hypothesis)	1.02 (0.90, 1.13)	0.07 (0.01, 0.19)	0.06 (0.01, 0.27)	1.02 (0.77, 1.28)	1.02 (0.73, 1.36)
Observer outcomes (excluding a harms hypothesis)	1.04 (0.84, 1.25)	0.08 (0.01, 0.23)	0.14 (0.01, 0.57)	1.04 (0.58, 1.81)	1.04 (0.56, 1.87)
Subjective observer outcomes	1.11 (0.86, 1.44)	0.09 (0.01, 0.42)	0.13 (0.01, 0.61)	1.11 (0.62, 2.03)	1.11 (0.56, 2.23)
Lack of blinding personnel (versus blinding personnel)					
Healthcare provider outcomes (IIa)	1.01 (0.84, 1.19)	0.06 (0.01, 0.30)	0.06 (0.01, 0.26)	1.01 (0.75, 1.30)	1.01 (0.68, 1.43)
Healthcare provider outcomes when the outcome assessor is blinded (IIb)	0.97 (0.64, 1.45)	0.10 (0.01, 0.59)	0.13 (0.01, 0.82)	0.97 (0.46, 2.08)	0.97 (0.39, 2.34)

Figure 4.4: Plot of the predictive distributions of the effect of bias from MetaBLIND with the mean effects of bias from other meta-epidemiological studies

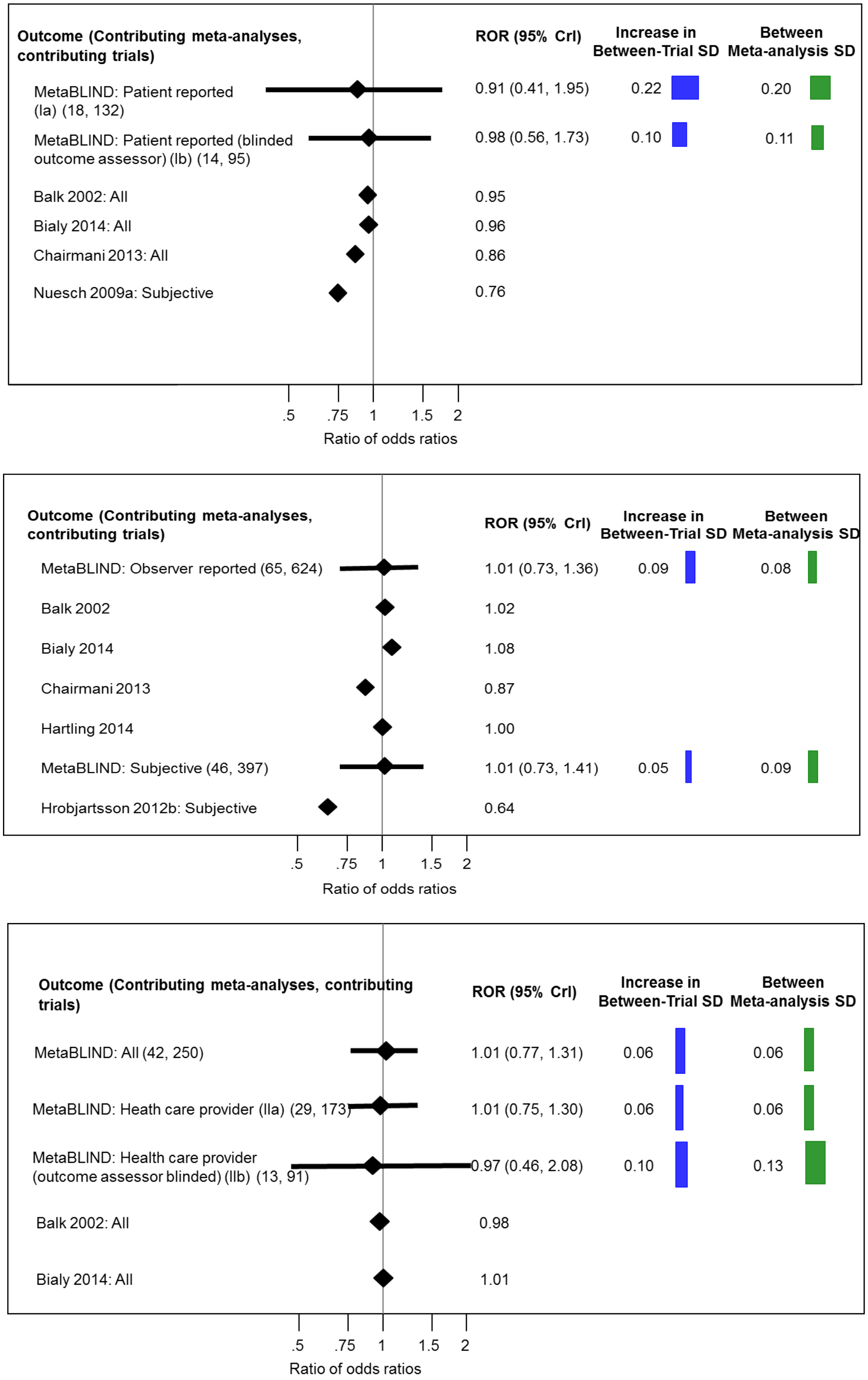


Table 4.5: Comparison with other studies

Study design characteristic	**Average bias (95% CI)**	**Increase in between-trial heterogeneity (95% CrI)**	**Variation in average bias (95% CrI)**
Lack of/unclear blinding of participants (versus blinding)			
Balk 2002: All	ROR 0.95 (0.70, 1.13)		
Bialy 2014: All	ROR 0.96 (0.78, 1.19)		
Chairmani 2013: All	ROR 0.86 (0.70, 1.05)		
Chairmani 2013: Mortality	ROR 0.85 (0.31, 2.13)		
Hróbjartsson 2014b: Other objective	dSMD -0.02 (-0.22, 0.18)		
Hróbjartsson 2014b: Subjective	dSMD -0.56 (-0.71, -0.41)	NA	I^2=60%
Nuesch 2009a: Subjective	dSMD -0.15 (-0.39, 0.09)	NA	0.26
Lack of/unclear blinding of outcome assessor (versus blinding)			
Balk 2002	ROR 1.02 (0.82, 1.22)		
Bialy 2014	ROR 1.08 (0.85, 1.33)		
Chairmani 2013	ROR 0.87 (0.63, 1.20)		
Hartling 2014	ROR 1.00 (0.82, 1.22)		
Hróbjartsson 2012: Subjective	ROR 0.64 (0.43, 0.96)	NA	I^2=45%
Hróbjartsson 2013: Subjective	dSMD -0.23 (-0.40, -0.06)	NA	I^2=46%
Lack of/unclear double blinding (versus double blinding)			
BRANDO (Savović 2012): All outcomes (95% CrI)	ROR 0.87 (0.79, 0.96)	0.20 (0.02, 0.39)	0.17 (0.03, 0.32)
Moher 1998	ROR 1.11 (0.76, 1.83)		
BRANDO (Savović 2012): Mortality (95% CrI)	ROR 0.92 (0.80, 1.04)].	0.09 (0.01, 0.44)	0.08 (0.01, 0.42)
Unverzagt 2013	ROR 0.84 (0.69, 1.02)		
BRANDO (Savović 2012): Other objective (95% CrI)	ROR 0.93 (0.74, 1.18)].	0.10 (0.01, 0.50)	0.20 (0.02, 0.85)
BRANDO (Savović 2012): Subjective (95% CrI)	ROR 0.78 (0.65, 0.92)].	0.24 (0.02, 0.45)	0.20 (0.04, 0.39)
Lack of blinding personnel (versus blinding personnel)			
Balk 2002	ROR 0.98 (0.75, 1.20)		
Bialy 2014	ROR 1.01 (0.82, 1.23)		
Lack of blinding patients/personnel (versus blinding patients/personnel)			
Hartling 2014	dSMD 0.00 (-0.09, 0.09)		

4.6.4 Conclusions and rationale for remainder of thesis

We found no evidence of a difference, on average, in estimated treatment effect between randomised clinical trials with blinded and non-blinded patients, between trials with blinded and non-blinded healthcare providers, and between trials with blinded and non-blinded outcome assessors. Lack of blinding of patients in trials with patient-reported outcomes was associated with some increase in between-trial heterogeneity. The apparent lack of a major average impact of blinding on estimated treatment effects is surprising and at odds with methodological standard practises and some empirical studies, but not with several more recent meta-epidemiological studies. It is therefore unclear to what extent our results reflect that blinding is less important than previously believed, or reflect limitations in the meta-epidemiological approach, for example residual confounding by other trial characteristics.

We have extended current statistical methodology in meta-epidemiology to (i) combine binary and continuous outcomes and (ii) use meta-analysis covariates in order to produce stratified analyses by particular groups. It is then possible for a trialist to use these outputs to conduct bias adjustment of their own trial; considering whether it should be based on the most relevant other trials/meta-analyses (similar interventions, outcomes etc). Thereby allowing the trialist to correct and down weight the results of their trial at a high RoB. However, we have treated 'high RoB' trials as 'high or unclear RoB' trials. Therefore, if a trialist wanted to use the results from these models to adjust the treatment effect in their trial, the effect of an unclear RoB trial on treatment effects is likely not applicable. A trialist, working on their own trial, will likely know which methodological components are at a high or low RoB. Further work should therefore explore how these models can account for unclear RoB trials. These findings are also consistent with the results from our qualitative work in Chapter 3 which suggested current meta-epidemiological methods needed extending for trialists to use them in practice.

We therefore extend current methodology of meta-epidemiology studies to categorical and numerical study characteristics in Chapter 5. We address the role of unclear RoB trials which have so far been treated as ‘high risk’. We explore how this can be modelled categorically so that a trialist can adjust their high RoB trial. Second, we extend these models to look at the association of continuous study characteristics on estimated treatment effects, which has so far not been considered.

5 Statistical considerations in meta-epidemiological evidence of non-dichotomous study characteristics

5.1 Context and overview

Current methodology for modelling meta-epidemiological evidence has looked at binary and dichotomous study characteristics; specifically, comparing high RoB trials to low RoB trials. However, most RoB assessments from individual trials are classified as high, low or unclear, and there is still no consensus on how unclear RoB trials should be classified in meta-epidemiological analyses [21].

Researchers conducting meta-epidemiological studies often group unclear RoB trials with high RoB trials [55, 64, 81]. However, the assumption that all unclear RoB trials are all at a high RoB could be too strong, as there could be other reasons why a RoB assessment is unclear, such as a trial being poorly reported [21]. Moreover, if a trialist wanted to use the results from such a model to adjust the treatment effect in their trial, the effect of an unclear RoB trial on treatment effects, is likely not applicable. Trialists working on their own trial will likely know if it is either at a high or low RoB, whether that be overall, or for specific methodological components, i.e. there is no 'unclear' category. For example, a trialist will know if their trial was unblind but not necessarily whether people sought additional healthcare. We would also expect estimated associations to be diluted by assuming all unclear RoB trials are high RoB trials, when in fact some of them will be low RoB trials. Another possible approach by those conducting meta-epidemiological research is to conduct two separate analyses: high vs low and unclear vs low RoB trials [64]. However, this often results in a loss of information and the inclusion of much fewer studies.

It can be hypothesised that smaller studies are at a higher RoB than larger studies because smaller studies could be conducted less rigorously [21]. Previous meta-epidemiological studies have explored this by dichotomising the continuous variable (study size) to define a binary characteristic [140]. Methods, such as the Welton *et al* model described in the previous chapter (Section 4.3.1) can then be applied [140]. However, dichotomisation can result in a loss of information. We may therefore prefer to model the study characteristic of interest, sample size, continuously, to make use of *all* available information. From an appraisal perspective, it provides an empirical association to allow a trialist to see whether smaller trials are associated with more extreme treatment effects.

In this chapter we develop and describe approaches to model the association between treatment effect estimates and (i) categorical study characteristics and (ii) continuous study-level characteristics. We use high vs unclear vs low RoB trials and sample size as examples. We make use of the database from the ROBES meta-epidemiological study [81]. We use the ROBES database, rather than the MetaBLIND database, to provide an exemplar of these methods using a more typical meta-epidemiological result. This database consists of 228 meta-analyses and 2443 trials each with a completed RoB table, extracted from the April 2011 issue of the Cochrane Database of Systematic Reviews.

The remainder of the chapter is structured as follows. Section 5.2 describes the meta-epidemiological models for categorical study characteristics, with an application to the role of unclear RoB trials. Section 5.3 similarly details the rationale and application of using a continuous study characteristic, with an application to the association between sample size and estimated treatment effects. This chapter concludes with discussion of the findings and highlights the potential advantages these models can have to inform bias adjustment for a new trial.

5.2 Meta-epidemiological considerations for categorical study characteristics with an application to the role of unclear risk of bias trials

5.2.1 Description of models

In this section a description of each of the models briefly introduced above are given.

A bivariate meta-epidemiological model

As described in the earlier section, an individual trial in a meta-analysis can be classified as either high, low or unclear RoB for a particular methodological characteristic. An analyst may want to model a study characteristic with three categories. One way to do this is to compare between two categories and have a common reference group. Therefore, correlation between the high vs low and unclear vs low bias estimates may arise from the common reference group of low RoB trials. We first describe the bivariate meta-epidemiological model, which is an extension of the Welton *et al* model described in Chapter 4.

We assume each study i, in meta-analysis m which has the baseline value of the categorical variable (i.e. in this case has been assessed as low RoB), provides an estimate of the underlying treatment effect $\delta_{i,m}$. As in the Welton *et al* model, we assume a normal random effects distribution for $\delta_{i,m}$ with mean d_m and variance τ_m^2 specific to meta-analysis m, given by equation (4.3). τ_m^2 represents the between study variability between the low RoB trials or those who have the baseline level of the variable. Since we have a three-level categorical variable, two indicator variables are created for the unclear RoB trials, denoted $IU_{i,m}$ and high RoB trials, denoted $IH_{i,m}$. When $IU_{i,m} = 1$ and $IH_{i,m} = 0$ this indicates the unclear RoB trials. When $IU_{i,m} = 0$ and $IH_{i,m} = 1$ this indicates the high RoB trials.

The outcome $r_{a,i,m}$ for arm a of trial i in meta-analysis m is assumed to have a binomial likelihood (for given denominator $n_{a,i,m}$), given by equation (4.1), in Section 4.3.1. The probability of success, $p_{a,i,m}$, is modelled by a logistic regression, with reference to the regression equation (4.2) where the binary study characteristic has been replaced by two indicator variables:

$$logit(p_{a,i,m}) = \begin{cases} \mu_{i,m} & \text{control arm} \\ \mu_{i,m} + \delta_{i,m} + \beta_{1i,m}IU_{i,m} + \ \beta_{2i,m}IH_{i,m} & \text{treatment arm} \end{cases} \tag{5.1}$$

To account for the correlation between the unclear vs low, $\beta_{1i,m}$, and high vs low, $\beta_{2i,m}$, RoB trials, we assume a bivariate normal distribution for these parameters.

We first outline the normal framework for a bivariate random effects model [141]:

$$\begin{pmatrix} \beta_{1i,m} \\ \beta_{2i,m} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} b_{1m} \\ b_{2m} \end{pmatrix}, \Omega\right) \text{ where } \Omega = \begin{pmatrix} \kappa_1^2 & \rho_w\kappa_1\kappa_2 \\ \rho_w\kappa_1\kappa_2 & \kappa_2^2 \end{pmatrix} \tag{5.2}$$

As in the Welton *et al* model, b_{1m} is the difference in treatment effects between the unclear and low RoB trials in the mth meta-analysis. Similarly, b_{2m} is the difference in treatment effects between the high and low RoB trials in the mth meta-analysis.

The average SD increase in between-trial heterogeneity among the unclear RoB trials and the high RoB trials is κ_1 and κ_2, respectively. In contrast to the Welton *et al* model, κ_1 and κ_2 are estimated in the within meta-analysis variance-covariance matrix, to account for the correlation between $\beta_{1i,m}$ and $\beta_{2i,m}$. The within meta-analysis variance-covariance matrix is given by Ω where $\kappa_1^2 = var(\beta_{1i,m})$ and $\kappa_2^2 = var(\beta_{2i,m})$ and $cov\,(\beta_{1i,m},\ \beta_{2i,m}) = \rho_w\kappa_1\kappa_2$. ρ_w indicates whether $\beta_{1i,m}$ and $\beta_{2i,m}$ are correlated and is estimated from meta-analyses which include high, unclear and low RoB trials.

In the second part of the model, we use these ROR bias estimates per meta-analysis of, b_{1m}

and b_{2m} and model the correlation between them:

$$\begin{pmatrix} b_{1m} \\ b_{2m} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} b_{01} \\ b_{02} \end{pmatrix}, \Sigma\right) \text{ where } \Sigma = \begin{pmatrix} \varphi_1^2 & \rho_B\varphi_1\varphi_2 \\ \rho_B\varphi_1\varphi_2 & \varphi_2^2 \end{pmatrix} \tag{5.3}$$

Here, b_{01} is the average ROR of the unclear vs low RoB trials and b_{02} is the average ROR of the high vs low RoB trials (on the log scale). The between meta-analysis variance covariance matrix is given by Σ and contains the between meta-analysis variances of the high vs low and unclear vs low RoB effect estimates (in the diagonal φ_1^2, φ_2^2). The between meta-analysis covariance for the high vs low and unclear vs low $cov\left(b_{1m}, b_{2m}\right) = \rho_B\varphi_1\varphi_2$. ρ_w is interpreted as the correlation in the estimated study-specific biases in the unclear vs low RoB trials and high vs low RoB trials, within meta-analyses. ρ_B is interpreted as the correlation in the estimated meta-analysis specific biases in the unclear vs low RoB trials and high vs low RoB trials, between meta-analyses. Therefore, in the context of meta-epidemiological models, the index of w and B refer to the correlation in biases *within* meta-analyses and *between* meta-analyses.

Specification of a prior distribution for variance-covariance matrices (Ω and Σ) can be problematic [142, 143, 144, 145, 146]. The two most common ways to do this is to either use a conjugate prior on the variance-covariance matrix, given as a 2-dimensional Wishart prior [142, 146] or re-parametrise using a series of univariate distributions, where the mean of one of the distributions is conditional on the other, known as the product normal formulation [143, 144, 145]. To compare our results to the standard model, we use the product normal formulation model to allow the separation of the variance-covariance matrices and use the same prior distributions for a true comparison. It has also been shown that the inverse-Wishart prior is particularly sensitive its scale [128]. In either case, the prior distributions chosen must ensure the variance covariance matrix is positive semi-definite [144, 145]. We ensure the variance covariance matrix is positive semi-definite by using the same prior for the variances as in the standard Welton *et al* models, and by choosing a uniform prior distribution for the correlation which restricts it to values between -1 and $+1$. In a sensitivity analysis the correlation was restricted to values between 0 and $+1$.

Product normal formulation

The product normal is a re-parameterisation of the bivariate model. This is convenient for estimation purposes, as it allows priors to be placed on the SDs and correlations directly. It is therefore the same model, just written differently. We assume the estimated bias in the unclear RoB trials, $\beta_{1i,m}$, are normally distributed with a meta-analysis specific bias b_{1m} and between study variability κ_1^2 which is estimated as fixed across all unclear RoB trials.

$$\beta_{1i,m} \sim N(b_{1m}, \kappa_1^2) \tag{5.4}$$

We re-parametrise the bias parameters for the high RoB trials, $\beta_{2i,m}$ so that their distributions are conditional on the bias in the unclear RoB trials $\beta_{1i,m}$:

$$\beta_{2i,m}|\beta_{1i,m} \sim N(\eta_{2i,m}, \kappa.re_2^2) \tag{5.5}$$

$$\eta_{2i,m} = b_{2m} + \lambda_w(\beta_{1i,m} - b_{1m}) \tag{5.6}$$

$$\kappa.re_2^2 = \kappa_2^2(1 - \rho_w^2) \tag{5.7}$$

$$\lambda_w = \rho_w \frac{\kappa_2}{\kappa_1} \tag{5.8}$$

We similarly re-parametrise the mean bias parameters for the high RoB trials, b_{2m} so that their distributions are conditional on the mean bias in the unclear RoB trials b_{1m}:

$$b_{1m} \sim N(b_{01},\ \varphi_1^2) \tag{5.9}$$

$$b_{2m}|b_{1m} \sim N(\mu_{02,m},\ \varphi.re_2^2) \tag{5.10}$$

$$\mu_{02,m} = b_{02}{+}\lambda_B(b_{1m} - b_{01}) \tag{5.11}$$

$$\lambda_B = \rho_B\frac{\varphi_2}{\varphi_1} \tag{5.12}$$

$$\varphi.re_2^2 = \varphi_2^2(1 - \rho_B^2) \tag{5.13}$$

Any given meta-analysis within a meta-epidemiological study might include:

- (i) A combination of low, unclear and high RoB trials
- (ii) A combination of low and unclear RoB trials only
- (iii) A combination of low and high RoB trials only

In a Bayesian hierarchical modelling framework, the contrast (ROR) between high versus low RoB trials and between unclear versus low RoB trials can be estimated in each meta-analysis, regardless of which of these three groups the meta-analysis belongs to. By allowing for correlation between these contrasts within meta-analyses, we can 'borrow strength' across categories as well as across meta-analyses. As described in Chapter 4, Section 4.3.1, meta-analyses with only one high risk or only one low risk trial are prevented from contributing to the estimation of κ.

Probability model estimating the proportion of unclear RoB trials which are high risk

We previously developed a bivariate model which makes use of all available data in the meta-epidemiological study, whilst simultaneously estimating a high vs low average RoB estimate and an unclear vs low average RoB estimate. However, we can also model the unclear RoB trials in a different way and quantify how likely they are truly high RoB trials using a probability model [82].

In the context of a NMA, Dias *et al* hypothesised that some trials may be of a poor methodological quality and overestimating the true effect of some interventions [82]. However, a lot of the trials were assessed as an 'unclear RoB'. Therefore, Dias *et al* developed a model whereby any study with an unclear RoB classification has a probability p of being at RoB. This model provides a bias-adjusted analysis and can identify specific trials (with an 'unclear RoB') as having a high probability of bias.

We extend the probability of bias model applied to a NMA by Dias *et al* [82], for use in the context of meta-epidemiological data. Rather than categorising trials in the standard Welton *et al* model as high or unclear vs low RoB (equation (4.4)), we can extend this classification to include a probability coefficient for the unclear risk of trials, $B_{i,m}$ as in the Dias *et al* model, which is drawn from a Bernoulli distribution:

$$C_{i,m} = \begin{cases} 1 & \text{if study } i \text{ at RoB} \\ B_{i,m} & \text{if RoB of study } i \text{ is unclear} \\ 0 & \text{if study } i \text{ not at RoB} \end{cases} \tag{5.14}$$

where

$$B_{i,m} \sim Bernoulli(p) \tag{5.15}$$

$B_{i,m}$ is interpreted as the posterior probability that an unclear RoB study i in meta-analysis m is at RoB. Therefore, the overall probability that a study with unclear RoB is actually at risk is denoted by the posterior mean of p. Since this is a quantity that is estimated in the model, p has the following prior distribution:

$$p \sim uniform(0,1) \tag{5.16}$$

Using the same notation as earlier (equation (5.1)), we modify the regression to include

$B_{i,m}$:

$$logit(p_{a,i,m}) = \begin{cases} \mu_{i,m} & \text{control arm} \\ \mu_{i,m} + \delta_{i,m} + \beta_{i,m}(B_{i,m}IU_{i,m} + IH_{i,m}) & \text{treatment arm} \end{cases} \tag{5.17}$$

Since $IU_{i,m}$ and $IH_{i,m}$ are dummy variables, we link this to $C_{i,m}$ in equation (5.14). When study i is at RoB, then $IH_{i,m} = 1$ and $IU_{i,m} = 0$, the terms inside the brackets of equation (5.17) reduce to 1, that is, $C_{i,m} = 1$. When study i is at an unclear RoB, then $IH_{i,m} = 0$ and $IU_{i,m} = 1$, the terms inside the brackets of equation (5.17) reduce to $B_{i,m}$, that is, $C_{i,m} = B_{i,m}$. When study i is not at RoB, then $IH_{i,m} = 0$ and $IU_{i,m} = 0$, the terms inside the brackets of equation (5.17) reduce to 0, that is, $C_{i,m} = 0$.

We expect that for unclear RoB trials that have a treatment effect closer to the average treatment effect of the high RoB trials to have a higher probability of being truly high risk. Additionally, by allowing $B_{i,m}$ to be estimated from the unobserved data, we do not have to assume the proportion of unclear RoB trials is constant.

Probability model - extended to use sample size as a predictor of how likely an unclear trial is high risk

Rather than basing the probability that an unclear RoB trial is actually high risk on its treatment effect alone, we extend the model to predict the probability of whether a trial is high risk using its sample size. We hypothesise that smaller trials are more likely to have study limitations, such as inadequate blinding and could therefore be used as a predictor of whether an unclear RoB trial is actually at a high RoB. This may mean smaller trials will have more extreme or exaggerated treatment effects than larger trials.

Rather than $B_{i,m}$ being drawn from a Bernoulli distribution with mean p, we modify equation (5.15) to estimate a trial specific $p_{i,m}$, which is predicted by some transformation of sample size, denoted by $x_{i,m}$:

$$B_{i,m} \sim Bernoulli(p_{i,m}) \tag{5.18}$$

$$logit(p_{i,m}) = mx_{i,m} + c \tag{5.19}$$

This transformation could be $log(n_{i,m})$ or $\frac{1}{n_{i,m}}$ and is discussed in more detail in Section 5.3). The coefficient of $x_{i,m}$ is m which can be interpreted as the log odds ratio of being high risk compared to low risk for each unit increase in $x_{i,m}$. c is the intercept and the log odds of being high risk when $x_{i,m} = 0$.

In practice, $x_{i,m}$ will be centred for model efficiency and improved interpretation. The intercept c will then be interpreted as the log odds of being high risk for an average sized trial. The estimated $\beta_{i,m}$ is the bias effect for a high vs low RoB in study i in meta-analysis m.

However, it is not possible to estimate m based on the unobserved data in equation (5.18). We therefore borrow the 'slope' (m) from the observed high vs low RoB trials. Using the observed data, of the high and low risk of trials where $high_j$ is an indicator variable for j trials which are either high or low RoB. $high_j$ is therefore Bernoulli distributed with the probability that a trial is at a high RoB, denoted by, $probhigh_j$, when x_j is used as a predictor.

$$high_j \sim Bernoulli(probhigh_j) \tag{5.20}$$

x_j is a function of sample size.

$$logit(probhigh_j) = mx_j + \alpha \tag{5.21}$$

This allows the intercept, α, to be different to c (exchangeable) but the slope, m, (impact of trial size) is shared.

The WinBUGS code for all the models described can be found in Appendix B.

5.2.2 Case study application

We analysed data from 1678 trials included in 144 binary outcome meta-analyses that were informative to detect differences in intervention effects between either high vs low and/or unclear vs low RoB trials i.e. all meta-analyses in ROBES that were informative. This dataset was derived by creating two separate datasets and then combining these. We created a dataset containing informative meta-analyses to detect differences in intervention effects between either high vs low RoB trials i.e., meta-analyses that included at least one high risk and at least one low risk trial. There were 917 trials included in 97 meta-analyses. We similarly derived a dataset which included only meta-analyses which were informative to detect differences in intervention effects between either unclear vs low risk trials. This dataset consisted of 1029 trials included in 98 meta-analyses. When these two datasets were combined there were 51 meta-analyses consisting of 268 trials which were in both. Therefore 35% (51/144) of the meta-analyses in our analysis dataset include high and unclear (and low) risk trials, while 32% (46/144) contain only high and low and 33% (47/144) contain only unclear and low.

The same priors as in the univariate (Welton *et al*) model are used with the addition of a uniform prior (-1,1) for the correlations of ρ_w and ρ_B. Sensitivity analyses were conducted with a uniform prior (0,1), assuming a positive correlation between the unclear and high bias estimates.

5.2.3 Results

Results for each of the models described above are given in Table 5.1.

We found as in the original Welton model, when we code the dichotomous variable as high or unclear versus low RoB, that treatment effects were exaggerated by 14% (ROR of 0.86, 95% CrI, 0.80 to 0.93). The estimated increase in between trial heterogeneity amongst high or unclear vs low RoB trials was SD=0.11 (95% CrI, 0.02 to 0.25) and the between meta-analysis variation in average bias was estimated to be SD=0.09 (95% CrI, 0.01 to 0.20).

When we model only the high vs low risk trials and the unclear vs low risk trials, i.e. stratified analyses, we see that the average bias is slightly bigger in high vs low risk trials compared to the unclear vs low RoB estimates. In the high vs low RoB analysis, treatment effects were exaggerated by 18% on average (ROR of 0.82, 95% CrI, 0.74 to 0.92) whilst the average bias in the unclear vs low RoB trials was estimated closer towards the null (ROR of 0.89, 95% CrI, 0.74 to 0.92). However, in both cases the CrIs are wider, highlighting more uncertainty, as there are less trials and meta-analyses included. The estimated increase in between trial heterogeneity amongst the unclear RoB trials is lower SD=0.06 (95% CrI, 0.01 to 0.20), in contrast to, the between trial heterogeneity which is higher amongst the high risk trials SD=0.17 (95% CrI, 0.02 to 0.36). However, the CrIs do overlap.

Bivariate normal model

The bivariate normal model estimates the ROR as almost identical as the separate models when only the unclear RoB trials are included and when only the high RoB trials are included. However, there was no gain in precision in the average biases when modelled simultaneously. When a uniform (-1,1) was used, ρ_w was estimated to be -0.04 (95% CrI, -

0.94 to 0.97) and ρ_B was estimated as -0.08 (95% CrI -0.94, 0.91). Similarly, when a uniform (0,1) was used, this gave ρ_w an estimate of 0.51 (95% CrI 0.03, 0.98) and ρ_B an estimate of 0.45 (95% CrI 0.02, 0.95). However, the results from each of our parameters of interest remained the same.

The estimated between-study heterogeneity, κ, of the high vs low RoB trials remained the same in the bivariate normal model as it did when they were estimated univariately. This estimated increase in between trial heterogeneity was 0.18 (95% CrI (0.02, 0.35)). However, the estimated between-study heterogeneity of the unclear vs low RoB trials increased from 0.06 (0.01, 0.20) when modelled univariately to 0.14 (0.01, 0.29) in the bivariate normal model. The between meta-analysis variation in average bias,φ, of the unclear vs low RoB trials increased from 0.08 (0.01, 0.20) in the univariate model to 0.14 (0.02, 0.29) in the bivariate model. Similarly, the between meta-analysis variation in average bias, φ, of the high vs low RoB trials increased from 0.06 (0.01, 0.21) in the univariate model to 0.10 (0.01, 0.32) in the bivariate model. It is likely these differences can be attributed to the variance components being estimated poorly in meta-epidemiological models due to their wide CrIs.

Probability models

The probability model estimated the average bias for high vs low RoB trials as ROR=0.81 (95% CrI, 0.74 to 0.90). This gives a similar ROR and narrower CrI, in contrast to when the high vs low RoB trials are estimated univariately, ROR of 0.82 (95% CrI, 0.74 to 0.92). This may because the probability model includes all 1678 trials (144 meta-analyses) whilst the high vs low RoB univariate analysis includes 97 meta-analyses (917 trials).

The estimated overall probability is given by 0.44 (95% CrI, 0.13 to 0.88). This probability varied by trial, such that the more extreme unclear RoB trials with treatment effect estimate further from the null and closer to that of high risk trials had a higher probability. Figure 5.1 shows the distribution of individual probabilities from each of the unclear RoB trials, i.e. the probability an unclear RoB trial is high risk. Towards the right-hand side of Figure 5.1, we can see the more extreme unclear RoB trials, with individual treatment effects which are further from the null, have a higher probability of being at a higher RoB.

Figure 5.2 shows the distribution of study sizes from each of the RoB categories; highlighting that unclear RoB trials do have study sizes which are more similar to the high RoB trials and smaller than the low RoB trials.

When we use sample size as a predictor of whether an unclear RoB trial is actually high risk, we see the ROR is the same as in the probability model, given by ROR= 0.81 (95% CrI, 0.73 to 0.90). Figure 5.3 shows the distribution of the mean posterior distribution of B, when being predicted by sample size, however we see a slightly different pattern to Figure 5.1. To see this more clearly, we plot the individual probabilities of an unclear RoB trial being high risk from both the probability model (green) and the probability model (black) when predicted by sample size in Figure 5.4. We broadly see that as the individual study size of a trial increases, the probability that a trial is high risk decreases. We can also see the probabilities are higher for smaller sample sizes and lower for bigger sample size in the probability model predicted by sample size rather than the probability model.

We can also plot the slope of the sample size regression from Table 5.2. Figure 5.5 shows that as sample size increases the predicted treatment effect decreases.

Table 5.1: Posterior summaries from each of the models described, examining the influence of not blinding participants.

	Welton *et al* model			**Bivariate random effects model**		**Probability model**	**Probability model**
	High/unclear vs low	*Unclear vs low*	*High vs low*	*Unclear vs low*	*High vs low*	*p=0.44* *(95% CI, 0.13, 0.88)*	*Using sample size*
	N=1678, MA=144	**N=1029, MA=98**	**N=917, MA=97**	**N=1678, MA=144**		**N=1678, MA=144**	**N=1678, MA=144**
ROR	0.86	0.89	0.82	0.89	0.83	0.81	0.81
(95% CrI)	(0.80, 0.93)	(0.80, 0.97)	(0.74, 0.92)	(0.81, 0.98)	(0.75, 0.92)	(0.74, 0.90)	(0.73, 0.90)
κ	0.11	0.06	0.17	0.14	0.18	0.18	0.18
(95% CrI)	(0.02, 0.25)	(0.01, 0.20)	(0.02, 0.36)	(0.01, 0.29)	(0.02, 0.35)	(0.03, 0.35)	(0.03, 0.35)
φ	0.09	0.08	0.06	0.14	0.10	0.14	0.13
(95% CrI)	(0.01, 0.20)	(0.01, 0.20)	(0.01, 0.21)	(0.02, 0.29)	(0.01, 0.32)	(0.02, 0.30)	(0.02, 0.29)

All priors on the variance components κ, φ and τ were the modified Inverse Gamma (0.001, 0.001) distribution, as described in Section 4.3.1.

Figure 5.1: Plot of the individual probabilities of unclear RoB trials being high risk trials against their estimated treatment effects and 95% CrIs.

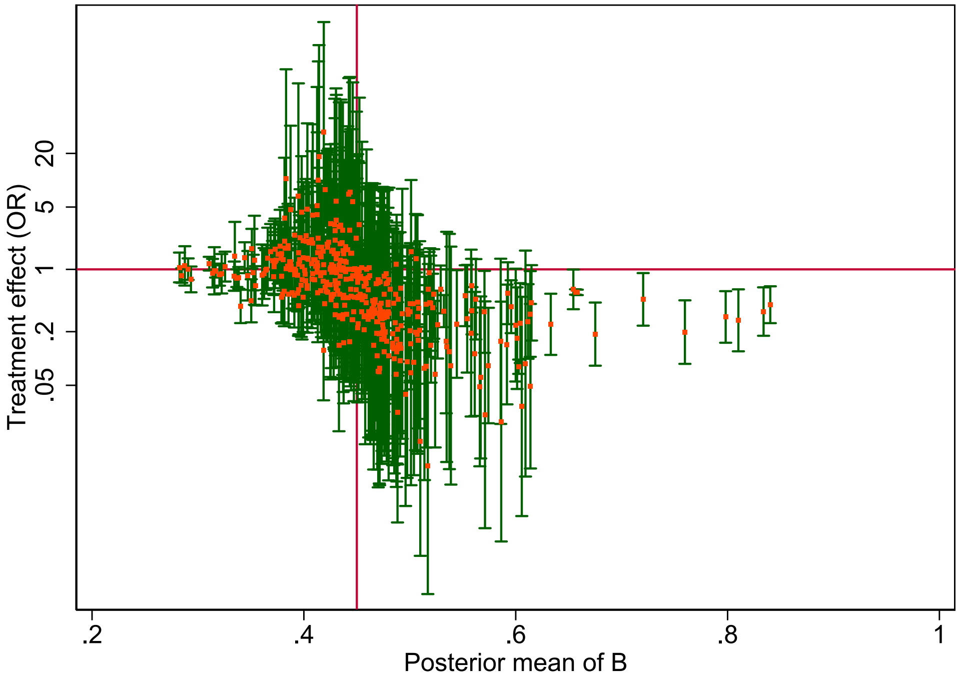
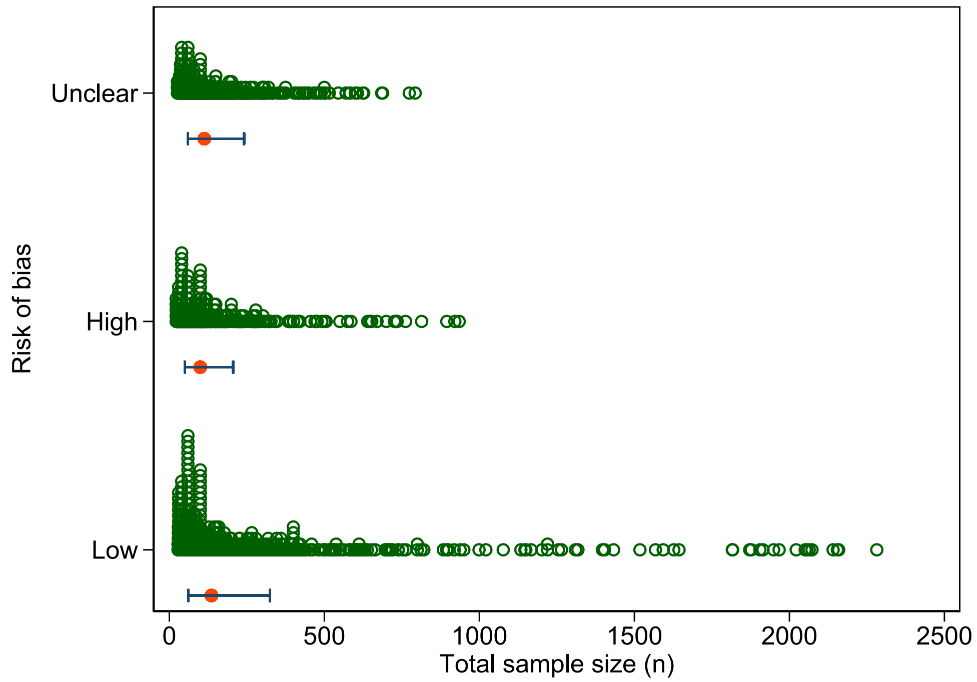


Figure 5.2: Summary of sample size for each RoB category.



Unclear RoB trials: median sample size is 113, IQR: 60 to 241, 95% percentile: 26-793.
High RoB trials: median sample size is 100, IQR: 50 to 206, 95% percentile: 22-935.
Low RoB trials: median sample size is 136, IQR: 61 to 324, 95% percentile: 28-2282.

Table 5.2: Probability model using sample size, results in more detail

	m **(95% CrI)**	c **(95% CrI)**	α **(95% CrI)**
$log(n)$	-0.26 (-0.36, -0.16)	-0.80 (-0.92, -0.68)	-0.09 (-5.14, 11.41)

from $m(x_{i,m}) + c$, where $x_{i,m}$ is centred, m is the slope (shared between the observed and unobserved); c is the intercept of the unobserved sample size regression and α is the intercept in the observed sample size regression. Reported on the log scale.

Figure 5.3: Plot of the individual probabilities of unclear RoB trials being high risk trials, when predicted by sample size against their estimated treatment effects and 95% CrIs.

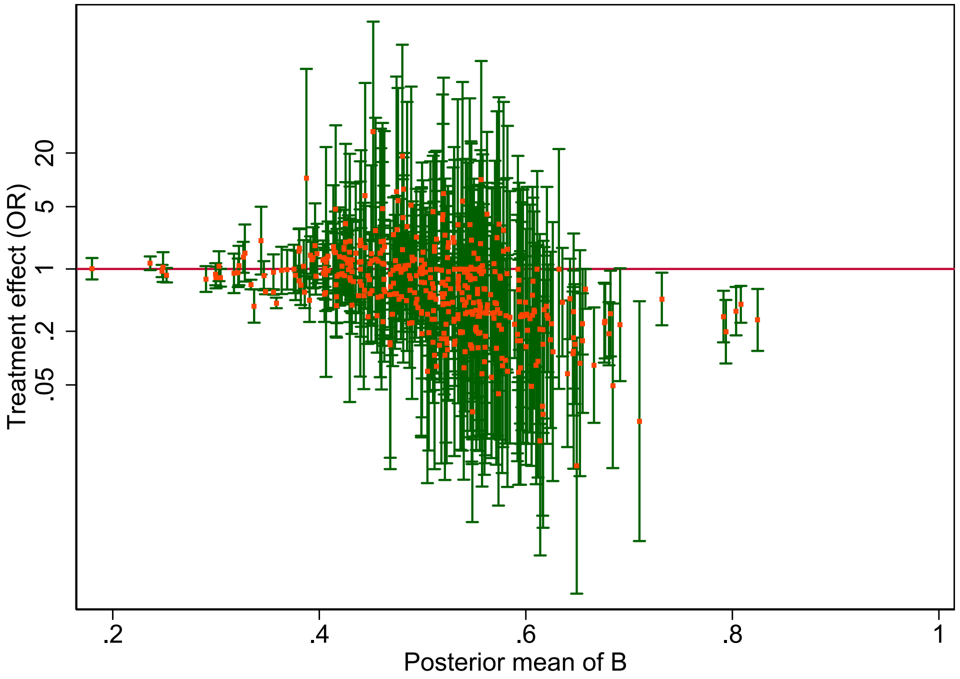


Figure 5.4: Plot of the mean posterior distribution of B against study size, for all unclear RoB trials, for both the probability model and the probability model when using sample size as a predictor. Shown for smaller study sizes.

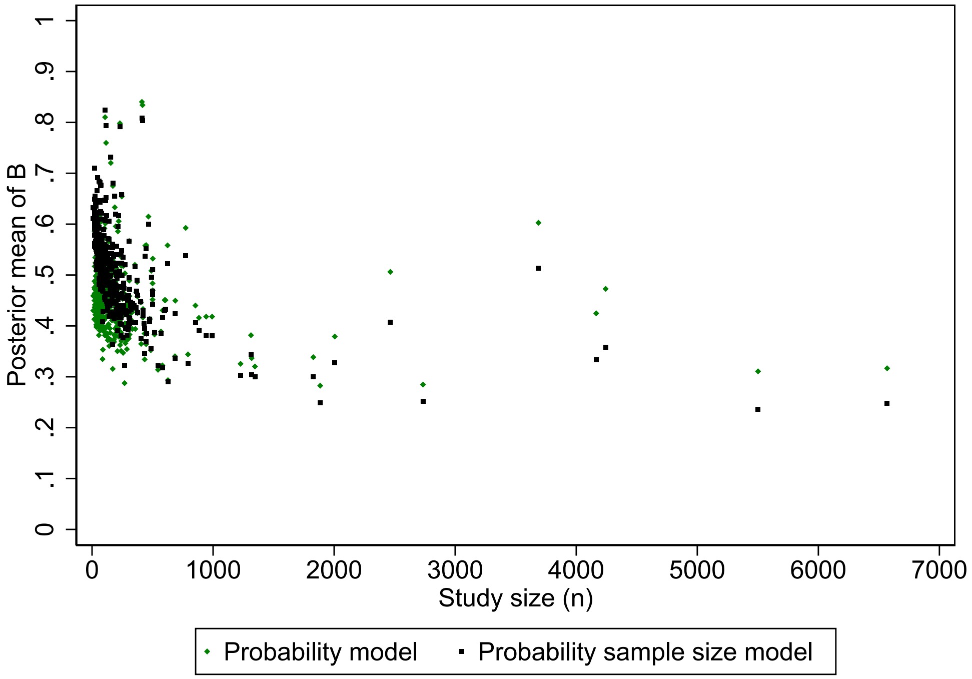
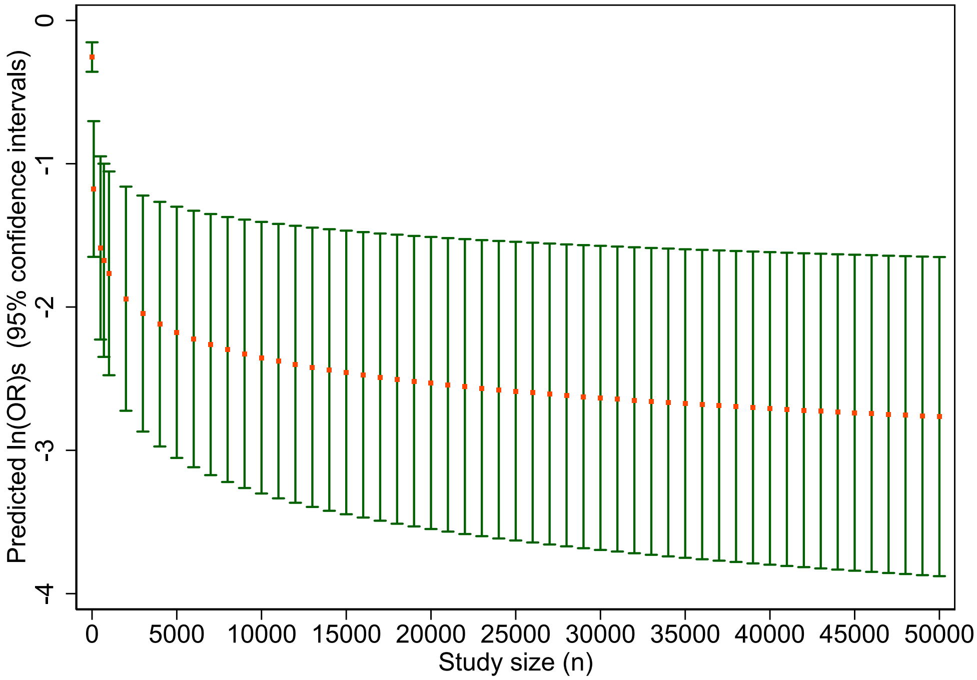


Figure 5.5: Plot of the individual probabilities of unclear RoB trials being high risk trials, when predicted by sample size against their study size.



5.2.4 Using the outputs of meta-epidemiological models to inform a new trial

The models developed in the first part of this chapter can be used to formulate a prior distribution for the likely amount of bias in a trial when analysed in a Bayesian framework. External information from these meta-epidemiological studies can be used to form the prior distribution, related to the bias characteristic and type of outcome. This adds uncertainty around the treatment effect to appropriately down-weight them. The estimated (posterior) intervention effect in a new study will change in light of which prior distribution is formulated and the assumptions made in each.

Bias adjustment in this way has not (to our knowledge) been used in an individual trial context and rarely used within a meta-analysis context [80, 140]. We assume that the high vs low bias estimate and its associated κ and φ are used from either the bivariate model or probability model to form a prior distribution on the amount of bias suspected to adjust the treatment effect in a new trial. There are several possible ways of forming the prior based on these outputs. We briefly describe possible prior distributions that could be considered. These include:

- the predictive distribution for bias in a new trial; a predictive distribution is drawn from the ROR distribution but additionally reflects the uncertainty as to where a randomly selected study setting might lie in this distribution, as well as the uncertainty in κ and φ [147]. This assumes that the target setting for the decision is "similar" to those in the studies included in the meta-epidemiological study. This is a prior based on the mean and all variation. Trialists may be happier to use this in practice as it includes all the uncertainty in the bias estimate, however, it will then also lead to imprecise estimates.

- Shrunken meta-analysis specific estimate; we can use the shrunken estimate drawn from the ROR (depending on the relative size of the study and the degree of heterogeneity) of a specific meta-analysis [147]. This meta-analysis could be chosen based on how closely related it is to a trial, for example based on the outcome measured or inclusion/exclusion criteria. The meta-analysis specific ROR will be more precisely estimated than the study estimates alone because it is "borrowing strength" from the other study estimates.

- a prior based only on the mean ROR; a prior based only on the mean will adjust the biased study and add only uncertainty from the ROR CrI. This may not be as extreme as some of the other options for bias adjustment and would allow the analyst to assess the sensitivity of their findings.

- a prior based on bias in an extreme meta-analysis (with large bias) and within-meta-analysis variation. This prior is likely to be used when a trialist has a very strong suspicion that their trial result is at a very high RoB. To obtain an extreme meta-analysis we take the 95% upper range of the between-meta-analysis variation. The variance is obtained by adding the variance (uncertainty) of the mean bias to the within-meta-analysis variance in bias. This prior should make a marked adjustment to the study and add considerable uncertainty.

5.2.5 Summary

In current meta-epidemiological methodology, a binary variable is modelled, usually 'unclear' is grouped with 'high' RoB. Therefore, the difference investigated is between a high or unclear RoB trial characteristic compared to the study without the characteristic. What may be more relevant to trialists working on individual trials is the difference between a high vs low RoB trial. Rather than excluding the unclear of bias trials, we developed a bivariate model which includes all the data and allows for high vs low estimates and an unclear vs low estimate. The key thing here is that it not only makes maximal use of the data but also in principle allows us to borrow strength between the high and unclear RoB estimates, which may be expected to be correlated. The bivariate (conditional product normal) model does have an extra computational complexity compared to the standard Welton *et al* model which took several more days for convergence to be achieved. This may be a significiant draw back to this model.

In the probability model, we estimated how likely a trial classified as unclear is in fact a high RoB trial. In our dataset, we found that the probability model gave an average ROR estimate which was consistent with the ROR from the high vs low RoB model but slightly more precisely estimated. The model estimates the probability that blinding was *not* adequate in each trial where reporting of blinding was unclear. In cases where the unclear RoB trials actually had the methodological characteristic, this will cause the average bias to be underestimated in the standard Welton model (i.e. when the unclear RoB trials are grouped as high RoB). Therefore, when there are many unclear RoB trials, the standard

Welton *et al* model could be either over estimating or underestimating the true amount of bias between high/unclear and low RoB trials.

5.3 Meta-epidemiological considerations for continuous study characteristics: examining the association of sample size on estimated treatment effects

5.3.1 Introduction and aims

Not all study characteristics of interest are binary or categorical, some are continuous. For example, more recent trials may be at less RoB than older ones due to improvements in trial conduct and methodology. We could therefore look at the impact ‘years’ has on estimated treatment effects. It can also be hypothesised that smaller studies are at a higher RoB than the larger studies. However, current meta-epidemiological models often dichotomise continuous study characteristics, using an arbitrary cut-off to characterise trials into ′high RoB′ and ′low RoB′ groups [148, 149, 140].

Therefore, our aim, in this section, is to develop meta-epidemiological models with a continuous study characteristic. A description of the models are given in Section 5.3.2. Section 5.3.3 is a case study using sample size, in which model fit and interpretation are discussed in relation to the Welton *et al* model with a binary study characteristic. The results and key findings are given in Sections 5.3.4 and 5.3.5.

5.3.2 Description of the models

Continuous study characteristic

We first introduce the model(s), describing analogous models to the Welton *et al* model with a binary covariate, and denoting the continuous covariate as $x_{i,m}$. The Welton *et al* model with a binary covariate is described explicitly in Chapter 4, where the average bias between the high and low RoB trials and the variability in the average bias is estimated (φ^2) and additionally the variability in the high RoB trials ($\tau^2 + \kappa^2$). For brevity, we first simplify this model by not estimating the variability in bias, i.e. not estimating κ^2.

We modify the Welton *et al* model, such that $\beta_{i,m}C_{i,m}$, in the regression equation (4.2) is replaced by $b_m x_{i,m}$. This fits a regression line with slope, b_m, for each meta-analysis, estimating the effect of $x_{i,m}$ on the estimated treatment effects:

$$logit(p_{a,i,m}) = \begin{cases} \mu_{i,m} & \text{control arm} \\ \mu_{i,m} + \delta_{i,m} & \text{treatment arm} \end{cases} \tag{5.22}$$

where

$$\delta_{i,\ m} \sim N(d_m{+}b_m x_{i,m}, \tau_m^2) \tag{5.23}$$

In contrast to the Welton *et al* model with a binary study characteristic, d_m is the intervention effect when $x_{i,m} = 0$ for meta-analysis m. In practice, $x_{i,m}$ is likely to be centred. b_m is the effect of $x_{i,m}$ in meta-analysis m, given by equation (4.6). τ_m^2 is the heterogeneity in meta-analysis m. b_0 is the average association between $x_{i,m}$ and effect size across meta-analyses and φ^2 is the variation in association of $x_{i,m}$ across meta-analyses.

Continuous study characteristic and between trial heterogeneity

Comparing the model described above with the structure of the Welton *et al* model for binary characteristics, there is a difference, which is allowing for the extra variability within the high RoB trials. This is fixed between meta-analyses, denoted by κ and explicitly described in Section 4.3.1. We therefore extend the model with the continuous covariate (5.22) to allow for a potential association of the continuous variable with heterogeneity, as well as with the average effect size. To account for the variability of the covariate, $x_{i,m}$ on the treatment effect, we multiply $x_{i,m}$ by κ^2:

$$\delta_{i,\ m} \sim N(d_m{+}b_m x_{i,m}, \tau_m^2 + \kappa^2 x_{i,m}) \tag{5.24}$$

This allows the heterogeneity to also depend on sample size. We therefore provide the analogous model with a continuous study characteristic to the Welton *et al* model with a binary study characteristic when the bias in the high risk trials is allowed to vary.

The WinBUGS code for all the models described can be found in Appendix B.

5.3.3 Case study application: sample size

It can be hypothesised that smaller studies are at a higher RoB than the larger studies. Here, the Welton *et al* meta-epidemiological model of a binary covariate (high vs low RoB) is extended to look at the association of a continuous covariate, sample size, on the treatment effect. To do this, we want a coefficient from each meta-analysis on how the treatment effect (log OR) is associated with sample size, if there is a relationship at all.

In deciding which function of sample size, n, to use as the covariate in our analyses, we draw on relevant literature [87, 150, 151, 152, 153]. Smaller studies could be done less rigorously and/or be at risk of publication bias, that is, they are only published if they find a significant treatment effect. Suppose in a (single) meta-analysis, study i contributes a treatment effect β_i such as a log of the OR, and an associated SE, $se\left(\beta_i\right)$ [151]. To look for an association between study size and treatment effect estimate in an individual meta-analysis, Egger *et al* regress $\frac{\beta_i}{se(\beta_i)}$ on $\frac{1}{se(\beta_i)}$. To examine if such studies were at risk of potential publication bias, Egger *et al* explored whether there was funnel plot asymmetry by using the following linear regression:

$$E\left(\frac{\hat{\beta}_i}{se\left(\hat{\beta}_i\right)}\right) = a + \frac{b}{se\left(\hat{\beta}_i\right)} \tag{5.25}$$

The underlying hypothesis was, the smaller the trial, the less precise it will be. In this regression, a is the intercept and if it is not equal to 0, this would indicate funnel plot asymmetry. However, it came to light that there could be artificial correlation between $\hat{\beta}_i$ and $se\left(\hat{\beta}_i\right)$ when $\hat{\beta}_i$ is on the log odds ratio scale.

In an attempt to correct this, Harbord *et al* used the variance rather than the standard error for when the log odds ratio is the outcome; but did still not advocate its use [87]. Peters *et al* then suggested using $\frac{1}{n}$ to see how sample size is associated with treatment effect estimates [154]. Moreno *et al* [152] compared each of these methods and found $\frac{1}{n}$ to be more useful than Eggers test. This motivates the use of $\frac{1}{n}$ and subsequent use of $\frac{1}{\sqrt{n}}$ [153]. This also has the nice property that when you extrapolate to an infinite sample size, the amount of bias tends to 0. We additionally use $log(n)$ and n as a comparison. To aid comparisons across models, we dichotomise sample size using a cut-off value previously used when the study characteristic was treated as binary [140].

Rhodes *et al* have previously used the ROBES database to look at the differences in intervention effect between trials with sample size less than 100 participants and those with larger sample sizes [140]. We therefore dichotomise the sample size of trials as less than 100 and greater than or equal to 100 in order to compare our results when we treat sample size as continuous. We analysed data from 2058 trials included in 180 binary outcome meta-analyses.

A uniform prior over the interval (0,5) was assumed for κ.

Model fit

The deviance information criterion (DIC) [155] was used to compare each of the models to help assess model fit by (i) comparing which study characteristic fits the dataset best and (ii) whether the extra extension to include the between-study variability (i.e. the inclusion of κ^2) is needed. The deviance is defined as – 2 * log(likelihood). The posterior mean of the residual deviance can be used to assess model fit, which is expected to be roughly equal to the number of unconstrained data points for non-hierarchical models, if the model fits well. The posterior mean of the total residual deviance is usually denoted by $\bar{D}$ and the point estimate of the deviance obtained by substituting in the posterior means is denoted by $\hat{D}$. P_D is the posterior mean of the deviance minus the deviance of the posterior means and defined as 'the effective number of parameters' [155], and is given by:

$$P_D = \bar{D} - \hat{D} \tag{5.26}$$

P_D should be approximately equal to the true number of parameters for more simple, non-hierarchical models with uninformative priors. The DIC is defined as

$$DIC = \bar{D} + P_D = \hat{D} + 2P_D \tag{5.27}$$

DIC is a generalisation of Akaike's Information Criterion, more commonly known as AIC. The model with the smallest DIC is said to be the model that would be the closest to predict a dataset with the same structure as the one observed [156]. It has been reported that models with differences of more than 10 may be viewed as significant whilst any models with differences smaller than 5 could mean there is little to choose from and focus should be more on inferences of the research question. Since $\bar{D}$ incorporates a degree of penalty for complexity it is viewed as a measure of model 'adequacy' rather than model fit [157].

Interpretation of continuous study characteristics analogous to binary study characteristics

Fitting a model with a continuous study characteristic means it can be difficult to compare our results to the standard Welton *et al* model which has a binary study characteristic. In order to get something comparable, the median sample size of the dichotomised larger vs smaller sample size was compared in order to get the ROR. The median sample size of the smaller trials (those with sample sizes less than 100) was 56 and the median sample size of the larger trials (those with sample sizes greater than 100) was 207.

We use the average estimated bias, ROR (b_0), from each of the models described to predict the estimated treatment effect, on the log odds ratio scale, for a trial with a study size of 56 and 207, respectively.

$$lnOR_{n=207} = b_0 x_{n=207} \tag{5.28}$$

$$lnOR_{n=56} = b_0 x_{n=56} \tag{5.29}$$

We then calculate the ROR, that is the predicted odds ratio for study of size 207 divided by the predicted odds ratio for a study of size 56. The log scale is additive and therefore we calculate:

$$ROR_{\frac{n=207}{n=56}} = e^{lnOR_{n=207} - lnOR_{n=56}} \tag{5.30}$$

Extrapolation to a new study

We can use the model to see what size a study would need to be, under the assumption that the larger studies become less biased, for the estimated treatment effect to have no association with its sample size. We therefore want to know what size a study will need to be for the amount of bias to be negligible. For an arbitrary effect size, δ (on the log odds ratio scale), we wish to see what size x_n will need to be for the average bias, b_0, to tend to zero in our regression:

$$\text{Predicted treatment effect} = \delta + b_0 x_n \tag{5.31}$$

We approximate the value of x_n that will give us a 'predicted treatment effect'$\approx \delta$. For the functions of sample size of $x_n = \frac{1}{n}$ and $x_n = \frac{1}{\sqrt{n}}$, as $x_n \rightarrow \infty$, $b_0 \rightarrow 0$. We plot this for a range of sample size values to graphically show the impact which sample size has on the estimated treatment effect. We choose an arbitrary effect size of $\delta = 0$, i.e. an OR=1. We could have chosen any effect size and the same relationship would be shown because we are interested in the relationship between b_0 and x_n when the 'predicted treatment effect'$\approx \delta$.

5.3.4 Results

The results from each of the models, with and without κ, and comparing to the binary model when sample size is dichotomised, are given in Table 5.3. The model with the lowest DIC was when sample size was modelled as a function of $\frac{1}{n}$ with DIC=21785. This model included κ^2 which is the increase in heterogeneity for each unit increase in $\frac{1}{n}$. However, we see that the DIC is similar between all three functions of $log(n)$, $\frac{1}{n}$ and $\frac{1}{\sqrt{n}}$ with a significantly smaller DIC to the original binary model. The model with sample size as a function of n had the worst model fit with DIC=21826.

Figure 5.6 suggest there is no association between sample size and treatment effect for sample sizes greater than approximately 10,000 participants when $\frac{1}{n}$ is used. However, Figure 5.7 shows a bigger sample size is needed when sample size is modelled as a function of $\frac{1}{\sqrt{n}}$.

When comparing to the results of the median sample size in each of the dichotomised groups, we found that when taking the median sample size in each group it was also consistent. For example, when the function $\frac{1}{n}$ was used, the ROR=1.21 (95% CrI 1.13 to 1.30), whilst in the standard Welton *et al* model, the ROR=1.19 (95% CrI 1.09 to 1.30).

Table 5.3: Posterior summaries from each of the models described, examining the influence of sample size on estimated treatment effects

Function	**ROR** **(95% CrI)**	φ **(95% CrI)**	κ **(95% CrI)**	$\bar{D}$	P_D	DIC	ROR$_{\frac{n=207}{n=56}}$
Binary, >100	1.19 (1.09, 1.30)	0.24 (0.03, 0.39)		4408.0 (4235.0, 4584.0)	2802.2	21818.7	
Binary, >100 with κ	1.19 (1.09, 1.30)	0.23 (0.04, 0.39)	0.05 (0.01, 0.13)	4378.0 (4211.0, 4549.0)	2819.7	21809.6	
n	1 (1.0, 1.0)	5.57 x 10-5 (1.30 x 10-5, 1.36 x 10-4)		4411.0 (4240.0, 4589.0)	2805.9	21825.6	1.01 (1.00, 1.02)
$log(n)$	1.11 (1.07, 1.16)	0.11 (0.05, 0.17)		4415.0 (4243.0, 4591.0)	2780.2	21802.7	1.15 (1.09, 1.21)
$log(n)$ with κ	1.11 (1.07, 1.16)	0.11 (0.05, 0.17)	0.016 (0.0007, 0.048)	4414.0 (4242.0, 4591.0)	2781.7	21803.6	1.15 (1.09, 1.21)
$\frac{1}{n}$	4.83 x 10-7 (2.02 x 10-9, 1.13 x 10-4)	20.3 (12.3, 28.5)		4406.0 (4235.0, 4582.0)	2783.8	21798.5	1.21 (1.13, 1.30)
$\frac{1}{n}$ with κ	4.92 x 10-7 (2.07 x 10-9, 1.12 x 10-4)	20.2 (12.5, 28.4)	3.04 (2.22, 3.72)	4303.0 (4131.0, 4481.0)	2872.9	21785.0	1.21 (1.13, 1.30)
$\frac{1}{\sqrt{n}}$	0.041 (0.01, 0.12)	3.8 (1.8, 5.4)		4417.0 (4245.0, 4594.0)	2779	21800.8	1.23 (1.14, 1.32)
$\frac{1}{\sqrt{n}}$ with κ	0.041 (0.01, 0.12)	3.8 (1.7, 5.4)	0.80 (0.43, 1.04)	4359.0 (4184.0, 4537.0)	2832.7	21801.7	1.23 (1.14, 1.32)

$\bar{D}$ is the posterior mean of the total residual deviance, P_D is the effective number of parameters, DIC is the deviance information criterion

Figure 5.6: Plot of the association between sample size and the predicted treatment effect when $\frac{1}{n}$ is used as the function to represent the relationship which sample size has on treatment effects.

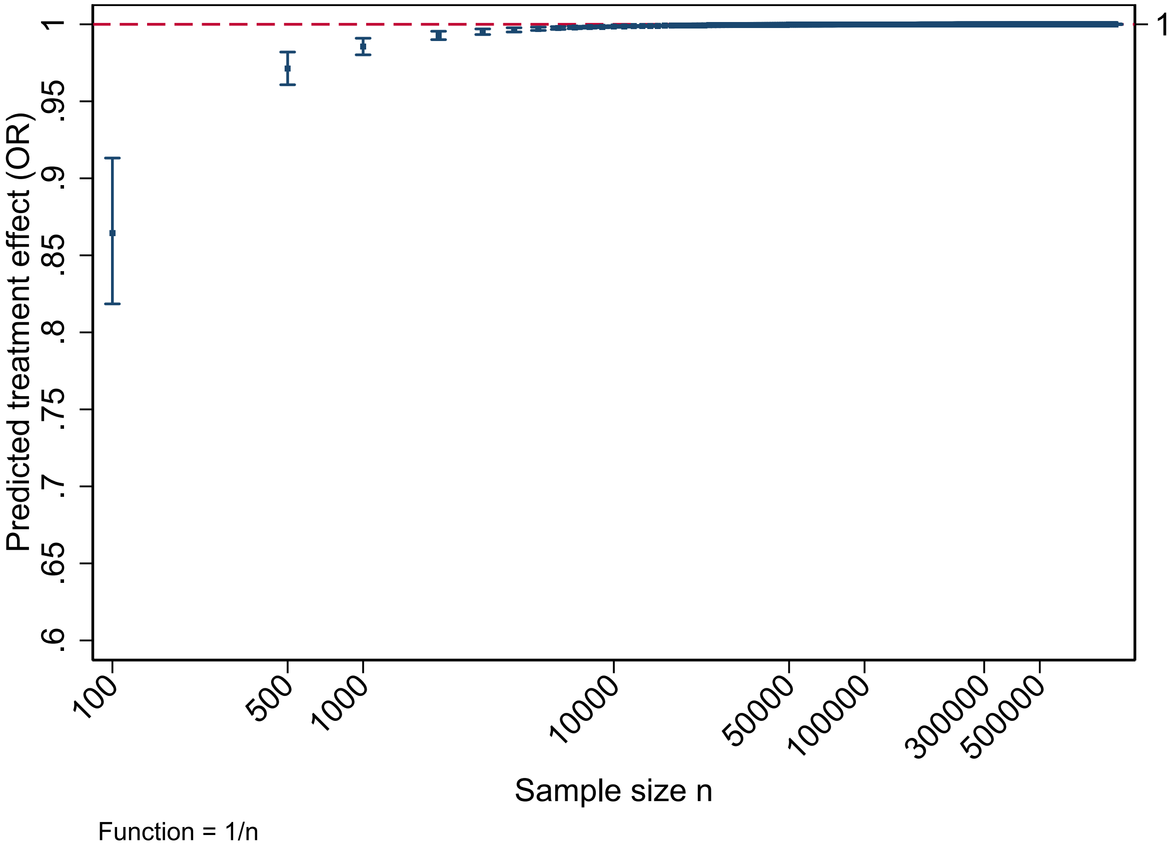
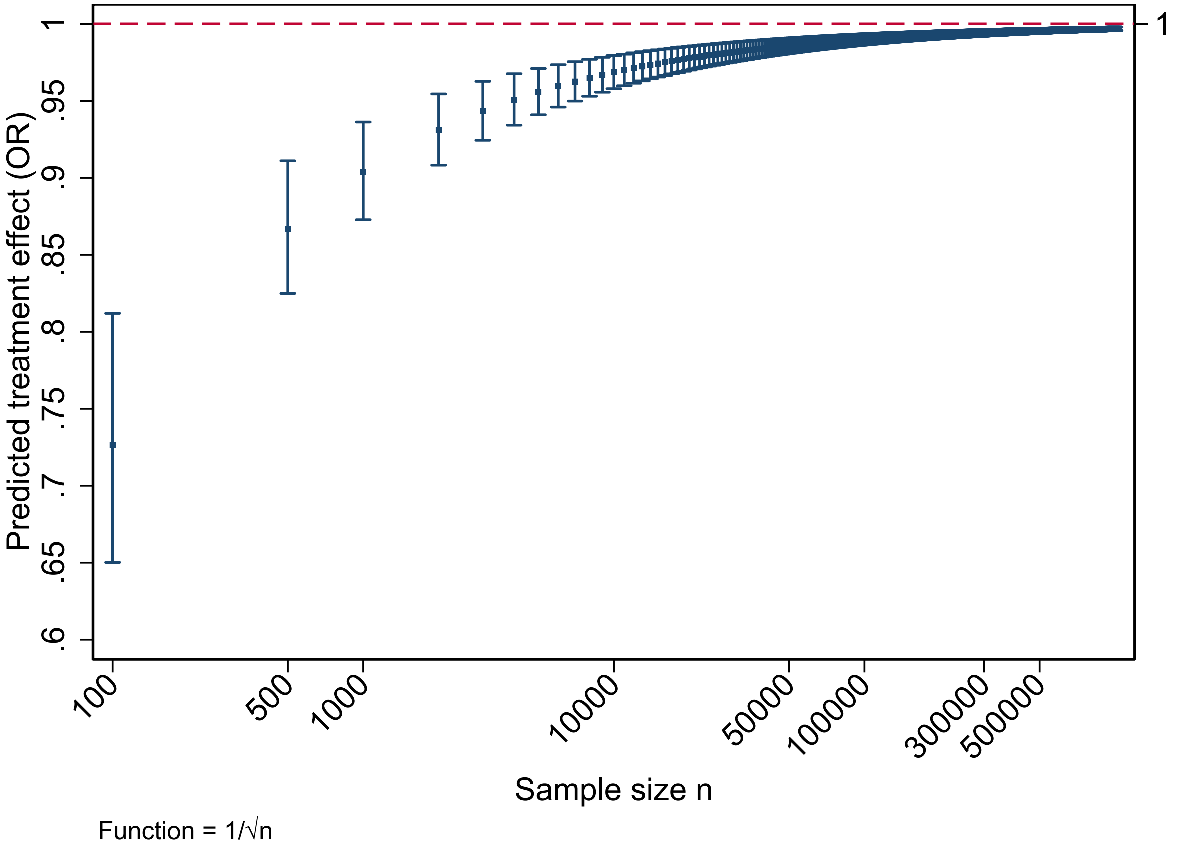


Figure 5.7: Plot of the association between sample size and the predicted treatment effect when $\frac{1}{\sqrt{n}}$ is used as the function to represent the relationship which sample size has on treatment effects.



5.3.5 Summary

Previous meta-epidemiological studies have so far only looked at the association between sample size and estimated treatment effects using either quartiles [149] or a dichotomised study characteristic [140]. We have provided the analogous model with a continuous study characteristic to the Welton *et al* binary study characteristic model which allows for the treatment effect to depend on sample size. It can also be used to predict what value a new study needs to have of this variable to provide an unbiased estimate of the treatment effect.

In each of the models we assume linearity between the outcome and covariate which means for each increase in the covariate we therefore expect the outcome to increase by the same amount. It may therefore be expected that the model which included the total sample size n as a covariate fit the data poorly.

To look at the association between treatment effect and sample size, and treating sample size as continuous, we found the best fitting model (in this case study) modelled the intervention effect as a linear function of $\frac{1}{n}$ as the study characteristic.

5.3.6 Comparison to other studies

In the Rhodes *et al* paper, the authors looked at the association of sample size on treatment effect, dichotomising sample size as a binary study characteristic, greater than or equal to 100 vs less than 100 and found a ROR=1.17 (95% CrI, 1.08 to 1.25) [140]. Using almost the same dataset, this is consistent with our final model result treating sample size as continuous.

Dechartres *et al* looked at the influence of trial size on estimated treatment effects in a meta-epidemiology study consisting of 735 trials from 93 meta-analyses across various medical conditions [158]. To do this, trials were split by quartiles and in a separate analysis split according to the number of patients in a trial, within each meta-analysis. In the analysis using quartiles: quarter 4 (containing 25% of the largest trials) was compared to quarter 1 (containing 25% of the smallest trials) using multilevel logistic regression. In quarter 4 vs quarter 1 they found a ROR=0.68 (95% CI, 0.57 to 0.82). In the analysis split using the size of the trial: trials of 1000 patients or more (large trials) were used as a reference group – to compare with trials of less than 50 patients (small size trials) and secondly trials of size 500-999 patients (moderate sized trials) . In secondary analyses they also compared trials <100 vs $\geq$ 100 and found a ROR=0.74 (95% CI, 0.65 to 0.85) which shows slightly more of an association than we found and is estimated less precisely. This is consistent to our study when we compared trials >100 vs trials $\geq$ 100 giving a ROR of 1.19 (95% CI, 1.09 to 1.30), as the analogous to this is <100 vs $\geq$ 100 giving a ROR of 0.84 (95% CI, 0.77 to 0.92).

5.4 Conclusions

In this chapter, extensions of the standard meta-epidemiology model have been described which can account for common scenarios of categorical and continuous study characteristics to allow for greater flexibility and understanding of meta-epidemiological data.

We have discussed the possible model implications for how to treat unclear RoB trials in meta-epidemiological studies so that trialists can estimate the average bias in high vs low RoB trials which can then be used for bias adjustment. We have shown that estimating the average bias attributed to high vs low and unclear vs low RoB trials in the bivariate model gives the same results as modelling the contrasts separately, in this case study data set. Therefore, in this dataset, there does not seem to be an advantage to using a bivariate model compared to modelling high vs low RoB trials separately.

We also modelled the probability that an unclear RoB trial is truly high risk which provides an estimate for the average effect of high risk trials. In this dataset we found the probability model estimates the average bias in high risk trials more precisely. More case studies are therefore needed using these models (bivariate and probability), given the extra computational time over the standard models.

To our knowledge, this is the first meta-epidemiological study to use a continuous study characteristic. Using a continuous study characteristic, rather than dichotomising at an arbitrary value, can make more use of the data. This would allow the analyst to look at the association between a continuous study characteristic and the estimated treatment effect. For example, whether there was an association between the percentage of missing data and the estimated treatment effect.

6 Comparing methods for incorporating external evidence on the effect size in sample size calculations

This chapter was a collaboration with Professor Nicky Welton and Dr Marta Soares. All work in this chapter is my own except the actual running of the EVSI analyses (in Section 6.4.3) which were run by Marta Soares.

6.1 Introduction and aims

The results of the INVEST survey in Chapter 2 highlighted that trialists were not using evidence synthesis methods to inform sample size calculations as much as they would like. Yet, funders often highlight the importance of considering existing evidence during the design of a RCT [3]. The justification of a sample size for a new trial is usually given in a grant application, a study protocol and/or in the statistical methods when the results of a trial are published; and these are almost always based on 'traditional' power calculations [159, 160]. Ultimately, what is often of interest is what size a new trial needs to be in order to update the existing evidence base and impact on policy and/or clinical practice. It may therefore be beneficial to conduct a meta-analysis of existing evidence and work out what size a new trial would need to be to demonstrate an intervention is effective, based on the totality of evidence. Several methods have been suggested for explicitly incorporating information from an existing meta-analysis into power or sample size calculations [6, 13, 73, 74], but it is unclear which of these (if any) to use in practice.

Notably, no one has compared all of these approaches in a single case study before. Therefore, the aim of this chapter is to describe and compare the following methods: (1) 'traditional' power calculations; (2) calculations based on the power of a Bayesian analysis of the new trial with an informative prior distribution based on a meta-analysis using: (i) expected and (ii) conditional power; (3) calculations based on the power of an updated meta-analysis using: (i) expected and (ii) conditional power; (4) EVSI calculations (first described in Section 1.3.3), a type of VoI calculation based on the ability of the new trial to change the decision based on a cost-effectiveness model [161]. All of the approaches are used to determine optimal sample size, but EVSI is not based on 'power'. EVSI is used to determine the optimal sample size, through means other than power and is therefore not directly comparable to the other methods. Advantages and limitations of each method, and additionally their differing assumptions and interpretations, will be discussed.

These methods are illustrated using one case study meta-analysis: an existing meta-analysis looking at the effectiveness of steroids for traumatic brain injury. The information available in 2002 showed inconclusive results. Based on this, the international, multi-centre trial of steroids for traumatic brain injury, CRASH [83, 162], was funded. We take a retrospective look and compare how information from the available meta-analysis could have informed the sample size calculation for the CRASH trial.

The remainder of the chapter is structured as follows. Section 6.2 gives some background to the role of previous evidence in power or sample size calculations, Section 6.3 summarises the case study for CRASH and Section 6.4 introduces each of the methods to be compared. These methods are contrasted in Section 6.5 and the advantages and limitations are discussed in Section 6.6, along with recommendations for use in practice and further research.

6.2 Background

In this section we first describe the power of a trial and the current role that existing evidence has in such a calculation. We then explicitly describe the null and alternative hypothesis, which is shared across approaches (1) to (3). Classical power is then described.

Finally, the concepts of conditional and expected power are introduced.

We focus on calculating sample size/power for a standard parallel clinical trial in which we sample from two groups of patients, say the experimental and control arm, assuming an equal number of patients in each. A power calculation calculates the probability of rejecting the null hypothesis for a specific sample size. Therefore, power depends crucially on sample size and we can set the same sample size, such that the power is 80%, for example. There are various possible extensions for different trial designs, for example, in cluster RCTs sample size also depends on the ICC [163, 164]. However, in this chapter we focus on a standard two arm trial.

The 'traditional' use of existing evidence in sample size calculations is to inform parameters about which we need to make assumptions, for example the likely size of the event rate in the control group or the typical standard deviation [160]. More specifically, the event rate hypothesised in the control group is likely to come from previous literature, whilst the clinician provides the minimal clinically important difference (MCID), which determines the event rate or risk in the treatment group [165]. The SD in patient outcomes, is the variability within individuals in which a value also needs to be assumed. A standard sample size calculation requires assumed values of these parameters. A meta-analysis of results of a set of similar, existing trials could provide information on these parameters. Reviews of trials funded by the NIHR HTA programme found that although 94% (32/34) of the trials examined used a referenced systematic review to justify the treatment comparison in the new trial, only 9% used it to inform the standard deviation (3/34) and none to estimate the control group event rate [57]. Jones *et al* found that 16.2% (6/37) of applications used a systematic review to inform the likely effect size [11]. Our INVEST survey found that 57% (59/103) of trialists surveyed use information from a meta-analysis to inform a sample size calculation. However, we did not collect precisely which parameters were informed by the meta-analysis.

We will use normal approximations throughout this chapter, to facilitate close form formulae and comparisons across approaches. The formulae we describe are based on using a normal likelihood to model the distribution of the data rather than a binomial likelihood for binary outcomes. Although this will often require some approximation it allows us to produce closed form formulae to calculate the required sample size or power methods. These methods can be extended to model exact count data using a binomial likelihood.

Suppose data will be collected for a new trial to estimate the intervention effect, say Y_n, where Y_n is, for example, a log odds ratio, based on a hypothetical sample size of n patients. We assume that Y_n will be normally distributed with mean equal to the true treatment effect δ and variance, $\frac{\sigma^2}{n}$ for some typical standard deviation σ:

$$Y_n \sim N\left(\delta,\ \frac{\sigma^2}{n}\right) \tag{6.1}$$

As in any standard sample size or power calculation, we re-express the hypothesised variance in the new trial as $\frac{\sigma^2}{n}$, where n is the total number of patients in the new trial. σ^2 in this equation is defined as the variability in outcomes between two randomly selected patients, it is not the between patient variability in outcomes. Our notation here is based on Spiegelhalter *et al* [13].

The null and alternative hypotheses

For a given outcome, we express the null hypothesis to be no effect ($\delta = 0$) and the two-sided alternative hypothesis [166]. We therefore define the two-sided null hypothesis $H_0 : \delta = 0$ *vs.* the alternative hypothesis $H_1 : \delta \neq 0$. When we test this hypothesis, we can make one of two decisions; either we reject the null hypothesis, or we do not reject it. However, as this is a sample from the population that we are trying to make inferences from, there is a possibility that we make the wrong decision. We can quantify the level of error we are willing to accept. We do this by setting the significance level to an arbitrary level of $\alpha = p\,(reject\ H_0|H_0\ true)$ (usually $(\alpha = 0.05)$, that is, there a maximum probability of 5% that we reject the null hypothesis is true, when in fact there is no difference).

Alternative hypotheses might be tested, however, in order to compare all approaches, we assume throughout the chapter that this is the hypothesis of interest. This hypothesis is shared across approaches (1) to (3). In this next section, we derive the 'classical' power calculation and introduce the test statistic needed to determine whether to reject the null hypothesis (or not). The test statistic compares our future data Y_n with what is expected under the null hypothesis. We then introduce the concept of 'conditional' and 'expected' power frameworks.

6.2.1 Classical power - (1)

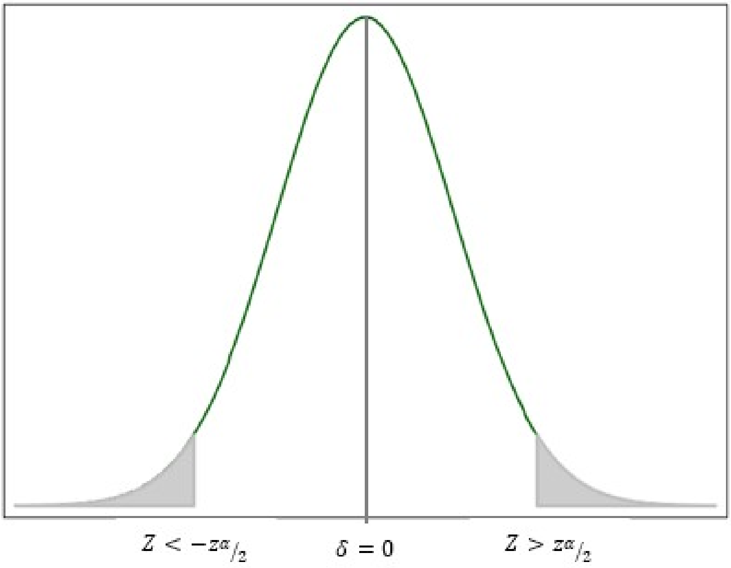
Rejecting of the null based on the CI or (equivalently) Z statistic

In a traditional analysis, the result will be declared statistically significant (H_0 rejected) if a $(1-\alpha)\%$ confidence interval (CI) around the estimate Y_n lies wholly above 0 or below 0, (i.e. does not include the null value) [29]. From equation (6.1) we define the 'standardised Z statistic':

$$Z = \frac{Y_n}{\sqrt{\frac{\sigma^2}{n}}} \sim N[0,1] \text{ under } H_0 : \delta = 0 \tag{6.2}$$

Since the Z statistic has a standard normal distribution, we can visualise this under the null hypothesis in Figure 6.1. Since the alternative hypothesis is two-sided we divide α evenly between the tails of the distribution, given by $\frac{\alpha}{2}$. If the null hypothesis is true, there will be only a 100×$\alpha\%$ chance that the Z values are in the tails of the distribution. Values far out in the tail areas therefore give strong evidence against the null hypothesis. The value from the standard normal distribution $z_{\frac{\alpha}{2}}$ such that $Pr\big(Z > z_{\frac{\alpha}{2}}\big) = \frac{\alpha}{2}$ (i.e. the

Figure 6.1: The standard normal distribution



area under the curve to the right of $z_{\frac{\alpha}{2}}$ is equal to $\frac{\alpha}{2}$) is defined as the critical value at the upper boundary or right tail. It is defined as the critical value because it separates where we reject/do not reject the null hypothesis. Similarly, for the lower boundary or left tail. From equation (6.2) we can define these as boundaries $\pm\, z_{\frac{\alpha}{2}}$, such that we reject the null hypothesis if $Z > z_{\frac{\alpha}{2}}$ or $Z < -z_{\frac{\alpha}{2}}$.

We also make use of some useful properties of the standard normal distribution function. For any tail area, $Pr\big(Z \leq -z_{\frac{\alpha}{2}}\big) = \frac{\alpha}{2}, z_{\frac{\alpha}{2}} = \Phi^{-1}\left(\frac{\alpha}{2}\right)$ where Φ is used to denote the cumulative distribution function of the normal distribution. Hence, $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ because the standard normal distribution is symmetrical around 0. Similarly, $\Phi\left(z_{\frac{\alpha}{2}}\right) = Pr(Z \leq z_{\frac{\alpha}{2}})$ is the probability that Z is less than or equal to $z_{\frac{\alpha}{2}}$ and, $\Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \Phi\left(-z_{\frac{\alpha}{2}}\right)$.

Test statistic

To get a significant result at the upper boundary,

$$Y_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} > 0$$

That is,

$$Y_n > \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}} \tag{6.3}$$

Similarly, at the lower boundary

$$Y_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} < 0 \tag{6.4}$$

That is,

$$Y_n < -\frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}} \tag{6.5}$$

Conditional power:

The probability of rejecting the null, given a hypothetical effect size

In traditional power calculations, a true value of the intervention effect, $\delta = \delta^*$ is assumed. We will refer to this as 'conditional' power, since the calculated power is conditional on this assumed value. For a particular true value of $\delta : Y_n \sim N\left(\delta^*, \; \frac{\sigma^2}{n}\right)$ where $\delta = \delta^*$, the event at the upper boundary (equation (6.3)), will occur with probability:

$$\mathbf{P}_{\text{Clinical Trial (CT)}}\left(\delta^*\right) = P\left(Y_n > \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}} | \delta = \delta^*\right) \tag{6.6}$$

$$\mathbf{P}_{CT}\left(\delta^*\right) = 1 - \Phi\left(\frac{\left[\frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right] - \delta^*}{\sqrt{\frac{\sigma^2}{n}}}\right) \tag{6.7}$$

$$\mathbf{P_{CT}}\left(\delta^*\right) = \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{\delta^*\sqrt{n}}{\sigma}\right) \tag{6.8}$$

Similarly, for the lower boundary (equation (6.5)):

$$\mathbf{P_{CT}}\left(\delta^*\right) = \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\delta^*\sqrt{n}}{\sigma}\right) \tag{6.9}$$

Therefore, the probability of rejecting the null hypothesis, given a hypothetical true effect of δ^*, ('classical power') is given by:

$$\mathbf{Power_{CT}}\left(\delta^*\right) = \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{\delta^*\sqrt{n}}{\sigma}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\delta^*\sqrt{n}}{\sigma}\right) \tag{6.10}$$

The total sample size can also be adjusted for expected dropout, n*=n/percentage expected to complete the trial [167], but we will not explore this further in this chapter.

Consideration of external evidence in sample size calculations

As described above, the 'classical' power approach is within the 'conditional' power framework as we use a pre-specified effect size. A 'conditional' power calculation calculates the probability of rejecting the null hypothesis, given a hypothetical effect size for a specific sample size. Sample size calculations for new trials are only informally guided by previous evidence in the classical power approach. However, more recently proposed methods explicitly incorporate external information from an existing meta-analysis with the assumption it will be used in the analysis stage of the trial [6, 13, 73, 74]. This previous evidence, based on a meta-analysis, can be incorporated by the use of a prior distribution in a Bayesian framework [13]. However, meta-analyses are rarely used in this way in practice [78]. In this chapter, we focus on the use of external evidence synthesis (through a meta-analysis) on the intervention effect rather than the SD or control group event rate.

As described in Section 1.3.2, meta-analyses use either a fixed effect or a random effects statistical model which are based on different assumptions. When there is substantial heterogeneity, a random effects meta-analysis should be used, however, this can have implications for the appropriate use of a meta-analysis to inform sample size calculations. We focus on the case of an inverse-variance weighted meta-analysis (Section 1.3.2) throughout this chapter.

We also introduce the concept of "expected power". This is an inherently Bayesian approach, as it involves averaging across a prior distribution. We distinguish between the situation when the target of inference is the new trial and when the target of inference is an updated meta-analysis in both the ′expected′ and ′conditional′ frameworks.

6.3 Case study: The CRASH trial

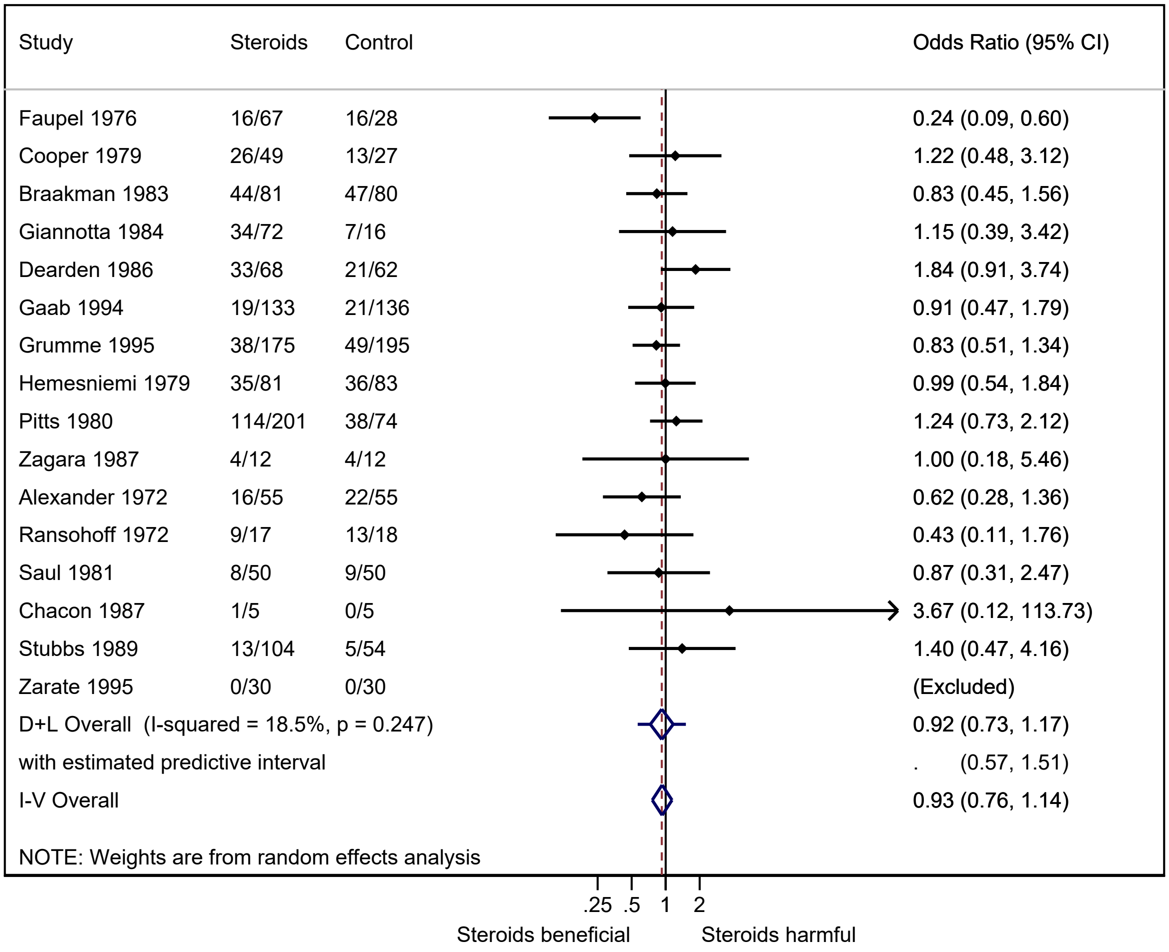
The CRASH trial was funded in 2002, to look at the effect of cortisterioids vs placebo after traumatic brain injury, following almost 25 years of inconclusive evidence and the most recent study showing a potentially beneficial effect of steroids in acute spinal cord injury [83]. We take a retrospective look at the evidence available before the CRASH trial was conducted to compare the required sample sizes according to each of the methods we describe in Section 6.4. Figure 6.2 shows the results from a random effects meta-analysis synthesising results from trials which were available before the CRASH trial was designed.

The ‘pooled’ estimate of the OR is 0.92 (95% CI 0.73 to 1.17) indicating an 8% reduction in the odds of death but with a wide, inconclusive, CI. However, there is evidence that the true intervention effect varies across studies, as shown by the between-study standard deviation $\hat{\tau}$ estimate of 0.19. The 95% prediction interval has a much wider range of intervention effects from 0.57 to 1.51 reflecting uncertainty as to where in the distribution of study effects a new study population might lie [42]. Since both of these intervals overlap an OR of 1, we cannot rule out the possibility that steroids may be harmful.

The *original* CRASH sample size calculation

> *"If the real mortality difference is 15% vs 13% then there is about a 65% chance that a trial involving 10,000 patients will achieve 2P<0.01, and a 95% chance that a trial involving 20,000 patients will do so [CRASH trial protocol [ISRCTN74459797]]."*[168]

Figure 6.2: Results from a random effects meta-analysis looking at the effect of cortisteri-oids versus placebo after traumatic brain injury. The black diamonds represent the odds ratios of the individual studies, and the horizontal lines their 95% confidence intervals. The grey squares represent the weight which each study contributes to the overall pooled result. The results of the 16 trials have been pooled in an inverse-variance (I-V) weighted random effects meta-analysis to give an overall weighted average of the treatment effect.



N.B 'P' is α in our notation. Replicating the initial sample size calculation for CRASH which assumed that steroids would reduce the mortality difference from 15% to 13%, at a power of 95% and an α of 0.01, the trialists aimed to include 21,442 patients. This power calculation is based on binary outcomes but all of the approaches in this chapter are based on normal approximations. The trialists calculated the classical power using the following equation (6.11), which is based on the same principles as (6.10) but uses binary outcomes directly.

The following equation is therefore equivalent to (6.10).

$$Power\left(\delta^{*}\right)_{prop} = \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}\right) \tag{6.11}$$

To compare the methods published [13, 73, 74], we present formulae based on the total sample size of a trial needed in order to detect a hypothetical treatment effect. In contrast to presenting the control group event rate and then the hypothesised event rate in the treatment group to detect the MCID, as is typically done in traditional sample size calculations, we present the hypothesised treatment effect between the two groups. Here, we present the odds ratio but the difference in means or risk ratio could also be presented.

We replicate the initial CRASH sample size calculation by using odds ratios and normal approximations to facilitate a comparison across methods. We convert the event rates in the control and treatment groups, from the original sample size calculation, to an odds ratio:

$$\text{Odds ratio} = \frac{\frac{p_1}{(1-p_1)}}{\frac{p_2}{(1-p_2)}} = \frac{\frac{0.13}{(1-0.13)}}{\frac{0.15}{(1-0.15)}} = 0.85 \tag{6.12}$$

We then need to hypothesise some uncertainty in this hypothetical treatment effect via its variance (σ^2) in equation (6.1). A value for σ^2 needs to be assumed. We assume the median value of $se_i\sqrt{n_i}$ across i previous trials [169]. To do this, we use the trials which were available before CRASH and are part of the meta-analysis. Using this, we obtain an estimate $\hat{\sigma} = 4.52$. Using equation (6.10) as above and assuming p-value of $0.05(\alpha)$, (which is more typical than 0.01 used in the original sample size calculation), we find that a sample size of 6000 is required to have an 80% chance of detecting an effect, the true effect is an OR of 0.85.

6.4 A comparison of methods for power calculations

In this section, a brief description of each of methods (2) - (4) is outlined.

Inference is based on the new clinical trial

We have already described the first approach when inference is based on the new clinical trial in Section 6.2.1, that is, classical power. The traditional approach to sample size calculations, as shown in (1) (Equation 6.10) only *informally* uses previous evidence [165] to provide information on the size of the effect that the trial is powered to detect and/or the typical standard deviation. It is also assumed that the results from the new trial will be analysed in isolation, using a hypothesis test. Therefore, the test statistic in classical power, given by equation (6.3) at the upper boundary, does not include any prior information.

In the next section, we introduce two ways, through the expected and conditional frameworks, in which prior information can be incorporated into a sample size calculation when inference is based on the new clinical trial. We also highlight the differences in the test statistic compared to the classical power derivation (method (1)).

6.4.1 Bayesian analysis of the new trial with an informative prior distribution - (2)

Lau *et al* first suggested expressing the results of a meta-analysis in the form of a prior distribution to inform a power calculation [6]. They argue that since a meta-analysis allows you to see if an intervention effect exists, it is possible to allow this evidence (via the prior) to be incorporated in some way when determining the sample size of a new trial. The main idea is to reduce research waste and use the results from a meta-analysis to formulate a sensible distribution of likely treatment effects in terms of a prior distribution [170], which adequately reflects the evidence base and its relationship to the new trial. There are several possible options for this prior distribution, for example it might be based on the predictive distribution or the shrinkage estimate from the meta-analysis for the most relevant previous trial [169]. Jones *et al* [169] suggest the predictive distribution is most generally applicable if you have no information about the potential causes of heterogeneity.

Following a fully Bayesian approach, inference will be based on the prior information from the meta-analysis and the estimate of the treatment effect of the new study. Thus, the prior will contribute to whether the null hypothesis is rejected or not. We use prior information for the overall treatment effect δ, based on an existing meta-analysis (assuming one does exist):

$$\delta \sim N\left(\delta_0, V\left(\delta_0\right)\right) \tag{6.13}$$

Where δ_0 is the mean and $V\left(\delta_0\right)$ is the variance of the prior distribution for δ. $V\left(\delta_0\right)$ can also be re-parameterised such that $V\left(\delta_0\right) = \frac{\sigma^2}{n_0}$ for some n_0 in which n_0 is the 'effective sample size' [13]. Here, δ_0 and $V\left(\delta_0\right)$ can take on any value, based on some output from the meta-analysis. Following a fully Bayesian approach, inference will be based on the posterior distribution of δ given the new trial data Y_n instead of the data alone:

$$\delta|Y_n \sim N\left(\frac{Y_n n + \delta_0 n_0}{n+n_0}, \frac{\sigma^2}{n+n_0}\right) \tag{6.14}$$

The posterior mean is a weighted average of the prior estimate (δ_0) and trial estimate (Y_n). We define $\hat{\delta}_1 = \frac{Y_n n+\mu_0 n_0}{n+n_0}$ and $V(\hat{\delta}_1) = \frac{\sigma^2}{n+n_0}$.

Test statistic

We may then wish to calculate the probability of obtaining a 'significant' posterior result when testing the same null and alternative hypothesis as before. We will reject H_0 if the $(1-\alpha)\%$ posterior CrI does not cross 0, i.e. upper bound of interval < 0 or lower bound of interval > 0 [13]. That is, a two-sided test at some pre-specified α rejects the null hypothesis $H_0 : \delta = 0$ if $Pr(\delta > 0|Y_n) < \frac{\alpha}{2}$ or $Pr(\delta < 0|Y_n) < \frac{\alpha}{2}$. To get a significant result at the upper boundary, $Pr(\delta > 0|Y_n) < \frac{\alpha}{2}$:

$$\hat{\delta}_1 - z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\delta}_1)} > 0 \tag{6.15}$$

Substituting $\hat{\delta}_1$ and $V(\hat{\delta}_1)$ and re-arranging for Y_n we have:

$$Y_n > \frac{z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0} - \hat{\delta}_0 n_0}{n} \tag{6.16}$$

Note that equation (6.16) reduces to equation (6.3) in Section 6.2.1 where there is no existing evidence, i.e. $\hat{\delta}_0 = 0$ and $n_0 = 0$.

Similarly, at the lower boundary, $Pr(\delta < 0|Y_n) < \frac{\alpha}{2}$

$$\hat{\delta}_1 + z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\delta}_1)} < 0 \tag{6.17}$$

$$Y_n < \frac{-z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0} - \hat{\delta}_0 n_0}{n} \tag{6.18}$$

We now describe the probability of rejecting the null hypothesis (i.e. probability of equations (6.16) or (6.18)) in the case of (i) expected power (first described by Spiegelhalter [13]) and (ii) conditional power (methodological novelty in this context).

2(i) Expected power

Spiegelhalter *et al* suggest averaging across the prior distribution to calculate the unconditional probability of a significant result, known as the 'expected power' [13], rather than setting a value for the alternative hypothesis. When inference is based on the new trial, the predictive distribution is chosen as the prior distribution to summarise the existing meta-analysis, as by definition it is the predicted intervention effect in a new study. The predictive distribution of Y_n is used to average over the prior and evaluate the chance of the event, and is given by (from equations (6.1) and (6.13)):

$$Y_n \sim N\left(\hat{\delta}_0, \sigma^2 \, \left(\frac{1}{n} + \frac{1}{n_0}\right)\right) \tag{6.19}$$

Recall at the upper boundary, given by equation number (6.16), under the predictive distribution, the probability of rejecting the null hypothesis is given by:

$$\mathbf{EP_{CT}} = P\left(Y_n > \frac{z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0} - \hat{\delta}_0 n_0}{n}\right) \tag{6.20}$$

Averaging over the predictive distribution, given by equation (6.19):

$$\mathbf{EP_{CT}} = 1 - \Phi\left(\frac{\left[\frac{z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0} - \hat{\delta}_0 n_0}{n}\right] - \hat{\delta}_0}{\sqrt{\sigma^2 \, \left(\frac{1}{n} + \frac{1}{n_0}\right)}}\right) \tag{6.21}$$

$$\mathbf{EP_{CT}} = \Phi\left(\sqrt{\frac{n_0}{n}}\left(-z_{1-\frac{\alpha}{2}} + \frac{\hat{\delta}_0\sqrt{n+n_0}}{\sigma}\right)\right) \tag{6.22}$$

Similarly, for the lower boundary:

$$\mathbf{EP_{CT}} = \Phi\left(\sqrt{\frac{n_0}{n}}\left(-z_{1-\frac{\alpha}{2}} - \frac{\hat{\delta}_0\sqrt{n+n_0}}{\sigma}\right)\right) \tag{6.23}$$

Therefore, the (two-sided) expected power of the new trial is given by:

$$\mathbf{EP_{CT}} = \Phi\left(\sqrt{\frac{n_0}{n}}\left(-z_{1-\frac{\alpha}{2}} + \frac{\hat{\delta}_0\sqrt{n+n_0}}{\sigma}\right)\right) + \Phi\left(\sqrt{\frac{n_0}{n}}\left(-z_{1-\frac{\alpha}{2}} - \frac{\hat{\delta}_0\sqrt{n+n_0}}{\sigma}\right)\right) \tag{6.24}$$

The formula described above is derived in the same way as that in Jones *et al* [169] but based on a two-sided instead of one-sided hypothesis test and based on a different null (resulting in precise formula being slightly different). O'Hagan *et al* have suggested a similar approach of averaging across a prior for sample size calculations but not using the prior in analysis [171].

2(ii) Conditional power

Rather than averaging across values of the prior distribution, we might base inference on the posterior distribution (6.14) but still calculate 'power' conditional on some prespecified effect size, $\delta = \delta^*$, more similar to a classical power calculation (Section 6.2.1). This allows us to answer the question, given the data so far, quantified by the prior distribution (predictive distribution/random effects mean distribution), what is the chance of getting a 'significant' result, if the true effect is at least as big as the MCID?

Although conditional power has not to our knowledge been derived within the context of using the results of an existing meta-analysis to predict whether the new trial will result in a posterior probability for the null hypothesis being rejected, it has been widely used in the context of interim analyses [13]. In an interim analysis trialists are looking at, given the accumulation of evidence in the trial so far (essentially the prior), what is the conditional probability that the future data collected in the trial will result in the posterior including the MCID and the null hypothesis being rejected?

For a particular true value of $\delta = \delta^*$: $Y_n \sim N\left(\delta^*,\ \frac{\sigma^2}{n}\right)$, the event at the upper boundary, given by equation (6.16), will occur with probability:

$$\mathbf{CP_{CT}}\left(\hat{\delta}_{\mathbf{0}}, \delta^*\right) = P\left(Y_n > \frac{z_{1-\frac{\alpha}{2}}\sqrt{n+n_0} - \hat{\delta}_0 n_0}{n} | \delta = \delta^*\right) \tag{6.25}$$

$$\mathbf{CP_{CT}}\left(\hat{\delta}_{\mathbf{0}}, \hat{\delta}_0^{\;*}\right) = 1 - \Phi\left(\frac{\left[\frac{z_{1-\frac{\alpha}{2}}\sqrt{n+n_0} - \delta_0 n_0}{n}\right] - \delta^*}{\sqrt{\frac{\sigma^2}{n}}}\right) \tag{6.26}$$

$$\mathbf{CP_{CT}}\left(\hat{\delta}_{\mathbf{0}}, \delta^*\right) = \Phi\left(\frac{-z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0} + \delta_0 n_0 + \delta^* n}{\sigma\sqrt{n}}\right) \tag{6.27}$$

Similarly, for the lower boundary:

$$\mathbf{CP_{CT}}\left(\hat{\delta}_{\mathbf{0}}, \delta^*\right) = \Phi\left(\frac{-z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0} - \hat{\delta}_0 n_0 - \delta^* n}{\sigma\sqrt{n}}\right) \tag{6.28}$$

Therefore, the two-sided conditional power of the new trial to detect the hypothetical treatment effect $\delta = \delta^*$ is given by:

$$\mathbf{CP_{CT}}\left(\hat{\delta}_{\mathbf{0}}, \delta^*\right) = \Phi\left(\frac{-z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0} + \hat{\delta}_0 n_0 + \delta^* n}{\sigma\sqrt{n}}\right) + \Phi\left(\frac{-z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0} - \hat{\delta}_0 n_0 - \delta^* n}{\sigma\sqrt{n}}\right) \tag{6.29}$$

When there is no prior information, i.e. when $\hat{\delta}_0 = 0$ and $n_0 = 0$, equation (6.29) is the traditional sample size calculation or classical power, given in equation (6.10).

Inference is based on the updated random effects mean

6.4.2 Updated meta-analysis - (3)

So far in this chapter we have assumed that the focus of inference is on the intervention effect in the new trial. Now suppose we are interested in whether a new trial can impact an existing meta-analysis. Inference will therefore be based on the updated meta-analysis including the new trial data, the focus of which is usually the updated random effects mean [73, 74]. This assumes that the intervention effect of the new trial will be drawn from the random effects distribution.

Suppose that the current estimate of the random effects mean, μ, (based on existing data from the current meta-analysis), equivalent to equations (1.4) and (1.6) in Section 1.3.2, is:

$$\hat{\mu}_0 = \frac{\sum \frac{y_i}{{s_i}^2+\tau^2}}{\sum \frac{1}{{s_i}^2+\tau^2}} \tag{6.30}$$

with

$$V\left(\hat{\mu}_0\right) = \frac{1}{\sum \frac{1}{{s_i}^2+\tau^2}} \tag{6.31}$$

Taking this as the basis of a prior distribution for μ, we have:

$$\mu \sim N\left(\frac{\sum \frac{y_i}{{s_i}^2+\tau^2}}{\sum \frac{1}{{s_i}^2+\tau^2}}, \frac{1}{\sum \frac{1}{{s_i}^2+\tau^2}}\right) \tag{6.32}$$

Then, if we assume that the true treatment effect in the new trial will be drawn from the random effects distribution, then the predictive distribution of Y_n, given μ and a fixed value of the between-study variance τ^2, is:

$$Y_n \sim N\left(\mu, \frac{\sigma^2}{n} + \tau^2\right) \tag{6.33}$$

Inference will be based on the updated meta-analysis, that is, the updated random effects mean μ including the new trial data Y_n, i.e. the posterior distribution of μ given new data Y_n (6.32):

$$\mu|Y_n \sim N\left(\frac{Y_n V(\hat{\mu}_0) + \hat{\mu}_0 V_p\left(Y_n\right)}{V(\hat{\mu}_0) + V_p\left(Y_n\right)}, \frac{V(\hat{\mu}_0) V_p\left(Y_n\right)}{V(\hat{\mu}_0) + V_p\left(Y_n\right)}\right) \tag{6.34}$$

Where $V_p\left(Y_n\right) = \frac{\sigma^2}{n} + \tau^2$ and written within a Bayesian framework. We denote $\hat{\mu}_1 = \frac{Y_n V(\hat{\mu}_0) + \hat{\mu}_0 V_p(Y_n)}{V(\hat{\mu}_0) + V_p(Y_n)}$ and $V(\hat{\mu}_1) = \frac{V(\hat{\mu}_0) V_p(Y_n)}{V(\hat{\mu}_0) + V_p(Y_n)}$.

Test statistic

We may then wish to calculate the probability of obtaining a classically 'significant' result when testing the two-sided null hypothesis $H_0 : \mu = 0$ vs. the alternative hypothesis $H_1 : \mu \neq 0$ for some $\alpha = Pr\left(reject\ H_0|H_0\right)$. This is equivalent to the $(1-\alpha)\%$ CrI for μ not crossing 0, i.e. upper bound of interval < 0 or lower bound of interval > 0. That is, a two-sided test at some pre-specified α rejects the null hypothesis $H_0 : \mu = 0$ if $Pr(\mu > 0|Y_n) < \frac{\alpha}{2}$ or $Pr(\mu < 0|Y_n) < \frac{\alpha}{2}$. To get a significant result at the upper boundary, $Pr(\mu > 0|Y_n) < \frac{\alpha}{2}$ then:

$$\hat{\mu}_1 - z_{1-\frac{\alpha}{2}} \sqrt{V(\hat{\mu}_1)} > 0 \tag{6.35}$$

Substituting $\hat{\mu}_1$ and $V\left(\hat{\mu}_1\right)$ and re-arranging for Y_n we have:

$$Y_n > \left(V_p\left(Y_n\right)\right)\left(z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} - \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)}\right) \tag{6.36}$$

Similarly, at the lower boundary $Pr(\mu < 0|Y_n) < \frac{\alpha}{2}$:

$$\hat{\mu}_1 + z_{1-\frac{\alpha}{2}} \sqrt{V(\hat{\mu}_1)} < 0 \tag{6.37}$$

$$Y_n \ < \left(V_p\left(Y_n\right)\right)\left(-z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} - \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)}\right) \tag{6.38}$$

Two methods in each of the 'expected' [74] and 'conditional' [73] power frameworks have been proposed to calculate the power of the new trial to impact upon the meta-analysis mean.

3(i) Expected power

The expected power approach was first described by Sutton *et al* [74] and calculates the unconditional probability of a significant result of the updated random effects meta-analysis mean with the new trial. Sutton *et al* [74] used a binomial likelihood rather than a normal approximation. This requires a simulations-based approach but it is more exact and should perform better for small counts (in particular, can handle counts of 0). Their approach also accounts for the full uncertainty in the between-study standard deviation τ. The authors assumed the control group event rate in a future study is equal to the (unweighted) average of those in the existing trials of the meta-analysis. A new study is simulated from the predictive distribution (6.32). The simulated study is then included in the meta-analysis and a rule used to establish whether the meta-analysis is conclusive, where 'conclusive' is defined as statistical significance. This is done N times. The expected or average power is estimated by calculating what proportion of the N simulations are deemed to give conclusive results. Jones *et al* [169] present a normal approximation version to produce a closed form solution, which is equivalent to the following derivation.

To make predictions concerning future values of Y_n (and therefore predict the unconditional chance of a 'significant result', in a meta-analysis), considering the uncertainty about its mean μ, we can average over the prior $\mu \sim N\left(\mu_0, V\left(\mu_0\right)\right)$ so that the predictive distribution of Y_n is given by:

$$Y_n \sim N\left(\hat{\mu}_0, V\left(\hat{\mu}_0\right) + V_p\left(Y_n\right)\right) \tag{6.39}$$

Then under this distribution, equation (6.39), the probability of rejecting the null hypothesis at the upper boundary, given by equation (6.36) is:

$$\mathbf{EP_{REMA}} = P\left(Y_n > \left(V_p\left(Y_n\right)\right)\left(z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} - \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)}\right)\right) \tag{6.40}$$

$$\mathbf{EP_{REMA}} = 1 - \Phi\left(\frac{\left[\left(V_p\left(Y_n\right)\right)\left(z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V(\hat{\mu}_0)} + \frac{1}{V_p(Y_n)}} - \frac{\hat{\mu}_0}{V(\hat{\mu}_0)}\right)\right] - \hat{\mu}_0}{\sqrt{V_p\left(Y_n\right) + V(\hat{\mu}_0)}}\right) \tag{6.41}$$

$$\mathbf{EP_{REMA}} = \Phi\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{V_p\left(Y_n\right)}{V(\hat{\mu}_0)}} + \frac{\hat{\mu}_0}{\sqrt{V(\hat{\mu}_0)}}\sqrt{1 + \frac{V_p\left(Y_n\right)}{V\left(\hat{\mu}_0\right)}}\right) \tag{6.42}$$

Similarly, at the lower boundary:

$$\mathbf{EP_{REMA}} = \Phi\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{V_p\left(Y_n\right)}{V(\hat{\mu}_0)}} - \frac{\hat{\mu}_0}{\sqrt{V(\hat{\mu}_0)}}\sqrt{1 + \frac{V_p\left(Y_n\right)}{V\left(\hat{\mu}_0\right)}}\right) \tag{6.43}$$

Therefore, the (two-sided) expected power of the new trial to impact the updated meta-analysis is given by:

$$\begin{aligned}\mathbf{EP_{REMA}} = \Phi\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{V_p\left(Y_n\right)}{V(\hat{\mu}_0)}} + \frac{\hat{\mu}_0}{\sqrt{V(\hat{\mu}_0)}}\sqrt{1 + \frac{V_p\left(Y_n\right)}{V\left(\hat{\mu}_0\right)}}\right) + \\ \Phi\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{V_p\left(Y_n\right)}{V(\hat{\mu}_0)}} - \frac{\hat{\mu}_0}{\sqrt{V(\hat{\mu}_0)}}\sqrt{1 + \frac{V_p\left(Y_n\right)}{V\left(\hat{\mu}_0\right)}}\right)\end{aligned} \tag{6.44}$$

When $\tau = 0$ equation (6.44) refers to the expected power of an updated fixed effect meta-analysis.

The formula given in the Jones *et al* paper again refers to testing the one-sided null hypothesis $H_0 : \mu > \theta$ vs. the alternative hypothesis $H_1 : \mu < \theta$ for some $\alpha = p\left(reject\ H_0|H_0\right)$ but is otherwise equivalent to 6.42.

3(ii) Conditional power

Similar to method 2(ii), Roloff *et al* [73] suggest rather than averaging across values of a prior distribution, the hypothetical effect size can be set, say $\mu = \mu^*$, to calculate the probability of rejecting the null hypothesis, assuming $\mu = \mu^*$. Under this assumption, $Y_n \sim N\left(\mu^*,\ \frac{\sigma^2}{n} + \tau^2\right)$ such that the event at the upper boundary (equation 6.36), will occur with probability:

$$\mathbf{CP}_{\mathbf{REMA}}\left(\hat{\mu}_{\mathbf{0}}, \mu^*\right) = P\left(Y_n > \left(V_p\left(Y_n\right)\right)\left(z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} - \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)} \vert\ \mu = \mu^*\right)\right) \tag{6.45}$$

$$\mathbf{CP}_{\mathbf{REMA}}\left(\hat{\mu}_{\mathbf{0}}, \mu^*\right) = 1 - \Phi\left(\frac{\left[\left(V_p\left(Y_n\right)\right)\left(z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} - \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)}\right)\right] - \mu^*}{\sqrt{V_p\left(Y_n\right)}}\right) \tag{6.46}$$

$$\mathbf{CP}_{\mathbf{REMA}}\left(\hat{\mu}_{\mathbf{0}}, \mu^*\right) = \Phi\left(-\left(\sqrt{V_p\left(Y_n\right)}\right)\left(z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} - \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)}\right) + \frac{\mu^*}{\sqrt{V_p\left(Y_n\right)}}\right) \tag{6.47}$$

Similarly, at the lower boundary:

$$\mathbf{CP}_{\mathbf{REMA}}\left(\hat{\mu}_{\mathbf{0}}, \mu^*\right) = \Phi\left(\left(\sqrt{V_p\left(Y_n\right)}\right)\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} - \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)}\right) - \frac{\mu^*}{\sqrt{V_p\left(Y_n\right)}}\right) \tag{6.48}$$

Therefore, the (two-sided) conditional power of the updated meta-analysis (with the new trial) to detect a hypothetical effect size $\mu = \mu^*$ is given by:

$$\begin{aligned}\mathbf{CP}_{\mathbf{REMA}}\left(\hat{\mu}_{\mathbf{0}}, \mu^*\right) = {} & \Phi\left(\sqrt{V_p\left(Y_n\right)}\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} - \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)}\right) - \frac{\mu^*}{\sqrt{V_p\left(Y_n\right)}}\right) + \\ & \Phi\left(\sqrt{V_p\left(Y_n\right)}\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V\left(\hat{\mu}_0\right)} + \frac{1}{V_p\left(Y_n\right)}} + \frac{\hat{\mu}_0}{V\left(\hat{\mu}_0\right)}\right) + \frac{\mu^*}{\sqrt{V_p\left(Y_n\right)}}\right)\end{aligned} \tag{6.49}$$

When $\tau = 0$ equation (6.49) refers to the conditional power of an updated fixed effect meta-analysis, having observed the fixed effect meta-analysis result.

Although inference is made on the updated meta-analysis result, Roloff *et al* do not strictly perform this within a Bayesian framework [73]. In contrast, Roloff *et al* has used information size or precision based on variance of the log odds ratio using assumptions for the event rate in the control group based on previous studies, rather than our normal approximation where we make an assumption regarding σ^2.

Assumption of heterogeneity parameter for methods (3(i) and 3(ii))

The authors [73, 74] additionally give a formula for calculating power given multiple additional studies and they note that this will increase power to impact on the random effects mean (rather than one large study) when existing between study heterogeneity is moderate to high. When this is the case, heterogeneity can be partitioned into the between study heterogeneity from the existing studies and the new estimate of the between study heterogeneity from the new study/studies. The extent of future heterogeneity (τ^2_{new}) needs to be specified. Two possibilities suggested are (i) to assume τ^2_{new} is it's the same as in existing studies, which is the approach we have used above or (ii) τ^2_{new} could be larger than existing studies if different subgroups of patients are studied.

Inference is based on the updated cost effectiveness model

6.4.3 Updated cost-effectiveness analysis, EVSI - (4)

Decisions on which treatments to recommend in clinical guidelines and guidance consider both costs and benefits of interventions, usually based on a cost-effectiveness model [52]. We may be uncertain as to which is the most cost-effective intervention and a new RCT will increase our certainty as to which is the best intervention. EVSI measures the benefit we gain from being able to reduce the uncertainty in a decision by adding to the evidence base through running a new RCT of a given sample size. The EVSI can be compared with the cost of running an RCT with such a sample size to find whether such a study would be an efficient use of research resources, and also to identify the optimal sample size. To calculate the sample size based on the results of an updated cost effectiveness analysis (assuming a decision model is already in place):

1. A prior distribution is specified for the treatment effect (6.13). Prior information for the treatment effect can be given by the predictive distribution of the random effects meta-analysis (although any prior can be used, predictive may be more applicable), assuming a fixed value of the between-study variance, τ^2, given by equation (6.13).

2. A hypothetical sample size is given to a new study (size n) which will provide new data, D. Following a fully Bayesian approach, inference will be based on the posterior distribution of δ, given the new trial data, Y_n, instead of the data alone, given by equation (6.14).

3. We then use this posterior distribution to update the cost-effectiveness model.

4. If the optimal decision changes, there is a gain in net benefit (NB), which is the difference in incremental benefit and cost from using the new optimal treatment. If the optimal decision is unchanged there is no gain in NB.

5. This is done multiple times over future possible datasets D to calculate the average gain and obtain the EVSI, defined as the expected NB from a particular study design minus the expected NB of current information. This requires posterior updating within simulation over new data. It is only possible to get a closed form formula for EVSI for very simple situations because it depends on the NB function. We then find the expected NB averaging over the posterior distribution of all the parameters conditional on new data. If the NB is linear in all parameters, then we can use conjugacy arguments to find approximate normal posteriors and plug their expectations into the formula [52]. However, the function is often non linear and depends

on products of parameters that are correlated. In this case it is not possible to plug in the means.

6. This process (steps 2-5) is repeated for multiple hypothesised sample sizes.

7. Then for each sample size the difference between the EVSI and the costs of sampling (i.e. monetary values) for specific designs gives the expected net benefit of sampling (ENBS), allowing the optimal sample size to be calculated [52]. Therefore, when the benefits of collecting more information are bigger than the costs of doing so, ENBS will be positive.

8. The optimal sample size design can then be determined from the ENBS which measures the difference between the cost and benefit of designs with different sample sizes [172]. The benefits are always in terms of the NB function, however that is defined. In health economics net benefit is usually defined in terms of monetary units, and health benefits are converted to monetary units via quality adjusted life years (QALYs) for a given willingness to pay per QALY.

Again, for this case study analysis Marta Soares provided the EVSI results but not the ENBS results. This would have to be contrasted with the cost of sampling to determine the optimal sample size. However, EVSI can still show where further gains are minimal.

Table 6.1 summarises each of the methods described.

Table 6.1: Summary of the closed form equations for methods (1) – (3). Method (4) does not give a closed form solution because the posterior distribution is updated using MCMC sampling and is therefore not included in this table.

	Method	Target of inference	Formula
1	Traditional sample size to detect a hypothetical effect size $\delta = \delta^*$	Trial	$\text{Power}_{\text{CT}}\left(\delta^*\right) = \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{\delta^*\sqrt{n}}{\sigma}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{\delta^*\sqrt{n}}{\sigma}\right)$
2(i)	Expected power of a new trial using an informative prior	Trial	$\text{EP}_{\text{CT}} = \Phi\left(\sqrt{\frac{n_0}{n}}\left(-z_{1-\frac{\alpha}{2}} + \frac{\hat{\delta}_0\sqrt{n+n_0}}{\sigma}\right)\right) + \Phi\left(\sqrt{\frac{n_0}{n}}\left(-z_{1-\frac{\alpha}{2}} - \frac{\hat{\delta}_0\sqrt{n+n_0}}{\sigma}\right)\right)$
2(ii)	Conditional power of a new trial to detect a hypothetical effect size $\delta = \delta^*$ using an informative prior	Trial	$\text{CP}_{\text{CT}}\left(\hat{\delta}_0, \delta^*\right) = \Phi\left(\frac{-z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0}+\hat{\delta}_0 n_0+\delta^* n}{\sigma\sqrt{n}}\right) + \Phi\left(\frac{-z_{1-\frac{\alpha}{2}}\sigma\sqrt{n+n_0}-\hat{\delta}_0 n_0-\delta^* n}{\sigma\sqrt{n}}\right)$
3(i)	Expected power of an updated meta-analysis using an informative prior	Updated meta-analysis mean	$\text{EP}_{\text{REMA}} = \Phi\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{V_p(Y_n)}{V(\hat{\mu}_0)}} + \frac{\hat{\mu}_0}{\sqrt{V(\hat{\mu}_0)}}\sqrt{1+\frac{V_p(Y_n)}{V(\hat{\mu}_0)}}\right) +$ $\Phi\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{V_p(Y_n)}{V(\hat{\mu}_0)}} - \frac{\hat{\mu}_0}{\sqrt{V(\hat{\mu}_0)}}\sqrt{1+\frac{V_p(Y_n)}{V(\hat{\mu}_0)}}\right)$
3(ii)	Conditional power of an updated meta-analysis to detect a hypothetical effect size $\mu = \mu^*$ using an informative prior	Updated meta-analysis mean	$\text{CP}_{\text{REMA}}\left(\hat{\mu}_0, \mu^*\right) = \Phi\left(\sqrt{V_p\left(Y_n\right)}\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V(\hat{\mu}_0)}+\frac{1}{V_p(Y_n)}} - \frac{\hat{\mu}_0}{V(\hat{\mu}_0)}\right) - \frac{\mu^*}{\sqrt{V_p(Y_n)}}\right) +$ $\Phi\left(\sqrt{V_p\left(Y_n\right)}\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{V(\hat{\mu}_0)}+\frac{1}{V_p(Y_n)}} + \frac{\hat{\mu}_0}{V(\hat{\mu}_0)}\right) + \frac{\mu^*}{\sqrt{V_p(Y_n)}}\right)$

$V_p\left(Y_n\right) = \frac{\sigma^2}{n} + \tau^2.$

6.5 Results

For each of the methods in Table 6.1, we give the results of the two-sided hypothesis test in Section 6.4 including the ability for a new trial to update a *fixed* and *random* effects meta-analysis.

We set the two-sided alternative hypothesis that there is a non-zero effect and the MCID to be OR=0.85 (i.e. δ^* or $\mu^* = log(0.85)$) as in the original CRASH sample size calculation and use a significance level of $\alpha = 0.05$, that is, we calculate the power of the new trial to detect a 15% reduction in odds of death in the treatment group compared to the control group. When applying the methods to CRASH we need to choose our prior distribution. We choose the predictive distribution from the meta-analysis as the prior throughout. We therefore take $\hat{\mu}_0 = ln(0.92) = -0.0833$, and assuming $\sigma = 4.52$ and $\tau = 0.19$ this gives $n_0 = 331$. Figure 6.3 displays the relationship between the new sample size, n, and the power for each of the four methods. Figure 6.4 displays the relationship between the new sample size, n and the EVSI.

Random effects meta-analysis

Figure 6.3 shows that the traditional power calculation (1) suggested 6000 patients were required for a power of 80%.

6.5.1 Bayesian analysis of the new trial with an informative prior distribution - (2)

For the same sample size, the 'expected power' of a Bayesian analysis (2(i)) with an informative prior based on the predictive distribution of the existing meta-analysis was 74%. However, the expected power of a Bayesian analysis (method 2(i)) is greater than the classical power (1) of smaller sample sizes. For example, for a total sample size of 1500 patients the traditional power calculation gives only 29% compared to an ‘expected’ power of 50%.

Assuming a true effect of $\delta = \delta^* = log(0.85)$, the conditional power of a Bayesian analysis of the new trial incorporating the prior (2(ii)), is visibly lower for a sample size of 1000 patients than the traditional power calculation (1). This seems slightly unintuitive as we are incorporating prior information equivalent to data on an additional 331 patients. However, note that the effect we wish to detect of (OR=)0.85 is more extreme than the prior mean (OR=0.92 (95% CI, 0.57 to 1.51)). As the sample size increases, the data from the future trial dominates the posterior and any influence of the prior becomes negligible.

6.5.2 Updated meta-analysis - (3)

The expected power (3(i)) and the conditional power of the random-effects meta-analysis to detect an effect at least as big as (OR=) 0.85 having observed a difference of OR = 0.92 [0.57; 1.51] (3(ii)) is <10% across all possible sample sizes.

6.5.3 Updated cost-effectiveness analysis, EVSI - (4)

In contrast, the EVSI (4) calculation in Figure 6.4 shows that, to reduce the uncertainty in the decision model, regarding whether or not cortisterioids should be used in practice, a 2-arm randomised trial with approximately 2000 patients would provide a net health gain of almost 15000 QALYs. We see that there is only minimal gains in QALYs for a sample size greater than 2000 patients (and costs would increase). The uncertainty in the treatment effect as seen in the wide prediction interval (Figure 6.2) may be driving the impact on the decision model as EVSI measures the value of collecting evidence from a given study design to reduce decision uncertainty [162].

Figure 6.3: Power curves for the two-sided test $H_0 : \delta = 0$ vs. the alternative hypothesis $H_1 : \delta \neq 0$. Methods 2(i)) and 3(i) require a prior (to average across) and uses the predictive distribution. Method 3(ii) doesn't require a prior and instead updates the meta-analysis. Method 2(ii) uses the predictive distribution as a prior. We assume $\alpha = 0.05$ and a 'typical' standard deviation σ of 4.52 (estimated as the median across studies). Additionally, for methods (1), (2(ii)) and (3(ii)) a hypothetical effect is set to detect an OR=0.85.

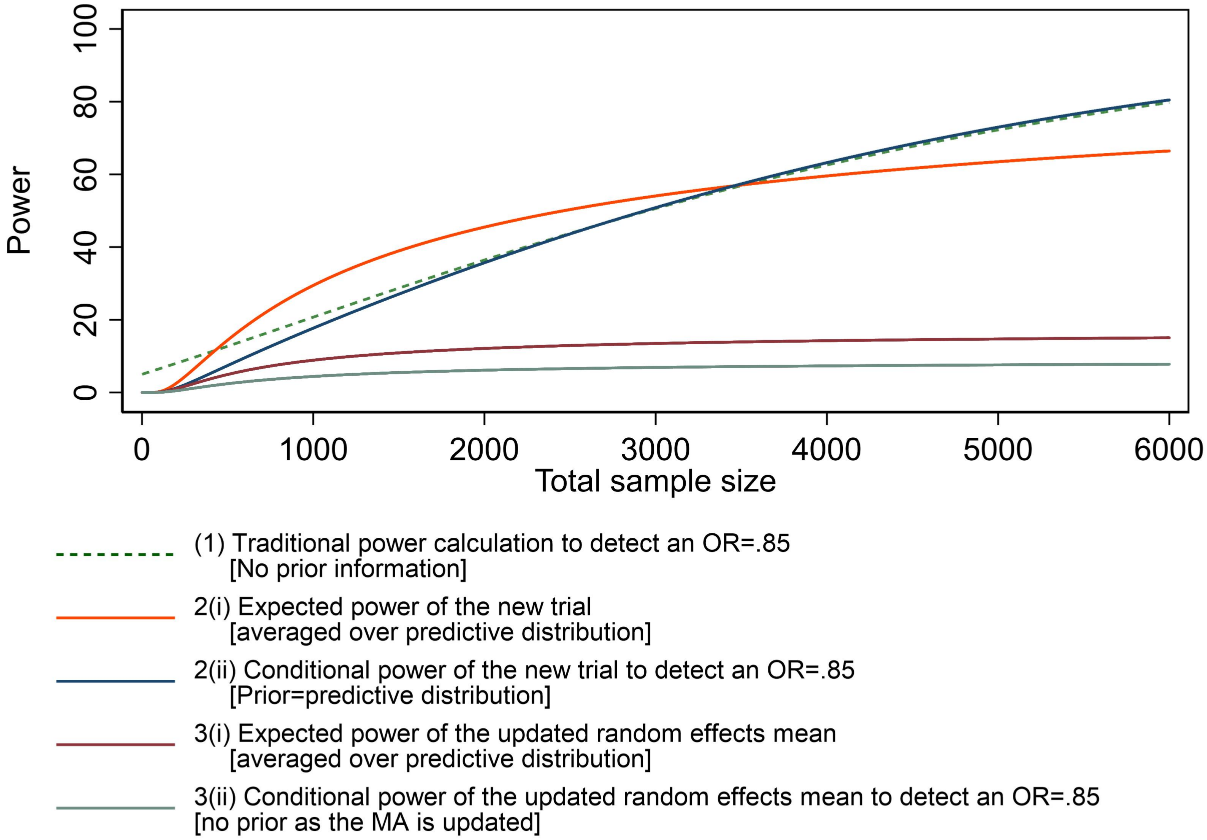
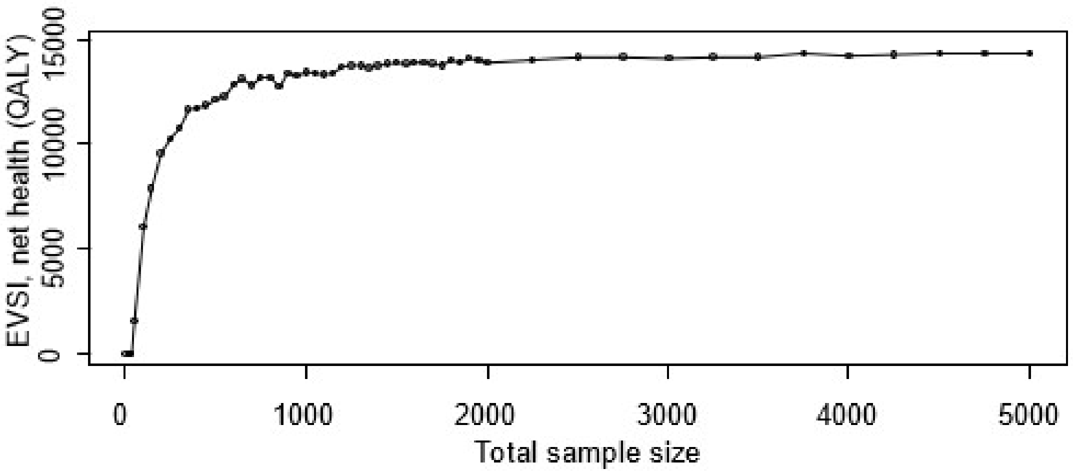


Figure 6.4: EVSI values according to total sample size of a new trial informing probability of the event and treatment effect in both groups (in terms of quality adjusted life years (QALYs) , ENBS is not worked out here and is therefore not given in terms of cost

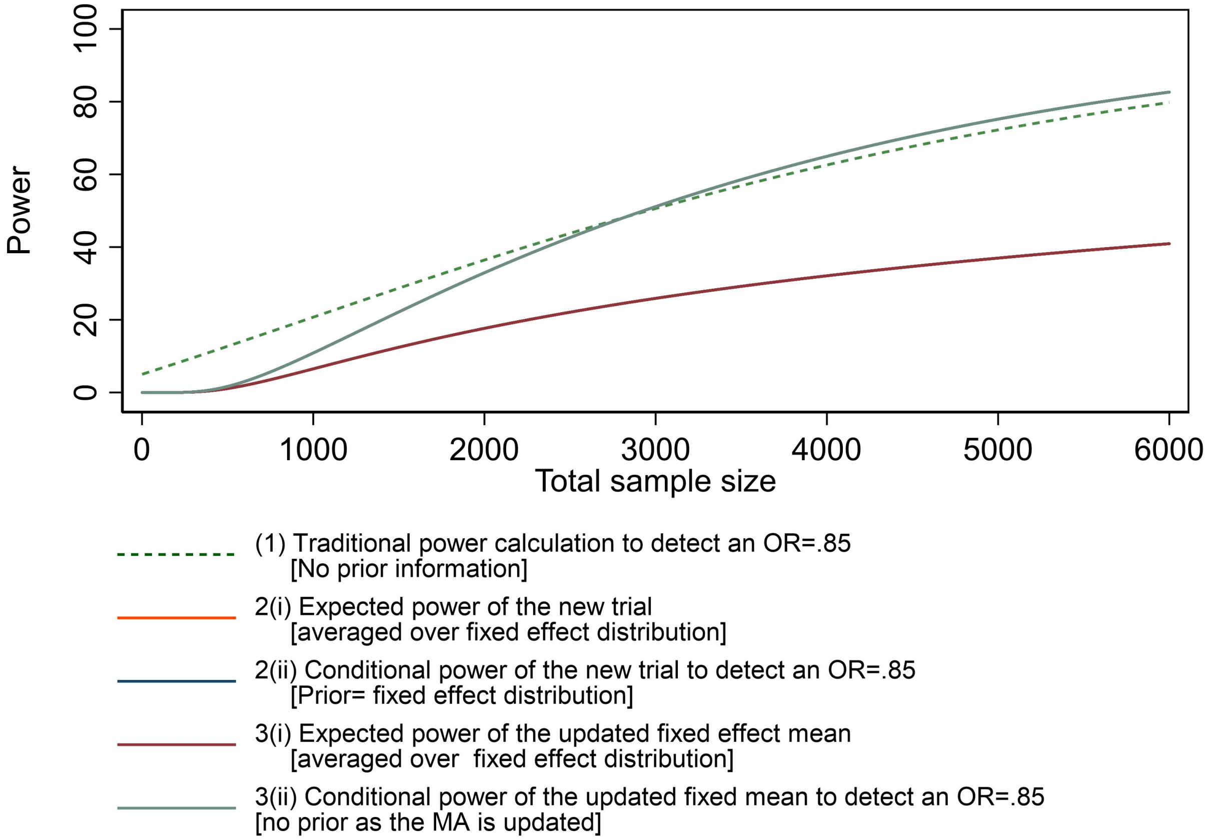


Fixed effect meta-analysis

We now choose to investigate the ability a new trial has to impact a fixed effect meta-analysis result, i.e. when equations 3(i) and 3(ii) have $\tau = 0$. When $\tau = 0$, methods 3(i) and 3(ii) reduce to 2(i) and 2(ii) respectively. The fixed effect meta-analysis distribution now becomes the prior distribution: OR=0.93 (95% CI, 0.76 to 1.14). We therefore take $\hat{\mu}_0 = ln(0.93) = -0.0833$, and assuming $\sigma = 4.52$ (as earlier) and $\tau = 0$. This gives $n_0 = 1446$.

As in the random effects case, we find that the conditional power of the Bayesian analysis (incorporating prior information equivalent to data on 1446 patients) is *less* than the classical power (ignoring the previous data) across all sample sizes. Again, this is because the prior mean is much closer to the null than the effect we are trying to detect. In contrast, to the random effects case, power for methods 3(i) and 3(ii) is higher (i.e. with no heterogeneity).

Figure 6.5: Power curves for the two-sided test $H_0 : \delta = 0$ vs. the alternative hypothesis $H_1 : \delta \neq 0$. Methods 2(i)) and 3(i) require a prior (to average across) and uses the predictive distribution. Method 3(ii) doesn't require a prior and instead updates the fixed effect meta-analysis. Method 2(ii) uses the predictive distribution as a prior. We assume $\alpha = 0.05$ and a 'typical' standard deviation σ of 4.52 (estimated as the median across studies). Additionally, for methods (1), 2(ii) and 3(ii) a hypothetical effect is set to detect an OR=0.85. When $\tau = 0$, methods 3(i) and 3(ii) reduce to 2(i) and 2(ii) respectively and therefore curves 3(i)= 2(i) and similarly 3(ii)=2(ii).



6.6 Discussion

In this chapter, we have compared methods for using external evidence on the effect size in sample size calculations, using an application to a real case study of CRASH [83]. A meta-analysis of results from existing RCTs can help determine whether there is a need for a new RCT and inform the sample size calculations. Although many studies have identified funders require a systematic review and/or meta-analysis [3], little analytical attention has been paid to how this previous evidence should be used to inform sample size calculations. We have addressed this issue by comparing recently proposed methods that use information from an existing meta-analysis in order to show what size a new trial would need to be, to show the intervention is effective, depending on whether the target of inference is the new trial or the updated meta-analysis mean. This is in contrast to traditional sample size calculations, where previous evidence is used to make assumptions about key parameters, but under the assumption that the trial will be analysed in isolation (not in combination with previous evidence). The key findings are displayed in Figure 6.6 and discussed in more detail below.

6.6.1 Overview of key findings

We found that in the case study of steroids for traumatic brain injury, a standard sample size calculation suggested approximately 6000 patients were required to detect a 15% reduction in odds of mortality for a power of 80%.

The predictive estimate from the existing meta-analysis is 0.92 (95% predictive interval, 0.57 to 1.51) indicating an 8% reduction in the odds of death in a new population. Since we are trying to detect a 15% reduction in the odds of death (OR=0.85) this will have some effect on the sample size needed. For this example, we found the classical power is greater than the power of a Bayesian analysis of the new trial incorporating the prior. This is likely to be because the prior mean is closer to the null, than the effect we are trying to detect, so that the prior pulls the posterior distribution closer to the null. For the same sample size, the 'expected power' of a Bayesian analysis with informative prior based on the predictive distribution from the meta-analysis was only 74%.

The power of an updated meta-analysis was <10% across all sample sizes: a new trial had practically no ability to impact upon the estimated mean in a random-effects meta-analysis, due to high levels of heterogeneity. This was the same in both the 'expected' and 'conditional' power frameworks, although they cannot be strictly compared as expected power is estimating the average power by averaging over the predictive distribution whilst conditional power conditions on a hypothetical effect size to detect. This has been noted previously [38, 74, 73]: that if there is high heterogeneity in a meta-analysis, no new trial is likely to change the random effects mean. In contrast, the EVSI calculations suggested that the required sample size was lower.

We have seen that each of these approaches can lead to very different conclusions for trialists. Trials are conducted for a variety of reasons: such as (i) to obtain licensing for a drug; (ii) to identify doses/settings/different populations where the treatment is effective; (iii) to get the treatment adopted by NICE or similar organisations; (iv) to confirm the results of previous/smaller trials [173]. When deciding whether to run a new trial, and the appropriate trial design, it is important to consider how the results of that trial will be used. In order for the results to change clinical practice, there needs to be a change in clinical guidelines and policy, and these need to be adopted by clinical practitioners. This can occur through 3 main routes. First, clinicians could be convinced by a journal article publishing the trial results. Second, the trial could have influence as a contributing study in a meta-analysis. Third, the influence could be through the trial's contribution to a cost-effectiveness analysis model. Meta-analyses are carried out in the development of all clinical guidelines developed by NICE in the UK, and similar organisations worldwide [174]. Cost-effectiveness models typically use the results of a meta-analysis to inform the effectiveness parameter, but also combine this information with other information on longer-term outcomes, costs, utilities for various outcomes (including adverse events), and natural history parameters describing disease progression. Cost-effectiveness models are used to inform reimbursement decisions by NICE in the UK, and similar organisations worldwide.

Figure 6.6: Key findings from a case study analysis

- The traditional power calculation (method (1)) suggested 6000 patients were required for a power of 80%.
- Sample size calculations based on the conditional power of a Bayesian analysis of the new trial, using an informative prior (method (2(ii)) gave lower power than method (1). This is likely due to the prior mean in this case study being closer to the null than the effect size we were trying to detect.
- A new trial had practically no ability to impact upon the estimated mean in a random-effects meta-analysis. This was likely due to high levels of heterogeneity in the meta-analysis.
- In contrast, the EVSI calculations, which are measured in QALYs and not power, suggested that the required sample size was lower.
- As the different approaches lead to *very* different conclusions, it is important for a trialist to decide their perspective: whether the target of inference is the (i) trial (ii) meta-analysis or (iii) cost-effectiveness analysis.

6.6.2 Are these methods implementable in routine trial design?

Methods (2) to (4) all require a systematic review and meta-analysis as an input for forming a prior distribution. However, funders often require a systematic review or meta-analysis to have been done to justify the need for the trial [9]. In these situations, it could be viewed as not much extra effort to formulate a prior distribution and, as part of a sensitivity analysis, to see the impact prior evidence can have compared with a traditional power calculation. Expected power estimates the average power whereas the conditional power framework conditions on the hypothetical effect to detect. Therefore, expected power averages over all possible values of the prior distribution rather than detecting a MCID. In an interim analysis trialists are looking at whether given the accumulation of evidence in the trial so far, how likely future data collected will result in finding the MCID (and the null hypothesis being rejected). Although conditional power isn't routinely used in the context of using information from an existing meta-analysis, it can be seen as a natural extension of interim analyses [13] with design considerations the same as when conducting a conventional power calculation and may be more applicable than the expected power approach. The conditional power approach has also been extended to a NMA setting [175].

We think this could have particular application for trials which are expected to have a small number of events, i.e. for rare event data. If there has been a meta-analysis looking at the comparison of interest it may be intuitive to incorporate this information in a Bayesian analysis of a new trial, when the target of inference is the new trial, rather than the updated meta-analysis. Whether there is a gain in power or reduction in sample size will be determined by the relationship between the MCID and prior information from the meta-analysis.

6.6.3 Towards recommendations

If the plan is for the results of the trial to be placed in the context of existing evidence at the end of the trial, then it seems sensible to conduct the meta-analysis in the planning of the trial and include it in the power calculation. Caution should however be exercised here as the appropriate use of these methods is heavily dependent on the purpose of the trial. For example, different stakeholders will always use results differently and the new trial will need to be powered according to how the results will be used. From the perspective of a trialist there is often no consideration of the meta-analysis at design stage of component studies, even though the updated meta-analysis may be more influential than any of the studies in isolation, in order to change policy and practice [176]. This was also consistent with the findings in our qualitative study: trialists recognised it was often the body of evidence that changes practice but did not think about whether their trial results would end up in a future meta-analysis. It is therefore possible that from an individual trial perspective that meta-analyses are considered by-product of research. However, it may be more coherent from a trialists perspective to design a trial based on results of the updated meta-analysis [5, 74].

Currently, all these methods are rarely used in practice [10, 11]. A possible recommendation is to make conditional or expected power calculations a standard output of Cochrane reviews, NIHR HTA monographs and/or in grant applications to HTA. This would allow trialists to assess the impact a new trial would have on the existing meta-analysis. Similarly, VoI analyses could be a required output from NIHR HTA monographs that have cost-effectiveness models, which are becoming more increasingly included. NICE technology assessments also often perform VoI analyses and make recommendations for priorities for further research, but this doesn't link directly to what is prioritised by funding bodies such as the NIHR HTA.

6.6.4 Areas for further research and consideration

Methodology for EVSI is relatively new and still developing. Throughout applying these methods, we have focussed on random effects rather than a fixed effect meta-analysis being used. One of the main reasons for this, is that a fixed effect meta-analysis makes a rather strong assumption, that are trials are estimating the same single treatment effect. However, random effects models have their own interpretability issues; one of the major challenges is how to interpret and incorporate heterogeneity between existing studies. Assumptions of exchangeability between previous studies and new study need consideration. Additionally, under the power of updated meta-analyses methods, conclusions can be that lots of small studies are needed instead of one large "definitive" study in order to change statistical significance [38, 74, 73]. This can again be unintuitive and an area for further research.

Ultimately, a trialist is likely to only focus on the trial in front of them and not think in the wider context of how the results will be used in a meta-analysis or cost-effectiveness analysis. In some situations, a trialist may conduct such an analysis which concludes that their study is not worth doing. As such, educating trialists and panel members on these methods might be useful so that a priority is put forward to do these analyses as part of a feasibility phase; enabling efforts to be focussed on more impactful research. The main area for future work is on training with a focus on trialists performing different sample size calculations as part of a sensitivity analysis and to assess how the incorporation of previous evidence could impact upon the required sample size.

Multi-arm multi-stage (MAMS) trials allow several new interventions to be tested against a single control within one trial and at predefined stages, each new intervention is analysed, much like an interim analysis, and dropped if the intervention does not show enough of an effect. New interventions can also be added in order to answer research questions more quickly and efficiently. Adaptive trial designs such as MAMS are being increasingly used with results from earlier phases in the trial to determine the design in later phases, including early-stopping decisions. As such this is a natural place for evidence synthesis to be used and a potential area for future research [177].

We have focussed on sample size determination, but other design features which address a much wider range of design questions can be considered using the EVSI approach, such as which patient subgroups to include, which interventions to include (and how many treatment arms), and which outcomes to measure [178]. The other methods only focus on uncertainty in effectiveness, which won't necessarily translate into decision uncertainty (for example, there may be high degree of uncertainty between two interventions, but neither likely to be cost-effective. Or it may be very clear which intervention is most effective, but on a surrogate outcome, so that a trial is warranted to establish longer-term effects) [53]. However, EVSI requires a cost-effectiveness model to be developed, which takes time and resources to build [94], and computations can be intensive and require specialist expertise.

RCTs require equipoise from both the clinician recruiting and the patient being randomised. However, if there is an existing body of evidence on the comparative effectiveness and/or cost-effectiveness of the interventions being trialled, then this may raise ethical questions for the randomisation to that trial [177]. As is briefly mentioned in Section 3.4.5, it is usually the case that there is some evidence available, and that neither the clinician or patient is in 50:50 ambivalence between two treatments. However, there may be a lot of uncertainty around this preference, and a meta-analysis and/or VOI analysis can quantify this uncertainty, and therefore has an important role to play in the interpretation of equipoise. This is an area for further consideration and development.

6.6.5 Conclusions

It seems logical to use previous evidence in the form of a prior distribution explicitly in sample size calculations rather than analysing a trial in isolation. We have shown there are several methods to do this, however the choice of method should depend on the primary purpose of the trial. When designing and powering a trial, consideration should be given to how the study results may be used and interpreted: (i) in isolation, (ii) in terms of its impact upon an updated meta-analysis, or (iii) in terms of its impact upon a cost effectiveness analysis.

In order to increase the use of previous evidence in sample size calculations, we recommend funding bodies advise trialists to report how the results of their trial will be interpreted (in relation to scenarios (i), (ii) and/or (iii)) during a feasibility stage. This will require increased collaborations between funders, trialists and the NICE technology assessments. As these methods are still relatively new, assumptions of such statistical models should be transparent and further training undertaken. In addition, the input to these analyses require at the very least a systematic review and detailed guidelines for a framework on how existing evidence should be used in sample size calculations is encouraged to streamline its use in practice.

7 Using external evidence on adverse outcomes with an application in first in human studies

The work presented in this chapter was performed in collaboration with Laurence Colin, Baldur Magnusson, Yue Li and Asher Schachter at Novartis AG, Switzerland.

7.1 Context and overview

In a standard two arm RCT, the control arm acts as a comparator to the experimental arm. The control arm can be the intervention which is already current practice, or it can be a placebo. In Chapter 3 we saw that an area where trialists thought they could potentially benefit from the incorporation of external evidence synthesis was when trials were not adequately powered to compare adverse events between the two groups. In these situations, current practice relies heavily on the clinician or chief investigator working on the trial to decide whether the observed adverse event rate in the experimental arm is inline with what would be expected or if it is a potential cause for concern, resulting from the intervention.

In this chapter, we take a case study of using external evidence on particular outcomes in "first in human" “(FIH)” studies [179]. These trials have very small sample sizes and are therefore unlikely to be powered to detect adverse events. In FIH studies, an extremely small placebo arm (e.g. 2 patients) acts as the comparator. We use a synthesis of placebo data from a set of previous FIH studies to inform the expected adverse event rate in a new or current FIH study. As this project was a collaboration with Novartis, the placebo data synthesised were from Novartis sponsored healthy volunteer studies.

7.2 Introduction and aims

FIH studies offer the first opportunity to test a new treatment in healthy human volunteers (or patients in an oncology setting), bridging the gap between animal and human studies. They are part of the exploratory phase of drug development to investigate safety and tolerability. The primary objective is to identify a suitable dose or dose range. Once a safe dose is found, efficacy is assessed in phase I and/or II studies. They are usually placebo controlled with subjects randomised to either the active drug or a placebo [180]. Typically, a FIH study is split into groups of subjects each assigned to particular doses. Within each group, known as cohorts, there are usually 8 subjects of which 6 will be given the active drug and 2 a placebo. Due to the typically small sample sizes of cohorts, safety signals are often difficult to interpret, particularly in the absence of a robust placebo group. Therefore, when analysing a FIH study, those working on such studies often need to decide if an adverse, or safety event on a new drug is likely due to the drug or if the event is likely to occur by chance. It is difficult to judge whether certain adverse events or abnormal laboratory values are occurring because they are caused by the new drug or simply because these events do occur on placebo frequently in the general population [181].

A report published in 2014 revealed that about half of all FDA rejections and delayed approvals in recent years were due, at least in part, to safety deficiencies [182], with cardiovascular and hepatic issues being the most common concerns. One of the most common reasons a drug does not make it to the next phase, phase I of drug development, is if it is thought to be too toxic and likely to cause drug induced liver injury (DILI). Biomarkers, such as alanine aminotransferase (ALT), measured from a blood test, are used as a potential indication of DILI if they are outside the normal range. Kobayashi *et al* found elevated ALT levels in volunteers after administration of placebo in a phase I study in 1993. This suggested there is some background event rate to be monitored on healthy volunteers taking placebo. However, elevations in biomarkers can be caused by many other things such as concomitant medications, whether a subject is ill (short term such as a cold or longer term) or patient demographics. Clinicians working on FIH studies often need to decide if adverse events on a new drug are likely due to the drug or occurring by chance. They do not want to withdraw the drug if it is indeed safe but on the other hand do not want to ignore a potential safety concern such as DILI.

Quantitative tools that identify and characterise safety issues earlier in the life cycle of investigational compounds would have a large impact on the efficiency of drug development, since the costliest phases of drug development are phase II and phase III [183]. Yet, early phase studies often lack quantitative evaluations of safety data. Previous attempts at quantifying the background rate of liver enzyme elevations in placebo-treated healthy individuals [85, 184] were limited by small datasets and did not consider the demographic and background characteristics of the healthy volunteers in their estimation of the incidence rates. For most other safety events, no reference rates are available at all.

We first model the synthesised data to estimate the incidence of safety events accounting for potential study differences and subject characteristics. This chapter then explores how teams working on such trials can use existing data of placebo incidence rates to inform the analysis of a new FIH study. It is therefore important to know the expected incidence of safety signals, such as elevated biomarker levels in these populations, and in particular, how this incidence varies in certain populations with different demographics. Subsequently, how the results from the model are used to predict the probability of a given subject experiencing a safety event above the upper limit of the normal range are discussed.

7.3 Summary of the available empirical data

All clinical studies included in this chapter were sponsored by Novartis and reviewed by an institutional review board (IRB). We retrospectively reviewed *all* studies in the Novartis database that involved healthy volunteers and were completed before 2016. Of those, we excluded 67 studies that did not involve placebo and 44 studies that used a cross-over design. There were 11 studies for which the laboratory and vital sign data were not readily available. All the placebo data from the remaining 77 studies were pooled. Among these 77 studies, 10 were conducted in Japanese subjects and 1 in Chinese subjects. The number of placebo subjects per study varied between 3 and 57, with a mean of 16.03 per study. The number of post-baseline observations per study varied between 1 and 18, with a mean of 5.65.

Data on the following routinely measured safety parameters, was collected:

- Liver safety: ALT, aspartate aminotransferase (AST), bilirubin
- Cardiovascular safety: the Fridericia-corrected QT interval (QTcF), standing systolic blood pressure (SBP), heart rate (HR)
- Renal safety: serum creatinine
- Pancreatic safety: lipase, amylase

While normal laboratory ranges are known for all of the parameters listed above, incidences of randomly occurring values outside of the normal ranges for healthy subjects receiving placebo in the setting of a clinical study are not known. The pooled database includes 1234 subjects with available measurements in at least one of the safety parameters above.

Demographic and background information for each subject were collected:

- Gender
- Ethnicity
- Height
- Weight
- Baseline measurements of the above safety parameters

The demographic characteristics of the population are presented in Table 7.1. The median age of subjects was 32 years (IQR 18-43) and 82.4% (1017/1234) were male. Age had a slight bimodal distribution which may be explained by the inclusion of 5 studies which had a specific inclusion criterion of >60 years. These studies were excluded as part of a sensitivity analysis to check if any conclusions changed. The mean height and weight were 174cm (SD=9.2) and 77.1kg (SD=12.5) respectively.

We present the raw incidences of various safety events in the pooled database and explain how these can be used to give a preliminary assessment of whether signals observed during the use of an investigational drug are in line with the expected incidence on placebo. The raw incidence of various safety signals in our pooled database of healthy volunteers receiving placebo are shown in Table 7.2, by target organ. This information can be used to judge how frequently random safety findings occur in a healthy population. For example, we see that increases in HR by more than 20 beats per minute from baseline occur in about 14% of healthy subjects receiving placebo. Now suppose we observe heart rate increases of this magnitude in 2 subjects (out of a cohort of 6 subjects) receiving the active drug. However, this would not necessarily be a concern, since it is not unlikely to happen in the same population receiving placebo: the probability of observing at least 2 subjects with an event in a cohort of 6, if the event truly occurs with a 14% probability, is 20.3% (from the binomial distribution, see Table 7.3).

Table 7.1: Demographics of the placebo database.

N=1234 subjects	**n**	**%**
Sex: Male (n, %)	1017	82.4
Ethnicity (n, %):		
White	662	53.6
Hispanic or Latino	259	21
Asian	141	11.4
Black or African American	125	10.1
Other	47	3.8
Continent (n, %):		
America/Canada	707	57.3
Europe	399	32.3
Australia	40	3.2
Asia	88	7.1
Age (years) (Median, Q1-Q3)[1]	32	25 - 43
Height (cm) (Median, Q1-Q3), N=1212[2]	175	168.2 - 181.0
Weight at baseline (kg) (Median, Q1-Q3)[3]	77.1	68.0 - 85.8

[1] Q1 = first quartile (25th percentile), Q3 = third quartile (75th percentile). Age ranged from a minimum of 18 to a maximum of 78 years.

[2] Height ranged from a minimum of 143.8 to a maximum of 199.0 cm.

[3] Weight ranged from a minimum of 47.7 to a maximum of 116.1 kg.

Table 7.2: Raw incidence (unadjusted for study effect) of safety signals in pooled database of placebo-treated healthy volunteers.

Target organ/Safety event	**Raw incidence**[1]	**Estimated incidence rate (%)**
Liver:		
ALT > ULN	77/1234	6.24
ALT > 2 x ULN	10/1234	0.81
ALT > 3 x ULN	4/1234	0.32
Bilirubin > ULN	92/1180	7.80
Bilirubin > 2 x ULN	36/1180	3.05
Bilirubin > 3 x ULN	30/1180	2.54
ALT or AST > 3 x ULN; & Bilirubin > ULN	0/1234	0
Cardiovascular system[2]:		
QTcF change > 60 ms & QTcF < 500 ms	7/1028	0.68
QTcF change > 60 ms & QTcF $\geq$ 500 ms	0/1028	0
HR increase > 20 bpm	165/1165	14.16
Standing SBP increase > 20 mmHg	64/790	8.10
Kidney[3]:		
Serum creatinine increase > 50%	0/1234	0
Pancreas:		
Lipase > 1.5 x ULN	34/1125	3.02
Lipase > 3 x ULN	7/1125	0.62
Amylase > 2 x ULN	4/1195	0.33

ALT=alanine aminotransferase; AST=aspartate aminotransferase; QTcF=Fridericia-corrected QT interval; HR= heart rate; standing SBP= standing systolic blood pressure (SBP) which is when blood pressure is taken when the subject is standing up. Units: ULN=upper limit of normal; ms=milliseconds; bpm=beats per minute; mmHg= millimetre of mercury.

[1] (number of subjects with at least one event /total number of subjects) in pooled early safety studies.

[2] Baseline QTcF ranged from 347 to 481 ms, with a mean of 398.0 (SD=12.5). Baseline HR ranged from 37 to 125 bpm, with a mean of 62.3 (SD=10.7). Standing SBP ranged from 86 to 168 mmHg, with a mean of 119.5 (SD=11.9).

[3] Baseline serum creatinine ranged from 35 to 168 umol/L, with a mean of 81.2 (SD=14.4). The ULN varied across studies with a median of 112 umol/L (IQR 106 to 115).

In our second example, suppose we observe 2 events, of an ALT elevation above 1 times the upper limit of normal, out of 6 subjects (Table 7.3). From Table 7.2, we know elevations of ALT above 1 times the upper limit of normal only occur in 6.2% of healthy subjects receiving placebo. We can use this information to try and make some inferences about how likely 2 events in our hypothetical study would have occurred by chance. We assume the probability that each person has an event is 6.2%, based on the population average. Therefore, the probability of observing 2 or more events out of 6 subjects by chance under placebo is 4.9%. As this is quite low, a clinician may attribute the cause of the event to the drug.

Table 7.3: Hypothetical situations in FIH studies and the corresponding probability of observing the same events under placebo.

Safety event	**Number of subjects under active drug with an event**	**Rate of event occurrence under placebo**	**Probability of observing 2 or more out of 6 events under placebo**	**Conclusion**
HR > 20 bpm from baseline	2/6	14%	20.4%	Situation is *likely* to have happened under placebo
ALT > ULN	2/6	6.2%	4.9%	Situation is *unlikely* to have happened under placebo

Whilst these raw incidence rates are helpful in providing a quick assessment of the likelihood for a safety signal to be caused by the active compound under investigation, more accurate answers can be given with a model adjusting for differences between study and individual subject characteristics. We use the liver enzyme ALT units per litre (U/L) as an example. We explore this lab maker only for the remainder of the chapter. This is a lab marker in the blood and is essentially a measure of how toxic a person's liver is. If it is elevated outside the normal range, it is defined as an adverse safety event.

7.4 Methods

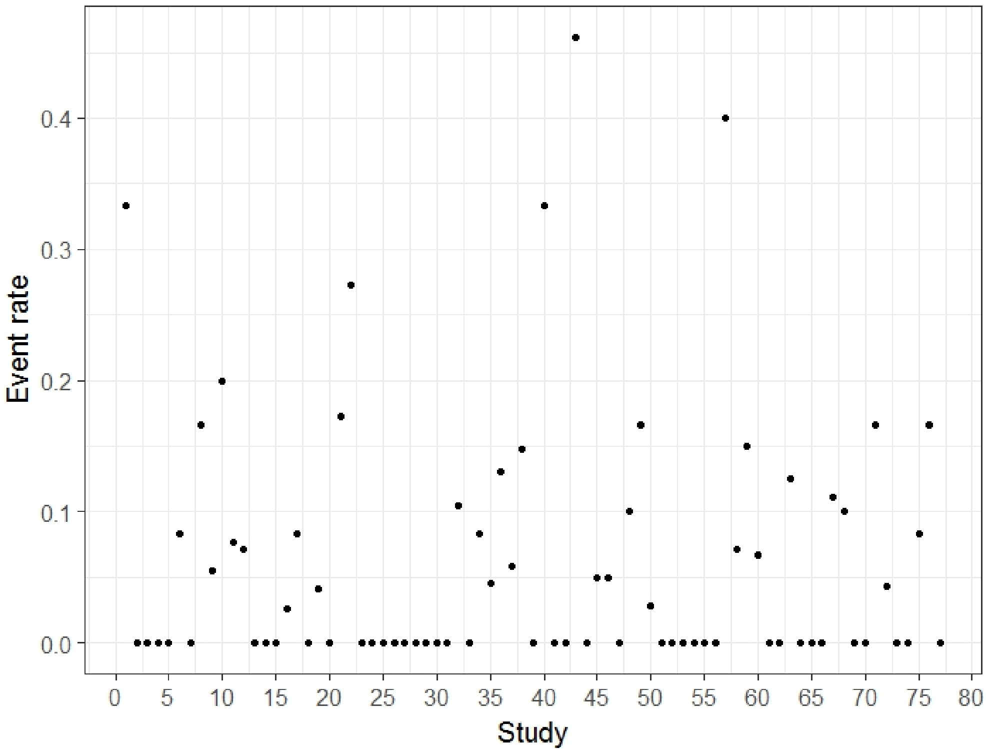
We have seen the raw incidences of various safety events; however, a model that adjusts for potential study differences and subject characteristics provides a more relevant assessment. We aimed to estimate, through a random effects regression model, how control arm ALT varies according to a number of demographic and baseline variables, including ALT at baseline.

7.4.1 Outcome

For each subject within a study, ALT was recorded at baseline and each time point for the duration of the study. Since each laboratory for a study (a study may have several laboratories) has different upper limits of normal due to the assays used, normalized data ("ALT/ULN") is used to allow for a standardized comparison across all studies. The standard approach to do this is to convert U/L to multiples of the ULN by dividing the ALT value by the ULN. For example, if the ULN was 45 U/L, then an ALT value of 12 U/L is converted to 12/45=0.27 ULN. ALT/ULN then becomes the response variable.

Baseline ALT varied from 4 U/L to 123 U/L. The median was 21 U/L (IQR 16-28). Similarly, the upper limit of normal varied. 77/1234 (6.2%) of subjects had at least one event of ALT>ULN. 67/1234 subjects had at least one event between 1 and 2 x ULN, 6/1234 had at least one event between 2 and 3 x ULN, 4/1234 had at least one event greater than 3 x ULN. As this is a safety event, our interest is in the maximum ALT value (for each person). 47% of studies had at least one patient with an event. By-study event rates are usually below 0.2, on rare occasion exceed 0.3, as shown in Figure 7.1.

Figure 7.1: Event rate by study.



Preliminary explorations

In the clinical trials community, clinicians are interested in whether ALT is greater than the upper limit of normal and therefore the binary outcome of an event. The most efficient way of modelling normalised ALT (not discarding information) is as a continuous variable. We can then calculate the probability that a subject has at least one ALT > ULN or 2 x ULN, which should be (if the model fits the observed data) approximately 6.2% and 1% respectively.

Our first approach was to model ALT/ULN as a continuous variable. Here we modelled the maximum ALT/ULN value for each person. We took the log of this value to make the outcome more normally distributed. Figure 7.2 shows the distribution of the maximum log(ALT/ULN) value of each subject. We fit a linear mixed model allowing for the clustering of subjects within each study (subjects i and study j). We therefore fit the following random intercept model, with a random effect on study:

$$(\log(\mathrm{ALT/ULN})_{ij} | x_{ij}, u_j) = \alpha + \beta_1 x_{ij} + u_j + \epsilon_{ij} \tag{7.1}$$

where $\epsilon_{ij} \sim \mathrm{Normal}(0, {\sigma_e}^2)$, and $u_j \sim \mathrm{Normal}(0, \tau^2)$. Here, x_{ij} is the baseline value of ALT for subject i in study j which is normalised (in units of ULN) and log transformed. However, when calculating the probability that a subject had at least one ALT > ULN the model predicted only 3%, compared with the observed value of 6.2%, i.e. the model did not fit the data well (Figure 7.3). We could not find a distribution that described the continuous data adequately (especially the tails, which are crucial for this exercise), providing unbiased predictions of the number of ALT > ULN events. In addition, the generalised mixed model type of approach tries to model the means, but our research question is about estimating the tails precisely which is difficult to do.

We therefore chose to dichotomise ALT, directly modelling the probability of an event defined as the maximum ALT>ULN or normalised ALT > 1. A random effects logistic regression model was fit with a random effect for study to allow for clustering of individuals by study, similar to equation (7.1) above.

Figure 7.2: Distribution of log(ALT/ULN).

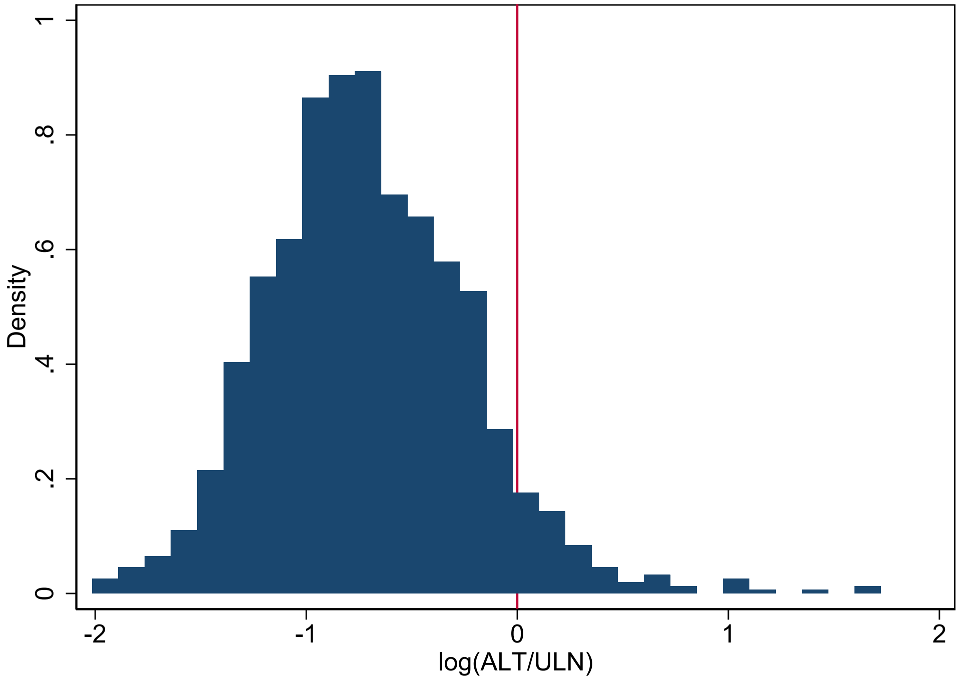
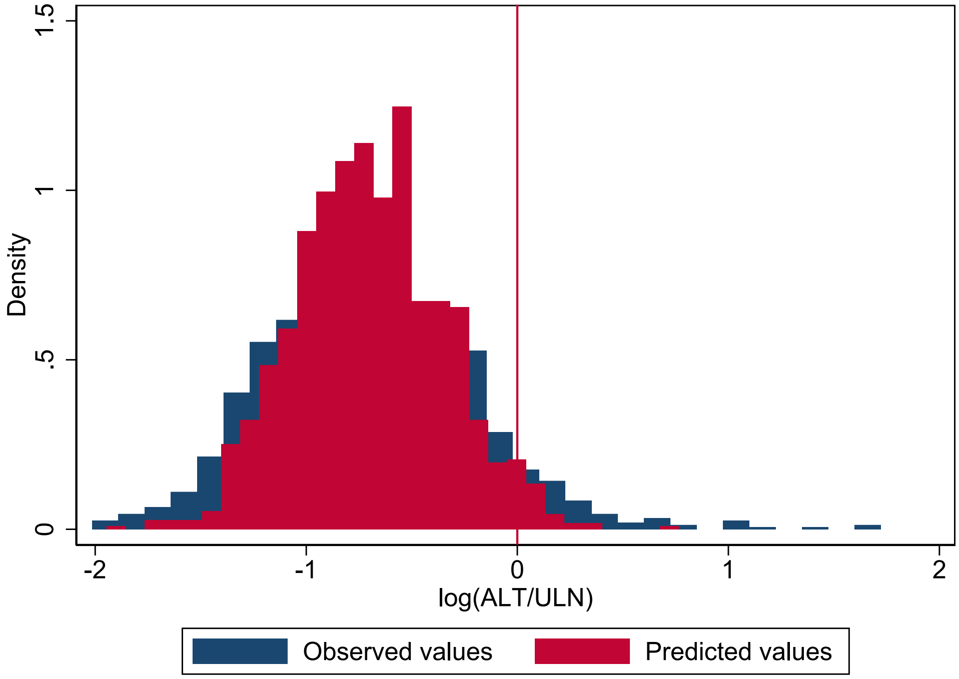


Figure 7.3: Distribution of observed log(ALT/ULN) values vs. the predicted log(ALT/ULN) values.



7.4.2 Description of the model

We fitted a multilevel logistic regression model in a Bayesian framework using non-informative priors as follows [185]. For subject i in study j we define the event $y_{ij} = 1$ if the subject had at least one ALT measurement exceeding the ULN during the study.

$$y_{ij} = \begin{cases} 1, & \text{if ALT} > \text{ULN} \\ 0, & \text{otherwise} \end{cases} \tag{7.2}$$

The probability of this event was modelled as $\mathrm{Bernoulli}\left[p_{ij}\right]$ with:

$$logit(p_{ij}) = \alpha_j + X_{ij}\beta \tag{7.3}$$

where α_j represents the study-specific intercept used to account for between-trial variation, X_{ij} is a vector of covariates specific to subject i (including the number of post-baseline samples for subject i), and β is a vector of covariate parameters. The model was fit in a Bayesian framework using the following weakly informative priors: $\alpha_j = \alpha + u_j$, $u_j \sim \mathrm{Normal}(0, \tau^2)$ with $\alpha \sim \mathrm{Cauchy}(0, 10)$, $\tau \sim \mathrm{Exponential}(1)$ and $\beta \sim \mathrm{Cauchy}(0, 2.5)$. A $\mathrm{Cauchy}(0, 2.5)$ prior assigns roughly 0.7 prior probability that the logit-coefficient is between -5 and 5. The prior for the between-study standard deviation assigns a prior probability of 0.95 that $\tau < 3$. Although only weakly informative priors are used, a Bayesian paradigm allows for full parameter uncertainty in comparison to the frequentist approach [13]. Since a logistic regression model is fit, model coefficient estimates are reported as the odds ratios based on posterior medians rather than means, as they are likely to be skewed.

7.4.3 Model covariates

Baseline ALT was taken post-randomization and prior to the placebo being given. Histograms of continuous covariates were produced to check for normality and potential outliers. Log transformation was applied to baseline ALT and age. Baseline ALT was calculated in the unit of ULN and then log transformed to obtain an approximately normally distributed variable. Age was also log transformed for the same reason. Log transformation was initially attempted for the number of samples; however, this did not improve model fit and therefore the original scale was used to aid interpretation. Weight was approximately normally distributed and therefore no transformation was considered necessary. All variables were standardized to a scale with mean of 0 and standard deviation of 1. The standardization was done via the typical approach, i.e. subtracting the mean of the variable from each of the individual (subject) values and dividing by the overall standard deviation. This means that each coefficient is unit less and therefore has a similar magnitude and better statistical properties [186].

7.4.4 Model selection and fit

The baseline covariates described in Table 7.1 were first included in random effects logistic regression models in univariable analyses, to identify any variables which were individually predictive of an elevated ALT event. Model fitting was done using Stan [187] via the R library (version 3.4.1) RStanArm (version 2.15.3) [188].

All of the baseline covariates were evaluated for inclusion in the final multivariable model, using a forward selection approach that calculates the difference in deviance for nested models [189]. Deviance was calculated using the leave-one-out (LOO) method which is a measure of how much the posterior distribution would change if a single observation were omitted [186]. The model with the smallest expected log point wise predictive density was selected [186]. Model fit was checked by comparing posterior predictive distributions to observed values. To check adequacy of model fit, we simulated observations from the posterior predictive distribution, the distribution of the outcome implied by the model after using the observed data to update our beliefs about the unknown parameters [13]. These are simulated datasets based on the same observations of our covariates that were used to estimate the model parameters and denoted by y_{rep} [186]. If the model fits the data well, then we would expect the simulated datasets from the model to be very similar to the observed y. The LOO method can also be seen as a check of potential influential observations or outliers which highlights any observations which are not predicted well by the model based on the other data points.

7.4.5 Predicting from the model

The aim is to provide a tool for clinicians who want to interpret emerging results in an ongoing FIH study. To do this, the final model is used to calculate conditional predictions of an elevated ALT>ULN for each of the active-treated subjects using their specific covariate values. If covariates are not available, the population average of 0.062 is used (from Table 7.2). From this model, we can derive the subject-specific probability of experiencing an event, conditionally on this subject's covariates. For subject i in our dataset, we would condition on the study-level effect α_j, while for a new subject this probability would be obtained by integrating over the study effect distribution [190, 191]:

$$p\left(y_{ij}|X_{ij} \ = x_{ij}\right) = \int_{-\infty}^{\infty} \frac{e^{\alpha+ \ u_j+x_{ij}\beta}}{1+e^{\alpha+ \ u_j+x_{ij}\beta}} \ f\left(\alpha,\beta,\ \tau,u_j\right) \ d(\alpha,\ \beta,\ \tau,u_j) \tag{7.4}$$

We substituted the subject's covariate values into the posterior draws of the linear predictor using the inverse logit function and conditioned on the study they belong to. This gives a matrix of size [N posterior draws] x [N subjects]. We therefore integrate with respect to the joint posterior of all the parameters. For a new subject, who does not belong to an existing study, we additionally integrate over the distribution of study effects. This is equivalent to deriving a marginal (population average) probability of a subject experiencing an event by averaging over the distribution of covariates. These probabilities were calculated in R version 3.4.1 using the functions posterior_predict and posterior_linpred, part of the R library RStanArm.

Due to the non-linearity of the model, the conditional probability of a subject with mean covariate values is not expected to match the population mean marginal probability [185]. In other words, taking the mean before or after the logit transformation will not yield the same results, as is always the case for generalized linear models with a non-linear link function. In this setting, because we are mostly interested in predictions for specific subjects in a new study (for whom we know the baseline covariates), the conditional probability is of most interest.

For a subject receiving the active drug in a new study, we produced the predicted probability of this subject experiencing an ALT > ULN event had he/she received placebo instead of active drug, based on the model described above. We will call this prediction the probability of a 'virtual placebo twin' to experience the event. Whether or not the virtual placebo twin is also likely to have experienced an event will determine whether the investigational drug is likely to have caused the event or not.

Finally, we combine the individual subjects' probabilities to estimate the probability that at least one subject experiences one event in a cohort of size n. This can be done by using conditional probabilities (a different one for each subject) if subject-specific covariates are available, or marginal probabilities (identical for each subject) if subject covariates are not available.

7.5 Results

7.5.1 Univariable analyses

Results from univariable analyses are shown in Table 7.4.

Table 7.4: Results of univariable analyses for each variable fit in a random effects logistic regression model with a random effect on study. Results are reported as odds ratios and the between study variance on the log odds scale.

Predictor	Odds ratio	95% CrI	Between study variance	elpd_loo	p_loo	looic
Baseline ALT/ULN (log transformed)	4.65	(3.38, 6.69)	1.14	-215.4	28.4	430.9
Number of post-baseline samples taken	1.68	(1.26, 1.87)	1.19	-264.6	31.1	529.1
Age in years (log transformed)	0.96	(0.74, 1.24)	1.17	-270.8	31.7	541.6
Height in cm	0.81	(0.68, 1.14)	1.13	-270	31	540
Weight in kg	1.14	(0.88, 1.49)	1.19	-270.2	32.1	540.4
Sex (male)	1.78	(0.89, 4.14)	1.2	-269.3	32.5	538.6

elpd_loo=the expected log predictive density; p_loo=the is the posterior mean of the deviance minus the deviance of the posterior means and defined as is 'the effective number of parameters looic=the leave one out information criterion which is used to help compare models; the model with the smallest looic is estimated to be the model that would best predict a replicate dataset which has the same structure as that currently observed.

Following the model selection procedure described earlier, the LOO method (Table 7.5) indicated that the best fitting model (i.e. that with the lowest leave-one-out information criteria (LOOIC)) included baseline ALT, age, weight, and number of samples in the study as fixed effects, and study as random effect (Table 7.6).

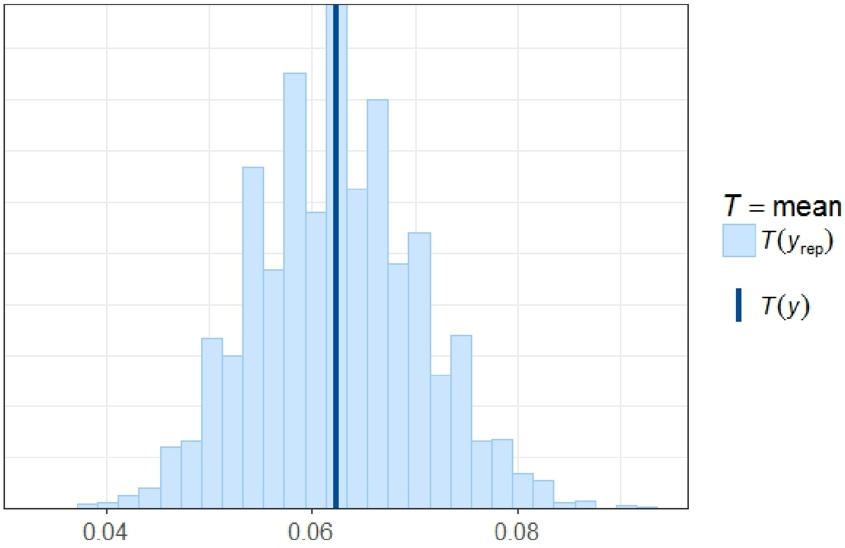
Table 7.5: LOO deviance table when deciding on the final model.

Model	elpd_loo	p_loo	looic
Null model	-269.4	30.6	538.9
Null model + log(baseline/uln)	-215.4	28.4	430.9
Null model + log(baseline/uln) +samples	-211.8	30.9	423.6
Null model + log(baseline/uln) +samples + log(age)	-210.9	31.8	421.8
Null model + log(baseline/uln) + samples + sex	-212.3	32.3	424.6
Null model + log(baseline/uln) + samples + weight	-211.6	32.2	423.1
Null model + log(baseline/uln) + samples + height	-212.5	32.1	424.9
Null model + log(baseline/uln) + samples + log(age) + weight	-210.4	32.8	420.4
Null model + log(baseline/uln) + samples + log(age) + height	-210.5	32	421
Null model + log(baseline/uln) + samples + log(age) + sex	-212.4	33.5	424.8
Null model + log(baseline/uln) + samples + log(age) + weight + sex	-211.3	34.6	422.5
Null model + log(baseline/uln) + samples + log(age) + weight + height	-211.7	33.7	423.4
Null model + log(baseline/uln) + samples + log(age) + weight + height + sex	-211.3	35	422.6
Null model + log(baseline/uln) + samples + age + weight	-209.9	33	420.4

7.5.2 Model diagnostics

Figure 7.4 shows the mean of each of the simulated datasets (light blue) are very similar to the mean observed proportion of events 77/1234=0.062 (dark blue).

Figure 7.4: Simulated mean proportion of events from the posterior predictive distribution (light blue) and the observed proportion of events (dark blue).



Conclusions from the model remained unchanged when the following sensitivity analyses were carried out: (i) removing studies with inclusion criteria of age greater than 60, (ii) height added back into the final model. The final model was also replicated in WinBUGS, with Normal (0, 1000) priors for all location paramters (regression coefficients) and a Uniform (0,5) for the between study SD, and each of the coefficients estimated were within 0.001 of each other.

7.5.3 Model estimates and interpretation

The final covariates and model coefficients are shown in Table 7.6.

The global/population mean and standard deviation for each of the predictors of the final model are given in Table 7.7.

Table 7.6: Final model coefficient estimates for the logistic model of the probability of a subject experiencing an event of ALT > ULN.

Model coefficients[1]	Posterior median (Odds ratio)	95% credible interval
Fixed effect coefficients:		
Intercept	0.016	(0.0076, 0.028)
Baseline ALT/ULN (log transformed)	5.21	(3.67, 7.77)
Number of postbaseline samples taken	1.7	(1.25, 2.36)
Age in years (log transformed)	0.72	(0.53, 0.97)
Weight in kg	0.73	(0.52, 1.00)
Random effects coefficient:		
Between-study variability $\hat{\tau}^2$	1.20	(0.39, 2.84)

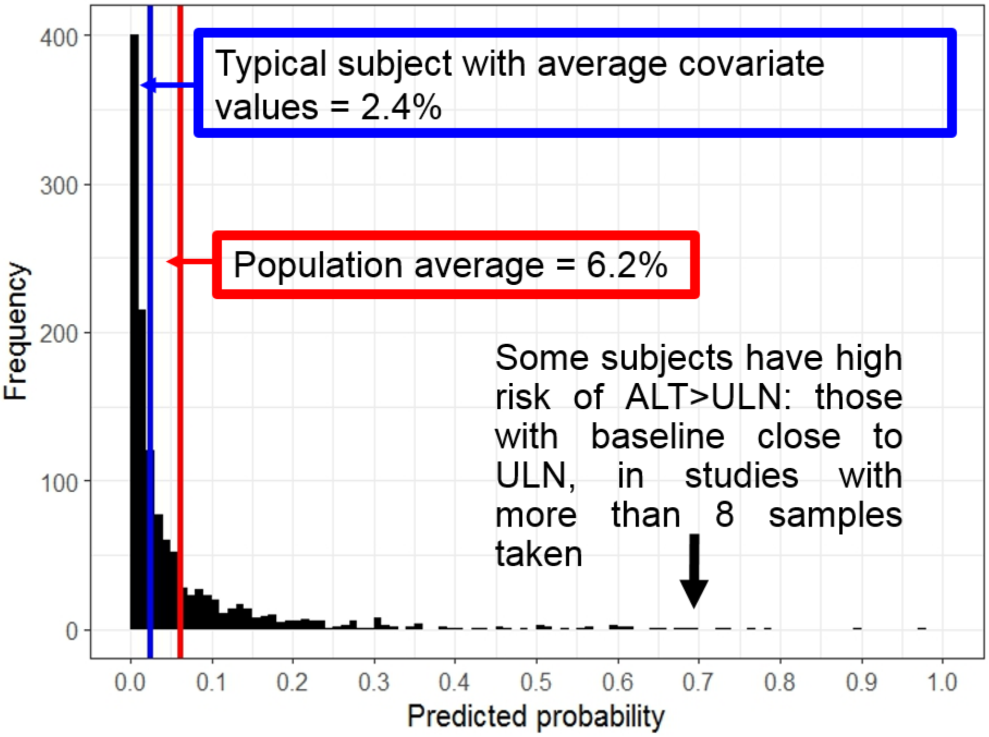
[1] All variables are standardised to a scale with a mean of 0 and a standard deviation of 1.

Table 7.7: Mean and standard deviations of the variables used in the model

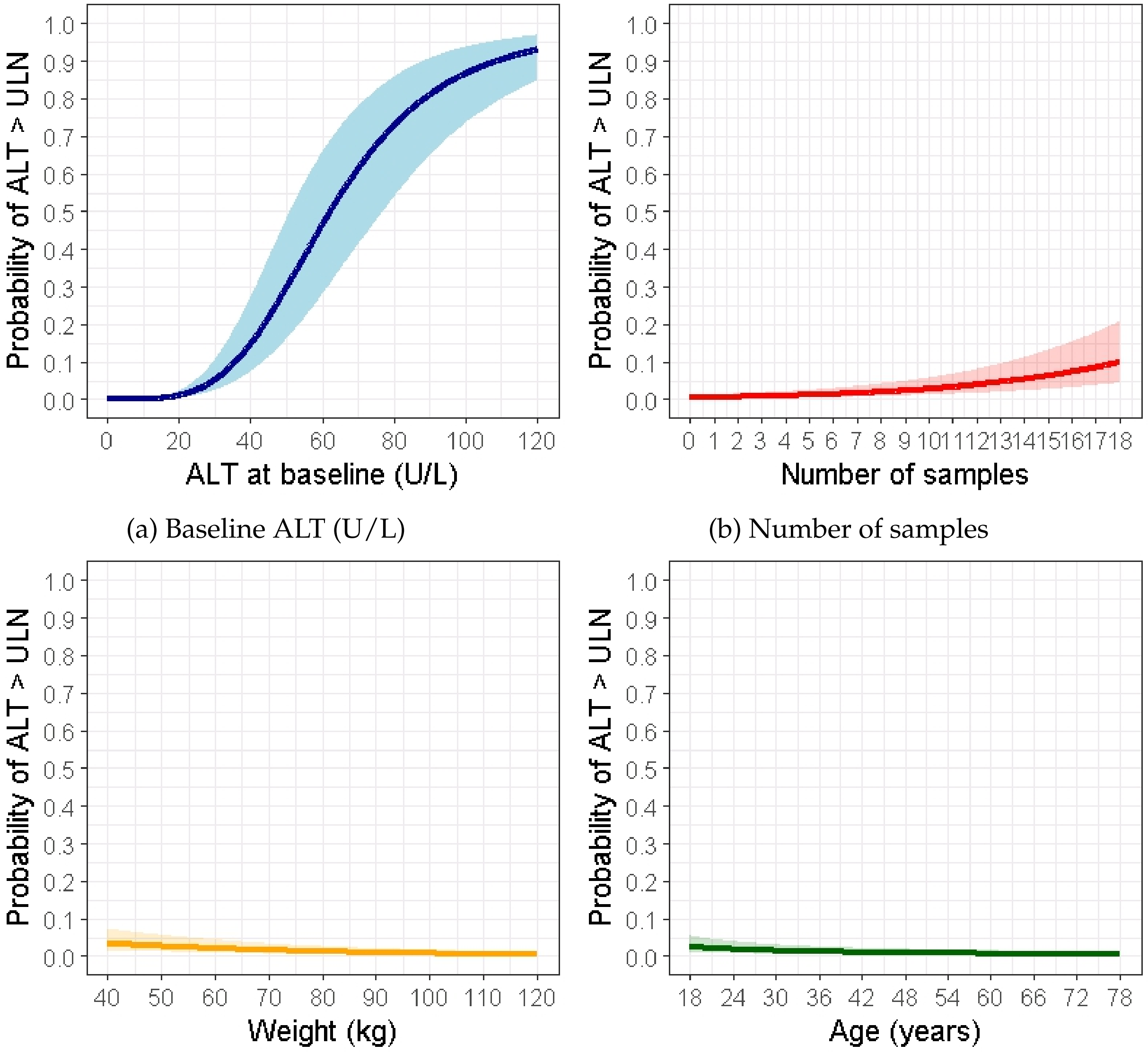
	Mean	SD
Log(baseline/uln)	-0.93	0.43
Number of measurements taken	5.65	3.13
Log(age)	3.50	0.32
Weight	77.11	12.50

Figure 7.5 shows the individual probabilities of ALT>ULN for all 1234 subjects in dataset conditional on their covariates and the study they are in. It also illustrates that some subjects have a high risk of ALT>ULN: those with baseline close to ULN, in studies with more than 8 samples taken. For a typical subject with average covariate values (baseline ALT value of 21.5 U/L, ULN of 55 U/L, aged 32.8 years with a weight of 77.1 kg, and 5.65 post-baseline observations over the study), the probability of developing an ALT > ULN is 2.4%. This is substantially lower than the population mean of 6.2%, because the distribution of probabilities is skewed to the right, and higher risk subjects drive the average up.

Figure 7.5: Model-based predicted probabilities of a subject experiencing an event of ALT > ULN



The most influential covariate is a subject's baseline ALT value and how close it is to the ULN. Figure 7.6a shows that the predicted probability of an ALT > ULN event varies from 2% for a baseline of 20 U/L to 19% for a baseline of 40 U/L, controlling for all other covariates.



(a) Baseline ALT (U/L)

(b) Number of samples

(c) Weight (kg)

(d) Age (years)

Figure 7.6: Predicted probability of an ALT>ULN for a subject, assuming other covariates are fixed at the mean population values. 25% and 75% quartiles are also displayed. The median probability is assumed.

The second most predictive covariate is the number of samples: subjects in studies with a higher than average number of samples taken have a higher probability of an ALT>ULN. The probability of an elevated ALT>ULN gradually decreases as weight and age increase. The between-study variance is 1.20 (on the log odds scale), which represents 27% of the residual variance. The estimated impact of age and weight on the probability of a subject developing an ALT elevation above ULN during a study is small in magnitude, and perhaps counter intuitively, is negative: increasing age and weight is associated with reduced risk of a random ALT elevation under placebo. Of note, including sex did not improve the model significantly; however, this could be due in part to the fact that 82% of the subjects in our pooled database were male.

Table 7.8 compares the characteristics of subjects who had an elevated ALT>ULN (N=77) compared to those who did not (N=1157). The distributions of age (Median=33, IQR: 26-39 vs Median=32, IQR: 25-43) and weight (Median=78.4, IQR: 69-86 vs Median=77, IQR: 68-85) are similar in subjects who experience an event of ALT>ULN and those who did not.

Table 7.8: Characteristics of subjects who had an elevated ALT>ULN (N=77) compared to those who did not (N=1157)

	Elevated ALT>ULN (N=77)		**No elevated ALT>ULN (N=1157)**	
Characteristics	**Median**	**IQR**	**Median**	**IQR**
Age (years)	33	26-39	32	25-43
Weight (kg)	78.4	69-86	77	68-85
Baseline (U/L)	34	27-46	20	15.5-27
Number of samples	6	4-8	5	3-7
Upper limit of normal (U/L)	50	45-60	55	45-60

7.5.4 Using the pooled placebo data to inform a new trial

Using the event ALT > ULN as an example, the raw incidence of elevations in healthy volunteers taking placebo is 6.2%, as shown in Table 7.2. However, this prediction varies substantially across individuals. We therefore use model based predictions which adjust for potential study differences and subject characteristics rather than the population average.

When subject characteristics are available

To illustrate how these findings can be applied in practice in the setting of a dose escalation study of a new investigational drug, consider a hypothetical cohort of 6 active-treated subjects. We create a hypothetical data set in which some subject's have high ALT values at baseline, to illustrate the substantial impact of this variable. Other baseline characteristics were selected at random within the interquartile range. Applying the model described above generates the probability that each subject's placebo twin would have had an event (i.e. the probability that this subject would have developed an event had he/she been receiving placebo), as shown in the last column of Table 7.9.

Table 7.9: Hypothetical cohort of active-treated subjects and individual model-based predictions for each individual's placebo twin

Subject ID	**Baseline ALT (U/L)**	**ULN (U/L)**	**Number of samples**	**Age (years)**	**Weight (kg)**	**Probability ALT>ULN**
1	25	55	6	22	75	0.045 (0.0036, 0.31)
2	29	55	6	28	72	0.066 (0.057, 0.40)
3	33	55	6	47	70	0.067 ((0.057, 0.41)
4	42	55	6	25	80	0.22 (0.023, 0.73)
5	45	55	6	52	76	0.16 (0.015, 0.66)
6	49	55	6	32	74	0.32 (0.038, 0.82)

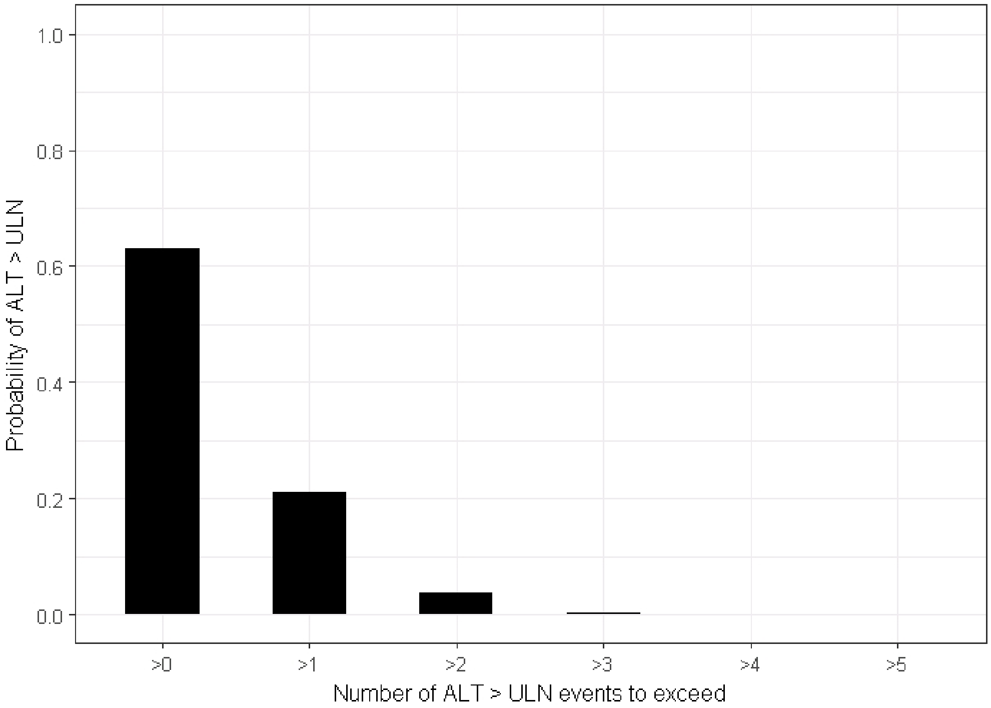
The probabilities are based on median values.

The probability of one or more of the subjects (with the same characteristics as those receiving the active intervention) experiencing an elevated ALT > ULN event in the 'hypothetical' placebo group is 0.63 (63%).

Pr(at least one event) = 1 – P(zero event) = 1 – (0.955 * 0.934 * 0.933 * 0.78 * 0.84 * 0.68) = 0.63.

Therefore, if one or two such events were observed in the cohort receiving the compound under investigation, the clinical team is likely to attribute this to chance. Figure 7.7 illustrates the probability of observing at least 1, 2, 3, 4, 5, or 6 placebo twins with an event in this cohort. However, if the event occurred in subject number 1 or 2 then we may be concerned. Furthermore, if a subject's model based prediction was low but did have the event, this may also raise safety concerns to the clinical team.

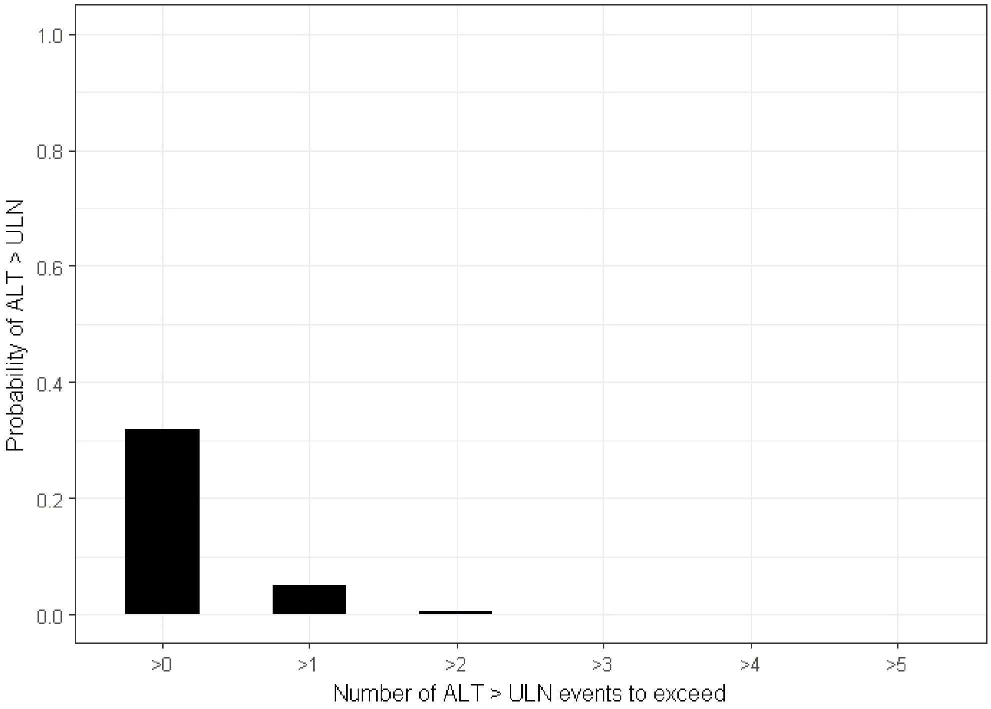
Figure 7.7: Probability of observing a given number of subjects with an event (ALT > ULN) by chance in the hypothetical cohort described by Table 7.9.



When subject characteristics are unavailable

When subject specific data is unavailable a clinician would use the population average of 6.2%. That is, on average, we expect a healthy volunteer to have a 6.2% chance of an ALT > ULN at any time during the study. In a cohort of 6 subjects, this translates into a 32% chance of observing at least one ALT>ULN in at least one of the six subjects.

Figure 7.8: Probability of observing a given number of subjects with an event (ALT > ULN) by chance using the population average probability



Scenario in practice

A common scenario in practice may be observing 1 out of 6 subjects on the experimental drug with an ALT elevation > ULN, whilst observing no events in the placebo arm, shown in Table 7.10.

Table 7.10: 2 by 2 analysis

	Experimental	**Control**	**Total**
Event	1	0	1
No event	5	2	7
Total	6	2	8

For the experimental subjects suppose we calculate the predicted probabilities of having an event under placebo, as shown in Table 7.9. We can translate this into an average probability to see if our observed proportion of events in the experimental arm (1/6=0.17) is consistent:

$$\begin{aligned}\text{Average predicted probability} &= \frac{0.045 + 0.066 + 0.067 + 0.22 + 0.16 + 0.32}{6} \\ &= \text{0.15 (95\% CrI: 0.016 to 0.55)}\end{aligned} \tag{7.5}$$

The observed number of events in the experimental arm is consistent with what would be expected under placebo, i.e. not a cause for concern.

7.6 Discussion

This is the first known large review of the expected frequency of random safety findings in placebo-treated healthy volunteers. We developed a model to predict how likely a safety event is due to occur by chance, conditional on the characteristics of the subject and the study. The key findings and how this study adds to current knowledge are displayed in Figure 7.9 and now discussed.

Figure 7.9: Key findings

- Little information is available in the literature on the expected variations in laboratory values and vital signs in a clinical study setting, when subjects receive placebo. A few reviews have mentioned unexpectedly high rates of ALT elevations in subjects taking placebo, but were based on small datasets and did not adjust for individual characteristics.
- Our goal was to provide a tool for clinicians working on FIH studies, to quantify whether safety signals observed were likely the result of chance or the compound under investigation.
- We synthesised data from a set of previous studies to provide estimates for the incidence of how likely an event is due to chance, conditional on the characteristics of the subject and the study.
- These could be used to formulate informative prior distributions through the (future development of) Bayesian approaches. This work should help teams identify safety signals earlier and with greater accuracy.

7.6.1 Overview of key findings

The reference event rates provided for parameters relating to the safety of the liver, cardiovascular system, kidney and pancreas will provide valuable insight for clinical teams assessing the safety of investigational compounds in early phase studies. Using the liver enzyme ALT as an example, we showed how predictive models can provide a more precise assessment of the chance occurrence of a safety signal in a subject. We developed a random effects logistic model for the probability of a subject developing an ALT > ULN event during a study, while taking placebo. This model showed that the most important factor influencing the chance of a random event is the ALT value at baseline. In our dataset, the mean probability estimates of a healthy volunteer to develop an ALT > ULN event under placebo is 6.2%. For a typical subject with average baseline characteristics (baseline ALT value of 21.5 U/L), this probability is 2.4%. If the baseline ALT doubles to 40 U/L, this probability increases to 19%. This illustrates that caution should be taken when interpreting ALT elevations in cases where the baseline value is higher than usual. Sponsors may elect to recruit only subjects with a predicted probability of a random elevation that is lower than 10% in FIH studies, since for those subjects, it will be easier to attribute any emerging liver safety signal to the investigational drug. If an individual has a low model-predicted probability under placebo but has the event, then the higher the chance that the drug under investigation is causing the issue.

7.6.2 Limitations

There are a few limitations with this work. The first one is the choice of modelling the ALT > ULN on a binary scale. We first tried to model ALT on a continuous scale, as this should theoretically make more efficient use of the data than dichotomising. This can be seen as a limitation although our final model did fit the data well. However, as documented in Section 7.4.1, modelling ALT as a continuous outcome did not fit the observed data well. Furthermore, our primary interest, was the probability of a safety event (of ALT>ULN) which is viewed as a binary variable in clinical practice.

Another limitation is that some of the events described in the text as ‘random elevations’ may be partially explained by factors that were not captured in the database. Differences in whether subjects were kept under controlled conditions (domiciled) and subject management between studies may explain some of these differences, and unmeasured medical history or other study specific design features (such as food intake, etc.) may explain others. For example, while our database did not capture this information, we know that most subjects in the dataset were domiciled at least for the first 3-5 days of the study and had normal access to food three times a day. Data collection in FIH studies is likely not optimal for this type of exercise. Each company may have different standards for first-in-human protocols and the numbers observed in studies sponsored by other companies than Novartis may look slightly different for this reason. Most FIH studies have samples collected in the first 3 to 5 days under domiciled conditions. Subjects under domiciled conditions are much less likely to have access to alcohol or drugs and as a result are less likely to have liver enzyme elevations due to that. As such it is possible these are true random fluctuations and random elevations outside this period could be incidences of alcohol or drug abuse. It is intuitive to assume subjects under domiciled conditions will have a lower incidence rate because of having less access to alcohol. If not, it is possible that elevations are driven by other internal biochemistry rather than the external factors, such as the environment. However, it is not routinely recorded whether an observation is under these conditions and so it is not possible to compare the incidence rate of elevations.

7.6.3 Previous studies looking at safety data in early phase trials

Previous attempts at quantifying the expected incidence of ALT elevations in healthy volunteers taking placebo [184] have reported a higher ALT > ULN event rate of 20.4%. They found that 19/93 subjects (20.4%) had at least one ALT value > 1 x ULN and 7/93 (7.5%) had at least one value >2 x ULN. They found the incidence was extremely rare in the first week of hospitalization and increased over time during the second week. ALT levels returned to normal after a few days of hospital discharge. The inclusion criteria included males aged 18-40 who were deemed healthy with no significant diseases after a physical examination however subjects with a high baseline were still enrolled in the studies. Between 2 and 7 samples were taken during the 14 days. Therefore, the difference is incidence is probably due to the small dataset used and the inclusion of less healthy subjects with higher baseline ALT levels. This highlights further the importance of controlling for baseline when making these predictions.

Kobayashi *et al* found a lower prevalence of 12.5% (13/104) in 104 health male volunteers in their phase 1 study compared to Rosenzweig *et al* but their review included studies with treatment duration of only 7 days. Similar to our pooled FIH studies, they found average ALT levels were higher at baseline in the group with elevated levels (13 subjects) compared to the group with non-elevated levels (91 subjects).

Informative priors have been used in phase 2 trials which have small populations like rare diseases and paediatrics, populations which are difficult to recruit [192]. Here, historical controls have been used to supplement the number of controls needed in the new trial and reduce the sample size needed. However, the focus is on aggregate level data summarised at the study level (rather than individual patient data) and using power priors [193].

7.6.4 Implications for routine safety data in practice and further research recommendations

We showed that a predictive model can be used to create 'virtual placebo twins', i.e. subjects with the same baseline characteristics, but who would have received placebo: for every subject experiencing an event under an investigational drug, the model will predict the likelihood of his/her 'virtual placebo twin' experiencing the same event. The lower this model-predicted probability, the higher the chance that the drug under investigation is causing the issue. By using the large amounts of placebo data collected in healthy volunteers over decades of clinical investigations, companies can contribute to increasing the quality of safety decision making in early phase clinical trials. We can quantify, with higher precision, the expected frequency of random safety signals in FIH studies and separate real signals from the noise.

In early phase studies, decisions are taken in the context of all available data (including pre-clinical evidence, pharmacokinetic data, etc.) and this tool is not intended to lead teams to ignore this complexity of information; rather, it should be viewed as a way to consider one piece of the complex array of data (namely, the rates of laboratory abnormalities) in a more objective and quantitative way.

Variability in the distribution of study effects does add a considerable amount of uncertainty to the predictions, however, it would be incorrect to ignore this when making predictions [191]. We have compared predictions from our model to observed events in a new study to see how likely they would have occurred. Future work could also investigate whether a particular study is more similar to the new study. In this case, a shrunken estimate from a similar study, in terms of its inclusion criteria, could be used to form a prior distribution. It may also not be clinically relevant to average across the random effects distribution. Instead, researchers could look at baseline risk in each study. Furthermore, in our scenario in practice, using Table 7.10 our simple approach does not account for the matching of predictions and observations. Further work should be conducted to explore how best to do this and then second how to make use of the two patients in the placebo arm.

A new policy should be put into practice to regularly update the database to get to the 'true' population. Such a model exercise could be repeated regularly e.g. annually to include more newly generated data. For implementation, R Shiny apps could be developed to translate these findings into user-friendly tools for clinical teams. In particular, to produce model-based predictions with new covariate values based on the active-cohort in a new FIH study. This will aid interpretation of emerging results in the ongoing study. Future work should include looking at the difference of incidence under domiciled and non-domiciled conditions. Further work should also include modelling other safety parameters, such as AST.

Finally, this work could be extended to patient populations (in particular for rare populations [194], for example paediatric studies [195, 196]). Hampson *et al* [196] have also found that there is a lack of data in paediatrics and potential for use of informative priors. This would come with additional complications, since patient studies often do not share similar designs and inclusion criteria, unlike FIH studies of healthy volunteers. Nevertheless, we think the outcome of this exercise could help distinguishing effects related to the drug under study from the underlying disease, in populations where the effect of placebo has been poorly studied. In addition, this work could also be extended by using historical control data from patient studies in the same or similar disease area to inform a new patient study. For example, when information is available on the expected event rate in the control population but the control group in the new study may not be powered to detect rare or adverse events. As seen in the literature (Section 1.4), external data could be used to inform the trialist of the expected population rate of particular outcomes in the control population, rather than the placebo population here. Furthermore, where IPD are accessible, for example, collated data within a trials unit, trialists may refine the population estimate by deriving individual expected predictions of an event using our model based approach. The trialist would then be better equipped to interpret the observed adverse event rate.

7.6.5 Conclusions

Liver toxicity is a widespread problem in drug development. An elevated ALT level is specific to liver toxicity which usually results in the development of a compound being stopped. However, what is unclear is why an elevated ALT has occurred. Since ALT elevations can occur in the general population it is not always a sign that the active drug is toxic. Therefore, in FIH studies of healthy volunteers, it is imperative to know whether an elevated ALT level is caused by the active drug or if it is likely to be caused by chance. One way to answer this question is to find the expected incidence rate in the healthy volunteer population taking placebo.

Little information is available in the literature on the expected variations in laboratory values and vital signs in a clinical study setting, when subjects receive placebo. A few reviews have mentioned unexpectedly high rates of ALT elevations in subjects taking placebo but were based on small datasets and did not adjust for individual characteristics. Our goal was to provide a tool for clinicians working on FIH studies, to quantify whether safety signals observed were likely the result of chance or the compound under investigation. As such this study provides reference incidence rates under placebo for commonly measured safety parameters.

We have developed models to appropriately synthesise relevant previous data to provide predictions. By modelling existing placebo data, we can make predictions about the chance occurrence of safety signals for individuals, rather than using the population average. We built a predictive model for ALT elevations which can be used to quantify precisely how likely an event is due to chance, conditionally on the characteristics of the subject and the study. These predictions could be used informally (compare observed events vs predictions from model) or more formally through (the future development of) Bayesian approaches. This work should help teams identify safety signals earlier and with greater accuracy. For pharmaceutical companies in particular, this data is already easily accessible and allows this array of data to be viewed in a more objective and quantitative way, and ultimately increase the quality of safety decision making in FIH studies.

8 Discussion

This thesis has explored areas in which external evidence syntheses can be used to inform the design and analysis of a new clinical trial, with specific attention to our three case studies of using external evidence to inform: (i) the likely amount of bias in an estimated treatment effect (using information from meta-epidemiological evidence), (ii) the likely treatment effect, for use in sample size calculations (using information from a meta-analysis), and (iii) the likely control group event rate when the sample size is small and events are rare (using individual patient information from a synthesis of FIH placebo arm studies) to potentially increase precision in analysis of FIH studies. We have compared these to the traditional approaches to determine potential advantages, key differences, and limitations. This chapter summarises the key findings and implications from the thesis and discusses areas for further research.

8.1 Key findings

The *incorporation* of external evidence syntheses to inform the design and analysis of a new RCT has long been discussed; mainly due to the intuition that we can learn from something that has been done in a similar setting [7, 197]. Traditionally, trials are undertaken in isolation and analysed within a frequentist framework, that is, trial results do not usually explicitly incorporate prior information about the potential effectiveness of two treatments (or any other parameters). In regard to *using* existing evidence, this is done [78, 198], but more descriptively rather than through Bayesian analyses [71].

In **Chapter 2**, our 2015 survey ('INVEST') investigated how trialists were using evidence synthesis in trials. We found they were using evidence synthesis informally in multiple ways to make inferences about the design and analysis of a trial. For example, using a systematic review to justify the intervention comparison, determine the most appropriate outcome or duration of follow-up or inform the size of the intervention that the trial is powered to detect. We also found trialists would like to be making more use of existing data but were not, due to time pressures and concerns over the relevance of such information to their trial.

Finding: The results of our INVEST survey, in Chapter 2, highlighted that trial teams responding to the survey at the ICTMC generally reported they are using evidence synthesis in trial design and analysis more than we might have expected, but less than they might like to.

Given the high proportions of participants reporting the use of external evidence in trial analysis, we felt that some participants may have interpreted our question about this differently from how we had intended: potentially referring to updating a meta-analysis rather than explicitly incorporating information from the meta-analysis in a Bayesian framework.

Therefore, a qualitative study was undertaken in **Chapter 3** to explore precisely *how* trialists were using existing external data when analysing trials, and whether there were areas in which such evidence could be advantageous over current methods, through semi-structured one to one interviews. Our findings were consistent with our INVEST survey, in that, trialists were using evidence synthesis in several ways; and time pressures and concerns over the relevance of previous data to their own trial were barriers to the use of evidence synthesis in practice. However, we found trialists rarely (never in our sample) incorporated external evidence with data from their own trial through Bayesian analyses.

Finding: Our qualitative study in Chapter 3 confirmed trialists do use existing evidence a lot to inform different aspects of their trial but only ever informally, and trialists rarely (never in our sample) explicitly combined this previous evidence statistically with data from their own trial.

We additionally found that barriers to explicitly incorporating external evidence syntheses into trials were (i) personal feelings: a lack of confidence in Bayesian methods; (ii) practical challenges such as accessing data with no infrastructure, subsequent anonymisation issues and an unfamiliarity with Bayesian software and (iii) concerns about a lack of acceptance due to the perceived negative views of Bayesian methods.

Finding: We found that although trialists want to make more use of existing data, their biggest concern regarding formally incorporating existing data in a Bayesian framework was their trust and relevance in/of external data; and how that could potentially impact their own trial. We also found that trialists did not feel confident in Bayesian methods and there were practical issues (hard to access data, anonymisation issues, Bayesian software).

Since RCTs can have methodological limitations [21, 63], a trialist may want to adjust the treatment effect estimate in a new study for potential bias to assess the sensitivity of their findings. A Bayesian approach can easily be used to adjust a treatment effect estimate suspected to be biased, by forming a prior distribution for the bias. In **Chapters 4 and 5**, we have developed meta-epidemiological models which can be used to inform this prior distribution on the bias parameter. Previous meta-epidemiological studies have, to our knowledge, only looked at the impact of double-blinding. However, this is an ambiguous term which does not describe precisely what party is blinded to the intervention. Therefore, in **Chapter 4**, a new meta-epidemiological study was conducted, the first to separate performance and detection bias. We found no evidence of an average difference in estimated treatment effect between randomised clinical trials with blinded and non-blinded patients, between trials with blinded and non-blinded healthcare providers, and between trials with blinded and non-blinded outcome assessors, which was surprising. We also extended current statistical methodology to (i) combine binary and continuous outcomes and (ii) use meta-analysis covariates in order to produce stratified or tailored analyses by those groups.

In **Chapter 5**, we extended current meta-epidemiological methods further, which had so far only used a binary study characteristic, looking at the average difference in treatment effect estimates between high versus low RoB trials. The unclear RoB trials are usually grouped with the high RoB trials, or separate analyses of high vs low and unclear vs low RoB contrasts are conducted. However, a trialist will likely know whether their trial was at a high or low RoB, meaning the outputs from an analysis in which high and unclear RoB trials are grouped together would not be relevant and separate analyses often reduce the sample size. We therefore developed a bivariate model and a probability model to include all the data and assess how likely an unclear RoB trial was to be indeed high risk. However, we saw no gain in precision for the average bias estimates in this case study, given the extra computational time. In the probability model we found the estimate of high vs low RoB was *slightly* more precisely estimated than when only the informative high vs low RoB trials was analysed univariately. We extended the probability model further by using the sample size to predict how likely an unclear risk trial would be high risk.

Finding: We developed novel meta-epidemiological methods to address the issue of dichotomising study characteristics in to high and low RoB classifications. This would allow future researchers to use the extended versions of the Welton *et al* models.

Previous work also hypothesised that smaller studies were more likely to estimate larger treatment effects than bigger studies. Rhodes *et al* found evidence that trials with sample sizes less than 100 tended to estimate larger treatment effects on average than trials with sample sizes greater than 100 [140]. We extended the current Welton *et al* model to treat sample size as a continuous characteristic rather than dichotomising it. We have provided the analogous model with a continuous study characteristic to the Welton *et al* binary study characteristic model which allows for the estimated treatment effect to vary with, for example, sample size. Results from the model can be used to predict what value a new study needs to have of this variable to provide an unbiased estimate of the treatment effect. To look at the association between treatment effects and sample size, and treating sample size as continuous, we found the best fitting model modelled the intervention effect as a linear function of $\frac{1}{n}$.

In our qualitative study we found trialists were wary about bias adjustment. Most were not aware of any methods for bias adjustment whilst some did not even think about it.

Qualitative finding: Trialists were hesitant about bias adjustment. Most were not aware of any methods for bias adjustment and were wary of the relevance of other trials to their own.

In **Chapter 6**, existing methods were compared for using a meta-analysis to inform a sample size or power calculation. These methods vary in terms of their assumptions and whether the focus of inference is the new trial or an updated meta-analysis. We found that when an existing meta-analysis has moderate heterogeneity, a new trial had practically no ability to impact upon the estimated mean in a random effects meta-analysis. When powering the new trial to update the existing meta-analysis it is likely a much larger sample size will be needed. This was the same for the 'expected' and 'conditional' frameworks. In our case study, calculations based on the conditional power of a Bayesian analysis of the new trial, using an informative prior distribution, gave lower power than classical power calculations. This was likely driven by the relationship between the MCID and the prior distribution based on the meta-analysis: if the prior distribution is closer to the null compared to the MCID, then the power of a Bayesian analysis of the new trial, incorporating the prior, may be less than classical power. In this case study, it is likely that there will be little gain (reduction in required sample size), from using the meta-analysis as a prior distribution for a new trial. The conditional power of the new trial can be seen as a natural extension of interim analyses [13] with design considerations the same as when conducting a conventional power calculation and may be more applicable than the expected power approach.

Finding: We provided a comparison of methods which explicitly incorporate information from an existing meta-analysis into power or sample size calculations, allowing a trialist to see what size a new trial needs to be to impact the current evidence base.

In our qualitative study, we found that the concept of using an existing meta-analysis to power a new trial based on its ability to impact an existing meta-analysis was unfamiliar to all participants. Having briefly explained to participants that it is possible to power a new trial in this way, the majority of participants saw it as an advantage and many thought it was a very attractive idea and could make the trial more efficient.

Qualitative finding: Although trialists do not think about how their trial results will impact a future meta-analysis, many recognised it was the potential meta-analysis that would change practice.

In **Chapter 7**, we proposed pooling information from previous trials to assess 'normal' rates of adverse events, in situations where the sample size is small, or events are rare. We used pooled placebo data from Novartis sponsored FIH studies to obtain individual subject predictions of how likely a safety event would occur, based on participant demographic and study characteristics. A multilevel logistic regression model was fit and then used to predict how likely an adverse event, in any given patient, was to occur by chance in a new FIH study. This approach had not previously been used in this context and was shown to help the decision-making process.

Finding: Our proposed method to synthesise control arm data from multiple studies allows predictions of how likely an adverse event would occur by chance.

In our qualitative study we found trialists felt a key area where external evidence could be useful was when it was difficult to compare adverse event rates (in particular when the number of events are small) between two groups. Participants also described how they relied on the clinician to pick up any safety concerns, which some viewed as not ideal.

Qualitative finding: Trialists felt they could be making more use of existing data to inform adverse event rates rather than relying on clinicians to determine if there are any potential safety concerns.

8.2 Implications/recommendations

Throughout this thesis, an important consideration was how likely these methods would be used in practice and therefore questions regarding this were asked throughout our qualitative study. We asked trialists what the key things were that needed to change for these methods to be used in practice, with specific attention to our three case studies. In the remainder of this final chapter we build upon our key findings, with specific applications for their use in practice. We then make suggestions for potential future research to address the remaining questions unanswered in the thesis.

Extensions of meta-epidemiological models can increase the scope of their use to inform bias parameters

Given that a key finding was that people were wary about bias adjustment, an implication is that models and research in meta-epidemiological methods should work towards tailored estimates. Whilst bias adjustment has been proposed in a meta-analysis context [55], trialists can in principle adjust for biases in the same way in an RCT context. In Chapter 4, we demonstrated how covariates (e.g. level of subjectivity) can be included in meta-epidemiological research to provide such tailored estimates, which are potentially more acceptable to trialists.

Similarly, by separating the impact that unclear and high RoB trials have in meta-epidemiological models in the bivariate and probability models, this allows the interpretation of such models to be more applicable to trialists for bias adjustment. When trialists are concerned their trial is at a high RoB, they can use the outputs from these extended versions of the Welton *et al* model to correct and down weight the results of trials at high RoB of a new RCT.

Implication: Using our extensions to the Welton *et al* model, researchers could assess how their conclusions change when they down weight their trial's results based on meta-epidemiological evidence.

Trialists could consider how a meta-analysis can be used to inform power or sample size calculation, dependent upon their perspective

When designing a new trial, determining the sample size needed is a crucial part of the design stage. It is at this point the trials team have already gathered their external evidence as part of the case to justify a gap in the evidence base or help determine which outcomes have previously been used. Lau *et al* [6] have suggested this can be used to make additional inferences in power, or similarly, sample size calculations. Given that a key finding in our qualitative study was that people do not usually think of the impact their trial has on the overall evidence base, but are potentially interested, trialists should consider some of the methods available, as described in Chapter 6.

Each of the methods compared makes different assumptions and, as discussed in Section 6.6.2, the choice of method should depend on how the study results may be used and interpreted: (i) in isolation, (ii) in terms of its impact upon an updated meta-analysis, or (iii) in terms of its impact upon a cost effectiveness analysis.

Implication: When a relevant meta-analysis exists, the impact of a new trial could be assessed, even as a sensitivity analysis. The trialist will need to assess the relevance of information in the meta-analysis to the trial of interest and decide upon primary use of the new trial.

In cases where a relevant meta-analysis has been used to justify a new trial, conditional and expected power calculations could be used alongside a traditional power calculation, as part of a sensitivity analysis. We therefore suggest that education or a greater awareness is needed to highlight that this is one path which allows changes to be made to existing policies and treatments adopted, and can be, at the very least, considered in the design stage of a new RCT.

The largest barriers to the use of incorporating external evidence identified by both our qualitative study and the INVEST survey were time constraints and the need to use specialist Bayesian software, such as WinBUGS. The work conducted in Chapter 6 was based on normal approximations and therefore closed form solutions which are implementable by hand and consequently easy to code in standard statistical software.

External evidence could be used to make evidence-based predictions of adverse events

Methods were explored in Chapter 7 for the analysis of FIH studies, which are small in sample size and therefore underpowered to detect rare events, meaning the response rate in the control arm is inadequate. It can therefore be difficult for a trialist to statistically compare two groups and make any quantitative inferences. In turn, this means that clinicians are heavily relied upon for potential rates of adverse events in particular populations. This is consistent with our qualitative study: many of the trialists interviewed stated that there was often no quantitative evidence behind adverse event rates, and whether an observed rate was ′normal′ was assessed fairly subjectively, with quite often the Chief Investigator making the final decision.

In Chapter 7, it was shown how a synthesis of placebo arm data from multiple studies can be used to make predictions on how likely an individual with particular characteristics (model-based covariates) would have an adverse event by chance in the placebo arm and this was compared with their observed event rate in the experimental arm. If their model based prediction was low but the patient did have the event, this would raise safety concerns to the clinical team. Therefore, given that trialists felt they could be making more use of external data, this method may generally be more acceptable for use in practice.

Although there have been calls for a global database and emerging guidance to help enable data sharing, it is likely this is still some time away. To bridge this gap, trials units could collate databases of their own trials. It is likely that most trials units specialise in certain disease areas, such as cardiac trials, and so other trials ran by the same trials unit may have similar populations to the control arm of their new trial. We suggest using in-house databases of individual patient data and filtering by similar populations i.e. those with similar baseline demographics to see what the rates of these expected events have been and see if this matches with what the clinician might expect. An extension to this would also be to use individual patient characteristics to predict how likely each person was to experience an event based on the information in the database, as demonstrated in Chapter 7.

Implication: To make more use of existing data to inform adverse event rates, individual trials units could set up a platform or database for their own trials which could be used to make inferences about adverse event rates for new trials. These trials are likely to be more similar in terms of their inclusion criteria to the new trial.

8.3 External evidence in a *routine* clinical trial setting: does it have a place?

> "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."
>
> *Ronald Fisher 1890-1962*

From our INVEST survey and qualitative study we have seen that Bayesian methods are rarely used in clinical trials. We also know evidence synthesis methods are more likely to be conducted to inform aspects of the design, rather than in the analysis stage. For example, in our INVEST survey and qualitative study, we found that a systematic review or meta-analysis is often used by trialists in numerous ways to inform aspects of trial design. Trialists also felt they could be making more use of evidence synthesis methods but because of time pressures and relevance of external data to their trial were not. In this thesis we have shown three areas in which trialists could consider explicitly incorporating external evidence.

The potential advantages of using informative prior distributions need to be counter balanced by whether the clinical trials community will ever accept the formal incorporation of, what is likely to be, someone else's data. Even if the prior information is based on previous 'similar' trials there are still different ways in which this information could be used to form a prior distribution, depending on various assumptions regarding the relationship between the previous trials and the new trial. The relevance of the prior data will need to be carefully assessed by individuals who are closely related to the research question of the new trial. There are two disadvantages to a Bayesian analysis of a trial. First, it is generally more complex than a classical analysis due to (usually) specialist software being required. Second, informative priors are subjective: there are multiple ways in which the same external information could be used to define an informative prior distribution for a parameter. As seen in our qualitative study, this is a source of concern to trialists.

The methods described for FIH studies in Chapter 7 could in principle be applied to Phase III trials, but we suggest that formal incorporation of external information on adverse event rates would generally be considered less acceptable in these trials. However, external information could be used in two less formal ways. First, as seen in the literature (Section 1.4), external data could be used to inform the trialist of the expected population rate of particular outcomes in the control population. Second, where IPD are accessible, for example, collated data within a trials unit, trialists may refine the population estimate by deriving individual expected predictions of an event (using a model approach as shown in Chapter 7). The trialist would then be better equipped to interpret the observed adverse event rate.

8.4 Future research

This thesis offers contributions to the Bayesian analysis of clinical trials and consequently the use of informative priors. However, there are several aspects that remain unaddressed and these are presented in Figure 8.1.

Given a key finding in our qualitative study was a general lack of confidence in and negative perception of Bayesian methods, this is something which will need to be addressed in future work. This may include more case studies exploring the ways in which incorporation of external data might be advantageous, together with clear methodological guidelines. Additionally, training courses in Bayesian methods, focusing on potential uses in trials, could promote use and increase confidence. The remaining barriers and additional future work regarding statistical methodology are discussed below.

Our MetaBLIND study in Chapter 4, which separates the impact of not blinding patients, healthcare providers and outcome assessors, gave surprising but inconclusive results. As discussed in Section 4.6.2, a potential reason for this is the sample of trials analysed. Further work could try and replicate this result in other samples [199, 200].

In Chapter 5, we briefly discussed the possible ways of forming a prior distribution on the bias parameter based on the outputs from meta-epidemiological models (Section 5.2.4). Future work should be conducted to explore the imapct of bias-adjustment based on these priors to trials in practice. We also recognise that methodological guidelines are needed to increase the uptake of bias adjustment in trials. This could include a tutorial paper in a general medical journal on bias adjustment and the role of meta-epidemiological methods.

In Chapter 6, we compared methods to see what size a new trial needs to be when inference is based on the new clinical trial, the updated random effects mean or on the updated cost-effectiveness model. Given we only looked at one case study, properties of these various methods should be investigated through more case studies or a simulation study.

In Chapter 7, we first modelled external data about adverse event rates in the placebo population of FIH studies. Second, due to the small sample sizes in FIH studies (making it difficult for a clinician to determine if a new drug was harmful), we used the model to predict how likely an individual in a new study would have an event. Future work should consider how best to analyse a new FIH study, formally accounting for this external evidence. This could include matching an individual's prediction to their observed data.

Figure 8.1: Key future research needs.

- Our MetaBLIND study had potentially lower power than we anticipated, as we had to disregard lots of meta-analyses which were non-informative. A future meta-epidemiology study investigating the impact of separate types of blinding should be conducted, in order to replicate our findings.
- Based on our findings in the INVEST survey and the qualitative study, further guidance is required on bias adjustment. This could include a general medical journal on bias adjustment and the role of meta-epidemiological methods in this.
- Given we only looked at one case study, properties of the various methods that use information from a meta-analysis in sample size calculations should be investigated.
- Further work is required to explore how best to incorporate external evidence on adverse event rates in trials, given the limited number of controls, including how to match individual model based predictions with event data.

8.5 Final conclusions

In conclusion, there are many arguments *both* for and against the use of Bayesian methods and more specifically informative priors, based on the current evidence base, in the design and analysis of clinical trials. The constant push to make more use of existing data by methodologists versus the practical issues and hesitation of doing so in a clinical trials setting has long been debated. Almost everyone agrees that existing evidence should be used, in some way, to inform the design and analysis of a clinical trial. This can range from using it to justify a gap in the evidence base (such as using a systematic review to show there is an unanswered clinical question) or using it to inform parameters in the sample size calculations such as the expected control group event rate. However, at present trialists rarely explicitly, statistically combine this previous evidence with their actual trial.

The work in this thesis has explored in depth the concerns about and barriers to the use of Bayesian methods in practice, by interviewing the trialists who have the ability to implement such methods in practice in order to better understand their views. We found that although trialists want to make more use of existing data, their biggest concern regarding formally incorporating existing data in a Bayesian framework was their trust in and perceived lack of relevance of external data; and how that could potentially impact their own trial.

In this thesis we explored how existing data could be synthesised and translated into prior information to inform aspects of the design or analysis of a new clinical trial. Trials are subject to biases, despite best efforts. We have shown how information from meta-epidemiological evidence can be used to quantify the likely amount of bias in the treatment effect of a trial, allowing the analyst to assess the sensitivity of the findings. Incorporating information from an existing meta-analysis into sample size calculations could lead to appropriately reduced sample sizes in some situations, but this depends on how the trial evidence will be used in practice. In cases where the number of patients or events in the control arm is small, individual patient adverse event rate data can be synthesised to help the clinical team quantify the expected event rate in the control group.

Trials are a huge burden to society; costing a huge amount of private and public money, as well as the time and resources by medical researchers and patients [183]. There must be a focus on making the incorporation of evidence implementable in a trial setting by methodologists who advocate such methods. Unless the necessary guidance is provided or there is a requirement by funders, it is likely the prevalence of use of informative priors will remain low in practice. Despite the challenges regarding applying Bayesian methods in trials, this thesis has shown that it is possible for trialists to make inferences based on all available evidence, or to assess how their conclusions might change if they incorporated relevant external evidence.

References

[1] C. Young and R. Horton. Putting clinical trials into context. *Lancet*, 366(9480):107–108, 2005. ISSN 0140-6736. doi: 10.1016/s0140-6736(05)66846-8.

[2] J. A. Berlin and R. M. Golub. Meta-analysis as evidence building a better pyramid. *Jama-Journal of the American Medical Association*, 312(6):603–605, 2014. ISSN 0098-7484. doi: 10.1001/jama.2014.8167.

[3] M. Nasser, M. Clarke, I. Chalmers, K. G. Brurberg, H. Nykvist, H. Lund, and P. Glasziou. What are funders doing to minimise waste in research? *Lancet (London, England)*, 389(10073):1006–1007, 2017. doi: 10.1016/s0140-6736(17)30657-8.

[4] R. Martina, D. Jenkins, S. Bujkiewicz, P. Dequen, K. Abrams, and W. Getreal. The inclusion of real world evidence in clinical development planning. *Trials*, 19, 2018. ISSN 1745-6215. doi: 10.1186/s13063-018-2769-2.

[5] M. L. Ferreira, R. D. Herbert, M. J. Crowther, A. Verhagen, and A. J. Sutton. When is a further clinical trial justified? *British Medical Journal*, 345, 2012. ISSN 1756-1833. doi: 10.1136/bmj.e5913.

[6] J. Lau, C. H. Schmid, and T. C. Chalmers. Cumulative metaanalysis of clinical-trials - builds evidence for exemplary medical-care. *Journal of Clinical Epidemiology*, 48(1): 45–57, 1995. ISSN 0895-4356. doi: 10.1016/0895-4356(94)00106-z.

[7] M. Clarke. Doing new research? don't forget the old - nobody should do a trial without reviewing what is known. *Plos Medicine*, 1(2):100–102, 2004. ISSN 1549-1277. doi: 10.1371/journal.pmed.0010035.

[8] A. J. Sutton, N. J. Cooper, and D. R. Jones. Evidence synthesis as the key to more coherent and efficient research. *Bmc Medical Research Methodology*, 9, 2009. ISSN 1471-2288. doi: 10.1186/1471-2288-9-29.

[9] I. Chalmers and P. Glasziou. Avoidable waste in the production and reporting of research evidence. *Lancet*, 374(9683):86–89, 2009. ISSN 0140-6736. doi: 10.1016/s0140-6736(09)60329-9.

[10] N. J. Cooper, D. R. Jones, and A. J. Sutton. The use of systematic reviews when designing studies. *Clinical Trials*, 2(3):260–264, 2005. ISSN 1740-7745. doi: 10.1191/1740774505cn090oa.

[11] A. P. Jones, E. Conroy, P. R. Williamson, M. Clarke, and C. Gamble. The use of systematic reviews in the planning, design and conduct of randomised trials: a retrospective cohort of NIHR HTA funded trials. *Bmc Medical Research Methodology*, 13, 2013. ISSN 1471-2288. doi: 10.1186/1471-2288-13-50.

[12] M. Clarke, S. Hopewell, and I. Chalmers. Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: a status report. *Journal of the Royal Society of Medicine*, 100(4):187–190, 2007. ISSN 0141-0768. doi: 10.1258/jrsm.100.4.187.

[13] D. J. Spiegelhalter. Incorporating bayesian ideas into health-care evaluation. *Statistical Science*, 19(1):156–174, 2004. ISSN 0883-4237. doi: 10.1214/088342304000000080.

[14] D. A. Berry. Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1):27–36, 2006. ISSN 1474-1776. doi: 10.1038/nrd1927.

[15] R. G. G. Russell, N. B. Watts, F. H. Ebetino, and M. J. Rogers. Mechanisms of action of bisphosphonates: Similarities and differences and their potential influence on clinical efficacy. *Osteoporosis International*, 19(6):733–759, 2008. ISSN 0937-941X. doi: 10.1007/s00198-007-0540-8.

[16] A. C. Leon, L. L. Davis, and H. C. Kraemer. The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research*, 45(5):626–629, 2011. ISSN 0022-3956. doi: 10.1016/j.jpsychires.2010.10.008.

[17] N. Feeley, S. Cossette, J. Cote, M. Heon, R. Stremler, G. Martorella, and M. Purden. The importance of piloting an rct intervention. *The Canadian journal of nursing research = Revue canadienne de recherche en sciences infirmieres*, 41(2):85–99, 2009. ISSN 0844-5621.

[18] N. S. Blencowe, J. M. Blazeby, S. Strong, A. Torrance, T. D. Pinkney, and G. Clayton et al. Feasibility work to inform the design of a randomized clinical trial of wound dressings in elective and unplanned abdominal surgery. *British Journal of Surgery*, 103(12):1738–1744, 2016. ISSN 0007-1323. doi: 10.1002/bjs.10274.

[19] B. C. Reeves, L. Andronis, J. M. Blazeby, N. S. Blencowe, M. Calvert, and J. Coast et al. A mixed-methods feasibility and external pilot study to inform a large pragmatic randomised controlled trial of the effects of surgical wound dressing strategies on surgical site infections (bluebelle phase b): Study protocol for a randomised controlled trial. *Trials*, 18, 2017. ISSN 1745-6215. doi: 10.1186/s13063-017-2102-5.

[20] K. F. Schulz, I. Chalmers, R. J. Hayes, and D. G. Altman. Empirical-evidence of bias - dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Jama-Journal of the American Medical Association*, 273(5):408–412, 1995. ISSN 0098-7484. doi: 10.1001/jama.273.5.408.

[21] J. P. T. Higgins, D. G. Altman, P. C. Gotzsche, P. Juni, and D. Moher et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, 343, 2011. ISSN 1756-1833. doi: 10.1136/bmj.d5928.

[22] L. C. Gurrin, J. J. Kurinczuk, and P. R. Burton. Bayesian statistics in medical research: an intuitive alternative to conventional data analysis. *Journal of Evaluation in Clinical Practice*, 6(2):193–204, 2000. ISSN 1356-1294. doi: 10.1046/j.1365-2753.2000.00216.x.

[23] M. M. Bennett, B. J. Crowe, K. L. Price, J. D. Stamey, and J. W. Seaman. Comparison of bayesian and frequentist meta-analytical approaches for analyzing time to event data. *Journal of Biopharmaceutical Statistics*, 23(1):129–145, 2013. ISSN 1054-3406. doi: 10.1080/10543406.2013.737210.

[24] D. A. Berry. Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science*, 19(1):175–187, 2004. ISSN 0883-4237. doi: 10.1214/088342304000000044.

[25] S. Greenland. Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International Journal of Epidemiology*, 35(3):765–775, 2006. ISSN 0300-5771. doi: 10.1093/ije.dyi312.

[26] S. Greenland. Bayesian perspectives for epidemiological research. ii. regression analysis. *International Journal of Epidemiology*, 36(1):195–202, 2007. ISSN 0300-5771. doi: 10.1093/ije/dyl289.

[27] J. R. Carpenter. Commentary: On Bayesian perspectives for epidemiological re-

search. *International Journal of Epidemiology*, 35(3):775–777, 2006. ISSN 0300-5771. doi: 10.1093/ije/dyl055.

[28] A. Racine, A. P. Grieve, H. Fluhler, and A. F. M. Smith. Bayesian methods in practice - experiences in the pharmaceutical-industry. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 35(2):93–150, 1986. ISSN 0035-9254. doi: 10.2307/2347264.

[29] J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference Part I. *Biometrika*, 20A:175–240, 1928. ISSN 0006-3444. doi: 10.2307/2331945.

[30] P. Brocklehurst, D. Elbourne, and Z. Alfirevic. Role of external evidence in monitoring clinical trials: experience from a perinatal trial. *British Medical Journal*, 320(7240): 995–998, 2000. ISSN 0959-8138. doi: 10.1136/bmj.320.7240.995.

[31] R. DerSimonian. Meta-analysis in the design and monitoring of clinical trials. *Statistics in Medicine*, 15(12):1237–1248, 1996. ISSN 0277-6715. doi: 10.1002/(sici)1097-0258(19960630)15:12<1237::aid-sim301>3.0.co;2-n.

[32] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS - a Bayesian Modelling Framework: Concepts, Structure, and Extensibility. *Statistics and Computing*, 10(4):325–337, 2000. ISSN 0960-3174. doi: 10.1023/a:1008929526011.

[33] L. Bero and D. Rennie. The Cochrane Collaboration - preparing, maintaining, and disseminating systematic reviews of the effects of health-care. *Jama-Journal of the American Medical Association*, 274(24):1935–1938, 1995. ISSN 0098-7484. doi: 10.1001/jama.274.24.1935.

[34] K. Abrams and D. R. Jones. Meta-analysis and the synthesis of evidence. *Ima Journal of Mathematics Applied in Medicine and Biology*, 12(3-4):297–313, 1995. ISSN 0265-0746.

[35] D. Moher, D. J. Cook, S. Eastwood, I. Olkin, D. Rennie, D. F. Stroup, and Q. Grp. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet*, 354(9193):1896–1900, 1999. ISSN 0140-6736. doi: 10.1016/s0140-6736(99)04149-5.

[36] S. L. T. Normand. Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18(3):321–359, 1999. ISSN 0277-6715. doi: 10.1002/(sici)1097-0258(19990215)18:3<321::aid-sim28>3.3.co;2-g.

[37] L. V. Hedges and T. D. Pigott. The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, 9(4):426–445, 2004. ISSN 1082-989X. doi: 10.1037/1082-989x.9.4.426.

[38] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010. ISSN 1759-2879. doi: 10.1002/jrsm.12.

[39] K. Rice, J. P. T. Higgins, and T. Lumley. A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 181(1): 205–227, 2018. ISSN 0964-1998. doi: 10.1111/rssa.12275.

[40] A. Deaton and N. Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine (1982)*, 2017. doi: 10.1016/j.socscimed.2017.12.005.

[41] R. D. Riley, J. P. T. Higgins, and J. J. Deeks. Interpretation of random effects meta-analyses. *Bmj-British Medical Journal*, 342, 2011. ISSN 1756-1833. doi: 10.1136/bmj.d549.

[42] J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter. A re-evaluation of random-

effects meta-analysis. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 172:137–159, 2009. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2008.00552.x.

[43] C. S. Berkey, D. C. Hoaglin, F. Mosteller, and G. A. Colditz. A random-effects regression-model for metaanalysis. *Statistics in Medicine*, 14(4):395–411, 1995. ISSN 0277-6715. doi: 10.1002/sim.4780140406.

[44] J. P. T. Higgins. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, 37(5):1158–1160, 2008. ISSN 0300-5771. doi: 10.1093/ije/dyn204.

[45] J. P. T. Higgins and S. G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002. ISSN 0277-6715. doi: 10.1002/sim.1186.

[46] A. J. Sutton and J. P. T. Higgins. Recent developments in meta-analysis. *Statistics in Medicine*, 27(5):625–650, 2008. ISSN 0277-6715. doi: 10.1002/sim.2934.

[47] R. M. Turner, D. Jackson, Y. H. Wei, S. G. Thompson, and J. P. T. Higgins. Predictive distributions for between-study heterogeneity and simple methods for their application in bayesian meta-analysis. *Statistics in Medicine*, 34(6):984–998, 2015. ISSN 0277-6715. doi: 10.1002/sim.6381.

[48] T. C. Smith, D. J. Spiegelhalter, and A. Thomas. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, 14(24):2685–2699, 1995. ISSN 0277-6715. doi: 10.1002/sim.4780142408.

[49] J. P. Jansen, R. Fleurence, B. Devine, R. Itzler, A. Barrett, N. Hawkins, K. Lee, C. Boersma, L. Annemans, and J. C. Cappelleri. Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research

Practices: Part 1. *Value in Health*, 14(4):417–428, 2011. ISSN 1098-3015. doi: 10.1016/j.jval.2011.04.002.

[50] D. M. Caldwell, A. E. Ades, and J. P. T. Higgins. Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *British Medical Journal*, 331 (7521):897–900, 2005. ISSN 0959-8146. doi: 10.1136/bmj.331.7521.897.

[51] A. H. Briggs. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics*, 17(5):479–500, 2000. ISSN 1170-7690. doi: 10.2165/00019053-200017050-00006.

[52] A. E. Ades, G. Lu, and K. Claxton. Expected value of sample information calculations in medical decision modeling. *Medical Decision Making*, 24(2):207–227, 2004. ISSN 0272-989X. doi: 10.1177/0272989x04263162.

[53] J. Bindels, B. Ramaekers, I. C. Ramos, L. Mohseninejad, J. Grutters S. Knies, M. Postma, T. Feenstra M. Al, and M. Joore. Use of value of information in healthcare decision making: Exploring multiple perspectives. *Pharmacoeconomics*, 34(3): 315–322, 2016. ISSN 1170-7690. doi: 10.1007/s40273-015-0346-z.

[54] S. D. Ramsey, R. J. Willke, H. Glick, S. D. Reed, F. Augustovski, B. Jonsson, A. Briggs, and S. A. Sullivan. Cost-Effectiveness Analysis Alongside Clinical Trials II-An ISPOR Good Research Practices Task Force Report. *Value in Health*, 18(2):161–172, 2015. ISSN 1098-3015. doi: 10.1016/j.jval.2015.07.001.

[55] N. J. Welton, A. E. Ades, J. B. Carlin, D. G. Altman, and J. A. C. Sterne. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 172:119–136, 2009. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2008.00548.x.

[56] I. Chalmers. Randomized controlled trials of aprotinin in cardiac surgery: could

clinical equipoise have stopped the bleeding? comment. *Clinical Trials*, 2(3):229–231, 2005. ISSN 1740-7745.

[57] A. Bhurke, A. Cook, A. Tallant, E. Williams A. Young, and J. Raftery. Using systematic reviews to inform NIHR HTA trial planning and design: a retrospective cohort. *Bmc Medical Research Methodology*, 15, 2015. ISSN 1471-2288. doi: 10.1186/s12874-015-0102-2.

[58] D. L. Burke, L. J. Billingham, A. J. Girling, and R. D. Riley. Meta-analysis of randomized phase II trials to inform subsequent phase III decisions. *Trials*, 15, 2014. ISSN 1745-6215. doi: 10.1186/1745-6215-15-346.

[59] J. F. Tierney, J. P. Pignon, F. Gueffyier, M. Clarke, L. Askie, C. L. Vale, S. Burdett, and I. P. D. M. a. M. G. Cochrane. How individual participant data meta-analyses have influenced trial design, conduct, and analysis. *Journal of Clinical Epidemiology*, 68(11):1325–1335, 2015. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2015.05.024.

[60] G. Salanti, J. P. T. Higgins, A. E. Ades, and J. P. A. Ioannidis. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research*, 17(3):279–301, 2008. ISSN 0962-2802. doi: 10.1177/0962280207080643.

[61] R. M. Turner, S. G. Thompson, and D. I. Spiegelhalter. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials*, 2(2):108–118, 2005. ISSN 1740-7745. doi: 10.1191/1740774505cn072oa.

[62] R. Peto, J. Emberson, M. Landray, C. Baigent, R. Collins, R. Clare, and R. Califf. Analyses of cancer data from three Ezetimibe trials. *New England Journal of Medicine*, 359(13):1357–1366, 2008. ISSN 0028-4793. doi: 10.1056/NEJMsa0806603.

[63] L. L. Gluud. Bias in clinical intervention research. *American Journal of Epidemiology*, 163(6):493–501, 2006. ISSN 0002-9262. doi: 10.1093/aje/kwj069.

[64] J. Savovic, H. E. Jones, D. G. Altman, R. J. Harris, P. Juni, J. Pildal, B. Als-Nielsen, E. M. Balk, C. Gluud, L. L. Gluud, J. P. A. Ioannidis, K. F. Schulz, R. Beynon, N. J. Welton, L. Wood, D. Moher, J. J. Deeks, and J. A. C. Sterne. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Annals of Internal Medicine*, 157(6):429–U97, 2012. ISSN 0003-4819. doi: 10.7326/0003-4819-157-6-201209180-00537.

[65] M. J. Page, J. P. T. Higgins, G. Clayton, J. A. C. Sterne, A. Hrobjartsson, and J. Savovic. Empirical evidence of study design biases in randomized trials: Systematic review of meta-epidemiological studies. *Plos One*, 11(7), 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0159267.

[66] R. M. Turner, M. Lloyd-Jones, D. O. C. Anumba, G. C. S. Smith, D. J. Spiegelhalter, H. Squires, J. W. Stevens, M. J. Sweeting, S. J. Urbaniak, R. Webster, and S. G. Thompson. Routine Antenatal Anti-D Prophylaxis in Women Who Are Rh(D) Negative: Meta-Analyses Adjusted for Differences in Study Design and Quality. *Plos One*, 7(2), 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0030711.

[67] S. J. Pocock. Combination of randomized and historical controls in clinical-trials. *Journal of Chronic Diseases*, 29(3):175–188, 1976. ISSN 0021-9681. doi: 10.1016/0021-9681(76)90044-8.

[68] G. B. Lu and A. Ades. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805, 2009. ISSN 1465-4644. doi: 10.1093/biostatistics/kxp032.

[69] X. Zhou, S. Y. Liu, E. S. Kim, R. S. Herbst, and J. L. Lee. Bayesian adaptive design for

targeted therapy development in lung cancer - a step toward personalized medicine. *Clinical Trials*, 5(3):181–193, 2008. ISSN 1740-7745. doi: 10.1177/1740774508091815.

[70] J. P. T. Higgins and A. Whitehead. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*, 15(24):2733–2749, 1996. ISSN 0277-6715. doi: 10.1002/(sici)1097-0258(19961230)15:24<2733::aid-sim562>3.0.co;2-0.

[71] L. A. Moye. Bayesians in clinical trials: Asleep at the switch. *Statistics in Medicine*, 27(4):469–482, 2008. ISSN 0277-6715. doi: 10.1002/sim.2928.

[72] G. D. Smith and M. Egger. Commentary: Incommunicable knowledge? Interpreting and applying the results of clinical trials and meta-analyses. *Journal of Clinical Epidemiology*, 51(4):289–295, 1998. ISSN 0895-4356. doi: 10.1016/s0895-4356(97)00293-x.

[73] V. Roloff, J. P. T. Higgins, and A. J. Sutton. Planning future studies based on the conditional power of a meta-analysis. *Statistics in Medicine*, 32(1):11–24, 2013. ISSN 0277-6715. doi: 10.1002/sim.5524.

[74] A. J. Sutton, N. J. Cooper, D. R. Jones, P. C. Lambert, J. R. Thompson, and K. R. Abrams. Evidence-based sample size calculations based upon updated meta-analysis. *Statistics in Medicine*, 26(12):2479–2500, 2007. ISSN 0277-6715. doi: 10.1002/sim.2704.

[75] C. Pope and N. Mays. Reaching the parts other methods cannot reach - an introduction to qualitative methods in health and health-services research. *British Medical Journal*, 311(6996):42–45, 1995. ISSN 0959-8138.

[76] V. Clarke and V. Braun. Teaching thematic analysis. *Psychologist*, 26(2):120–123, 2013. ISSN 0952-8229.

[77] A. Kuper, S. Reeves, and W. Levinson. Qualitative research - an introduction to

reading and appraising qualitative research. *Bmj-British Medical Journal*, 337(7666), 2008. ISSN 1756-1833. doi: 10.1136/bmj.a288.

[78] G. L. Clayton, I. L. Smith, J. P. T. Higgins, B. Mihaylova, B. Thorpe, R. Cicero, K. Lokuge, J. R. Forman, J. F. Tierney, I. R. White, L. D. Sharples, and H. E. Jones. The invest project: investigating the use of evidence synthesis in the design and analysis of clinical trials. *Trials*, 18, 2017. ISSN 1745-6215. doi: 10.1186/s13063-017-1955-y.

[79] M. T. Haahr and A. Hrobjartsson. Who is blinded in randomized clinical trials? a study of 200 trials and a survey of authors. *Clinical Trials*, 3(4):360–365, 2006. ISSN 1740-7745. doi: 10.1177/1740774506069153.

[80] J. Wolf, A. Pruess-Ustuen, O. Cumming, J. Bartram, S. Bonjour, S. Cairncross, T. Clasen, Jr. J. M. Colford, V. Curtis, J. De France, L. Fewtrell, M. C. Freeman, B. Gordon, P. R. Hunter, A. Jeandron, R. B. Johnston, D. Maeusezahl, C. Mathers, M. Neira, and J. P. T. Higgins. Assessing the impact of drinking water and sanitation on diarrhoeal disease in low- and middle-income settings: systematic review and meta-regression. *Tropical Medicine & International Health*, 19(8):928–942, 2014. ISSN 1360-2276. doi: 10.1111/tmi.12331.

[81] J. Savovic, R. Turner, D. Mawdsley, J. Higgins, and J. Sterne. A new large-scale meta-epidemiological study on bias in randomized trials using routinely collected risk-of-bias assessments by cochrane reviewers: results from the robes study. *Trials*, 16, 2015. ISSN 1745-6215.

[82] S. Dias, N. J. Welton, V. C. C. Marinho, G. Salanti, J. P. T. Higgins, and A. E. Ades. Estimation and adjustment of bias in randomized evidence by using mixed treatment comparison meta-analysis. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 173:613–629, 2010. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2010.00639.x.

[83] P. Edwards, M. Arango, L. Balica, R. Cottingham, and H. El-Sayed et al. Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury - outcomes at 6 months. *Lancet*, 365(9475): 1957–1959, 2005. ISSN 0140-6736.

[84] V. Braet, G. Mazairac, J. Jacques, M. F. Arango, P. Svoboda, and J. Ochmann et al. The MRC CRASH Trial: study design, baseline data, and outcome in 1000 randomised patients in the pilot phase. *Emergency Medicine Journal*, 19(6):510–514, 2002. ISSN 1472-0205.

[85] M. Merz, M. Seiberling, G. Hoxter, M. L. Holting, and H. P. Wortha. Elevation of liver enzymes in multiple dose trials during placebo treatment: Are they predictable. *Journal of Clinical Pharmacology*, 37(9):791–798, 1997. ISSN 0091-2700.

[86] N. J. Welton, A. E. Ades, D. M. Caldwell, and T. J. Peters. Research prioritization based on expected value of partial perfect information: a case-study on interventions to increase uptake of breast cancer screening. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 171:807–834, 2008. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2008.00558.x.

[87] R. M. Harbord, M. Egger, and J. A. C. Sterne. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, 25 (20):3443–3457, 2006. ISSN 0277-6715. doi: 10.1002/sim.2380.

[88] D. J. Spiegelhalter. Bayesian graphical modelling: a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 47:115–133, 1998. ISSN 0035-9254. doi: 10.1111/1467-9876.00101.

[89] R. Ganann, D. Ciliska, and H. Thomas. Expediting systematic reviews: methods and

implications of rapid reviews. *Implementation Science*, 5:10–19, 2010. ISSN 1748-5908. doi: 10.1186/1748-5908-5-56.

[90] S. Khangura, K. Konnyu, R. Cushman, J. Grimshaw, and D. Moher. Evidence summaries: the evolution of a rapid review approach. *Systematic reviews*, 1:10–10, 2012. doi: 10.1186/2046-4053-1-10.

[91] S. Khangura, J. Polisena, T. J. Clifford, K. Farrah, and C. Kamel. Rapid review: An emerging approach to evidence synthesis in health technology assessment. *International Journal of Technology Assessment in Health Care*, 30(1):20–27, 2014. ISSN 0266-4623. doi: 10.1017/s0266462313000664.

[92] M. K. B. Parmar, J. Carpenter, and M. R. Sydes. More multiarm randomised trials of superiority are needed. *Lancet*, 384(9940):283–284, 2014. ISSN 0140-6736.

[93] N. J. Cooper, J. Peters, M. C. W. Lai, P. Juni, S. Wandel, S. Palmer, M. Paulden, S. Conti, N. J. Welton, K. R. Abrams, S. Bujkiewicz, D. Spiegelhalter, and A. J. Sutton. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? *Value in Health*, 14(2):371–380, 2011. ISSN 1098-3015. doi: 10.1016/j.jval.2010.09.001.

[94] K. P. Claxton and M. J. Sculpher. Using value of information analysis to prioritise health research - some lessons from recent uk experience. *Pharmacoeconomics*, 24(11): 1055–1068, 2006. ISSN 1170-7690. doi: 10.2165/00019053-200624110-00003.

[95] N. Mays and C. Pope. Rigour and qualitative research. *British Medical Journal*, 311 (6997):109–112, 1995. ISSN 0959-8138.

[96] N. Mays and C. Pope. Qualitative research in health care - assessing quality in qualitative research. *British Medical Journal*, 320(7226):50–52, 2000. ISSN 0959-8138. doi: 10.1136/bmj.320.7226.50.

[97] R. Fitzpatrick and M. Boulton. Qualitative methods for assessing health care. *Quality in health care : QHC*, 3(2):107–113, 1994. ISSN 0963-8172.

[98] L. Locock and L. Smith. Personal experiences of taking part in clinical trials – a qualitative study. *Patient Education and Counseling*, 84(3):303–309, 2011. ISSN 0738-3991. doi: https://doi.org/10.1016/j.pec.2011.06.002.

[99] B. G. Glaser and A. L. Strauss. *The discovery of grounded theory : strategies for qualitative research*. 1967. ISBN 9780202300283 0202300285 0202302601 9780202302607.

[100] M. D. Hughes. Reporting Bayesian Analyses of Clinical-Trials. *Statistics in Medicine*, 12(18):1651–1663, 1993. ISSN 0277-6715. doi: 10.1002/sim.4780121802.

[101] I. T. Coyne. Sampling in qualitative research. purposeful and theoretical sampling; merging or clear boundaries? *Journal of Advanced Nursing*, 26(3):623–630, 1997. ISSN 0309-2402. doi: 10.1046/j.1365-2648.1997.t01-25-00999.x.

[102] L. A. Goodman. Snowball sampling. *Annals of Mathematical Statistics*, 32(1):148–170, 1961. ISSN 0003-4851. doi: 10.1214/aoms/1177705148.

[103] G. Guest, E. Namey, J. Taylor, N. Eley, and K. McKenna. Comparing focus groups and individual interviews: findings from a randomized study. *International Journal of Social Research Methodology*, 20(6):693–708, 2017. ISSN 1364-5579. doi: 10.1080/13645579.2017.1281601.

[104] S. Thorne. Data analysis in qualitative research. *Evidence Based Nursing*, 3(3):68, 2000. doi: 10.1136/ebn.3.3.68.

[105] H. Boeije. A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and Quantity*, 36(4):391–409, 2002. ISSN 1573-7845. doi: 10.1023/A:1020909529486.

[106] A. J. B. Fugard and H. W. W. Potts. Supporting thinking on sample sizes for thematic analyses: a quantitative tool. *International Journal of Social Research Methodology*, 18 (6):669–684, 2015. ISSN 1364-5579. doi: 10.1080/13645579.2015.1005453.

[107] M. Sandelowski. Sample size in qualitative research. *Research in Nursing & Health*, 18(2):179–183, 1995. ISSN 0160-6891. doi: 10.1002/nur.4770180211.

[108] T. Lorenc, L. Felix, M. Petticrew, G. J. Melendez-Torres, J. Thomas, S. Thomas, A. O'mara-Eves, and M. Richardson. Meta-analysis, complexity, and heterogeneity: a qualitative interview study of researchers' methodological values and practices. *Systematic reviews*, 5(1):192–192, 2016.

[109] C. Snowdon. Qualitative and mixed methods research in trials. *Trials*, 16, 2015. ISSN 1745-6215. doi: 10.1186/s13063-015-1084-4.

[110] L. M. Connelly and J. N. Peltzer. Underdeveloped themes in qualitative research relationship with interviews and analysis. *Clinical Nurse Specialist*, 30(1):51–57, 2016. ISSN 0887-6274. doi: 10.1097/nur.0000000000000173.

[111] M. Sandelowski and J. Leeman. Writing usable qualitative health research findings. *Qualitative Health Research*, 22(10):1404–1413, 2012. ISSN 1049-7323. doi: 10.1177/1049732312450368.

[112] A. Tong, P. Sainsbury, and J. Craig. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6):349–357, 2007. ISSN 1353-4505. doi: 10.1093/intqhc/mzm042.

[113] K. El Emam, S. Rodgers, and B. Malin. Anonymising and sharing individual patient data. *Bmj-British Medical Journal*, 350, 2015. ISSN 1756-1833. doi: 10.1136/bmj.h1139.

[114] K. Spencer, C. Sanders, E. A. Whitley, D. Lund, J. Kaye, and W. G. Dixon. Patient perspectives on sharing anonymized personal health data using a digital system for dynamic consent and research feedback: A qualitative study. *Journal of Medical Internet Research*, 18(4), 2016. ISSN 1438-8871. doi: 10.2196/jmir.5011.

[115] P. Brown, K. Brunnhuber, K. Chalkidou, I. Chalmers, M. Clarke, M. Fenton, C. Forbes, J. Glanville, N. J. Hicks., J. Moody, S. Twaddle, H. Timimi, and P. Young. Health research - how to formulate research recommendations. *British Medical Journal*, 333(7572):804–806, 2006. ISSN 0959-8146. doi: 10.1136/bmj.38987.492014.94.

[116] S. Greenland. Response: Bayesian perspectives for epidemiological research. *International Journal of Epidemiology*, 35(3):777–778, 2006. ISSN 0300-5771. doi: 10.1093/ije/dyl081.

[117] R. J. Lilford. Ethics of clinical trials from a bayesian and decision analytic perspective: whose equipoise is it anyway? *British Medical Journal*, 326(7396):980–981, 2003. ISSN 0959-535X. doi: 10.1136/bmj.326.7396.980.

[118] J. Donovan, N. Mills, M. Smith, L. Brindle, A. Jacoby, T. Peters, S. Frankel, D. Neal, F. Hamdy, and Grp Protect Study. Quality improvement report - improving design and conduct of randomised trials by embedding them in qualitative research: Protect (prostate testing for cancer and treatment) study. *British Medical Journal*, 325 (7367):766–769, 2002. ISSN 0959-535X. doi: 10.1136/bmj.325.7367.766.

[119] M. Caliendo and S. Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, 2008. ISSN 0950-0804. doi: 10.1111/j.1467-6419.2007.00527.x.

[120] D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. Consort 2010 explanation and elabora-

tion: updated guidelines for reporting parallel group randomised trials. *Bmj-British Medical Journal*, 340, 2010. ISSN 1756-1833. doi: 10.1136/bmj.c869.

[121] C. Tudur Smith, C. Hopkins, M. R. Sydes, K. Woolfall, M. Clarke, G. Murray, and P. Williamson. How should individual participant data (ipd) from publicly funded clinical trials be shared? *BMC Medicine*, 13(1): 298, 2015. ISSN 1741-7015. doi: 10.1186/s12916-015-0532-z. URL `https://doi.org/10.1186/s12916-015-0532-z`.

[122] J. Pildal, A. Hrobjartsson, K. J. Jorgensen, J. Hilden, D. G. Altman, and P. C. Gotzsche. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *International Journal of Epidemiology*, 36(4):847–857, 2007. ISSN 0300-5771. doi: 10.1093/ije/dym087.

[123] L. Wood, M. Egger, L. L. Gluud, K. F. Schulz, P. Juni, D. G. Altman, C. Gluud, R. M. Martin, A. J. G. Wood, and J. A. C. Sterne. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *British Medical Journal*, 336(7644):601–605, 2008. ISSN 0959-8146. doi: 10.1136/bmj.39465.451748.AD.

[124] P. J. Devereaux, B. J. Manns, W. A. Ghali, H. Quan, C. Lacchetti, V. M. Montori, M. Bhandari, and G. H. Guyatt. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. *Jama-Journal of the American Medical Association*, 285(15):2000–2003, 2001. ISSN 0098-7484. doi: 10.1001/jama.285.15.2000.

[125] H. C. Van Houwelingen, L. R. Arends, and T. Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4): 589–624, 2002. ISSN 0277-6715. doi: 10.1002/sim.1040.

[126] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10 (4):325–337, 2000. ISSN 0960-3174. doi: 10.1023/a:1008929526011.

[127] J. Savovic, H. E. Jones, D. G. Altman, R. J. Harris, P. Juni, J. Pildal, B. Als-Nielsen, E. M. Balk, C. Gluud, L. L. Gluud, J. P. A. Ioannidis, K. F. Schulz, R. Beynon, N. Welton, L. Wood, D. Moher, J. J. Deeks, and J. a. C. Sterne. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: Combined analysis of meta-epidemiological studies. *Health Technology Assessment*, 16(35):1–+, 2012. ISSN 1366-5278. doi: 10.3310/hta16350.

[128] A. Gelman. Prior distributions for variance parameters in hierarchical models(comment on an article by browne and draper). *Bayesian Analysis*, 1(3):515–533, 2006. ISSN 1931-6690.

[129] E. A. Akl, X. Sun, J. W. Busse, B. C. Johnston, M. Briel, S. Mulla, J. J. You, D. Bassler, F. Lamontagne, C. Vera, D. Heels-Ansdell M. Alshurafa, C. M. Katsios, Q. Zhou, E. Mills, and G. H. Guyatt. Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. *Journal of Clinical Epidemiology*, 65(3):262–267, 2012. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2011.04.015.

[130] J. Anzures-Cabrera, A. Sarpatwari, and J. P. T. Higgins. Expressing findings from meta-analyses of continuous outcomes in terms of risks. *Statistics in Medicine*, 30 (25):2967–2985, 2011. ISSN 0277-6715. doi: 10.1002/sim.4298.

[131] S. Chinn. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, 19(22):3127–3131, 2000. ISSN 0277-6715. doi: 10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M. URL `https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.C`

[132] S. G. Thompson and J. P. T. Higgins. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21(11):1559–1573, 2002. ISSN 0277-6715. doi: 10.1002/sim.1187.

[133] A. Hrobjartsson, A. S. S. Thomsen, F. Emanuelsson, B. Tendal, J. Hilden, I. Boutron, P. Ravaud, and S. Brorson. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *British Medical Journal*, 344, 2012. ISSN 1756-1833. doi: 10.1136/bmj.e1119.

[134] E. M. Balk, P. A. L. Bonis, H. Moskowitz, C. H. Schmid, J. P. A. Ioannidis, C. Wang, and J. Lau. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA*, 287(22):2973–2982, 2002. ISSN 0098-7484. doi: 10.1001/jama.287.22.2973.

[135] L. Bialy, B. Vandermeer, T. Lacaze-Masmonteil, D. M. Dryden, and L. Hartling. A meta-epidemiological study to examine the association between bias and treatment effects in neonatal trials. *Evidence-Based Child Health: A Cochrane Review Journal*, 9(4): 1052–1059, 2014. ISSN 1557-6272. doi: 10.1002/ebch.1985.

[136] A. Chaimani, C. H. Schmid, H. S. Vasiliadis, N. J. Welton, N. Pandis, and G. Salanti. Effects of study precision and risk of bias in networks of interventions: a network meta-epidemiological study. *International Journal of Epidemiology*, 42(4):1120–1131, 2013. ISSN 0300-5771. doi: 10.1093/ije/dyt074.

[137] E. Nüesch, S. Reichenbach, S. Trelle, A. W. S. Rutjes, K. Liewald, R. Sterchi, D. G. Altman, and P. Jüni. The importance of allocation concealment and patient blinding in osteoarthritis trials: A meta-epidemiologic study. *Arthritis Care & Research*, 61(12): 1633–1641, 2009. ISSN 0004-3591. doi: 10.1002/art.24894.

[138] L. Hartling, A. Milne, L. Tjosvold, D. Wrightson, J. Gallivan, and A. S. Newton.

A systematic review of interventions to support siblings of children with chronic illness or disability. *Journal of Paediatrics and Child Health*, 50(10):E26–E38, 2014. ISSN 1034-4810. doi: 10.1111/j.1440-1754.2010.01771.x.

[139] S. Armijo-Olivo, J. Fuentes, B. R. da Costa, H. Saltaji, C. Ha, and G. G. Cummings. Blinding in physical therapy trials and its association with treatment effects: A meta-epidemiological study. *American Journal of Physical Medicine & Rehabilitation*, 96(1), 2017. ISSN 0894-9115.

[140] K. M. Rhodes, D. Mawdsley, R. M. Turner, H. E. Jones, J. Savovic, and J. P. T. Higgins. Label-invariant models for the analysis of meta-epidemiological data. *Statistics in Medicine*, 37(1):60–70, 2018. ISSN 0277-6715. doi: 10.1002/sim.7491.

[141] R. D. Riley, K. R. Abrams, P. C. Lambert, A. J. Sutton, and J. R. Thompson. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, 26(1):78–97, 2007. ISSN 0277-6715. doi: 10.1002/sim.2524.

[142] S. Bujkiewicz, J. R. Thompson, A. J. Sutton, N. J. Cooper, M. J. Harrison, D. P. M. Symmons, and K. R. Abrams. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Statistics in Medicine*, 32(22):3926–3943, 2013. ISSN 0277-6715. doi: 10.1002/sim.5831.

[143] Y. Wei and J. P. T. Higgins. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(7):1191–1205, 2013. ISSN 0277-6715. doi: 10.1002/sim.5679.

[144] S. Bujkiewicz, J. R. Thompson, R. D. Riley, and K. R. Abrams. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Statistics in medicine*, 35(7):

1063–1089, 2016. ISSN 1097-0258 0277-6715. doi: 10.1002/sim.6776. URL https://www.ncbi.nlm.nih.gov/pubmed/26530518 https://www.ncbi.nlm.nih.gov/pmc/PMC4950070/.

[145] S. Bujkiewicz, J. R. Thompson, E. Spata, and K. R. Abrams. Uncertainty in the bayesian meta-analysis of normally distributed surrogate endpoints. *Statistical Methods in Medical Research*, 26(5):2287–2318, 2015. ISSN 0962-2802. doi: 10.1177/0962280215597260. URL https://doi.org/10.1177/0962280215597260.

[146] R. D. Riley, K. R. Abrams, A. J. Sutton, P. C. Lambert, and J. R. Thompson. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *Bmc Medical Research Methodology*, 7, 2007. ISSN 1471-2288. doi: 10.1186/1471-2288-7-3.

[147] N. J. Welton, M. O. Soares, A. E. Ades, D. Harrison, M. Shankar-Hari, and M. K. Rowan. Accounting for heterogeneity in relative treatment effects for use in cost-effectiveness models and value-of-information analyses. *Medical Decision Making*, 35(5):608–621, 2015. doi: 10.1177/0272989X15570113.

[148] D. G. Contopoulos-Ioannidis, S. M. Gilbody, T. A. Trikalinos, R. Churchil, K. Wahlbeck, J. P. A. Ioannidis, and E.-P. Project. Comparison of large versus smaller randomized trials for mental health-related interventions. *American Journal of Psychiatry*, 162(3):578–584, 2005. ISSN 0002-953X. doi: 10.1176/appi.ajp.162.3.578.

[149] A. Dechartres, L. Trinquart I. Boutron, P. Charles, and P. Ravaud. Single-center trials show larger treatment effects than multicenter trials: Evidence from a meta-epidemiologic study. *Annals of Internal Medicine*, 155(1):39–+, 2011. ISSN 0003-4819. doi: 10.7326/0003-4819-155-1-201107050-00006.

[150] J. A. C. Sterne, A. J. Sutton, J. P. A. Ioannidis, N. Terrin, D. R. Jones, J. Lau, J. Car-

penter, G. Rucker, R. M. Harbord, C. H. Schmid, J. Tetzlaff, J. J. Deeks, J. Peters, P. Macaskill, G. Schwarzer, S. Duval, D. G. Altman, D. Moher, and J. P. T. Higgins. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal*, 343, 2011. ISSN 1756-1833. doi: 10.1136/bmj.d4002.

[151] J. A. C. Sterne, D. Gavaghan, and M. Egger. Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, 53(11):1119–1129, 2000. ISSN 0895-4356. doi: 10.1016/s0895-4356(00)00242-0.

[152] S. G. Moreno, A. J. Sutton, A. E. Ades, T. D. Stanley, K. R. Abrams, J. L. Peters, and N. J. Cooper. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *Bmc Medical Research Methodology*, 9, 2009. ISSN 1471-2288. doi: 10.1186/1471-2288-9-2.

[153] S. G. Moreno, A. J. Sutton, E. H. Turner, K. R. Abrams, N. J. Cooper, T. M. Palmer, and A. E. Ades. Novel Methods to Deal with Publication Biases: Secondary Analysis of Antidepressant Trials in the FDA Trial Registry Database and Related Journal Publications. *British Medical Journal*, 339, 2009. ISSN 0959-8146. doi: 10.1136/bmj.b2981.

[154] J. L. Peters, A. J. Sutton, D. R. Jones, K. R. Abrams, and L. Rushton. Comparison of two methods to detect publication bias in meta-analysis. *Jama-Journal of the American Medical Association*, 295(6):676–680, 2006. ISSN 0098-7484. doi: 10.1001/jama.295.6.676.

[155] D. J. Spiegelhalter, N. G. Best, B. R. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64:583–616, 2002. ISSN 1369-7412. doi: 10.1111/1467-9868.00353.

[156] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 76(3):485–493, 2014. ISSN 1369-7412. doi: 10.1111/rssb.12062.

[157] A. Van Der Linde. DIC in variable selection. *Statistica Neerlandica*, 59(1):45–56, 2005. ISSN 0039-0402.

[158] A. Dechartres, L. Trinquart, I. Boutron, and P. Ravaud. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. *Bmj-British Medical Journal*, 346, 2013. ISSN 1756-1833. doi: 10.1136/bmj.f2304.

[159] P. Charles, B. Giraudeau, A. Dechartres, G. Baron, and P. Ravaud. Reporting of sample size calculation in randomised controlled trials: review. *British Medical Journal*, 338, 2009. ISSN 0959-8146. doi: 10.1136/bmj.b1732.

[160] D. Moher, C. S. Dulberg, and G. A. Wells. Statistical power, sample-size, and their reporting in randomized controlled trials. *Jama-Journal of the American Medical Association*, 272(2):122–124, 1994. ISSN 0098-7484. doi: 10.1001/jama.272.2.122.

[161] C. Mckenna and K. Claxton. Addressing adoption and research design decisions simultaneously: The role of value of sample information analysis. *Medical Decision Making*, 31(6):853–865, 2011. ISSN 0272-989X. doi: 10.1177/0272989x11399921.

[162] C. Mckenna, S. Griffin, H. Koffijberg, and K. Claxton. Methods to place a value on additional evidence are illustrated using a case study of corticosteroids after traumatic brain injury. *Journal of Clinical Epidemiology*, 70:183–190, 2016. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2015.09.011.

[163] R. A. Turner, R. Z. Omar, and S. G. Thompson. Constructing intervals for the intracluster correlation coefficient using Bayesian modelling, and application in cluster

randomized trials. *Statistics in Medicine*, 25(9):1443–1456, 2006. ISSN 0277-6715. doi: 10.1002/sim.2304.

[164] M. Rotondi and A. Donner. Sample size estimation in cluster randomized trials: An evidence-based perspective. *Computational Statistics & Data Analysis*, 56(5):1174–1187, 2012. ISSN 0167-9473. doi: 10.1016/j.csda.2010.12.010.

[165] J. Lerman. Study design in clinical research: Sample size estimation and power analysis. *Canadian Journal of Anaesthesia-Journal Canadien D Anesthesie*, 43(2):184–191, 1996. ISSN 0832-610X.

[166] W. D. Dupont and W. D. Plummer. Power and sample-size calculations - a review and computer-program. *Controlled Clinical Trials*, 11(2):116–128, 1990. ISSN 0197-2456. doi: 10.1016/0197-2456(90)90005-m.

[167] S. E. Maxwell, K. Kelley, and J. R. Rausch. Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59:537–563, 2008. ISSN 0066-4308. doi: 10.1146/annurev.psych.59.103006.093735.

[168] CRASH trial collaborators. Effect of intravenous corticosteroids on death within 14 days in 10,008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. *Lancet*, 364:1321–1328, 2004.

[169] H. E. Jones, A. E. Ades, A. J. Sutton, and N. J. Welton. Use of a random effects meta-analysis in the design and analysis of a new clinical trial. *Statistics in medicine*, 2018. doi: 10.1002/sim.7948.

[170] W. G. Henderson, T. Moritz, S. Goldman, J. Copeland, and G. Sethi. Use of cumulative metaanalysis in the design, monitoring, and final analysis of a clinical-trial - a case-study. *Controlled Clinical Trials*, 16(5):331–341, 1995. ISSN 0197-2456. doi: 10.1016/0197-2456(95)00071-2.

[171] A. O'hagan, J. W. Stevens, and M. J. Campbell. Assurance in clinical trial design. *Pharmaceutical Statistics*, 4(3):187–201, 2005. ISSN 1539-1604. doi: 10.1002/pst.175.

[172] M. O. Soares, N. J. Welton, D. Harrison, P. Peura, and M. Shankar-Hari. An evaluation of the feasibility, cost and value of information of a multicentre randomised controlled trial of intravenous immunoglobulin for sepsis (severe sepsis and septic shock): incorporating a systematic review, meta-analysis and value of information analysis. *Health Technol Assess*, 16(7), 2012.

[173] R. Collins, R. Peto, M. Flather, S. Parish, and P. Sleight et al. ISIS-4 - a randomized factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium-sulfate in 58,050 patients with suspected acute myocardial-infarction. *Lancet*, 345(8951):669–685, 1995. ISSN 0099-5355.

[174] A. E. Ades, G. Lu, and J. P. T. Higgins. The interpretation of random-effects meta-analysis in decision models. *Medical Decision Making*, 25(6):646–654, 2005. ISSN 0272-989X. doi: 10.1177/0272989x05282643.

[175] A. Nikolakopoulou, D. Mavridis, and G. Salanti. Using conditional power of network meta-analysis (nma) to inform the design of future clinical trials. *Biometrical Journal*, 56(6):973–990, 2014. ISSN 0323-3847. doi: 10.1002/bimj.201300216.

[176] A. Sutton, N. Cooper, and K. Abrams. Evidence based sample size calculations for future trials based on results of current meta-analyses. *Controlled Clinical Trials*, 24: 88S–88S, 2003. ISSN 0197-2456.

[177] L. M. Kunz, R. W. Yeh, and S.-L. T. Normand. Comparative effectiveness research: does one size fit all? *Statistics in Medicine*, 31(25):3062–3065, 2012. ISSN 0277-6715. doi: 10.1002/sim.5482.

[178] S. Griffin, N. J. Welton, and K. Claxton. Exploring the research decision space: The

expected value of information for sequential research designs. *Medical Decision Making*, 30(2):155–162, 2010. ISSN 0272-989X. doi: 10.1177/0272989x09344746.

[179] G. L. Clayton, A. D. Schachter, B. Magnusson, Y. Li, and L. Colin. How often do safety signals occur by chance in first-in-human trials? *Cts-Clinical and Translational Science*, 11(5):471–476, 2018. ISSN 1752-8054. doi: 10.1111/cts.12558.

[180] C. Buoen, O. J. Bjerrum, and M. S. Thomsen. How first-time-in-human studies are being performed: A survey of phase I dose-escalation trials in healthy volunteers published between 1995 and 2004. *Journal of Clinical Pharmacology*, 45(10):1123–1136, 2005. ISSN 0091-2700. doi: 10.1177/0091270005279943.

[181] C. Buoen, S. Holm, and M. S. Thomsen. Evaluation of the cohort size in phase I dose escalation trials based on laboratory data. *Journal of Clinical Pharmacology*, 43 (5):470–476, 2003. ISSN 0091-2700. doi: 10.1177/0091270003252243.

[182] L. V. Sacks, H. H. Shamsuddin, Y. I. Yasinskaya, K. Bouri, M. L. Lanthier, and R. E. Sherman. Scientific and Regulatory Reasons for Delay and Denial of FDA Approval of Initial Applications for New Drugs, 2000-2012. *Jama-Journal of the American Medical Association*, 311(4):378–384, 2014. ISSN 0098-7484. doi: 10.1001/jama.2013.282542.

[183] L. Martin, M. Hutchens, C. Hawkins, and A. Radnov. How much do clinical trials cost? *Nature Reviews Drug Discovery*, 16(6):381–382, 2017. ISSN 1474-1776. doi: 10.1038/nrd.2017.70.

[184] P. Rosenzweig, N. Miget, and S. Brohier. Transaminase elevation on placebo during Phase I trials: prevalence and significance. *British Journal of Clinical Pharmacology*, 48 (1):19–23, 1999. ISSN 0306-5251.

[185] G. Molenberghs and G. Verbeke. *Models for discrete longitudinal data*. Springer, New York, 2005.

[186] A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017. ISSN 0960-3174. doi: 10.1007/s11222-016-9696-4.

[187] B. Carpenter, A. Gelman, M. D. Hoffman, B. Goodrich D. Lee, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–29, 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01.

[188] Stan Development Team. Rstanarm: Bayesian applied regression modeling via stan. r package version 2.15.3. Report, 2017. URL `http://mc-stan.org`.

[189] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135, 2009. ISSN 0169-5347. doi: 10.1016/j.tree.2008.10.008.

[190] M. Pavlou, G. Ambler, S. Seaman, and R. Z. Omar. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *Bmc Medical Research Methodology*, 15, 2015. ISSN 1471-2288. doi: 10.1186/s12874-015-0046-6.

[191] A. Skrondal and S. Rabe-Hesketh. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 172:659–687, 2009. ISSN 0964-1998.

[192] B. Neuenschwander, G. Capkun-Niggli, M. Branson, and D. J. Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, 2010. ISSN 1740-7745. doi: 10.1177/1740774509356002.

[193] H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'hagan, D. Spiegelhalter, and

B. Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014. ISSN 0006-341X. doi: 10.1111/biom.12242.

[194] C. Brard, G. Le Teuff, M.-C. Le Deley, and L. V. Hampson. Bayesian survival analysis in clinical trials: What methods are used in practice? *Clinical trials (London, England)*, 14(1):78–87, 2017. doi: 10.1177/1740774516673362.

[195] L. V. Hampson, J. Whitehead, D. Eleftheriou, and P. Brogan. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, 33(24):4186–4201, 2014. ISSN 0277-6715. doi: 10.1002/sim.6225.

[196] L. V. Hampson, J. Whitehead, D. Eleftheriou, C. Tudur-Smith, and R. Jones et al. Elicitation of expert prior opinion: Application to the mypan trial in childhood polyarteritis nodosa. *Plos One*, 10(3), 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0120981.

[197] A. C. Goudie, A. J. Sutton, D. R. Jones, and A. Donald. Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. *Journal of Clinical Epidemiology*, 63(9):983–991, 2010. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2010.01.022.

[198] K. A. Robinson and S. N. Goodman. A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Annals of Internal Medicine*, 154(1):50–U187, 2011. ISSN 0003-4819. doi: 10.7326/0003-4819-154-1-201101040-00007.

[199] S. A. Iqbal, J. D. Wallach, M. J. Khoury, S. D. Schully, and J. P. A. Ioannidis. Reproducible research practices and transparency across the biomedical literature. *Plos Biology*, 14(1), 2016. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002333.

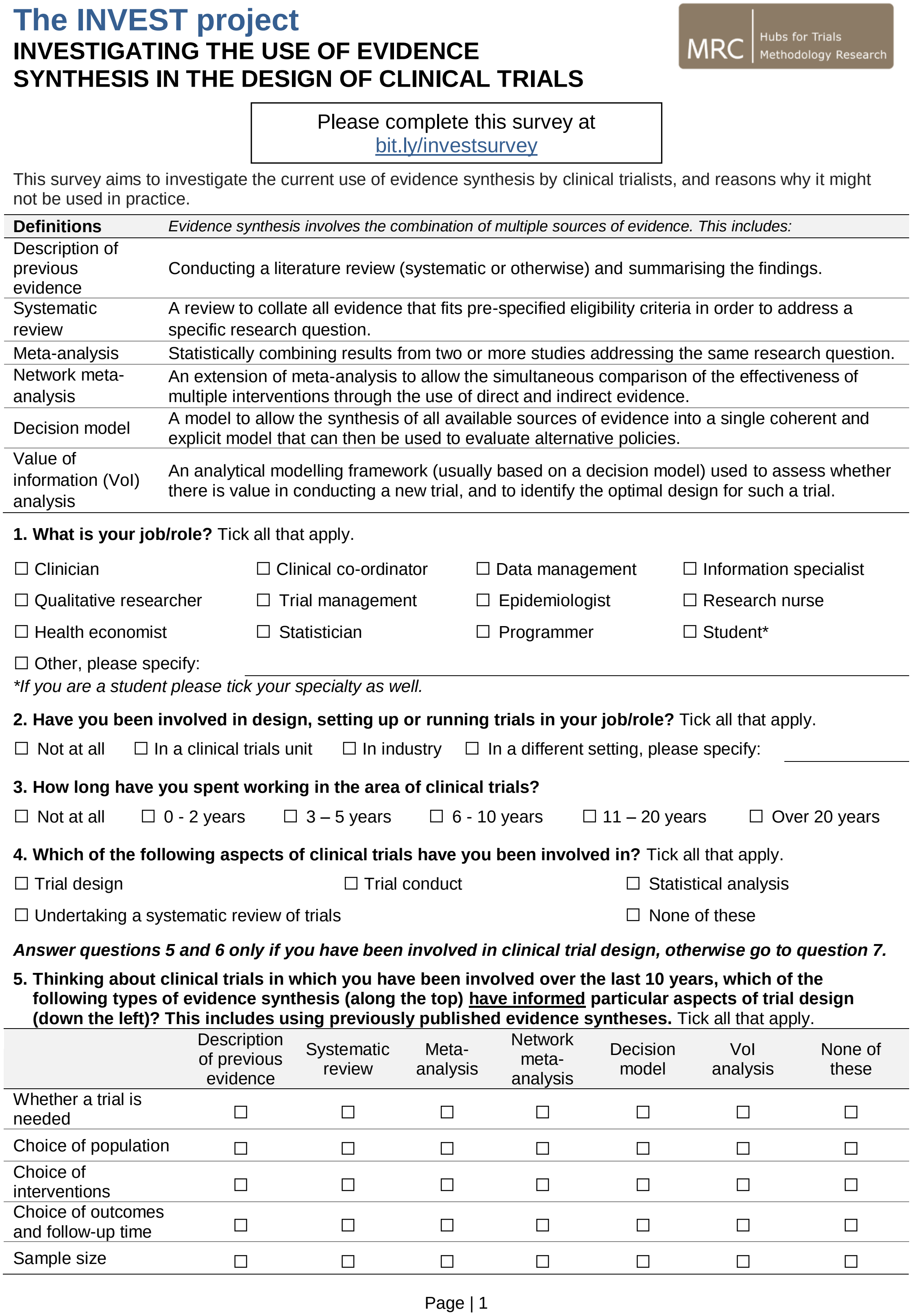
[200] M. J. Page, D. G. Altman, L. Shamseer, J. E. Mckenzie, N. Ahmadzai, D. Wolfe, F. Yazdi, F. Catala-Lopez, A. C. Tricco, and D. Moher. Reproducible research practices are underused in systematic reviews of biomedical interventions. *Journal of Clinical Epidemiology*, 94:8–18, 2018. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2017.10.017.

A Documents and Tables

Table A.1: Differences between the topic guide for statisticians and clinicians.

Topic guide question for statisticians	**Topic guide question for clinicians**
Can you describe the types of methods you are using to analyse trials?	What is your role in terms of your input into the analysis of the trial results?
What are your thoughts on the use of Bayesian methods to analyse a clinical trial?	What is it you understand by the use of Bayesian methods to analyse a clinical trial?

Figure A.1: The INVEST survey.



The INVEST project

INVESTIGATING THE USE OF EVIDENCE SYNTHESIS IN THE DESIGN OF CLINICAL TRIALS

MRC Hubs for Trials Methodology Research

Please complete this survey at
bit.ly/investsurvey

This survey aims to investigate the current use of evidence synthesis by clinical trialists, and reasons why it might not be used in practice.

Definitions	*Evidence synthesis involves the combination of multiple sources of evidence. This includes:*
Description of previous evidence	Conducting a literature review (systematic or otherwise) and summarising the findings.
Systematic review	A review to collate all evidence that fits pre-specified eligibility criteria in order to address a specific research question.
Meta-analysis	Statistically combining results from two or more studies addressing the same research question.
Network meta-analysis	An extension of meta-analysis to allow the simultaneous comparison of the effectiveness of multiple interventions through the use of direct and indirect evidence.
Decision model	A model to allow the synthesis of all available sources of evidence into a single coherent and explicit model that can then be used to evaluate alternative policies.
Value of information (VoI) analysis	An analytical modelling framework (usually based on a decision model) used to assess whether there is value in conducting a new trial, and to identify the optimal design for such a trial.

1. What is your job/role? Tick all that apply.

☐ Clinician ☐ Clinical co-ordinator ☐ Data management ☐ Information specialist

☐ Qualitative researcher ☐ Trial management ☐ Epidemiologist ☐ Research nurse

☐ Health economist ☐ Statistician ☐ Programmer ☐ Student*

☐ Other, please specify: ____________________

**If you are a student please tick your specialty as well.*

2. Have you been involved in design, setting up or running trials in your job/role? Tick all that apply.

☐ Not at all ☐ In a clinical trials unit ☐ In industry ☐ In a different setting, please specify: __________

3. How long have you spent working in the area of clinical trials?

☐ Not at all ☐ 0 - 2 years ☐ 3 – 5 years ☐ 6 - 10 years ☐ 11 – 20 years ☐ Over 20 years

4. Which of the following aspects of clinical trials have you been involved in? Tick all that apply.

☐ Trial design ☐ Trial conduct ☐ Statistical analysis

☐ Undertaking a systematic review of trials ☐ None of these

Answer questions 5 and 6 only if you have been involved in clinical trial design, otherwise go to question 7.

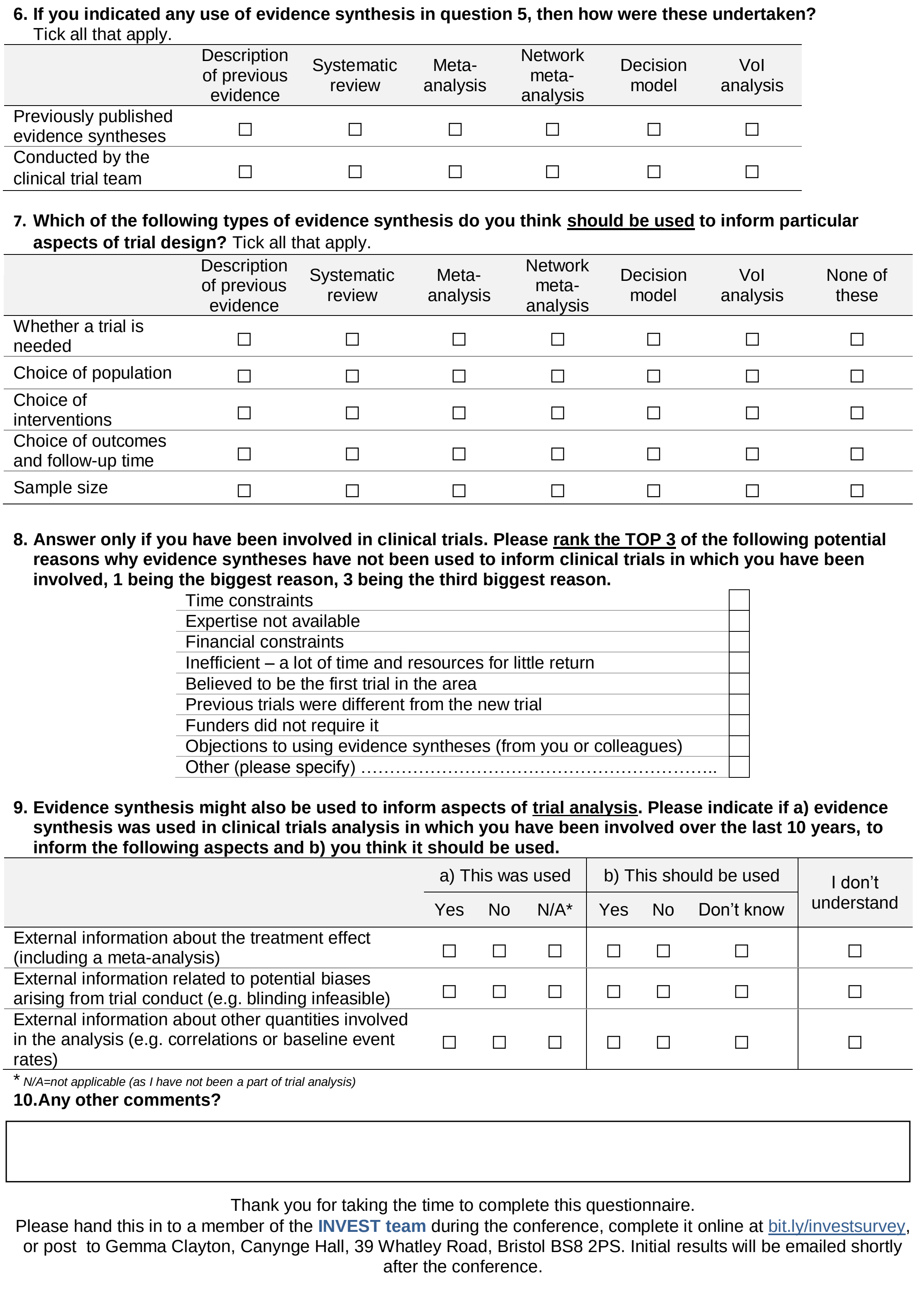
5. Thinking about clinical trials in which you have been involved over the last 10 years, which of the following types of evidence synthesis (along the top) <u>have informed</u> particular aspects of trial design (down the left)? This includes using previously published evidence syntheses. Tick all that apply.

	Description of previous evidence	Systematic review	Meta-analysis	Network meta-analysis	Decision model	VoI analysis	None of these
Whether a trial is needed	☐	☐	☐	☐	☐	☐	☐
Choice of population	☐	☐	☐	☐	☐	☐	☐
Choice of interventions	☐	☐	☐	☐	☐	☐	☐
Choice of outcomes and follow-up time	☐	☐	☐	☐	☐	☐	☐
Sample size	☐	☐	☐	☐	☐	☐	☐

Page | 1

(a) Page 1 of 2

Figure A.1: The INVEST survey. (cont.)



6. If you indicated any use of evidence synthesis in question 5, then how were these undertaken? Tick all that apply.

	Description of previous evidence	Systematic review	Meta-analysis	Network meta-analysis	Decision model	VoI analysis
Previously published evidence syntheses	□	□	□	□	□	□
Conducted by the clinical trial team	□	□	□	□	□	□

7. Which of the following types of evidence synthesis do you think <u>should be used</u> to inform particular aspects of trial design? Tick all that apply.

	Description of previous evidence	Systematic review	Meta-analysis	Network meta-analysis	Decision model	VoI analysis	None of these
Whether a trial is needed	□	□	□	□	□	□	□
Choice of population	□	□	□	□	□	□	□
Choice of interventions	□	□	□	□	□	□	□
Choice of outcomes and follow-up time	□	□	□	□	□	□	□
Sample size	□	□	□	□	□	□	□

8. Answer only if you have been involved in clinical trials. Please <u>rank the TOP 3</u> of the following potential reasons why evidence syntheses have not been used to inform clinical trials in which you have been involved, 1 being the biggest reason, 3 being the third biggest reason.

Time constraints	□
Expertise not available	□
Financial constraints	□
Inefficient – a lot of time and resources for little return	□
Believed to be the first trial in the area	□
Previous trials were different from the new trial	□
Funders did not require it	□
Objections to using evidence syntheses (from you or colleagues)	□
Other (please specify) ……………………………………………………..	□

9. Evidence synthesis might also be used to inform aspects of <u>trial analysis</u>. Please indicate if a) evidence synthesis was used in clinical trials analysis in which you have been involved over the last 10 years, to inform the following aspects and b) you think it should be used.

	a) This was used			b) This should be used			I don't understand
	Yes	No	N/A*	Yes	No	Don't know	
External information about the treatment effect (including a meta-analysis)	□	□	□	□	□	□	□
External information related to potential biases arising from trial conduct (e.g. blinding infeasible)	□	□	□	□	□	□	□
External information about other quantities involved in the analysis (e.g. correlations or baseline event rates)	□	□	□	□	□	□	□

* *N/A=not applicable (as I have not been a part of trial analysis)*

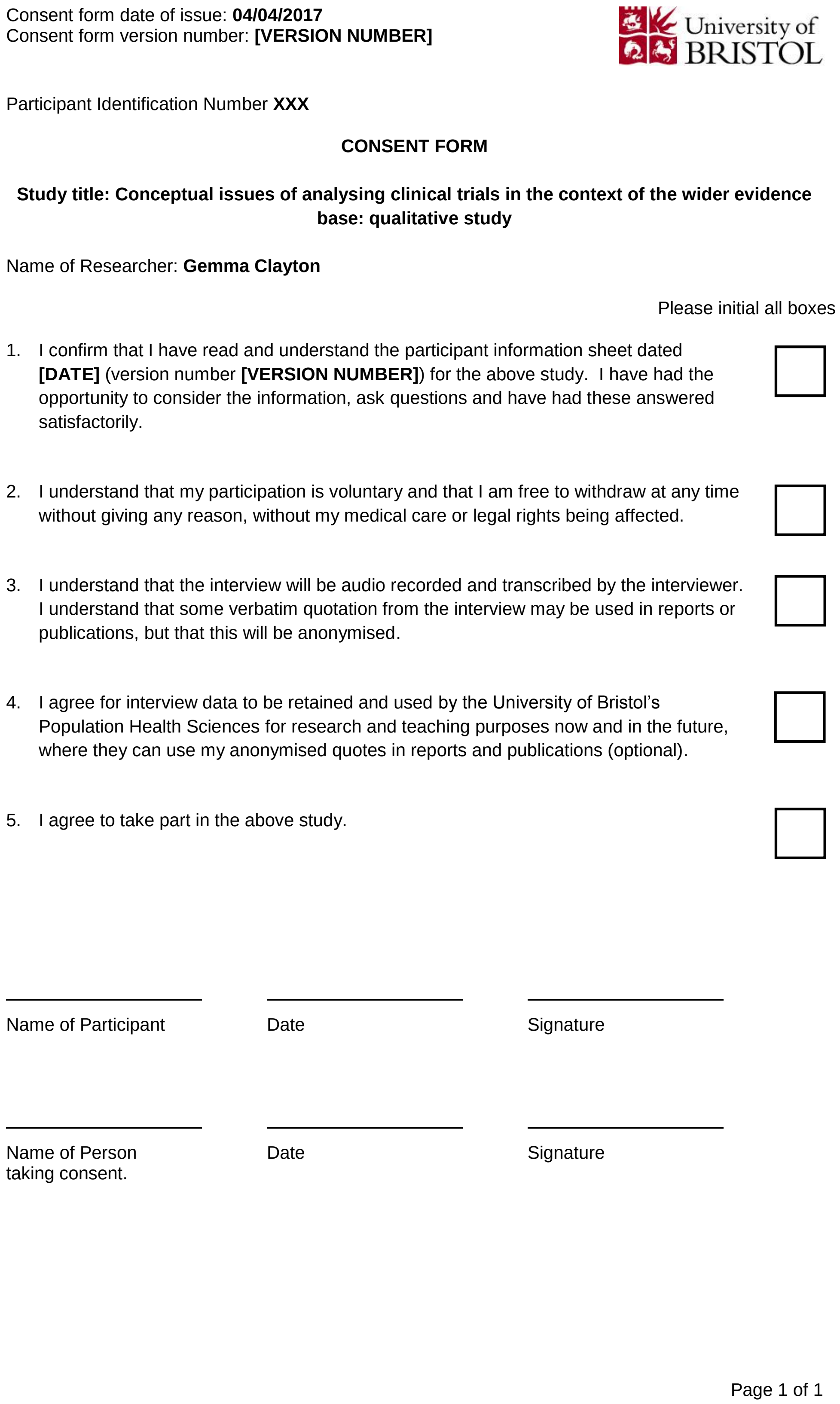
10.Any other comments?

Thank you for taking the time to complete this questionnaire.
Please hand this in to a member of the **INVEST team** during the conference, complete it online at bit.ly/investsurvey, or post to Gemma Clayton, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS. Initial results will be emailed shortly after the conference.

Page | 2

(b) Page 2 of 2

Figure A.2: Consent form.



Consent form date of issue: **04/04/2017**
Consent form version number: **[VERSION NUMBER]**

University of BRISTOL

Participant Identification Number **XXX**

CONSENT FORM

Study title: Conceptual issues of analysing clinical trials in the context of the wider evidence base: qualitative study

Name of Researcher: **Gemma Clayton**

Please initial all boxes

1. I confirm that I have read and understand the participant information sheet dated **[DATE]** (version number **[VERSION NUMBER]**) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. □

2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my medical care or legal rights being affected. □

3. I understand that the interview will be audio recorded and transcribed by the interviewer. I understand that some verbatim quotation from the interview may be used in reports or publications, but that this will be anonymised. □

4. I agree for interview data to be retained and used by the University of Bristol's Population Health Sciences for research and teaching purposes now and in the future, where they can use my anonymised quotes in reports and publications (optional). □

5. I agree to take part in the above study. □

________________ ________________ ________________

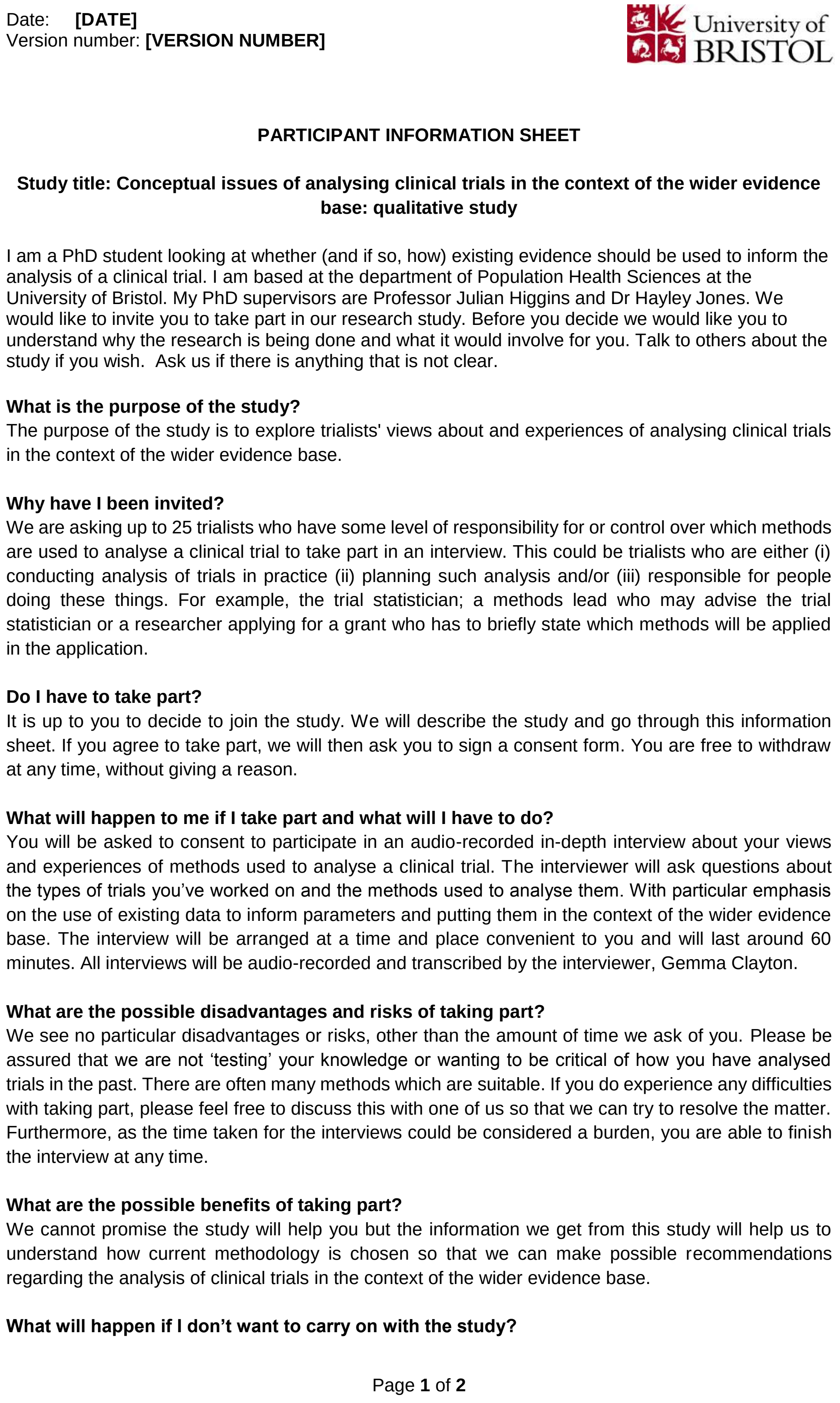
Name of Participant Date Signature

________________ ________________ ________________

Name of Person taking consent. Date Signature

Page 1 of 1

Figure A.3: Participant information sheet.



Date: **[DATE]**
Version number: **[VERSION NUMBER]**

PARTICIPANT INFORMATION SHEET

Study title: Conceptual issues of analysing clinical trials in the context of the wider evidence base: qualitative study

I am a PhD student looking at whether (and if so, how) existing evidence should be used to inform the analysis of a clinical trial. I am based at the department of Population Health Sciences at the University of Bristol. My PhD supervisors are Professor Julian Higgins and Dr Hayley Jones. We would like to invite you to take part in our research study. Before you decide we would like you to understand why the research is being done and what it would involve for you. Talk to others about the study if you wish. Ask us if there is anything that is not clear.

What is the purpose of the study?
The purpose of the study is to explore trialists' views about and experiences of analysing clinical trials in the context of the wider evidence base.

Why have I been invited?
We are asking up to 25 trialists who have some level of responsibility for or control over which methods are used to analyse a clinical trial to take part in an interview. This could be trialists who are either (i) conducting analysis of trials in practice (ii) planning such analysis and/or (iii) responsible for people doing these things. For example, the trial statistician; a methods lead who may advise the trial statistician or a researcher applying for a grant who has to briefly state which methods will be applied in the application.

Do I have to take part?
It is up to you to decide to join the study. We will describe the study and go through this information sheet. If you agree to take part, we will then ask you to sign a consent form. You are free to withdraw at any time, without giving a reason.

What will happen to me if I take part and what will I have to do?
You will be asked to consent to participate in an audio-recorded in-depth interview about your views and experiences of methods used to analyse a clinical trial. The interviewer will ask questions about the types of trials you've worked on and the methods used to analyse them. With particular emphasis on the use of existing data to inform parameters and putting them in the context of the wider evidence base. The interview will be arranged at a time and place convenient to you and will last around 60 minutes. All interviews will be audio-recorded and transcribed by the interviewer, Gemma Clayton.

What are the possible disadvantages and risks of taking part?
We see no particular disadvantages or risks, other than the amount of time we ask of you. Please be assured that we are not 'testing' your knowledge or wanting to be critical of how you have analysed trials in the past. There are often many methods which are suitable. If you do experience any difficulties with taking part, please feel free to discuss this with one of us so that we can try to resolve the matter. Furthermore, as the time taken for the interviews could be considered a burden, you are able to finish the interview at any time.

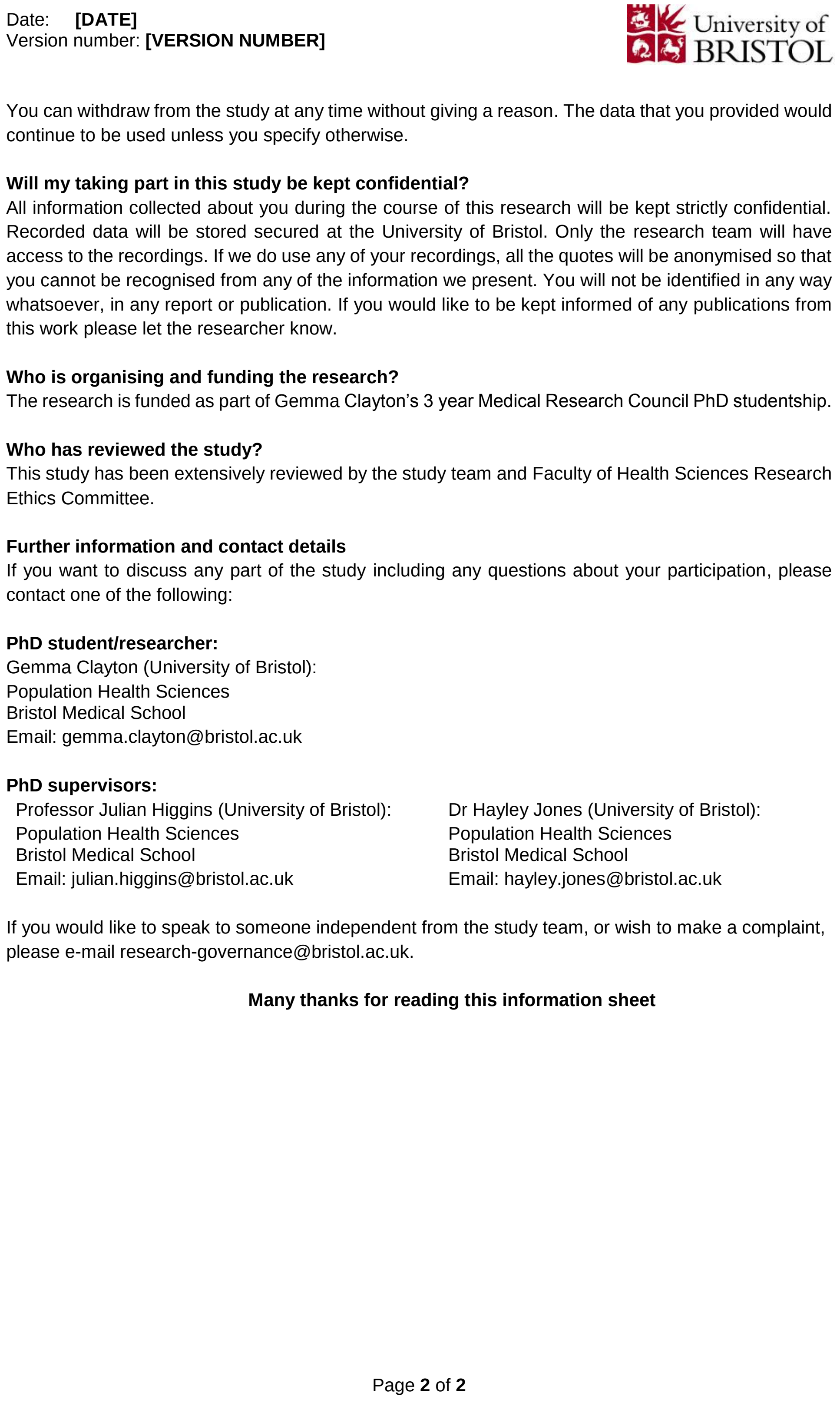
What are the possible benefits of taking part?
We cannot promise the study will help you but the information we get from this study will help us to understand how current methodology is chosen so that we can make possible recommendations regarding the analysis of clinical trials in the context of the wider evidence base.

What will happen if I don't want to carry on with the study?

Page **1** of **2**

(a) Page 1 of 2

Figure A.3: Participant information sheet (cont.)



Date: **[DATE]**
Version number: **[VERSION NUMBER]**

You can withdraw from the study at any time without giving a reason. The data that you provided would continue to be used unless you specify otherwise.

Will my taking part in this study be kept confidential?
All information collected about you during the course of this research will be kept strictly confidential. Recorded data will be stored secured at the University of Bristol. Only the research team will have access to the recordings. If we do use any of your recordings, all the quotes will be anonymised so that you cannot be recognised from any of the information we present. You will not be identified in any way whatsoever, in any report or publication. If you would like to be kept informed of any publications from this work please let the researcher know.

Who is organising and funding the research?
The research is funded as part of Gemma Clayton's 3 year Medical Research Council PhD studentship.

Who has reviewed the study?
This study has been extensively reviewed by the study team and Faculty of Health Sciences Research Ethics Committee.

Further information and contact details
If you want to discuss any part of the study including any questions about your participation, please contact one of the following:

PhD student/researcher:
Gemma Clayton (University of Bristol):
Population Health Sciences
Bristol Medical School
Email: gemma.clayton@bristol.ac.uk

PhD supervisors:

Professor Julian Higgins (University of Bristol):
Population Health Sciences
Bristol Medical School
Email: julian.higgins@bristol.ac.uk

Dr Hayley Jones (University of Bristol):
Population Health Sciences
Bristol Medical School
Email: hayley.jones@bristol.ac.uk

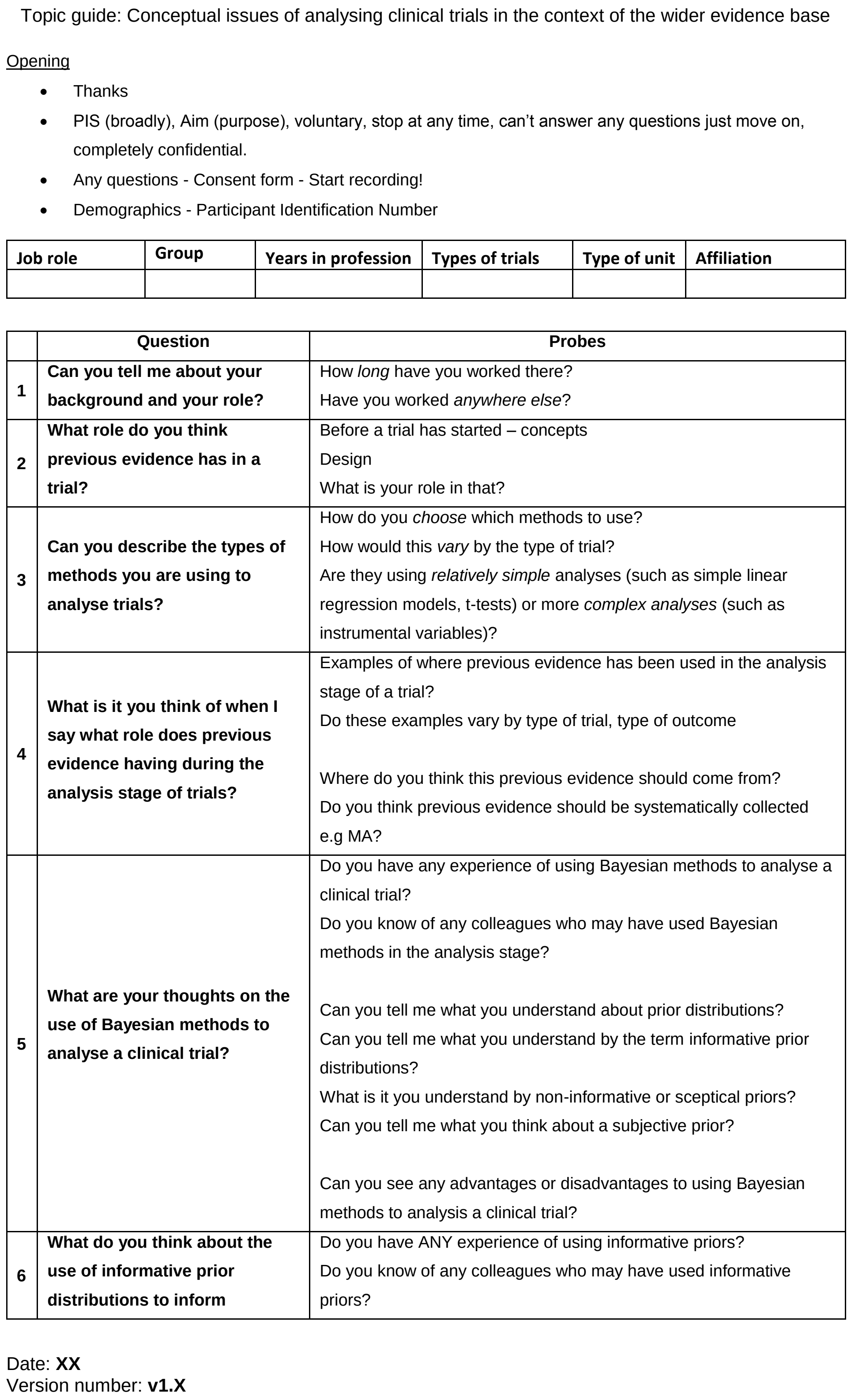
If you would like to speak to someone independent from the study team, or wish to make a complaint, please e-mail research-governance@bristol.ac.uk.

Many thanks for reading this information sheet

Page **2** of **2**

(b) Page 2 of 2

Figure A.4: Example of topic guide



Topic guide: Conceptual issues of analysing clinical trials in the context of the wider evidence base

Opening

- Thanks
- PIS (broadly), Aim (purpose), voluntary, stop at any time, can't answer any questions just move on, completely confidential.
- Any questions - Consent form - Start recording!
- Demographics - Participant Identification Number

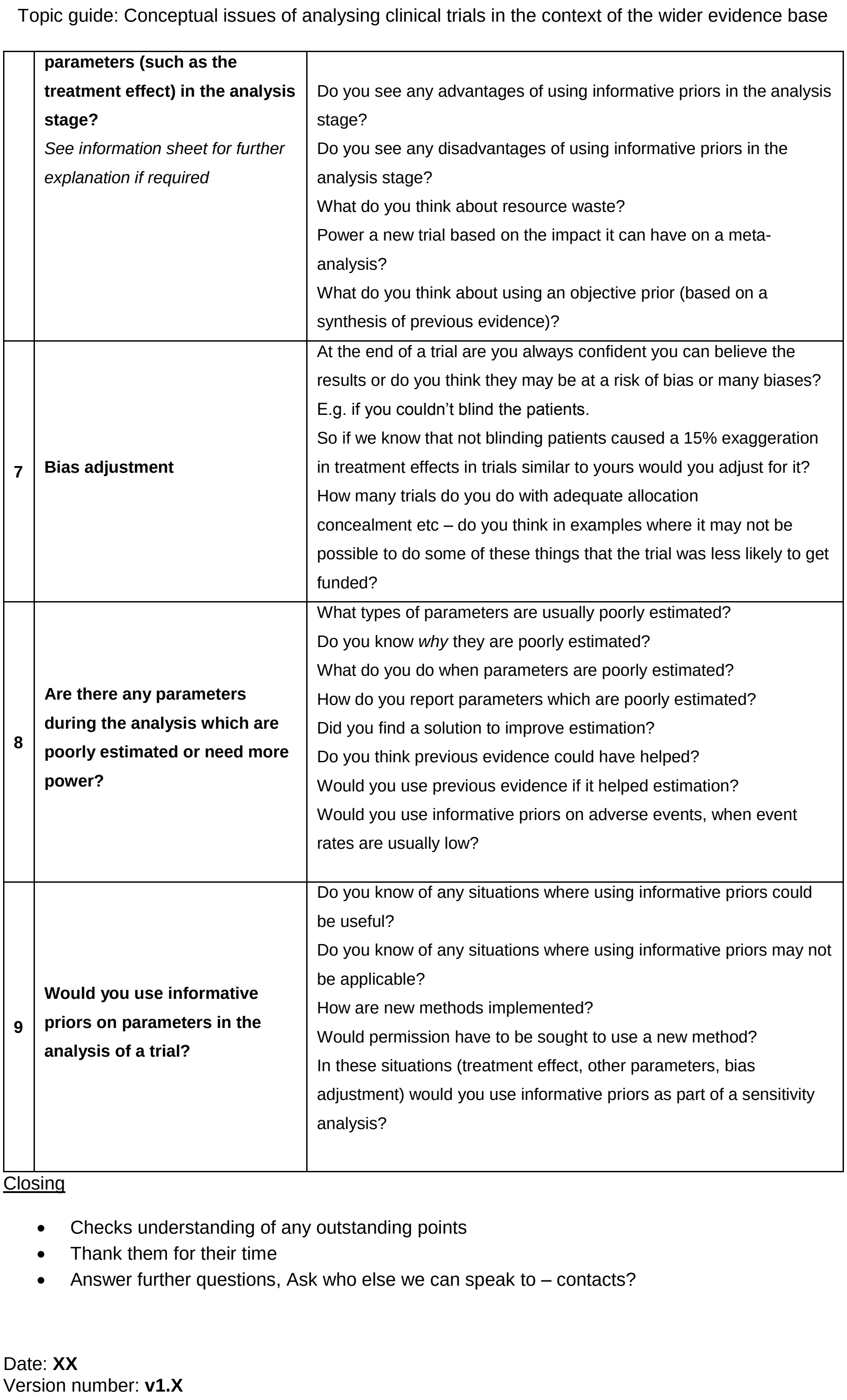
Job role	Group	Years in profession	Types of trials	Type of unit	Affiliation

	Question	Probes
1	**Can you tell me about your background and your role?**	How *long* have you worked there? Have you worked *anywhere else*?
2	**What role do you think previous evidence has in a trial?**	Before a trial has started – concepts Design What is your role in that?
3	**Can you describe the types of methods you are using to analyse trials?**	How do you *choose* which methods to use? How would this *vary* by the type of trial? Are they using *relatively simple* analyses (such as simple linear regression models, t-tests) or more *complex analyses* (such as instrumental variables)?
4	**What is it you think of when I say what role does previous evidence having during the analysis stage of trials?**	Examples of where previous evidence has been used in the analysis stage of a trial? Do these examples vary by type of trial, type of outcome Where do you think this previous evidence should come from? Do you think previous evidence should be systematically collected e.g MA?
5	**What are your thoughts on the use of Bayesian methods to analyse a clinical trial?**	Do you have any experience of using Bayesian methods to analyse a clinical trial? Do you know of any colleagues who may have used Bayesian methods in the analysis stage? Can you tell me what you understand about prior distributions? Can you tell me what you understand by the term informative prior distributions? What is it you understand by non-informative or sceptical priors? Can you tell me what you think about a subjective prior? Can you see any advantages or disadvantages to using Bayesian methods to analysis a clinical trial?
6	**What do you think about the use of informative prior distributions to inform**	Do you have ANY experience of using informative priors? Do you know of any colleagues who may have used informative priors?

Date: **XX**
Version number: **v1.X**

(a) Page 1 of 2

Figure A.4: Example of topic guide (cont.)



Topic guide: Conceptual issues of analysing clinical trials in the context of the wider evidence base

	parameters (such as the treatment effect) in the analysis stage? *See information sheet for further explanation if required*	Do you see any advantages of using informative priors in the analysis stage? Do you see any disadvantages of using informative priors in the analysis stage? What do you think about resource waste? Power a new trial based on the impact it can have on a meta-analysis? What do you think about using an objective prior (based on a synthesis of previous evidence)?
7	**Bias adjustment**	At the end of a trial are you always confident you can believe the results or do you think they may be at a risk of bias or many biases? E.g. if you couldn't blind the patients. So if we know that not blinding patients caused a 15% exaggeration in treatment effects in trials similar to yours would you adjust for it? How many trials do you do with adequate allocation concealment etc – do you think in examples where it may not be possible to do some of these things that the trial was less likely to get funded?
8	**Are there any parameters during the analysis which are poorly estimated or need more power?**	What types of parameters are usually poorly estimated? Do you know *why* they are poorly estimated? What do you do when parameters are poorly estimated? How do you report parameters which are poorly estimated? Did you find a solution to improve estimation? Do you think previous evidence could have helped? Would you use previous evidence if it helped estimation? Would you use informative priors on adverse events, when event rates are usually low?
9	**Would you use informative priors on parameters in the analysis of a trial?**	Do you know of any situations where using informative priors could be useful? Do you know of any situations where using informative priors may not be applicable? How are new methods implemented? Would permission have to be sought to use a new method? In these situations (treatment effect, other parameters, bias adjustment) would you use informative priors as part of a sensitivity analysis?

Closing

- Checks understanding of any outstanding points
- Thank them for their time
- Answer further questions, Ask who else we can speak to – contacts?

Date: **XX**
Version number: **v1.X**

(b) Page 2 of 2

Table A.2: Log of amendments to topic guide for statisticians.

Old version number	New version number	Changes
1.0	1.1	After the first two interviews, we added the question "What role do you think previous evidence has in a trial?"
1.1	1.2	Added the question: "What do you think about explicitly using prior information to inform a sample size calculation?" following recent work regarding the use of previous evidence in sample size calculations during the design of a new trial. Secondly, a potential use of existing evidence that came up was on missing data, so this was added as a prompt.
1.2	1.3	Access to the data and where you should get this from is an emerging theme which has come from discussing how these methods would be implemented in practice. As such the following question has been added: "How would you access data needed [prompt: would such data need to be collated by you or another colleague?]" Similarly, some participants have suggested it may be harder to publish Bayesian work and therefore the following question has been added: "How do you think publishing a Bayesian analysis with previous evidence will be received? [prompt: do you think it is more or less likely to get published]" Some participants mentioned that for this to be used more in practice it would have to become a requirement by funders and added to guidelines such as CONSORT. The following was added "What would make these methods be used in practice?"

Table A.3: COREQ checklist.

No	Item	Guide questions/ description	Comment
		Domain 1: Research team and reflexivity	
Personal characteristics			
1	Interviewer/facilitator	Which author/s conducted the interview?	GC
2	Credentials	What were the researcher's credentials? e.g. PhD, MD	MSc
3	Occupation	What was their occupation at the time of the study?	PhD Student in Trials Methodology
4	Gender	Was the researcher male or female?	Female
5	Experience and training	What experience or training did the researcher have?	GC has 2 years' experience working as a trial statistician at a clinical trials unit. GC has also been qualitative workshops and short courses. GC is also worked with an experienced qualitative researcher DE.
Relationship with participants			

Continued on next page

Table A.3 – continued from previous page

No	Item	Guide questions/description	Comment
6	Relationship established	Was a relationship established prior to study commencement?	Yes. Since GC has worked in clinical trials for 2 years prior to starting the PhD she was able to invite several people to interview who met the inclusion criteria.
7	Participant knowledge of the interviewer	What did the participants know about the researcher? e.g. personal goals, reasons for doing the research	GC introduced herself, explained the purpose of the research and provided an information leaflet about the study. GC did not go into detail about her PhD until after the interview so that her views were not pushed onto the participants.
8	Interviewer characteristics	What characteristics were reported about the interviewer/facilitator? e.g. Bias, assumptions, reasons and interests in the research topic	GC stated that all the answers given could be positive or negative regarding the use of previous evidence in trials.
		Domain 2: Study design	
Theoretical framework			

Continued on next page

Table A.3 – continued from previous page

No	Item	Guide questions/description	Comment
9	Methodological orientation and theory	What methodological orientation was stated to underpin the study? e.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis	Data were analysed thematically using techniques of constant comparison derived from grounded theory methodology.
Participant selection			
10	Sampling	How were participants selected? e.g. purposive, convenience, consecutive, snowball	Purposeful and snowball.
11	Method of approach	How were participants approached? e.g. face-to-face, telephone, mail, email	GC contacted researchers by email
12	Sample size	How many participants were in the study?	13 statisticians and 3 clinical academics
13	Non-participation	How many people refused to participate or dropped out? Reasons?	Of the people asked to participate, one person who was asked and said they would take part, then could not find time.
Setting			

Continued on next page

Table A.3 – continued from previous page

No	Item	Guide questions/description	Comment
14	Setting of data collection	Where was the data collected? e.g. home, clinic, workplace	Interviews were held at the participants workplace.
15	Presence of non-participants	Was anyone else present besides the participants and researchers?	No
16	Description of sample	What are the important characteristics of the sample? e.g. demographic data, date	Participants' full details are provided in Table 1, and key information is provided in the methods section
Data collection			

Continued on next page

Table A.3 – continued from previous page

No	Item	Guide questions/description	Comment
17	Interview guide	Were questions, prompts, guides provided by the authors? Was it pilot tested?	Topic guides were developed (based on the study aims and relevant literature along with DE) to ensure that discussions covered the same basic issues but with sufficient flexibility to allow new issues of importance to the informants to emerge. A pilot interview was conducted, however as all questions were deemed relevant, this was used as participant data. As analysis progressed, the topic guide adapted to enable exploration of emerging themes and a log of amendments was recorded. A separate topic guide was used for clinicians which included similar question but rephrased.
18	Repeat interviews	Were repeat interviews carried out? If yes, how many?	No repeat interviews were carried out

Continued on next page

Table A.3 – continued from previous page

No	Item	Guide questions/description	Comment
19	Audio/visual recording	Did the research use audio or visual recording to collect the data?	Interviews were audio-recorded.
20	Field notes	Were field notes made during and/or after the interview?	GC kept a few notes about each interview detailing the overall tone and the key points.
21	Duration	What was the duration of the interviews?	Interviews lasted an average of 54 minutes (range = 37 - 79 minutes).
22	Data saturation	Was data saturation discussed?	Data collection continued until GC and DE were confident that saturation had been reached.
23	Transcripts returned	Were transcripts returned to participants for comment and/or correction?	Transcripts were not returned to participants for comments or corrections
Domain 3: Analysis and findings			
Data analysis			

Continued on next page

Table A.3 – continued from previous page

No	Item	Guide questions/description	Comment
24	Number of data coders	How many data coders coded the data?	GC initially coded the data, and emerging themes were discussed with DE, with reference to the raw data and coding frame. Double coding between GC and HJ was used for three transcriptions in total, with reference to the raw data. In that time, we were happy that concerns in interpretation of coding frame was reached.
25	Description of the coding tree	Did authors provide a description of the coding tree?	A description of the coding tree is not provided in the article
26	Derivation of themes	Were themes identified in advance or derived from the data?	Themes were derived from the data
27	Software	What software, if applicable, was used to manage the data?	NVivo (version 11) was used to analyse the data
28	Participant checking	Did participants provide feedback on the findings?	Results were not sent out for response validation
Reporting			

Continued on next page

Table A.3 – continued from previous page

No	Item	Guide questions/description	Comment
29	Quotations presented	Were participant quotations presented to illustrate the themes / findings? Was each quotation identified? e.g. participant number	The interpretation of each theme is supported by illustrative quotes. Each quote is identified by a participant code
30	Data and findings consistent	Was there consistency between the data presented and the findings?	There is consistency between the data presented and the findings (see Table3.2)
31	Clarity of major themes	Were major themes clearly presented in the findings?	The themes are clearly presented in the findings
32	Clarity of minor themes	Is there a description of diverse cases or discussion of minor themes?	Description of diverse cases and where minor themes occurred between participant groups are discussed

Table A.4: Characteristics of included trials.

	(All) N=1153		(Ia) N=132		(Ib) N=95		(IIa) N=173		(IIb) N=91		(III) N=397	
	n	%	n	%	n	%	n	%	n	%	n	%
Blinding status												
Patients blind												
Definitely no	66	5.7	16	12.1	5	5.3	6	3.5	3	3.3	24	6
Definitely yes	170	14.7	15	11.4	24	25.3	40	23.1	29	31.9	38	9.6
Probably no	589	51.1	73	55.3	32	33.7	73	42.2	18	19.8	250	63
Probably yes	274	23.8	18	13.6	33	34.7	46	26.6	41	45.1	69	17.4
Unclear	54	4.7	10	7.6	1	1.1	8	4.6	0	0	16	4
Healthcare providers blinded												
Definitely no	94	8.2	20	15.2	8	8.4	8	4.6	8	8.8	32	8.1
Definitely yes	100	8.7	6	4.5	11	11.6	40	23.1	14	15.4	22	5.5
Probably no	591	51.3	78	59.1	37	38.9	64	37	21	23.1	248	62.5
Probably yes	312	27.1	21	15.9	37	38.9	53	30.6	47	51.6	79	19.9
Unclear	56	4.9	7	5.3	2	2.1	8	4.6	1	1.1	16	4
Outcome assessors blinded												
Definitely no	21	1.8	0	0	0	0	0	0	0	0	18	4.5

Continued on next page

Table A.4 – continued from previous page

	n	%	n	%	n	%	n	%	n	%	n	%
Definitely yes	202	17.5	0	0	76	80	0	0	54	59.3	128	32.2
Probably no	290	25.2	0	0	0	0	0	0	0	0	160	40.3
Probably yes	181	15.7	0	0	19	20	0	0	19	20.9	71	17.9
Unclear	38	3.3	0	0	0	0	0	0	0	0	20	5
N/A	421	36.5	132	100	0	0	173	100	18	19.8	0	0
Double-blind explicitly mentioned												
Yes	402	34.9	28	21.2	49	51.6	81	46.8	64	70.3	100	25.2
No	750	65	103	78	46	48.4	92	53.2	27	29.7	297	74.8
Unclear	1	0.1	1	0.8	0	0	0	0	0	0	0	0
All groups described as blinded/ double-blinded/ triple-blinded												
Yes	412	35.7	29	22	49	51.6	87	50.3	65	71.4	102	25.7
No	740	64.2	102	77.3	46	48.4	86	49.7	26	28.6	295	74.3
Unclear	1	0.1	1	0.8	0	0	0	0	0	0	0	0
Risk of bias												
Concealment of allocation												
High risk	127	11.1	15	11.5	19	20	9	5.2	11	12.1	67	16.9

Continued on next page

Table A.4 – continued from previous page

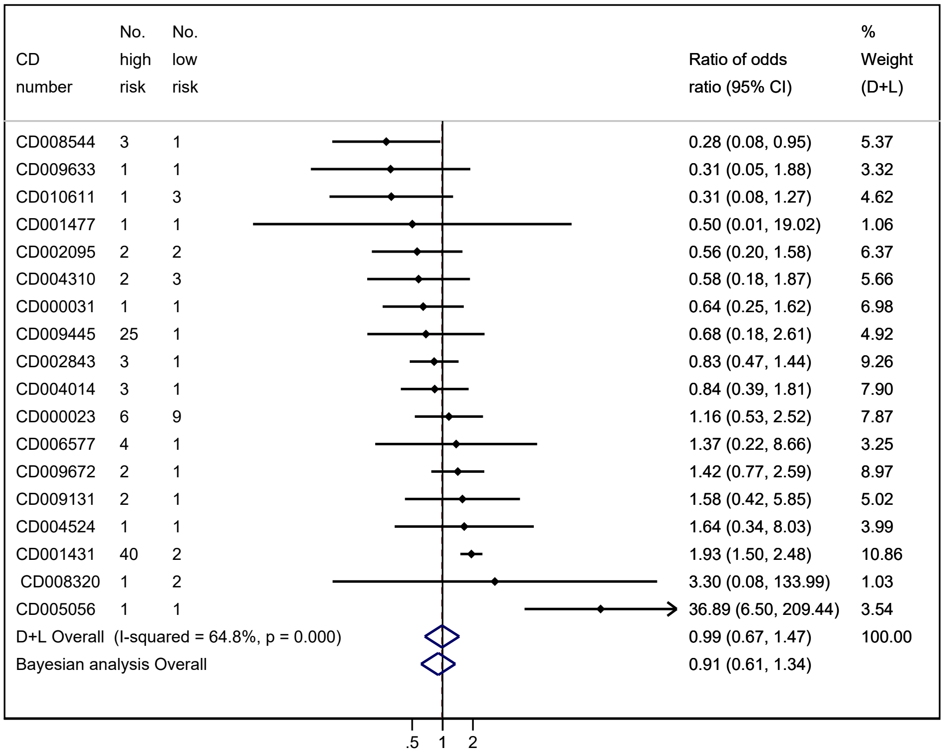
	n	%	n	%	n	%	n	%	n	%	n	%
Low risk	510	44.4	69	52.7	46	48.4	110	64	57	62.6	151	38.1
Unclear	512	44.6	47	35.9	30	31.6	53	30.8	23	25.3	178	44.9
Incomplete outcome data												
High risk	177	15.8	6	4.5	9	11.1	12	6.9	8	9.1	69	18.3
Low risk	771	68.7	103	78	59	72.8	143	82.7	69	78.4	228	60.3
Unclear	175	15.6	23	17.4	13	16	18	10.4	11	12.5	81	21.4
Drug trial*	753	65.3	48	36.4	77	81.1	127	73.4	81	89	205	51.6
Funding												
Profit organisations	251	21.8	16	12.1	29	30.5	40	23.1	36	39.6	61	15.4
Non- profit organisations	364	31.6	63	47.7	36	37.9	48	27.7	23	25.3	144	36.3
Both	108	9.4	10	7.6	5	5.3	13	7.5	10	11	36	9.1
Unclear	430	37.3	43	32.6	25	26.3	72	41.6	22	24.2	156	39.3
Trial design												
Cluster randomisation	20	1.7	2	1.5	0	0	0	0	0	0	12	3

Continued on next page

Table A.4 – continued from previous page

	n	%	n	%	n	%	n	%	n	%	n	%
Cross-over	7	0.6	0	0	0	0	0	0	0	0	4	1
Cross-over trial used as parallel group trial in meta-analysis	9	0.8	3	2.3	3	3.2	0	0	0	0	7	1.8
Parallel	1112	96.4	125	94.7	92	96.8	173	100	91	100	374	94.2
Split body	0	0	0	0	0	0	0	0	0	0	0	0
Unclear	5	0.4	2	1.5	0	0	0	0	0	0	0	0

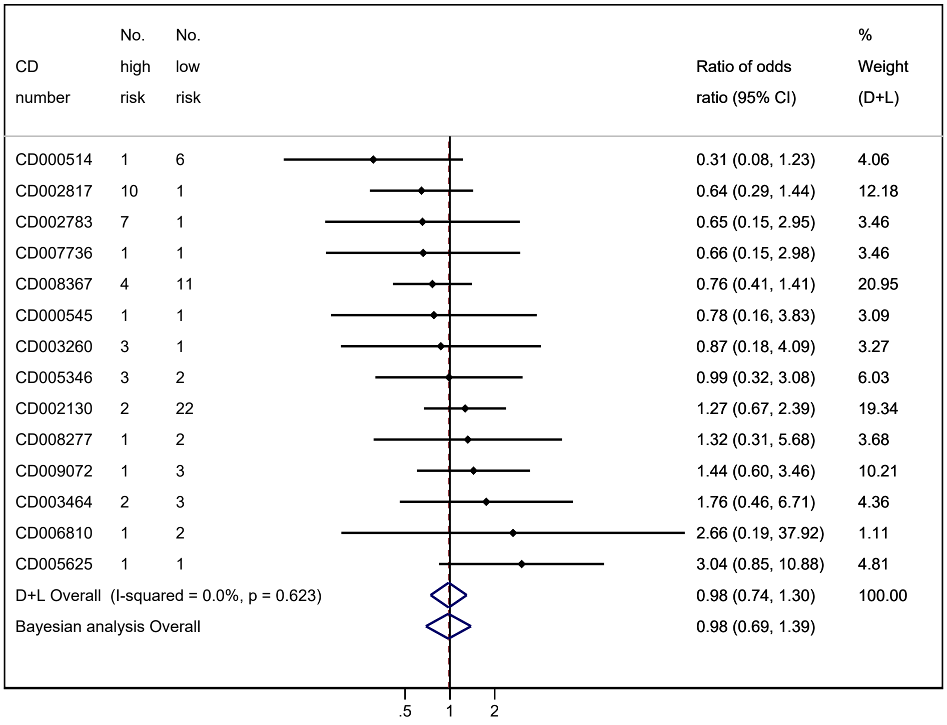
Figure A.5: RORs from individual meta-analyses and from analyses combined across all meta-analyses. Results for individual meta-analyses are frequentist estimates with confidence intervals, based on comparing the summary odds ratio from studies with the study characteristic of interest with the summary odds ratio from studies without the characteristic. The overall estimates of RORs are results based on the Bayesian hierarchical model described in Section 4.3.1. CD numbers are identifiers of individual Cochrane reviews, from the Cochrane Database of Systematic Reviews.



CD number	No. high risk	No. low risk	Ratio of odds ratio (95% CI)	% Weight (D+L)
CD008544	3	1	0.28 (0.08, 0.95)	5.37
CD009633	1	1	0.31 (0.05, 1.88)	3.32
CD010611	1	3	0.31 (0.08, 1.27)	4.62
CD001477	1	1	0.50 (0.01, 19.02)	1.06
CD002095	2	2	0.56 (0.20, 1.58)	6.37
CD004310	2	3	0.58 (0.18, 1.87)	5.66
CD000031	1	1	0.64 (0.25, 1.62)	6.98
CD009445	25	1	0.68 (0.18, 2.61)	4.92
CD002843	3	1	0.83 (0.47, 1.44)	9.26
CD004014	3	1	0.84 (0.39, 1.81)	7.90
CD000023	6	9	1.16 (0.53, 2.52)	7.87
CD006577	4	1	1.37 (0.22, 8.66)	3.25
CD009672	2	1	1.42 (0.77, 2.59)	8.97
CD009131	2	1	1.58 (0.42, 5.85)	5.02
CD004524	1	1	1.64 (0.34, 8.03)	3.99
CD001431	40	2	1.93 (1.50, 2.48)	10.86
CD008320	1	2	3.30 (0.08, 133.99)	1.03
CD005056	1	1	36.89 (6.50, 209.44)	3.54
D+L Overall (I-squared = 64.8%, p = 0.000)			0.99 (0.67, 1.47)	100.00
Bayesian analysis Overall			0.91 (0.61, 1.34)	

(a) (Ia) The effect of blinding patients in trials with patient-reported outcomes

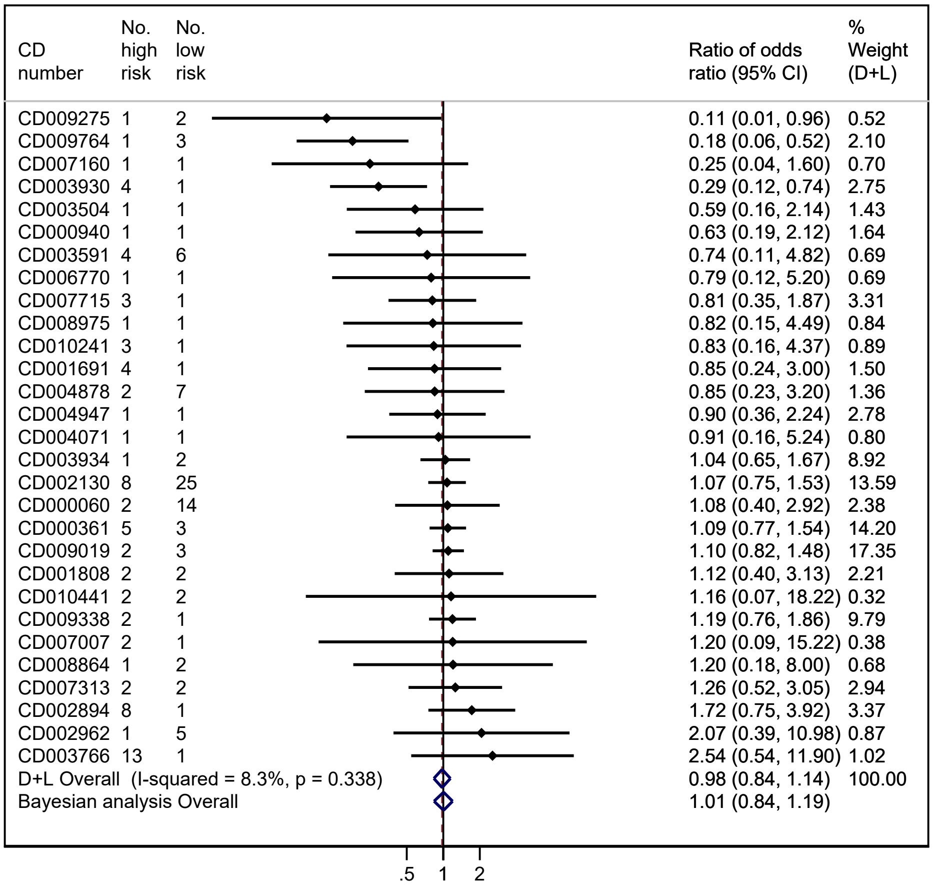
Figure A.5: (cont.)



CD number	No. high risk	No. low risk	Ratio of odds ratio (95% CI)	% Weight (D+L)
CD000514	1	6	0.31 (0.08, 1.23)	4.06
CD002817	10	1	0.64 (0.29, 1.44)	12.18
CD002783	7	1	0.65 (0.15, 2.95)	3.46
CD007736	1	1	0.66 (0.15, 2.98)	3.46
CD008367	4	11	0.76 (0.41, 1.41)	20.95
CD000545	1	1	0.78 (0.16, 3.83)	3.09
CD003260	3	1	0.87 (0.18, 4.09)	3.27
CD005346	3	2	0.99 (0.32, 3.08)	6.03
CD002130	2	22	1.27 (0.67, 2.39)	19.34
CD008277	1	2	1.32 (0.31, 5.68)	3.68
CD009072	1	3	1.44 (0.60, 3.46)	10.21
CD003464	2	3	1.76 (0.46, 6.71)	4.36
CD006810	1	2	2.66 (0.19, 37.92)	1.11
CD005625	1	1	3.04 (0.85, 10.88)	4.81
D+L Overall (I-squared = 0.0%, p = 0.623)			0.98 (0.74, 1.30)	100.00
Bayesian analysis Overall			0.98 (0.69, 1.39)	

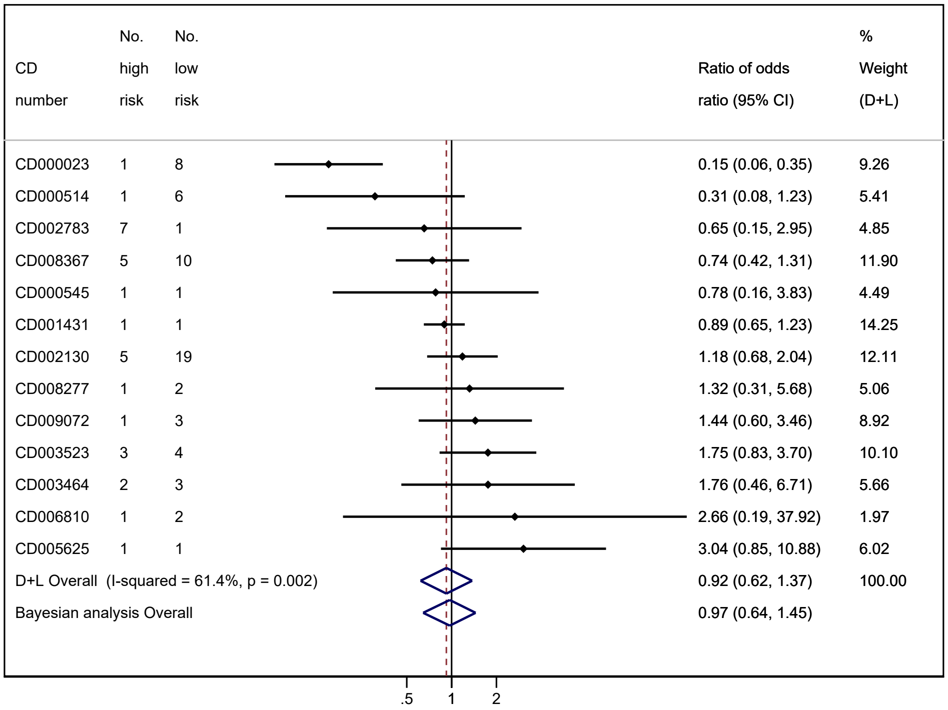
(b) (Ib) The effect of blinding patients in trials with blinded observer-reported outcomes

Figure A.5: (cont.)



(c) (IIa) The effect of blinding healthcare providers in trials with healthcare provider decision outcomes

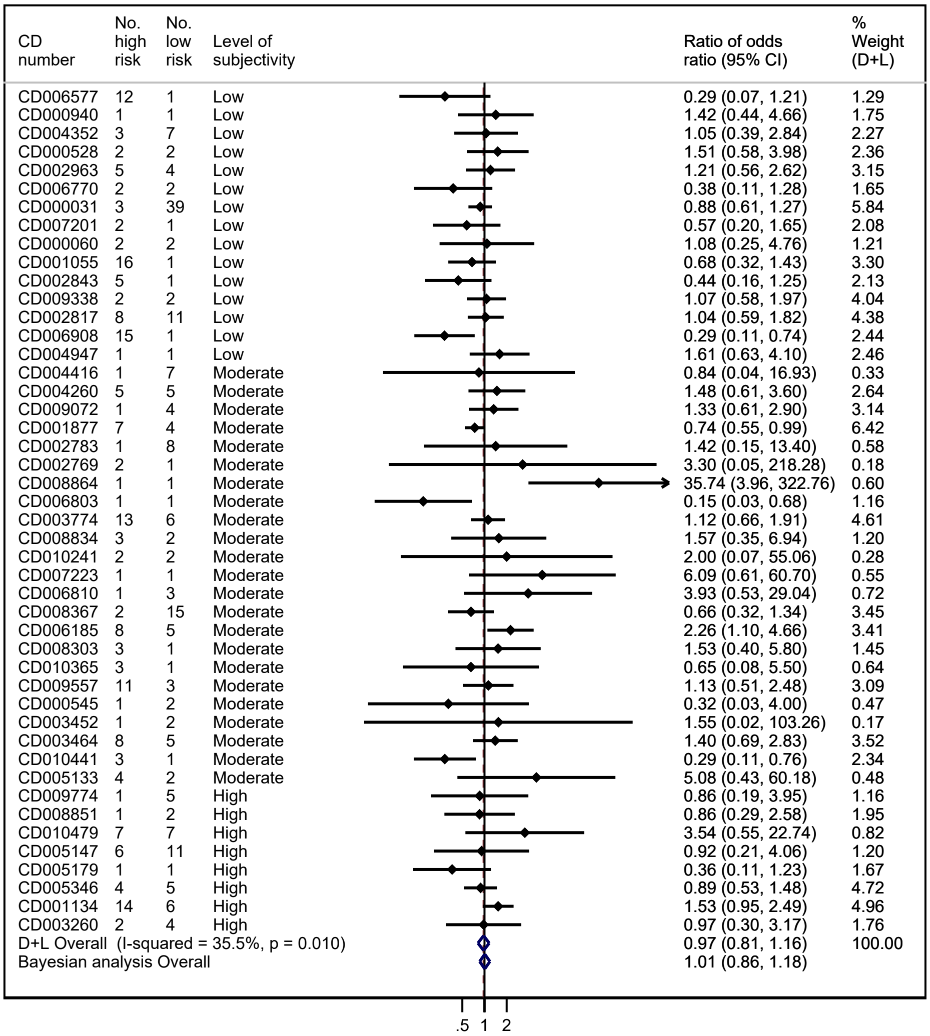
Figure A.5: (cont.)



CD number	No. high risk	No. low risk	Ratio of odds ratio (95% CI)	% Weight (D+L)
CD000023	1	8	0.15 (0.06, 0.35)	9.26
CD000514	1	6	0.31 (0.08, 1.23)	5.41
CD002783	7	1	0.65 (0.15, 2.95)	4.85
CD008367	5	10	0.74 (0.42, 1.31)	11.90
CD000545	1	1	0.78 (0.16, 3.83)	4.49
CD001431	1	1	0.89 (0.65, 1.23)	14.25
CD002130	5	19	1.18 (0.68, 2.04)	12.11
CD008277	1	2	1.32 (0.31, 5.68)	5.06
CD009072	1	3	1.44 (0.60, 3.46)	8.92
CD003523	3	4	1.75 (0.83, 3.70)	10.10
CD003464	2	3	1.76 (0.46, 6.71)	5.66
CD006810	1	2	2.66 (0.19, 37.92)	1.97
CD005625	1	1	3.04 (0.85, 10.88)	6.02
D+L Overall (I-squared = 61.4%, p = 0.002)			0.92 (0.62, 1.37)	100.00
Bayesian analysis Overall			0.97 (0.64, 1.45)	

(d) (IIb) The effect of blinding healthcare providers in trials with blinded observers/patients assessing the outcome

Figure A.5: (cont.)



CD number	No. high risk	No. low risk	Level of subjectivity	Ratio of odds ratio (95% CI)	% Weight (D+L)
CD006577	12	1	Low	0.29 (0.07, 1.21)	1.29
CD000940	1	1	Low	1.42 (0.44, 4.66)	1.75
CD004352	3	7	Low	1.05 (0.39, 2.84)	2.27
CD000528	2	2	Low	1.51 (0.58, 3.98)	2.36
CD002963	5	4	Low	1.21 (0.56, 2.62)	3.15
CD006770	2	2	Low	0.38 (0.11, 1.28)	1.65
CD000031	3	39	Low	0.88 (0.61, 1.27)	5.84
CD007201	2	1	Low	0.57 (0.20, 1.65)	2.08
CD000060	2	2	Low	1.08 (0.25, 4.76)	1.21
CD001055	16	1	Low	0.68 (0.32, 1.43)	3.30
CD002843	5	1	Low	0.44 (0.16, 1.25)	2.13
CD009338	2	2	Low	1.07 (0.58, 1.97)	4.04
CD002817	8	11	Low	1.04 (0.59, 1.82)	4.38
CD006908	15	1	Low	0.29 (0.11, 0.74)	2.44
CD004947	1	1	Low	1.61 (0.63, 4.10)	2.46
CD004416	1	7	Moderate	0.84 (0.04, 16.93)	0.33
CD004260	5	5	Moderate	1.48 (0.61, 3.60)	2.64
CD009072	1	4	Moderate	1.33 (0.61, 2.90)	3.14
CD001877	7	4	Moderate	0.74 (0.55, 0.99)	6.42
CD002783	1	8	Moderate	1.42 (0.15, 13.40)	0.58
CD002769	2	1	Moderate	3.30 (0.05, 218.28)	0.18
CD008864	1	1	Moderate	35.74 (3.96, 322.76)	0.60
CD006803	1	1	Moderate	0.15 (0.03, 0.68)	1.16
CD003774	13	6	Moderate	1.12 (0.66, 1.91)	4.61
CD008834	3	2	Moderate	1.57 (0.35, 6.94)	1.20
CD010241	2	2	Moderate	2.00 (0.07, 55.06)	0.28
CD007223	1	1	Moderate	6.09 (0.61, 60.70)	0.55
CD006810	1	3	Moderate	3.93 (0.53, 29.04)	0.72
CD008367	2	15	Moderate	0.66 (0.32, 1.34)	3.45
CD006185	8	5	Moderate	2.26 (1.10, 4.66)	3.41
CD008303	3	1	Moderate	1.53 (0.40, 5.80)	1.45
CD010365	3	1	Moderate	0.65 (0.08, 5.50)	0.64
CD009557	11	3	Moderate	1.13 (0.51, 2.48)	3.09
CD000545	1	2	Moderate	0.32 (0.03, 4.00)	0.47
CD003452	1	2	Moderate	1.55 (0.02, 103.26)	0.17
CD003464	8	5	Moderate	1.40 (0.69, 2.83)	3.52
CD010441	3	1	Moderate	0.29 (0.11, 0.76)	2.34
CD005133	4	2	Moderate	5.08 (0.43, 60.18)	0.48
CD009774	1	5	High	0.86 (0.19, 3.95)	1.16
CD008851	1	2	High	0.86 (0.29, 2.58)	1.95
CD010479	7	7	High	3.54 (0.55, 22.74)	0.82
CD005147	6	11	High	0.92 (0.21, 4.06)	1.20
CD005179	1	1	High	0.36 (0.11, 1.23)	1.67
CD005346	4	5	High	0.89 (0.53, 1.48)	4.72
CD001134	14	6	High	1.53 (0.95, 2.49)	4.96
CD003260	2	4	High	0.97 (0.30, 3.17)	1.76
D+L Overall (I-squared = 35.5%, p = 0.010)				0.97 (0.81, 1.16)	100.00
Bayesian analysis Overall				1.01 (0.86, 1.18)	

(e) (III) The effect of blinding outcome assessors (i.e. observers) in trials with subjective outcomes

Table A.5: Associations between reported study characteristics

Study characteristic 1	**Study characteristic 2**	**All trials (n, %)**	**(Ia) (n, %)**	**(Ib) (n, %)**	**(IIa) (n, %)**	**(IIb) (n, %)**	**(III) (n, %)**
Patients	**Healthcare provider**						
Blinded	Blinded	399 (34.6)	26 (19.7)	48 (50.5)	84 (48.6)	61 (67.0)	99 (24.9)
Blinded	Non-blinded	45 (3.9)	7 (5.3)	9 (9.5)	2 (1.2)	9 (9.9)	8 (2.0)
Non-blinded	Blinded	13 (1.1)	1 (0.8)	0 (0)	9 (5.2)	0 (0)	2 (0.5)
Non-blinded	Non-blinded	696 (60.4)	98 (74.2)	38 (40.0)	78 (45.1)	21 (23.1)	288 (72.5)
OR		474.7	364.0	388.0	364.0	271.9	1782.0
95% CI		(253.0, 890.7)	(42.8, 3092.2)	(21.9, 6880.2)	(76.3, 1737.2)	(15.2, 4876.1)	(372.1, 8533.4)
Patients	**Outcome assessor**						
Blinded	Blinded	264 (35.9)	0 (0)	57 (60.0)	0 (0)	52 (71.2)	103 (25.9)
Blinded	Non-blinded	10 (1.4)	0 (0)	0 (0)	0 (0)	0 (0)	4 (1.0)
Non-blinded	Blinded	120 (16.3)	0 (0)	38 (40.0)	0 (0)	21 (28.8)	96 (24.2)
Non-blinded	Non-blinded	341 (46.4)	0 (0)	0 (0)	0 (0)	0 (0)	194 (48.9)
OR		75.0		1.5		2.4	52.0
95% CI		(38.6,145.8)		(0.0, 76.9)		(0.0, 127.1)	(18.6,145.5)

Continued on next page

Table A.5 – continued from previous page

Study characteristic 1	Study characteristic 2	All trials (n, %)	(Ia) (n, %)	(Ib) (n, %)	(IIa) (n, %)	(IIb) (n, %)	(III) (n, %)
Patients	**Allocation concealment**						
Blinded	Yes	261 (22.7)	23 (17.4)	38 (40.0)	68 (39.3)	52 (57.1)	56 (14.1)
Blinded	No	183 (15.9)	10 (7.6)	19 (20.0)	18 (10.4)	18 (19.8)	51 (12.8)
Non-blinded	Yes	249 (21.7)	46 (34.8)	8 (8.4)	42 (24.3)	5 (5.5)	95 (23.9)
Non-blinded	No	456 (39.7)	53 (40.2)	30 (31.6)	45 (26.0)	16 (17.6)	195 (49.1)
OR		2.6	2.7	7.5	4.0	9.2	2.3
95% CI		(2.0, 3.3)	(1.1, 6.1)	(2.9, 19.5)	(2.1, 7.9)	(3.0, 28.9)	(1.4, 3.5)
Patients	**Incomplete outcome data**						
Blinded	Complete	338 (30.1)	28 (21.2)	44 (46.3)	74 (42.8)	56 (61.5)	78 (19.6)
Blinded	Incomplete	100 (8.9)	5 (3.8)	13 (13.7)	12 (6.9)	14 (15.4)	29 (7.3)
Non-blinded	Complete	433 (38.6)	75 (56.8)	15 (15.8)	69 (39.9)	13 (14.3)	150 (37.8)
Non-blinded	Incomplete	252 (22.4)	24 (18.2)	23 (24.2)	18 (10.4)	8 (8.8)	140 (35.3)

Continued on next page

Table A.5 – continued from previous page

Study characteristic 1	Study characteristic 2	All trials (n, %)	(Ia) (n, %)	(Ib) (n, %)	(IIa) (n, %)	(IIb) (n, %)	(III) (n, %)
OR		2.0	1.8	5.2	1.6	2.5	2.5
95% CI		(1.5, 2.6)	(0.6, 5.2)	(2.1, 12.7)	(0.7, 3.6)	(0.9, 7.1)	(1.5, 4.1)
Healthcare provider	**Outcome assessor**						
Blinded	Blinded	248 (33.7)	0 (0)	48 (50.5)	0 (0)	48 (65.8)	100 (25.2)
Blinded	Non-blinded	1 (0.1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.3)
Non-blinded	Blinded	136 (18.5)	0 (0)	47 (49.5)	0 (0)	25 (34.2)	99 (24.9)
Non-blinded	Non-blinded	350 (47.6)	0 (0)	0 (0)	0 (0)	0 (0)	197 (49.6)
OR		638.2		1.0		1.9	199.0
95% CI		(88.7, 4594.5)		(0.0, 52.5)		(0.0, 98.7)	(27.4, 1447.8)
Healthcare provider	**Allocation concealment**						
Blinded	Yes	243 (21.1)	16 (12.1)	33 (34.7)	74 (42.8)	45 (49.5)	54 (13.6)
Blinded	No	168 (14.6)	11 (8.3)	15 (15.8)	19 (11.0)	16 (17.6)	47 (11.8)
Non-blinded	Yes	267 (23.2)	53 (40.2)	13 (13.7)	36 (20.8)	12 (13.2)	97 (24.4)

Continued on next page

Table A.5 – continued from previous page

Study characteristic 1	Study characteristic 2	All trials (n, %)	(Ia) (n, %)	(Ib) (n, %)	(IIa) (n, %)	(IIb) (n, %)	(III) (n, %)
Non-blinded	No	471 (41.0)	52 (39.4)	34 (35.8)	44 (25.4)	18 (19.8)	199 (50.1)
OR		2.6	1.4	5.8	4.8	4.2	2.4
95% CI		(2.0, 3.3)	(0.6, 3.4)	(2.4, 13.9)	(2.4, 9.3)	(1.7, 10.7)	(1.5, 3.7)
Healthcare provider	**Incomplete outcome data**						
Blinded	Complete	319 (28.4)	23 (17.4)	40 (42.1)	81 (46.8)	49 (53.8)	77 (19.4)
Blinded	Incomplete	88 (7.8)	4 (3.0)	8 (8.4)	12 (6.9)	12 (13.2)	24 (6.0)
Non-blinded	Complete	452 (40.2)	80 (60.6)	19 (20.0)	62 (35.8)	20 (22.0)	151 (38.0)
Non-blinded	Incomplete	264 (23.5)	25 (18.9)	28 (29.5)	18 (10.4)	10 (11.0)	145 (36.5)
OR		2.1	1.8	7.4	2.0	2.0	3.1
95% CI		(1.6, 2.8)	(0.6, 5.7)	(2.8, 19.2)	(0.9, 4.4)	(0.8, 5.5)	(1.8, 5.1)
Outcome assessor	**Allocation concealment**						
Blinded	Yes	200 (27.2)	0 (0)	46 (48.4)	0 (0)	41 (56.2)	99 (24.9)
Blinded	No	184 (25.1)	0 (0)	49 (51.6)	0 (0)	32 (43.8)	100 (25.2)

Continued on next page

Table A.5 – continued from previous page

Study characteristic 1	Study characteristic 2	All trials (n, %)	(Ia) (n, %)	(Ib) (n, %)	(IIa) (n, %)	(IIb) (n, %)	(III) (n, %)
Non-blinded	Yes	94 (12.8)	0 (0)	0 (0)	0 (0)	0 (0)	52 (13.1)
Non-blinded	No	256 (34.9)	0 (0)	0 (0)	0 (0)	0 (0)	146 (36.8)
OR		3.0		0.9		1.3	2.8
95% CI		(2.2, 4.0)		(0.0, 48.3)		(0.0, 66.1)	(1.8, 4.2)
Outcome assessor	**Incomplete outcome data**						
Blinded	Complete	267 (37.7)	0 (0)	59 (62.1)	0 (0)	55 (75.3)	126 (31.7)
Blinded	Incomplete	103 (14.5)	0 (0)	36 (37.9)	0 (0)	18 (24.7)	73 (18.4)
Non-blinded	Complete	190 (26.8)	0 (0)	0 (0)	0 (0)	0 (0)	102 (25.7)
Non-blinded	Incomplete	149 (21.0)	0 (0)	0 (0)	0 (0)	0 (0)	96 (24.2)
OR		2.0		1.6		3.0	1.6
95% CI		(1.5, 2.8)		(0.0, 84.0)		(0.1, 156.6)	(1.1, 2.4)

B WinBUGS code

WinBUGS code: Combining binary and continuous outcomes

```
model {
for (i in 1:Nb) {
    rc[i] ~ dbin(pc[i],nc[i])
   rt[i] ~ dbin(pt[i],nt[i])
logit(pc[i]) <- mu[i]  # model for binary outcomes (logit link)
logit(pt[i]) <- mu[i] + delta[i]  + beta[i]*C[i]
 mu[i] ~ dnorm(0,.0001)
}

for (i in Nb+1:Nc+Nb) {
var[i] <- pow(se[i],2)   # calculate variances
prec.smd[i] <- 1/var[i]     # set precisions
# likelihood for continuous outcomes on log odds ratio scale
   lnor[i] ~ dnorm(nu[i],prec.smd[i])
# model for continuous outcomes (identity link)
  nu[i] <- delta[i] + beta[i]*C[i]
}

for (i in 1:Nc+Nb) {
# between study, within MA, variation in bias
 beta[i]~dnorm(b[ma[i]],p.k2[ma[i]])I(-10,10)
#RE for treatment effect within meta-analysis
   delta[i]~dnorm(d[ma[i]],p.d[ma[i]])I(-10,10)
}

for (m in 1:N_ma) {
# priors for true fixed (unrelated) treatment effects
d[m] ~ dnorm(0,.0001)
#between meta-analysis variation in mean bias
         b[m] ~ dnorm(b0,p.phi)
p.d1[m]~dgamma(.001,.001)
p.d[m]<-p.d1[m]/(1-patom.d[m])
patom.d[m]~dbeta(1,1)
}
  # vague prior for overall mean bias
b0 ~ dnorm(0,.0001)

p.k1~dgamma(.001,.001)
kappa <- pow(p.k,-0.5)
p.k<-p.k1/(1-patom.k)
patom.k~dbeta(1,1)
for (m in 1:N_kappa_ok){
p.k2[kappa_ok[m]]<-p.k
}
for (m in 1:N_kappa_cut){
p.k2[kappa_cut[m]]<- cut(p.k)
}

p.phi1~dgamma(.001,.001)
```

```
phi <- pow(p.phi,-0.5)
p.phi<-p.phi1/(1-patom.phi)
patom.phi~dbeta(1,1)
#predictive distn for mean bias in new meta-analysis
b.new~dnorm(b0,p.phi)
 #predictive distn for bias in new study in a new meta-analysis
beta.new~dnorm(b.new,p.k)
lkappa<-log(kappa)
lphi<-log(phi)
dum<-s[1]
}
```

WinBUGS code: Stratifying the average magnitude of bias at the meta-analysis level

```
model {
for (i in 1:Nb) {
    rc[i] ~ dbin(pc[i],nc[i])
   rt[i] ~ dbin(pt[i],nt[i])
logit(pc[i]) <- mu[i]
logit(pt[i]) <- mu[i] + delta[i]  + beta[i]*C[i]
 mu[i] ~ dnorm(0,.0001)
}

for (i in Nb+1:Nc+Nb) {
var[i] <- pow(se[i],2)   # calculate variances
prec.smd[i] <- 1/var[i]     # set precisions
   lnor[i] ~ dnorm(nu[i],prec.smd[i]) # likelihood
  nu[i] <- delta[i]  + beta[i]*C[i] # model
}

for (i in 1:Nc+Nb) {
 beta[i]~dnorm(b[ma[i]],p.k2[ma[i]])I(-10,10)
   delta[i]~dnorm(d[ma[i]],p.d[ma[i]])I(-10,10)
}

for (m in 1:N_ma) {
d[m] ~ dnorm(0,.01)
        b[m]<-b0l[m]+b1*dum2[m]  + b2*dum3[m]
        b0l[m] ~ dnorm(b0,p.phi)
p.d1[m]~dgamma(.001,.001)
p.d[m]<-p.d1[m]/(1-patom.d[m])
patom.d[m]~dbeta(1,1)
}
# vague prior for overall mean bias at each stratification level
b0 ~ dnorm(0,.0001)
b1 ~ dnorm(0,.0001)
b2 ~ dnorm(0,.0001)

p.k1~dgamma(.001,.001)
kappa <- pow(p.k,-0.5)
p.k<-p.k1/(1-patom.k)
patom.k~dbeta(1,1)
```

```
for (m in 1:N_kappa_ok){
p.k2[kappa_ok[m]]<-p.k
}
for (m in 1:N_kappa_cut){
p.k2[kappa_cut[m]]<- cut(p.k)
}

p.phi1~dgamma(.001,.001)
phi <- pow(p.phi,-0.5)
p.phi<-p.phi1/(1-patom.phi)
patom.phi~dbeta(1,1)
#predictive distn for mean bias in new meta-analysis
b.new~dnorm(b0,p.phi)
#predictive distn for bias in new study in new meta-analysis
beta.new~dnorm(b.new,p.k)
lkappa<-log(kappa)
lphi<-log(phi)

dum<-s[1]
}
```

WinBUGS code: Product normal model

```
model {
for (i in 1:N) {
rc[i] ~ dbin(pc[i],nc[i])
```

```
   rt[i] ~ dbin(pt[i],nt[i])
logit(pc[i]) <- mu[i]
logit(pt[i]) <-mu[i] +delta[i] + beta1[i]*iu[i]+ beta2[i]*ih[i]
delta[i]~dnorm(low[ma[i]],p.tau[ma[i]])I(-10,10)
mu[i] ~ dnorm(0,.001)
beta1[i]~dnorm(b1[ma[i]], p.k21[ma[i]])I(-10,10)
beta2[i]~dnorm(b.re[i], p.k22[ma[i]])I(-10,10)
b.re[i]<-b2[ma[i]]+rho.kappa*(kappa2/kappa1)*(beta1[i]-b1[ma[i]])
}
for (m in 1:N_ma) {
low[m] ~ dnorm(0,.0001)

p.tau1[m]~dgamma(.001,.001)
p.tau[m]<-p.tau1[m]/(1-patom.tau[m])
patom.tau[m]~dbeta(1,1)
sd.tau[m] <- pow(p.tau[m],-0.5)
b1[m]~dnorm(b01,p.phi1)
b2[m]~dnorm(b0.re[m], p.phi.re)
b0.re[m]<-b02+rho.phi*(phi2/phi1)*(b1[m]-b01)

}

b01 ~ dnorm(0,.0001)
  b02 ~ dnorm(0,.0001)
rho.kappa~dunif(-1,1)
condvar.beta2 <- (1 - pow(rho.kappa, 2))*pow(kappa2, 2)
p.kre <- 1/condvar.beta2
```

```
kappa2 <- pow(p.k2,-0.5)
p.kz2~dgamma(.001,.001)
p.k2<-p.kz2/(1-patom.k2)
patom.k2~dbeta(1,1)
for (m in 1:N_kappa_ih_ok){
p.k22[kappa_ih_ok[m]]<-p.kre
}
for (m in 1:N_kappa_ih_cut){
p.k22[kappa_ih_cut[m]]<- cut(p.kre)
}
#KAPPA
#unclears
p.kz1~dgamma(.001,.001)
kappa1 <- pow(p.k1,-0.5)
kappa1.sq<-kappa1*kappa1
p.k1<-p.kz1/(1-patom.k1)
patom.k1~dbeta(1,1)
for (m in 1:N_kappa_iu_ok){
p.k21[kappa_iu_ok[m]]<-p.k1
}
for (m in 1:N_kappa_iu_cut){
p.k21[kappa_iu_cut[m]]<- cut(p.k1)
}
#high risk
#link between unclears and high
phiv.re<-phi2.sq*(1-pow(rho.phi,2))
```

```
p.phi.re<-1/phiv.re

rho.phi~dunif(-1,1)

#PHI
#unclears
p.phiz1~dgamma(.001,.001)
phi1 <- pow(p.phi1,-0.5)
phi1.sq<-phi1*phi1
p.phi1<-p.phiz1/(1-patom.phi1)
patom.phi1~dbeta(1,1)

#high risk
p.phiz2~dgamma(.001,.001)
phi2 <- pow(p.phi2,-0.5)
phi2.sq<-phi2*phi2
p.phi2<-p.phiz2/(1-patom.phi2)
patom.phi2~dbeta(1,1)

#covariances
cov.kappa<-kappa1*-kappa2*rho.kappa
cov.phi<-phi1*phi2*rho.phi

B01<-exp(b01)
B02<-exp(b02)
dum1<-s[1]
dum2<-CDnumber[1]
```

```
}
```

WinBUGS code: Probability model

```
model {
for (i in 1:N) {
rc[i] ~ dbin(pc[i],nc[i])
   rt[i] ~ dbin(pt[i],nt[i])
B[i] ~ dbern(p)
logit(pc[i]) <- mu[i]
logit(pt[i]) <-mu[i] + delta[i]  + beta[i]*(ih[i] + iu[i]*B[i])
mu[i] ~ dnorm(0,.0001)
 beta[i]~dnorm(b[ma[i]],p.k2[ma[i]])I(-10,10)
 delta[i]~dnorm(d[ma[i]],p.tau[ma[i]])I(-10,10)

}
p ~ dunif(0,1)
for (m in 1:N_ma) {
# priors for true fixed (unrelated) treatment effects
d[m] ~ dnorm(0,.0001)
#between meta-analysis variation in mean bias
b[m] ~ dnorm(b0,p.phi)
p.tau1[m]~dgamma(.001,.001)
p.tau[m]<-p.tau1[m]/(1-patom.tau[m])
patom.tau[m]~dbeta(1,1)
sd.tau[m] <- pow(p.tau[m], -0.5)
}
# vague prior for overall mean bias
b0 ~ dnorm(0,.0001)
B0<- exp(b0)
p.k1~dgamma(.001,.001)
kappa <- pow(p.k,-0.5)
p.k<-p.k1/(1-patom.k)
patom.k~dbeta(1,1)
for (m in 1:N_kappa_ih_ok){
p.k2[kappa_ih_ok[m]]<-p.k1
}
for (m in 1:N_kappa_ih_cut){
p.k2[kappa_ih_cut[m]]<- cut(p.k1)
}
p.phi1~dgamma(.001,.001)
phi <- pow(p.phi,-0.5)
p.phi<-p.phi1/(1-patom.phi)
patom.phi~dbeta(1,1)
dum1<-s[1]
dum2<-CDnumber[1]
}
```

Example data:

```
CDnumber[] ma[] s[] ih[] iu[] rt[] rt[] nt[] rc[] nc[]
8603 1 1 0 1 11 11 102 13 104
8603 1 2 0 1 7 7 163 5 84
8603 1 3 0 1 9 9 156 6 156
8603 1 4 0 0 30 30 538 27 539
```

```
8603 1 5 0 1 5 5 124 2 16

8603 1 6 0 1 3 3 119 2 79

8603 1 7 0 1 9 9 190 2 82

8603 1 8 0 1 0 0 40 1 41

8603 1 9 0 1 2 2 25 2 25

8603 1 10 0 1 2 2 178 1 171

8603 1 11 0 1 18 18 155 12 148

8603 1 12 0 0 2 2 43 1 42

7228 2 13 1 0 3 3 62 4 65
```

WinBUGS code: Probability model with sample size

```
model {

for (i in 1:N) {

rc[i] ~ dbin(pc[i],nc[i])

   rt[i] ~ dbin(pt[i],nt[i])

n[i]<-nt[i]+nc[i]

x[i] <- log(n[i])

B[i] ~ dbern(p[i])

logit(p[i]) <- m*(x[i] - mean(x[]))+ c

logit(pc[i]) <- mu[i]

logit(pt[i]) <-mu[i] + delta[i]  + beta[i]*(ih[i] + iu[i]*B[i])

  mu[i] ~ dnorm(0,.0001)

 beta[i]~dnorm(b[ma[i]],p.k2[ma[i]])I(-10,10)

    delta[i]~dnorm(d[ma[i]],p.tau[ma[i]])I(-10,10)


}
```

```
for(j in 1:M){
high[j] ~ dbern(probhigh[j])
 xobs[j] <- log(nobs[j])
logit(probhigh[j]) <- m*(xobs[j]-mean(xobs[]))+ alpha
}
alpha ~ dnorm(0,.0001)
c~dnorm(0,.0001)
m~ dnorm(0,.0001)
for (m in 1:N_ma) {
d[m] ~ dnorm(0,.0001)
# priors for true fixed (unrelated) treatment effects
        b[m] ~ dnorm(b0,p.phi)
#between meta-analysis variation in mean bias
p.tau1[m]~dgamma(.001,.001)
p.tau[m]<-p.tau1[m]/(1-patom.tau[m])
patom.tau[m]~dbeta(1,1)
sd.tau[m] <- pow(p.tau[m], -0.5)
}
b0 ~ dnorm(0,.0001)
# vague prior for overall mean bias
B0<- exp(b0)
p.k1~dgamma(.001,.001)
kappa <- pow(p.k,-0.5)
p.k<-p.k1/(1-patom.k)
patom.k~dbeta(1,1)
for (m in 1:N_kappa_ih_ok){
p.k2[kappa_ih_ok[m]]<-p.k1
}
for (m in 1:N_kappa_ih_cut){
p.k2[kappa_ih_cut[m]]<- cut(p.k1)
}
p.phi1~dgamma(.001,.001)
phi <- pow(p.phi,-0.5)
p.phi<-p.phi1/(1-patom.phi)
patom.phi~dbeta(1,1)
dum1<-s[1]
dum2<-CDnumber[1]
}
```

Example data:

```
CDnumber[] ma[] s[] ih[] iu[] rt[] rt[] nt[] rc[] nc[] highrisk[]
8603 1 1 0 1 11 11 102 13 104 NA
8603 1 2 0 1 7 7 163 5 84 NA
8603 1 3 0 1 9 9 156 6 156 NA
8603 1 4 0 0 30 30 538 27 539 0
8603 1 5 0 1 5 5 124 2 16 NA
8603 1 6 0 1 3 3 119 2 79 NA
8603 1 7 0 1 9 9 190 2 82 NA
8603 1 8 0 1 0 0 40 1 41 NA
8603 1 9 0 1 2 2 25 2 25 NA
8603 1 10 0 1 2 2 178 1 171 NA
8603 1 11 0 1 18 18 155 12 148 NA
8603 1 12 0 0 2 2 43 1 42 0
```

```
7228 2 13 1 0 3 3 62 4 65 1
```

```
list(M=1241)

high[] nobs[]

0 1077

0 85

1 127

0 1518

0 181
```

WinBUGS code: Meta-epidemiological model wth continuous study characteristic

```
model {
for (i in 1:N) {
rc[i] ~ dbin(pc[i],nc[i])
   rt[i] ~ dbin(pt[i],nt[i])
logit(pc[i]) <- mu[i]
logit(pt[i]) <- mu[i] + delta[i]
c[i] <- 1/sqrt(n[i])
delta[i]~dnorm(new.m[i],new.p[i])I(-10,10)
new.m[i]<- d[ma[i]]+b[ma[i]]*(c[i]- mean(c[]))
new.var[i]<-pow(tau[ma[i]], 2) + kappa.sq*(c[i])
new.p[i]<-1/new.var[i]

mu[i] ~ dnorm(0,.0001)
rhat.c[i] <-pc[i]*nc[i] #calculate residual deviance
rhat.t[i]<-pt[i]*nt[i]
dev.t[i] <- 2 * (rt[i] * (log(rt[i])-log(rhat.t[i]))  +
(nt[i]-rt[i]) * (log(nt[i]-rt[i]) - log(nt[i]-rhat.t[i])))
dev.c[i] <- 2 * (rc[i] * (log(rc[i])-log(rhat.c[i]))  +
(nc[i]-rc[i]) * (log(nc[i]-rc[i]) - log(nc[i]-rhat.c[i])))
}
resdev <-sum(dev.t[])+sum(dev.c[])
for (m in 1:N_ma) {
d[m] ~ dnorm(0,.0001)
        b[m] ~ dnorm(b0,p.phi)
   log.tau[m] ~ dnorm(mean.lt, prec.lt)I(-5,5)
log(tau[m]) <- log.tau[m]
p.d[m] <- pow(tau[m], -2)

}

b0 ~ dnorm(0,.0001)
      mean.lt ~ dnorm(0, 0.0001)
sd.lt ~ dunif(0, 2)
prec.lt <- pow(sd.lt, -2)

sd.k ~ dunif(0, 5)
p.k<- pow(sd.k, -2)
kappa.sq<-sd.k*sd.k

p.phi1~dgamma(.001,.001)
phi <- pow(p.phi,-0.5)
```

```
p.phi<-p.phi1/(1-patom.phi)

patom.phi~dbeta(1,1)

dum1<-s[1]

log.tau2~dnorm(mean.lt, prec.lt)

B0<-exp(b0)

}
```