Borrowing strength from ‘indirect’ evidence:

Methods and policy implications for Health Technology Assessment

Georgios Nikolaidis

A thesis presented for the degree of
Doctor of Philosophy

University of York
Department of Health Sciences

July 2020

- *“Atticus told me to delete the adjectives and I’d have the facts”*

Jean Louise (Scout) Finch
To Kill a Mockingbird
by Harper Lee

Abstract

Sparse relative effectiveness evidence is a frequent problem in Health Technology Assessment (HTA). For example, evidence on a particular comparator or randomised evidence in a specific population (e.g. paediatric) may be limited. Where evidence directly pertaining the decision problem is sparse, one could expand the evidence base to include studies relating to the decision problem only indirectly: for instance, when there is no evidence on a specific comparator, evidence from other treatments of the same molecular class could be used; similarly, a decision on children may borrow strength from evidence on adults. Usually, in HTA, such indirect evidence is either included by ignoring any differences ('lumping') or is not allowed to influence the decision ('splitting'). However, more sophisticated methods exist in the literature which, rather than lumping or splitting, borrow strength from the indirect evidence by imposing more moderate, and perhaps more appropriate, degrees of information-sharing.

This thesis commences with a systematic review that sought to identify methods to combine evidence directly and indirectly relating to a research question. A classification of Information-sharing methods (ISMs) according to the main assumption employed to facilitate information-sharing is proposed. Subsequently, detailed descriptions of methods' assumptions and implementation suggestions are provided in the context of a specific synthesis problem. To aid transparency in selecting ISMs, a step-by-step approach that could be useful for HTA analysts and policy-makers is proposed. Then, all applicable methods are used to borrow strength from indirect evidence on relative effectiveness in a case-study. Findings imply that the choice of method can affect how much strength is borrowed, impact relative effectiveness estimates, and influence adoption and research recommendation decisions. Then, the strength of information-sharing imposed by the various methods is investigated using probabilistic scenarios. Finally, lessons learned throughout the thesis are distilled into a set of recommendations for HTA practice.

Contents

List of Tables

List of Figures

Acknowledgements

I am eternally thankful to my primary supervisor, Dr. Marta Soares for her invaluable advice, continuous help, and insightful guidance throughout the whole process of this thesis. I would also like to thank my two other supervisors, Mrs. Beth Woods for devoting a substantial amount of time to help me with both general and technical aspects of this work as well as Prof. Stephen Palmer for his astute advice to ensure that this work is relevant to the policy-context and remains useful for decision-makers. Also, I am deeply grateful to my Thesis Advisory Panel (TAP), Dr. Sylwia Bujkiewicz for devoting time, both in TAP and other meetings, to bring in her statistical expertise and provide me with useful comments and suggestions, as well as Dr. Mona Kanaan for overseeing the overall process and offering important statistical advice on all matters.

I would also like to thank all staff and students in the Centre for Health Economics who were always available for a friendly chat and made me feel welcome and supported during these three years in York.

I acknowledge proofreading assistance kindly provided by Ewan Tomeny, Dr. David Glynn, Michail Prapas, and Sophia Nikolaidou.

This thesis would have never been completed without the selfless help of my life-partner and best friend, Dr. Chrysoula Rizava. I hope life will give me the chance to reciprocate.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, university. All sources are acknowledged as references.

My studies were funded by the Centre for Health Economics by the means of a PhD Studentship.

Chapter 1

Introduction

Health Technology Assessment (HTA) is the process of systematically evaluating the properties of a particular health technology (World Health Organization, 2018). These properties relate not only to the technology's effects on health (e.g. clinical benefits), but also to wider concerns such as its cost and economic impact on the healthcare system (Drummond et al., 2015). In some cases, the perspective is further expanded to include societal welfare concerns such as patients' or even carers' productivity. However, for the purposes of this thesis the public-sector perspective (including the National Health Service (NHS) and personal social services) is adopted in accordance with National Institute for Health and Care Excellence (NICE) guidance. This includes the costs borne by the public-sector, service users, and their families (NICE, 2013).

The term *'health technology'* might refer to any application of practical knowledge that has the potential to improve health (Goodman, 2014). For instance, health technologies are not only limited to pharmaceutical agents (e.g. monoclonal antibodies, vaccines), but may also include procedures (e.g. surgery), medical devices (e.g. diagnostic tests), and social interventions (e.g. educational or behavioural). The variety of existing technologies combined with the rapid advent of new innovative expensive treatments implies that not all technologies can be implemented, and thus not all needs can be met by a health care system (Drummond et al., 2015). Regardless of whether or not a system functions under explicit budget constraints, resources spent on a technology could have been used on alternative options. Therefore, policy-makers considering the implementation of health technologies are always faced with unavoidable decisions associated with *opportunity costs* and benefits (Claxton et al., 2015b). It follows that a set of tools ought to be used so that policy-makers can rationally and transparently decide about the adoption of specific health technologies (Drummond et al., 2015).

Decision analysis offers a quantitative framework that brings together evidence on all relevant parameters (e.g. natural history of disease, Relative Treatment Effects (RTEs), Quality of Life (QoL), resources use, costs etc.), quantifies their relationships, and, after making explicit judgements about social value and modelling structure, produces outputs that can be useful in informing decision-makers regarding the value of alternative policy choices (Briggs et al., 2006). The process of constructing a Decision Analytic Model (DAM) that accurately reflects technologies‘ relative performance, accounting for costs and consequences of all alternative courses of action while producing outputs which directly relate to explicit decision rules, is termed *cost-effectiveness analysis*. A broad summary of the steps involved in cost-effectiveness analysis (Warner, 1989) is provided below:

1. Define the decision problem in terms of the population, the intervention(s), the comparator(s), and the outcome(s) of interest.

2. Identify alternative strategies (courses of action) and construct a DAM.

3. Identify evidence on all relevant model parameters (i.e. natural history, RTEs, costs, resource use, QoL).

4. Synthesise multiple evidence sets on single parameters or groups of related parameters accounting for parameter uncertainty.

5. Bring together the evidence on all relevant parameters quantitatively in a decision-model.

6. Discount future costs and benefits.

7. Appropriately reflect uncertainty in model inputs and outputs.

8. Produce outputs (e.g. total costs and total effects) which will inform a deliberative decision-making process.

1.1. Background

A key component of cost-effectiveness analysis is the process of synthesising the identified evidence. This entails the combination of the various sources of information to produce estimates for single parameters or groups of parameters, and is hence different from pooling the evidence together in a model where mathematical relationships are defined among model inputs. Even though formal synthesis is rare for some decision model inputs such as costs and QoL, for which only a few studies are usually available, it is much more common for RTEs (Drummond et al., 2015). For this purpose, Meta-Analysis (MA) is usually used. MA is a formal process that enables the statistical synthesis of several independently conducted RCTs, and is considered in the top of the evidence hierarchy due to its potential to minimise bias when the whole existing body of research is considered (Haidich, 2010). Since the advent of the first meta-analytic methods which aimed to facilitate the combination of multiple studies comparing the same two treatments (DerSimonian and Laird, 1986), several extensions have been proposed to accommodate particular evidence synthesis challenges and data structures. For instance, Network Meta-Analysis (NMA), enables the simultaneous synthesis of several studies which do not necessarily evaluate the same treatments, allowing treatment comparisons across the whole network of treatments to be made (Lu and Ades, 2004) and the best treatment to be identified. As a result, NMA has been widely used for policy-making where decision-makers are often confronted with making decisions between multiple treatment options in the absence of head-to-head studies that include all relevant treatments.

In evidence synthesis for decision-making, the amount of available evidence is of paramount importance. To assist study selection in systematic reviews, guidance from the Centre for Reviews and Dissemination (CRD) suggests that the research question should be made explicit by defining its PICOS (P : Population, I : Intervention, C : Comparator, O : Outcome, S : Study-design) (Centre for Reviews and Dissemination, 2006). The set of evidence that comprises studies which investigate the research question as defined by all dimensions of PICOS are henceforth termed *direct evidence*. Ideally, the direct evidence on relative effectiveness for an HTA comprises of a collection of comparative studies, appropriately randomised, evaluating all of the interventions under assessment, recruiting patients from the population of interest, and measuring effects on all relevant outcomes. Where such evidence is available, the aforementioned standard MA and NMA methods may be used to synthesise the evidence base and provide decision models with the appropriate relative effectiveness inputs and appropriate characterisation of uncertainty.

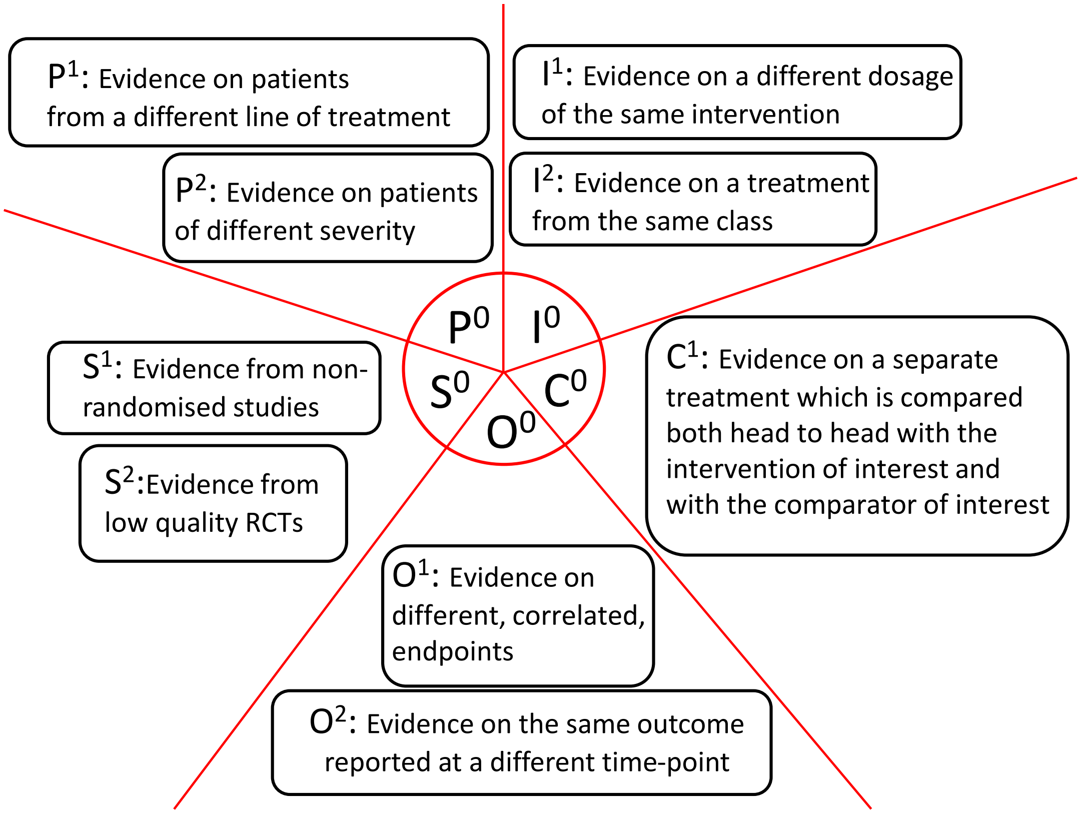
Often though, direct evidence in HTA is sparse and/or heterogeneous and synthesis becomes challenging. *Evidence sparsity* is defined as the situation where sparse event rates are observed for one or more of the treatments under consideration, because either only a few patients were recruited in the available trials, or studies reported at a short follow-up (Dias et al., 2018). This can create several problems for HTA and estimates of RTE. First, the required RTE estimates cannot often be obtained, and even when they can, they are surrounded by high uncertainty and are not considered adequately robust to inform an un-deferrable decision (Ades and Sutton, 2006). Second, evidence sparsity may prevent appropriate exploration of heterogeneity because small studies are at higher risk of enrolling unrepresentative populations (IntHout et al., 2015) and inappropriately reporting patient characteristics (Soares, 2017). Finally, a sparse evidence base may complicate the justification of distributional approximations for certain parameters of interest (Sweeting et al., 2004).

A primary example concerns paediatric indications of health technologies where the evidence base is typically sparse due to the regulatory restrictions on trials which restrict evidence development in children. Hence, HTA is facing important barriers because the absence of evidence prevents the precise quantification of the magnitude of effect (English et al., 2010). Such situations have been considered by the U.S. Food and Drugs Administration (FDA) (Food and Drug Administration and Center for Devices and Radiological Health, 2016) and the European Medicines Agency (EMA). Specifically, EMA has suggested that *"The evidence needed to address the research questions that are important for marketing authorisation of a given product in the target population might be modified based on what is known for other populations"* (European Medicines Agency, 2016). Therefore equivalent evidence on adult patients may be considered for decision-making, while acknowledging potential differences between the two populations. The proposal here is to expand the evidence base (i.e. extrapolate) to include other evidence that relate to the research question only indirectly, but may still be judged as relevant (i.e. *indirect evidence*).

Similarly, since the decision to include indirect evidence is based on a judgement of its relevance to the research question, it is possible that evidence may be judged relevant when the indirectness pertains to another PICOS dimension. For example, indirect evidence pertaining to a study-design or a treatment that is not directly considered in the decision research question may still be able to contribute relevant information. Figure 1.1 shows several cases where the direct evidence base may be expanded to consider indirectly related evidence on any PICOS level. It becomes immediately apparent that if interest lies in a particular treatment comparison even NMA can be considered as borrowing strength from evidence on indirectly relevant treatment comparisons through the consistency assumption to inform the treatment effect of primary focus (Ohlssen et al., 2014).

An important concern with the use of indirect evidence is the imposed level of borrowing (i.e. the extent to which the indirect evidence is allowed to affect the estimates

Figure 1.1: *Extended evidence base.*


P1: Evidence on patients from a different line of treatment
I1: Evidence on a different dosage of the same intervention
P2: Evidence on patients of different severity
I2: Evidence on a treatment from the same class
P0
I0
S0
C0
O0
S1: Evidence from non-randomised studies
C1: Evidence on a separate treatment which is compared both head to head with the intervention of interest and with the comparator of interest
S2:Evidence from low quality RCTs
O1: Evidence on different, correlated, endpoints
O2: Evidence on the same outcome reported at a different time-point

The direct evidence, characterised by P^0, I^0, C^0, O^0, S^0 is included in the small circle in the centre. All the evidence sets outside the small circle address a slightly different research question and may be considered only indirectly relevant.

obtained by using only the direct evidence). In England and Wales, Technology Appraisals (TAs) are conducted by NICE to assess whether new and existing technologies should be routinely funded in the NHS. Typically, when indirect evidence is used in TAs, evidence sets are either considered perfectly generalisable (*lumping*), or separated according to a suspected source of heterogeneity and independently analysed (*splitting*). For example, Duarte et al., 2017, expanded a paediatric evidence base by adding studies that enrolled adult patients and lumped all trials in a single analysis. Also, Faria et al., 2016 generalised relative effectiveness across subgroups of different Hepatitis C genotype and Rodgers et al., 2011 combined evidence from studies that report at different follow-up periods without accounting for the potential impact of the length of follow-up on relative effectiveness. In contrast, Corbett et al., 2017 -TA 445- used a splitting approach and separately analysed the evidence that relate to two population subgroups and four outcome measures, conducting eight different analyses.

Despite the fact that the majority of HTAs usually either lump or split, there are examples of HTAs which use more sophisticated synthesis methods. These methods may impose different levels of information-sharing between direct and indirect evidence. For instance, Corbett et al., 2016 —TA 383 —not only assumed a 'class-effect' (i.e. exchangeability) among the RTEs of interventions that share the same molecular pathway, but also functionally related two outcomes so that information on any outcome contributes to inferences on the other outcomes. Also, McDaid et al., 2009 —TA139 —and Burch et al., 2008 —TA168 —simultaneously modelled two outcomes by allowing their correlation structure to share information between outcome measures. Finally, Riemsna et al., 2011 —TA244 —modelled a network of interventions with multiple treatment components assuming that the RTE of an intervention is the sum of the RTEs of its comprising components.

Overall, the use of indirectly related evidence necessarily implies that there is some influence of the indirect evidence on the estimates of interest —that is, there is *sharing of information*. On the one hand, borrowing strength enables the use of all available relevant evidence, providing a coherent way to interpret the data from the synthesis to the economic model. It may also yield more precise estimates and allow better characterisation of heterogeneity and uncertainty, providing a more comprehensive basis for decision-making. On the other hand, incorporating indirectly related evidence may introduce bias, inflate heterogeneity, impose difficult-to-verify assumptions, and raise questions around relevance. Crucially, underlying the use of indirect evidence is always a judgement of relevance which is untestable, and can therefore be challenged. As a result, borrowing strength from indirect evidence should not be viewed as a substitute for high-quality direct evidence, but instead only as a way of making better use of all the available evidence by explicitly describing assumptions that relate to information-sharing in order to produce more appropriate inputs for decision-models.

To date, no work has attempted to bring together all the methods that can facilitate information-sharing, and explore their implications for decision-making. Instead, existing work broadly falls into two categories: the first category is methodological papers that develop models for very specific synthesis problems, such as the combination of randomised and observational evidence (Verde and Ohmann, 2015), the simultaneous analysis of evidence pertaining to multiple dosages of a particular treatment (Welton et al., 2008), and bias-adjustment (Turner et al., 2009). Although these papers deal with information-sharing, they lack the generality required to approach information-sharing comprehensively. The second category includes papers that describe methods for *multi-parameter evidence synthesis* (Ades and Sutton, 2006; Ades et al., 2008), discussing their potential applicability and value for decision-making (Ades et al., 2006). However, these papers were published more than 10 years ago, and do not include recent developments, or explore the policy implications of information-sharing. Hence, further research that aims to produce more general guidance regarding the use of Information-sharing methods (ISMs) in HTA and their implications for cost-effectiveness is warranted.

1.2. Thesis aims

This thesis is concerned with the use of evidence to support decision-making. In particular, it is focused on issues relevant to the use of indirect evidence and methods that borrow strength from *indirect evidence* to assist inference. Importantly, the notion of indirect evidence is used in a broader manner than in the NMA literature[1] and hence evidence may be indirect to any PICOS level (see Figure 1.1).

The aims of this thesis are:

1. To identify and classify evidence synthesis methods that have been used in the literature to combine evidence that directly and indirectly relate to a research question.

2. To comprehensively describe the different methods that can be used to borrow strength in a specific context (from an indirectly related population), identify explicitly their underlying assumptions, and show how they can be implemented.

3. To develop a framework for the identification and selection of applicable methods that borrow strength in order to systematise the process of methods choice and in this way aid transparency.

4. To illustrate the use of ISMs in an applied case-study.

5. To highlight the impact of using different ISMs on adoption and further research recommendation decisions.

6. To understand which features of the evidence determine the extent of borrowing of strength that each method imposes.

7. To produce recommendations for HTA practice and further methods research.

[1] In the NMA literature the term *indirect evidence* is only used to illustrate how the evidence that pertain to the various treatment comparisons are allowed to influence one another through the consistency equations.

1.3. Structure of the thesis

This thesis is structured in the following manner:

Chapter 2 starts by introducing the standard methodology for HTA focusing primarily on evidence synthesis methods such as pairwise MA, NMA, and ways to account for and explain heterogeneity. Subsequently, standard decision-modelling methods are described, including decision rules, and ways to characterise uncertainty in decision-making and prioritise further research towards where it is most needed.

Chapter 3 is a systematic review that used citation-mining techniques to identify methods that have been used in the literature to combine evidence directly and indirectly relating to a research question. A categorisation of the identified methods according to the main assumption imposed to facilitate information-sharing is provided.

Chapter 4 considers the problem of synthesising independent studies conducted on two different yet related populations, allowing information on relative effectiveness to be shared between the two sets of studies. This chapter describes in detail the different NMA methods that may be used for a specific synthesis problem, thoroughly discussing the assumptions underpinning each approach. A step-by-step framework is suggested to systematise the process of choosing ISMs, and aid transparency.

Chapter 5 is a case-study on the use of intravenous immunoglobulin (IVIG/IVIGAM) for adults with severe sepsis and septic shock. The adult evidence base is expanded to include evidence from paediatric patients. Several ISMs are used to combine the two evidence sets. Methods are compared according to the degree of information-sharing impose based on three different measures that aim to capture the impact on the point estimate and the precision of the adult relative effect estimate.

In Chapter 6 the RTE estimates produced from the application of the various ISMs in Chapter 5 are used in a decision model that evaluates the cost-effectiveness of IVIG/IVIGAM. This work reveals the implications of information-sharing for policy.

Chapter 7 describes a simulation using probabilistic scenarios to compare ISMs in terms of the degree of information-sharing they impose. The methods described in Chapter 4 are reduced to the pairwise MA case. Several different scenarios are constructed differing in the characteristics of the indirect evidence and in how the indirect evidence relates to the direct. This shows how the nature of the direct and indirect evidence affects how methods compare to lumping and splitting. The simulation is run both under FE and RE to contrast methods' behaviour under both approaches.

Finally, Chapter 8 provides an overall summary and discussion of the thesis, drawing attention to the main findings and contributions of each chapter, and highlighting a number of practical recommendations and directions for future research.

Chapter 2

Synthesis and modelling methods in Health Technology Assessment

This chapter introduces the basic methods used in HTA to obtain the most contemporary reflection of the costs and consequences of the alternative choices that decision-makers are faced with. Section 2.1 summarises the main statistical methods that are used for evidence synthesis. These methods primarily consider the synthesis of relative effectiveness evidence as the use of formal evidence synthesis methods is much less common for other types of parameters (e.g. costs, utilities). The methods summarised here will serve as a foundation for the ISMs that will be described in this thesis. Subsequently, Section 2.2 summarises quantitative methods that are used to combine all the evidence that is relevant to a decision within a decision analytic modelling framework and appropriately reflect uncertainty.

2.1. Evidence synthesis methods for binary data

This section describes the basic statistical methods used for evidence synthesis in HTA. Initially, pairwise MA for binary data is introduced (DerSimonian and Laird, 1986), as a means of combining several studies that assess the effectiveness of the same two interventions in separate arms. MA provides a vehicle that enables the synthesis of multiple studies, and the estimation of an overall summary effect that has the potential to be more precise than the results of single studies. Subsequently, pairwise MA methods are extended to accommodate the inclusion of multiple treatments, which have not necessarily been compared head-to head in a network meta-analytic framework. Finally, the approach adopted in this thesis for model fitting and model comparison is described along with issues that arise when between-trial differences are present, and methods that have been suggested to account for and/or explain heterogeneity. Although evidence synthesis methods can be implemented under both a Bayesian and a frequentist framework, the primary focus here will be on the former, because the Bayesian framework naturally lends itself to information-sharing.

2.1.1. Pairwise meta-analysis

Consider a set of n studies ($i = 1, ..., n$) comparing the same two alternative treatment options in the population of interest. If all studies report the same dichotomous (i.e. binary) outcome, the data generation process is commonly assumed to follow a binomial distribution so that:

$$r_{i,k} \sim Binomial(p_{i,k}, n_{i,k}) \tag{2.1}$$

$$logit(p_{i,k}) = ln(\frac{p_{i,k}}{1 - p_{i,k}}) = \theta_{i,k} \tag{2.2}$$

where $r_{i,k}$, $n_{i,k}$, and $p_{i,k}$ are the number of events, the total number of patients, and the probability of an event in study i and arm k respectively. $\frac{p_{i,k}}{1-p_{i,k}}$ represents the odds of the event and $\theta_{i,k}$ the logit transformed probability from the $0 \leqslant p \leqslant 1$ range to the real line $(-\infty, +\infty)$.

Depending on the nature of the data, different likelihoods and link functions can be used. For example, continuous data are commonly modelled using a normal likelihood and an identity link, count data using a poisson likelihood and a log-link, and competing risk data[1] using a multinomial likelihood log-link function (Dias et al., 2011a).

In order to synthesise the evidence provided by all the studies, typically the Contrast-based Model (CBM) is used. For the binomial likelihood model, the CBM looks like:

$$\theta_{i,k} = \mu_i + \delta_i \tag{2.3}$$

where μ_i is the log-odds of the baseline treatment, and δ_i is the RTE between the baseline and the active treatment in the form of the log-odds ratio. Even though the study-specific baselines are commonly treated as nuisance parameters, further assumptions are usually imposed on the RTEs. Under fixed-effect models, the studies are assumed to be estimating the same underlying 'true' RTE, and therefore differences across studies are purely attributed to random noise (i.e. sampling error). Under RE, studies are assumed to be estimating only exchangeable RTEs, and hence discrepancies are also attributed to differences between study-specific 'true' effects (Figure 2.1). Mathematically, under a FE model:

$$\delta_i = d \tag{2.4}$$

[1] This is special data structure where a patient can reach several endpoints, but once they reach one they cannot reach any other. As a result, negative correlations are induced between the different endpoints and those need to be appropriately reflected.

where d is the underlying true RTE (Equation 2.4). Alternatively, under a RE model,

$$\delta_i \sim (d, \tau^2) \tag{2.5}$$

where d is the hyperparameter of the normal distribution from which the underlying study-specific true RTEs are drawn from, and τ^2 is the variance of that distribution which is indicative of the heterogeneity across the RTEs of the included studies (Equation 2.5).

Figure 2.1: *FE (A) and RE (B).*

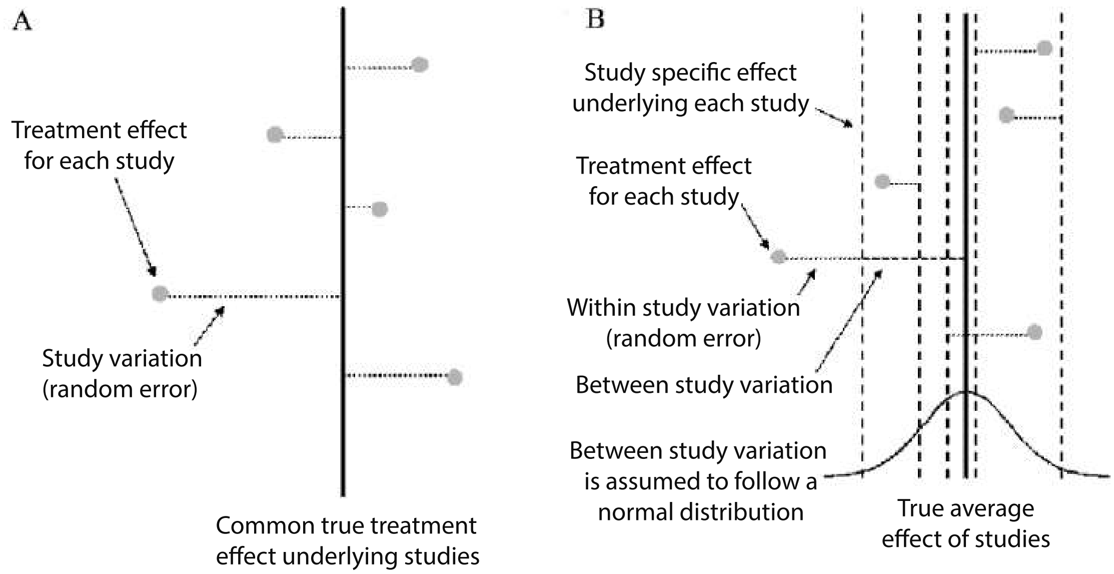


Figure adapted from Lee, 2008.

2.1.2. Network meta-analysis

Suppose now that more than two treatments, forming a connected network (e.g. Figure 2.2), can be used to treat the population of interest, and that several studies which assess the effectiveness of all or any subset of the alternative treatments are available.

Figure 2.2: *An example of a connected network.*

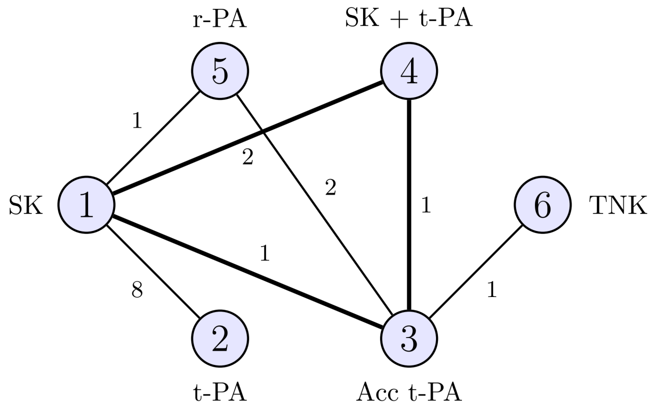


Figure adapted from Caldwell et al., 2005.

To accommodate the simultaneous analysis of all treatments, Lu and Ades, 2004, motivated by Lumley, 2002, suggested methods that extend the CBM from pairwise MA to NMA whilst retaining randomisation. Essentially,

$$r_{i,k} \sim Bin(p_{i,k}, n_{i,k}) \tag{2.6}$$

$$logit(p_{i,k}) = \theta_{i,k} = \mu_{i_b} + \delta_{i,bk} \cdot I_{\{k \neq b\}} \tag{2.7}$$

$$\delta_{i,bk} = d_{bk} \quad \text{(FE)} \tag{2.8}$$

$$\delta_{i,bk} \sim N(d_{bk}, \tau^2_{bk}) \quad \text{(RE)} \tag{2.9}$$

$$d_{bk} = d_{1k} - d_{1b} \tag{2.10}$$

$$d_{AA} = 0 \tag{2.11}$$

where now μ_{i_b} is the study-specific baseline log-odds of the reference treatment b in trial i (which is not necessarily the reference treatment of the whole network), and $\delta_{i,bk}$ is the study-specific RTE between the baseline treatment in study i and the treatment in arm k.

Under a FE model, d_{bk} is the study-invariant RTE between treatments b and k (Equation 2.8), whilst under a RE model, the study-specific $\delta_{i,bk}$ are assumed to be drawn from a normal distribution with a common mean d_{bk} and a between-trials variance parameter τ^2_{bk} (Equation 2.9). Between-trial variances are typically assumed invariant across comparisons (i.e. $\tau^2_{bk} = \tau^2$) to assist identification. This assumption implies that the correlation between any two treatment comparisons is 0.5 (Higgins and Whitehead, 1996). However, if either patient populations are not similar across the included studies, or study-designs are not similar across comparisons, it may not be reasonable to assume a common heterogeneity parameter. For such cases, alternative modelling approaches for the comparison-specific heterogeneities have been suggested (Lu and Ades, 2009).

The main assumption of NMA is embedded in the consistency equations (Equation 2.10). These describe a set of functional relationships that express any comparison-specific RTE mean as a linear function between two basic parameters (i.e. RTE means between a treatment and the reference treatment in the network —denoted as treatment 1 —). *Consistency* implies that 'direct' evidence (i.e. evidence from studies directly comparing two treatments in separate arms) and 'indirect' evidence (i.e. evidence from studies comparing two treatments via a third anchor treatment) are in agreement (Figure 2.3). Essentially, consistency assumes that all studies can be considered multi-arm studies that would have assessed all treatments included in the network, but the arms not included are missing at random (Salanti, 2012). Methods for evaluating consistency include the Bucher method (Bucher et al., 1997) and node-splitting (Dias et al., 2010a).

Figure 2.3: *An illustration of direct and indirect evidence in NMA.*

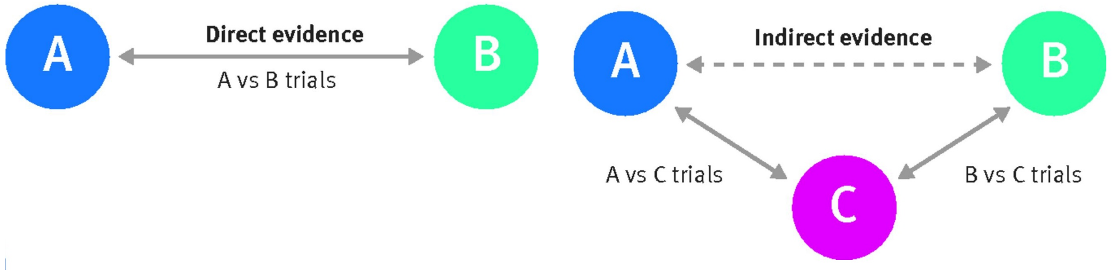


Figure adapted from Riley et al., 2017.

2.1.3. Model fitting and model comparison

Both MA and NMA can be implemented using either a frequentist or a Bayesian approach. Under the former, all information is contained within the data, and model parameters are estimated using maximum likelihood functions. Under the latter, Bayes' theorem (Bayes, 1763) is used to combine previous information with the available study data (Ntzoufras, 2008). Essentially,

$$posterior \propto likelihood \cdot prior \tag{2.12}$$

where the posterior probability distribution represents the updated belief once previous information, expressed in the prior distribution, is combined with the available data contained in the Bayesian likelihood. It follows that under a Bayesian framework priors need to be specified for all parameters that are estimated by the model. Most often, researchers opt for 'uninformative' prior distributions which contain minimal information, in order to allow the information contained within the data to dominate the posterior. For example, in MA and NMA models, common choices of uninformative priors for log odds-ratios RTEs and baseline log-odds are $Normal(0, 10^4)$, whilst for between-trial variances $Uniform(0, 2)$ or, less often, $Inverse.Gamma(10^{-2}, 10^{-2})$ (Dias et al., 2011a).

All models described in this thesis adopt a Bayesian approach primarily due to its flexibility and ease of implementation even with complex models (Dias et al., 2018), but also because the Bayesian framework naturally lends itself to information-sharing. To fit the various Bayesian models, most common software choices are *WinBUGS* and *OpenBUGS*, both of which use Markov Chain Monte Carlo (MCMC) sampling algorithms to estimate posterior distributions (Lunn et al., 2013). For model comparison, information criteria, and in particular Deviance Information Criterion (DIC) is used here. This is the sum of the posterior mean of the residual deviance and the effective number of parameters (McCullagh and Nelder, 1989). Hence, DIC incorporates both model fitting and model complexity considerations, and therefore provides a principled way to balance over-fitting and under-fitting (McElreath, 2016). In this work it is assumed that differences larger than three units of DIC are important (Spiegelhalter et al., 2002).

2.1.4. Heterogeneity and meta-regression

It is inevitable that studies combined in a meta-analysis will often differ. The Cochrane handbook (Higgins and Green, 2011) distinguishes between two main sources of between-trials variation. First, clinical heterogeneity is attributed to differences across studies in the characteristics of the enrolled populations, the interventions, or the outcomes used. Second, methodological heterogeneity relates to differences in study-design and conduct. Clinical and methodological diversity between studies lead to variability in the study-specific treatment effects which is termed *statistical heterogeneity* or simply heterogeneity. Tests for heterogeneity include Cochran's Q test (a chi-squared test that evaluates the null hypothesis that all studies estimate the same underlying treatment effect) (Cochran, 1950), and Higgin's I^2 (estimates the proportion of variability that is beyond what can be attributed to chance) (Higgins and Thompson, 2002).

In a RE model, heterogeneity is accounted for by assuming that the study-specific RTEs are not identical across studies, but only exchangeable, and are hence drawn from a normal distribution (see Equation 2.5, Equation 2.9). Essentially, τ^2 reflects the extent of heterogeneity, with higher values indicating a more heterogeneous evidence base. It follows that when $\tau^2 = 0$, a RE model reduces to FE. Importantly, even though the RE model accounts for heterogeneity, it does not specifically model its source, and therefore it does not explain it.

When a study-level covariate is suspected to be a source of heterogeneity, analyses can be adapted to reflect the covariate's effect. In the simplest case where a covariate is categorical, the evidence base can be split into subsets of studies according to the levels of the variable, and separate analyses can be conducted for each subset of studies. However, in such subgroup analyses, less data is used in each analysis rendering the estimation of the between-trial variances harder —particularly in subgroups with fewer studies. Also, subgroup analyses do not provide a test of interaction; hence, it cannot be confirmed that the suspected covariate is indeed an important source of heterogeneity (Dias et al., 2011b). Alternatively, to simultaneously model all studies while accounting for potential heterogeneity caused by a study-level covariate, the pairwise meta-analysis model (Equation 2.3) can be extended to a meta-regression model so that:

$$\theta_{i,k} = \mu_i + \delta_i + \beta \cdot \mathrm{X}_i \tag{2.13}$$

where X_i is the study-level covariate. If X is binary, β is the additional RTE in the subset of studies for which $X = 1$. If X is continuous, β is the slope that represents the additional RTE for every additional unit of X. The δi represent the study-specific RTEs not attributed to the covariate, and may follow either a FE or a RE model according to Equation 2.4, Equation 2.5 respectively. In a RE meta-regression model, τ^2 represents the additional heterogeneity that cannot be explained by the covariate effect and is expected to decrease compared to the null model that did not include any covariates. In a Bayesian framework, β is usually assigned a vague prior such as $\sim N(0, 10^4)$ (Dias et al., 2011b). The model naturally extends to multiple study-level covariates. However, it should be noted that several studies are required to detect such interaction effects with adequate power (Borenstein et al., 2009), and the existence of a relationship at the study-level (i.e. across studies) does not mean that the same relationship also applies at the individual level (i.e. across individuals within a each study) (Higgins and Green, 2011). This is termed 'ecological fallacy' and to avoid it one would have to use Individual-patient data (IPD) in order to establish a relationship at the within-study level. Finally, meta-regression models can be extended to the NMA context, however comparison-specific slopes are then used and assumptions may need to be made regarding how those are related across treatment comparisons (Cooper et al., 2009).

For decision-making, heterogeneity mainly matters for two reasons: first, if benefits differ due to some patient characteristic, then the estimates of the benefit that are used in cost-effectiveness analyses must reflect the expected benefit in the target population; that is the population considered by the decision. Second, recognising differences in the expected benefits across subgroups might justify 'optimising' decisions instead of making 'one size fits all' recommendations. Importantly, it is necessary to ensure that the health gain for all subgroups for which a treatment is recommended is sufficient to offset the potential health lost from a reduction in the provision of services elsewhere in the health system necessary to fund the new treatment.

2.2. Decision modelling methods

This section explains basic methods used in HTA for constructing and analysing decision analytic models. In particular, it starts by describing the various model structures, their main benefits, and cases where each model type may be more appropriate than others. Subsequently, the main model inputs are presented along with sources typically employed to identify relevant evidence, and quantities used to inform resource allocation decisions are described. Then, the importance of uncertainty is highlighted, and ways to evaluate and present the uncertainty that surrounds decisions are explained. Finally, the established methods for assessing the value of obtaining further information are detailed along with methods for determining what type and design for future research may be most valuable.

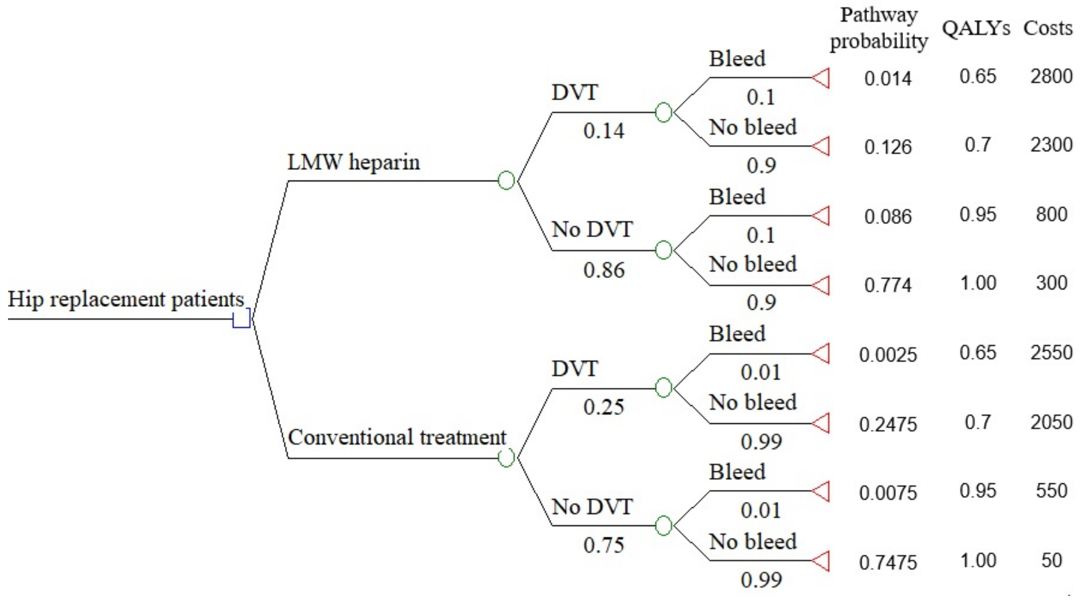
2.2.1. Model structures

Once the research question has been defined, a decision model is often constructed to represent the decision problem, and provide a quantitative framework to bring together all relevant evidence. The choice of decision model type is not necessarily straightforward, however, often, a particular decision model type is better suited to represent diseases with specific characteristics. Examples of key considerations include whether we need to model recurring events, whether event probabilities vary with time in the model, and whether patients' prognoses depend on events that have already happened in the past (Drummond et al., 2015). This section briefly describes the different types of decision models used in HTA, focusing primarily on decision trees and Markov models.

2.2.1.1 Decision trees

Decision trees represent patients' possible prognoses using a set of alternative pathways (Figure 2.4). Each pathway describes a sequence of events, often chronologically ordered (Briggs et al., 2006). Three types of nodes can be distinguished (Gray et al., 2010): first, decision nodes represent potential policy questions that decision-makers may be faced with (e.g. should we give low molecular weight heparin or conventional treatment to patients who receive a hip replacement?). Second, chance nodes indicate that a set of mutually exclusive events (i.e. events which are characterised by probabilities that sum to one) occur (e.g. the patient either experiences a deep vein thrombosis event or not). Finally, terminal nodes signal the end of a pathway. Pathway-specific probabilities are calculated by multiplying the initial branch probability with the subsequent conditional probabilities (Drummond et al., 2015). Finally, pathway-specific costs and payoffs are assigned to each of the terminal nodes, and the model is 'averaged out' by weighting costs and payoffs with their corresponding pathway probabilities. Though the simplicity of decision trees has led to their widespread use for acute conditions featuring short time horizons, these models are unable to accommodate time-dependency, and tend to get 'bushy' and cumbersome in more complex diseases with recurring events; hence, their use in chronic conditions is limited (Petrou and Gray, 2011).

Figure 2.4: *An example of a decision tree.*



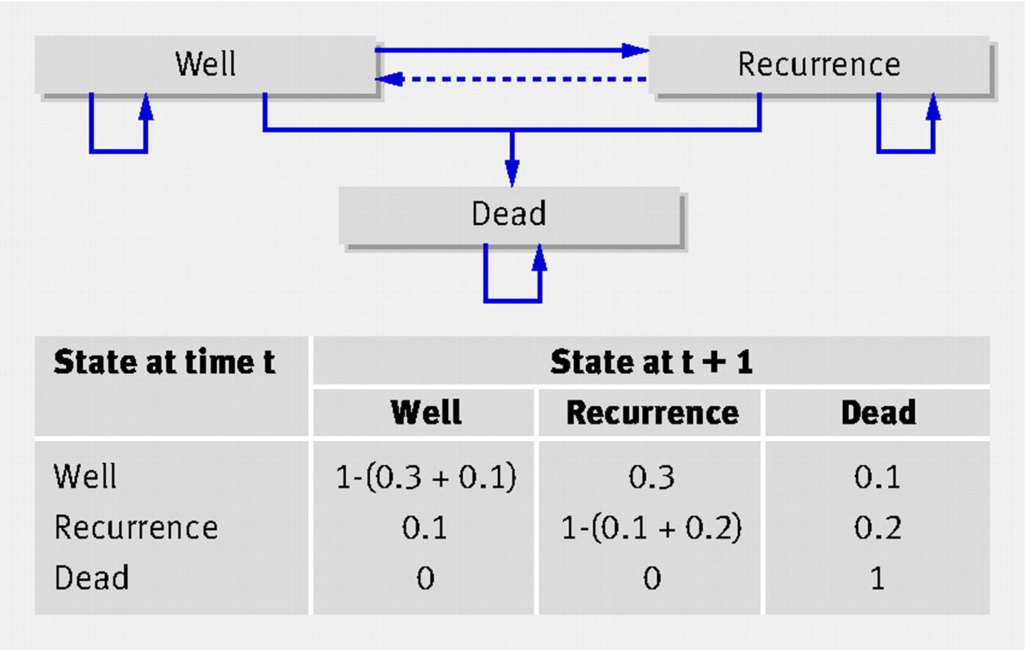
The model compares the use of low molecular weight (LMW) heparin and Conventional treatment for preventing deep vein thrombosis in patients who undergo hip replacement. Adapted from Sutton, 2016.

2.2.1.2 Markov models

In contrast to decision trees which do not explicitly model time, Markov models use a finite number of 'states' in which patients reside at any point in time. Time is explicitly modelled and discretised into cycles of particular length (e.g. one month). Between cycles, patients either remain in the same state or transition across states according to specified transition probabilities. The model comes to an end either when all patients have arrived in an absorbing state from which they cannot 'escape' (e.g. death), or when a pre-specified number of cycles has elapsed (Petrou and Gray, 2011). Costs and benefits associated with each cycle are calculated by weighing the state-specific costs and benefits by the proportion of patients that occupy each state in that cycle. Total costs/benefits can be estimated as the sum of cycle-specific costs/benefits over the time horizon of the model.

The main limitation of Markov models is their 'memorylessness' which means that the transition probabilities are usually fixed through time; hence, a patient has the same probability of moving from State A to State B, regardless of the cycle the model is currently in, or the time the patient has already spent in a state (Drummond et al., 2015). This limitation can be overcome by introducing 'tunnel' states that patients are required to occupy in a particular sequence, or states that patients can only reside in for one cycle (Briggs et al., 2006).

Figure 2.5: *An example of a Markov model.*



State at time t	State at t + 1		
	Well	Recurrence	Dead
Well	1-(0.3 + 0.1)	0.3	0.1
Recurrence	0.1	1-(0.1 + 0.2)	0.2
Dead	0	0	1

Figure is adapted from Petrou and Gray, 2011.

2.2.1.3 Other model structures

To overcome the limitations of cohort models (e.g. decision trees, Markov models), approaches that model individuals separately have been suggested (Siebert et al., 2012; Karnon et al., 2012). For instance, patient-level simulation models track how each individual transits among several states while accumulating costs and benefits. Such models provide analysts with increased flexibility by allowing each patient's prognosis to depend on their history. Hence, complex dependencies between risks of clinical events and patient histories can be easily modelled, and the overall expected values can be calculated by averaging costs and benefits across patients. Discreet-event simulation models also simulate patients individually, however they also model the time until the next event for each patient separately, avoiding the disadvantages of fixed cycle lengths (Petrou and Gray, 2011). Despite their benefits, the aforementioned models can be hard to inform as they may be more heavily parametrised than cohort models, and thus require more evidence to be populated (Drummond et al., 2015).

Another recent development is Partition Survival Models (partSA). Despite their conceptual similarity with state-transition models, partSA models directly utilise survival curves (e.g. overall survival, progression-free survival) to determine state membership. On the one hand, partSA models can be very useful when only survival data is available, and therefore there is not enough evidence to calculate transition probabilities and the effect of treatments on these probabilities (as is usually the case for anti-cancer treatments). On the other hand, partSA models, consider outcomes independently and cannot accommodate cases where there is a structural dependency across endpoints (Woods et al., 2017).

All model types described above are static and therefore do not allow for interactions across patients. This means that each patient's health is assumed to be independent of other patients' health status. Despite the fact that this assumption may be true for non-communicable diseases, it does not hold for infectious diseases which may require dynamic modelling. For instance, the probability of being infected by a disease may be contingent on the proportion of people who have been immunised against it (herd-immunity). Dynamic transition models can be used in these cases to allow appropriate estimation of cost-effectiveness that accounts for such interactions (Drummond et al., 2015).

2.2.2. Model inputs

Because decision-makers are accountable, decision models need to be fit-for-purpose, and able to reasonably approximate incremental costs and effects. Hence, the evidence used in the model has to be comprehensive. Consequently, evidence needs to be systematically identified and synthesised in order to produce reliable model inputs. Even though the principles of Evidence-based Medicine (EBM) (Centre for Reviews and Dissemination, 2006) are generally followed for RTE parameters, this is not necessarily also the case for other parameters such as the natural history of the disease, costs, resource use, and QoL. This is because there are rarely many studies providing us with information on these parameters to justify any synthesis. Also, it is necessary to ensure that model inputs constitute the most contemporary reflection of the current clinical practice in the country that faces the decision problem, and therefore evidence pertaining to other jurisdictions may be deemed overly unrepresentative and inappropriate. As a result, it is not uncommon to see natural history, costs, or QoL being informed by observational studies conducted in the country of interest. Crucially, it is paramount to ensure that even when RCTs are not used to derive the required model inputs, a transparent process is followed which is in line with the objectives of the decision model (Philips et al., 2004). A summary of the main decision model inputs and usual sources of evidence is supplied in Table 2.1.

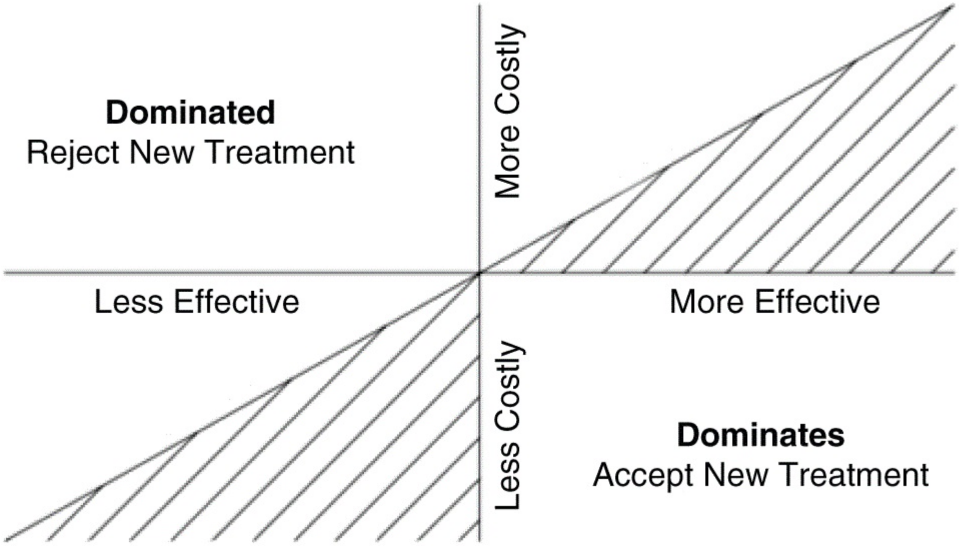
Table 2.1: *Decision model inputs and usual sources of evidence.*

Parameter	Usual sources of evidence	Notes
Natural history	Control arms of clinical trials or observational studies.	It is very important to use the best reflection of the current clinical practice. This means obtaining evidence representative of the context of interest.
Relative effects	Randomised clinical trials.	Formal synthesis methods are commonly used to synthesise the available RCTs. Usual issues include the lack of adequate studies which lead to evidence sparsity problems.
Quality of life	Randomised or observational studies.	Commonly, QoL weights, estimated using instruments such as EQ-5D (EuroQol Research Foundation, 2019), are combined with time, to calculate Quality-Adjusted Life-Year (QALY)s.
Costs and resources use	Randomised or observational studies.	The actual resources that are included depend on the perspective that is adopted in the analysis. NICE suggests that the perspective of the decision-maker (i.e. the payer) is used (NICE, 2013).

2.2.3. Making decisions

The main question that economic evaluation is primarily concerned with is whether the additional health offered by a new technology is sufficient to justify any additional costs (Drummond et al., 2015). Figure 2.6 illustrates the possible situations that analysts may be faced with. It is immediately obvious that under some circumstances decisions can be straightforward. For example, when the new technology falls in the north-west (NW) quadrant, it is dominated by the old technology because it is more costly, and leads to reductions in health. Similarly, when it falls in the south-east (SE) quadrant it dominates the old technology because it increases health, and also yields cost-savings.

Figure 2.6: *The cost-effectiveness plane.*



The shaded area shows the region below the willingness-to-pay threshold that the technology would be considered cost-effective. Figure is adapted from (Savitz and Savitz, 2016).

However, when the new treatment is either more costly and leads to health gains (i.e. north-east quadrant), or less costly and results in health losses (i.e. south-west quadrant), it is less clear whether or not it should be considered cost-effective. The trade-off between costs and health is usually illustrated with the Incremental Cost-Effectiveness Ratio (ICER) statistic. Essentially,

$$ICER = \frac{\Delta C}{\Delta H} = \frac{C' - C}{H' - H} \tag{2.14}$$

where C' and H' represent the costs and benefits of the new treatment, whilst C and H the costs and benefits of the comparator treatment.

A higher ICER implies that we need to devote more resources in order to get the same amount health, and therefore the profile of the technology becomes less favourable. It follows that in order to make an adoption decision we need to be able to specify a cost-effectiveness threshold k (i.e. a cut-off value), also termed Willingness to Pay (WTP) threshold. An ICER that is lower than k would mean that the new technology offers an acceptable trade-off and is therefore recommended, whilst when the ICER is higher than k it would imply that the additional cost of the new technology is not justified by the additional health it offers. In Figure 2.6, for a threshold depicted by the diagonal line, ICERs that fall in the shaded area or in the SE quadrant represent favourable options. Effectively, k represents an expectation of the amount of resources that displaces a single unit of health elsewhere in the health care system (i.e. opportunity cost), and is therefore indicative of the system's productivity (Claxton et al., 2015a). In the United Kingdom (UK), NICE has been using a threshold between £$20,000 - 30,000$ per QALY (McCabe et al., 2008); however, this estimate has little empirical basis, and does not align with more recent empirical threshold estimates.

Even though ICERs are straightforward to use when there are only two alternative treatments, they become complicated when decision-makers are confronted with multiple alternatives (Drummond et al., 2015). This is because many different comparisons can be made and it may be challenging to establish dominance and extended dominance. Under these circumstances, a more convenient way that also avoids the disadvantages of ratio statistics (Hoch et al., 2002), is to summarise cost-effectiveness using the Net-Benefit (NB) statistic (Stinnett and Mullahy, 1998). Essentially, NBs directly translate each technology's associated costs and benefits to overall health or monetary gains by specifying a particular threshold k. The NBs of the various competing alternatives can then be directly compared without the need to calculate increments. Two types of NBs can be specified according to whether we prefer to express NBs in health (Equation 2.15) or monetary terms (Equation 2.16).

$$Net\ Health\ Benefit = H - \frac{C}{k} \tag{2.15}$$

$$Net\ Monetary\ Benefit = H * k - C \tag{2.16}$$

where H is the health that the technology in question offers, C its costs, and k the specified threshold. If net health (or monetary) benefits are positive, we can conclude that a technology is cost-effective as its health gains exceed the opportunity cost.

2.2.4. Evaluating uncertainty

Inevitably, most decisions are associated with some level of uncertainty. Decision uncertainty relates to the fact that we do not know exactly what the actual costs and effects of the use of an intervention to the population of interest will be in any future 'rollout'. Two main sources of uncertainty can be distinguished (Briggs et al., 2012): first, *parameter uncertainty* refers to the input parameters of the decision model, and second *structural uncertainty* relates to the assumptions and judgements that were made in the process of constructing the decision model (Drummond et al., 2015).

Despite the fact that adoption decisions are typically based on the expected costs and benefits, an assessment of the uncertainty surrounding the decision is crucial to evaluate whether the existing evidence is sufficient to inform the decision, or further evidence should be sought. The simplest way of assessing uncertainty is by using Deterministic Sensitivity Analysis (DSA) where one input parameter is varied to determine the sensitivity of the decision to that parameter(s) values (Briggs and Sculpher, 1995). However, DSA is limited because it does not indicate the level of uncertainty that the decision is associated with, and also tends to underestimate uncertainty by ignoring that in reality multiple parameters vary simultaneously (Drummond et al., 2015). Alternatively, PSA can be used (Claxton et al., 2005; Claxton, 2008). In PSA, probability distributions are assigned to all model input parameters, and then random samples are repeatedly drawn for all parameters and combined to calculate iteration-specific NBs.

Figure 2.7: *An example of how the results of the PSA can be used in the incremental cost-effectiveness plane to represent joint uncertainty in costs and effects.*

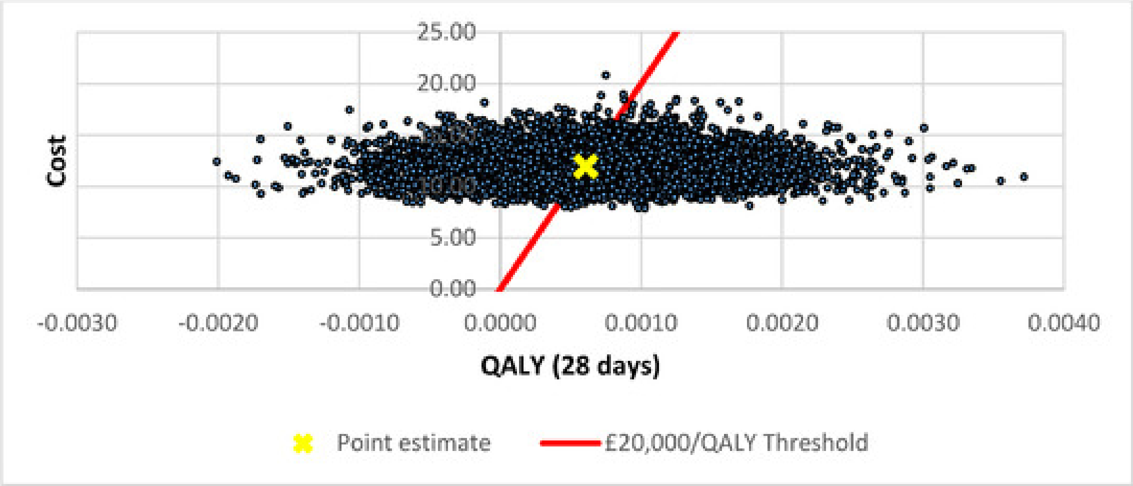


Figure is adapted from (Holmes et al., 2018).

To represent the joint uncertainty in costs and effects the PSA output can directly be illustrated in the cost-effectiveness plane. For instance, in Figure 2.7 each dot represents the incremental costs (y-axis) and effects (x-axis) calculated in an iteration of the PSA. The size of 'cloud' is indicative of the joint uncertainty in incremental costs and effects. All iterations falling below the threshold indicated by the red line suggest that the new treatment should be adopted; hence, the proportion of these iterations indicates the probability that the new treatment is cost-effective. Sensitivity analysis is typically performed for different threshold values, and the probability of the new treatment being cost-effective for each threshold is represented in CEACs (see Figure 2.8). There is a direct link between the quadrants that the iterations' probability mass is located in the cost-effectiveness plane and the shape of the CEACs, and a detailed discussion of the various situations that might occur can be found in Fenwick et al., 2004. Overall, PSA is a more appropriate estimation of decision uncertainty (Briggs et al., 2006), and is recommended by NICE (NICE, 2013).

Figure 2.8: *An example of a CEAC.*

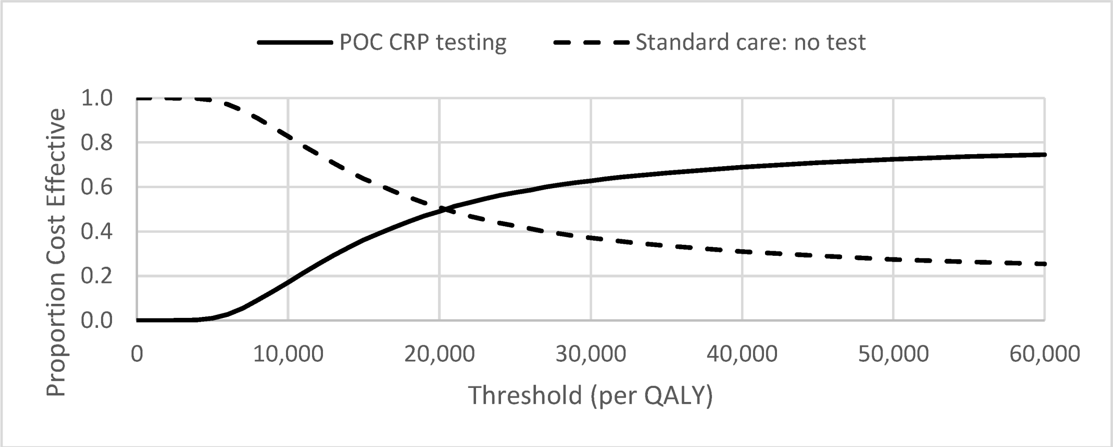


Figure is adapted from (Holmes et al., 2018).

2.2.5. Value of information

Given the presence of decision uncertainty, it is reasonable to infer that decision-makers should also consider the need for and value of further research. Crucially, if the ultimate aim is to improve overall health of existing and future patients, an assessment of the potential value of acquiring further information and resolving uncertainty of decision-model inputs is also central to policy-making (Drummond et al., 2015). Such an assessment would evaluate the expected cost of current uncertainty by considering both the probability and the magnitude of the consequences of making a wrong decision with the existing evidence. The resulting estimate would represent the maximum amount that decision-makers should be willing to invest to reduce uncertainty in the decision at hand.

In what follows, three main concepts of value-of-information analysis are introduced. These are the Expected Value of Perfect Information (EVPI), the Expected Value of Perfect Parameter Information (EVPPI), and the Expected Value of Sample Information (EVSI). The first, EVPI, calculates the maximum value of further research that resolves all parameter uncertainty; EVPPI extends EVPI to identify the particular type of research that would be more useful (i.e. which parameters should we focus on acquiring more information for); finally, EVSI aims to inform the design of further research.

2.2.5.1 Expected Value of Perfect Information

Assume that there is a set of alternative interventions t and that the net-benefit (B) of each is dependent on uncertain parameters that may take a range of possible values (θ). With current information, the best treatment is that which maximises the expected net-benefits, B (Ades et al., 2004). i.e.

$$max_t E_\theta B(t, \theta) \tag{2.17}$$

where max_t indicates that we are seeking the treatment t which maximises the expression that follows, $E(\theta)$ denotes the expectation of θ, and $B(t, \theta)$ the net-benefits produced by treatment t for parameters taking the value of θ. However, the treatment decision that is based on the average values of θ (i.e. $E(\theta)$) is not necessarily the decision that we would have made under every possible combination of the model inputs. In other words, had we known the current state of the world, we would only need to choose the intervention that maximises the NBs of a given value of θ (i.e. $max_t B(t, \theta)$) and we may have made a different decision. However, when decisions about further research are made, θ is typically unknown. Hence, to calculate the maximum NB under every possible combination of the model inputs, we need to average across the joint distribution of θ, i.e.

$$E_\theta max_t B(t, \theta) \tag{2.18}$$

EVPI represents the additional benefit that could be gained if all uncertainty surrounding the treatment choice decision was resolved and therefore it can be used as the upper-bound for the potential benefits of new research. In other words, it is the difference between the net-benefits which can be achieved under perfect and current information and is derived by subtracting Equation 2.18 from Equation 2.17, so that:

$$EVPI = E_{\theta} max_t B(t, \theta) - max_t E_{\theta} B(t, \theta) \tag{2.19}$$

EVPI can be calculated from the PSA output by directly applying the aforementioned rationale.

Since the information acquired by further research can be valuable to more than one patient, to appropriately calculate the upper bound for the cost of research we need to account for the whole current and future patient population. This is achieved by applying the individual EVPI estimate across the whole set of patients expected to be affected by the disease for as long as the technology is expected to be relevant i.e.

$$Pop.EVPI = EVPI \cdot \sum_{y=1}^{Y} \frac{I_y}{(1+r)^y} \tag{2.20}$$

where I_y is the incidence in year y, Y is the total number of years that additional information will be useful for, and r is the discount rate.

2.2.5.2 Expected Value of Perfect Parameter Information

Despite that EVPI is a valuable measure that can be used to eliminate research suggestions expected to cost more than this upper bound, it does not help in prioritising different types of research. In other words, it does not provide insight into the contribution of the various model inputs towards the decision uncertainty and consequences. Such information is obtained by calculating the partial EVPI of a single or a subset of parameters (θ_I), and can be valuable in tailoring further research. For example, if the majority of decision uncertainty and consequences is attributed to relative effectiveness, a randomised controlled trial may be prioritised to better characterise this parameter.

EVPPI (Equation 2.21) is calculated in a similar manner to EVPI (Equation 2.19), only now in the first component we need to average over the joint distribution of the subset of parameters under evaluation i.e.

$$EVPPI = E_{\theta_I} max_t E_{\theta|\theta_I} B(t, \theta) - max_t E_{\theta} B(t, \theta) \tag{2.21}$$

where θ_I denotes a subset of parameters of θ and θ_I^c its complement. $E(\theta|\theta_I)$ represents the expectation of θ conditional on a given value for θ_I. Note that the first component may become computationally intensive as it requires an inner simulation to estimate the NB of each θ_I value (i.e. $E_{\theta|\theta_I}B(t,\theta)$) and an outer integration to sample all possible values for θ_I. However, if $B(t,\theta)$ is linear or multi-linear in θ_I^c, then $E_{\theta|\theta_I}B(t,\theta)$ can be re-written as $B(t,\theta_I,E(\theta_I^c))$ (Thompson and Evans, 1997). For non-linear models, efficient methods that partly alleviate the computational burden have also been suggested (Strong et al., 2014). Population EVPPI is calculated in the same manner as population EVPI.

2.2.5.3 Expected Value of Sample Information

Once the maximum acceptable cost of future research and the type of research that will most significantly reduce the existing uncertainty have been established, we need to better define the characteristics of this research. These may include a trial's sample size and allocation of patients or the appropriate outcome and its optimal length of follow-up. Since we are never going to fully resolve all uncertainty regarding any parameter, the extent of the information that we will gain will directly relate to its characteristics (Drummond et al., 2015).

The development of EVSI follows that of EVPPI (Briggs et al., 2006). Assuming that a new study of sample size n is undertaken, it will provide us with sufficient statistics D relating to the subset of parameters θ_I [2]. If we knew what D was going to be, then we would just $max_t E_{\theta_I^c,(\theta_I|D)}B(\theta_I,\theta_I^c,t)$. However, since we do not know D, we need to average across its distribution. Therefore, EVSI is expressed as:

$$EVSI = E_D max_t E_{\theta_I^c,(\theta_I|D)}B(\theta_I,\theta_I^c,t) - max_t E_\theta B(\theta_I,\theta_I^c,t) \tag{2.22}$$

Finally, to calculate the optimal sample size, the cost of each potential trial of sample size n needs to be deducted from its corresponding population EVSI, calculated in the same way as population EVPI, to produce the Expected Net Benefit of Sample (ENBS):

$$ENBS(n) = Pop.EVSI(n) - Cost(n) \tag{2.23}$$

The sample size n that maximises ENBS is the optimal sample size of the new trial.

In summary, this chapter reviewed the basic methods used in HTA to synthesise relative efficacy evidence as well as methods used to bring together all relevant types of evidence (RTEs, costs, utilities) in a decision-model and characterise uncertainty. These methods serve as the foundation for the work that is undertaken in subsequent chapters.

[2] It is assumed here that θ_I and its complement θ_I^c are independent.

Chapter 3

Classifying information-sharing methods: a citation-mining review

3.1. Chapter aims and structure

As introduced in Section 1.1, usually HTAs either adopt a splitting approach, whereby indirect evidence is completely disregarded (e.g. Corbett et al., 2017), or lump direct and indirect evidence as if they do not differ in any respect (e.g. Duarte et al., 2017; Faria et al., 2016). However, these approaches represent only the two extremes of the 'information-sharing spectrum' and more options are usually available. For instance, some recent HTAs have used more sophisticated synthesis methods, initially developed in the biostatistics literature, which impose more moderate assumptions and perhaps more appropriate degrees of information-sharing between direct and indirect evidence (e.g. Corbett et al., 2016; McDaid et al., 2009; Burch et al., 2008).

This chapter aims to look in the biostatistics literature, and in particular into the field of MA/NMA, in order to identify and classify methods that have been used to share information among multiple populations, treatment comparisons, outcomes, study-designs and more generally between evidence directly and indirectly relating to a research question. Importantly, due to the vastness of the field, this chapter is neither meant to provide an exhaustive list of ISMs, nor describe the details underpinning all those methods, because such aims would far exceed the scope of a single chapter. Instead, its main contribution is the classification of Information-sharing methods (ISMs) according to the main assumption of each. This classification provides a conceptual way of thinking more generally around methods and assumptions that could be used for information-sharing synthesis challenges. To date, no work has been conducted with a primary focus on 'information-sharing' or 'borrowing strength' and the aim of categorising the breadth of the available methods. On the contrary, the existing literature consists of either old studies that try to introduce and discuss methods for multi-parameter evidence synthesis (Ades and Sutton, 2006; Ades et al., 2006), or more recent reviews that discuss methods for MA/NMA (e.g. Efthimiou et al., 2016) without focusing on information-sharing.

This chapter is organised in the following manner: in Section 3.2 the citation-mining methods which were used to identify relevant papers are explained. Section 3.3 starts by providing the characteristics of the included articles and introduces the four 'core' assumptions, at least one of which is employed by any ISM. The identified methods are then explained under each 'core' assumption. Finally, in Section 3.4, strengths, limitations and areas for future research are discussed.

3.2. Methods

This work aims to answer the following research question: *What methods have been used in the literature to combine evidence directly and indirectly relating to a research question and how can those be classified according to the assumptions that they impose?*

Given the lack of consistent terminology in the literature referring to methods that combine direct and indirect evidence, keyword-based search methods (Higgins and Green, 2011) were impractical and could not be used. Instead, 'citation-mining' methods (Grandage et al., 2002), which are efficient (Badampudi et al., 2015) and have been used for similar reviews (Verde and Ohmann, 2015) were preferred. The process consisted of the following steps: initially, a list of seminal/influential papers was compiled and articles that either cited the seminal papers (forwards citation-mining) or were cited by the seminal papers (backwards citation-mining) were identified; subsequently, as in classical systematic reviews, the list of identified papers was screened according to inclusion and exclusion criteria and data were extracted from the included studies.

Seminal papers were chosen after an initial scoping review of the literature and in consensus with the supervisory team to represent a variety of fields including MA, NMA, multi-parameter evidence synthesis, and the incorporation of evidence of historical controls in trial-design. Even though the last category was outside the scope of this review, it may have inspired the extension of methods from that field into MA/NMA. Overall, the citations of 7 seminal papers were searched in the Web of Science (WoS) on 20-Feb-2019. The number of papers cited in each seminal paper and the number of times each seminal paper has been cited on the day of the search, according to WoS, are shown in Table 3.1.

Table 3.1: *'Pearls' (i.e. seminal papers) used for forwards and backwards citation-mining.*

#	'Pearl'	Citations	Cited by
1	Higgins and Whitehead, 1996. *Borrowing strength from external trials in a meta-analysis.*	33	309
2	Ades and Sutton, 2006. *Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches.*	109	82
3	Ades et al., 2006. *Bayesian methods for evidence synthesis in cost-effectiveness analysis.*	79	210
4	Jackson et al., 2011. *Multivariate meta-analysis: Potential and promise.*	74	148
5	Efthimiou et al., 2016. *GetReal in network meta-analysis: a review of the methodology.*	193	37
6	Hobbs et al., 2011. *Hierarchical commensurate and power-prior models for adaptive incorporation of historical information in clinical trials.*	16	64
7	Schmidli et al., 2014. *Robust meta-analytic-predictive priors in clinical trials with historical control information.*	50	42

Subsequently, articles that cited (forwards citation-mining) or were cited by the seminal papers (backwards citation-mining) were identified. Articles were included if they mathematically specified MA or NMA models , either Bayesian or frequentist, that combined information from comparative studies pertaining to multiple populations, interventions, outcomes, study-designs or if they utilised evidence from an external source such as previous meta-analyses. Importantly, papers that used only standard NMA methods were excluded even though they shared information across treatment comparisons, because such methods are well established in the literature.

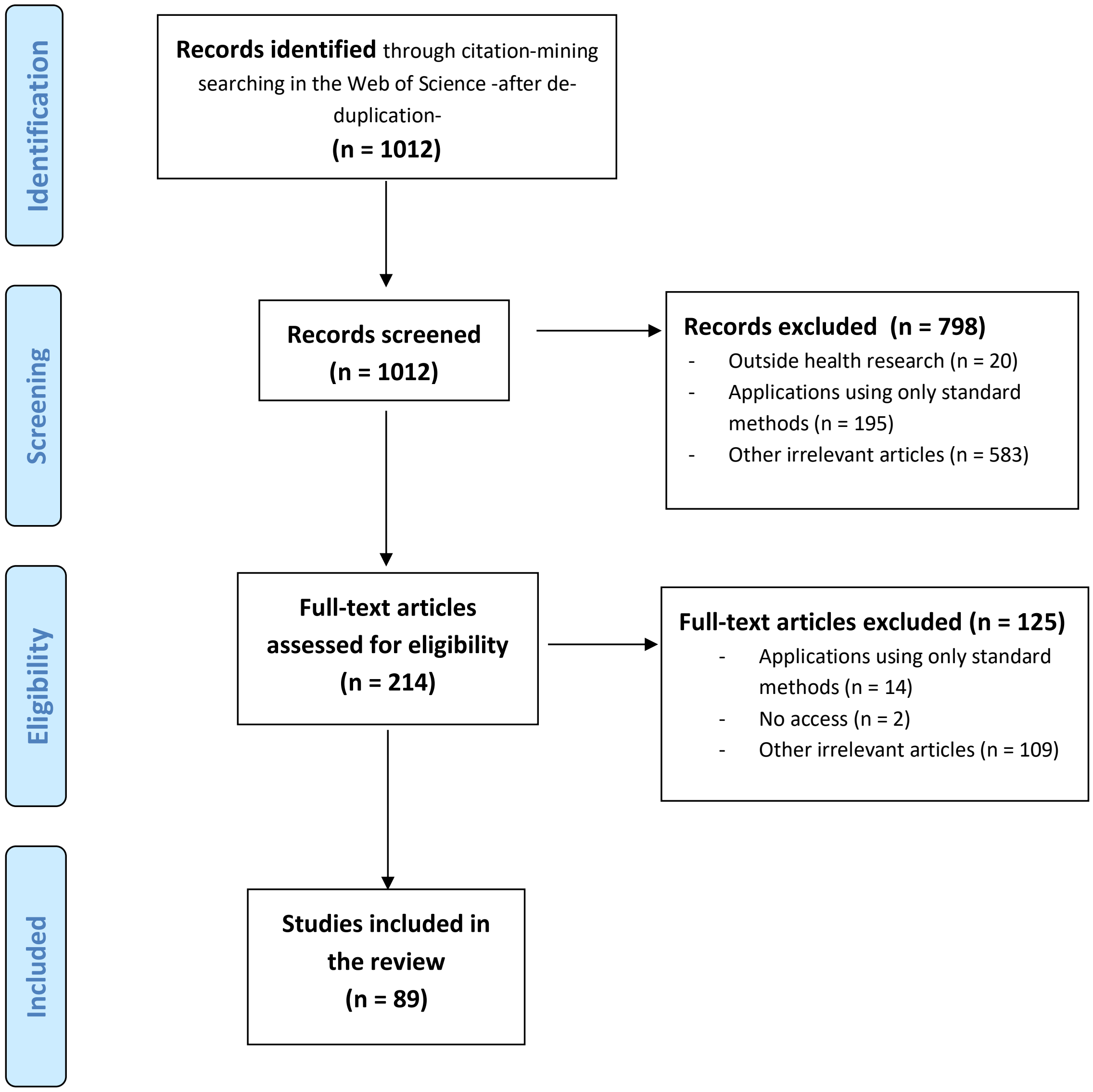
From each included paper, the evidence synthesis model was isolated and from within it, methods facilitating information-sharing were extracted. Methods were subsequently categorised according to the 'core' relationship that they used to enable information-sharing. When papers tackled multiple synthesis challenges simultaneously (e.g. Madan et al., 2014; Welton et al., 2010), the issues they dealt with were isolated along with the method used to address each. The PICOS level of indirectness was also extracted. The search was conducted in Zotero version 5.0.69 using various 'tags' which enabled the categorisation of the identified studies according to their characteristics.

3.3. Results

3.3.1. Characteristics of the included studies

The review identified 89 papers in total, as shown in Figure 3.1.

Figure 3.1: *Citation-mining flow chart.*


Identification
Screening
Eligibility
Included
Records identified through citation-mining searching in the Web of Science -after de-duplication- (n = 1012)
Records screened (n = 1012)
Records excluded (n = 798)
- Outside health research (n = 20)
- Applications using only standard methods (n = 195)
- Other irrelevant articles (n = 583)
Full-text articles assessed for eligibility (n = 214)
Full-text articles excluded (n = 125)
- Applications using only standard methods (n = 14)
- No access (n = 2)
- Other irrelevant articles (n = 109)
Studies included in the review (n = 89)

After removing duplicates, 1012 records were screened, 798 of which were excluded based on their title and abstract, and 214 were assessed for eligibility by full text. A total of 125 studies were subsequently excluded because they only used standard MA/NMA methods without any methodological developments ($n = 14$), or the University of York could not provide access ($n = 2$), or they were irrelevant to information-sharing and borrowing strength ($n = 109$) (e.g. they described methods to assess quality of methods used in MA/NMA or to evaluate network consistency).

The database with all included papers can be found in the following url `https://www.zotero.org/groups/2360368/citation-mining_included-studies`[1]. Most papers ($n = 84$) developed ISMs in a Bayesian framework. The majority of papers ($n = 79$) described methods that shared information on relative treatment effects. Other studies shared information on the comparison-specific meta-regression slopes ($n = 4$), the comparison-specific between studies heterogeneities ($n = 6$), or the study-specific baselines ($n = 2$). Regarding the PICOS level, 39 papers shared information across multiple outcomes, 23 across multiple treatments, 10 across study-designs and 6 across populations. Note that the numbers do not necessarily add up to the total number of identified studies (i.e. 89), because some of the identified papers described methods that shared information on several types of parameters and across more than one PICOS level (e.g. Dakin et al., 2011). Overall, there was a balance amongst papers that developed methods within MA ($n = 45$) and within NMA ($n = 44$). The reader can identify the references pertaining to each category by navigating through the library of the included papers using the relevant tags (link supplied above). A full list of the included papers along with a description of how information was shared within each paper can be found in Table A.2.1.

Table 3.2 lists the most common synthesis challenges addressed by the included papers. Synthesis Challenge 1 is the most common issue relating to multiple population subgroups and considers the combination of adult and paediatric evidence. Challenges 2-4 relate to models that allow the evidence on some treatment to affect the estimation of parameters specific to other treatments. Challenges 5-6 consider the simultaneous synthesis of studies of different designs that may be of different quality and therefore prone to different types and levels of bias. Challenges 7-9 relate to the synthesis of multiple outcomes that may only be correlated or dependent in a more complex manner. Finally, Challenge 10 considers methods for utilising meta-epidemiological evidence to strengthen analyses and assist estimation in conditions of data sparsity.

[1] The included papers are organised according to various tags such as the type of the relationship they use, the synthesis issue they address, whether they developed models for MA or NMA, and the parameter on which information was shared.

Table 3.2: *Main synthesis challenges identified and relevant references.*

Synthesis challenge 1 ($n = 3$): *Synthesis of adult and paediatric evidence*.
Duarte et al., 2017; Gamalo-Siebers et al., 2017; Roever et al., 2019
Synthesis challenge 2 ($n = 7$): *Synthesis of multiple dosages of the same treatment*.
da Costa et al., 2017; Del Giovane et al., 2013; Langford et al., 2018; Mawdsley et al., 2016; Owen et al., 2015; Warren et al., 2014; Wu et al., 2018
Synthesis challenge 3 ($n = 7$): *Synthesis of drugs falling under the same 'class'*.
Dakin et al., 2011; Dominici et al., 1999; Moreno et al., 2011; Nixon et al., 2007; Owen et al., 2015; Soares et al., 2014a; Warren et al., 2014
Synthesis challenge 4 ($n = 5$) : *Synthesis of complex interventions*.
Madan et al., 2014; Melendez-Torres et al., 2015; Mills et al., 2012; Nixon et al., 2007; Welton et al., 2009b
Synthesis challenge 5 ($n = 4$): *Synthesis of randomised and non-randomised evidence*.
Efthimiou et al., 2017; Prevost et al., 2000; Rietbergen, 2016; Schmitz et al., 2013
Synthesis challenge 6 ($n = 12$): *Synthesis of potentially biased and unbiased studies*.
Chaimani and Salanti, 2012; Dias et al., 2010b; Eddy et al., 1990; Efthimiou et al., 2017; Mavridis et al., 2013; Salanti et al., 2010; Schmitz et al., 2013; Spiegelhalter and Best, 2003; Trinquart et al., 2012; Turner et al., 2009; Welton et al., 2009a; Wolpert and Kerrie, 2004
Synthesis challenge 7 ($n = 2$): *Synthesis of structurally related outcomes*.
Welton et al., 2008, 2010
Synthesis challenge 8 ($n = 5$): *Synthesis of studies reporting on multiple follow-ups*.
da Costa et al., 2017; Ding and Fu, 2013; Jackson et al., 2014; Lu et al., 2007; Musekiwa et al., 2016
Synthesis challenge 9 ($n = 22$): *Synthesis of correlated outcomes*.
Achana et al., 2014; Ades et al., 2010; Bujkiewicz et al., 2014, 2016; Daniels and Hughes, 1997; Efthimiou et al., 2014; Hong et al., 2016; Jackson et al., 2011, 2013; Jackson and Riley, 2014; Jackson et al., 2018; Madan et al., 2014; Mavridis and Salanti, 2013; Nam et al., 2003; Riley et al., 2007a, 2008; Van Houwelingen et al., 1993; van Houwelingen et al., 2002b; Wei and Higgins, 2013a,b; Welton et al., 2008, 2010
Synthesis challenge 10 ($n = 5$): *Incorporating evidence from previous meta-analyses*.
Higgins and Whitehead, 1996; Pullenayegum, 2011; Rhodes et al., 2015; Turner et al., 2015; Welton et al., 2009a

3.3.2. ‘Core’ assumptions of information-sharing

As the primary purpose of this review was to classify the breadth of methods into a handful of meaningful categories, efforts were made to identify information-sharing patterns. Across all the included papers, one or more of four ‘core’ methods was used to relate direct and indirect evidence. Each ‘core’ method was using a different underlying assumption. Those four method categories are described below:

The first type is functional relationships which include deterministic functions among model parameters that pertain to the direct and indirect evidence. They range from simple relationships such as lumping (i.e. equal effects), which mainly aim to reduce the number of parameters estimated by the model and thus assist parameter identifiability, to complicated functionals (e.g. dose-response curves) which might introduce additional parameters and may not be estimable under data sparsity conditions. The central assumption is the validity of the imposed function which is often untestable. When d_{Dir} is the parameter that relates to the direct evidence and d_{Indir_v} the parameter that relates to the $v-th$ indirect source, functional dependence can be expressed in the following way:

$$d_{Dir} = f(d_{Indir_1}, d_{Indir_2}, ..., d_{Indir_v}) \tag{3.1}$$

The second type is exchageability-based relationships, where a common distribution (usually normal) is imposed on a set of source-specific parameters which are hence treated as random draws from that distribution and shrink towards its mean (m) so that:

$$(d_{Indir_1}, d_{Indir_2}, ..., d_{Indir_v}) \sim N(m, \sigma) \tag{3.2}$$

The variance of the estimated distribution (σ) provides an indication of the extent of heterogeneity between evidence sets. Importantly, the source of heterogeneity is not explicitly modelled and heterogeneity is not explained, but only accounted for. The critical underlying assumption (i.e. exchangeability) assumes that the set of parameters on which the distribution is imposed do differ, but in a non-systematic way.

The third type is prior-based relationships, which aim to regularise the estimation of the parameters that pertain to the policy research question (i.e. the direct evidence) by using ‘informative’ prior distributions derived from the indirectly related evidence. This method is usually a two-step process where the indirectly related evidence is initially analysed to generate an informative prior which is then combined with the direct evidence in a Bayesian framework. Even though the assumption is similar to lumping (i.e. there are no differences between direct and indirect evidence), the effect of the prior on the posterior distribution will decrease as more direct information becomes available (Gelman et al., 2013). Importantly, prior-based methods allow the analyst to specify the perceived similarity between direct and indirect evidence and vary the degree of similarity by down-weighting the derived 'informative' priors if necessary. For D_v and $L(d_v|D_v)$ being, respectively the data and the likelihood of evidence set v, these relationships can be expressed as follows:

$$p(d_{Dir}|D_{Dir}, D_{Indir_1}, D_{Indir_2}, \cdots, D_{Indir_v}) \propto L(d_{Dir}|D_{Dir})\ x\ \pi_0$$

where

$$\pi_0 \propto f(L(d_{Indir_1}|D_{Indir_1}), L(d_{Indir_2}|D_{Indir_2}), \cdots, L(d_{Indir_v}|D_{Indir_v}))$$

The final type is multi-variate relationships. These assume that both direct and indirect parameters are correlated and thus their relative effects are multi-variately distributed (often- but not necessarily- multivariately normally distributed). Information-sharing is, hence, achieved through the flow of information between all correlated variables. For $d_{n,i}$, being the source (v) and study (i) specific parameter of interest, τ_n the source-specific between-studies variance in parameters $d_{n,i}$, and $\rho_{n,k}$ the between-studies correlation across the parameters of interest of sources n and k, multi-variate relationships can be expressed as:

$$\begin{pmatrix} d_{Dir_i} \\ \vdots \\ d_{Indir_{v,i}} \end{pmatrix} \sim MVN\left[\begin{pmatrix} d_{Dir} \\ \vdots \\ d_{Indir_v} \end{pmatrix}, \begin{pmatrix} \tau^2_{Dir} & \cdots & \rho_{Dir,Ind_v} \cdot \tau_{Dir} \cdot \tau_{Indir_v} \\ \vdots & \ddots & \vdots \\ \rho_{Dir,Ind_v} \cdot \tau_{Dir} \cdot \tau_{Indir_v} & \cdots & \tau^2_{Indir_v} \end{pmatrix}\right]$$

Table 3.3 classifies papers according to the 'core' relationship that they used and the PICOS level that the additional evidence was indirect to.

Table 3.3: *A categorisation of papers that share information on the relative effectiveness parameter according to the 'core' method that they use and the PICOS level that direct and indirect evidence differ in.*

Multiple Treatments
Functional ($n = 21$): Chaimani and Salanti, 2012; Cooper et al., 2009; da Costa et al., 2017; Dakin et al., 2011; Del Giovane et al., 2013; Dias et al., 2011a,b; Langford et al., 2018; Lu and Ades, 2009; Madan et al., 2014; Mawdsley et al., 2016; Melendez-Torres et al., 2015; Mills et al., 2012; Nixon et al., 2007; Owen et al., 2015; Soares et al., 2014a; Thorlund et al., 2013; Warren et al., 2014; Welton et al., 2009b,a; Wu et al., 2018
Exchangeability-based ($n = 14$): Achana et al., 2013; Chaimani and Salanti, 2012; Cooper et al., 2009; da Costa et al., 2017; Dakin et al., 2011; Del Giovane et al., 2013; Dominici et al., 1999; Lu and Ades, 2009; Moreno et al., 2011; Nixon et al., 2007; Owen et al., 2015; Soares et al., 2014a; Thorlund et al., 2013; Warren et al., 2014
Prior-based ($n = 0$): No references
Multivariate ($n = 1$): Nixon et al., 2007
Multiple Populations
Functional ($n = 2$): Duarte et al., 2017; Soares et al., 2014a
Exchangeability-based ($n = 3$): Achana et al., 2013; Dias et al., 2011c; Gamalo-Siebers et al., 2017
Prior-based ($n = 3$): Achana et al., 2013; Gamalo-Siebers et al., 2017; Roever et al., 2019
Multivariate ($n = 0$): No references
Multiple Outcomes
Functional ($n = 4$): Dakin et al., 2011; Ding and Fu, 2013; Lu et al., 2007; Welton et al., 2008
Exchangeability-based ($n = 2$): da Costa et al., 2017; Lu et al., 2007
Prior-based ($n = 0$): No references
Multivariate ($n = 30$): Achana et al., 2014; Ades et al., 2010; Bujkiewicz et al., 2014, 2016; Copas et al., 2018; Daniels and Hughes, 1997; Efthimiou et al., 2014, 2015; Hong et al., 2016, 2018b; Hwang and DeSantis, 2018; Jackson et al., 2011, 2018, 2014, 2013; Jackson and Riley, 2014; Liu et al., 2018; Lu et al., 2014; Madan et al., 2014; Mavridis and Salanti, 2013; Musekiwa et al., 2016; Nam et al., 2003; Riley et al., 2008, 2007a; Tan et al., 2018; Van Houwelingen et al., 1993; van Houwelingen et al., 2002b; Wei and Higgins, 2013a,b; Welton et al., 2008, 2010
Multiple Designs
Functional ($n = 12$): Chaimani and Salanti, 2012; Dias et al., 2010b; Eddy et al., 1990; Mavridis et al., 2013; Salanti et al., 2010, 2009; Spiegelhalter and Best, 2003; Trinquart et al., 2012; Turner et al., 2009; Wolpert and Kerrie, 2004; Welton et al., 2009a; Moreno et al., 2011
Exchangeability-based ($n = 4$): Efthimiou et al., 2017; McCarron et al., 2010; Prevost et al., 2000; Schmitz et al., 2013
Prior-based ($n = 7$): Efthimiou et al., 2017; Mak et al., 2009; McCarron et al., 2010, 2011; Rietbergen, 2016; Schmitz et al., 2013; Welton et al., 2009b
Multivariate ($n = 0$): No references

3.3.3. Information-sharing methods

This section presents the findings of the citation-mining review, categorising the identified papers and methods under the four ‘core’ relationships of information-sharing.

3.3.3.1 Functional relationships

The simplest functional relationship is lumping (i.e. common effects) where all data points inform a single parameter independently of whether the evidence is direct or indirect. Examples include pooling RTEs across time-points (Dakin et al., 2011) or (sub-)populations (Soares et al., 2014a; Duarte et al., 2017) as well as pooling between-trial heterogeneity parameters (Dias et al., 2011a) or meta-regression slopes (Cooper et al., 2009).

Constraints impose a strict inequality among parameters, facilitating information-sharing by preventing MCMC simulations that do not conform to the specified constraint. Such methods have been used to relate RTEs across dosages, expressing that higher dosages are expected to exhibit larger RTE (Owen et al., 2015; Del Giovane et al., 2013), describe structurally-related outcomes (Welton et al., 2008) and specify second-order consistency equations that impose a triangle inequality on the comparison-specific between-trial variances (Lu and Ades, 2009; Thorlund et al., 2013).

Meta-regression-type methods have also been suggested. In the examples found, the relationships were usually linear —on the modelling scale —with one RTE component independent and another RTE component dependent on a particular study characteristic (see e.g. Equation 2.13). The most common example of methods in this category is bias-adjustment, primarily used to synthesise studies of different designs. Bias-adjustment methods broadly fall into two categories: general frameworks that adjust the RTE for biases affecting internal and external validity provided that the extent of bias can be either estimated from empirical evidence or elicited from experts (Eddy et al., 1990; Wolpert and Kerrie, 2004; Turner et al., 2009; Spiegelhalter and Best, 2003), and approaches that adjust for bias due to particular study-level characteristics (considered proxies for study quality such as their size (Chaimani and Salanti, 2012; Trinquart et al., 2012; Mavridis et al., 2013; Salanti et al., 2010; Moreno et al., 2011), publication year (Salanti et al., 2009), or risk-of-bias (Dias et al., 2010b; Welton et al., 2009a). Meta-regression-type relationships have also been used for complex interventions (i.e. treatments that comprise of multiple components). In their simplest form, they model the cumulative RTE of a complex intervention as the sum of RTEs of its treatment components (Nixon et al., 2007; Madan et al., 2014; Mills et al., 2012; Melendez-Torres et al., 2015), whilst more sophisticated approaches allow for synergistic or antagonistic relationships by suggesting functions that also contain treatment interaction RTE components (Welton et al., 2009b). Other applications include approaches that model the RTEs measured in two survival outcomes (e.g. time-to-mortality and time-to-progression) by assuming that they only differ by a constant component which is invariant across treatment comparisons (Welton et al., 2010), methods that assume a linear relationship between dosage and RTE (Warren et al., 2014), and methods for baseline-risk adjustment (Achana et al., 2013).

Finally, more complex, non-linear, relationships have also been presented in the literature, namely those enabling the synthesis of RTEs across a range of dosages using the *Emax* model (Mawdsley et al., 2016; Wu et al., 2018; Langford et al., 2018) commonly employed in pharmacokinetics and those enabling the sharing of information across follow-up periods (Lu et al., 2007; Ding and Fu, 2013).

3.3.3.2 Exchangeability-based relationships

The simplest exchangeability-based relationship relates a set of parameters using a RE model which accounts for heterogeneity without explicitly modelling its source(s). The RE model assumes that all parameters are drawn from a distribution, implying that individual parameters are shrunk towards the RE mean; this can happen to a greater or lesser extent, depending on each individual estimate's precision and discrepancy from the RE mean. Examples include pooling RTEs of different dosages of the same treatment (Del Giovane et al., 2013), comparison-specific meta-regression slopes (Cooper et al., 2009; Achana et al., 2013; Chaimani and Salanti, 2012; Moreno et al., 2011), comparison-specific between-trial variances (Lu and Ades, 2009; Thorlund et al., 2013), and study-specific baseline-risks (Dias et al., 2011c; Achana et al., 2013).

Random-walks express the assumption that data points which are more similar with respect to a particular characteristic are expected to exhibit more similar RTEs. Examples include approaches assuming that the RTE of a particular dosage comes from a distribution that is centred around the RTE of its adjacently lower or higher dosage (Del Giovane et al., 2013) and models assuming that the RTE of a particular follow-up period is more similar and hence centred around the RTEs of adjacent follow-up periods than to RTEs of more 'distant' follow-ups (Lu et al., 2007; da Costa et al., 2017).

Multi-level models also impose the assumption of exchangeability, but also account for the hierarchical/clustered structure of the available data. As such, exchangeability is imposed once within specific groups of parameters (i.e. conditional on some characteristic) and across the group-specific hyper-parameters. For example, in the bottom level, studies may be categorised according to a characteristic and a different random-effect may be imposed within every category, producing group-specific means and heterogeneities. In the top-level, exchangeability may also be assumed across the group-specific means which are shrunk towards an overall, global, group-independent, hyper-mean. Examples include ‘class-effects’ models where, on top of the classical RE NMA models, the basic parameters of treatments that function through the same mechanism are assumed to be drawn from a common distribution with an overall ‘class’ mean and an across-treatments, within-class, heterogeneity (Soares et al., 2014a; Owen et al., 2015; Nixon et al., 2007; Dakin et al., 2011; Dominici et al., 1999; Warren et al., 2014). Class-effect approaches have also been imposed across comparison-specific meta-regression slopes (Cooper et al., 2009; Moreno et al., 2011). Multi-level models have been suggested to combine adult and paediatric evidence (Gamalo-Siebers et al., 2017) and studies of different designs (Prevost et al., 2000; Efthimiou et al., 2017; Schmitz et al., 2013; McCarron et al., 2010).

3.3.3.3 Prior-based relationships

Direct and indirect evidence can also be combined through the use of prior distributions. The process usually consists of two steps where initially the indirect evidence is analysed and subsequently the resulting distribution is used as a prior in the analysis of the direct evidence. Examples include the combination of adult and paediatric evidence (Gamalo-Siebers et al., 2017) or randomised and non-randomised evidence (Efthimiou et al., 2017; Schmitz et al., 2013; McCarron et al., 2010, 2011). The prior can be adjusted for bias or have its precision inflated (Efthimiou et al., 2017). Alternative ways to inform the prior include using meta-epidemiological evidence or expert elicitation. The former has been used primarily for bias-adjustment (Welton et al., 2009b), whilst both the former (Higgins and Whitehead, 1996; Turner et al., 2015; Pullenayegum, 2011) and the latter (Ren et al., 2018) have been used to define a prior distribution for the between-trials heterogeneity.

More nuanced prior-based approaches such as mixture priors have also been used. Here, the distribution representing the indirect evidence is not used at face value, but instead combined with a ‘vague’ component. The informative and vague parts are mixed according to weights that may be specified by the analyst or determined within the synthesis model. The resulting informative prior is typically heavy-tailed, and allows for *adaptive* information-borrowing which has been argued to be desirable under prior data conflict conditions (Roever et al., 2019); that is, when the direct and indirect evidence turn out to be substantially different from one another. Mixture priors have been used to combine evidence on RTE and between-studies heterogeneity across adults and children (Roever et al., 2019) and to analyse the study-specific baseline parameters from studies that enrol populations with different baseline risk (Achana et al., 2013). The use of mixture priors has also been discussed for the synthesis of randomised and non-randomised evidence (Efthimiou et al., 2017).

Finally, a flexible method, initially proposed by Ibrahim and Chen, 2000, is the power-prior. In this method, direct and indirect evidence are simultaneously analysed but the Bayesian likelihood of the indirect evidence is raised to a power scalar $0 \leqslant \alpha \leqslant 1$ which reflects the perceived similarity between the two evidence sets. When $\alpha = 1$ the results are equivalent to lumping and when $\alpha = 0$ it is identical to splitting. The power parameter, α, needs to be specified, and can be elicited (Rietbergen et al., 2016) or varied in sensitivity analyses (Spiegelhalter et al., 2004). Power-priors have been used to combine observational and randomised evidence whilst regulating the impact of observational evidence (Rietbergen, 2016) and their use has also been described for the simultaneous synthesis of adult and paediatric evidence (Gamalo-Siebers et al., 2017).

3.3.3.4 Multi-variate relationships

Multi-variate relationships have primarily been used to borrow strength across multiple outcomes. Multivariate meta-analysis correlates the various outcomes and may separate within- and between-studies correlations (Mavridis and Salanti, 2013). At the within-studies level, the study-specific correlations arise due to differences among the included patients, and indicate how the outcomes co-vary across individuals within the study. For example, patients who —due to a baseline characteristic that makes their disease more severe —show high values for outcome A, are also more likely to yield high values for outcome B. At the between-studies level, correlations arise mainly due to study-level differences such as the distribution of patient-level characteristics across studies. For instance, studies that enrol more severe cases and show high values for the mean of outcome A, are also more likely to result in high values for the mean of outcome B. These models are argued to potentially produce more precise estimates (Riley et al., 2007a) and mitigate outcome reporting bias (Hwang and DeSantis, 2018; Kirkham et al., 2012).

Multivariate methods have been developed to consider two (Van Houwelingen et al., 1993; van Houwelingen et al., 2002b; Nam et al., 2003), three, or more correlated outcomes (Wei and Higgins, 2013b; Jackson et al., 2011), accommodate the simultaneous analyses of multiple treatments (Efthimiou et al., 2014; Achana et al., 2014), and assess the relationship between surrogate and final outcomes (Daniels and Hughes, 1997; Bujkiewicz et al., 2016). Given that within-trial correlations are commonly unknown, authors have suggested using external data to inform prior distributions for these parameters (Bujkiewicz et al., 2014) or, when external data are not available, methods that approximate within-study co-variances (Wei and Higgins, 2013a). Further extensions can also handle missing data (Jackson et al., 2013) and allow modelling of heterogeneity and inconsistency using two different variance components (Jackson et al., 2018).

To accommodate cases where the within-trials correlations are unavailable and cannot be otherwise obtained, alternative models which require the same data as a univariate approach and do not separate within- and between-trials correlations have been suggested for MA (Riley et al., 2008; Hong et al., 2018a) and NMA (Efthimiou et al., 2014). When the overall correlation is not very strong, these models perform very similarly with their counterpart that separates the two correlations whilst preserving their benefits against the univariate approach.

Finally, some models only account for either the within- or the between- studies correlations. For example, to model mutually exclusive outcomes, it has been suggested to only account for the within-trials negative correlations which are induced by the competing risks structure of the data (i.e. the more patients that reach one outcome, the fewer that reach another) (Ades et al., 2010). Also, other approaches have only modelled the between-studies covariance matrix to allow simultaneous synthesis of multiple outcomes (Welton et al., 2008, 2010; Madan et al., 2014; Hong et al., 2016), accommodate outcomes reported at several follow-up periods (Jackson et al., 2014; Musekiwa et al., 2016) and enable information-sharing across different treatment components of complex interventions (Nixon et al., 2007).

3.4. Discussion

The aim of this study was to identify and classify evidence synthesis methods that have been used to combine evidence directly and indirectly relating to a research question. Given that there was a pattern in the main assumptions/mathematical relationships that the various methods utilised in order to facilitate information-sharing, the results are presented in four main categories. These are functional relationships, relationships based on exchangeability, prior-based relationships, and multivariate relationships. In addition, the most common evidence synthesis challenges were identified and listed in Table 3.2 along with the papers that developed methods to address them.

Interestingly, most of the identified ISMs were developed in a Bayesian setting. This is perhaps because the Bayesian framework naturally lends itself to information-sharing. Specifically, in Bayesian inference parameter estimation requires that the observed data are combined with a prior distribution to derive posterior conclusions. The prior distribution reflects any previous beliefs/knowledge about the parameters, and therefore provides an obvious 'vehicle' to load indirect evidence. However, ISMs that do not share information though prior distributions can also be implemented in a frequentist setting without any particular implications and there are several examples in the literature (Jackson et al., 2011, 2013; Wei and Higgins, 2013a; Riley et al., 2007a; Mavridis and Salanti, 2013).

The findings of this review have a number of important implications. As shown in Table 3.3, some 'core' relationships are preferred when information is shared across specific PICOS levels. For instance, most of the identified papers which share information across treatment comparisons either use functional or exchangeability-based relationships, and no example using priors was found. Also, papers that used multivariate relationships, did so to borrow strength across related outcomes and no paper used multivariate methods to borrow strength across populations or study-designs. This may be partly because the information required to implement multivariate methods for multiple populations or study-designs is usually unavailable. For instance, to estimate the between-study correlation across populations we would require evidence from studies enrolling and separately reporting on each population. Still, methods that were originally developed to share information on one PICOS level may be transferable to other levels.

This review highlights the breadth of methods that can facilitate information-sharing. Although, typically, particular relationships are used preferentially to share information on specific information-sharing contexts, it is likely that several methods are applicable and analysts would need to choose which method is more appropriate. This work highlights that appropriate considerations need to be made when choosing 'core' relationships and methods, as choices are likely to influence the degree of information-sharing. Specifically, method selection may be informed by the following considerations; the first is the plausibility of the assumptions imposed by the methods in the context of interest. The classification of methods according to the 'core' relationship that enables information-sharing, is expected to facilitate a clearer discussion about the plausibility of these assumptions in the decision context of interest.

The second is the degree of information-sharing that is imposed between direct and indirect evidence. There has been limited exploration of how much different methods borrow-strength from indirect evidence, though for multivariate methods, it has been noted that information-sharing is 'usually modest' (Jackson et al., 2011; Copas et al., 2018) and, sometimes, instead of 'borrowing-strength', multi-variate methods may end up 'borrowing-weakness' (Bujkiewicz et al., 2013). The few studies that have assessed the degree of information-sharing typically consider only the degree of precision gains (Jackson et al., 2017) rather than also examining how the point estimate changes —which is also important for decision-making. Further research to understand the extent to which different methods share information is warranted.

Finally, decision-makers may be interested in exploring different levels of information-sharing. One way to do that is by using prior-based methods that allow some control on the degree of information-sharing. For instance, an informative prior may use either the posterior distribution of the mean, or the predictive distribution of the indirect evidence. The former is equivalent to lumping, whilst the latter imposes less information-sharing. Similarly, power-priors allow a range of values to be used for α which determines the extent of information-sharing.

Whilst it is expected that the above identification of 'core' relationships is exhaustive, the use of citation-mining techniques may have missed relevant methods, particularly those outside of health research. Additionally, this review only looked for methods that shared information between evidence sets that address different research questions. Hence, methods such as commensurate priors which have been used to combine individual-patient data and aggregate-level evidence on the same research question (Hong et al., 2018b) could also be useful for combining evidence sets that pertain to different research questions, but were here considered outside of the scope of the search.

Overall, this is the first attempt to summarise and categorise the existing literature by classifying methods according to the 'core' assumption that they use to facilitate information-sharing. Despite the challenges described above, the identified papers allowed borrowing-of-strength patterns to emerge. Further research could explore the following questions: first, how can we determine whether indirect evidence is relevant? Second, how can the appropriateness of each ISM be assessed for the synthesis problem at hand? Finally, can the extent of information-sharing be quantified to assist transparent decision-making?

Chapter 4

Network meta-analytic methods that borrow strength from aggregate-level binary evidence from indirectly related populations

4.1. Chapter aims and structure

In the previous chapter, the literature was searched in order to identify evidence synthesis models which can be used to facilitate information-sharing across evidence sets that investigate different, yet related, research questions. Furthermore, the identified ISMs were classified into four 'core' relationships, each one using a different main assumption to relate direct and indirect evidence. However, the identified ISMs were not explained in detail, because they were applied for a variety of synthesis challenges, data structures, and outcome types. This chapter aims to provide a thorough description of the details underpinning the methods identified in Chapter 3, in order to aid transparency in method choice. Specifically, this chapter has the following main aims:

1. To describe in detail the ISMs that were identified in Chapter 3 in the context of a simple, commonly encountered, synthesis problem. This is the simultaneous synthesis of two sets of evidence, each one including a number of studies enrolling a different patient population, whilst allowing for information to be shared on the relative effectiveness parameter. It will be assumed here that studies provide only aggregate-level evidence.

2. To thoroughly explain and discuss the assumptions underpinning ISMs in the context of the synthesis problem at hand.

3. To provide programming and coding suggestions for *WinBUGS* (MRC Biostatistics Unit, 2010) with the purpose of advancing the accessibility and applicability of ISMs.

4. To provide a step-by-step process for the identification of applicable ISMs that can be used by others facing similar synthesis problems.

The remainder of this chapter is structured as follows. In Section 4.2, the synthesis problem that will be central to this chapter is described, and in Section 4.3 the classical NMA model is extended to account for multiple evidence sets, though without imposing any degree of information-sharing. Subsequently, in Section 4.4 models that borrow information from the indirect evidence to 'strengthen' the relative treatment effect estimate of the direct evidence are mathematically developed. Note here that multi-variate relationships are not included because the required information for such relationships is not provided in the synthesis problem under consideration. Finally, in Section 4.5, a process to identify applicable ISMs for similar synthesis problems is explained, while in Section 4.6 the strengths and limitations of this work are discussed along with directions for future research.

4.2. Definition of the synthesis problem

This section describes the synthesis problem that will be considered throughout this chapter. As a starting point, it is assumed that there are only two sets of evidence: one which provides information directly relevant to the research question, and another which provides indirectly related information, because it enrolled a different, yet related, population. The focal question which is then explored here is: *'What methods can be used to share information between direct and indirect evidence on relative effectiveness?'*

Except for the enrolled population, the two sets of evidence are identical in all other PICOS aspects. Specifically, they include only two-arm studies and test the same set of interventions, thus producing the same basic parameters. Also, both evidence sets assess effectiveness based on the a binary outcome (e.g. mortality) and there are no studies including patients from both populations while reporting separately for each. In order for the models to be widely applicable, it is assumed that only aggregate-level evidence is available; however, all models can be easily extended to incorporate IPD as well. Finally, it is beyond the scope of this chapter to present ways to explore heterogeneity within the evidence sets, and hence only FE and RE models are considered.

The data take the form shown in Table 4.1. Essentially, a study-level covariate X is constructed to indicate whether a study provides direct or indirect information. Hence, studies with the same covariate value belong to the same evidence set. If multiple indirect evidence sets were available, a categorical variable could be used and only one of its values would correspond to the directly relevant population, whilst all remaining values would pertain to different indirect evidence sets. Finally, if the indirect evidence sets can further be ordered according to their expected RTEs, the value of the categorical variable assigned to each evidence set should reflect the a priori expected ordering.

Table 4.1: *An example of an extended evidence base. A binary variable X is used to indicate the source that each study pertains.*

Study index	na[]	t[,1]	r[,1]	n[,1]	t[,2]	r[,2]	n[,2]	X
1	2	1	13	27	3	8	29	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	0
n_{dir}	2	1	10	22	3	9	30	0
n_{dir} +1	2	2	9	28	3	1	27	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_{dir} + n_{indir}$	2	1	3	19	3	4	19	1

n_{dir}: number of studies in the direct evidence set, n_{indir}: number of studies in the indirect evidence set, na: number of arms, $t[,k]$ treatment indicator for arm k, $r[,k]$: number of events in arm k, $n[,k]$: number of patients enrolled in arm k.

The synthesis problem described above was chosen for three main reasons: First, it is a simple problem allowing for the focus to remain on the explanation of the methods. Second, it is relatively common in HTA and therefore analysts may be able to use the methods and the provided code with little to no adaptation. Finally, this synthesis problem links with the applied work that is undertaken in the next chapters, and is hence convenient in the sense that methods will not have to be repeated, with only the unavoidable modifications requiring further explanation.

4.3. The ‘splitting-model’

To accommodate the simultaneous synthesis of both direct and indirect evidence without imposing any information-sharing, for population j, the basic NMA model described on page 27 is extended here so that:

$$r_{i,k} \sim Bin(p_{i,k}, n_{i,k}) \tag{4.1}$$

$$logit(p_{i,k}) = \theta_{i,k} = \mu_{i_b} + \delta_{i,bk} \tag{4.2}$$

where under a FE model:

$$\delta_{i,bk} = d^j_{bk}$$

whereas under a RE model:

$$\delta_{i,bk} \sim N(d^j_{bk}, {\tau^j}^2)$$

$$d^j_{bk} = d^j_{1k} - d^j_{1b} \tag{4.3}$$

$$d^j_{11} = 0 \tag{4.4}$$

where d^j_{bk}, d^j_{1k}, τ^j are the evidence set-specific relative treatment effects, basic parameters, and between-studies heterogeneity respectively. Since no studies report for both populations, the population index is nested within the study index and hence parameters that have the study index (i) do not need a population index (j). Therefore, by defining population specific relative effects d^j_{bk} and basic parameters d^j_{1k} no information-sharing is allowed across populations; vague priors are assigned to all parameters.

The model can be applied in *WinBUGS* using the code developed by Dias et al., 2011a as shown below. However, two separate loops are defined, one for each evidence set, and different names are being used for parameters that represent population-specific quantities and are not study-specific. For instance, $\delta_{i,k}$ are study-specific and the evidence set that each $\delta_{i,k}$ pertains to is nested within the i index. In contrast, d^j_{1k}, τ^j are assigned names specific to each evidence set.

```
model{ # *** Program starts

    for(i in 1:ns.dir){                                            # LOOP THROUGH DIRECT STUDIES
w[i,1] <- 0              # multi-arm adjustment is zero for control arm
delta[i,1] <- 0          # treatment effect is zero for control arm
mu[i] ~ dnorm(0,.0001) # vague priors for all trial baselines
    for (k in 1:na[i]) {                                                     # Loop through arms
r[i,k]   dbin(p[i,k],n[i,k])        # binomial likelihood
logit(p[i,k]) <- mu[i] + delta[i,k] # model for linear predictor
            }                                                                  # Arms loop closes
    for (k in 2:na[i]){                                                      # Loop through arms
delta[i,k] ~ dnorm(md[i,k],precd[i,k])          # trial-specific LOR
md[i,k] <- d.dir[t[i,k]] - d.dir[t[i,1]] + sw[i,k]    # mean of LOR
precd[i,k] <- prec.dir *2*(k-1)/k                # precision of LOR
w[i,k] <- (delta[i,k] - d.dir[t[i,k]] + d.dir[t[i,1]]) # multi-arm adjustment
sw[i,k] <- sum(w[i,1:k-1])/(k-1)                 # cumulative multi-arm adjustment
            }                                                                  # Arms loop closes
            }                                                      # DIRECT STUDIES LOOP CLOSES
d.dir[1]<-0                                # tr.effect is zero for ref.treatment in direct studies
for (k in 2:nt){ d.dir[k] ~ dnorm(0,.0001) } # vague priors for direct basic params
tau.dir ~ dunif(0,2)                       # between-trial SD
prec.dir <- pow(tau.dir,-2)                # between-trial precision

    for(i in ns.dir+1:ns.dir+ns.indir){                          # LOOP THROUGH INDIRECT STUDIES
w[i,1] <- 0              # multi-arm adjustment is zero for control arm
delta[i,1] <- 0          # treatment effect is zero for control arm
mu[i] ~ dnorm(0,.0001) # vague priors for all trial baselines
    for (k in 1:na[i]) {                                                     # Loop through arms
r[i,k]   dbin(p[i,k],n[i,k])          # binomial likelihood
logit(p[i,k]) <- mu[i] + delta[i,k]   # model for linear predictor
            }                                                                  # Arms loop closes
    for (k in 2:na[i]){                                                      # Loop through arms
delta[i,k] ~ dnorm(md[i,k],precd[i,k])              # trial-specific LOR
md[i,k] <- d.indir[t[i,k]] - d.indir[t[i,1]] + sw[i,k]     # mean of LOR
precd[i,k] <- prec.indir *2*(k-1)/k                 # precision of LOR
w[i,k] <- (delta[i,k] - d.indir[t[i,k]] + d.indir[t[i,1]]) # multi-arm adjustment
sw[i,k] <- sum(w[i,1:k-1])/(k-1)                    # cumulative multi-arm adjustment
            }                                                                  # Arms loop closes
            }                                                    # INDIRECT STUDIES LOOP CLOSES
d.indir[1]<-0                                # tr.effect is zero for ref.treatment in indirect studies
for (k in 2:nt){ d.indir[k] ~ dnorm(0,.0001) } # vague priors for indirect basic params
tau.indir ~ dunif(0,2)                       # between-trial SD
prec.indir <- pow(tau.indir,-2)              # between-trial precision

} # *** Program ends
```

4.4. Information-sharing methods

This section describes models that can be used to share information on RTEs for the synthesis problem described in Section 4.2. This is sharing information between two evidence sets enrolling different populations; one directly related to the research question and one only indirectly relevant. The models consider information-sharing on the RTE mean, d, and between-studies heterogeneity, τ, (under RE) and are summarised in Table 4.2 and Table 4.3 respectively. The evidence set that each parameter pertains to is denoted in their subscript (e.g. $d^{Dir}_{1k}, d^{Indir}_{1k}$). Where possible, it is also shown how the methods can be extended to consider the case of one direct and multiple indirect sources (i.e. $d^{j}_{1k} = d^{Dir}_{1k}, d^{Indir_1}_{1k}, d^{Indir_2}_{1k}, d^{Indir_3}_{1k}, ..., d^{Indir_J}_{1k}$). *WinBUGS* code for all the models can be found in `https://github.com/NikolaidisGFZ/PHD.git` Models are organised under the 'core' relationship that they impose to facilitate information-sharing as introduced on page 49. Note that multi-variate relationships are not included here, because the necessary information that is required for such relationships is not provided in the synthesis problem under consideration. This is further discussed in Section 4.6.

4.4.1. Functional relationships

This section covers relationships that take the form of deterministic functions among parameters that pertain to the direct and the indirect evidence. These include lumping, constraints, and meta-regression/bias-adjustment approaches. Their main assumption is the validity of the imposed deterministic relationship which is often un-testable.

4.4.1.1 Lumping

Being the easiest to implement and most commonly used ISM, lumping simply ignores any differences between direct and indirect evidence. Under FE models, this is achieved by assuming that $d^{Dir}_{1k} = d^{Indir}_{1k}$, which is equivalent to dropping the j index and analysing all studies as if they pertain to the same set of evidence. Under RE, however, both d and τ are components of the RTE and lumping can be imposed on either or both. For instance, if the two sets of evidence are expected to exhibit the same point estimate for the relative effect but one set of evidence is evidently more heterogeneous than the other, then we could just assume that $d^{Dir}_{1k} = d^{Indir}_{1k}$ and allow for population-specific heterogeneities τ^{j}. In contrast, the two sources may only be equally heterogeneous without equal RTE means, so that only $\tau^{Dir} = \tau^{Indir}$ and the point estimate is population-specific d^{j}_{1k}. Finally, if the direct and indirect evidence sets are expected to yield both equal d and τ, a RE can be imposed across all studies.

4.4.1.2 Constraints

When it can be reasonably assumed that the RTEs of the two populations should follow a particular ordering (e.g. that the RTE is expected to be larger in the indirect population), this assumption can be explicitly reflected in the model. As suggested by Owen et al. (2015), this can be achieved by defining a new variable γ which takes the form of a step function $I(x)$, so that:

$$\gamma = \begin{cases} I(x) = 0, \text{ if } x \leqslant 0 \\ I(x) = 1, \text{ if } x > 0 \end{cases}$$

The expected ordering can then be expressed by appropriately defining x and forcing γ to take a specific value. For example, if we expect a larger reduction in mortality in the direct population than in the indirect population (i.e. a lower log-odds ratio), then we can define $x = d_{1k}^{Indir} - d_{1k}^{Dir}$ and $\gamma = 1$, so that $d_{1k}^{Indir} - d_{1k}^{Dir} > 0$ and therefore $d_{1k}^{Indir} > d_{1k}^{Dir}$. To apply a constraint in *WinBUGS* the code displayed below can be added to the splitting model, outside the population-specific loops, for every basic parameter $1k$ the constraint is applied on.

```
b <- 1
b ~ dbern(constraint)
constraint <- step(d.ind[1k] - d.dir[1k]) # e.g. for basic parameter 1k
```

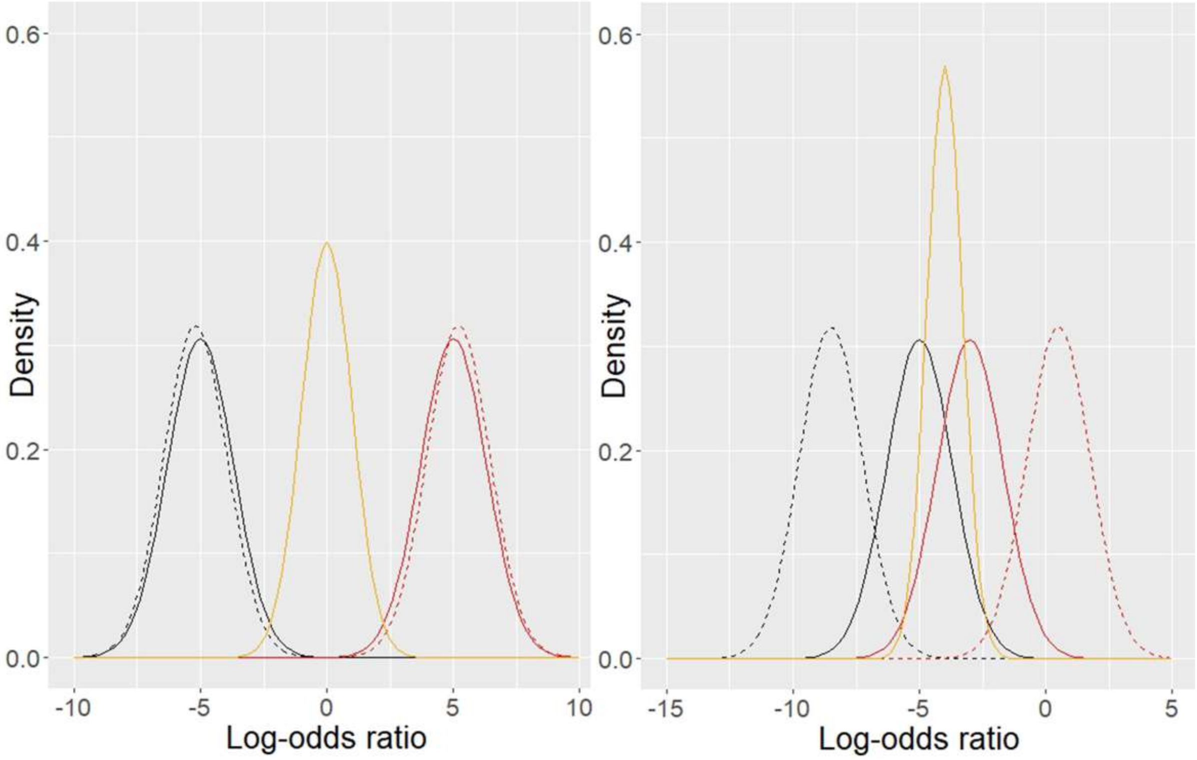
If there were more sets of indirect evidence, the RTEs of which were expected to follow an underlying ordering, the model could be extended by using an ordinal distinguishing variable. For instance, if we simultaneously analyse evidence that pertain to patient subgroups of differing disease severity, one of which is the population directly considered, and we expect the RTE to increase for more severe disease (i.e. $d_{1k}^{j=1} < d_{1k}^{j=2} < ... < d_{1k}^{j=J}$), we can express this belief by defining the following function $\gamma = 1$:

$$\gamma = \prod_{j=1}^{j=J} I(d_{1k}^{j+1} - d_{1k}^{j})$$

where d_{1k}^{j+1} is the RTE that pertains to the more severe patients than d_{1k}^{j}. By assuming that the product of all $I(x)$ is equal to 1, we are effectively assuming all $I(x) = 1$ and hence all $x = d_{1k}^{j+1} - d_{1k}^{j} > 0$.

It is worth noting that this model does not imply that the posterior distributions for d_{1k}^{Dir} and d_{1k}^{Indir} cannot overlap, but only that the expressed ordering is preserved within each MCMC iteration. Crucially, the notion of information-sharing may seem counter-intuitive in this case, because increasing information-sharing implies that parameters become more distant instead of more similar. For instance, as illustrated in Figure 4.1, when direct and indirect evidence suggest very different evidence set-specific RTEs (left hand-side graph), imposing a constraint of the correct direction will have minimal effect on the RTE estimates, whilst lumping would majorly affect estimates. In contrast, when direct and indirect evidence suggest very similar RTEs (right hand-side graph), lumping will lead to minimal changes, but the constraint is expected to produce a major shift in the RTE means of the two evidence sets.

Figure 4.1: *An example of how constraints can be very informative under particular circumstances.*



Black and red solid lines correspond to estimates of the direct and the indirect relative effects under a splitting approach, whilst dotted lines correspond to estimates using a model that imposes constraints. The solid orange line represents lumping estimates. On the left, the direct and indirect evidence are very different -distant- and imposing a constraint of the correct direction results in only minor changes in the estimates RTEs. On the contrary, on the right, direct and indirect evidence are more similar and overlap considerably. As a result, imposing a constraint results in larger changes in the estimated mean effects.

4.4.1.3 Meta-regression (and bias-adjustment)

Since the population indicator is a binary study-level variable, meta-regression models can also be applied. For this, the j index is dropped from the splitting model and Equation 4.2 is modified so that:

$$logit(p_{i,k}) = \theta_{i,k} = \mu_i + \delta_{i,1k} + \beta_{1k} \cdot X_i \tag{4.5}$$

where β_{1k} is a treatment comparison-specific component that represents the additional RTE (log-odds ratio) that is exhibited in the indirect population and is assigned a vague prior. Most often β_{1k} is assumed to be comparison-invariant (i.e. $\beta_{1k} = \beta$) to assist parameter identification, however alternative modelling approaches are explained in detail in (Cooper et al., 2009). This model effectively assumes that there is an additional RTE component in the indirect population which is the same for any treatment comparison and can be implemented using material that is supplied in (Dias et al., 2011b).

Despite the fact that this approach allows us to test the interaction between the relative effect and the population covariate, it does not facilitate any information-sharing between direct and indirect evidence as long as β parameters remain comparison-specific. This is because no studies report for both populations, and therefore the indirect evidence will provide information for the estimation of β, whilst the direct evidence will provide information for the estimation of δ. However, if evidence from multiple patients subgroups is available and X can be used as a cardinal variable, then information would be shared across all evidence sets in which $X \neq 0$ in the estimation of β. The model would then assume that as X increases, the additional relative treatment effect remains constant or, in other words, that the relative effect increases linearly with X.

Since this model is essentially a bias-adjustment approach, ideas that were developed in Welton et al., 2009a could potentially be used to enable further information-sharing if appropriate meta-epidemiological data were available. Such an approach would facilitate information-sharing through the use of a prior distribution by substituting the vague prior on β (i.e. the difference between direct and indirect evidence) with an informative prior derived from previous meta-analyses. Another approach could be to treat indirect studies as being externally biased, extend the model to accommodate study-specific bias components β_i and elicit a distribution for the study-specific bias components from experts according to methods that were suggested by Turner et al., 2009.

4.4.2. Exchangeability-based relationships

This sections covers methods that employ the assumption of exchangeability. That is the assumption that a set of parameters do not differ in a systematic manner. Methods include multi-level models and random-walks.

4.4.2.1 Multi-level model

When the same treatments have been evaluated in different evidence sets and the expected RTEs cannot be distinguished with an ordinal variable[1], but only with a categorical —unordered—variable[2], multi-level models are still applicable. Under this approach, the 'splitting-model' is extended so that:

$$d^{j}_{1k} \sim N(D_{1k}, \phi_{1k}) \tag{4.6}$$

where population- and comparison-specific basic parameters are normally distributed with a 'global', population-invariant, comparison-specific mean D_{1k}, and a comparison-specific, between-populations, variance ϕ_{1k}. The variance is indicative of the between-populations heterogeneity on the basic parameters and may be assumed equivalent across treatment comparisons (i.e. $\phi_{1k} = \phi$). At the bottom level this method performs NMA separately across studies within each population, whilst at the top level it performs a meta-analysis of the population-specific basic parameters applying a random-effect (see Figure 4.2). Under RE, the model contains three-levels overall, whilst under FE two levels. The model assumes that evidence set-specific basic parameters d^{j}_{1k} do not systematically differ across populations and therefore allows them to shrink towards a comparison-specific hyper-mean that is independent of population D_{1k}. It is worth noting that this approach requires that the same basic parameters can be obtained from every evidence set and that it is not necessarily applicable when different evidence sets compare different interventions. Efthimiou et al., 2017 discuss approaches for which be more appropriate under such circumstances.

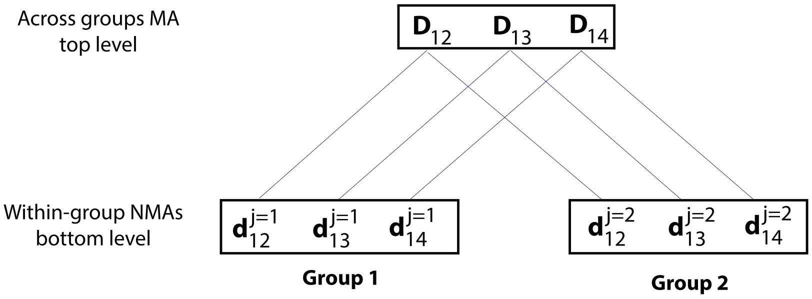
Importantly, multi-level models not only require enough studies within each population in order for the population-specific between-trial variance to be adequately identified, but also require evidence from multiple different patient groups so that the between-populations heterogeneity is identified. Gelman, 2006a suggested that when a RE is

[1] e.g. If the two evidence sets relate to a more severe subgroup and a less severe subgroup of patients, it may be reasonable to expect a priori that the RTE pertaining to the more severe subgroup is larger than that of the less severe subgroup.

[2] e.g. if the evidence sets relate to different disease sub-types such as different gene mutations, it is unlikely that there will be a rationale to expect, a priori, that the RTE pertaining to the subgroup of one mutation is higher or lower than the RTE of the other mutation subgroup, and hence they cannot be ordered.

used, at least four or five data-points should be available to prevent implausibly high or low variance values. However, here, at the top level, we pool population-specific means instead of study-specific observations, which may be less erratic and hence we may be able to obtain realistic estimates of the between-populations heterogeneity with less than five population groups. Alternative options for the application of multi-level models with a small number of groups include the use of an informative prior for ϕ that is produced either by eliciting from experts the expected dispersion of the results across population groups, or by using meta-epidemiological data; although, no such approach came up in the review of Chapter 3.

Figure 4.2: *An illustration of the multi-level model.*


Across groups MA
top level
D_{12} D_{13} D_{14}
Within-group NMAs
bottom level
$d_{12}^{j=1}$ $d_{13}^{j=1}$ $d_{14}^{j=1}$
$d_{12}^{j=2}$ $d_{13}^{j=2}$ $d_{14}^{j=2}$
Group 1
Group 2

At the bottom level, NMA is performed within each evidence set to produce evidence set-specific basic parameters, d^{j}_{1k}. At the top level, basic parameters are pooled across evidence sets using an additional random-effect. This process results in the estimation of evidence set-independent basic parameter D_{1k} and a between-sets heterogeneity.

The *WinBUGS* code that extends the 'splitting model' to impose a multi-level model across four evidence sets is given below. All source-specific basic parameters are assumed to be drawn from a common overarching normal distribution, the hyperparameters of which are independent of the source and are assigned vague priors.

```
for(k in 2:nt) {
d.dir[k] ∼ dnorm(d[k], prec.pop) # d[k] is hyper-mean
d.indir1[k] ∼ dnorm(d[k], prec.pop)
d.indir2[k] ∼ dnorm(d[k], prec.pop)
d.indir3[k] ∼ dnorm(d[k], prec.pop)
d[k] ∼ dnorm(0, .001 ) # vague prior for every basic parameter hyper-mean
}
prec.pop <- pow(tau.pop, -2)
tau.pop ∼ dunif(0,2) # vague prior for between populations variance
```

4.4.2.2 Random-Walk

Approaches based on random-walks have also been suggested in the literature. Under this method, if there is just one indirect evidence set, the following relationship is assumed:

$$d_{1k}^{Dir} \sim N(d_{1k}^{Indir}, \eta) \tag{4.7}$$

where the RTE of the direct evidence is drawn from a normal distribution centered around the RTE of the indirect evidence set. The variance of this distribution, η, is estimated within the model, typically as a comparison-independent parameter, and represents the plausibility of sharing information on relative effectiveness between the two evidence sets. Crucially, for the case of only two evidence sets, the random-walk model is very similar to the multi-level model and is expected to yield comparable results because both models shrink the direct and indirect RTEs closer to each other.

Despite the fact that this method can be used when only one indirect evidence set exists, it was initially developed —and is better suited —for cases where there are multiple indirect evidence sets which can be ordered in terms of their expected RTE according to some ordinal characteristic. The model then assumes that the RTE that pertains to one group is more similar to the RTEs of adjacent groups (i.e. groups that are more similar with respect to the characteristic in question) than to the RTE of groups that are more 'distant'. This is reflected by extending Equation 4.7 so that:

$$d_{1k}^{j+1} \sim N(d_{1k}^{j}, \eta) \tag{4.8}$$

where, if population groups are for example ordered according to severity, then d_{1k}^{j+1} is the basic parameter for comparison $1k$ of the population group with higher severity which is drawn from a normal distribution with a mean that is the basic parameter $1k$ of the adjacently lower severity group i.e. d_{1k}^{j}. The variance, η, is here assumed common across comparisons and groups and assigned a vague prior. If, however, the difference between d_{1k}^{j} and d_{1k}^{j+1} can be expressed based on a continuous characteristic, the variance may depend on this difference so that it is larger when the distance between d_{1k}^{j} and d_{1k}^{j+1} increases (Del Giovane et al., 2013). Finally, a vague prior needs to be assigned to the basic parameter that pertains to the population group with the lowest j. The *WinBUGS* code that can be used to extend the 'splitting-model' in order to implement a random-walk across four evidence sets is shown below.

```
# Say that there are 4 population groups that can be ordered
# according to some categorical characteristic (e.g. severity)
# in the following manner: d.indir1, d.indir2, d.indir3, d.dir
for(k in 2:nt){
d.dir[k] ∼ dnorm(d.indir3[k], prec.pop)
d.indir3[k] ∼ dnorm(d.indir2[k], prec.pop)
d.indir2[k] ∼ dnorm(d.indir1[k], prec.pop)
d.indir1[k] ∼ dnorm(0, .001 ) # vague prior for last group
}
prec.pop <- pow(eta.pop, -2)
eta.pop ∼ dunif(0,2)
```

4.4.3. Prior-based relationships

This section covers relationships that are based on priors. These employ a Bayesian framework to 'load' the indirect evidence on priors that inform parameters of the direct evidence. They include commensurate priors, standard informative priors, mixture priors, and power-priors.

4.4.3.1 Commensurate prior

Commensurate priors, which were recently applied in order to simultaneously synthesise individual-level and aggregate-level evidence (Hong et al., 2018b), can also be adapted for the case of multiple population groups. In this approach, the priors for the basic parameters of the direct evidence are centred around the basic parameters of the indirect evidence and the variance of the prior controls the extent of borrowing of strength. Essentially,

$$d_{1k}^{Dir} \sim N(d_{1k}^{Indir}, \eta_{1k}) \tag{4.9}$$

$$\text{where} \quad \frac{1}{\eta_{1k}} \sim \begin{cases} N(20,1) & ,\text{if } c_{1k} = 0 \\ Gamma(0.1,0.1)I(0.1,5) & ,\text{if } c_{1k} = 1 \end{cases} \tag{4.10}$$

$$\text{and} \quad c_{1k} \sim Bernoulli(p_{1k}) \tag{4.11}$$

where η_{1k} are the comparison-specific variances that can be assumed invariant across comparisons (i.e. $\eta_{1k} = \eta$). When η_{1k} is very low, it imposes strong information-sharing by requiring that d_{1k}^{Dir} and d_{1k}^{Indir} be very similar (thus forcing commensurability), whilst when η is very high, it imposes minimal information-sharing, if any, by effectively disengaging d_{1k}^{Dir} and d_{1k}^{Indir}. This is achieved by imposing a 'spike-and-slab' hyper-prior on the precision (i.e. $\frac{1}{\eta_{1k}}$) which puts a probability p_{1k} on a high precision value 'spike', forcing strong borrowing of strength from the indirect evidence, and a probability of $1 - p_{1k}$ on a very low precision value 'slab' that imposes minimal information-sharing. The first is here expressed using a normal distribution arbitrarily centered around 20 and given a low standard deviation of 1, whilst the latter using a truncated Gamma distribution with shape and rate parameters of 0.1[3]. c_{1k} are independent Bernoulli trials i.e. $c_{1k} \sim Bernoulli(p_{1k})$. Since the probability placed on the spike controls the extent of commensurability, it also determines the strength of information-sharing between direct and indirect evidence, and hence the strength of the assumption that the method imposes. One option is for p_{1k}

[3] Indicatively, a random simulation of 1000 samples from $\sim Gamma(shape = 0.1, rate = 0.1)$ yields a collection of values with the following quantiles $0.05\% = 2.45 \cdot 10^{-13}, 0.25\% = 8.18 \cdot 10^{-6}, 0.5\% = 5.24 \cdot 10^{-3}, 0.75\% = 3.4 \cdot 10^{-1}, 0.95\% = 6.9$.

to be fixed to an arbitrary value (e.g. $p_{1k} = 0.5$). Alternatively, they can be assigned a vague hyper-prior, such as $p_{1k} \sim Beta(1,1)$, in order to be estimated within the model and hence potentially facilitate adaptive information borrowing. However, such an approach estimates more parameters and may lead to increased uncertainty.

When there are only two evidence sets, the commensurate-prior is similar to the random-walk in the sense that they both assume that d^{Dir} is drawn from a distribution that is centered around d^{Indir}. However, unless p_{1k} is set to a value close to 1, the commensurate prior does not estimate the variance within the model, but imposes additional assumptions on it, encouraging further information-sharing.The commensurate prior is therefore expected to share information more 'strongly' than the random-walks. The model can be extended to accommodate multiple indirect evidence sets in the same way as done for the random-walks i.e. by extending Equation 4.9 so that $d_{1k}^{j+1} \sim N(d_{1k}^{j}, \eta_{1k})$.

WinBUGS code that extends the splitting model to implement the commensurate-prior is shown below. Vague priors are assigned to the basic parameters of the indirect evidence, whilst for the basic parameters of the direct evidence commensurate priors are used.

```
for(k in 2:nt) { d.dir[k] ~ dnorm(d.indir[k], ssprec[k]) }
# ssprec = 1000 would force commensurability
# ssprec = 0.001 would disconnect d.dir and d.indir
for (k in 2:nt) {
tee[k,1] ~ dnorm(20,1) # Spike
tee[k,2] ~ dgamma(0.1, 0.1)I(0.1, 5) # Slab
flip[k] ~ dbern(prob) # $p_k$ is considered an uncertain parameter and is estimated
within the model. Silence if p deterministic
# flip[k] ~ dbern(0.5) # $p_k$ is assumed 0.5. Un-silence for deterministic p
pick[k] <- flip[k] + 1 # convert 0,1 to 1,2 to match tee indexing
ssprec[k] <- tee[k,pick[k]]
sstau[k] <- sqrt(1/ssprec[k])
}
prob ~ dbeta (a, b) # Silence if p deterministic
a ~ dunif (0, 1000) # vague prior for hyperparam a. Silence if p deterministic
b ~ dunif (0, 1000) # vague prior for hyperparam b. Silence if p deterministic
for (k in 2:nt) { sstau[k]<-sstau.com } # use for comparison-independent variance
```

4.4.3.2 Informative prior

Informative priors can also be used to inform the RTE and between-studies heterogeneity of the direct evidence. Although all previous models simultaneously analysed all evidence sets, this does not have to be the case. Two-step approaches can also be used, whereby in the first step the indirect evidence is analysed using NMA with vague priors, followed by a second step which takes these posterior estimates and uses them as prior information in an analysis of the direct evidence. In the simplest case, the evidence of the directly relevant population is analysed with informative priors placed on the basic parameters, defined on the basis of a separate previous analysis of the indirect data, so that:

$$d_{1k}^{Dir} \sim N(d_{1k}^{Indir}, V_{1k}^{Indir}) \tag{4.12}$$

where d_{1k}^{Indir} is the posterior mean of the basic parameters that result from the NMA of the indirect evidence, and V_{1k}^{Indir} their corresponding variance. If the indirect evidence has been analysed using a FE model, then the variance would be the square of the standard error of the posterior mean and the model would be expected to yield results equal to those under lumping.

If a RE approach is employed for the analysis of the indirect evidence to form a prior for d_{1k}^{Dir}, we can use either the posterior mean and its associated uncertainty for d_{1k}^{Indir} i.e. $\sim N(d_{1k}^{Indir}, se^2{}_{1k}^{Indir})$, or its predictive distribution i.e. $\sim N(d_{1k}^{Indir}, se^2{}_{1k}^{Indir} + \tau^2)$. Since we are not trying to predict the relative effect of a future 'rollout', but rather trying to characterise d_{1k}^{Dir}, which is the random-effect mean (i.e. the hyperparameter), the former is more appropriate than the latter. Although the posterior distribution would serve as the appropriate prior for the RE mean, the predictive distribuition has been used before (Efthimiou et al., 2017) to impose less information-sharing (due to its larger variance).

A similar approach can be undertaken if we wish to share information only on the heterogeneity component. Once again, in this case we have to start with an analysis of the indirect evidence, and subsequently use the posterior distribution of the heterogeneity parameter τ^{Indir} as a prior for τ^{Dir} in the analysis of the direct evidence. Importantly, the type of prior distribution would need to be chosen so that it respects the nature of the variance parameter, which must be strictly positive. Following previous work by Turner et al., 2015, a *log-normal* prior is suggested here, so that:

$$\tau_{Dir} \sim Lognormal(\mu_{Indir}, {\sigma_{Indir}}^2) \tag{4.13}$$

where μ_{Indir} and σ_{Indir} are the mean and the standard deviation in the logarithmic scale and can be obtained by extracting the CODA for τ_{Indir} and fitting a log-normal distribution.

Other prior distributions, such as the inverse gamma *half-Cauchy* and *half-t*, have also been used for τ (Gelman, 2006b). If there are multiple sources of indirect evidence or several historical meta-analyses including studies with similar characteristics, we could perform a meta-analysis of meta-analyses, just as Turner et al., 2015 did, in order to estimate the predictive distribution for heterogeneity (τ_{new}) and use it as prior information for heterogeneity in the analysis of the direct evidence.

Finally, when there are multiple sets of indirect evidence, the analyst would have to choose how to analyse all indirect sets as part of the first step, in order to produce an informative prior that can be used in the second step. A simple option might be to lump all indirect evidence, whilst a more complicated approach could potentially analyse the indirect evidence in a multi-level framework and use the predictive distribution of the top-level, incorporating the uncertainty between evidence-sets in the RTE hyper-means as informative priors for the analysis of the direct evidence.

4.4.3.3 Mixture prior

An extension of the informative prior is the mixture of prior. In this approach, the posterior estimates of the initial analysis of the indirect evidence are mixed with a non-informative prior to form comparison-specific mixture priors. These are then imposed on the basic parameters of the direct evidence, so that:

$$d_{1k}^{Dir} \sim \nu \cdot N(d_{1k}^{Indir}, V_{1k}^{Indir}) + (1-\nu) \cdot N(0, 10^4) \tag{4.14}$$

where, as before, d_{1k}^{Indir} and V_{1k}^{Indir} are the posterior mean and variance of the basic parameters that result from the NMA of the indirect evidence under a FE model, or, under a RE model, the corresponding parameters of either the predictive distribution or the posterior mean (see relevant discussion on page 74). The parameter ν is the weight that is placed on the informative component, and $(1-\nu)$ the weight of the vague component which ensures that the weights sum to 1. The value of ν reflects the plausibility of sharing information between the direct and the indirect evidence sets on relative effectiveness. It follows that for $\nu = 1$, the analysis becomes equivalent to the previous described classical informative prior, whilst for $\nu = 0$, equivalent to splitting.

The model can be extended to accommodate multiple sources of indirect evidence as long as the same basic parameters can be estimated in separate evidence set-specific NMAs. Essentially, each source is reflected in a separate component and a vague component is added in the following way:

$$d_{1k}^{Dir} \propto \nu_1 \cdot N(d_{1k}^{Indir_1}, V_{1k}^{Indir_1}) + \cdots + \nu_J \cdot N(d_{1k}^{Indir_J}, V_{1k}^{Indir_J}) + [1 - \sum_{i=1}^{J} \nu_i] \cdot N(0, 10^4) \qquad (4.15)$$

where all parameters are specific to indirect evidence set j.

```
# For comparison 12
lambda2[1] ~ dnorm(d.ind[2], prec.d.ind[2]) # Informative component
lambda2[2] ~ dnorm(0, .001) # Flat component
P2[1:2] ~ ddirch(alpha[]) # uncertain probabilities vector
# P2[1] <- 0.5 # Un-silence for deterministic probabilities vector
# P2[2] <- 0.5 # Un-silence for deterministic probabilities vector
T2 ~ dcat(P2[]) # Draw a categorical indicator according to P
d.dir[2] <- lambda2[T2] # Prior

# For comparison 13
lambda3[1] ~ dnorm(d.ind[3], prec.d.ind[3])
lambda3[2] ~ dnorm(0, .001)
P3[1:2] ~ ddirch(alpha[]) # uncertain probabilities vector
# P3[1] <- 0.5 # Un-silence for deterministic probabilities vector
# P3[2] <- 0.5 # Un-silence for deterministic probabilities vector
T3 ~ dcat(P3[])
d.ad[3] <- lambda3[T3]
```

Regarding the choice of weights, analysts have several options. One option is to arbitrarily specify some weight to be placed on the informative component. For example, the vague and informative components can each be set to 50%, with sensitivity analyses then conducted to explore the impact of different weights. However, if results are indeed sensitive to the choice of weights, this approach does not provide any insight into which set of weights is the most credible. Another option is to elicit from experts how relevant each evidence set is to the research question, and then to normalise weights accordingly. Finally, a Dirichlet prior can be placed on the weights to allow them to be estimated within the model, potentially providing an adaptive way of information-sharing. This is achieved by estimating the weights according to the degree of comparability between the evidence sets, and results in lower weights for the informative component when the data suggest that the two evidence sets are in 'disagreement'. The *WinBUGS* code for the implementation of these approaches is shown above. Essentially, an NMA is conducted only on the direct evidence and mixture priors are placed on all basic parameters. The informative component of these mixture priors is derived from a previous step (NMA of indirect evidence with vague priors) and is inserted here as data. Despite that both mixture prior and commensurate priors have the potential to facilitate 'adaptive' borrowing, it is worth noting that the former requires two separate steps whilst the latter analyses all evidence simultaneously.

4.4.3.4 Power-prior

Another modelling approach, initially introduced by Ibrahim and Chen, 2000 to facilitate flexible information-sharing, is the power-prior. This model down-weights indirect evidence by raising its likelihood to a power α. To put into context the formula introduced in Equation 2.12 on page 28, in a standard Bayesian analysis that uses only direct evidence and vague priors, the posterior distribution of the basic parameters arises as:

$$\pi(d_{1k}^{Dir}|D^{Dir}) \propto \prod_{i=1}^{N_{Dir}} L(d_{1k}^{Dir}|D^{Dir}) \cdot \pi_0(d_{1k}^{Dir})$$

where D^{Dir} denotes the direct data so that $\pi(d_{1k}^{Dir}|D^{Dir})$ are the posterior distributions of the basic parameters which are proportional to the likelihood of the direct evidence, $L(d_{1k}^{Dir}|D^{Dir})$, and to a vague prior $\pi_0(d_{1k}^{Dir})$.

Under a Power-prior approach, $\pi_0(d_{1k}^{Dir})$ becomes informative by incorporating the indirect information so that:

$$\pi_0(d_{1k}^{Dir}) \propto \prod_{i=1+N_{Dir}}^{N_{Dir}+N_{Indir}} L(d_{1k}^{Indir}|D^{Indir})^{\alpha} \cdot \pi_0(d_{1k}^{Indir}) \tag{4.16}$$

where D^{Indir} denotes the indirect data provided by N_{Indir} indirect studies, $L(d_{1k}^{Indir}|D^{Indir})^{\alpha}$ is the likelihood of the indirect evidence raised to the power of α, and $\pi_0(d_{1k}^{Indir})$ is a vague prior for the basic parameters of the indirect evidence. Therefore the posterior distribution of the basic parameters of the direct evidence becomes:

$$\pi(d_{1k}^{Dir}|D^{Dir}, D^{Indir}, \alpha) \propto \prod_{i=1}^{N_{dir}} L(d_{1k}^{Dir}|D^{Dir}) \cdot \underbrace{\prod_{i=1+N_{Dir}}^{N_{Dir}+N_{Indir}} L(d_{1k}^{Indir}|D^{Indir})^{\alpha} \cdot \pi_0(d_{1k}^{Indir})}_{\text{Power-prior}}$$

with α regulating the influence of the indirect evidence. The variable α may be interpreted as the relative precision of the indirect evidence with lower values of α producing priors with heavier tails. The model can be extended to accommodate both study- and comparison-specific discounting powers (i.e $\alpha_{i,k}$). When $\alpha = 1$, the power-prior is equivalent to lumping, while when $\alpha = 0$ the indirect evidence is effectively disregarded and the approach is equivalent to splitting. As with the weights of mixture priors, the value of α could be arbitrarily defined, in which case extensive sensitivity analyses should be conducted to explore the impact of α (Spiegelhalter et al., 2004). Alternatively, α could be based on subjective expert opinion on the relevance of the indirect evidence in determining the relative effects of the direct evidence. The model can also accommodate the inclusion of multiple indirect evidence sets by defining evidence set-specific likelihoods and powers α_j, so that the power-prior becomes:

$$\pi_0(d_{1k}^{Dir}) \propto \prod_{i=1+N_{dir}}^{N_{Indir_1}} L(d_{1k}^{Indir_1}|D^{Indir_1})^{\alpha_1} \cdot ... \cdot \prod_{i=N_{Indir_{j-1}}+1}^{N_{Indir_j}} L(d_{1k}^{Indir_j}|D^{Indir_j})^{\alpha_j} \cdot \pi_{vague} \tag{4.17}$$

where π_{vague} is a vague prior, D^{indir_j} is the data provided by studies $N_{indir_{j-1}}, ..., N_{indir_j}$ and $d_{1k}^{Indir_j}$ the evidence set-specific basic parameter estimates.

To code the power-prior model, relevant guidance for the specification of an arbitrary sampling distribution can be used (Lunn et al., 2013). Specifically, we can use the 'zeros' trick to produce a custom 'down-weighted' likelihood. An example of a custom binomial down-weighted likelihood is shown below. Initially, we create a set of $z_{i,k}$ that are assumed to be drawn from a $Poisson(\phi_{i,k})$ distribution. Each observation then has a likelihood contribution $e^{-\phi_{i,k}}$, and therefore by defining $\phi_{i,k} = \alpha \cdot (-loglikelihood_{i,k})$ we obtain the correct discounted likelihood contribution (because $log(L^{\alpha}) = \alpha \cdot log(L)$). $\phi_{i,k}$ is the mean of the Poisson distribution and therefore needs to be positive; this is ensured by adding an arbitrary constant. The $-loglikelihood_{i,k}$ part is subsequently defined based on the binomial likelihood function ($\frac{n!}{r!(n-r)!} \cdot p^r \cdot (1-p)^{n-r}$). Importantly, if there are multiple indirect evidence sets, each one should be analysed in a separate loop and a custom likelihood should be specified for every source.

```
constant<-10000 # arbitrary to ensure phi[i,k] is positive
for(i in n.dir+1 : n.dir + n.indir) {
for (k in 1:na[i]) {
z[i,k] <- 0
z[i,k] ~ dpois(phi[i,k])
phi[i,k] <- alpha * neg.LL[i,k] + constant
neg.LL[i,k] <- - logfact(n[i,k]) + logfact(r[i,k]) + logfact(n[i,k] - r[i,k])
- r[i,k]*log(p[i,k]) - (n[i,k] - r[i,k])*log(1-p[i,k])
logit(p[i,k]) <- mu[i] + delta[i,k] # model for linear predictor
} }
```

Finally, although it is expected that as α increases the posterior estimates will move from those obtained under splitting to those obtained under lumping, the relationship between α values and the posterior estimates is not necessarily strictly increasing or decreasing. Instead, a non-monotonous relationship between α and posterior estimates may be observed when the indirect evidence comprises studies of very different sample sizes that suggest considerably different relative effects. Under such circumstances, as discounting happens on the studies' likelihood, when α is low, studies of high sample size will exert disproportionately more influence than studies of low sample size.

4.5. A 'methods identification' framework

The previous sections highlighted that several methodological options exist for the combination of direct and indirect evidence. That said, the methods presented are not applicable across all instances. This is because their applicability depends on both the nature of the available indirect evidence and the plausibility of the assumptions that they impose in the context of the synthesis problem at hand. In this section, a simple set of steps is suggested that can assist analysts in identifying applicable ISMs for their own synthesis problems (Figure 4.3). Note that this process can also be used when indirectness stems from a non-population level of PICOS such as the 'Intervention' or the 'Study-design' level. However, this process is not applicable for the synthesis of indirectly related outcomes. More details on this matter are provided in the discussion.

The process begins with the *'Identification'* of direct and indirect evidence. In this step, directly and indirectly relevant evidence are identified using classical systematic review methods which have been described in Higgins and Green, 2011. To date, specific routines for the identification of indirect evidence have only been described for the case of evidence pertaining to indirectly related treatments (Hawkins et al., 2009).

In the *'Parametrisation'* step, the analyst has to decide how the sources will be grouped. In the simplest case where only one indirect source exists, a study-level binary variable can be used to indicate to which evidence set each study pertains. However, if more indirect sources exist, a categorical variable may be more appropriate. The analyst then would have to decide whether or not there is a way of ordering the various sources according to their expected size of the RTE (i.e. the variable is ordinal) and whether the magnitude of the difference is meaningful (i.e. the variable is cardinal). For instance, if the difference among the various sources is that they pertain to patients who suffer from different disease severities, it may be reasonable to assume —based on empirical evidence or expert opinion —that as the disease severity increases we would expect to observe a higher relative effectiveness; yet it may not be the case that the RTE increases linearly.

Figure 4.3: *A step-by-step process to identify applicable ISMs.*

‘Identification’: Identification of direct and indirect evidence sets and studies.

⇓

‘Parametrisation’: Based on the nature and quantity of indirect evidence, decide how the various evidence sets can be grouped (e.g. using a categorical, ordinal or cardinal variable).

⇓

‘Base-model selection’: Explore heterogeneity within each evidence set and across evidence sets to understand which effect modifiers influence which evidence sets ultimately selecting the model that best describes the evidence (hereafter termed base-model).

⇓

‘Eligibility’: Determine which ISMs can be used given the nature of the variable used to describe the differences between the direct and indirect datasets. Table 4.2 and Table 4.3 can be useful.

⇓

‘Plausibility’: Examine each method’s assumptions in the context of the particular synthesis problem and eliminate methods which impose unrealistic assumptions. Table 4.2 and Table 4.3 can be useful.

⇓

‘Implementation’: Implement the remaining methods.

The next step is *‘Base-model selection’*. At this point, heterogeneity should be explored in the extended evidence base in order to decide which synthesis model best describes each evidence set. This is similar to the classical model selection process that is commonly followed to identify effect modifiers and determine the model that best describes the data and accounts/explains between-study heterogeneity when there is no indirect evidence. The difference here is that this process needs to be conducted for both evidence sets, leading to the development of a *Base-model* which best describes both evidence sets. This can then be the starting point for the application of ISMs. This step allows us to investigate whether the same effect modifiers apply to the various evidence sets. Where this is the case, the extent of the effect modification can be compared across evidence sets with a view to assess the appropriateness of sharing information across treatment effect modifiers. Meta-regression models with a common effect modification coefficient in all evidence sets can be used for this purpose. Where this is not the case, different base-models can be used for each evidence set. Importantly, at this step no information-sharing is imposed across different evidence sets unless a base-model with a common effect modification coefficient is chosen, in which case there is already some information-sharing when this model parameter is estimated.

In the *'Eligibility'* step, a list of applicable methods is obtained. Table 4.2 and Table 4.3 show methods that may potentially be applied for each type of variable that may be chosen to distinguish between direct and indirect evidence in the *'Parametrisation'* step. These tables also raise additional points for consideration. For example, the type of the selected base-models should be considered (i.e. FE or RE), because under RE, several lumping approaches may be applicable as well as methods that share information only on the heterogeneity component.

Subsequently, in the *'Plausibility'* step, the assumptions underlying each method should be examined for the specific synthesis problem at hand and a judgement regarding their plausibility should be made. The reader may again find Table 4.2, Table 4.3 helpful in this process. It follows that methods which impose implausible assumptions should be eliminated. Furthermore, the number of indirect evidence sets is important in determining whether model parameters such as the across-sources variance of the multi-level models can be appropriately estimated and therefore, in this step, it should also be judged whether there are enough data to implement potentially applicable methods.

Finally, all the remaining methods should be applied in the *'Implementation'* and their results should be compared.

Table 4.2: *A summary of the methods described in this chapter that share information only on the RTE mean d.*

Method	Relationship	Assumption	Can be used for multiple indirect evidence sets?
Lumping	$d_{1k}^{Dir} = d_{1k}^{Indir}$	Relative effectiveness is identical in the direct and the indirect evidence.	Yes
Constraint	$d_{1k}^{Dir} \leqslant d_{1k}^{Indir}$	Within each MCMC iteration there is a strict direction of the evidence set-specific RTEs.	Yes, as long as an ordinal or cardinal distinguishing variable can be used.
Meta-regression	$d_{1k}^{Dir} = d_{1k}^{Indir} + \beta$	There is an additional, fixed, RTE component in the indirect evidence.	Yes. In fact it imposes information-sharing only when a cardinal distinguishing variable can be used. ∗
Multi-level	$(d_{1k}^{Dir}, d_{1k}^{Indir}) \sim N(D_{1k}, \phi)$	The source-specific RTEs are exchangeable (i.e. they do not systematically differ).	Yes ‡
Random-walk	$d_{1k}^{Dir} \sim N(d_{1k}^{Indir}, \eta)$	The RTE of evidence set is more similar to that of its adjacent evidence set than to more ‘distant’ sets.	Yes, as long as an ordinal or cardinal distinguishing variable can be used. ‡

∗ If the direct evidence pertain to the lowest value of the cardinal variable i.e. $X = 0$, then information-sharing will only take place among the indirect evidence sets. This is because the regression coefficient will only hold as a non-zero quantity for $X \geqslant 1$, which would correspond to the indirect sources.

‡ The variance of the normal distribution may not be appropriately estimated if the number of indirect evidence sets is low. This might result in unrealistic variance values (low or high). When that is the case an informative prior elicited or estimated from meta-epidemiological data could be used.

Table 4.2: *A summary of the methods described in this chapter that share information only on the RTE mean d (continued).*

Method	Relationship	Assumption	Can be used for multiple indirect evidence sets?
Commensurate prior	$d_{1k}^{Dir} \sim N(d_{1k}^{Indir}, \eta_{1k})$ $\frac{1}{\eta_{1k}} \sim \begin{cases} N(20,1) & ,\text{if } c_{1k}=0 \\ Gamma(0.1,0.1) & ,\text{if } c_{1k}=1 \end{cases}$ $c_{1k} \sim Bernoulli(p_{1k})$	The direct RTE is similar to the indirect RTE. The strength of this assumption is controlled by η_{1k}, the components of which are mixed according to a probability that is either fixed or estimated in the model.	Yes, as long as an ordinal or cardinal distinguishing variable can be used. ∗
Informative prior	$d_{1k}^{Dir} \sim N(d_{1k}^{Indir}, V_{1k}^{Indir})$	The RTE of the direct evidence is more likely to take values from the posterior estimates of the indirect RTE than other values.	Yes, as long as all indirect evidence sets can be initially analysed and summarised in a single prior distribution. ∗
Mixture prior	$d_{1k}^{Dir} \propto$ $\nu \cdot N(d_{1k}^{Indir}, V_{1k}^{Indir}) + (1-\nu) \cdot N(0,10^4)$	The RTE of the direct evidence is more likely to take values from a ‘hybrid’ prior, which comprises of a mixture of the posterior estimates of the indirect RTE and a vague component, than other values.	Yes. Either priors are estimated for each indirect evidence set and subsequently mixed with set-specific weights or all indirect evidence sets are initially analysed and summarised in a single prior.
Power-prior	$\pi_0(d_{1k}^{Dir}) \propto$ $\prod_{i=1+N_{dir}}^{N_{indir}} L(d_{1k}^{Indir} \| D^{Indir})^{\alpha} \cdot \pi_0(d_{1k}^{Indir})$	The likelihood of the direct evidence is lumped with the discounted likelihood of the indirect evidence.	Yes. A separate discounting power α_j would need to be used for each indirect evidence set.

∗ An additional decision needs to be made about the way that the indirect evidence sets are initially synthesised. For example, they might be lumped or they might be analysed using a multi-level model.

Table 4.3: *A summary of methods that share information on τ and can therefore be used under RE base-models.*

Method	Relationship	Assumption	Can be used for multiple indirect evidence sets?
Lumping heterogeneities ‡	$\tau^{Dir} = \tau^{indir}$	The between-studies variance in the direct and the indirect evidence is identical.	Yes
Informative prior on heterogeneity ‡	$\tau^{Dir} \sim Lognormal(\mu_{Indir}, \sigma^2_{Indir})$	The between-studies heterogeneity of the direct evidence is more likely to take values from the posterior distribution of the indirect between-studies heterogeneity than other values.	Yes. The indirect evidence sets would need to be initially analysed and their predictive distribution of the heterogeneity can be used as the prior distribution. ∗

∗ An additional decision needs to be made about the way that the indirect evidence sets are initially synthesised. For example, they might be lumped or they might be analysed using a multi-level model.
‡ Only applicable for RE models.

4.6. Discussion

This chapter utilised the findings of Chapter 3 to describe ISMs that can be implemented for a specific synthesis problem. That is, when we wish to borrow strength from evidence pertaining to a population that is different —yet related —to that considered by the policy research question. Thorough mathematical descriptions and explanations of the different methods were provided along with coding suggestions. Therefore, researchers facing similar synthesis issues can consult this chapter to both get a theoretical understanding of, and practical implementation suggestions for the various ISM options. This chapter serves as the methodological foundation for subsequent chapters.

Given the plethora of available methods and that it is unclear how methods compare in terms of the extent of information-sharing that they impose, method choice should be justified carefully. To that end, this chapter thoroughly explains the assumptions underpinning all the specified methods, raising important points for consideration and discussing ways to extend the methods for cases where several indirect sets of evidence exist. Furthermore, a step-by-step framework is described which attempts to systematise the process of choosing ISMs. This framework can be used by both researchers who seek to find methods applicable to their own synthesis issues, and by appraisers who need to ensure that no applicable method is unjustifiably excluded. To my knowledge, this is the first attempt to describe a process of identifying and choosing amongst alternative ISMs. The use of this process not only raises awareness around the existence of several ISMs, but also aids transparency by discouraging the use of sole methods that produce 'convenient' results.

The suggested methods identification process does not provide a way of choosing the most appropriate amongst the applicable methods. This is because the actual relationship between the true RTEs of the two populations is typically unknown in decision-making and the plausibility of each ISM cannot be solely determined by the data alone or the relationships imposed analytically. Hence, it is likely that a judgement will be required about the plausibility of each method's assumptions and the 'appropriate' degree of information-sharing.

It should be clarified at this point that information criteria, such as DIC and residual deviance ($\overline{D_{res}}$) are not appropriate means of comparing ISMs. First, some methods analyse the extended evidence base in two steps (first the indirect evidence and then the direct) and therefore their DICs are not comparable with methods that analyse all evidence in a single step. Second, perhaps more importantly, models that impose stronger information-sharing (e.g. lumping) probably yield larger $\overline{D_{res}}$ and DICs, because increasing information-sharing implies that estimates increasingly differ from the no sharing case and therefore the model fits to the data less well. Hence, increasing conflict between direct and indirect evidence implies higher $\overline{D_{res}}$ and DICs for methods that impose stronger information-sharing. However, if, for instance, we expect the direct evidence set to suffer from biases and the indirect evidence set to be of higher quality, conflict may justified and strong information-sharing desirable. Finally, given that model fitting quantities such as $\overline{D_{res}}$ and DICs are based on information theory and the concept of divergence, they would be primarily driven by the evidence set which contains the most information and would be indicative of the best fitting model to the richer evidence set instead of the Information-sharing method (ISM) that imposes the most appropriate degree of information-sharing. Overall, although it may be useful to see which models may not fit well to the available data, methods choice should not be based on statistical fit. Instead, choice of model is likely best made in the context of a deliberative process that takes into account the characteristics of the available evidence sets (e.g. patient characteristics, study-design) as well as clinical opinion, in order to determine the appropriate degree of information-sharing. However, given that such assessments may not always be straightforward, a judgement regarding the desired degree of information-sharing is perhaps a good starting point in refining the list of applicable ISMs. If there is no basis on which to reduce the number of models, implementation of all applicable models may still be useful as it can be viewed as a sensitivity analysis to ISMs.

This chapter only considers the case where the indirect evidence pertains to a different, yet relevant, population (e.g. patient group) on which the same interventions are used. This implies that the evidence sets can be distinguished using a study-level variable to which patients have not been randomised —and hence findings cannot be interpreted as causal. The methods are directly transferable with minor modifications to cases where indirectness stems from some other non-population study-level characteristic. For example, if the various evidence sets pertain to studies of different quality, or to randomised and observational evidence, the same models are applicable without any modifications. However, if evidence is indirect to the intervention level and we want to relate the basic parameters pertaining to different interventions, direct and indirect evidence would then be distinguished by an arm-level variable to which patients have been randomised , and hence any conclusions would retain causal character. Finally, if indirectness stems from the outcome level -and hence the indirect evidence provide information on an outcome that is not directly considered by the policy question but is relevant to one of the considered outcomes- then other methods may be applicable. For instance, the direct and indirect outcomes may be functionally related, or multivariate relationships could be used model the correlation structure of direct and indirect outcomes.

The ISMs explained in Section 4.4 consider only simple FE and RE models without any effect modifiers. Extending the ISMs to the meta-regression case may not be straightforward if the extent of the effect modification is different in the various evidence sets. This is because imposing relationships among the direct and indirect basic parameters would effectively only share information on the part of the relative effect that the effect modifier is *not* responsible for. Therefore, to properly extend these methods to meta-regression, we would need to relate the adjusted relative effects in the direct and the indirect evidence. This issue is revisited in Chapter 5, where the circumstances under which the existing methods can be used for meta-regression models are explained.

Lastly, multi-variate relationships have not been described in this chapter. This is because it is assumed that the two evidence sets include studies that do not enrol from and report for both populations. Therefore, the information that is required to calculate within- and between- studies correlations is not available. Specifically, for the between-study correlations we would require several studies that enrol from and separately report for both population groups. Even more restrictively, for the within-study correlations we would require that the nature of the two populations is such that patients move from belonging in one population to belonging in the other, and also that there is IPD evidence reporting the outcome for each patient being in each population. Albeit rare, there are examples where such evidence could be obtained. For instance, progression in a patient cohort may be observed for a considerable length of time, as their condition becomes more severe or they switch to a later line of treatment. Then, within-patient correlations across outcomes pertaining to a different disease severity or a different line of treatment could be estimated. An example where registry data is used to inform the correlation between treatment effects in two different lines of therapy in a multi-variate model can be found in Abrams et al., 2017.

Future work can try to address the limitations of this chapter by explicitly describing how the models and the step-by-step 'methods identification process' can be adapted to address cases where indirectness stems from a non-population level. Furthermore, extending these methods to meta-regression, allowing for covariates to be considered as effect modifiers, can considerably increase their usefulness and make them applicable to a wider set of circumstances. Finally, given that no general conclusions can be made from this chapter about how methods compare to each other, further work —such as that undertaken in Chapter 7 —can try to identify which factors determine the degree of information-sharing that the various models impose. This might assist model choice by mapping judgements regarding the desired degree of information-sharing to specific methods, and thus provide methods guidance to decision-makers.

Chapter 5

Borrowing strength from paediatric patients to inform relative effectiveness in adults: a case-study

5.1. Chapter aims and structure

The previous chapters provided the necessary foundation for the use of ISMs in HTA. Chapter 3 identified the statistical models that have been used to combine information directly and indirectly relating to a research question, and classified them according to the core-relationships imposed. Chapter 4 detailed methods applicable for the case where sharing happens across studies enrolling different populations and developed a step-by-step process for determining which methods are applicable for such synthesis problems.

This chapter aims to illustrate the impact of using different ISMs on RTE estimates using a case-study. The following questions are explored:

1. How can the indirect evidence help in explaining heterogeneity?
2. How can the applicable ISMs influence relative effectiveness estimates and how do these estimates compare in terms of the imposed strength-of-sharing?

It should be noted that this chapter does not aim to produce conclusions about how the results of the various ISMs compare in general and such a task would require simulation experiments.

The remainder of this chapter is structured as follows. Section 5.2 provides the necessary background to the case-study describing the decision problem and the synthesis approach that was originally adopted. Section 5.3 explains the motivation for borrowing strength from indirect evidence and Section 5.4 describes methods and findings of a review that sought to systematically update the searches for direct evidence and identify indirect evidence. Subsequently, in Section 5.5 the identified direct and indirect evidence is initially 'naively' combined using lumping and splitting approaches. Then, in Section 5.6, heterogeneity is explored in the extended evidence base in order to understand whether the same covariates influence the relative effect in the direct and the indirect evidence and identify the best-fitting base-models. In Section 5.7, given the best-fitting base-models, the applicable methods to share information on relative effectiveness between direct and indirect evidence are identified using the framework that was developed in Section 4.5. These are then implemented, and their results are compared and contrasted in terms of a set of strength of sharing measures. Finally, in Section 5.8 the findings of this chapter are discussed along with its strengths and limitations.

5.2. Background to the case-study

This section provides details on the decision problem which was initially explored in NIHR funded secondary research by Soares et al., 2012. A detailed description of the evidence synthesis undertaken as part of that work was also reported in Welton et al., 2015, and a detailed description of the cost-effectiveness evaluation and policy implications was separately published in Soares et al., 2014b. In what follows, that piece of work is briefly described along with the main challenges that the authors encountered in their analysis.

5.2.1. Decision problem

Sepsis is an inflammatory response caused by a serious infection of the bloodstream that can rapidly develop to a life-threatening condition (Hall et al., 2011). Recent UK estimates suggest that, each year, there are at least 260,000 new episodes of sepsis, claiming around 44,000 lives and costing around £15.6 billion to the NHS (Daniels and Nutbeam, 2017; Hex et al., 2017). Standard treatment includes antibiotics to target the infection, fluids to tackle the symptoms of septic shock, and occasionally albumin (ALB) serum to boost the immune system (National Institute for Health and Care Excellence, 2016).

Randomised trials have suggested that adjuvant Intravenous Immunoglobulin (IVIG) may be more effective than ALB in adults (Rodriguez et al., 2005; Werdan et al., 2007), but meta-analyses have failed to produce conclusive results (Alejandria and Marissa, 2013), possibly due to the fact that the evidence base was comprised of only a few high quality trials and was heterogeneous. Secondary research by Soares et al., 2012, which is also the case-study used in this chapter, reviewed the literature to obtain relevant studies, synthesised the evidence base to estimate relative effectiveness, and constructed a decision model to answer the following policy-relevant questions:

- Is IVIG a cost-effective adjuvant therapy for adults with severe sepsis or septic shock in terms overall mortality?
- Does the potential value of a new randomised clinical trial exceed the cost of conducting it? If so, what is the optimal sample size of such a trial?

5.2.2. Previous work on clinical effectiveness

In this section, the evidence identification and synthesis process followed in Soares et al., 2012 is summarised and the main findings and key challenges are highlighted.

The authors conducted a systematic review (up to 2nd October 2009) to look for evidence from randomised studies on the effectiveness of IVIG in adults (Table 5.1). All comparators to IVIG in RCTs were included. This evidence is here considered to be *direct*. They identified 17 head-to-head RCTs, conducted between 1981 and 2007, comparing various preparations of IVIG on top of Standard of Care (SoC) against SoC alone, or SoC combined with ALB , considered in terms of all-cause mortality. SoC comprised of antibiotics only. The main characteristics of these studies can be found in (Table B.1.1). Additional information about these studies was also extracted such as the duration of therapy, the dosage of IVIG, and the overall study quality proxied by Jadad score (0: lowest quality, 5: highest quality) (Jadad et al., 1996).

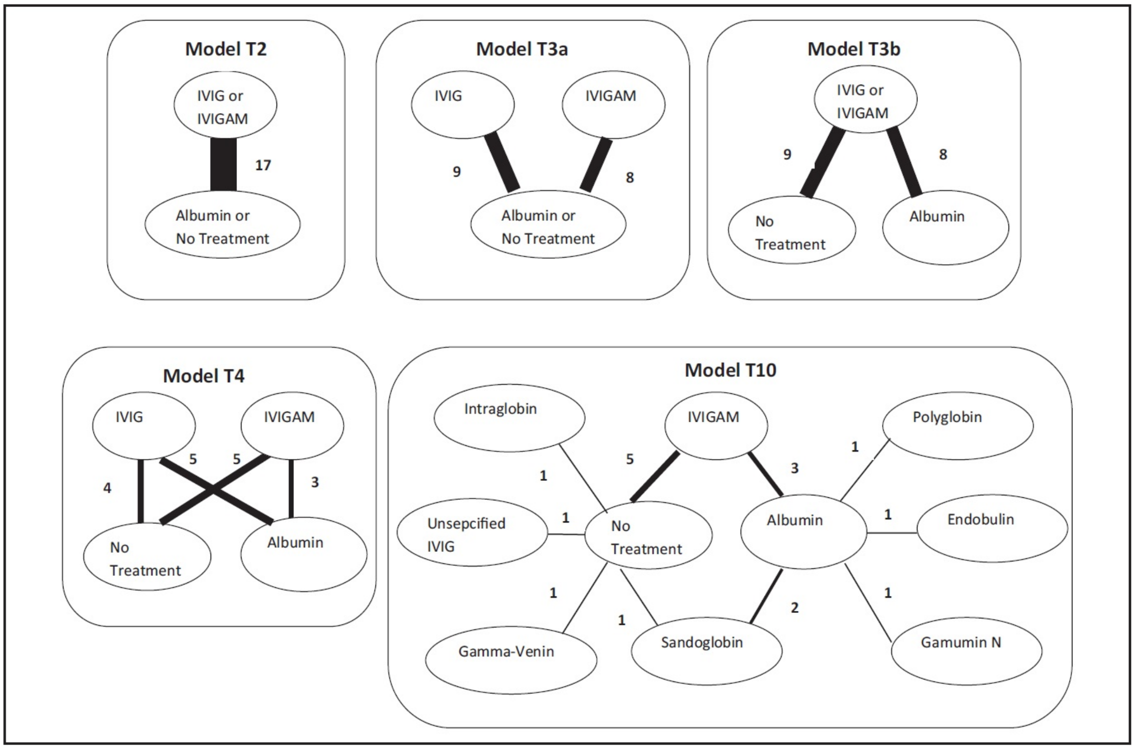
Table 5.1: *Research question investigated in Soares et al., 2012.*

Population	Adults with severe sepsis and septic shock
Intervention	IVIG or IVIGAM (in addition to SoC)
Comparator(s)	Placebo + SoC, Albumin + SoC
Outcome	All-cause mortality
Study-design	RCTs

When synthesising the 17 RCTs together to obtain a summary odds-ratio estimate, significant statistical heterogeneity was identified. First, the authors explored alternative treatment parametrisations to attempt to resolve heterogeneity due to the treatment definition. Five different treatment parametrisations were compared, which either combined or separated control and active treatments (Figure 5.1). The authors concluded that network T3B fitted the best because it achieved the best balance between resolving treatment heterogeneity and retaining adequate volume of evidence to inform treatment comparisons. The T3b network pooled immunoglobulin treatments (i.e. IVIG/IVIGAM + SoC), hereafter termed simply IVIG/IVIGAM, whilst retaining two separate treatment nodes for the comparators (i.e. ALB + SoC and SoC), hereafter termed ALB and PLA respectively. However, when covariates were added, network T2 -which pooled comparator as well as treatment arms- fitted similarly to T3b for most models, and was therefore preferred on parsimony grounds.

Additionally, a number of effect-modifiers relating to the treatment characteristics and risk-of-bias were explored using meta-regression models. The authors found duration of treatment to be important in explaining heterogeneity, however the clinical experts

Figure 5.1: *Different treatment parametrisations explored in Soares et al., 2012.*



could not determine a clear clinical rationale for it. Furthermore, covariates that relate to study-quality, including sample size[1] and year of publication, were found to significantly modify the treatment effect when included alone. However, neither the combination of the study quality covariates, nor the combination of any study quality covariate with duration of treatment explained a larger proportion of heterogeneity than that explained by using duration of treatment alone. The meta-regression model that used the 'publication year' covariate returned similar estimates with that which used the Jadad score, and since the latter provides more useful predictions (by referring to a study of very good quality rather than to a study published in a particular year), it was preferred.

[1] Taken as $1/\sqrt{N}$, where N is the number of patients in the treatment arm.

Despite the thorough heterogeneity exploration process, there was significant remaining heterogeneity under most models, as well as uncertainty regarding the optimal treatment parametrisation. As a result, the authors did not select just one model, but considered the implications of a set of best-performing base-models (Table 5.2). Except for the prediction for a study of infinite sample size, all model predictions suggested that the treatment with IVIG reduces mortality when compared to ALB or PLA alone. Within models, relative effectiveness predictions were very uncertain, with the 95% credible intervals of the odds ratios ranging considerably both below and above 1. They also differed in terms of their point estimates, which ranged from 0.68 in M2 —suggesting that IVIG is more effective than ALB —to 1.27 in M4b —suggesting that IVIG is less effective than ALB —(Table 5.2).

Table 5.2: *Final set of synthesis models considered in Soares et al., 2012.*

Model	Odds Ratio (95% CrI)
M1 : T3b FE Meta-regression on duration of treatment *(Prediction for 3 days of treatment)*	0.75 (0.58, 0.96)
M2 : T3b RE *(Predictive distribution)*	0.68 (0.16, 1.83)
M3 : T2 RE Meta-regression on Jadad *(Predictive distribution. Prediction for Jadad = 5)*	0.83 (0.18, 2.13)
M4a : T2 RE Meta-regression on $1/\sqrt{N}$ *(Predictive distibution. Prediction for sample = 339)*	0.92 (0.23, 2.10)
M4b : T2 RE Meta-regression on $1/\sqrt{N}$ *(Predictive distibution. Prediction for sample = ∞)*	1.27 (0.25, 3.17)

The Odds Ratio estimates reported correspond to the relative effect of ALB *vs* IVIG/IVIGAM for the T3b models, and to the ALB/PLA *vs* IVIG/IVIGAM for the T2 models. Credible Interval (CrI) is the Bayesian analogue of the Confidence Interval (CI).

5.3. Motivation for this work

As mentioned in the previous section, despite the meticulous analyses undertaken by Soares et al., 2012 using the adult evidence —here considered direct evidence —, significant uncertainties remain. Particularly, the clinical effectiveness of IVIG/IVIGAM as an adjunctive therapy to SoC remains unclear with relative effectiveness estimates surrounded by considerable uncertainty. Also, the authors could not conclude with certainty the best treatment parametrisation nor fully explain heterogeneity. As a result, several synthesis models —which used different covariates and suggested considerably different RTEs —were used and brought forward to the economic analysis.

Intravenous immunoglobulin has also been suggested as an adjunctive treatment for paediatric patients with severe sepsis and septic shock (Samatha et al., 1997; Akdag et al., 2014; Kola et al., 2014) with several studies, including a recent large multi-centre study that enrolled more than 3,000 patients (Brocklehurst et al., 2011), assessing its relative effectiveness against the current standard treatment (i.e. ALB).

To date, adult and paediatric evidence has not been considered together. However, a recent study by Capasso et al., 2017 used the effectiveness evidence of IVIG in adults to support the potential value of IVIG in paediatric patients, implying that evidence may be partially transferable between adults and paediatric patients with severe sepsis. In this chapter, the adult and paediatric evidence are combined using a range of different ISMs. The adult population will still remain the population of interest, while the paediatric evidence will only support inference for adults.

Importantly, the appropriateness of using paediatric evidence to inform inferences on adults needs to be judged. To that end, the European Medicines Agency, 2016 suggests that an explicit *extrapolation concept* should be developed to identify whether there are adequate data to justify extrapolation, and an *extrapolation plan* should be undertaken to address existing data gaps and uncertainties relating to the similarity of the populations and the transferability of the evidence. Crucially, for the purposes of this case-study, even if evident differences between the two populations exist, it may still be appropriate to combine relative effectiveness evidence. Hence, a judgement on the commensurability of evidence sets for the specific parameter(s) that information is to be shared on is still required. Moreover, whilst it might still not be acceptable to consider the two evidence sets perfectly generalisable, some level of 'sharing' may still be appropriate. Here, as Chapter 4 illustrated, we have plenty of ISMs in our arsenal, each imposing assumptions of varying strengths, including the power-prior that allows the analyst to specify the desired strength of information-sharing. Nevertheless, it should be acknowledged that whilst the two evidence sets are here combined for the purpose of methodological research, the usefulness of findings for health policy and clinical practice hinges critically on a judgement on the appropriateness of borrowing strength from the paediatric population.

5.4. Systematic review update

5.4.1. Methods

This component of work aimed to i) update the review of RCTs on the adult population undertaken within the HTA (Soares et al., 2012), and ii) expand the evidence base by including studies that enrolled paediatric patients. Therefore, the resulting studies will reflect the current *evidence totality* on both populations. The process comprised of the following steps:

1. The identified studies on adults from Soares et al., 2012 (up to December 2009) along with the identified studies on both adults and paediatric patients from Alejandria and Marissa, 2013 (up to December 2012), which was a systematic review that identified studies in both adults and children, were directly included.
2. The search strategy employed by Soares et al., 2012 to search for adult studies up to December 2009 was used to:
   (a) Identify citations before December 2009 that pertained only to paediatric patients. This was possible because the search strategy did not apply any population criteria, and only excluded studies on non-adult patients during the screening process
   (b) Update the search for both adult and paediatric patients by restricting searches to between 1st January 2010 and 1st August 2018.

The search strategies for MEDLINE and EMBASE are included in the Appendix (Table B.2.1 and Table B.2.2).

Inclusion Criteria

The inclusion criteria were the same as those used in Soares et al., 2012, with an important difference being that studies enrolling participants of *any age* were included. In particular:

- Population(s) : Patients of any age with severe sepsis or septic shock
- Intervention(s) : Any preparation of polyclonal IVIG or IVIGAM (i.e. IgM-enriched IVIG)
- Comparator(s) : No treatment (Placebo), Standard of Care (SoC) i.e. antibiotics, or Albumin (ALB) serum
- Outcome(s) : All-cause mortality
- Setting : Critical-care unit
- Study-design : Randomised controlled trials

All studies which investigated the use of IVIG/IVIGAM for *prevention* of sepsis were excluded, along with those studies which had enrolled patients with suspected but unconfirmed sepsis.

Data Extraction

Data were extracted using the template that was developed in Soares et al., 2012. The following information was extracted:

- Population: Whether the enrolled patient population was paediatric or adult. If paediatric, information on population age (young children, full- or pre-term neonates etc.) was also extracted.
- Intervention: The specific IVIG/IVIGAM product used in the treatment arm, including days of treatment duration, and total dosage (in mg/kg). For control interventions, data extracted predominantly concerned the type of treatment (e.g. no treatment, antibiotics, albumin-serum).
- Outcome: The number of patients enrolled in each arm, along with the number of events (deaths).
- Quality: Allocation concealment, blinding, randomisation, intention-to-treat analysis, missing data; from these, Jadad scores were subsequently calculated (Jadad et al., 1996).
- Other details : Year of publication, setting (e.g. Intensive Case Unit).

5.4.2. Results

The number of studies identified by each step of the systematic review update are illustrated in Figure 5.2. In brief, from previous reviews (Soares et al., 2012; Alejandria and Marissa, 2013) 17 studies in adults and 9 studies in paediatric patients were included. By updating the search strategy of Soares et al., 2012, no further RCTs enrolling adults were identified. However, when their search strategy was used without population restrictions two further studies enrolling paediatric patients were found. The list of all the identified studies with the 'filter' by which they were identified is shown in the Appendix in Table B.2.3, and the full data set is reproduced in Table 5.3.

Figure 5.2: *Flow chart. Results of the systematic review.*

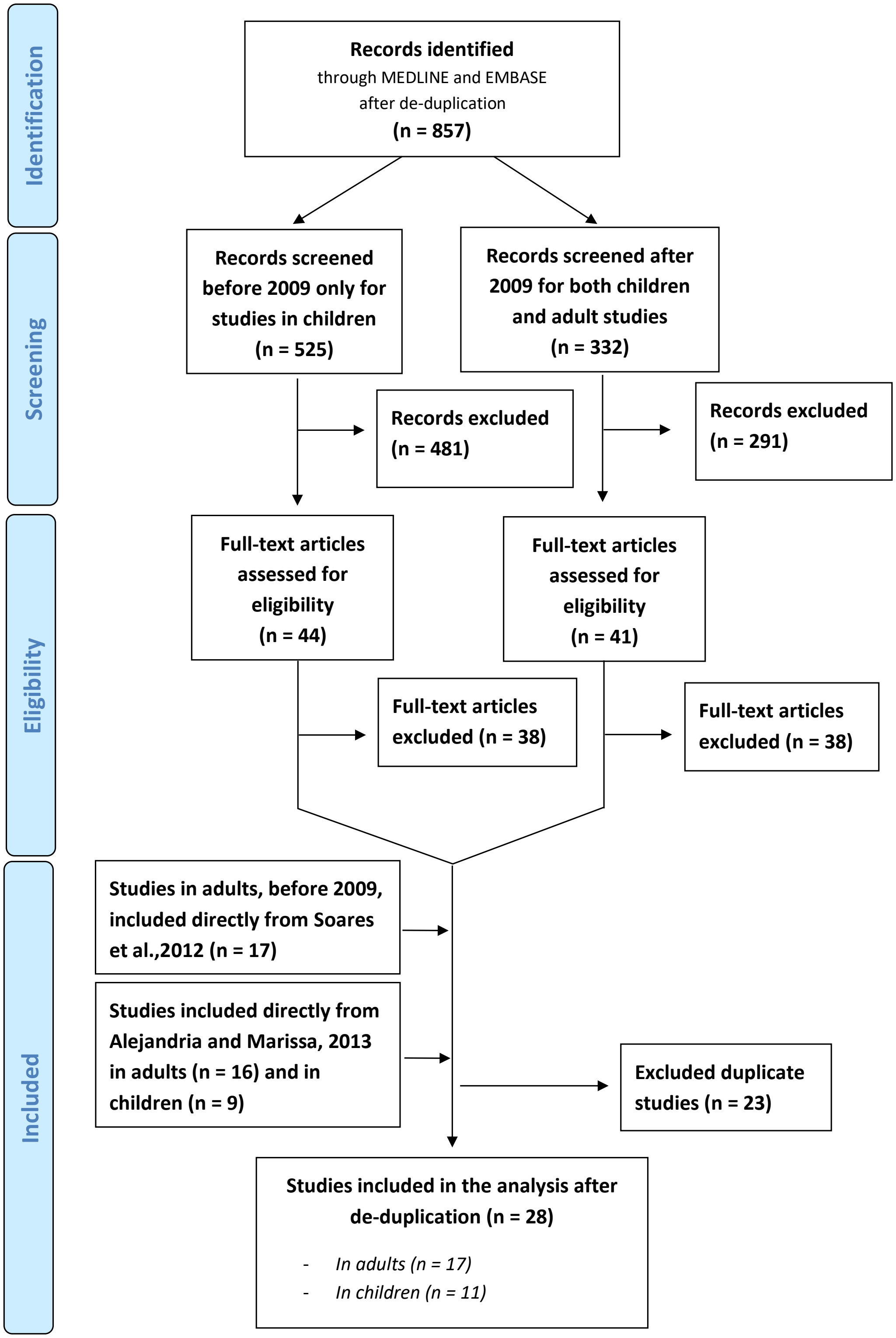

Identification
Screening
Eligibility
Included
Records identified
through MEDLINE and EMBASE
after de-duplication
(n = 857)
Records screened before 2009 only for studies in children (n = 525)
Records screened after 2009 for both children and adult studies (n = 332)
Records excluded (n = 481)
Records excluded (n = 291)
Full-text articles assessed for eligibility (n = 44)
Full-text articles assessed for eligibility (n = 41)
Full-text articles excluded (n = 38)
Full-text articles excluded (n = 38)
Studies in adults, before 2009, included directly from Soares et al.,2012 (n = 17)
Studies included directly from Alejandria and Marissa, 2013 in adults (n = 16) and in children (n = 9)
Excluded duplicate studies (n = 23)
Studies included in the analysis after de-duplication (n = 28)
- In adults (n = 17)
- In children (n = 11)

Table 5.3: *Data used for all the models.*

St.ID	na	t1	r1	n1	t2	r2	n2	pub.year	Jadad	dosage	duration	population	$1/\sqrt{N}$	Study
1	2	ALB	13	27	IVIGAM	8	29	2005	5	1.75	5	adults	0.19	Rodriguez 2005
2	2	ALB	29	103	IVIGAM	27	103	2006	3	0.93	3	adults	0.1	Hentrich 2006
3	2	PLA	14	34	IVIGAM	8	34	2002	2	0.75	3	adults	0.17	Karatzsas 2002
4	2	PLA	7	21	IVIGAM	5	21	2002	3	0.75	3	adults	0.22	Tugrul 2002
5	2	ALB	10	22	IVIGAM	9	30	1995	1	0.93	3	adults	0.18	Behre 1995
6	2	PLA	9	28	IVIGAM	1	27	1991	3	0.855	3	adults	0.19	Shedel 1991
7	2	PLA	13	17	IVIGAM	8	18	1990	1	0.75	3	adults	0.24	Wesoly 1990
8	2	PLA	11	25	IVIGAM	6	25	1987	1	0.45	3	adults	0.2	Spannbruker 1987 + Vogel 1987
9	2	ALB	36	56	IVIG	19	57	1996	3	1	5	adults	0.13	Dominioni 1996
10	2	ALB	3	19	IVIG	4	19	1991	5	1.2	3	adults	0.23	Burns 1991
11	2	PLA	9	12	IVIG	7	12	1988	1	1	5	adults	0.29	De Simone 1988
12	2	ALB	113	303	IVIG	126	321	2007	5	0.9	2	adults	0.06	Werdan 2007
13	2	PLA	19	22	IVIG	15	24	1988	2	0.5	2	adults	0.2	Grundmann 1988
14	2	ALB	4	11	IVIG	1	10	2003	5	2	3	adults	0.32	Darenberg 2003
15	2	PLA	1	74	IVIG	1	74	1981	3	0.45	3	adults	0.12	Lindquist 1981
16	2	PLA	10	343	IVIG	3	339	2000	3	0.21	3	adults	0.05	Masaoka 2000
17	2	ALB	9	19	IVIG	3	21	1998	3	1.8	7	adults	0.22	Yakut 1998
18	2	PLA	1	28	IVIG	2	28	1996	4	0.5	1	neonates	0.19	Chen 1996
19	2	PLA	9	24	IVIGAM	6	20	1993	1	0.6	3	neonates	0.22	Erdem 1993
20	2	PLA	6	30	IVIGAM	1	30	1988	4	0.5	1	neonates	0.18	Haque 1988
21	2	PLA	2	18	IVIG	2	19	1992	3	0.5	1	neonates	0.23	Mancilla 1992
22	2	PLA	8	30	IVIGAM	5	30	1997	1	0.6	3	neonates	0.18	Samatha 1997
23	2	PLA	7	25	IVIG	7	25	1999	1	0.15	3	neonates	0.2	Shenoi 1999
24	2	ALB	5	17	IVIG	2	14	1992	5	0.5	1	neonates	0.27	Weisman 1992
25	2	ALB	677	1734	IVIG	686	1759	2011	5	1	3	neonates	0.02	Brocklehurst 2011
26	2	PLA	2	51	IVIGAM	4	51	2014	5	0.75	3	neonates	0.14	Akdag 2014
27	2	PLA	14	39	IVIGAM	5	39	2014	5	0.6	3	children	0.16	Kola 2014
28	2	PLA	10	30	IVIG	8	30	2005	3	2	2	children	0.18	Yildizdas 2005

$r1$ and $r2$ correspond to the number of deaths in the control and treatment arms which are of size $n1$ and $n2$ respectively. Dosage is measured in mg/kg and duration of treatment in days. na is the number of treatment arms in each study and N to the number of patients in the active treatment arm.

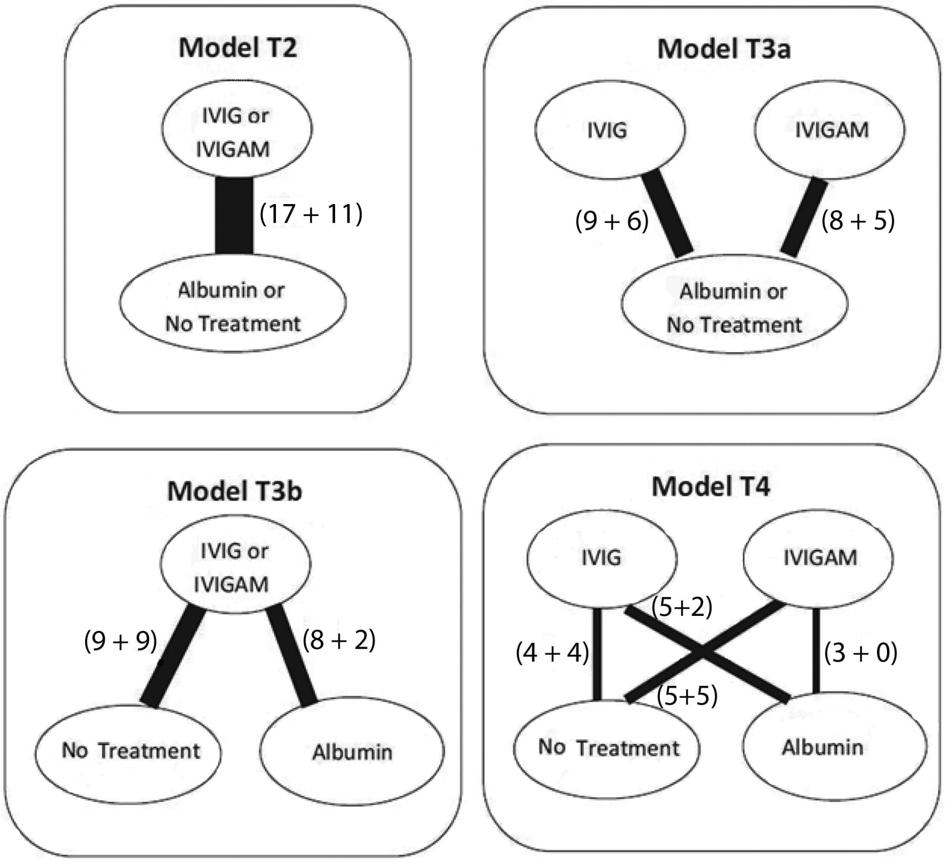
Overall, 28 studies were included. Among these 17 studies enrolled adults (2300 patients), and 11 studies enrolled children[2] (4071 patients). Despite the larger number of adult studies, more paediatric patients are included in the analysis due to a single large trial of almost 3,500 paediatric patients (Brocklehurst et al., 2011). As illustrated in Table 5.4 the median sample sizes are similar between adult and paediatric studies (52 and 60 patients respectively). Across populations, a similar proportion of studies used IVIG instead of IVIGAM in the treatment arm ($9/17 = 53\%$ in adults and $6/11 = 55\%$ in paediatric patients). In contrast, for the control treatment, a much smaller proportion of paediatric studies used ALB ($\approx$ 18%) compared with the adult studies ($\approx$ 47%). With respect to total dosage of IVIG/IVIGAM, paediatric studies used, on average, slightly lower dosages per kg than adult studies. This is because paediatric studies used IVIG/IVIGAM for relatively smaller time periods. Paediatric studies seem to be of better quality overall with with 55% of the paediatric studies achieving a Jadad score of 4 or more, compared to only 24% of the adult studies. The total number of studies of the full evidence base informing each comparison in the various potential treatment parametrisations is illustrated in Figure 5.3.

Table 5.4: *Characteristics of the direct (adult) and the indirect (paediatric) evidence bases.*

Quantity	Adults (N=17)	Paediatric patients (N=11)
Median sample size (interquartile range $Q_{25\%}$ - $Q_{75\%}$)	52 (40-113)	60 (47 - 69)
Average total dosage across studies, mg/kg (standard deviation)	0.95 (0.49)	0.7 (0.48)
Average treatment duration, in days (standard deviation)	3.4 (1.2)	2.2 (1.0)
Number of studies using IVIG (%) -in the treatment arm-	9 (53%)	6 (55%)
Number of studies using Albumin (%) -in the control arm-	8 (47%)	2 (18%)
Number of studies of good quality (%) i.e. Jadad score $\geqslant 4$	4 (24%)	6 (55%)

[2] Neonates and young children will be lumped henceforth.

Figure 5.3: *Updated graphs of networks.*



In the parentheses, the first number indicates the number of adult studies providing evidence for the comparison in question, while the second number indicates the number of paediatric studies.

5.5. Naive analyses

Commonly, HTAs either discard indirect evidence as if it is irrelevant or lump them with the direct evidence as if it does not differ from the direct evidence in any respect (Duarte et al., 2017; Faria et al., 2016; Rodgers et al., 2011). In this section, lumping and splitting approaches are presented as a starting point to motivate the use of more sophisticated ISMs covered later in the chapter. Despite that the comparison of interest is ALB against IVIG/IVIGAM, here the T2 treatment parametrisation is used. T2 is more parsimonious than T3b and provides the ALB/PLA against IVIG/IVIGAM comparison. Direct and indirect studies are either separately analysed or pooled with a RE meta-analysis (Figure 5.4). A FE meta-analysis is included in the Appendix (Figure B.3.1), although given the considerable between-studies heterogeneity in the adult population, this model may be considered inappropriate. The funnel-plot (Figure B.3.2) shows that the addition of the paediatric studies alleviates publication bias, as revealed by a gap in the bottom right part of the graph.

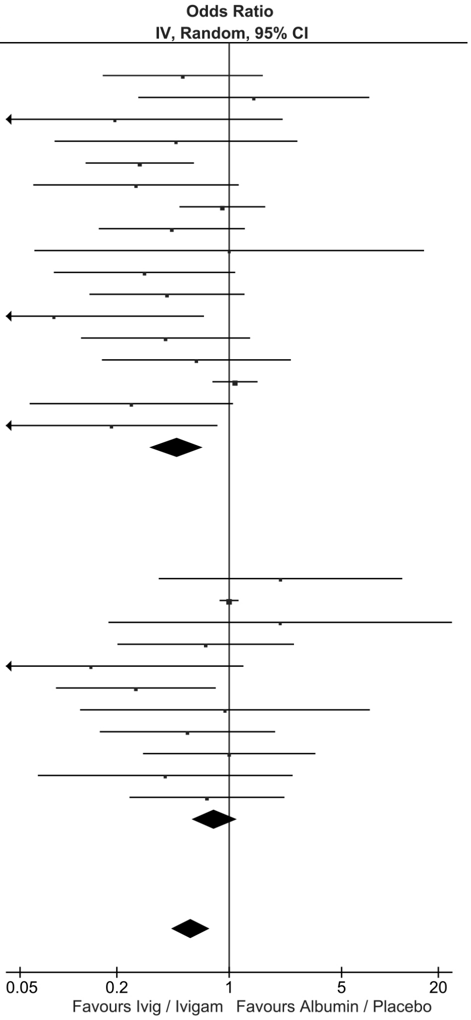
When the two sources are analysed separately, a significant relative effect is estimated for adults, favouring IVIG/IVIGAM. However, the same is not observed for paediatric patients where the relative effect crosses the 'line of no effect'. Also, the between-studies heterogeneities seem to differ substantially between adults and paediatric patients. In particular, across adult studies, the no heterogeneity hypothesis is rejected at a 95% confidence level (p-value = 0.02), yielding a $\tau^2_{AD} = 0.23$ and $I^2 = 68\%$. In contrast, paediatric studies are much less heterogeneous ($\tau^2_{PE} = 0.04$ and $I^2 = 13\%$) failing to reject the null hypothesis of no significant between-studies heterogeneity (p-value = 0.32).

When direct (adults) and indirect (paediatric patients) evidence is analysed under a common RE, a significant overall effect is estimated. This is considerably smaller and more precise than the effect of adults alone. The combined effect is more precise than either the adult or the paediatric effect, epitomising a potential benefit of using indirect evidence. However, pooling all studies together as if they are exchangeable might be considered inappropriate. Hence, lower levels of sharing may be more acceptable. Optimally, analysts should consider several methods which impose varying degrees of information-sharing, and check whether all methods consistently lead to similar conclusions.

Given that in the paediatric evidence there is one large, high-quality study, including 85% of the paediatric patients, it could be of interest to consider this study in isolation in order to strengthen the adults evidence base. As shown in Figure 5.5, the RTE is very similar with that of the analysis that combines all studies, albeit slightly less precise. Importantly, when only Brocklehurst et al., 2011 is added, the between-studies heterogeneity is higher that when all studies indirect studies are included. The RE model

Figure 5.4: *RE meta-analysis separately within each population and across both populations.*

	IVIG / IVIGAM		ALB / PLAC			Odds Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	IV, Random, 95% CI
1.1.1 Adults						
Behre 1995	9	30	10	22	3.8%	0.51 [0.16, 1.62]
Burns 1991	4	19	3	19	2.1%	1.42 [0.27, 7.44]
Darenberg 2003	1	10	4	11	1.1%	0.19 [0.02, 2.15]
De Simone 1988	7	12	9	12	2.0%	0.47 [0.08, 2.66]
Dominioni 1996	19	57	36	56	6.2%	0.28 [0.13, 0.60]
Grundmann 1988	15	24	19	22	2.6%	0.26 [0.06, 1.15]
Hentrich 2006	27	103	29	103	7.7%	0.91 [0.49, 1.68]
Karatzas 2002	8	34	14	34	4.3%	0.44 [0.15, 1.25]
Lindquist 1981	1	74	1	74	0.8%	1.00 [0.06, 16.29]
Masaoka 2000	3	339	10	343	3.2%	0.30 [0.08, 1.09]
Rodriguez 2005	8	29	13	27	4.0%	0.41 [0.14, 1.25]
Shedel 1991	1	27	9	28	1.4%	0.08 [0.01, 0.70]
Spannbruker 1987	6	25	11	25	3.5%	0.40 [0.12, 1.35]
Tugrul 2002	5	21	7	21	3.0%	0.63 [0.16, 2.42]
Werdan 2007	126	321	113	303	11.0%	1.09 [0.79, 1.50]
Wesoly 1990	8	18	13	17	2.6%	0.25 [0.06, 1.06]
Yakut 1998	3	21	9	19	2.5%	0.19 [0.04, 0.85]
Subtotal (95% CI)		**1164**		**1136**	**61.8%**	**0.47 [0.32, 0.69]**
Total events	251		310			
Heterogeneity: Tau² = 0.23; Chi² = 30.13, df = 16 (P = 0.02); I² = 68%						
Test for overall effect: Z = 3.88 (P = 0.0001)						
1.1.2 Paediatric patients						
Akdag 2014	4	51	2	51	2.0%	2.09 [0.36, 11.93]
Brocklehurst 2011	686	1759	677	1734	12.8%	1.00 [0.87, 1.14]
Chen 1996	2	28	1	28	1.1%	2.08 [0.18, 24.31]
Erdem 1993	6	20	9	24	3.3%	0.71 [0.20, 2.53]
Haque 1988	1	30	6	30	1.3%	0.14 [0.02, 1.23]
Kola 2014	5	39	14	39	3.8%	0.26 [0.08, 0.82]
Mancilla 1992	2	19	2	18	1.4%	0.94 [0.12, 7.50]
Samatha 1997	5	30	8	30	3.3%	0.55 [0.16, 1.93]
Shenoi 1999	7	25	7	25	3.4%	1.00 [0.29, 3.44]
Weisman 1992	2	14	5	17	1.8%	0.40 [0.06, 2.48]
Yildizdas 2005	8	30	10	30	4.0%	0.73 [0.24, 2.21]
Subtotal (95% CI)		**2045**		**2026**	**38.2%**	**0.81 [0.59, 1.12]**
Total events	728		741			
Heterogeneity: Tau² = 0.04; Chi² = 11.51, df = 10 (P = 0.32); I² = 13%						
Test for overall effect: Z = 1.26 (P = 0.21)						
Total (95% CI)		**3209**		**3162**	**100.0%**	**0.58 [0.44, 0.75]**
Total events	979		1051			
Heterogeneity: Tau² = 0.14; Chi² = 48.94, df = 27 (P = 0.006); I² = 45%						
Test for overall effect: Z = 4.05 (P < 0.0001)						
Test for subgroup differences: Chi² = 4.57, df = 1 (P = 0.03), I² = 78.1%						

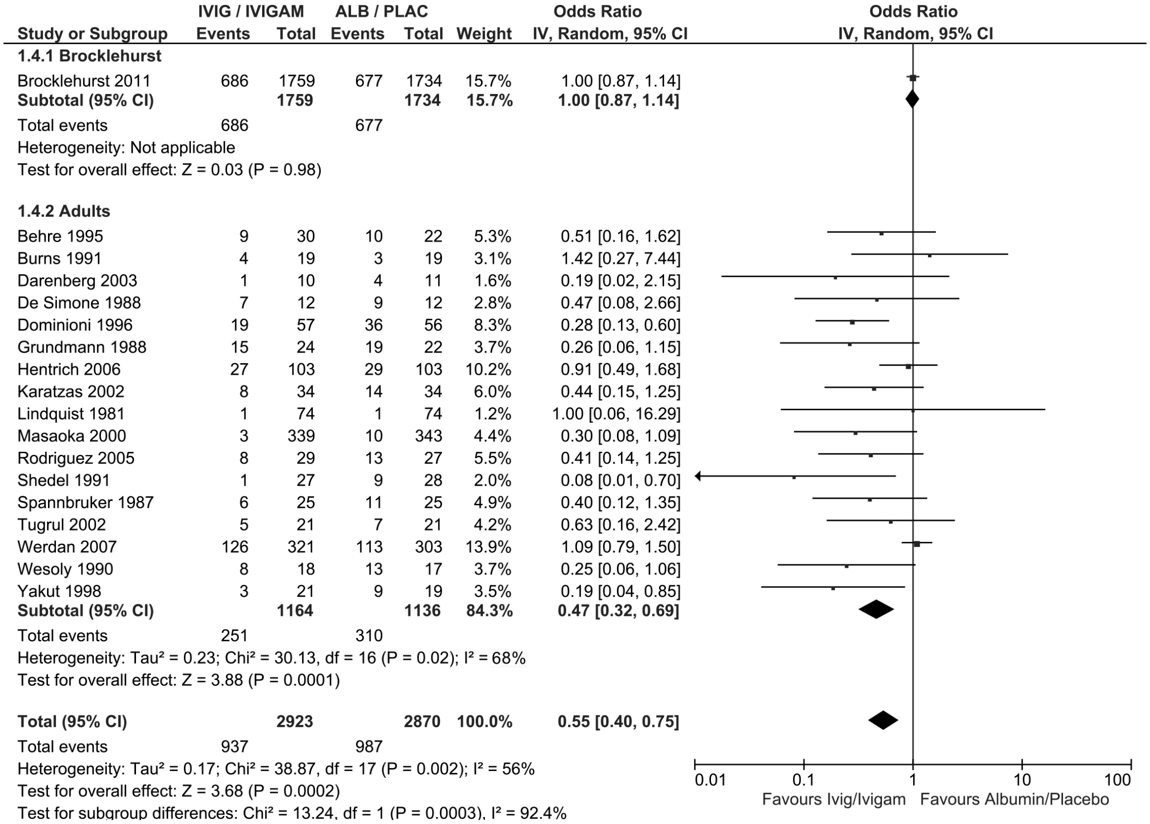


The plot was created using Review Manager 5.4 (Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014).

using only Brocklehurst et al., 2011 estimates that 56% of the between-studies variance cannot be explained by chance compared with the all-studies analysis that estimates 45%. Furthermore, this approach does not allow us to explore the heterogeneity in the two populations separately, nor estimate the extent to which the potential effect modifiers explain the between-trials heterogeneity in the two populations. Lastly, including only one indirect study restricts our list of ISMs because the between-indirect-trials heterogeneity cannot be estimated[3] and shared among the two populations.

[3] The between-paediatric-studies heterogeneity cannot be estimated when there is only one study on

Figure 5.5: *RE meta-analysis combining adult evidence with the large paediatric study only.*



Study or Subgroup	IVIG / IVIGAM Events	IVIG / IVIGAM Total	ALB / PLAC Events	ALB / PLAC Total	Weight	Odds Ratio IV, Random, 95% CI
1.4.1 Brocklehurst						
Brocklehurst 2011	686	1759	677	1734	15.7%	1.00 [0.87, 1.14]
Subtotal (95% CI)		**1759**		**1734**	**15.7%**	**1.00 [0.87, 1.14]**
Total events	686		677			
Heterogeneity: Not applicable						
Test for overall effect: Z = 0.03 (P = 0.98)						
1.4.2 Adults						
Behre 1995	9	30	10	22	5.3%	0.51 [0.16, 1.62]
Burns 1991	4	19	3	19	3.1%	1.42 [0.27, 7.44]
Darenberg 2003	1	10	4	11	1.6%	0.19 [0.02, 2.15]
De Simone 1988	7	12	9	12	2.8%	0.47 [0.08, 2.66]
Dominioni 1996	19	57	36	56	8.3%	0.28 [0.13, 0.60]
Grundmann 1988	15	24	19	22	3.7%	0.26 [0.06, 1.15]
Hentrich 2006	27	103	29	103	10.2%	0.91 [0.49, 1.68]
Karatzas 2002	8	34	14	34	6.0%	0.44 [0.15, 1.25]
Lindquist 1981	1	74	1	74	1.2%	1.00 [0.06, 16.29]
Masaoka 2000	3	339	10	343	4.4%	0.30 [0.08, 1.09]
Rodriguez 2005	8	29	13	27	5.5%	0.41 [0.14, 1.25]
Shedel 1991	1	27	9	28	2.0%	0.08 [0.01, 0.70]
Spannbruker 1987	6	25	11	25	4.9%	0.40 [0.12, 1.35]
Tugrul 2002	5	21	7	21	4.2%	0.63 [0.16, 2.42]
Werdan 2007	126	321	113	303	13.9%	1.09 [0.79, 1.50]
Wesoly 1990	8	18	13	17	3.7%	0.25 [0.06, 1.06]
Yakut 1998	3	21	9	19	3.5%	0.19 [0.04, 0.85]
Subtotal (95% CI)		**1164**		**1136**	**84.3%**	**0.47 [0.32, 0.69]**
Total events	251		310			
Heterogeneity: Tau² = 0.23; Chi² = 30.13, df = 16 (P = 0.02); I² = 68%						
Test for overall effect: Z = 3.88 (P = 0.0001)						
Total (95% CI)		**2923**		**2870**	**100.0%**	**0.55 [0.40, 0.75]**
Total events	937		987			
Heterogeneity: Tau² = 0.17; Chi² = 38.87, df = 17 (P = 0.002); I² = 56%						
Test for overall effect: Z = 3.68 (P = 0.0002)						
Test for subgroup differences: Chi² = 13.24, df = 1 (P = 0.0003), I² = 92.4%						

The plot was created using Review Manager 5.4 (Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014).

Optimally, before combining the evidence on the two populations, heterogeneity should be explored in the full evidence base, identifying the effect-modifiers separately in each population and then, for common effect-modifiers, also checking whether or not the use of the indirect evidence can lead to more precise estimation of the extent of the effect-modification. This process is explained in more detail in the next section.

paediatric patients. As a result, all the methods that share the heterogeneity parameter (i.e. common heterogeneity, prior on heterogeneity) cannot be used.

5.6. Re-exploration of heterogeneity

5.6.1. Methods

In Soares et al., 2012, the authors developed and implemented a step-by-step framework to identify important effect modifiers and select the best-fitting models in an evidence base that comprised only of direct evidence. Here, this framework is extended to also include the indirect evidence from paediatric patients. The proposed process not only explores potential effect modifiers and alternative treatment parametrisations separately within each population, but also identifies if and how indirect evidence might help in explaining the heterogeneity among the direct studies.

The extended framework developed here consists of the following steps:

1. FE and RE models without covariates: For every possible treatment parametrisation[4], fit FE and RE models separately in each population without imposing any information-sharing between direct and indirect evidence[5]. Record population-specific residual deviances and between-studies heterogeneities as well as overall DIC and residual deviance[6]. This step provides an initial understanding of the heterogeneity within each population, the extent to which alternative treatment parametrisations can partly explain heterogeneity, and whether or not the relative effects seem to be similar among the two populations.

2. Adding covariates: Subsequently, for each potential effect modifier fit the following four meta-regression models:

   (a) FE with separate, population-specific, effect modification
   (i.e. $logit(p_{i,k}) = \mu_i + d_{k,pop} + \beta_{pop} \times cov$)

   (b) FE with common effect modification across the two populations
   (i.e. $logit(p_{i,k}) = \mu_i + d_{k,pop} + \beta \times cov$)

   (c) RE with separate, population-specific, effect modification
   (i.e. $logit(p_{i,k}) = \mu_i + \delta_{i,k} + \beta_{pop} \times cov$)

   (d) RE with common effect modification across the two population
   (i.e. $logit(p_{i,k}) = \mu_i + \delta_{i,k} + \beta \times cov$)

[4] Note that the 10 treatments network is not considered here because there would be only 1 or 2 studies informing each treatment comparison.

[5] This is achieved by specifying separate parameters for each population. The only quantities that refer to the full evidence base are the DIC and the overall residual deviance.

[6] That is simply the sum of the population-specific residual deviances.

In the equations above, pop indexes the population, k the treatment, and i the study. Note that because no study provides information on both populations, pop is nested in i. Study-specific random-effects are assumed to follow treatment- and population-specific normal distributions ($\delta_{i,k} \sim N(d_{k,pop}, \tau_{pop})$). All remaining parameters in the aforementioned models are defined as in Chapter 2 (Section 2.1) and vague priors are applied to all hyperparameters. Finally, it is important to highlight that models (a) and (c) do not impose any information-sharing among the direct and indirect evidence, whilst models (b) and (d) impose some information-sharing because common effect modification is assumed across the two populations. The process is repeated for all treatment parametrisations.

This step allows us to compare the direction and estimated magnitude of the effect modification for each potentially important covariate in the two populations to assess whether it is statistically reasonable to impose a common effect modification coefficient. We can also obtain additional information regarding the optimal treatment parametrisation, and confirm whether the results are consistent with the previous step, and that the inclusion of covariates has not changed the best-fitting treatment parametrisation.

3. Combining covariates: For the identified important effect modifiers and the best performing treatment parametrisation, we can repeat the process of Step 2, using combinations of covariates. If the results of Step 2 suggest that different covariates are important in the two populations, models which use different effect modifiers for each population can be fit i.e. *'hybrid'* models.

For this analysis, four different treatment parametrisations were explored: T2, T3a, T3b and T4 (see Figure 5.1). The five covariates which had been found to explain some of the heterogeneity in Soares et al., 2012 were also considered. These were: duration of treatment, Jadad score, $1/\sqrt{N}$ (sample size), Dosage of IVIG/IVIGAM, and year of study publication.

Models were implemented in *WinBUGS* (MRC Biostatistics Unit, 2010), through *R* (R Development Core Team, 2010) using the R2WinBUGS package (Sturtz et al.); a Bayesian framework was adopted. Three MCMC chains with different starting values were used for all models, and Gelman-Rubin statistics were used to assess model convergence. For model comparison, deviance information criterion (DIC) and posterior mean residual deviance (D_{res}) were used. Models were considered to fit significantly worse when differences in DIC were larger than 3 points (Spiegelhalter et al., 2002). (For an example of a *WinBUGS* RE model which assumes separate effect modification coefficients for each evidence set, see `https://github.com/NikolaidisGFZ/PHD.git`)

5.6.2. Results

The results are presented in the same sequence as the steps were described in the methods section. This is because the conclusions of each step feed into the next, until we reach the last step and decide on the final list of base-models. The base-models will subsequently be the starting point for sharing information among adults and paediatric patients on relative effectiveness.

5.6.2.1 Step 1 : FE and RE models without any covariates

Table 5.5 illustrates the results of the application of FE and RE models, without any covariates (null models), in the four treatment parametrisations being assessed.

Table 5.5: *Results of Step 1 of the re-exploration of heterogeneity.*

Network	Model	τ_{AD}	τ_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
T2	FE	n/a	n/a	77.02	51.43	25.59	303.36
	RE	0.56*	0.47	51.97	30.82	21.14	289.31
T3a	FE	n/a	n/a	71.11	50.12	20.99	299.45
	RE	0.60*	0.35*	51.89	31.23	20.67	289.50
T3b	FE	n/a	n/a	65.61	42.76	22.84	293.91
	RE	0.49*	0.47	52.98	31.57	21.41	290.28
T4	FE	n/a	n/a	65.57	43.58	22.00	295.92
	RE	0.53*	0.46*	52.99	31.90	21.08	291.76

Blue colour indicates a low within-column value, red a high within-column value, and yellow similar within-column values. The asterisk ($*$) indicates a significant value at the 95% confidence level. τ refers to the between-studies-heterogeneity and D_{res} to the residual deviance. Subscripts (AD, PE) represent whether a measure only refers to adult or paediatric studies, while when there is no subscript the measure refers to the whole database (adults and paediatric patients).

In all networks, the DIC and residual deviance for RE models are lower than for FE models. The breakdown in residual deviance between the adult and paediatric patients shows that the decrease in residual deviance is mainly driven from the adult evidence (the difference between the paediatric residual deviance across FE and RE models within each network is very small and this is consistent across networks). Across adult studies, heterogeneity (τ_{AD}) is significant regardless of treatment parametrisation, implying that heterogeneity is not adequately explained by the network structure, and therefore covariates will need to be considered.

Across networks, all RE models fit similarly based on both Total Residual Deviance and DIC. However, based on the heterogeneity estimates (τ_{AD}, τ_{PA}), T3b network is the best for adults and T3a for the paediatric studies. However, it should be noted that in the paediatric network only 2 out of the 11 studies use ALB in the control arm, so when

ALB and PLA are separated (as in T3b), only a limited amount of evidence informs the IVIG/IVIGAM *vs* ALB comparison. With regards to T4, DIC and D_{res} are similar to T3b, but heterogeneity estimates are significant for both evidence sets and the model is less parsimonious. Overall, based on DIC values and given the fact that the primary focus here is on adults, T3b is chosen as the best treatment parametrisation.

5.6.2.2 Step 2 : Adding covariates

Table 5.6 shows the results of the meta-regression models, under T3b parametrisation, on a collection of variables which where shown by Welton et al., 2015 to influence the relative effect.

Table 5.6: *Step 2c. Results of meta-regression models on various covariates in network T3b.*

Covariate	Model	τ_{AD}	τ_{PE}	β_{AD}	β_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
NULL	FE	n/a	n/a	-	-	65.61	42.76	22.84	293.91
	RE	0.49^*	0.47	-	-	52.98	31.57	21.41	290.28
Duration	FE sep	n/a	n/a	-0.40^*	0.54	50.88	27.95	22.94	281.23
	FE com	n/a	n/a	-0.36^*	-0.36^*	53.25	27.99	25.26	282.58
	RE sep	0.19	0.50	-0.40^*	0.54	50.03	28.43	21.60	284.81
	RE com	0.20	0.57	-0.36^*	-0.36^*	50.50	28.32	22.17	285.26
Jadad	FE sep	n/a	n/a	0.26^*	1.97	61.07	38.23	22.85	290.39
	FE com	n/a	n/a	0.26^*	0.26^*	61.09	38.20	22.88	290.41
	RE sep	$0.44^{\ddagger}$	0.45	0.18	-3.18	53.86	32.43	21.43	291.07
	RE com	$0.43^{\ddagger}$	0.47	0.19	0.19	53.84	32.46	21.38	291.10
Sample	FE sep	n/a	n/a	-7.49^*	-4.48	55.98	33.09	22.89	286.28
	FE com	n/a	n/a	-6.70^*	-6.70^*	55.56	32.95	22.61	284.92
	RE sep	0.29	0.50	-6.70^*	-4.51	53.22	31.57	21.65	289.36
	RE com	0.27	0.45	-6.22^*	-6.22^*	52.78	31.47	21.31	287.71
Dosage	FE sep	n/a	n/a	-1.44^*	2.06	58.65	35.75	22.90	288.97
	FE com	n/a	n/a	-1.19^*	-1.19^*	60.61	35.80	24.80	289.95
	RE sep	$0.37^{\ddagger}$	0.49	$-1.25^{\ddagger}$	2.02	52.88	31.26	21.62	290.32
	RE com	$0.38^{\ddagger}$	0.54	-1.01	-1.01	53.05	31.16	21.90	290.31
Year	FE sep	n/a	n/a	0.08^*	0.05	57.45	34.65	22.80	287.68
	FE com	n/a	n/a	0.08^*	0.08^*	56.86	34.47	22.39	286.19
	RE sep	0.31	0.48	0.06	0.06	54.62	32.98	21.63	290.97
	RE com	0.29	0.44	$0.07^{\ddagger}$	$0.07^{\ddagger}$	54.08	32.89	21.19	289.24

Blue colour indicates a low value with darkest shading showing the lowest values. The ‡ symbol indicates significance at the 90% confidence level. τ refers to the between-studies-heterogeneity, D_{res} to the residual deviance and β to the meta-regression coefficient of the control variable in question which is modelled in the log-odds ratio scale. Subscripts (AD, PE) represent whether a measure refers to only adult or paediatric studies, while when there is no subscript the measure refers to the whole database (adults and paediatric patients).

The first feature to notice is that in contrast to the null models, FE models here perform better than RE with the exception of the meta-regression model on Jadad which struggles to explain any heterogeneity. All other models seem to at least partly explain heterogeneity, and improve the fit according to DIC.

The covariate which produces the best performing meta-regression models is duration of treatment, though this improvement seems to be driven only by the adult evidence[7]. This is further supported by the fact that compared to the model that imposes separate effect modification coefficients, when a common effect modification is imposed, the direction of β_{PE} changes and its magnitude becomes very similar to the adult one. As a result of this difference between the two populations, the FE model with population-specific coefficients fits the best in terms of both DIC and residual deviance.

FE meta-regression models on sample size also fit well and marginally better than random-effects. In this case, the magnitude of effect modification is similar among adult and paediatric studies and when a common coefficient is imposed, its estimate becomes more precise (CrI not shown in Table 5.6). The meta-regression models on year of publication provide a very similar fit with those on sample size, albeit slightly worse in terms of heterogeneity, residual deviance, and DIC.

[7] The fact that duration of treatment does not seem to be an important effect modifier in the paediatric studies may confirm the clinicians' suspicion about this variabe in Soares et al., 2012 where they were unable to intuitively explain the reason that it was the main source of heterogeneity in adults —see Soares et al., 2012 page 38.

Regarding alternative treatment parametrisations, the results of the meta-regression models on all covariates are provided in the Appendix in Table B.4.1, Table B.4.2, Table B.4.3. For illustrative purposes though, the models on duration of treatment are provided in Table 5.7 and for sample size and Jadad in the Appendix (Table B.4.4, Table B.4.5). Just as in Step 1, T3b is again much better than T2 and T3a, but not significantly better than T4. However, T3b may be preferred over T4 on parsimony grounds.

Table 5.7: *Results of meta-regression models on duration for all network parametrisations.*

covariate	Model	τ_{AD}	τ_{PE}	β_{AD}	β_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
T2	FE sep	n/a	n/a	-0.38^*	0.36	61.39	37.11	24.27	289.70
	FE com	n/a	n/a	-0.27^*	-0.27^*	69.01	38.12	30.90	296.34
	RE sep	0.37	0.46	-0.29	0.25	54.01	32.40	21.61	290.33
	RE com	0.43	0.57	-0.18	-0.18	53.28	31.71	21.57	290.52
T3a	FE sep	n/a	n/a	-0.37^*	0.29	58.24	37.57	20.67	288.55
	FE com	n/a	n/a	-0.27^*	-0.27^*	63.53	38.25	25.28	292.85
	RE sep	0.42	0.37	-0.28	0.31	53.14	32.63	20.51	290.34
	RE com	0.49	0.44	-0.14	-0.14	53.55	31.90	21.65	291.56
T3b	FE sep	n/a	n/a	-0.40^*	0.54	50.88	27.95	22.94	281.23
	FE com	n/a	n/a	-0.36^*	-0.36^*	53.25	27.99	25.26	282.58
	RE sep	0.19	0.50	-0.40^*	0.54	50.03	28.43	21.60	284.81
	RE com	0.20	0.57	-0.36^*	-0.36^*	50.50	28.32	22.17	285.26
T4	FE sep	n/a	n/a	-0.41^*	0.54	50.88	28.80	22.08	283.23
	FE com	n/a	n/a	-0.36^*	-0.36^*	53.29	28.85	24.44	284.61
	RE sep	0.21	0.50	-0.40^*	0.54	50.42	29.15	21.27	286.72
	RE com	0.22	0.57	-0.36^*	-0.36^*	51.10	29.14	21.97	287.43

Blue colour indicates a low value with darkest shading showing the lowest values. τ refers to the between-studies-heterogeneity, D_{res} to the residual deviance and β to the meta-regression coefficient of the control variable in question which is modelled in the log-odds ratio scale. Subscripts (AD, PE) represent whether a measure only refers to adult or paediatric studies, while when there is no subscript the measure refers to the whole database (adults and paediatric patients).

5.6.2.3 Step 3 : Combining covariates

The results of Step 2 suggest that duration of treatment is the most important predictor of the relative effect, followed by sample size and year of publication. In this step, these covariates are combined to assess whether the models would fit better and explain a larger part of heterogeneity. The models that combined year with sample size and all three variables together did not converge[8] despite additional efforts such as centering or thinning the MCMC chains. The results of the two remaining combination meta-regression models are shown in Table 5.8.

Table 5.8: *Step 3. Meta-regression models with multiple covariates in T3b network. The first two columns of β correspond to the adult and paediatric coefficients for the first variable and the next two columns for the second variable.*

Covariate	Model	τ_{AD}	τ_{PE}	$\beta_{1,AD}$	$\beta_{1,PE}$	$\beta_{2,AD}$	$\beta_{2,PE}$	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
Duration + Sample	FE sep	n/a	n/a	-0.35^{*}	-0.18	-1.6	-5.7	51.74	28.76	22.98	283.14
	FE com	n/a	n/a	-0.26^{*}	-0.26^{*}	$-3.9^{\ddagger}$	$-3.9^{\ddagger}$	51.42	28.69	22.72	281.8
	RE sep	0.20	0.48	-0.34^{*}	-0.04	-2.04	-4.7	50.69	29.12	21.56	286.32
	RE com	0.20	0.47	$-0.27^{\ddagger}$	$-0.27^{\ddagger}$	-3.85	-3.85	50.11	28.77	21.34	284.76
Duration + Year	FE sep	n/a	n/a	-0.36^{*}	6.0	0.015	-0.57	51.82	28.78	23.03	283.26
	FE com	n/a	n/a	-0.27^{*}	-0.27^{*}	$0.05^{\ddagger}$	$0.05^{\ddagger}$	51.67	28.84	22.82	281.99
	RE sep	0.22	0.50	-0.36^{*}	-2.5	0.01	0.32	50.63	29.13	21.5	286.3
	RE com	0.21	0.48	-0.29^{*}	-0.29^{*}	0.04	0.04	50.78	29.38	21.4	285.7

Blue colour indicates a low value with darkest shading showing the lowest values. τ refers to the between-studies-heterogeneity, D_{res} to the residual deviance, and β_1 and β_2 to the meta-regression coefficient of the first and the second control variable in question which is modelled in the log-odds ratio scale. Subscripts (AD, PE) represent whether a measure only refers to adult or paediatric studies, while when there is no subscript the measure refers to the whole database (adults and paediatric studies).

The first thing to note is that the effects of sample size, modelled as $1/\sqrt{N}$, and the year of publication are not significant at the 95% level when separate coefficients are assumed in each population. In the first set of models (Duration + Sample), common coefficient models perform slightly better than those with separate coefficients. This may be because the coefficient for duration across paediatric studies has the same sign as in adults, and as a result when a common coefficient is imposed, the residual deviance does not increase and a better fit is observed overall. Heterogeneity in adults is comparable to that estimated with the T3b meta-regression solely on duration (Table 5.6), indicating that sample size does not explain any additional heterogeneity in adults. However, it does on paediatric studies as can be observed by comparing the paediatric-specific heterogeneity estimates in Table 5.8 with the corresponding estimates in Table 5.6, and this leads to lower total and paediatric-specific residual deviance values.

[8] According to Gelman-Rubin convergence diagnostics.

As mentioned above, the model that included both the year of publication and the sample size did not converge. This may indicate that these variables explain the same component of heterogeneity, implying that only one is enough to explain the heterogeneity due to study quality. This hypothesis is further supported by the fact that the last meta-regression models that use the duration of treatment and year of publication perform similarly to the models that use the duration of treatment and the sample size.

Finally, given the conclusions from all steps, 'hybrid' models were attempted (Table 5.9), using different specifications for each evidence set. Note that the first model with FE meta-regression on duration in adults and a FE model without any covariates in paediatric patients provides the best fit amongst all attempted models in this section.

Table 5.9: *'Hybrid' models that do not use the same covariates in the two populations. T3b network parametrisation.*

covariate	Model ad	Model paed	τ_{AD}	τ_{PE}	β_{AD}	β_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
Duration (adults)	FE M-regr	FE	n/a	n/a	-0.4^{*}	-	50.82	27.96	22.87	280.1
Duration (adults)	FE M-regr	RE	n/a	0.47	-0.4^{*}	-	49.34	27.96	21.39	281.7
Dur (ad); Sample (paed)	FE M-regr	FE Mregr	n/a	n/a	-0.4^{*}	-5	51.28	27.96	23.32	281.9

Blue colour indicates a low value with darkest shading showing the lowest values. τ refers to the between-studies-heterogeneity, D_{res} to the residual deviance and β to the meta-regression coefficient of the control variable in question which is modelled in the log-odds ratio scale. Subscripts (AD, PE) represent whether a measure only refers to adult or paediatric studies, while when there is no subscript the measure refers to the whole database (adults and paediatric studies).

5.6.2.4 Model selection

The previous steps identified duration of treatment as the covariate that explained the largest part of heterogeneity, particularly under a T3b treatment parametrisation. However, there is no evidence of such an effect on paediatric patients, and hence a FE meta-regression 'hybrid' model was selected as the base-model. Despite the fact that this model fits significantly better than any other model that either uses another covariate or no covariate at all, more models need to be included in our final list. This is because, as explained in Soares et al., 2012, page 38, clinical experts did not identify a rationale for treatment duration to be an important effect-modifier. Instead, they thought that covariates that related to study quality were more important and should be considered. Therefore, here the list of final base-models is expanded to accommodate the experts' views on this matter, enabling both a more comprehensive comparison with the findings of Soares et al., 2012, and an assessment of whether conclusions regarding the performance of ISMs are consistent across different base-models.

Regarding the meta-regression models on sample size, which was the second most important effect modifier, T3b again yields the lowest DIC values (Table B.4.4). Here, the extent of the effect modification is similar among adults and paediatric patients, and the FE meta-regression model with the common coefficient fits the best. In Soares et al., 2012, a T2 RE meta-regression model on sample size was preferred. Therefore, by including the paediatric evidence here, we were able to estimate the sample size effect more precisely and move from a RE to a FE model, and from T2 network to T3b.

Based solely on DIC values, we could stop here and move forward only with these two models. However, to allow a more thorough comparison with the original HTA (Soares et al., 2012), we will include two more models that were included in Soares et al., 2012 after consultation with clinicians.

Among simple models without any covariates, T3b is preferred here because it yields the lowest DIC values under FE. However, since heterogeneity is evidently not negligible, RE models are more appropriate and fit better. Therefore, our final list also includes the simple T3b RE model, just as in Soares et al., 2012.

Finally, even though the meta-regressions on Jadad did not perform very well in Soares et al., 2012, the authors nevertheless considered a Jadad model[9]. In this work, according to Table B.4.5 (see Appendix, page 225), all FE and RE models across all treatment parametrisations fit very similarly, so T2 is preferred on parsimonious grounds. Within T2 meta-regression models on Jadad, there is significant remaining heterogeneity and RE models should be preferred. Among these, the one with the common coefficient seems to perform slightly, but not significantly, better and is therefore included in our list of base-models. The final list of base-models is shown in Table 5.10 and a comparison between the final base-models in this work and those in Soares et al., 2012 is provided in the Appendix in Table B.4.6 and Figure B.4.1.

Table 5.10: *The final list of models used in this work.*

Final Models for CEA	Posterior mean (adults)	se	DIC
T3b : FE + duration (adults), FE (paediatric patients)	-0.3	0.13	280.1
T3b : FE + sample size (common)	0.02	0.14	284.9
T3b : RE	-0.56	0.6	290.3
T2 : RE + Jadad (common)	-0.38	0.6	291.1

Estimates correspond to predictions for a duration of treatment of 3 days, a treatment arm sample size of n = 339, which is the largest amongst all adult studies, and Jadad = 5, which corresponds to a study of the best possible quality.

[9] This model was included for the following reason: the prediction of the Jadad model for a study of the best quality (i.e. Jadad = 5) provided an identical point estimate to the meta-regression model on the year of publication (Year = 2007) and it is much easier to suggest that the CEA should consider the relative estimate of a study of the best quality instead of a study published in 2007. Hence they preferred the model on Jadad instead of the model on the year of publication.

5.7. Sharing information between adult and paediatric evidence

5.7.1. Methods

This section comprises of two parts. First, the applicable methods that were used to share information among paediatric and adult studies on the relative effectiveness parameter are identified for each one of the base-models chosen in Section 5.6.2.4. This made use of the 'methods identification framework' introduced in Section 4.5. Second, the statistical measures that were used to both compare the ISMs and measure the strength of information-sharing that they imposed are described.

5.7.1.1 Information-sharing methods

Application of 'methods identification framework'

For the *Identification* step, the direct (adult) and indirect (paediatric) evidence was identified in the literature according to the methods described in Section 5.4. As described in Section 5.4.2, the extended evidence base comprised of 17 *'direct'* studies on adults (the same studies that were included in Soares et al., 2012) and 11 new *'indirect'* studies in paediatric patients including a large multi-center study that enrolled 85% ($n = 3493$) of all the paediatric patients.

In the *Parametrisation* step, a variable was chosen to distinguish between direct and indirect evidence. This was straightforward in our case, because the only option was a binary variable where one value (say 0) indicates that a study enrolled adults, whilst the other (say 1) indicates that a study enrolled paediatric patients. Since there was no prior expectation on which population would exhibit a larger relative effect, the binary variable was unordered. Had there been any evidence or biological justification to suggest that one of the two populations should exhibit a larger relative effect, an ordered variable that reflected this a priori assumption could have been chosen.

In the *Base-model selection* step, alternative model parametrisations and meta-regression models were fit to the extended evidence base to identify the best-fitting models for each evidence set. When the same effect modifiers were identified for both evidence sets, meta-regression models that assumed a common effect modification coefficient were implemented to achieve increased precision in the coefficient estimate. This process was undertaken in Section 5.6 and led to the selection of 4 base-models that are shown in Table 5.10.

In the *Eligibility* step, I went through all the models that were described in Chapter 4 and consulted Table 4.2 and Table 4.3 to identify which methods were eligible for the type of variable that we were using to distinguish between direct and indirect evidence.

Note that RE base-models would be eligible for more methods than FE base-models, because they provide the heterogeneity component that can be shared using alternative assumptions.

In the *Plausibility* step, methods that imposed unrealistic assumptions were eliminated, as well as methods that had data requirements which could not be accommodated by the available data. Regarding assumptions, this work considered all alternative models to make explicit the strength of information-sharing that would be imposed had all the models been considered plausible. Regarding data requirements, given that no single study reports results for both populations, meta-regression models would not impose any information-sharing and hence were not be considered here. Additionally, since an unordered variable was used to describe the types of evidence available, random-walks were not be applied because when only one indirect evidence set is used, the random-walk model is very similar to a multi-level model. Lastly, methods that shared information on the between-studies heterogeneity could only be used for the RE base-models.

Table 5.11: *Applicable methods for each base-model of Table 5.10.*

Method	T3b RE	T3b FE M-reg (duration)	T3b FE M-reg (sample size)	T2 RE M-reg (Jadad)
Lumping	✓	✓	✓	✓
Common Heterogeneity	✓	X	X	✓
Multi-level	✓	✓	✓	✓
Informative prior	✓	✓	✓	✓
Mixture prior	✓	✓	✓	✓
Power-prior	✓	✓	✓	✓
Commensurate prior	✓	✓	✓	✓
Prior on heterogeneity	✓	X	X	✓

Finally, in the *Implementation* step all the remaining methods shown in Table 5.11 were applied. Note that for RE base-models, the predictive distribution was used rather than the posterior distribution of the mean. This is because in the presence of heterogeneity, the predictive distribution better represents uncertainty[10] (Dias et al., 2011b). A plain language explanation of how the ISMs used here work, see (subsection B.5) on page 227 of the Appendix.

[10] It incorporates uncertainty due to heterogeneity on top of the uncertainty around the point estimate.

Specification of implemented models

As noted in Section 4.4, some of the ISMs allow further flexibility, by including parameters that influence information-sharing, and hence additional decisions were made regarding their specification. In particular, power-prior models were applied for all α values between 0 and 1 in 0.1 increments. Mixture priors were used assuming that the weights of the prior components are uncertain parameters estimated in the model. Commensurate priors were implemented assuming that the Bernoulli trials had a fixed 50% change of yielding the 'spike' or the 'slab' hyper-prior. Finally, for RE models, the predictive distribution was used as an informative prior for the two-step models instead of the posterior mean distribution, and the lumping model which analyses direct and indirect evidence under a single random-effect (i.e. with common d and τ) since it is the most commonly used in the literature.

It should also be highlighted that the ISMs developed in Chapter 4 consider only the simple case where no effect modifiers are taken into account in the direct or the indirect evidence. However, here, most base-models were meta-regression models. This complicates the use of ISMs because, under meta-regression models, the relative effect comprises of two parts; a component that is only due to the effect modifier considered (i.e. the β slope) and depends on its value, and another component that is independent of particular variables. The need to extend ISMs to appropriately accommodate this additional complexity was overcome here by choosing to center the covariate at the value on which we wanted to relate the two evidence sets. For instance, say we wanted to express the assumption that the predictions for the relative effect of the direct and indirect evidence were equal at the effect modifier value of $X = 5$. Then, instead of extending methods to share information on both relative effect components ($d + \beta$), we could just center the covariate at $X = 5$ and share information only on d using the methods developed in Chapter 4. This is further explained on page 228 in the Appendix. Here, the 'hybrid' base-model that considered the duration of treatment related the RTE of the indirect evidence with the predicted RTE of the direct evidence for three days of treatment, as this was the treatment duration in most adult studies. The base-model that considered Jadad score related the predicted RTE of the direct and indirect evidence for a study of the highest quality (i.e. $Jadad = 5$). Finally, the base-model that considered sample size related the predicted RTE of the direct and indirect evidence for a treatment arm of $n = 339$, which was the largest treatment arm found amongst the adult studies.

All models were programmed in *R* (R Development Core Team, 2010) and estimation was undertaken in *WinBUGS* (MRC Biostatistics Unit, 2010) using the R2Winbugs package (Sturtz et al.) and the coding developments that were introduced in Chapter 4.

5.7.1.2 Measuring the 'strength of information-sharing'

The inclusion of the indirectly related data in the analysis implies that we are allowing the paediatric evidence to influence the adults estimates. Here, the focus is only on the influence of the indirect evidence on the relative effectiveness parameter that pertains to the population considered by the decision. To quantify the extent of influence allowed by each ISM, deviations from the no-sharing/only adults splitting method, which is here considered the base-case, were measured. It should be highlighted that since we do not know what the true adult relative effect is, we have no way of assessing which estimate is closer to the truth. Hence, the only divergence that we can calculate is the distance from the estimate that is produced using only the adult information to the estimate that is produced using the extended evidence base after implementing each ISM. This section describes in detail the three strength-of-sharing measures that will subsequently be used in this chapter.

Point Estimate Divergence

The first measure is the Point Estimate Divergence (PED), evaluating the absolute difference in the adult relative effectiveness posterior mean between the no-sharing/only adults splitting method ($d_{ad}^{model_0}$) and each of the j alternative ISMs ($d_{ad}^{model_j}$). Mathematically, this is defined as:

$$PED_j = |d_{ad}^{model_j} - d_{ad}^{model_0}| \tag{5.1}$$

where a larger PED_j implies a larger difference in the adult point estimate between the no-sharing/only adults splitting method and ISM j.

Precision Increase

To capture changes in the standard error between the no-sharing/only adults splitting method ($sd_{d_{ad}}^{model_0}$) and each of the j ISMs ($sd_{d_{ad}}^{model_j}$), Precision Increase (PrI), which was introduced in Jackson et al., 2017[11], is used. This is defined as:

$$PrI_j = 1 - \frac{sd_{d_{ad}}^{model_j}}{sd_{d_{ad}}^{model_0}} \tag{5.2}$$

where $PrI \in (-\infty, 1]$. Negative PrI values indicate that information-sharing has led to increased uncertainty, and a value of 0 that information-sharing results in the same uncertainty as the no-sharing/only adults splitting method. In contrast, as information-sharing leads to more precision gains compared to splitting, PrI tends to 1.

[11] Note that in Jackson et al., 2017 this measure is termed BoS. However, here, multiple measures are used and hence this measure is renamed to better reflect the quantity that it is calculating.

Kullback-Leibler divergence

Finally, a measure that simultaneously considers changes in the posterior mean and standard deviation was used. This is based on the notion of 'divergence' used in information theory[12] to derive information criteria (McElreath, 2016) and in the context of our example can be defined as the additional uncertainty that is induced by using the probability distribution of $d_{ad}^{model_j}$ to describe the probability distribution of $d_{ad}^{model_0}$. A metric that suits this description is Kullback-Leibler Divergence[13] (Kullback and Leibler, 1951) and is defined as:

$$D_{KL}(p,q) = \int [p(x) \times log(p(x)) - p(x) \times log(q(x))] \; d(x) \tag{5.3}$$

where p is the target distribution that we aim to describe (here $d_{ad}^{model_0}$), and q the distribution that we use in doing so (here $d_{ad}^{model_j}$).

The integral is calculated in R with adaptive quadrature which is a method of numerical integration using the 'integrate' command of the 'stats' package. It is important to highlight that interpretation of KL differences is challenging (i.e. it is tough to establish whether a KL value is high or low in absolute terms), and it is here used only in relative terms to compare the KL of different ISMs. An illustration of KL-divergence from a standard normal distribution is provided in the Appendix in Figure B.5.1 and the R code, which is based on a previous application of this metric by Jackson, 2019, can be found in `https://github.com/NikolaidisGFZ/PHD.git`.

[12] Information entropy is defined as the uncertainty that is contained in a probability distribution and is described as $H(p) = -\sum_{i=1}^{n} p_i log(p_i)$. Cross-entropy is the additional uncertainty that is induced when events appear based on $p(x)$, but we are instead using $q(x)$ to predict them i.e. $H(p,q) = -\sum_{i=1}^{n} p_i log(q_i)$. The difference between cross-entropy $H(p,q)$ and the entropy $H(p)$ is the Kullback-Leibler divergence i.e. $D_{KL} = H(p,q) - H(p)$.

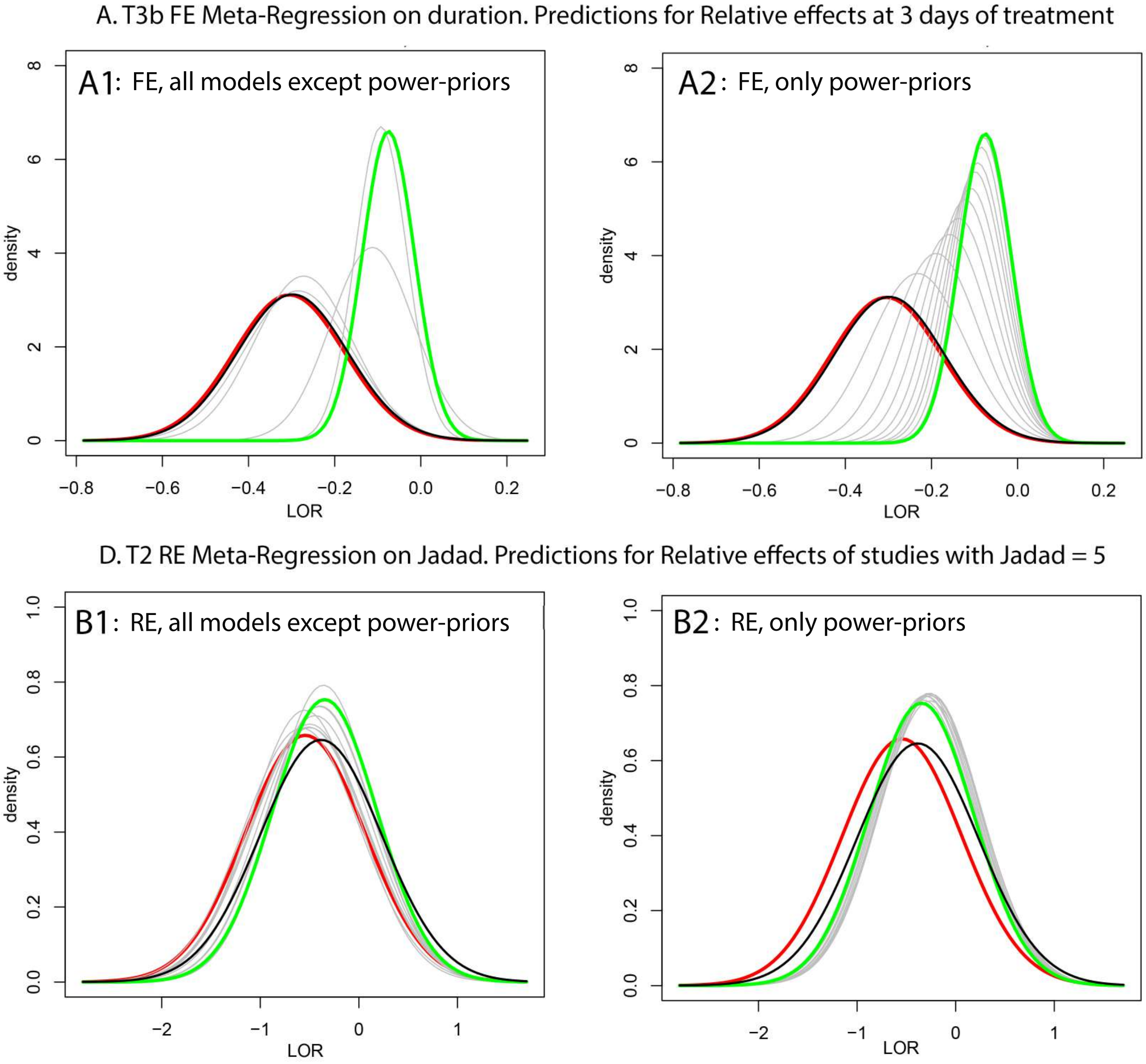
[13] Although, this was originally outlined by Harold Jeffreys in *Theory of probability* in 1948.

5.7.2. Results

The resulting log-odds ratio estimates across all ISMs are displayed in Figure 5.6. In the main text only two base-models are discussed: the first is the T3b FE meta-regression on the treatment duration (days), henceforth referred to as the *Duration FE base-model*. This base-model is shown because it is the best-fitting base-model. The second base-model shown here is the T2 RE meta-regression on Jadad score, henceforth referred to as the *Jadad RE base-model*. This base-model is shown because it allows us to see how the ISMs also impact RE base-models, and is also likely to be of interest to decision-makers since it can produce the predicted RTE for a study of the best possible quality (i.e. Jadad = 5). The remaining results of the T3b FE meta-regression on sample size and the T3b RE model are included in Table B.5.1 and Table B.5.2 of the Appendix respectively.

Note that in Figure 5.6, the x-axes of the two base-models are differently scaled; an illustration under a common scale is provided in Figure B.5.2 of the Appendix. Black lines correspond to the original estimate used in Soares et al., 2012; red lines to the no-sharing/only adults splitting method; green lines to the estimate of the most extreme ISM i.e. lumping; and grey lines to the estimates of the remaining ISMs. The graphs on the right (A2, B2) depict estimates produced using the Power-prior and α values from 0.1 to 1 in 0.1 increments, while the graphs on the left (A1, B1) show the estimates of the remaining ISMs. Despite the fact that the x-axes are not equally scaled, it can be inferred that for the Duration FE base-model results seem less similar across ISMs (A1) than for the Jadad RE base-model (B1). Furthermore, in the power-prior FE models (A2), as α increases, estimates 'smoothly' move from splitting to lumping, whilst in RE power-prior models (B2), estimates are always similar to lumping (even for small α values), and may actually exceed it (i.e. the resulting log-odds ratios of some ISMs as indicated by the grey lines are not contained in the range defined by splitting and lumping).

Figure 5.6: *Resulting estimates of all ISMs for the Duration FE and the Jadad RE base-models.*



Figures on the left illustrate all non-power-prior methods while the right figures all the Power-prior methods with $\alpha = 0.1, 0.2, ..., 1$. Black estimates correspond the original estimates from Soares et al., 2012; red to the no-sharing/only adults splitting method in this work; green to lumping; and grey to the various remaining ISMs.

The fact that power-prior models under RE can produce more extreme results (not contained in the spectrum defined by splitting and lumping) is also apparent in Table 5.12[14]. DICs are reported only for those ISMs that use the same data within a single model and can therefore be compared[15].

Table 5.12: *Predictions across all eligible ISMs for the T3b Duration FE and the T2 Jadad RE base-models.*

	T3b Duration FE base-model			T2 Jadad RE base-model		
ISM	RTE	sd	DIC	RTE	sd	DIC
Base-case (no sharing/only adults)	-0.31	0.13	280.1	-0.55	0.31	290.8
Lumping	-0.08	0.06	283.3	-0.36	0.24	289.8
Multi-level	-0.28	0.12	279.8	-0.50	0.29	290.5
Commensurate prior	-0.27	0.11	281.7	-0.44	0.28	293.1
Informative prior	-0.09	0.06	-	-0.50†	0.29†	-
Mixture prior	-0.11	0.10	-	-0.51†	0.29†	-
Common Heterogeneity	n/a			-0.55	0.29	289.3
Prior on Heterogeneity	n/a			-0.58	0.30	-
Power-prior (a=0.1)	-0.23	0.11	-	-0.26	0.27	-
Power-prior (a=0.2)	-0.19	0.10		-0.25	0.26	
Power-prior (a=0.3)	-0.16	0.09		-0.26	0.25	
Power-prior (a=0.4)	-0.14	0.08		-0.27	0.25	
Power-prior (a=0.5)	-0.12	0.08		-0.27	0.24	
Power-prior (a=0.6)	-0.11	0.07		-0.29	0.25	
Power-prior (a=0.7)	-0.10	0.07		-0.30	0.25	
Power-prior (a=0.8)	-0.09	0.07		-0.31	0.24	
Power-prior (a=0.9)	-0.08	0.06		-0.33	0.25	
Power-prior (a=1)	-0.08	0.06		-0.35	0.25	

The estimates correspond to studies of the best quality (i.e. $Jadad = 5$) and 3 days of treatment duration respectively. The symbol † indicates that the result pertains to the use of the predictive distribution of the indirect evidence. Light red shading indicates the base-case no-sharing/only adults splitting method for comparison purposes. ISM: Information-sharing method.

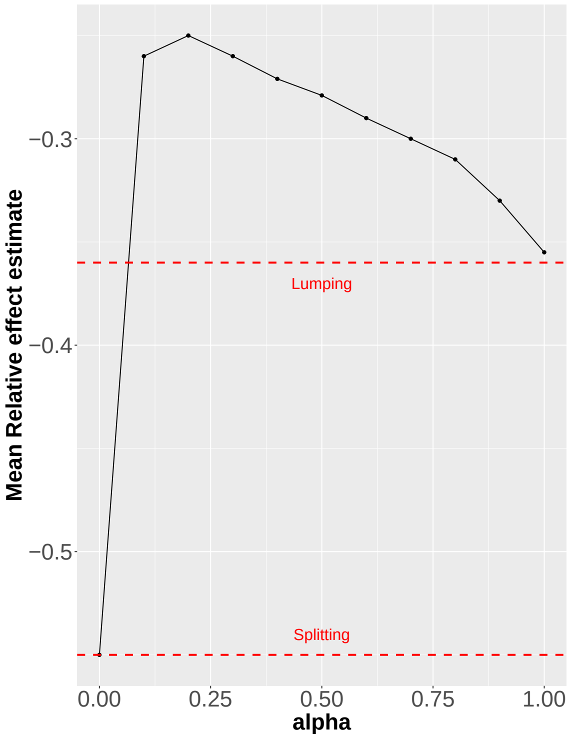
Interestingly, in the Jadad RE base-model, for low values of α, the power-prior yields results which share information more strongly than lumping, and as α gets closer to 1, the estimates get closer to lumping. This is here attributed to the way that the power-prior operates and to the nature of the indirect evidence. Specifically, there are only two paediatric studies that contribute directly to the comparison of interest (ALB *vs*

[14] The resulting estimates of the two remaining base-models can be found in the Appendix in Table B.5.1 and Table B.5.2.

[15] For example, informative priors cannot be compared with lumping because the first method is a two-step process where, in the first step, the relative effect is estimated for the indirect evidence and, in the second step, the direct evidence are solely analysed using a prior that was derived from the first step. Hence, the second analysis of the two-step process uses less data and its DIC cannot be compared with lumping that combines the whole of the extended evidence base in a single step.

IVIG/IVIGAM) (see Table 5.3) and their findings differ. In the big study (Brocklehurst et al., 2011), an odds-ratio of 1.00 was observed suggesting that there is no effect, and in the very small study (Weisman et al., 1992) with an Odds-ratio of 0.39 suggesting that IVIG is associated with a larger reduction in mortality than ALB. For small α values, the likelihood[16] of the big study gains some weight, whilst the likelihood of the small study remains negligible. Hence, as illustrated in Figure 5.7, the combined estimate is initially excessively pulled towards the neutral effect of the big study and we observe the most extreme estimate for $\alpha \approx 0.2$. As α values increase, the likelihood of the small study which suggests a large effect, becomes non-negligible and starts influencing the estimates pulling them towards the effect that is observed under lumping. This explanation perfectly aligns with the idea that RE models give higher weight to small studies. Essentially, for small α values the Brocklehurst et al., 2011 study is a small-study and is therefore assigned a disproportionately large weight, whilst the small paediatric studies are practically non-existent. A thorough explanation of this is provided in the Appendix on page 235.

Figure 5.7: *The predicted relative effects (IVIG/IVIGAM vs ALB) for Jadad = 5 of power-prior models with varying alpha values between 0 (no-sharing/only adults) and 1 (full sharing).*



[16] The word 'likelihood' here refers to the Bayesian 'likelihood function'.

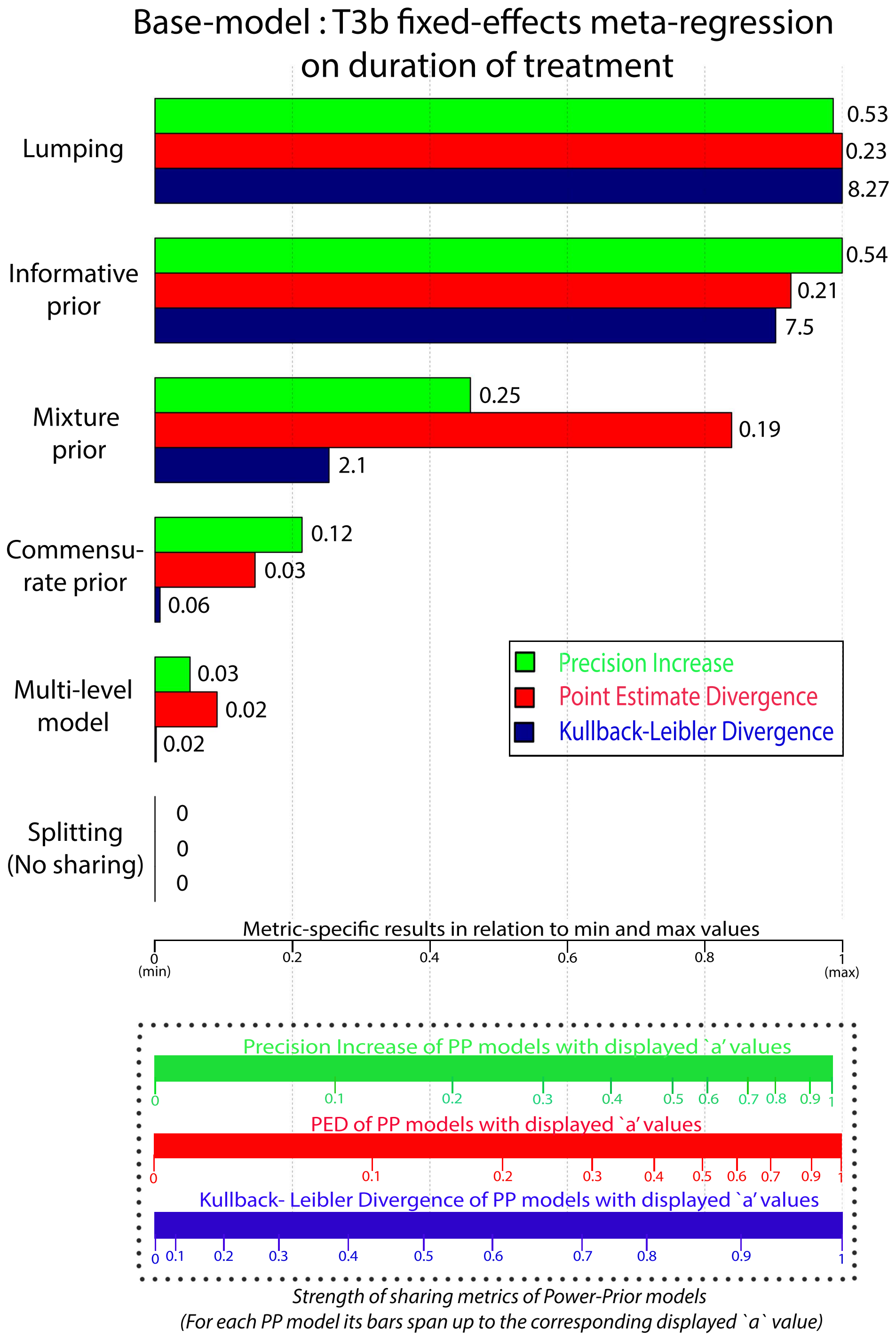
The relative and absolute values of the ‘strength-of-borrowing’ measures are depicted for the Duration FE base-model using coloured bars in Figure 5.8. ISMs are listed on the left in descending KL-divergence order. The x-axis displays relative values (i.e. a ratio of each model’s value divided by the maximum value observed for each measure) and therefore ranges from 0 (min value) to 1 (max value). Each bar represents the relative value for a strength of sharing measure. For example, in Figure 5.8 the precision increase (green bar) that is achieved by the commensurate prior model is 21% of the maximum precision increase across all models (which is observed for the informative prior). The absolute values of the three strength-of-sharing measures (i.e. PED, PrI, and KL) are noted in text to the right of each bar. For instance, the point estimate divergence (red) of the mixture prior method is 0.19 in the log-odds ratio scale.

The additional coloured axes at the bottom of the graph indicate the point at which each strength of sharing measure is estimated for a power-prior model with the corresponding α. For a power-prior model with any α, its corresponding bars span from the beginning of the axis up to the point where the α value of interest is displayed on the axis. For instance, PED of the power-prior model with $\alpha = 0.5$ would span from the beginning of the red axis up to the point where the value 0.5 is displayed on this axis. By representing power-prior models with varying α in this way, the reader can ‘map’ the various ISMs to particular α values of the power-prior models. For example, the mixture prior model achieves a KL-divergence (blue) similar to that of the power-prior model with $\alpha = 0.4$. Hence, in this case-study mixture priors share information as strongly as a power-Prior model with $\alpha = 0.4$ (this can be intuitively understood by drawing a vertical line from the end of the blue bar of the mixture prior model and noticing that it cuts the blue x-axis close to 0.4).

As expected, multi-level models and commensurate priors —both of which rely on the exchangeability assumption[17] —produce the lowest strength of sharing according to all three measures in this case-study. Also, as expected, for the Duration FE base-model, across all measures, the informative prior imposes a very similar strength of sharing with that of lumping. Finally, in this case-study mixture priors impose less information-sharing than informative priors, but the level of sharing is not consistent across metrics (i.e. they impose around 80% of lumping’s PED, but only around 40% of Lumping’s PrI).

[17] Recall that the commensurate prior is essentially a random-walk where a spike-and-slab hyperprior is imposed on the precision parameter, encouraging or discouraging information-sharing.

Figure 5.8: *Statistical measures of ISMs used in the FE meta-regression on duration base-model.*



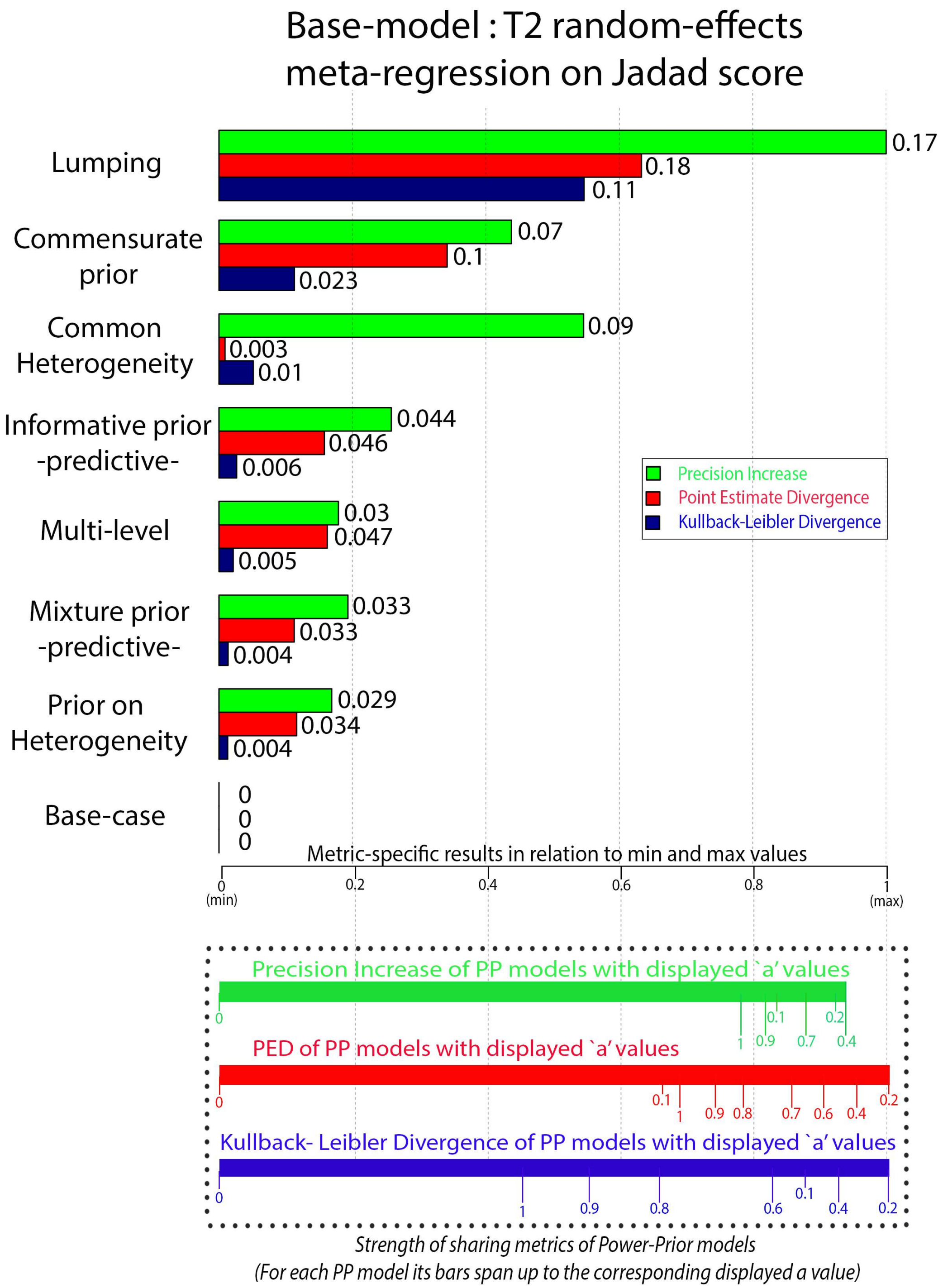
The additional coloured bars in the bottom display, in a compact way, the strength of sharing measures of all the power-prior models with α between 0 and 1. For example, a power-prior model with $\alpha = 0.5$ would produce a red bar spanning from the beginning of the current red bar up to the point where the value 0.5 is displayed.

The same chart for the Jadad RE base-model is displayed in Figure 5.9. Some counter-intuitive features are observed here. In particular, the power-prior model with $\alpha = 0.2$ achieves the maximum PED and KL divergence across all models. As a result, none of the ISMs listed on the left has a red or blue bar reaching up to 1. Furthermore, all power-prior models with $\alpha > 0.2$ exhibit lower values across all measures than the model with $\alpha = 0.2$. In particular, as α increases beyond 0.2 all measures decrease. This is indicated in Figure 5.9 by the longer lines used to display the α indicating that a power-prior using an $\alpha > 0.2$ achieves a lower absolute (and relative) value for the measure in question than the power-prior model with $\alpha = 0.2$. This feature is again a reflection of the issue that was highlithed in Table 5.12 and explained in Figure 5.7 and in the Appendix on page 235.

Not surprisingly, lumping the two evidence sets under a common random-effect leads to the highest precision gains (as indicated by PrI), although its PED and KL are surpassed by almost all power-prior models above with $\alpha > 0.1$. Lumping direct and indirect evidence only on heterogeneity seems to also considerably increase precision (PrI), with minimal changes in the posterior mean (PED). In this case-study, the mixture prior shares less information than the informative prior. This should be expected because once the informative part is mixed with the vague component, the resulting prior is less informative and therefore does not influence the results as much as when it is used on its own. With regards to the two models that share information on the between-studies heterogeneity, assuming a common parameter for direct and indirect evidence leads to higher KL and precision gains (PrI), but lower PED compared to using the heterogeneity estimate from the indirect evidence on a log-normal prior. All the remaining methods seem to impose similar information-sharing across all measures, except for power-priors which even for very small values of α impose comparably very strong information-sharing, as explained on page 121.

It is worth noting that under this RE base-model informative priors and mixture priors share much less information relative to lumping compared to the FE base-model (Figure 5.8). The reason for this somewhat counter-intuitive result is that the predictive distribution of the indirect evidence is used to form the prior distribution in the Jadad RE base-model, whilst the posterior mean distribution of the indirect evidence in necessarily used in the FE base-model.

Figure 5.9: *Statistical measures of ISMs used in the Jadad RE base-model.*



The additional coloured bars in the bottom display in a compact way the performance measures of all the Power-prior models with α between 0 and 1. For example, a Power-prior model with $\alpha = 0.5$, would produce a red bar that would span from the beginning of the red axis up to the point where the value 0.5 is displayed.

Table 5.13 ranks all ISMs except for power-priors[18] according to all three 'strength-of-sharing' measures, across all base-models. A lower number indicates that a method imposes stronger information-sharing in terms of the measure in question and hence ranks higher. For FE base-models, the results are consistent across the two base-models. The information-sharing spectrum is clearly defined by lumping and splitting and there is no discrepancy in how the various methods compare to the spectrum extremes. Specifically, informative priors —being almost equivalent to lumping —rank second, followed by mixture-priors that rank third. Commensurate priors rank consistently in the fourth place, imposing more information-sharing than multi-level models due to the additional assumption on the variance, followed by multi-level models that rank fifth and impose the least information-sharing.

Regarding RE base-models, amongst non-power-prior ISMs, the extremes of the spectrum are again clearly defined by lumping and splitting; however methods do not rank within that spectrum as consistently as under FE. Common heterogeneity seems to rank high in terms of precision gains (PrI), but very low in terms of changes in the point estimate (PED). However, this is not the case for the log-normal prior on heterogeneity, which leads similarly minor changes in the mean, but also lower precision gains. Furthermore, informative and mixture priors behave similarly with multi-level models and hence rank lower than commensurate priors. This could potentially be attributed to their use of the predictive rather than the posterior mean distribution, and is an issue we is explored in depth in Chapter 7.

Table 5.13: *Relative ranking of ISMs in descending order (1 is the highest).*

ISM	FE Mregr on duration	FE Mregr on sample	RE (Predictive distrib)	RE Mregr on Jadad (Predictive distrib)
Lumping	1-1-2	1-1-1	1-1-1	1-1-1
Common heterogeneity	-	-	2-6-2	3-7-2
Multi-level	5-5-5	5-5-5	4-3-7	5-3-6
Commensurate prior	4-4-4	4-4-4	3-2-3	2-2-3
Informative prior	2-2-1	2-2-2	5-4-4	4-4-4
Mixture prior	3-3-3	3-3-3	6-5-6	7-6-5
Prior on heterogeneity	-	-	7-7-5	6-5-7
Base-case (no-sharing)	6 - 6 - 6	6 - 6 - 6	8 - 8 - 8	8 - 8 - 8

The first number (blue) ranks methods according to their KL-divergence, the second (red) according to their PED, and the third (green) according to their PrI. Power-prior models are not listed here because they can accommodate various degrees of sharing. ISM: Information-sharing method.

[18] This is because they accommodate a range of possible degrees of information-sharing.

5.8. Discussion

This chapter describes a case-study in which strength is borrowed from indirect evidence, using several methods, with the aim of strengthening inferences. In particular, evidence from an indirect population (paediatric patients) is employed to estimate the relative effectiveness of the population of interest (adult patients). To achieve that, direct evidence from a previous secondary analysis by Soares et al., 2012 was used, indirectly related studies were identified by the means of a systematic review, heterogeneity in the extended evidence base was re-explored, and finally, the two evidence sets were combined using all the applicable ISMs introduced in Chapter 4. The various methods were compared according to a list of predefined 'strength-of-sharing' measures which sought to account for changes in the point estimate and its uncertainty separately and simultaneously.

To re-explore heterogeneity in the extended evidence base, the framework that was provided in Welton et al., 2015 was extended to allow for the simultaneous exploration of heterogeneity in multiple evidence sets. Interestingly, the inclusion of the indirect evidence enabled better characterisation of heterogeneity by leading to more precise estimation of the extent of the various covariates' effect-modification and consequently more definite conclusions about the best performing base-models. As a result, compared to Welton et al., 2015, a slightly modified list of best performing base-models was used.

The results of the application of power-prior models revealed that, even though their behaviour was as expected under FE base-models, this is not necessarily the case for RE base-models. In particular, we should not necessarily expect the estimated relative effect to move monotonously from the splitting to the lumping as α increases in the power-prior model (i.e. more information is borrowed from the indirect evidence). Scenario analyses showed that the sizes of the individual indirect studies need to be closely considered when using power-priors under RE base-models, because by discounting their likelihoods small studies may be negligible, whilst larger studies might still exert influence on the overall effect; particularly given that RE models assign disproportionately large weights to smaller studies.

This work also extends ISMs for meta-regression base-models by choosing to centre both evidence sets at the covariate value for which the predicted relative effects of the two sources are associated. However, this method does not generalise to other cases where we might want to relate predictions of different evidence sets that pertain to different covariate values (e.g. lumping the adult relative effect at three days of treatment duration with the paediatric relative effect at five days of treatment duration). Furthermore, it should be highlighted that for RE base-models, not all lumping options were attempted (see Section 4.4.1.1), but only those models that would be most likely to be used in the policy context (e.g. lumping all direct and indirect evidence under a common random-effect). Also, for RE base-models, two-step ISMs (i.e. informative priors, mixture priors) used only the predictive distribution (not the posterior mean) of the indirect evidence to inform the second step, as it incorporates uncertainty resulting from the between-indirect-studies heterogeneity. The impact of all the aforementioned limitations is explored in Chapter 7.

Overall, in this case-study, findings on how methods compare to each other in terms of the strength of sharing that they impose broadly align with expectations. In particular, it seems that lumping and splitting define a strength-of-sharing spectrum within which the various alternative ISMs can be positioned. However, under RE base-models, some exceptions were observed and were here attributed to the heterogeneity in study sizes and the RTEs suggested by the alternative studies. Also, informative priors seem to always impose stronger information-sharing than mixture priors which is expected based on their specification. Further, informative priors that use the posterior mean RTE of the indirect evidence (here implemented only under FE base-models) are almost equivalent to lumping. In addition, commensurate priors and multi-level models impose only subtle information-sharing, a feature that may be partially attributed to the fact that in this case-study there were only two evidence sets: one direct and one indirect. Finally, the value of α in power-priors does not necessarily relate to the imposed strength of sharing in a linear manner, and its interpretation may not be straightforward. Consequently, structured elicitation exercises that may potentially seek to elicit α can be challenging.

In conclusion, this chapter provided an example of the process that needs to be followed in order to combine direct and indirect evidence, it then implemented all possible ISMs, and finally provided insights into how the various methods compare under FE and RE, illustrating that the choice of ISM can have a significant impact on relative effectiveness estimates. However, it should be highlighted that the findings of this case-study do not necessarily generalise to other cases, and extensive simulation experiments need to be conducted to assess how different characteristics of the available evidence impact how methods compare in terms of the imposed strength of sharing. This is the focus of Chapter 7. Also, even more important than differences across methods on RTE estimates, is how these differences may impact policy, and in particular if they can lead to different adoption and/or research prioritisation decisions. This question will be the focus of Chapter 6.

Chapter 6

Policy-implications of information-sharing: cost-effectiveness and value of information analyses

6.1. Chapter aims and structure

In the previous chapter, the use of alternative ISMs was introduced in a case-study that sought to evaluate the effectiveness of IVIG/IVIGAM as an adjunctive therapy to the SoC for adult patients with severe sepsis and septic shock. The key features of the case-study relate to the unexplained heterogeneity within the available adult studies and the uncertainty associated with the effect modifiers and the RTEs. Indirect evidence was sought from studies exploring the use of the same intervention on paediatric patients. Heterogeneity was re-explored in the extended evidence base and applicable ISMs were identified using the step-by-step process introduced in Section 4.5. The findings suggested that different ISMs impose different degrees of information-sharing, and lead to different RTE estimates for adult patients. That said, whether or not the RTE differences are significant can only be ascertained by examining whether they could lead to different adoption decisions and/or further research recommendations.

This chapter uses a decision-model that was originally developed by Soares et al., 2012 in order to answer the following questions:

1. What is the impact of using different ISMs on adoption decisions?
2. What is the impact of using different ISMs on further research recommendation decisions?

The remainder of this chapter is structured as follows: In Section 6.2.1 the structure and parameters of the decision-model are detailed, while in Section 6.2.2 the policy measures that will be considered are explained. Section 6.3 describes the results of using estimates resulting from different ISMs on the policy measures, and differences in the results among methods are explored and explained. Finally, in Section 6.4 the findings of this chapter along with strengths and limitations are discussed.

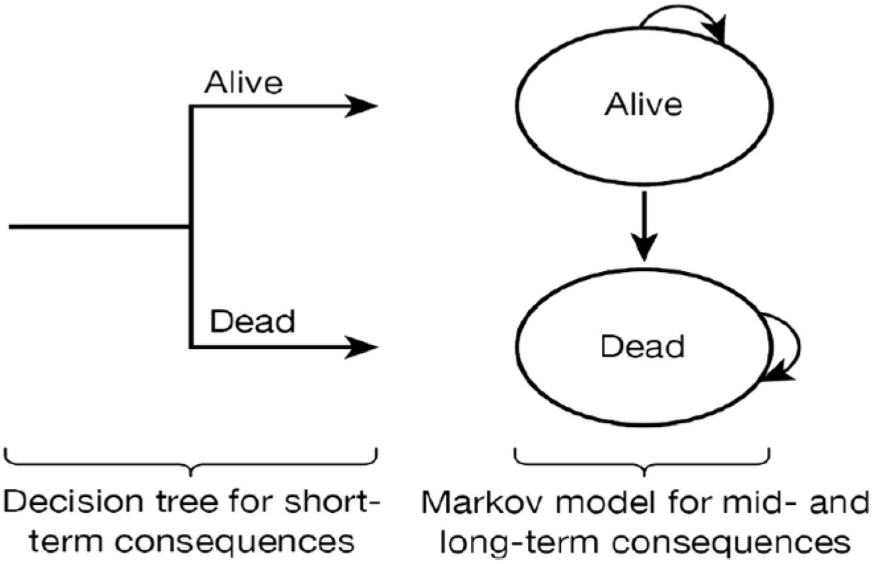
6.2. Methods

6.2.1. Decision model

This section summarises the basic characteristics of the decision model that was used in this work. The model was originally developed in Soares et al., 2012 and the reader is directed to Soares et al., 2012, (Chapter 5) for a more detailed description of the model's technical characteristics. The model sought to assess the cost-effectiveness of two alternative strategies for adult patients with severe sepsis and septic shock: IVIG/IVIGAM as an adjunctive therapy to standard of care (SoC), which includes antibiotics or ALB, *vs* SoC alone. Outcomes were expressed in Quality-adjusted Life Years (QALYs) and costs reflect 2009 UK prices from the perspective of the NHS. An annual discount rate of 3.5% was applied to both. Expected costs and outcomes were compared by the means of ICERs, effectively expressing the additional cost per QALY gained. ICER thresholds of both £20,000 and £30,000 per QALY gained were used as they represent the range of approval norms used by NICE.

The model evaluated the lifetime prognosis of severe sepsis and septic shock in order to capture the lifetime costs and consequences associated with the natural history in the absence of IVIG/IVIGAM. To achieve this the model is comprised of two distinct components which are shown in Figure 6.1:

Figure 6.1: *A simplified representation of the decision-model.*


Alive
Dead
Alive
Dead
Decision tree for short-term consequences
Markov model for mid- and long-term consequences

The decision tree that captures the short-term effects and on which the relative effect is applied, and the Markov-model that captures the lifetime prognosis conditional on survival in the short-term. This figure is adopted from Soares et al., 2014b.

1. A short-term decision tree: This component of the model evaluates the consequences of the initial hospitalisation period (in either critical on non-critical care) following the first sepsis episode, by quantifying the probability of an event (death) during this period. Baseline mortality for standard of care was sourced from the Intensive Care National Audit and Research Centre (ICNARC) Case Mix Program Database (CMPD) (Harrison et al., 2006). The relative treatment effect was only applied at this stage.

   For the T3b base-models, the relative effect of interest was IVIG/IVIGAM *vs* ALB whilst for the T2 base-models the comparison of interest was IVIG/IVIGAM *vs* ALB/PLA. Essentially, for each ISM estimated as part of every base-model, the corresponding log-odds ratio estimated in Section 5.7.2 was combined with the baseline mortality to produce an absolute probability of an all-cause mortality event under IVIG/IVIGAM.

2. A long-term Markov model: A simple Markov model with two mutually exclusive health states (dead & alive) and annual cycles was designed to capture the long-term consequences of sepsis survivors from the initial hospitalisation period. Even though the relative treatment effect was only applied in the short-term model, the long-term model allowed the lifetime costs and QALY consequences associated with mortality differences captured in the short-term model to be accumulated. Long-term survival estimates were based on a UK multicenter observational study of sepsis survivors (Cuthbertson et al., 2010). In Soares et al., 2012, several parametric survival models were fitted, assuming that the long-term mortality of a patient with sepsis cannot become lower than that of the general population at any point (see Figure C.1.1). This work considers only a Weibull model that controls for age at admission as this model provided the best fit in the original analysis.

Parameters where inserted into the model as probability distributions in order to appropriately reflect parameter uncertainty. Probabilistic Sensitivity Analysis[1] (PSA) was then used to jointly propagate uncertainty in the model inputs through the decision model and allow estimation of the uncertainty surrounding the results. A list of the most important inputs that where used to populate the model is provided in Table 6.1, but for a more detailed description the reader is referred to the original HTA report (Soares et al., 2012, chap. 4). In this work the decision model was re-evaluated for each of the relative effects estimated under each ISM of the four base-models included in the final list (see Table 5.11).

[1] 5000 Monte Carlo Samples.

Table 6.1: *Key decision-model inputs.*

Cohort characteristics		
Parameter	Value	Source
Mean age of severe sepsis patients at admission	63	ICNARC CMPD (Harrison et al., 2006)
Proportion of males	53%	ICNARC CMPD (Harrison et al., 2006)
Short-term outcomes		
Baseline-risk i.e. probability of dying in the hospital under treatment with SoC	40.6%	ICNARC CMPD Harrison et al., 2006
Odds-ratio for IVIG/IVIGAM compared with ALB (T3b) or ALB/PLA (T2)	All sharing method-specific estimates as presented in Table 5.12, Table B.5.1, Table B.5.2	Synthesis of direct and indirect evidence of Chapter 5 of this work
Long-term outcomes		
Probability of death conditional on survival	Time-dependent. Based on the Weibul model of Figure C.1.1	Cuthbertson et al., 2010
Costs		
Cost of SoC	£0.00	Assumed to be 0 as they are applied in the IVIG/IVIGAM arm as well
Cost of IVIG	£5,539.05	British National Formulary (BNF)
Cost of 1 day in ICU (for a severe sepsis patient)	£1,393.00	NHS reference costs (Department of Health)
Cost of 1 day in ward (for a severe sepsis patient)	£196.00	NHS reference costs (Department of Health)
Utilities		
HRQoL weight (for an in-hospital severe sepsis patient)	0.53	Drabinski et al., 2001
Other		
Incidence of severe sepsis	$66 \cdot 10^{-5}$	ICNARC CMPD (Harrison et al., 2006)

Table adopted from Soares et al., 2014b.

6.2.2. Policy-related outcome measures

To understand the policy-implications of the use of different ISMs, the following measures, which are commonly used to inform adoption decisions and further research prioritisation decisions, are used:

1. Incremental Cost-Effectiveness Ratio

   As explained in subsection 2.2.3 (Equation 2.14), the ICER provides an estimate of the value-for-money of the new intervention. Here, it represents the additional cost per QALY gained under the IVIG/IVIGAM strategy compared to ALB, and is using the mean costs and benefits of each intervention, across the 5000 PSA iterations.

2. Decision

   The final adoption decision is based on a simple decision rule that uses a cost-effectiveness threshold (cut-off value). This threshold (k) represents an expectation of the amount of resources that displace a single QALY elsewhere in the health care system. For example, an explicitly estimated threshold of $k = 30,000$ £/QALY means that the maximum price that the health care system should pay for a technology that offers one QALY, without displacing more health than the health gained, is £30,000 (i.e. this is the marginal productivity of the health care system). Here, both £20,000 and £30,000 thresholds were used to represent the threshold range that is adopted by NICE (Appleby et al., 2007). If the estimated ICER is below £$20,000$ the technology is considered cost-effective and therefore is assumed to be approved, while if the ICER is above £$30,000$ it is not cost-effective and therefore rejected. Finally, if the ICER falls between £$20,000$ and £$30,000$ the technology is considered borderline cost-effective.

3. Probability of IVIG/IVIGAM being cost-effective

   This is calculated here by counting the number of PSA iterations in which the Net Monetary Benefit of the IVIG/IVIGAM strategy (calculated with $k =$ £$30,000$) is larger than the Net Benefit of ALB, and then dividing that by 5000 (i.e. the total number of simulations 5000).

4. Population Expected Value of Perfect Information (EVPI)

   To estimate the EVPI for the whole adult population who may be affected by severe sepsis during the lifetime of the technology (assumed to be 10 years), we need to appropriately discount future costs and consequences, as explained in Section 2.2.5. This is done in the following manner:

$$Pop.EVPI = EVPI \sum_{t=1}^{T} \frac{I_t}{(1+r)^t} \tag{6.1}$$

   where T is the lifetime of the technology, I_t the incidence rate at year t (Table 6.1) and r the discount rate (3.5%). The 2009 estimate of the UK adult population (aged 16 years or over) was used (50,243,000 people) and an incidence of 66/100,000 people/year totalling 33,160 patients/year. Population EVPI calculations assumed that information generated by further research will be used to inform reimbursement decisions over a period of 10 years and will not be relevant thereafter.

5. Expected Value of Perfect Information of the relative effectiveness parameter

   Given that information-sharing here is only realised on the relative effectiveness parameter, we are primarily considered with the impact of using different methods on the EVPPI of relative effectiveness, and hence no EVPPI estimates for other parameters were calculated. A detailed explanation of the EVPPI calculation process is provided in subsection C.1 on page 236.

6. Expected Value of Sample Information

   As explained in Section 2.2.5.3 the process of calculating EVSI is computationally intensive, because it contains two nested expectations (i.e. an integration within a maximisation). However, given that the relative effect is only applied on the short-term model, the net-benefits of the two treatments can be expressed as a multi-linear function of the model inputs. Hence, one loop can be avoided by estimating the net-benefits using the expectations of the various model inputs. Further, to calculate the expected log-odds ratio and then the expected probability of death in the short-term model, a Taylor series approximation with two terms was used. Finally, in combining the new data with the prior on the relative treatment effect, normal-normal closed form solutions were used to derive the parameters of the posterior distributions. These methods closely follow the directions described previously in the literature by Ades et al., 2004. For more details regarding the derivations of the aforementioned quantities in the decision-model, the reader is referred to Soares et al., 2012, (Chapter 6, Appendix 5).

6.3. Results

In this section, the results of the decision-model calculations are presented for all ISMs that were applicable to the two following base-models: 1. the Duration FE base-model and 2. the Jadad RE base-model. The results for the remaining base-models are shown in the Appendix in Table C.2.1 and Table C.2.2. For RE base-models, the predictive distribution is used, because it incorporates the uncertainty that is due to between-studies heterogeneity.

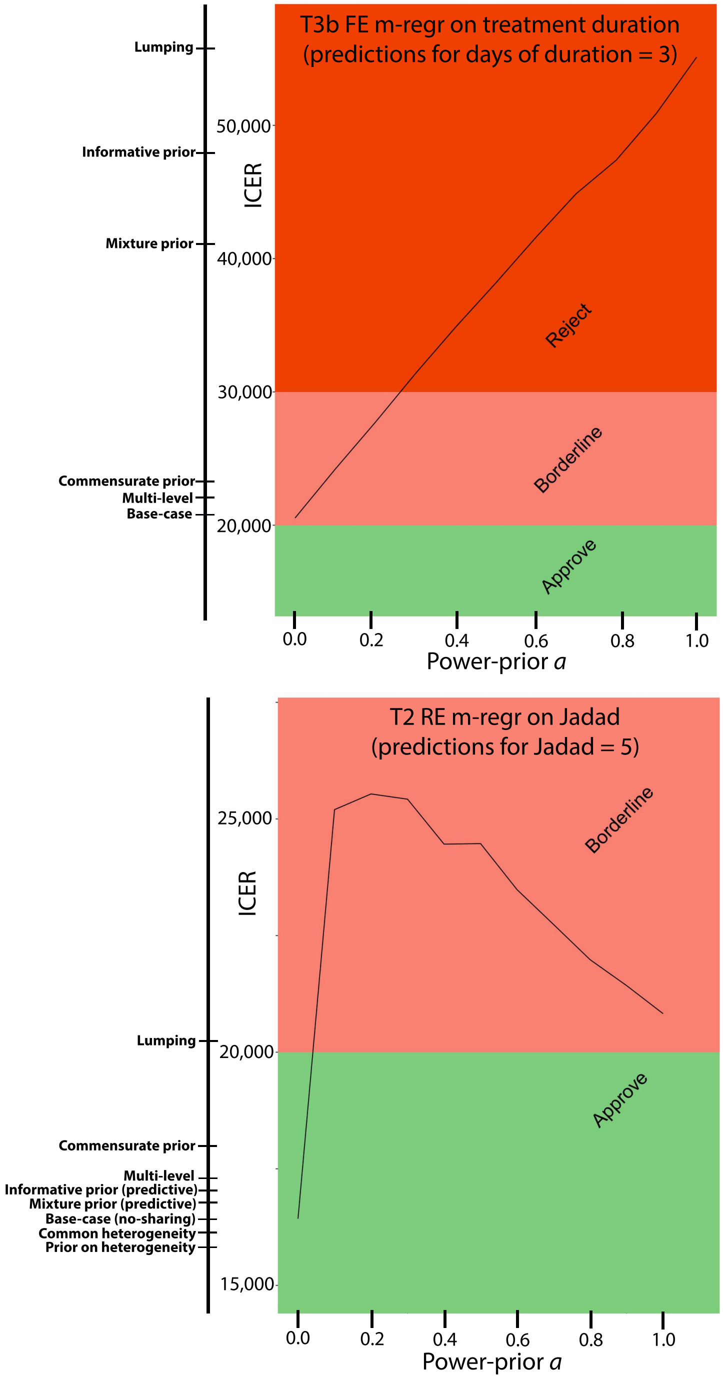
6.3.1. ICERs and decisions

Figure 6.2 illustrates the ICERs and decisions of all the applicable ISMs across the Duration FE (top) and the Jadad RE (bottom) base-models. The ICERs along with estimates of total costs and QALYs for IVIG/IVIGAM are also shown in Table 6.2. Note that differences across the ICERs of the various models are solely due to differences in the point estimate and are irrelevant to the methods' precision gains/losses. In the top graph, apart from the base-case (i.e. no-sharing/only adults), six ISMs were applied, whilst in the bottom graph eight ISMs were applied. Since the power-prior allows us to specify the power (α) that the likelihood of the indirect evidence is raised to, values of α between 0 and 1 were considered in increments of 0.1. The graphs' lines correspond to the ICERs (y-axis) that result from the use of power-prior models with various α weights (x-axis). The ICERs of the remaining ISMs are shown next to the main y-axis.

Regarding the Duration FE base-model, ICERs vary considerably from £20,542 per QALY gained under the base-case (no sharing/adults only) to £55,316 per QALY gained under lumping. ISMs that impose more moderate sharing assumptions (Multi-level models, Commensurate priors) produce ICERs closer to the base-case. This is also observed in the Power-prior models where ICERs increase with α and become identical to lumping for $\alpha = 1$. Given the variation in the ICER values, the fact that different ISMs lead to potentially different decisions is not surprising.

The results of the Jadad RE base-model paint a similar picture. However, even though more ISMs were applicable, ICERs varied less (from £15,900 per QALY gained when a prior is imposed on heterogeneity to £25,530 per QALY gained under the Power-prior with $\alpha = 0.2$). This might seem unexpected since the ranges of the estimated mean RTEs across ISMs are similar for the two base-models (Duration FE base-model [-0.31,-0.08], Jadad RE base-model [-0.55, -0.25]) according to Table 5.12 on page 120. However, under the Jadad RE base-model, estimates suggest larger relative effects. This implies that the probability of death under the new treatment is lower and the QALYs gained under IVIG/IVIGAM are overall higher. As a result, the denominator of the ICER (i.e. the QALYs) is larger and the resulting ICERs less variable.

Figure 6.2: *ICERs and Decisions for all applicable ISMs across the Duration FE and Jadad RE base-models.*



The results of the application of power-prior models are shown in the plotted line for varying values of α (x-axis). Results for other ISMs are marked in the y-axis.

Interestingly, in the Jadad RE base-model, some power-prior models result in ICERs that are greater than under lumping. This is a consequence of the very big paediatric study that showed no effect (see page 235). Furthermore, the Jadad RE base-model does not produce the same ICER value with the power-prior model where $\alpha = 1$. This is because the power-prior model only shares information on relative effectiveness, whilst the lumping model imposes a random-effect across all studies, effectively sharing information on relative effects, heterogeneity and imposing exchangeability across all studies included in the direct and the indirect evidence base. Hence, it imposes stronger information-sharing that the power-prior model with $\alpha = 1$, which is why the graph implies that lumping could map to a power-prior model with $\alpha > 1$. Lastly, the decision again varies among ISMs, although the results now suggest that IVIG/IVIGAM is either clearly cost-effective, or borderline cost-effective.

Table 6.2: *ICERs, total costs, and total QALYs for all applicable ISMs.*

T3b FE Meta-Regression (on duration of treatment)				T2 RE Meta-Regression (on Jadad score)			
ISM	ICER	Tot. C	Tot.Q	ISM	ICER	Tot.C	Tot.Q
Base-case	20,542	54,996	4.36	Prior on heterogeneity	15,910	57,413	4.65
Power-prior $\alpha = 0$	20,539	54,994	4.36	Common heterogeneity	16,216	57,169	4.62
Multi-level	21,392	54,744	4.33	Base-case	16,430	57,012	4.60
Commensurate prior	21,954	54,593	4.31	Power-prior $\alpha = 0$	16,458	57,084	4.59
Power-prior $\alpha = 0.1$	24,204	54,102	4.26	Multi-level	17,131	56,546	4.54
Power-prior $\alpha = 0.2$	27,723	53,562	4.19	Commensurate prior	18,066	56,029	4.48
Power-prior $\alpha = 0.3$	31,381	53,175	4.15	Mixture prior	19,071	55,574	4.43
Power-prior $\alpha = 0.4$	34,863	52,907	4.11	Informative prior	19,272	55,493	4.42
Power-prior $\alpha = 0.5$	38,176	52,710	4.09	Lumping	20,110	55,183	4.38
Mixture prior	40,809	52,586	4.08	Power-prior $\alpha = 1$	20,826	54,955	4.35
Power-prior $\alpha = 0.6$	41,603	52,548	4.07	Power-prior $\alpha = 0.9$	21,428	54,781	4.33
Power-prior $\alpha = 0.7$	44,871	52,421	4.06	Power-prior $\alpha = 0.8$	21,983	54,634	4.32
Informative prior	47,365	52,337	4.05	Power-prior $\alpha = 0.7$	22,737	54,454	4.29
Power-prior $\alpha = 0.8$	47,393	52,338	4.05	Power-prior $\alpha = 0.6$	23,484	54,292	4.27
Power-prior $\alpha = 0.9$	50,907	52,238	4.03	Power-prior $\alpha = 0.5$	24,460	54,104	4.25
Power-prior $\alpha = 1$	54,116	52,161	4.03	Power-prior $\alpha = 0.4$	24,472	54,101	4.25
Lumping	56,314	52,114	4.02	Power-prior $\alpha = 0.1$	25,200	53,976	4.24
				Power-prior $\alpha = 0.3$	25,424	53,939	4.23
				Power-prior $\alpha = 0.2$	25,536	53,921	4.23

ISM: Information-sharing method, ICER: Incremental cost-effectiveness ratio (in £/QALY), Tot.C: Total costs (in £), Tot.Q: Total QALYs. ISMs are ranked in ascending ICER values. The informative and mixture priors under the random-effect base-model use the predictive distribution of the indirect evidence. Base-case (no sharing/adults only) results are shaded in red for comparison purposes. ISM: Information-sharing method.

6.3.2. Probability of being cost-effective and EVPI

Table 6.3 ranks, in ascending order, the ISMs for each base-model according to the probabilities that IVIG/IVIGAM is cost-effective compared to ALB at a threshold of £30,000 per QALY gained and also displays their corresponding population EVPI estimates.

Table 6.3: *Probability of IVIG/IVIGAM being cost-effective (pCE) and population EVPI at 10 years.*

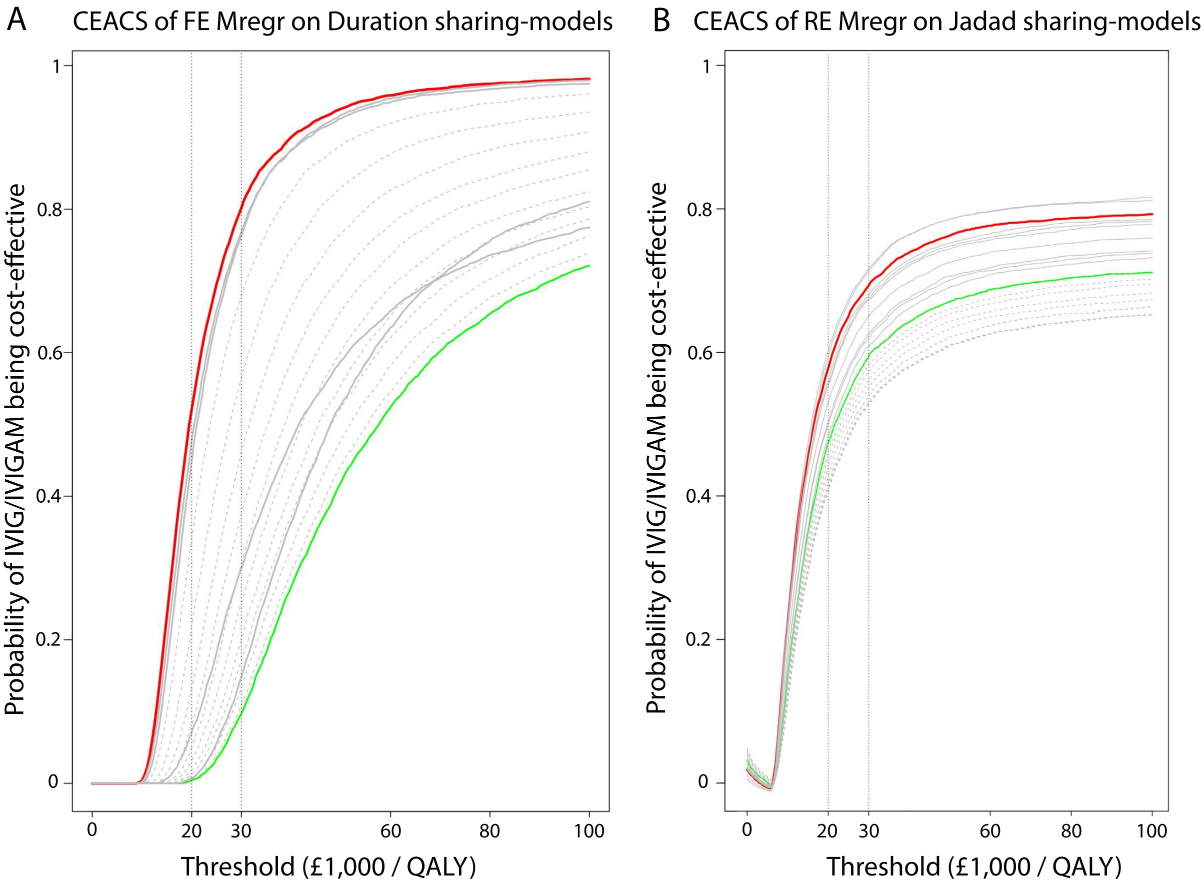
T3b FE Meta-Regression (on duration of treatment)			T2 RE Meta-Regression (on Jadad score)		
ISM	pCE	Pop.EVPI	ISM	pCE	Pop.EVPI
Lumping	0.10	33	Power-Pr. $\alpha = 0.2$	0.54	1709
Power-Pr. $\alpha = 1$	0.11	39	Power-Pr. $\alpha = 0.3$	0.54	1714
Power-Pr. $\alpha = 0.9$	0.13	51	Power-Pr. $\alpha = 0.1$	0.54	1732
Informative prior	0.15	55	Power-Pr. $\alpha = 0.4$	0.55	1642
Power-Pr. $\alpha = 0.8$	0.17	71	Power-Pr. $\alpha = 0.5$	0.55	1644
Power-Pr. $\alpha = 0.7$	0.20	90	Power-Pr. $\alpha = 0.6$	0.56	1584
Power-Pr. $\alpha = 0.6$	0.25	128	Power-Pr. $\alpha = 0.7$	0.57	1542
Mixture prior	0.30	209	Power-Pr. $\alpha = 0.8$	0.58	1480
Power-Pr. $\alpha = 0.5$	0.31	175	Power-Pr. $\alpha = 0.9$	0.59	1459
Power-Pr. $\alpha = 0.4$	0.37	252	Power-Pr. $\alpha = 1$	0.60	1419
Power-Pr. $\alpha = 0.3$	0.46	366	Lumping	0.62	1273
Power-Pr. $\alpha = 0.2$	0.57	373	Commensurate prior	0.66	1213
Power-Pr. $\alpha = 0.1$	0.68	286	Multi-level	0.68	1155
Commensurate prior	0.76	220	Informative prior	0.68	1125
Multi-level	0.77	226	Mixture prior	0.69	1112
Base-case	0.80	207	Base-case	0.70	1091
Power-Pr. $\alpha = 0$	0.80	206	Power-Pr. $\alpha = 0$	0.70	1091
			Common heterogeneity	0.72	908
			Prior on heterogeneity	0.72	955

All calculations assume a threshold of £30,000 per QALY gained and are rounded to the nearest million £. ISMs are ranked in ascending order of the pCE. The informative and mixture priors under the random-effect base-model use the predictive distribution of the indirect evidence. Base-case (no sharing/adults only) results are shaded in red for comparison purposes. ISM: Information-sharing method.

The choice of the ISM seems to significantly impact probabilities, however the results seem to be much less variant under the Jadad RE base-model than the Duration FE base-model. ISMs that impose stronger sharing assumptions do not necessarily result in lower probabilities as evidenced by the power-priors in the Jadad RE base-model. Under the Jadad RE base-model, ISMs that suggest a higher probability of IVIG/IVIGAM being cost-effective also see lower population EVPI estimates. In contrast, under the FE Duration base-model, the relationship between the probability of IVIG/IVIGAM being cost-effective and population EVPI is not monotonous. This is because all ISMs under the Jadad RE base-model suggest that IVIG/IVIGAM should be adopted , and pCE is consistently above 0.5, whilst under the Duration FE base-model the decision is not consistent across all ISMs , with pCE varying between 0.1 and 0.8.

Figure 6.3 depicts how the probabilities change as the threshold increases up to £100,000 per QALY gained. The red and the green lines correspond to the base-case (no sharing/only adults) and the lumping case respectively. Dotted grey lines correspond to power-prior models with varying α weights, and solid grey lines to the remaining non-power-prior ISMs.

Figure 6.3: *Cost-effectiveness acceptability curves of all the applicable ISMs in the FE meta-regression on treatment duration and the RE meta-regression on Jadad score.*



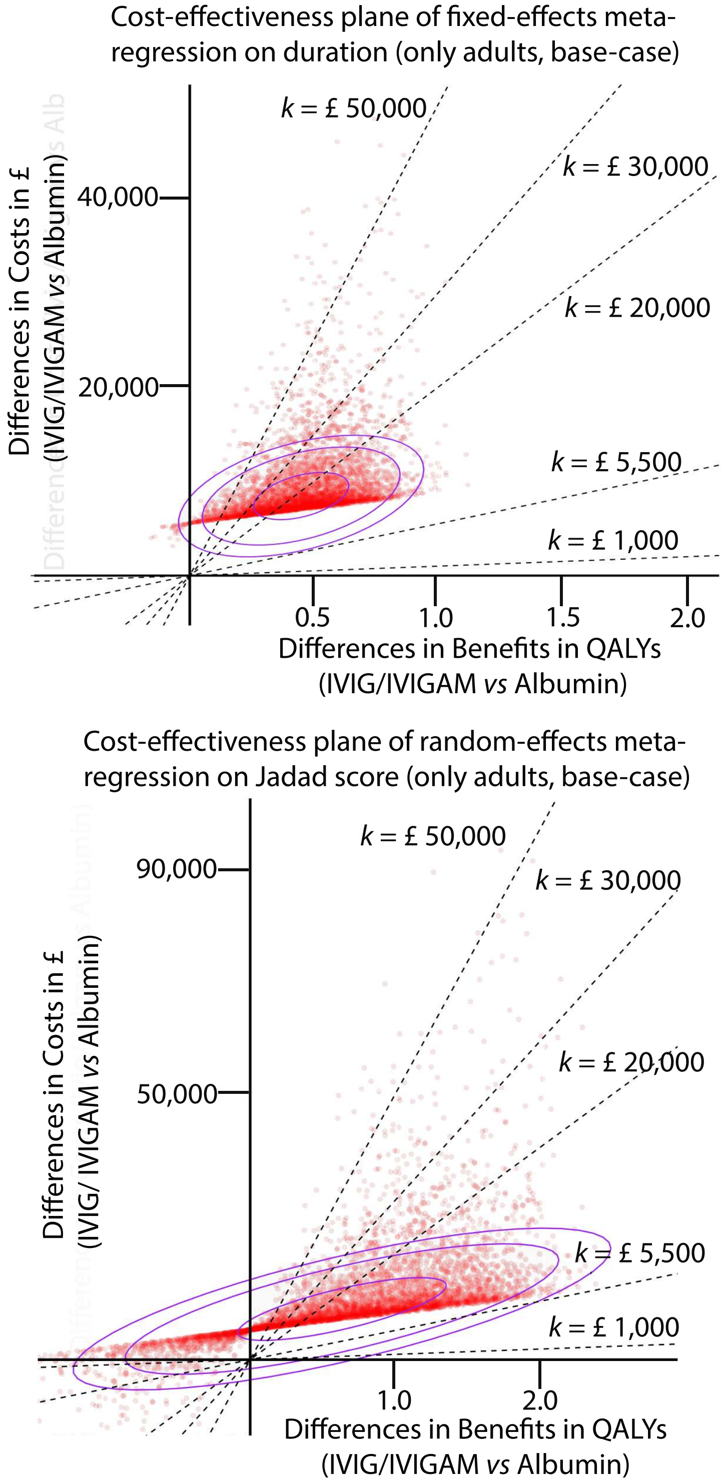
Red lines correspond to the base-case (no sharing/only adults); green to lumping; dotted grey lines to the power-prior models with varying α weights; solid grey lines to the remaining ISMs.

In the FE base-model (Figure 6.3A), as α increases CEACs move from the no sharing/ only adults (red) case to lumping (green). The two ISMs that are very close to the no sharing/only adults (red) CEAC are the multi-level model and the commensurate prior, whilst the two lines that are close to lumping (green) are the informative prior and the mixture prior. The CEACs of the RE model (Figure 6.3B) are somewhat more difficult to interpret. Firstly, two models produce CEACs which fall above the space that is defined by no sharing/only adults (red) and lumping (green). These are the two ISMs that do not share information directly on the relative effectiveness parameter but only on heterogeneity (i.e. common heterogeneity, and prior on heterogeneity). Hence, there is no particular reason to expect them to fall between the no sharing/only adults (red) and the lumping (green) models that share information only on relative effectiveness.

Furthermore, power-prior models produce CEACs which are below lumping (green) (i.e. information-sharing is stronger than lumping). These correspond to the power-prior models with low values of α, which disproportionately assign a high weight to the large Brocklehurst et al., 2011 study (which showed no effect of IVIG in paediatric patients) – see explanations on page 121 and page 235 for more details. The larger that α is, the closer the CEAC is to lumping (green). Finally, the CEAC shows a kink point that is observed for low threshold values close to £10,000 per QALY. This is attributed to the fact that for the Jadad RE base-model the joint distribution of the differences in costs and effects falls in all four quadrants (see bottom of Figure 6.4). North-east (NE) and south-west (SW) quadrants reflect potentially cost-effective probability mass (Fenwick et al., 2004). As the threshold increases from $k = £0$ / QALY to around $£5,500$ / QALY, cost-effective probability mass is lost from the SW quadrant without any gains in the NE. However, for larger threshold values, large amounts of cost-effective mass is gained very quickly in the NE rendering the losses in SW negligible.

Regarding EVPI, Table 6.3 shows the population EVPI estimates assuming a 10 year intervention lifetime. For the Duration FE base-model, models that impose stronger information-sharing result in lower EVPI estimates. This is because as more strength is borrowed, we estimate a lower probability of IVIG/IVIGAM being cost-effective, implying that it becomes increasingly more certain that IVIG/IVIGAM should be rejected.

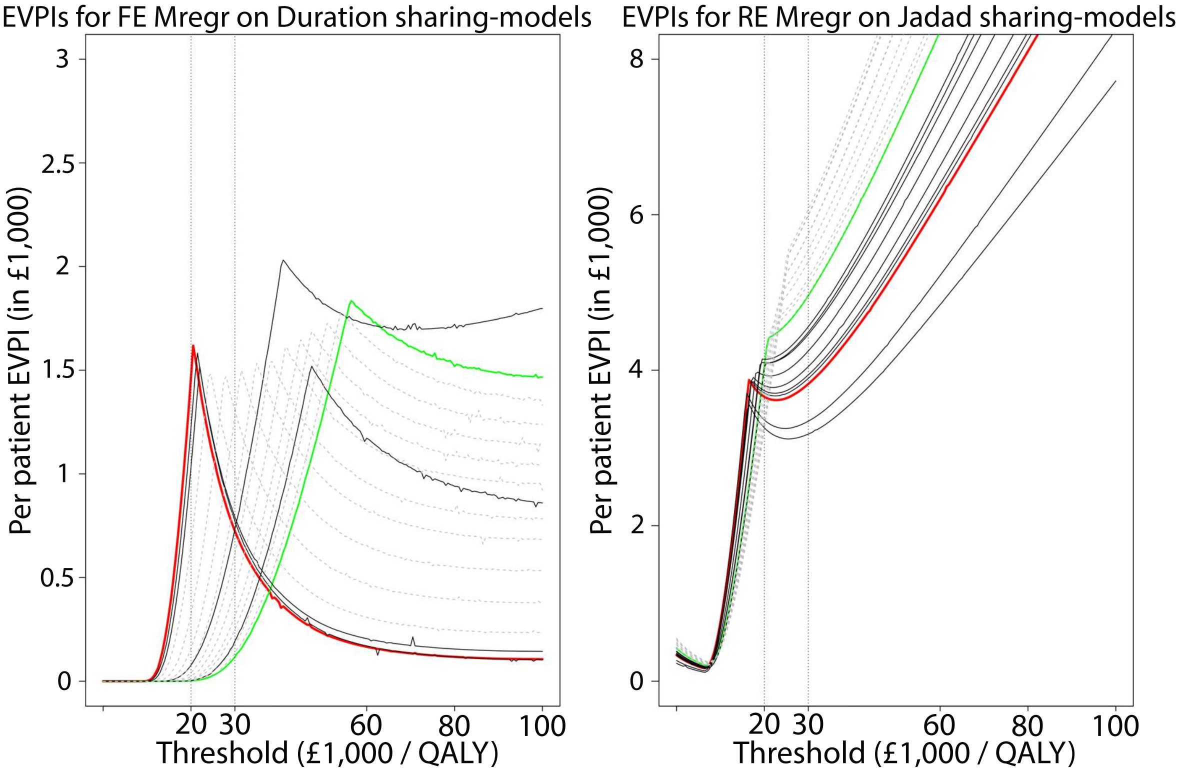
Figure 6.4: *Cost-effectiveness plane of the FE meta-regression on treatment duration FE and the Jadad RE base-models in the no sharing/only adults case (i.e. no information-sharing).*



Cost-effectiveness plane of fixed-effects meta-regression on duration (only adults, base-case)

Cost-effectiveness plane of random-effects meta-regression on Jadad score (only adults, base-case)

Figure 6.5: *Per person EVPIs of the various ISMs for the FE meta-regression on duration and the RE meta-regression on Jadad score base-models.*



Red lines correspond to the base-case (no sharing/only adults); green to lumping; dotted grey lines to power-prior models; black lines to the remaining ISMs.

In contrast to the Duration FE base-model, under the Jadad RE base-model, population EVPI estimates are more similar across ISMs with almost all models leading to EVPI estimates above £1 billion. This is because the CEACs of the Jadad RE base-model plateau around a probability of 0.8 (Figure 6.3B), implying that even for very high thresholds the model assigns a 20% probability of IVIG/IVIGAM being less cost-effective than ALB. This is expected given that Figure 5.6 B2 on page 119 showed that the log-odds ratio estimates span the region both below and above zero and there is always a chance that IVIG/IVIGAM is harmful compared to ALB. As a result, when the threshold increases, health is valued more, and the consequences of clinical uncertainty soar, producing EVPI curves that increase indefinitely (Figure 6.5 B).

Overall, it seems that the choice of ISM is not only strongly influencing the probabilities of a treatment strategy being cost-effective (with some ranging between 0.1 and 0.8 among different models), but is also significantly impacting estimates of the value of further research.

6.3.3. EVPPI and EVSI

In this section the focal parameter is the EVPPI of the relative effect, which is the main parameter we would gain information on if we prioritised an RCT that enrolled adult patients into two arms: one receiving IVIG/IVIGAM and one ALB.

The EVPPI estimates for the ISMs implemented under the Duration FE base-model are shown in Figure 6.6. The black line shows the EVPPI estimates from power-prior models with their corresponding α shown on the x-axis. The estimated EVPPIs using the remaining ISMs are also displayed on the y-axis. The red dashed line depicts the cost of a randomised clinical trial[2] of the optimal sample size (i.e. the sample size which maximises the expected net-benefit of the sample) for the corresponding power-prior model. The vertical distance between the red line and the black line corresponds to the maximum ENBS of the power-prior model that uses the α values on the x-axis. This is simply the difference between EVSI and the cost of the trial of the optimal sample size for a power-prior model with given α. The maximum ENBS and the optimal sample of all the ISMs is also listed in Table 6.4 in ascending optimal sample size order. Among ISMs suggesting that an RCT should be prioritised (i.e. $max.ENBS > 0$), the optimal sample sizes range between 1940 and 3400. Power-prior models with $\alpha > 0.7$ suggest that the cost of a new trial of any sample size is greater than the consequences of the existing uncertainty, suggesting that a trial should not be prioritised.

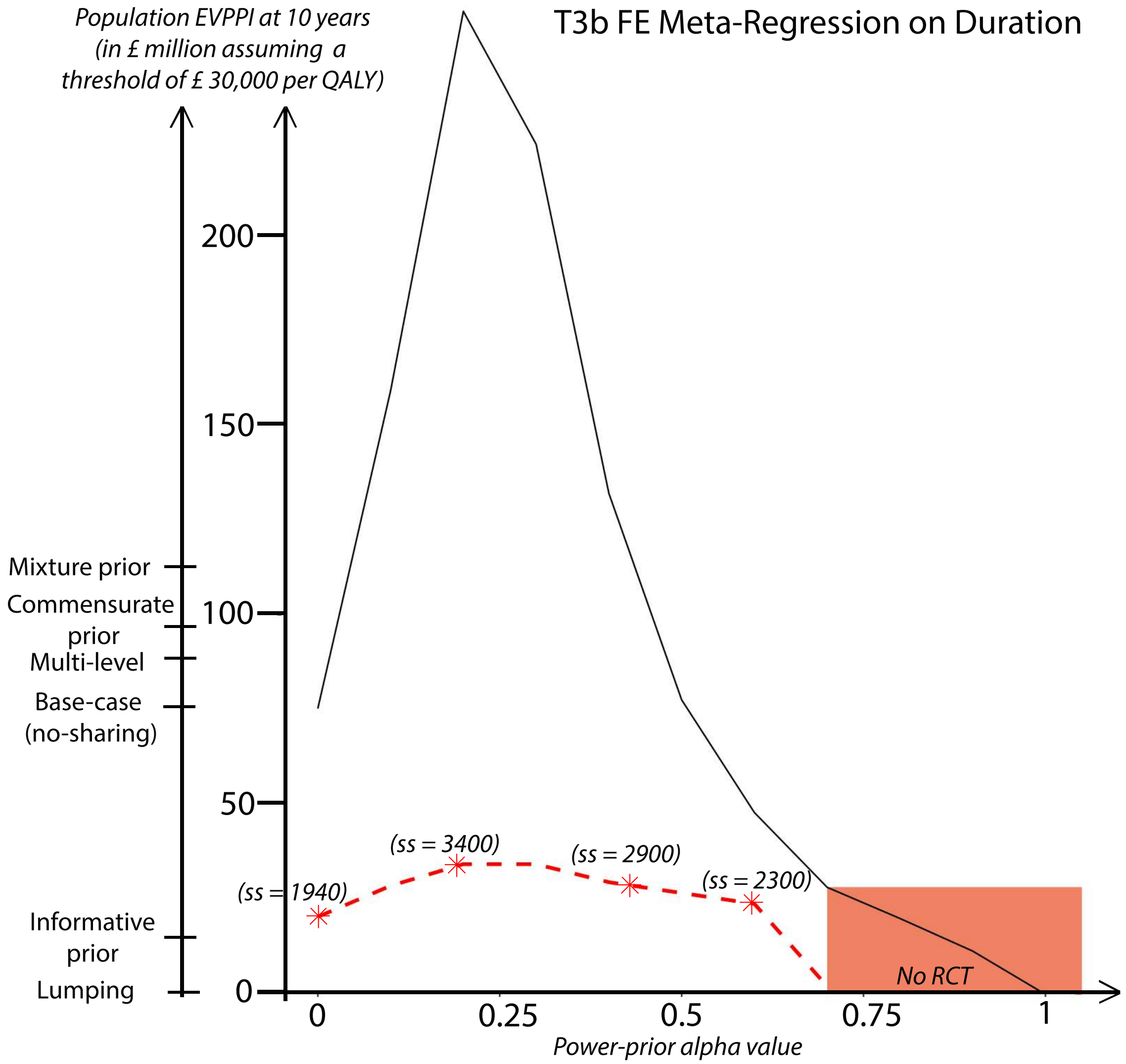
Contrary to the Duration FE base-model, under the Jadad RE base-model, all ISMs estimate that there is considerable value in prioritising a new trial in adults, with EVPPI estimates reaching well above £1 billion (Figure 6.7); effectively, rendering the costs[3] of that trial negligible. The estimated optimal arm sample sizes are more homogeneous than under the Duration FE base-model (Table 6.4), with estimates ranging between 1040 and 1500 per patient per arm. Contrary to the Duration FE base-model, under the Jadad RE base-model we see that ISMs imposing stronger assumptions suggest that trials of larger sizes should be conducted. This implies that in the former, as more strength is borrowed from the paediatric evidence we become increasingly confident about what the cost-effective strategy is, whilst in the latter significant uncertainty remains regardless of the imposed strength of sharing. As a result the cost of adding patients to the trial is overcome by the value of the potential uncertainty consequences.

In conclusion, the impact of different ISMs on further research prioritisation decisions, and on the design of a future trial again seems to be substantial, rendering the considerations and transparency surrounding the choice of ISM very important.

[2] Trial costs assume a fixed £2 million cost for the trial and a further £2,000 per patient enrolled. Treating with IVIG/IVIGAM incurs an additional cost of £5,500 per patient.

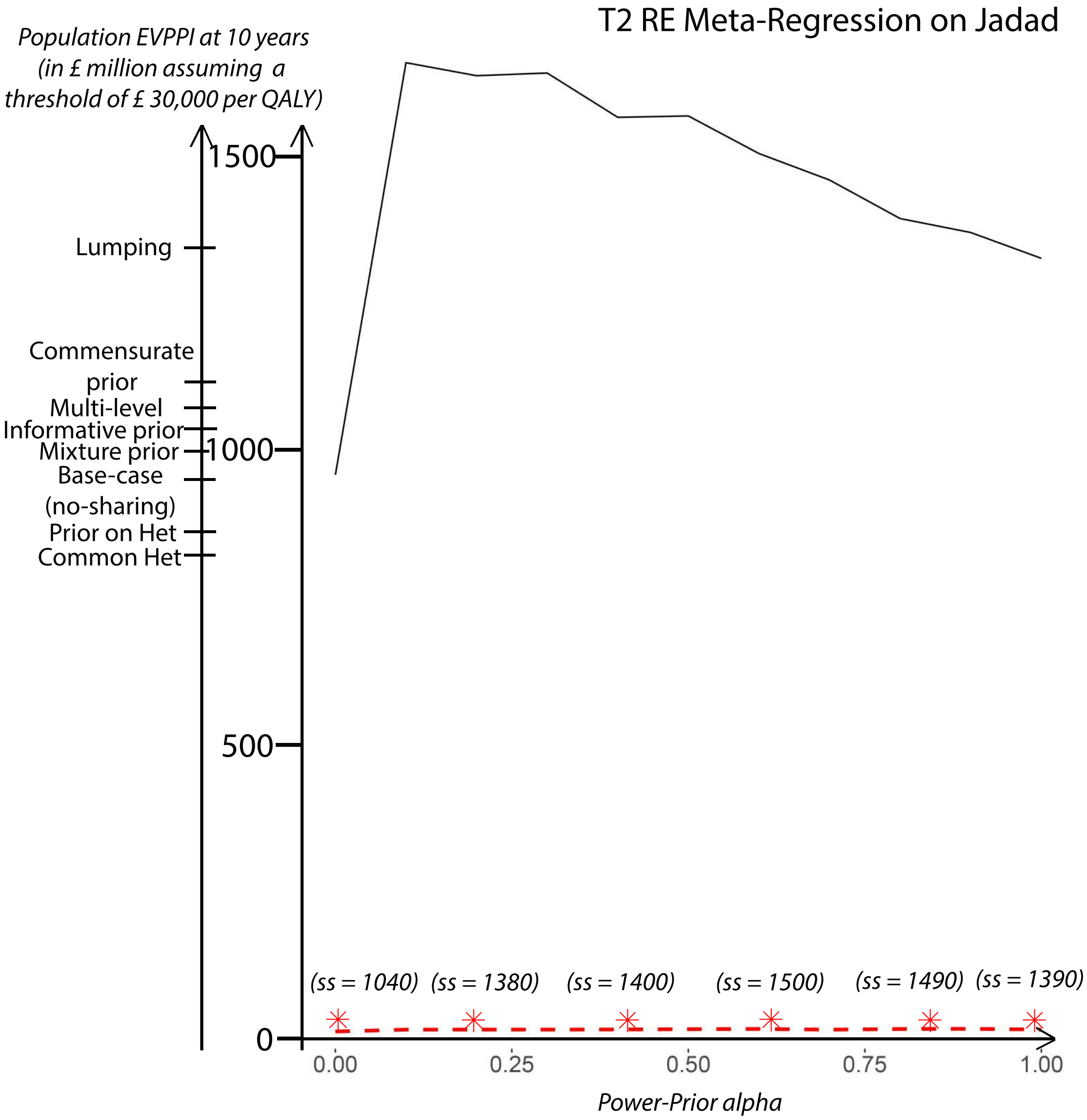
[3] A two-arm trial enrolling 1000 patients per arm is assumed to cost around £15 million.

Figure 6.6: *Population EVPPI at 10 years for the Duration FE base-model.*



Calculations assume $k = £30,000$. The black line corresponds to the EVPPI estimates (y-axis) of the Power-prior models that use the α weight in the x-axis. The red line is the cost of a trial of the optimal sample size (shown in parenthesis) for power-prior models of every alpha. Hence, the distance between the black and the red line corresponds to the ENBS. The EVPPI of the non-power-prior models is shown on the y-axis.

Figure 6.7: *Population EVPPI at 10 years for the Jadad RE base-model.*



Calculations assume $k = £30,000$. The black line corresponds to the EVPPI estimates (y-axis) of the Power-prior models that use the α weight in the x-axis. The red line is the cost of a trial of the optimal sample size (shown in parenthesis) for power-prior models of every alpha. The EVPPI of the non-power-prior models is shown on the y-axis. The informative and mixture priors use the posterior predictive distribution of the indirect evidence.

Table 6.4: *ISMs for each base-model ranked according to optimal sample size in ascending order.*

T3b FE Meta-Regression (on duration)			T2 RE Meta-Regression (on Jadad score)		
ISM	Optimal Sample	max. ENBS	ISM	Optimal Sample	max. ENBS
Lumping	No-RCT	<0	Base-case	1040	936
Informative prior	No-RCT	<0	Pr-Power $\alpha = 0$	1040	936
Pr-Power $\alpha = 0.7$	No-RCT	<0	Common Heterogeneity	1120	751
Pr-Power $\alpha = 0.8$	No-RCT	<0	Multi-level	1190	1008
Pr-Power $\alpha = 0.9$	No-RCT	<0	Mixture prior	1270	961
Pr-Power $\alpha = 1$	No-RCT	<0	Prior on Heterogeneity	1300	796
Base-case	1940	41.3	Informative prior	1350	976
Pr-Power $\alpha = 0$	1940	41.3	Pr-Power $\alpha = 0.1$	1370	1638
Multi-level	2150	60	Pr-Power $\alpha = 0.7$	1370	1435
Commensurate prior	2250	51.6	Pr-Power $\alpha = 0.2$	1380	1616
Pr-Power $\alpha = 0.6$	2300	9.64	Pr-Power $\alpha = 0.3$	1380	1620
Pr-Power $\alpha = 0.5$	2600	36.2	Pr-Power $\alpha = 0.4$	1400	1544
Mixture prior	2750	70.9	Pr-Power $\alpha = 0.5$	1460	1546
Pr-Power $\alpha = 0.1$	2800	118	Pr-Power $\alpha = 1$	1480	1132
Pr-Power $\alpha = 0.4$	2900	89.2	Lumping	1490	1150
Pr-Power $\alpha = 0.2$	3400	213	Commensurate prior	1490	1074
Pr-Power $\alpha = 0.3$	3400	179	Pr-Power $\alpha = 0.8$	1490	1369
			Pr-Power $\alpha = 0.6$	1500	1481
			Pr-Power $\alpha = 0.9$	1520	1345

Maximum ENBS, in millions Pounds Sterling (£) is displayed for each method. All calculations assume $k = 30,000$ £. The informative and mixture priors under the random-effect base-model use the predictive distribution of the indirect evidence. Base-case (no sharing/adults only) results are shaded in red for comparison purposes. ISM: Information-sharing method.

6.4. Discussion

This chapter explored the impact of the choice of ISM on decisions that relate to both the implementation of a technology and prioritisation of further research. Relative effectiveness was estimated using different ISMs described in Chapter 5 and applied here in a decision-model developed by Soares et al., 2012. Commonly used quantities for policy-making such as ICERs, the probability that the new intervention is cost-effective, and the optimal sample size of a future randomised trial that would seek to gain more information on the relative effectiveness parameter were then calculated. This work is the first to investigate the impact of a set of systematically identified ISMs on policy decisions.

Irrespective of policy measure, the findings consistently suggest that the choice of ISM can impact decisions. This implies that using only a subset of ISMs does not adequately capture the potential impact of using indirect evidence. Hence, it is vital that applicable ISMs are systematically identified. In the IVIG/IVIGAM case, the indirect evidence was sourced from a population (paediatric patients) on which the evidence suggested that IVIG/IVIGAM is less effective than in the directly relevant population (adults). Therefore, the inclusion of the paediatric evidence makes IVIG/IVIGAM less cost-effective. However, as seen for the Jadad RE base-model, we should not necessarily infer that the stronger information is borrowed from the indirect evidence, the higher the ICER. Instead, methods that impose moderate information-sharing can lead to relative effect estimates that are beyond the range defined by lumping and splitting, leading to ICER estimates outside of this range. It should also be noted that the impact of the ISM is very much dependent on the base-model which may imply that the impact of using different methods is context-specific and general conclusions cannot be drawn with respect to how different methods affect decision-making.

In terms of the power-prior models, it seems that the characteristics of the indirect studies should be closely considered when interpreting the results. Specifically, when the indirect evidence base consists of (one or more) large studies that present a more extreme result than the whole body of evidence together, interpretation can be challenging. If that is the case, at least under RE models, the analyst should expect larger studies to influence the overall results more heavily for low α values, and potentially yield estimates that fall beyond the range that is defined by splitting and lumping.

This case-study found that the implications of ISM choice on VoI and further research prioritisation can be substantial. Under the Duration FE base-model, some methods suggested that there is no value in a future RCT that would seek to resolve uncertainty in relative effectiveness, whilst other ISMs suggested that a trial can resolve uncertainty worth more than £10 million. The inconsistency among methods suggestions implies that it is paramount to ensure that all applicable ISMs have been implemented, and that decision-makers scrutinise the assumptions underpinning each method.

An important issue, which is not addressed here, relates to the appropriate calculation of the value of information when the evidence that is included in the analysis pertains to more than one decision. Here, although transferability of the evidence is assumed between adults and paediatric patients, the value of information calculations do not reflect how further research in adults may affect decisions in paediatric patients; instead, population EVPI only considers adult patients. However, since VoI seeks to estimate the value of resolving existing uncertainty in treatment choice, and further evidence in adults may also resolve uncertainty for the cost-effectiveness of treatments for paediatric patients, VoI calculations should, in principle, reflect the total benefits of collecting data. That is, they should take account of the contribution of the collected data to all decisions they will affect, directly or indirectly. In the simplest case, if information was completely transferable across the two populations the upper bound for future research could be simply calculated as the sum of the population-specific EVPPI of the subset of parameters for which uncertainty will be resolved by the research. However, information is likely to be only partially transferable and therefore appropriate VoI calculations would be more complicated. Further research could try to show how information-sharing considerations could be appropriately incorporated in VoI calculations. Such considerations are not only relevant for decisions that pertain to indirect populations/subgroups, but may also be relevant for decisions that consider the use of an intervention that belongs to the same class as the one assessed in the trial, and even to decisions made in other countries/jurisdictions (Woods et al., 2018).

Whilst the choice of ISM appears to have significant implications for policy-making in this case-study, the findings of this chapter are not adequate to result in more general conclusions. This is due to the fact that, whether policy recommendations are impacted by the inclusion of indirectly related evidence or not, relies on the characteristics of both the direct and indirect evidence. In particular, characteristics such as the number of direct and indirect studies, the total number of patients in the direct and indirect evidence base, the between-studies heterogeneity of each evidence set, and the disagreement in their summary estimate, can play a significant role in determining both the significance of method choice and how strongly alternative methods share information. These dimensions will be further explored in the next chapter with the purpose of drawing more generalisable conclusions about comparability of the alternative ISMs.

Chapter 7

Comparing information-sharing methods: a simulation

7.1. Chapter aims and structure

The results of previous chapters illustrated that when strength is borrowed from indirect evidence the choice of ISM is crucial. In Chapter 5, the combination of direct and indirect evidence under different methods led to significantly different RTE estimates. When these were used in a decision-model to inform a policy question (Chapter 6), they resulted in discrepant suggestions regarding the adoption of the technologies under consideration and the value of further research. Hence, ISMs that impose different assumptions and consequently borrow more or less strength from the indirect evidence can influence RTE estimates and policy recommendations.

However, it is not always easy to judge how methods compare to one another in terms of the strength of sharing they impose. Sometimes it is clear how the assumptions underlying the different methods compare. For instance, lumping imposes a stronger assumption (direct and indirect evidence inform the exact same RTE parameter) —and will impose larger degrees of information-sharing —than methods that make more moderate assumptions (e.g. direct and indirect evidence inform different, yet exchangeable, RTE parameters). In other cases, the relationships are not as clear. For example, it is unclear how methods using commensurate priors compare to multi-level models. Furthermore, in many cases, the level of borrowing may depend on the features of the datasets.

This chapter aims to use simulated scenarios to address the following question:
How do the following features of the evidence influence how methods compare in terms of the strength of sharing they impose? 1. the difference in RTE estimate mean between direct and indirect evidence; 2. their differences in the between-studies heterogeneity; and 3. their difference in the number of patients included in each evidence set.

The remainder of this chapter is structured as follows. In Section 7.2 the methods used in the simulation and synthesis of datasets are explained, along with the approach taken in comparing the results. Subsequently, the results of the analysis are described separately for the FE and RE models. Finally, in Section 7.4 a synopsis of the most important findings is provided, the limitations of the experiment are discussed, and directions for further research are suggested.

7.2. Methods

The purpose of this work was to firstly investigate the degree of sharing imposed by the various ISMs, and secondly to lay the ground work for the stimulation of further research. In particular, this study aimed to determine how the different ISMs rank according to how strongly they borrow strength from the indirect evidence. This was considered by the means of a number of carefully devised scenarios. The scenarios were constructed to explore how particular characteristics of the indirect evidence influence methods' ranking.

Importantly, this simulation used probabilistic scenarios to allow the estimation of credible intervals around the strength of sharing measures. In this way, it was feasible to assess whether methods were consistently ranked in particular positions, and gain insight into the way that sampling variation affects how methods compare to one another. Also, the use of probabilistic scenarios will assist the generalisability of findings in the context of sampling uncertainty.

Most ISMs can be used under FE or RE models and the case-study of Chapter 5 showed that the ranking of methods may differ under FE and RE base-models. Therefore, two separate simulation studies are conducted here: one under a FE and one under a RE base-model, to allow potential differences to be identified.

In this simulation study, the scenarios considered a direct evidence set that is loosely based on the adult evidence of the case-study in Chapter 5 (see page 98). In determining the strength of information-sharing, it is critical how the evidence sets compare to one another. Hence, the relative difference between the two evidence sets is a more important consideration than the absolute values assumed for each evidence set's underlying parameters). Therefore, for convenience, here the characteristics of the direct data were set to be similar to those in the case-study presented in Chapter 5 and, importantly, the data generating model for the indirect evidence was defined relative to the characteristics of the direct evidence set. Each scenario varied a characteristic of the indirect evidence such as the extent of heterogeneity, the sample size, and the point estimate, —always in relation to the characteristics of the direct evidence set. It was further assumed that there were only two competing treatments (i.e. one control and one treatment arm) to eliminate information-sharing occurring indirectly via the consistency equations.

This section follows the *'Aims, Data-generating mechanisms, Estimands, Methods, Performance measures'* (ADEMP) structure for simulation experiments, as suggested by Morris et al., 2019.

7.2.1. Simulation aims

1. To investigate how ISMs compare to each other according to the degree of information-sharing that they impose under a set of pre-defined scenarios. These scenarios would vary the following dimensions:

   (a) The distance between the RTE means of the direct and the indirect evidence sets.

   (b) The difference across the between studies heterogeneity of the direct and the indirect evidence sets.

   (c) The difference between the sample sizes of the direct and the indirect evidence.

2. To generalise conclusions regarding the comparison of ISMs in the context of sampling uncertainty.

3. To understand the properties of the various ISMs and check models' stability across various evidential scenarios in order to identify potential circumstances under which model fitting and model convergence could become challenging.

4. To stimulate further, more focused, simulation experiments that would seek to confirm tentative statements regarding the relative comparison of ISMs resulting from this simulation.

7.2.2. Data generating mechanisms

7.2.2.1 Direct evidence

The data-generating model for the direct evidence is outlined in eqs. (7.1) to (7.4). For each direct evidence dataset, study-specific relative treatment effects d_{dir_i} (log-odds ratios) were randomly drawn from a normal distribution with mean d_{dir} and between-studies standard deviation τ_{dir} (Equation 7.1). The parameter τ_{dir} was based on the estimated heterogeneity of a pairwise RE meta-analysis of the adult direct evidence in the case-study of Chapter 5. Furthermore, for each direct evidence dataset, study-specific baseline log-odds (μ_i) were drawn from a normal distribution with mean ($\mu_{dir} = -0.5$) and variance ($sd^2_{\mu_{dir}} = 0.5^2$) that were estimated by fitting a normal distribution to the control arm of the direct evidence in the case-study of Chapter 5 (Equation 7.2). The number of events in the control arm ($r_i^{control}$) was drawn from a binomial distribution in which the probability of an event was based on both the aforementioned randomly drawn baseline log-odds and the size of the control arm (Equation 7.3). A binomial distribution was also used to draw the number of events in the treatment arm, however the probability parameter was based on both the simulated baseline log-odds and the simulated relative treatment effects (Equation 7.4). No alternative scenarios were used for the direct evidence because the degree of sharing, which is the main consideration in this work, is expected to be driven by the relative difference between direct and indirect evidence sets. Hence, all direct evidence was generated based on the aforementioned process and properties.

$$d_{dir_i} \sim N(d_{dir}, \tau^2_{dir}) \tag{7.1}$$

$$\mu_i \sim N(\mu_{dir}, sd^2_{\mu_{dir}}) \tag{7.2}$$

$$r_i^{control} \sim Binomial(inverse.logit(\mu_i),\ \frac{ss_{dir}}{2 \cdot n_{dir}}) \tag{7.3}$$

$$r_i^{trt} \sim Binomial(inverse.logit(\mu_i + d_{dir_i}),\ \frac{ss_{dir}}{2 \cdot n_{dir}}) \tag{7.4}$$

In total 5000 direct datasets were simulated, each one including as many patients as the direct evidence of the case-study (i.e. $ss_{dir} = 2300$). The number of datasets was chosen to ensure that the standard deviation of the strength of sharing measures was stable (see Appendix on page 240 for more details). Each direct dataset comprised of $n_{dir} = 17$ studies (i.e. as many direct studies as in the case-study) of equal size (2300/17). Note that the study-specific sample sizes of the case-study were not used because the existence of small studies would require zero-cell adjustments which could affect the stability of the simulations, and could bias estimates (Higgins and Green, 2011). It was assumed that there were only two competing treatments (i.e. control and treatment) and that, across studies, control and treatment arms were of equal size (2300/ (17*2)). The point estimate d_{dir} was based on a FE meta-analysis without any covariates of the adult evidence used in the case-study of Chapter 5, where SoC and Albumin had been lumped as comparator treatments and IVIG and IVIGAM had been lumped as active treatments (i.e. network T2 of Figure 5.1). The uncertainty surrounding the point estimates was assumed to be different between the FE and the RE simulations. Specifically, the standard errors of the RTE of the direct evidence in the FE and RE simulations (i.e. $se_{dir_{FE}}$, $se_{dir_{RE}}$) were based on the corresponding standard errors of the RTEs of the pairwise FE and RE meta-analyses of the direct evidence in the case-study. Overall, the point estimates and their associated standard errors were assumed $\sim N(d_{dir} = -0.43, se_{dir_{FE}} = 0.11)$ for the FE simulation and $\sim N(d_{dir} = -0.43, se_{dir_{RE}} = 0.22)$ for the RE simulation.

7.2.2.2 Indirect evidence —alternative scenarios

Indirect datasets were comprised of $n_{indir} = 10$ studies[1], which were simulated using the same data-generating model as the direct evidence, and in accordance with the properties of the various scenarios that were defined. It was expected that the degree of information-sharing would be affected by the following three key features of the evidence base:

1. The point estimate of the indirect evidence
2. The number of patients included in the indirect evidence
3. The heterogeneity of the indirect evidence

In the *base-case*, the heterogeneity of the indirect evidence is assumed equivalent to the heterogeneity of the direct evidence (i.e. $\tau_{indir_{base-case}} = \tau_{dir} = 0.56$), and the number of patients equivalent to the number of patients in the direct evidence (i.e. $ss_{indir_{base-case}} = ss_{dir} = 2300$). In contrast, the point estimate of relative treatment effect $d_{indir_{base-case}}$ was not assumed equivalent to that of the direct evidence (i.e. $d_{dir} = -0.4$), because the two evidence sets would effectively be equivalent. Instead, the point estimate of the relative effect for the indirect evidence in the base-case was that which yielded a 50% overlapping coefficient[2] (OVL) when compared with d_{dir}, and in the direction of 'no effect' (i.e. FE: -0.281, RE: -0.13). Note that under RE the RTE estimate of direct evidence was assumed to have a larger standard error, and it is for this reason that a 50% OVL under RE produces an estimate that is further away from d_{dir} than under FE.

Additional levels for the heterogeneity were loosely based on the desired between-studies heterogeneity I^2 of the indirect evidence. General rules of thumb suggest that $I^2 \leqslant 25\%$ means that there is low heterogeneity, $25\% < I^2 \leqslant 50\%$ medium-low, $50\% < I^2 \leqslant 75\%$ medium-high, and $I^2 > 75\%$ very high (Higgins and Green, 2011). Here, two heterogeneity scenarios were defined; one where the indirect evidence was quite homogeneous (i.e. $I^2 \approx 12.5\%$) and one where it was quite heterogeneous (i.e. $I^2 \approx 87.5\%$). The corresponding τ values were 0.24 and 0.65 respectively. More details about the τ calculation process can be found in the Appendix on 245.

[1] The choice was made to include 10 direct studies. This was primarily because it is close to the number of indirect studies included in the case-study (i.e. 11), while it also allows us to easily control the proportion of indirect patients that will be included in each of the studies. The importance of this will become evident in the scenario run under the RE simulation where this proportion is modified.

[2] The overlapping coefficient is defined here according to Weitzman, 1970 as the area lying under both the density curves of d_{dir} and $d_{indir_{base-case}}$. Further details for the calculation of the overlapping coefficient, including R code, are provided in the Appendix on 241.

Regarding the number of patients included in the indirect evidence, two additional scenarios were considered. One in which the indirect evidence was modelled to have four times as many patients as the direct ($N = 2300 * 4 = 9200$) —and therefore expected to yield a RTE estimate with half the standard error of that of the direct evidence[3] —and the second scenario wherein the indirect evidence was modelled to have half as many patients as the direct evidence ($N = 2300/2 = 1150$). No scenarios were considered where the number of patients in the indirect evidence was fewer than half of the patients in the direct evidence, because such a scenario would be unlikely to motivate any borrowing strength from the indirect evidence.

Additional levels for the point estimate of the relative treatment effect where defined based on OVL (Weitzman, 1970). Specifically, in one scenario the direct and the indirect RTE estimates only minimally overlapped ($OVL = 5\%$), and in another scenario they majorly overlapped ($OVL = 75\%$). As mentioned above, the estimates that yielded the desired OVL were different across FE and RE simulations because a larger standard error for d_{dir} was used for the RE simulation.

Finally, for the RE model simulation, an extreme scenario was defined in which the indirect evidence included eight times as many patients as the direct evidence and one big study contained 85% of all the indirect patients, while the remaining nine studies were modelled with equal size. Additionally, the big study was assumed to suggest a very different relative effect to that of the small studies[4]. The purpose of this scenario was to mimic situations where a number of studies in an indirect population might motivate the conduct of a large multi-center randomised clinical trial that would go on to show no effect; just as in the case-study of Chapter 5.

Importantly, drawing random samples for such a scenario is challenging and requires some necessary simplifications. In particular, given that we want to preserve the overall properties of the indirect evidence (i.e. d_{indir}, τ_{indir}), we need to initially draw the relative effect of the big study ($d_{indir_{BIG}}$) from the right tail of the overall predictive distribution for the indirect evidence. We can subsequently 'back-calculate' the relative treatment effect of the remaining studies $d_{indir_{SMALL}}$ that preserves the overall d_{indir} and yields roughly the same τ_{indir}. To make such calculations easier, the assumption was made that all small indirect studies had exactly the same relative treatment effect $d_{indir_{SMALL}}$, and that there was no heterogeneity amongst them. The back-calculation process is explained in detail on page 243 of the Appendix.

[3] Because $se = \frac{\sigma}{\sqrt{N}}$ where se is the standard error, σ the population variance, and N the sample size.

[4] Here it is assumed that the big study is drawn from the right tail (beyond the 90% percentile) of the predictive distribution that is defined by d_{indir}, τ_{indir} and hence yields a positive log-odds ratio that suggests that the new treatment is less effective than the comparator.

Table 7.1: *Properties of direct and indirect evidence across scenarios in FE and RE simulations.*

Dimension	Direct Evidence	Indirect Evidence (Base-case)	Indirect Evidence (Additional Scenarios)
Fixed-effect simulation			
Number of studies	17	10	-
Number of patients	2300	2300	1. 2300*4, 2. 2300/2
Heterogeneity	$\tau = 0.56$	$\tau = 0.56$	1. $\tau = 0.24$ 2. $\tau = 0.65$
Point estimate	Log-odds ratio = -0.43 (i.e. OR = 0.65)	-0.281 i.e. OR = 0.76 (i.e. that which yields 50% overlapping coefficient)	1. 0.003 (5% OVL) i.e. OR=1.003 2. -0.36 (75% OVL) i.e. OR=0.7
Proportion of patients in one study	Each study contains 1/17 of the overall number of patients equally split across arms	Each study contains 1/10 of the overall number of patients equally split across arms	-
Random-effects simulation			
Number of studies	17	10	-
Number of patients	2300	2300	1. 2300*4, 2. 2300/2
Heterogeneity	$\tau = 0.56$	$\tau = 0.56$	1. $\tau = 0.24$ 2. $\tau = 0.65$
Point estimate	Log-odds ratio = -0.43 (i.e. OR = 0.65)	-0.13 i.e. OR = 0.87 (i.e. that which yields 50% overlapping coefficient)	1. 0.45 (5% OVL) i.e. OR= 1.57 2. -0.29 (75% OVL) i.e. OR= 0.74
Proportion of patients in one study	Each study contains 1/17 of the overall number of patients equally spread across arms	Each study contains 1/10 of the overall number of patients equally split across arms	One study contains 85% of all the indirect patients (i.e. 8*2300*0.85), and all the remaining studies contain 8*2300*0.15 equally split among them

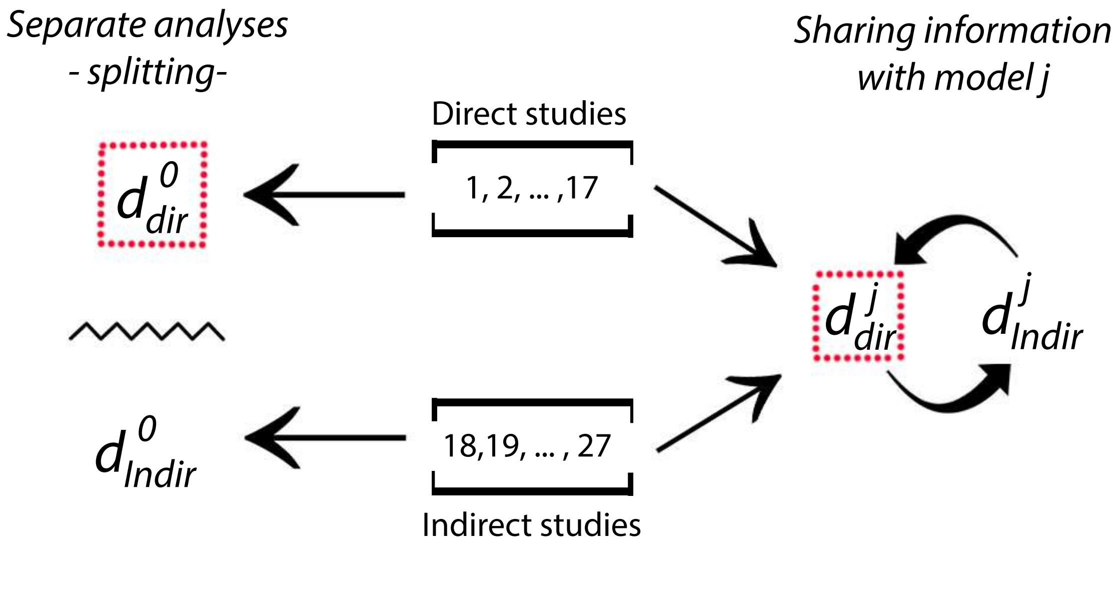
This scenario was only tested under the RE model because FE models give lower weights to smaller studies and therefore the back-calculation process sometimes results in no solution (i.e. there is no $d_{indir_{SMALL}}$ which when combined with $d_{indir_{BIG}}$ recovers the overall d_{indir}).

To obtain a feasible number of scenarios, dimensions were varied *one-by-one*, instead of factorially, despite this process not allowing us to explore the effect of dimension interactions. The simulation experiment was run twice: once under a FE model and once under a RE model. Overall, including the base-case, seven scenarios were run for the FE model and eight for the RE. A summary of the various scenarios tested under the FE and RE models is provided in Table 7.1.

7.2.3. Target quantity

We wish to compare the direct RTE estimate under the splitting method ($method_0$), d^0_{dir}, which does not borrow any strength from the indirect evidence with the direct RTE estimate from each one of the available ISMs ($method_1, ..., method_j$), d^j_{dir}, which combines the direct and the indirect evidence under different assumptions. Figure 7.1 shows in red the two relative effect quantities that we are interested in comparing (i.e. d^0_{dir} and d^j_{dir}).

Figure 7.1: *An illustration of the two RTE quantities of interest.*



Where d^0_{dir} corresponds to the 'un-strengthened' RTE estimate that pertains to the direct evidence and is produced by synthesising only the direct evidence, and d^j_{dir} corresponds to the 'strengthened' RTE estimate that pertains to the direct evidence and results from synthesising both direct and indirect evidence using information-sharing $model_j$.

7.2.4. Information-sharing methods

Table 7.2 lists separately all the ISMs that were used under the FE and RE simulations. More details regarding all these methods can be found in Chapter 4 where all methods are explained in detail in the context of NMA. In this simulation only two treatments are compared and therefore all methods are reduced to the pairwise meta-analysis case.

As explained in Chapter 4, not all ISMs are applicable under both FE and RE models. This is because RE models define the between-studies heterogeneity on which further ISMs can be imposed. In brief, the following methods may differ between FE and RE models:

1. Lumping: Even though under FE models only one lumping approach can be used (i.e. that which imposes a common RTE mean across direct and indirect evidence —i.e. lumping (d only) —), under RE models several lumping approaches may be adopted depending on which parameters of the direct and the indirect evidence are lumped. One approach is to lump only the RTE mean between direct and indirect evidence —i.e. lumping (d only). Another approach is to analyse all studies (direct and indirect) using a common random-effect allowing information-sharing on both the RTE mean and the between-studies heterogeneity, that is lumping (d & τ). Finally, one may choose to only share information on between-studies heterogeneity, that is lumping (τ only). All three of these approaches to lumping are implemented in the RE simulation.

2. Informative and mixture priors: When the indirect evidence is initially analysed using a RE model, the analyst can choose to use either the posterior distribution of the mean of the RTE, or its predictive distribution as a prior. Here, both approaches are used in order to assess how they compare to each other in terms of the strength of sharing that they impose.

Finally, it is noted that mixture priors were used with the assumption that the weights of the prior components were uncertain parameters estimated in the model, and commensurate priors were implemented assuming that the Bernoulli trials had a fixed 50% chance of yielding the 'spike' or the 'slab' hyper-prior.

Table 7.2: *ISMs used in the FE and RE simulations.*

Information-sharing method	FE	RE
Lumping (only d) Direct and indirect evidence are assumed to have the same RTE mean $d_{dir} = d_{indir}$	✓	✓
Lumping (d & tau) A random-effect is imposed across all studies regardless of whether they provide direct or indirectly related evidence $d_{dir} = d_{indir}; \tau_{dir} = \tau_{indir}$	X	✓
Constraint $d_{dir} < d_{indir}$	✓	✓
Multi-level model $d_{dir}, d_{indir} \sim N(d_{overall}, \tau^2_{overall})$	✓	✓
Random-walk $d_{dir} \sim N(d_{indir}, \eta^2)$ where a vague prior is imposed on η	✓	✓
Commensurate prior $d_{dir} \sim N(d_{indir}, \eta^2)$ where $\frac{1}{\eta} \sim \begin{cases} N(20,1) & \text{,if } c = 0 \\ Gamma(0.1,0.1)I(0.1,5) & \text{,if } c = 1 \end{cases}$ and $c \sim Bernoulli(p)$, with $p = 0.5$	✓	✓
Informative-prior on the relative effect of the direct evidence based on the estimated posterior mean distribution of the indirect evidence $d_{dir} \sim N(d_{indir}, se^2_{d_{indir}})$	✓	✓
Mixture-prior on the relative effect of the direct evidence based on the estimated posterior mean distribution of the indirect evidence and a vague component $d_{dir} \sim p \cdot N(d_{indir}, se_{d^2_{indir}}) + (1-p) \cdot N(0, 100^2)$ with p estimated within the model	✓	✓
Informative-prior on the relative effect of the direct evidence based on the estimated predictive distribution of the indirect evidence $d_{dir} \sim N(d_{indir}, \tau^2_{indir})$	X	✓
Mixture-prior on the relative effect of the direct evidence based on the estimated predictive distribution of the indirect evidence and a vague component $d_{dir} \sim p \cdot N(d_{indir}, \tau^2_{indir}) + (1-p) \cdot N(0, 100^2)$ with p estimated within the model	X	✓
Power-prior on the relative effect of the direct evidence based on the relative effect of the indirect evidence and $\alpha = 0.5$ $\pi(d \| Data_{dir}, Data_{indir}, \alpha) \propto$ $L(d \| Data_{dir}) \cdot \{\prod_{i=1st-indirect-study}^{last-indirect-study} L(d \| Data_{indir})^{\alpha}\} \cdot \pi_0(d)$	✓	✓
Lumping (only on τ) $\tau_{dir} = \tau_{indir}$	X	✓
Informative prior on the heterogeneity $\tau_{dir} \sim Lognormal(\mu_{indir}, \sigma^2_{indir})$ $\mu_{indir}, \sigma_{indir}$ are derived from fitting a log-normal distribution to the coda of τ_{indir}	X	✓

More details about the methods listed can be found in Chapter 4.

7.2.5. Strength-of-sharing measures

Simulation experiments usually follow a very specific process. They assume an underlying truth for a population parameter (say θ), they use θ in their data-generating model to produce a set of randomly drawn data, they then analyse the resulting data using a particular method (e.g. an estimator, say $\hat{\theta}$), and then finally they evaluate the performance of $\hat{\theta}$ using a set of measures such as bias, mean-squared error, or coverage (Morris et al., 2019).

Here, as is also the case in policy-making, we do not have access to the underlying true RTE of the population that we are directly interested in, and perhaps more importantly, we do not have an understanding of the nature of the true relationship between the indirect evidence and the true RTE for the direct evidence. Instead, we have only a set of studies conducted in the direct population and another set of studies conducted in an indirect population. Therefore, when we combine the two evidence sets using different ISMs, we cannot use the classic performance measures to evaluate how the estimate that results from the combination of the two evidence sets using ISM j (i.e. d^j_{dir}) compares to the true parameter. Instead, we can only evaluate how d^j_{dir} compares to the estimate that is produced when only the direct evidence is used (i.e. d^0_{dir}). In other words, we can only assess how much d^j_{dir} diverges from d^0_{dir}, or in other words, how much strength is borrowed from the indirect evidence.

The two measures that were previously introduced on page 116 are also used here to assess the degree of information-sharing imposed by ISM j compared to splitting. Namely, these measures are firstly the *Point estimate divergence* (PED), which is defined as $PED = |d^j_{dir} - d^0_{dir}|$, and secondly the *Precision Increase* (PrI) where $PrI = 1 - se_{d^j_{dir}} / se_{d^0_{dir}}$. Note that Kullback-Leibler divergence is not used here because, as shown on page 230, it is not transparent in the way that it weighs changes in the point estimate and the precision and here we want to make such trade-offs explicit.

However, it should be noted that PED and PrI are influenced by the absolute magnitude of the simulated RTEs and their uncertainty. For instance, when the direct and indirect RTEs happen to be simulated such that they have similar RTE point estimates, all methods will inevitably lead to low PED values. In contrast, when the direct and indirect RTEs happen to be simulated so that they are quite distant in their RTE point estimates, most methods will produce much higher PED values. However, it may be the case that a given method is imposing a similar degree of information-sharing in both cases i.e. in both cases a particular method may impose a PED that is 50% of the PED imposed by lumping. Hence, PED and PrI are affected by the variability across simulated datasets. Therefore, ranking ISMs according to PED and PrI would add unnecessary noise to conclusions.

Given that the primary aim here is to understand how methods performed relative to one another (while not directly interested in the variability across datasets beyond their influence on the degree of sharing), additional metrics were developed which were standardised to the anchor points of lumping and splitting.

To evaluate how each method compares to lumping in terms of the point estimate divergence that it imposes, the *PED-ratio* (i.e. proportion of lumping's PED) is used, defined as $PED - ratio_j = \frac{PED_j}{PED_{lumping}}$. When $PED - ratio_j = 1$, method *j* is leading to the same changes in the point estimate as lumping, whilst when it is close to 0, method *j* is producing very similar point estimates with a splitting approach. For PED-ratio values below 1, method *j* is causing smaller changes in the point estimate than lumping, whilst for values above 1, method *j* is leading to larger shifts in the point estimate than lumping.

In what concerns relationships between the precision of the estimates, a ratio of PrI cannot be used to assess how each method compares to lumping in terms of the precision of the resulting relative effect estimate. This is because PrI can be either positive or negative and hence a positive PrI-ratio could result from either both method *j* and lumping yielding positive PrI, or alternatively from both method *j* and lumping yielding negative PrIs. Instead, to enable comparison of the precision under lumping and ISM *j*, a simple ratio of the standard errors (SeR) under lumping and ISM *j* is defined as $LumpingSeR = \frac{se_{Lumping}}{se_j}$. When *lumping SeR* is equal to 1, ISM *j* results in RTE estimates of the same precision as lumping. When Lumping SeR is above 1 it means that ISM *j* results in more precise estimates than lumping, whilst when Lumping SeR is below 1 it indicates that ISM *j* leads to more uncertain estimates for the mean relative effect than lumping.

A disadvantage of using Lumping SeR is that we have no way of understanding where the precision that corresponds to splitting lies, and we would like to be able to identify cases where an ISM is producing more uncertain estimates than splitting i.e. when information-sharing leads to precision losses. To accommodate this objective a very similar *Splitting SeR* can be constructed that would compare the standard error of ISM *j* to that of splitting, so that $SplittingSeR = \frac{se_{splitting}}{se_j}$. When Splitting SeR is equal to 1 it indicates that when using ISM *j* the uncertainty surrounding the RTE mean is the same as in the splitting case. In contrast, a Splitting SeR above 1 indicates that ISM *j* is leading to precision gains, whilst a Splitting SeR below 1 that ISM *j* is leading to precision losses.

Overall, the following five strength of sharing measures are used: PED, PrI, PED-ratio, Lumping SeR, Splitting SeR.

7.2.6. Software and implementation

The simulation experiment was run in the *York Advanced Research Computing Cluster* (YARCC), using *R* version 3.5.1 (R Development Core Team, 2010) to simulate the datasets, and *OpenBUGS* (MRC Biostatistics Unit, 2010) version 3.2.3 to analyse the datasets using the various ISMs. Package R2OpenBUGS (Sturtz et al.) was used to call *OpenBUGS* from *R*. All models were run using three MCMC chains with different starting values. Model convergence was checked using the Gelman-Rubin diagnostic[5] and in particular the multivariate potential scale reduction factor (psrf statistic) (Gelman and Rubin, 1992). The experiment was divided into multiple 'jobs' which were then allocated to different cores in the cluster. The overall computing time in the cluster was 20,000 hours.

7.2.7. Presentational methods

This simulation aims to rank ISMs according to the strength of information-sharing that they impose. To do this, multiple datasets were simulated and subsequently analysed using all applicable ISMs. Each method's strength of sharing was then calculated by comparing its resulting 'strengthened' RTE estimate with the 'un-strengthened' splitting estimate that was obtained by solely analysing the direct evidence. The various methods were then ranked within each simulated dataset according to each strength of sharing measure, and the overall probability of each method being ranked at each position across all simulated datasets was calculated.

This process is analogous to the process commonly used to rank treatments in NMA models according to their effectiveness. Essentially, instead of ISMs, in NMA there are treatments, and instead of ranking methods according to the imposed strength of sharing, in NMA treatments are ranked according to their log-odds ratios (if a binary outcome is used). Several methods to rank treatments in NMA have been suggested in the literature including *rankograms* and surface under the cumulative ranking area (SUCRA) values (Salanti et al., 2011; Chaimani et al., 2013). Rankograms illustrate the probability of each of the treatments being ranked at each of the positions across MCMC iterations. Sucra values provide an overall measure of which treatment is the best by calculating the cumulative area under the ranking curves, with higher values indicating better treatments. Here, these methods are adapted to rank ISMs according to strength of sharing measures rather than treatments in terms of their effectiveness.

[5] The Gelman-Rubin diagnostic provides the scale reduction factor for each parameter specified in the model. Essentially, it compares the within-chains variance with the between-chains variance, with a scale reduction factor of 1 implying that within- and between- chains variances are equal. To avoid checking the reduction factor for each parameter, the multi-variate reduction factor was used, which was able to take into account all parameters specified in the model.

Table 7.3: *A summary of the main characteristics of the simulation.*

Aim	To understand how particular statistical characteristics influence the relative ranking of ISMs.
Dimensions	The following dimensions were varied only for the indirect evidence and defined in relation to the values implemented for the direct evidence: 1. Point estimate 2. Number of patients 3. Heterogeneity 4. Proportion of the overall patients included in a single study. (only applicable for RE) The levels that were used within each dimension can be found in Table 7.1.
Simulation characteristics	5000 datasets were simulated per scenario. Dimensions were varied 'one-by-one'. Two simulation experiments were run: one under FE; and another under RE base-models.
Target quantity	d^{j}_{dir} (see Figure 7.1)
Strength-of-sharing measures	1. Point Estimate Divergence (PED) 2. Precision Increase (PrI) 3. PED-ratio (i.e. Proportion of Lumping's PED) 4. Lumping Standard error Ratio (Lumping SeR) 5. Splitting Standard error Ratio (Splitting SeR)
ISMs	See Table 7.2
Presentational methods	1. Sucra values 2. Rankograms 3. Forest plots of PED-ratios, Lumping SeR and Splitting SeR.

ISM: Information-sharing method.

A summary of the main characteristics of the simulation is provided in Table 7.3 and a step-by-step explanation of the simulation process is detailed in the Appendix on page 246.

7.3. Results

7.3.1. Fixed-effect simulation

7.3.1.1 Base-case scenario

Table 7.4 shows the sucra values of PED and PrI for the base-case scenario of the FE simulation. Higher sucra values indicate that a method is more often ranked higher in terms of the corresponding strength of sharing measure. Lumping and the informative-prior yield the highest sucra values and are practically equivalent as discussed on page 74. The mixture prior borrows-strength relatively strongly and, compared to the power-prior, it leads to a slightly higher PED and somewhat lower PrI. Given that in the base-case scenario direct and indirect evidence overlap by an OVL coefficient of 50%, the constraint imposes non-negligible information-sharing. Despite ranking close to the bottom, commensurate priors borrow more strength than random-walks in terms of both measures due to the additional assumptions they impose on the variance component. Finally, the multi-level models rank last imposing the least information-sharing across both measures.

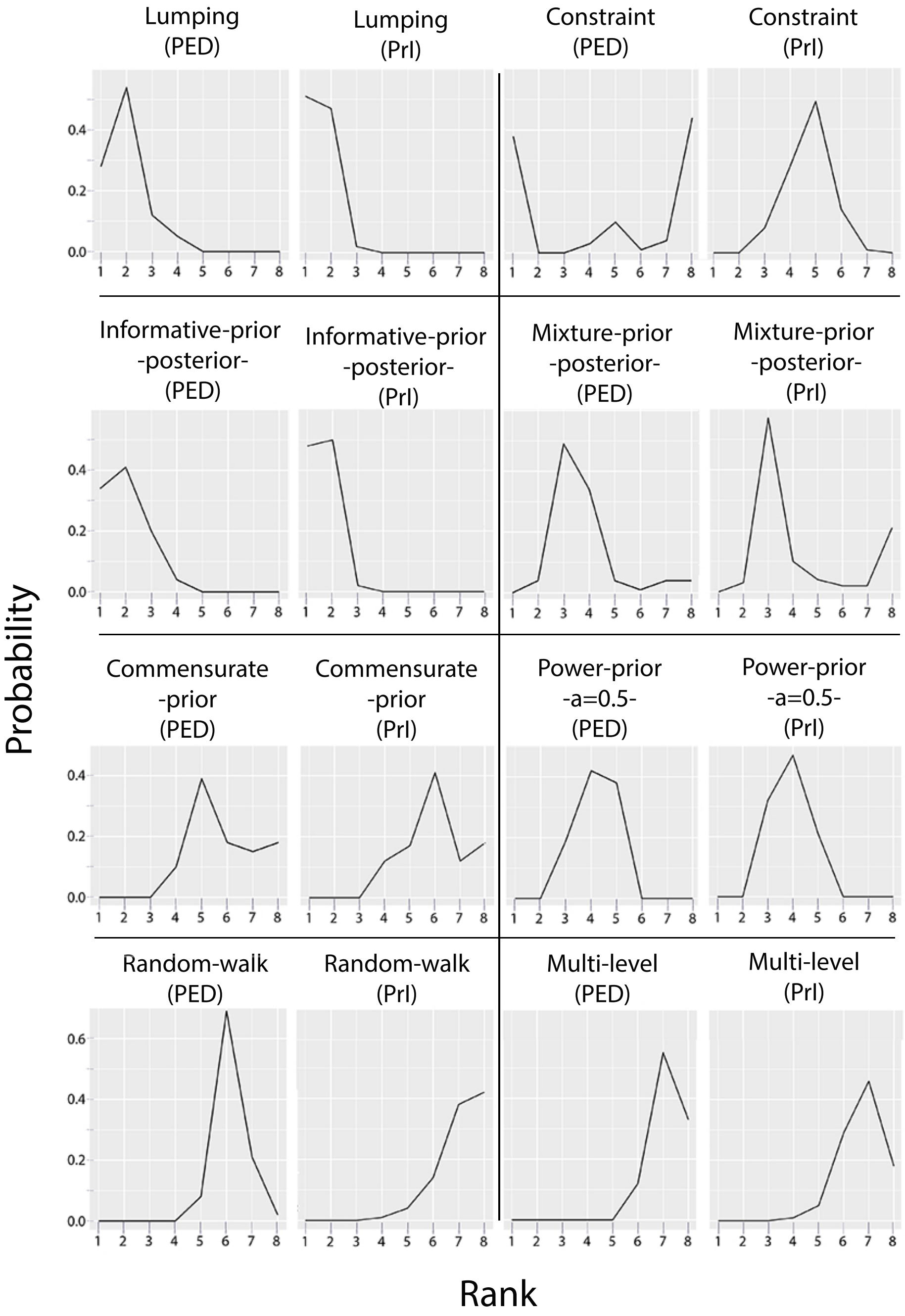
Table 7.4: *Sucra values of the two strength-of-sharing measures (PED, PrI) used in the analysis of the base-case scenario under a FE model.*

Information-sharing method	PED Sucras	PrI Sucras
Lumping (d only)	0.75	0.81
Informative-prior ‡	0.75	0.81
Mixture-prior ‡ (p estimated in model)	0.53	0.45
Power-prior ($\alpha = 0.5$)	0.47	0.51
Constraint	0.39	0.41
Commensurate prior (p fixed at 0.5)	0.26	0.24
Random-walk	0.23	0.15
Multi-level	0.1	0.1

Methods are arranged in a descending PED sucras. ‡: under FE only the posterior estimate for the RTE of the indirect evidence can be used as an informative prior for the analysis of the direct evidence.

Despite their conciseness, sucra values are only useful in providing an overall hierarchy of the various methods. To understand exactly how methods accumulate sucras, one has to look at the rankograms which reveal exactly how each method's ranking probability is distributed across different ranks. Figure 7.2 illustrates the rankograms of each method for both measures side by side. The y-axis is the probability of ranking at the position shown in the x-axis.

Figure 7.2: *Rankograms of ISMs used in the base-case scenario of the FE simulation.*



Graphs are organised in pairs, showing the rankings according to the two strength-of-sharing measures (PED and PrI) side by side to reveal commonalities and differences. The *y*-axis depicts the probability of ranking in the position shown in the *x*-axis.

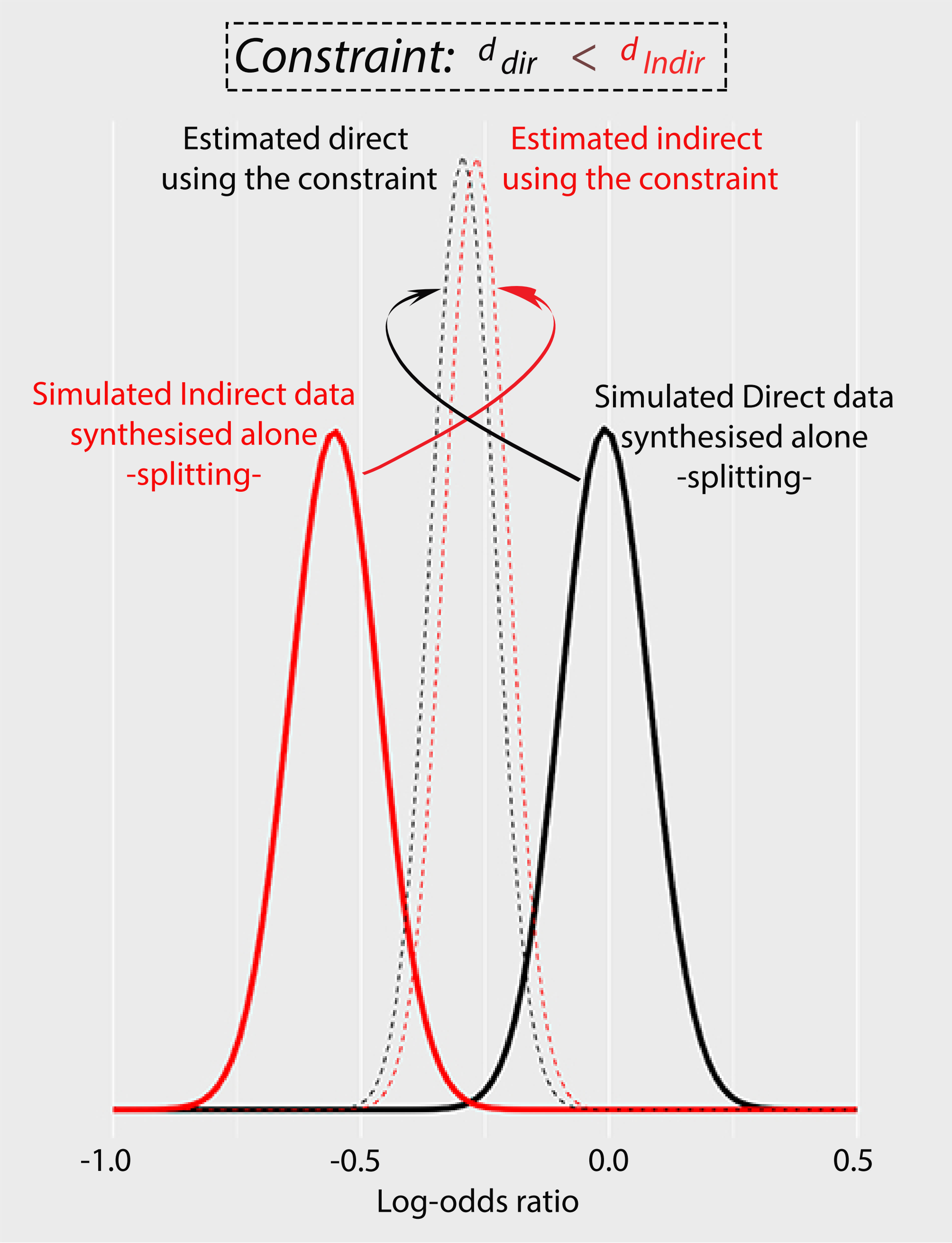
The rankograms show that methods do not always rank in the same position, but they fluctuate between similar/adjacent positions. In general, methods rank in similar positions for both measures which implies that PED and PrI are often in agreement. Two exceptions can be observed. The first is that mixture priors have around 20% probability of being ranked last in terms of PrI but not in terms of PED. This is because mixture priors have the capacity to lead to significant precision losses; this feature will be discussed in detail later in the chapter.

The second exception relates to the constraints (top-right in Figure 7.2). Constraints seem to most often rank in the extremes for PED but almost never for PrI, suggesting that they always offer some precision gains, but have the potential to either majorly shift the point estimate or leave it almost unaffected. The latter happens when the simulated direct and indirect evidence end up being even more separated, and the direction of the relative effects is that which is to be expected by the constraint (i.e that the direct evidence is more negative —i.e. on the left —of the indirect). The former case arises when the simulated direct and indirect evidence suggests that the *indirect* relative effect is more negative than the *direct*. This comes in contrast to the specified direction of the constraint, and in order to satisfy the specified direction of relative effects, the model majorly shifts direct and indirect means so that the specified constraint is only marginally satisfied.

An example relating to one of the simulated datasets is presented in Figure 7.3. This shows the simulated direct and indirect evidence of Dataset 38 of the base-case scenario. The solid lines depict the RTE estimates when separately analysing the direct (black) and the indirect (red) evidence with a simple FE model (i.e. splitting). Note that in this dataset the direct evidence indicates a more positive effect than the indirect, whilst the characteristics of the base-case scenario suggest the opposite. Therefore, when these two sources are simultaneously analysed imposing the constraint that expects the indirect to be more positive, the direction of effects is swapped and precision gains are preserved. The resulting RTE estimates are shown in dotted lines.

It is important to note at this point that it may be hard to justify imposing a constraint that is of the opposite direction of what the data suggest. However, the existing evidence may suggest a counter-intuitive direction due to sampling error and/or between-studies heterogeneity. Therefore, if there is a robust clinical rationale suggesting a particular direction of effects, it may still be reasonable to express it in the synthesis model. For instance, if we are analysing studies that used different dosages of the same drug, it may be reasonable to assume that as the dosage increases the relative effect also increases, even if this is not supported by the existing evidence, but rather from clinical experts only.

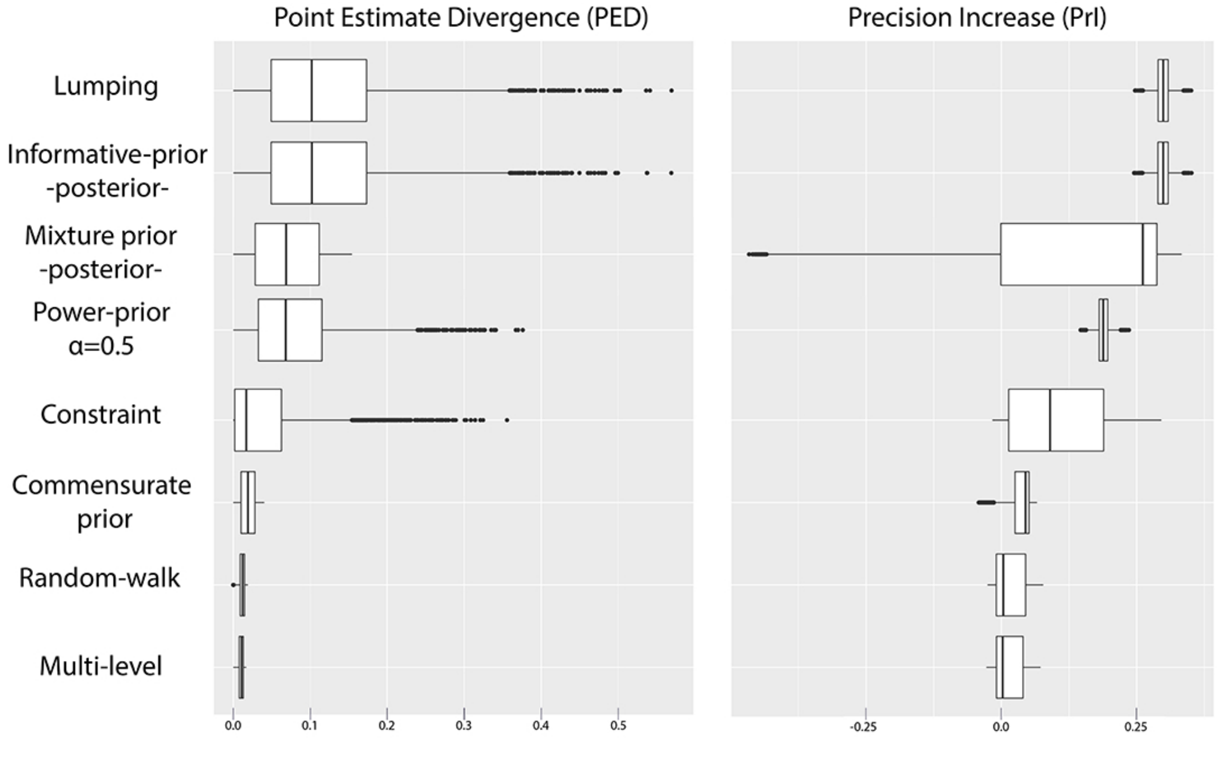
Figure 7.3: *A case where the simulated data suggest the opposite direction of relative effects than the specified constraint (Dataset 38 of the base-case scenario).*



When the simulated direct and indirect studies are analysed on their own (solid lines) they suggest a more positive relative effect for the direct evidence, even though the scenario characteristics specify the opposite. A simultaneous analysis of the two evidence sets using the expected constraint that is indicated by the scenario characteristics leads to a major shift in their means in order to conform to the specified direction of the constraint.

To get a sense of the magnitude of each method's information-sharing, boxplots for both measures calculated across all simulated datasets of the base-case scenario are shown (Figure 7.4). Note that a method that has a larger median PED and PrI than another method across all simulations, does not necessarily have larger PED or PrI within each simulation. The various methods result in median PEDs between $0.01 - 0.1$, with extreme outlier values reaching up to 0.6 in the log-odds ratio scale. In terms of the base-case scenario this means that when strength is borrowed from the indirect evidence the odds-ratio can change from $OR = 0.65$ up to $OR = 0.71$ on average if the two sources are lumped, but also up to $OR = 1.18$ in extreme cases. Regarding PrI, all methods yield positive median PrI, suggesting that, on average, they reduce the standard error for the RTE by 0.1% to 30%. Although infrequently, commensurate priors, random-walks, and multi-level models can result in minimal precision loses compared to splitting, while mixture priors may lead to considerable precision losses (this is more thoroughly discussed and illustrated in Figure 7.6).

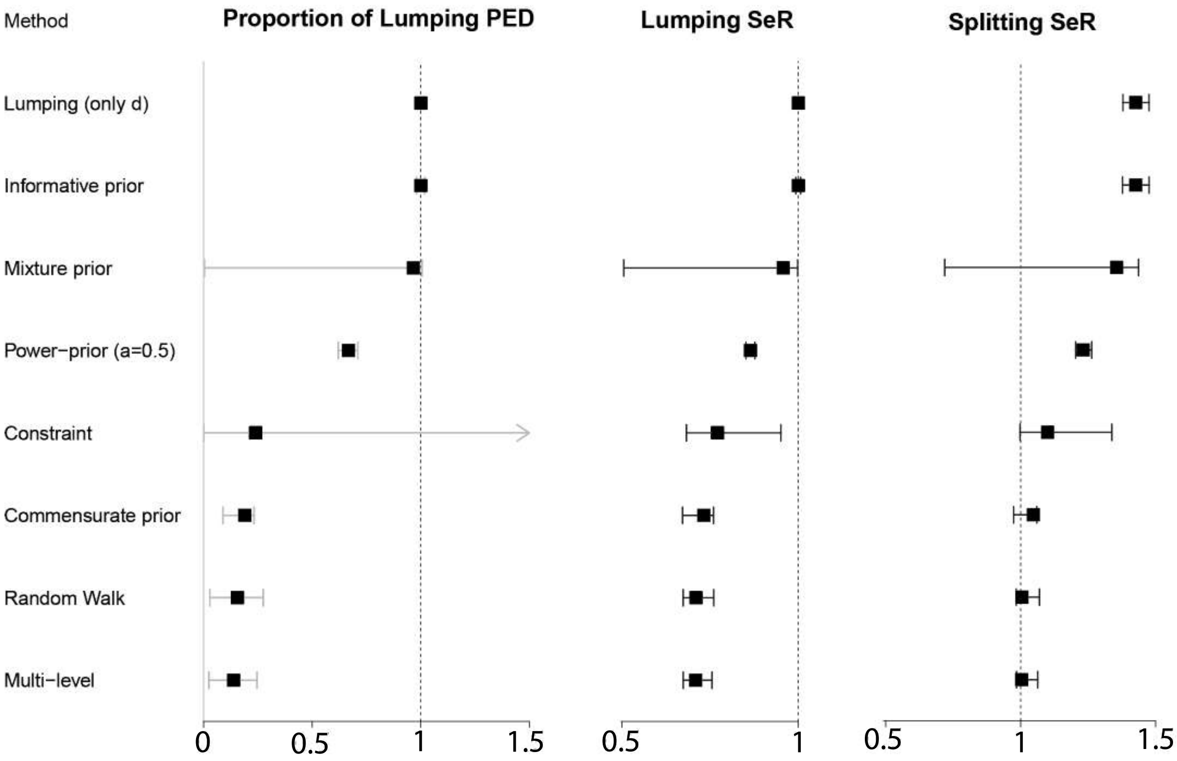
Figure 7.4: *Absolute values of PED and PrI across all simulations in the base-case scenario under FE.*



PED values are in the log-odds ratio scale.

Figure 7.5 shows how the various methods compare, within simulations, to lumping and splitting in terms of both PED and precision. The plot on the left depicts the PED-ratio (i.e. the ratio of a method's PED divided by lumping's PED) and the solid grey and dotted black lines illustrate where splitting and lumping fall respectively. The plot in the middle shows Lumping SeR (i.e. the ratio of lumping's standard error of the RTE divided by the standard error of the RTE which is estimated using each ISM) and the plot on the right displays the Splitting SeR (i.e. the ratio of splitting's standard error of RTE divided by the standard error of the RTE which is estimated using each ISM).

Figure 7.5: *Forest plots of PED-ratios, Lumping SeR, and Splitting SeR across ISMs.*



The 'Proportion of Lumping PED' (i.e. PED-ratio) reflects a methods PED divided by lumping's PED. 'Lumping SeR' is the ratio of a method's standard error divided by lumping's standard error. 'Splitting SeR' is the ratio of a method's standard error divided by splitting's standard error.

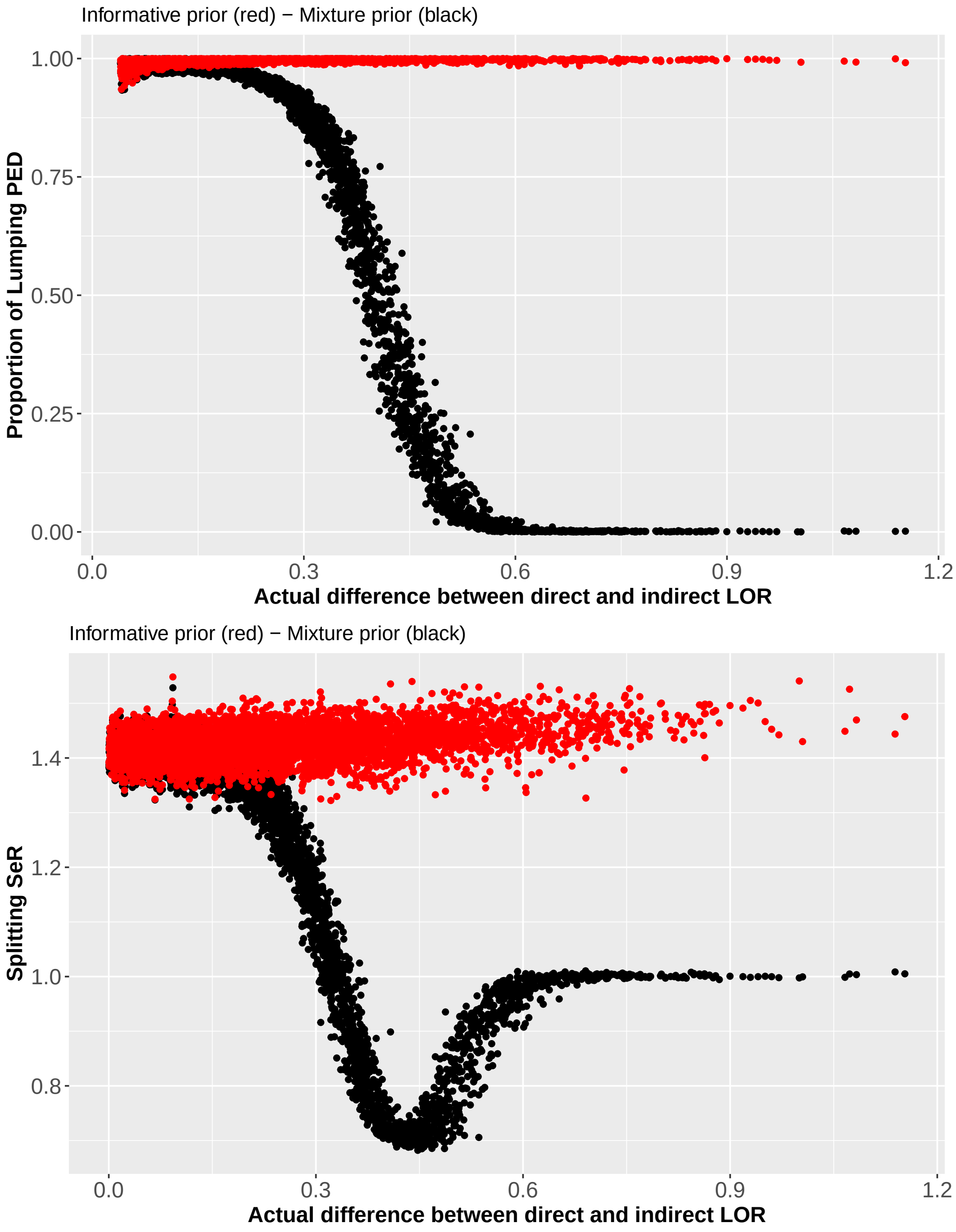
Except for mixture priors and constraints, most methods' credible intervals are quite narrow implying that, across simulations, they relate to lumping and splitting in a specific way that does not vary considerably across simulated datasets. Informative priors are practically equivalent to lumping across all measures, whilst power-priors ($\alpha = 0.5$) impose more moderate information-sharing; these do not however correspond to 50% of lumping. This suggests that interpreting the α value in power-priors may be challenging as the extent of information-sharing may not be linearly related to α. Commensurate-priors, random-walks and multi-level models impose minimal information-sharing, with their PED ranging between 0-25% of the lumping's PED, and their Splitting SeR falling very close to 1, indicating that they often do not lead to precision gains compared to splitting. Interestingly, the constraints show considerable density beyond 1 for the PED-ratio, suggesting that they can lead to larger changes in the point estimate than lumping (this is a reflection of the phenomenon discussed in Figure 7.3), but they never offer higher precision gains than lumping, nor precision losses.

Perhaps the most interesting feature of Figure 7.5 relates to the mixture priors. Even though mixture priors, on average, lead to very similar PED and precision gains to informative priors (and by extension to lumping), their confidence bands cover the whole range between lumping and splitting, and sometimes even result in more uncertain estimates than splitting. This means that information-sharing can vary considerably and the actual characteristics of the direct and indirect evidence may dictate how strongly mixture priors will actually share information. This finding is in line with previous work by Roever et al., 2019 which suggested that mixture priors are robust to *'prior data conflict'*, meaning that when direct and indirect evidence are substantially different, mixture priors do not share any information across the two sources.

Figure 7.6 explores further the features of mixture prior models. The top graph in Figure 7.6 plots for each simulation the PED-ratio against the actual difference between simulated direct and indirect point estimates. Informative priors (red), always impose the same degree of information-sharing as lumping regardless of how distant the simulated direct and indirect evidence is; hence PED-ratios are always close to 1. In contrast, mixture priors impose a degree of sharing that depends on the actual difference between the two evidence sets. When the actual difference is low (i.e. there is agreement between the two evidence sets), mixture priors behave like informative priors, imposing maximum information-sharing. However, as the actual difference increases, the degree of sharing that is imposed by mixture priors rapidly falls, up to the point where no information is borrowed at all. Regarding precision, in this scenario informative priors always yield more precise estimates than splitting (i.e. Splitting SeR $\geqslant 1$), reducing the width of credible intervals by around 30%[6]. In contrast, mixture priors offer precision gains when the direct and indirect evidence is similar, but these decline as the two sources become more distant. Interestingly, this relationship is not strictly decreasing, and the model is 'borrowing weakness' (i.e. leads to less precise estimates than no borrowing at all) up until the point that the two sources become distant enough to prevent any information-sharing.

[6] The informative prior reduces the standard error by $\approx$ 30% because $\frac{se_{sp}}{se_{inf-prior}} = 1.4$, hence $\frac{se_j}{se_{sp}} = 0.714$.

Figure 7.6: *Actual difference between direct and indirect LOR plotted against the PED-ratio (top) and Splitting SeR (bottom) for mixture priors under FE.*

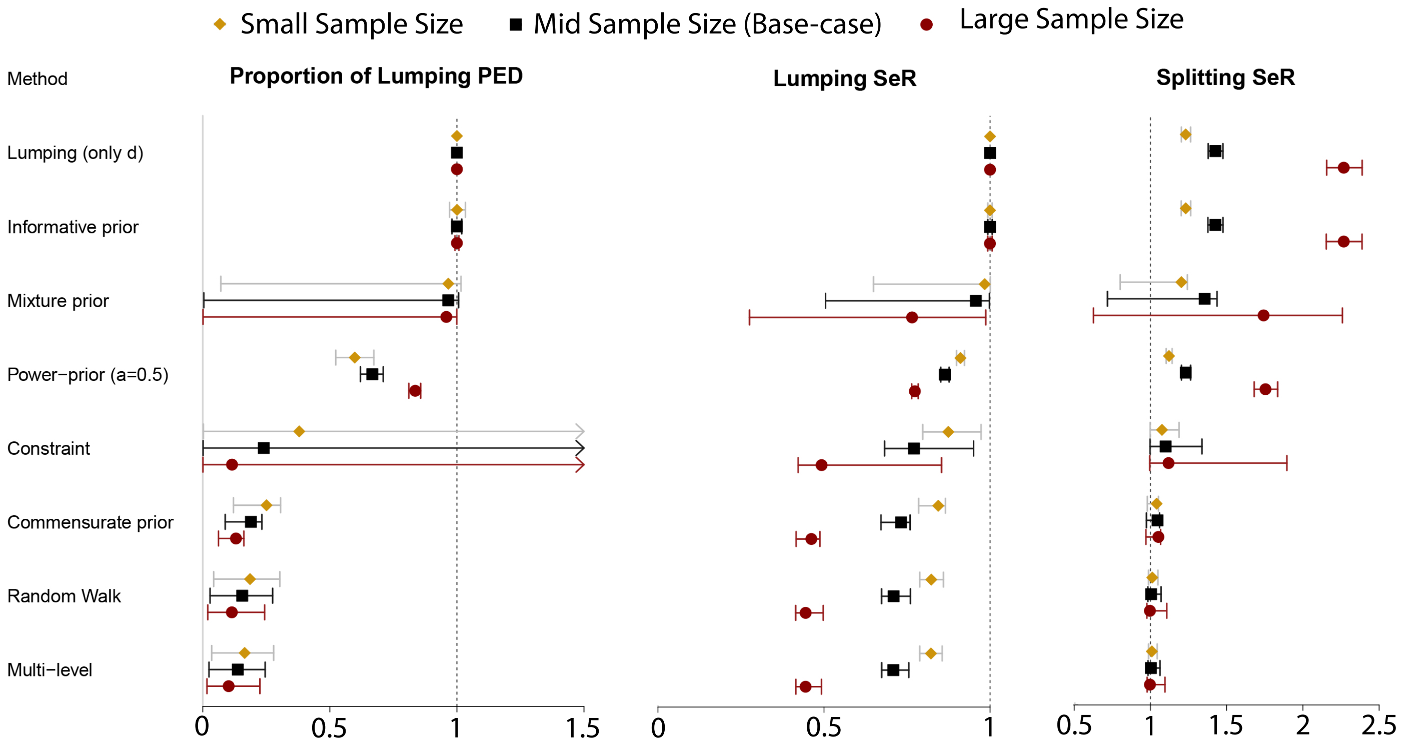


In the top, a scatter plot with the actual difference between direct and indirect evidence mean LORs (when analysed alone using a FE model) against the ratio of the PED that is imposed by a given method and the PED that is imposed by lumping. The red points pertain to the informative prior, and the black to the mixture of prior. In the bottom, the same graph is displayed for Splitting SeR i.e. the ratio of splitting's standard error of RTE to the corresponding method's standard error.

7.3.1.2 Alternative scenarios

With respect to the scenarios that assumed different levels of heterogeneity of the indirect evidence, Figure D.2.1 (see Appendix, page 247) illustrates that heterogeneity does not significantly influence the degree of sharing that the various methods impose. Contrariwise, for different sample sizes of the indirect evidence, Figure 7.7 shows that some measures are affected. In particular, across all models, as sample size increases, Lumping SeR decreases, suggesting that methods offer less increase in precision than what would be achieved by lumping. This is not surprising given that lumping imposes a much stronger assumption compared to most methods, and as the sample size of the indirect evidence increases, more information is contained in the indirect evidence, and lumping also borrows more strength. This can be observed in the Splitting SeR plot, where lumping achieves the largest increase in precision gains as sample size increases.

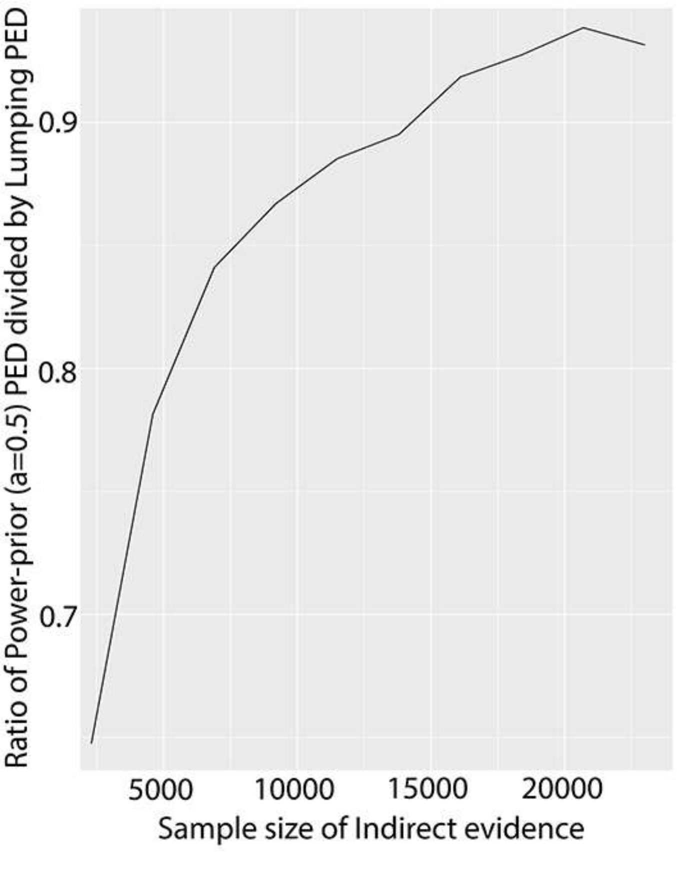
Figure 7.7: *Forest plots of PED-ratios, Lumping SeR, and Splitting SeR across all methods for three scenarios with different sample size of the indirect evidence.*



The 'Base-case scenario' corresponds to a sample size of 2300, the 'small sample size scenario' to a sample size of 2300/2, and the 'large sample size scenario' to a sample size of 2300 ∗ 4. The 'Proportion of Lumping PED' (i.e. PED-ratio) reflects a methods PED divided by lumping's PED. 'Lumping SeR' is the ratio of a method's standard error divided by lumping's standard error. 'Splitting SeR' is the ratio of a method's standard error divided by splitting's standard error.

Regarding PED-ratios, power-priors seem to be more sensitive to the sample size changes implemented in these scenarios. In fact, the results suggest that as sample size increases, a power-prior model with a given α value (here $\alpha = 0.5$) shares information more similarly with lumping. This result motivated a further post-hoc short simulation where for 10 different equidistant sample sizes between 2,300 and 23,000 the power-prior model with $\alpha = 0.5$ and lumping were applied and compared. As shown in Figure 7.8 the same pattern is observed and increasing sample size of the indirect evidence increases the proportional sharing of the power-prior model in the PED-ratio. In other words, as the sample size of the indirect evidence increases, the strength of sharing of the power-prior becomes more similar to that under lumping even though the value of α remains unchanged. This finding further supports that interpretation of α is not straightforward.

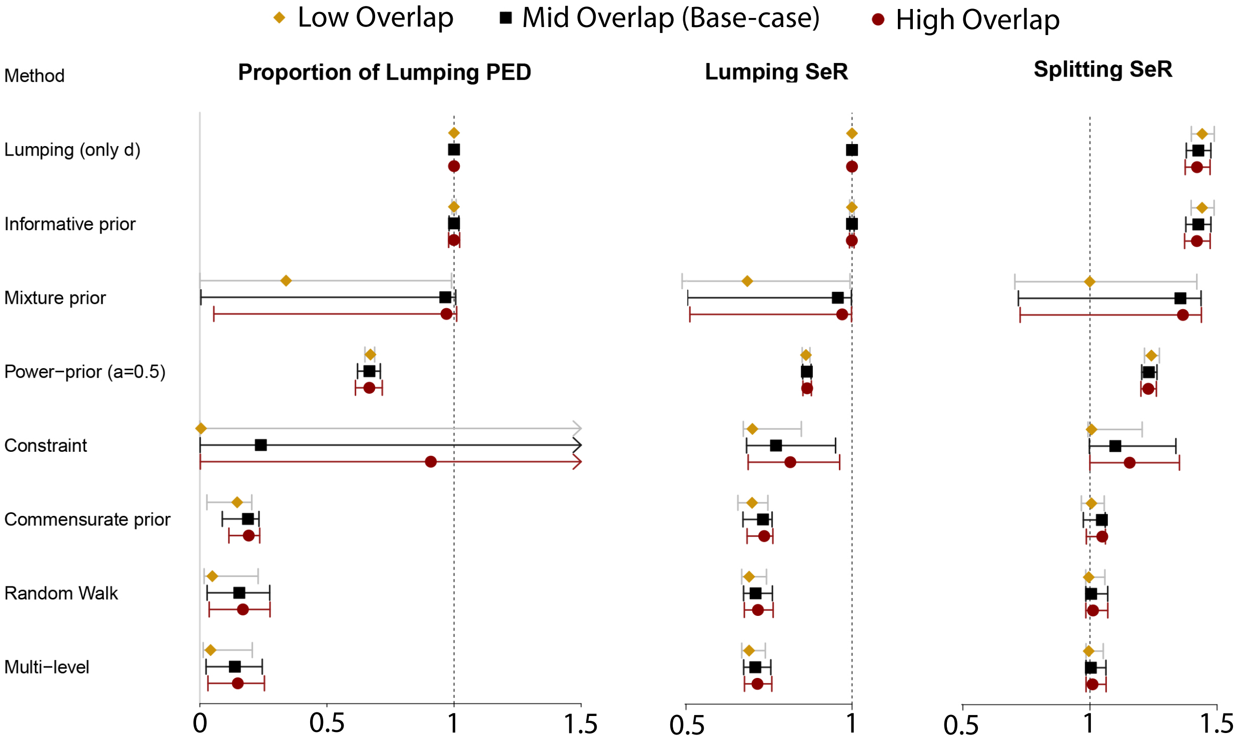
Figure 7.8: *The relationship between the degree of sharing imposed by the power-prior with $\alpha = 0.5$ and sample size of the indirect evidence for PED.*



The ‘Proportion of Lumping PED’ (i.e. PED-ratio) reflects the ratio of power-priors’s PED divided by lumping’s PED.

Finally, the extent of overlap between direct and indirect evidence seems to primarily affect constraints and mixture priors (Figure 7.9). As expected, when the overlap is high, on average constraints impose a higher PED-ratio, retaining the potential to exceed lumping's PED. This is because in this scenario, simulated datasets more often 'violate' the direction of relative effects suggested by the scenario characteristics. In these cases the need to conform to the direction dictated by the constraint —which is the opposite of the direction suggested by the data —leads to a major shift in the direct and indirect evidence means. This issue was also discussed in Figure 7.3. Regarding mixture priors, when direct and indirect evidence overlaps *minimally*, mixture priors yield very low PED-ratios and on average do not offer any precision gains from splitting. On the other hand, when direct and indirect evidence overlaps *majorly*, mixture priors are effectively equivalent to the informative priors and to lumping. This finding further supports the argument developed on page 170 regarding the robustness of mixture priors to 'prior data conflict'.

Figure 7.9: *Forest plots of PED-ratios, Lumping SeR, and Splitting SeR across all methods for three scenarios with different overlapping coefficient between direct and indirect evidence.*



'Base-case scenario' corresponds to an overlapping (OVL) coefficient of 50%, 'Low Overlap scenario' to an OVL of 5%, and 'High Overlap scenario' to an OVL of 75%. The 'Proportion of Lumping PED' (i.e. PED-ratio) reflects a methods PED divided by lumping's PED. 'Lumping SeR' is the ratio of a method's standard error divided by lumping's standard error. 'Splitting SeR' is the ratio of a method's standard error divided by splitting's standard error.

7.3.2. Random-effects simulation

7.3.2.1 Base-case scenario

As described in Section 7.2.4, under RE lumping can take different forms depending on which parameters of the direct and indirect evidence (d and τ) are pooled together. Table 7.5 shows sucra values for PED and PrI. In the base-case scenario, the method that only lumps the RTE mean —i.e. Lumping (d only) —shows a slightly higher PED and considerably lower PrI than the method that lumps all studies together under a single random-effect —i.e. Lumping (d & τ). Despite showing results for all lumping approaches, in this work the main lumping approach is Lumping (d only) because most methods only share information on the RTE mean; hence, they should be compared with the lumping approach that imposes the strongest assumption on the same parameter.

Table 7.5: *Sucra values for PED and PrI in the base-case scenario of the RE simulation.*

Information-sharing method	PED Sucra	PrI Sucra
Lumping (d only)	0.83	0.69
Lumping (d & τ)	0.8	0.83
Power-prior ($\alpha = 0.5$)	0.76	0.66
Informative prior —posterior —	0.75	0.82
Mixture prior —posterior —(p estimated in model)	0.64	0.64
Commensurate prior (p fixed at 0.5)	0.5	0.38
Random-walk	0.35	0.36
Informative prior —predictive —	0.33	0.35
Constraint	0.32	0.34
Mixture prior —predictive —(p estimated in model)	0.29	0.31
Multi-level	0.28	0.15
Lumping (τ only)	0.08	0.26
Prior on heterogeneity	0.04	0.2

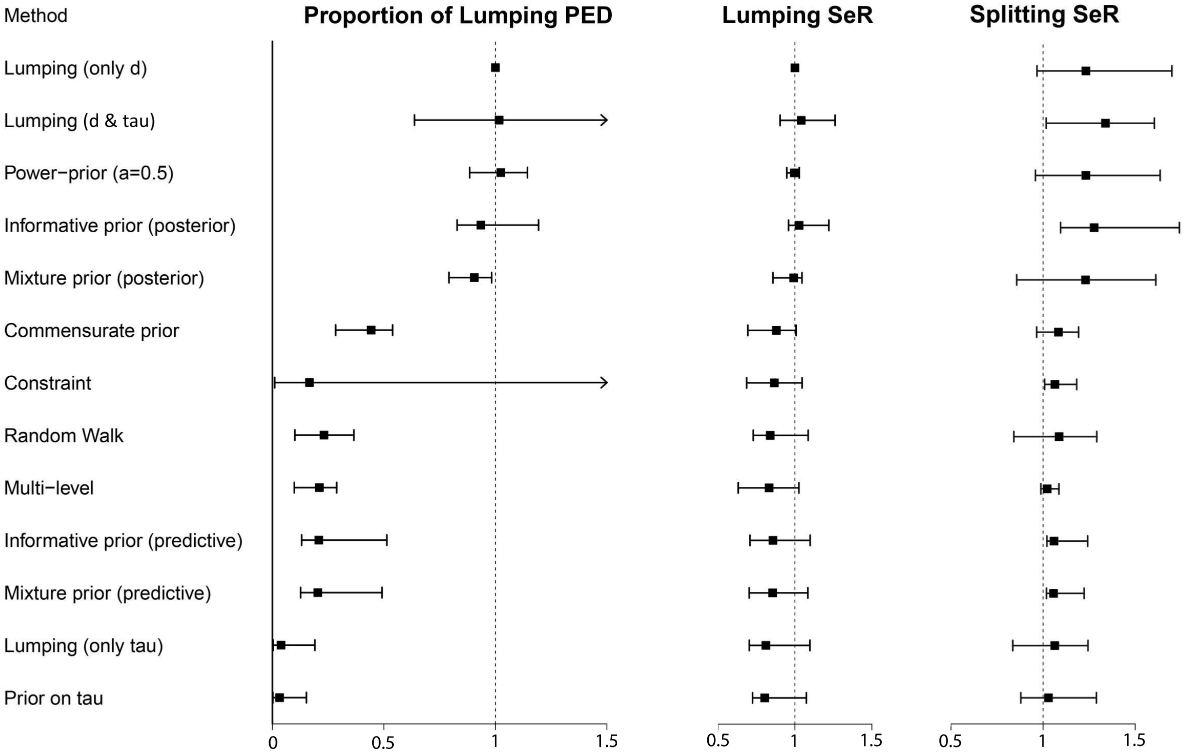
Methods arranged in descending PED Sucra values.

Contrary to the fixed-effect simulation, the informative prior that uses the posterior RTE mean distribution is not exactly equivalent to Lumping (d only) in terms of PED and PrI sucra values. However, the informative priors that use the predictive distribution of the RTE rank much lower in terms of both measures. The power-priors, despite using an $\alpha = 0.5$, share only slightly less information than lumping, and rank close to the top. Generally, as expected according to their assumptions, commensurate priors rank above random-walks, and informative priors above the mixture priors. Multi-level models seem to rank lower in terms of PrI than in terms of PED. However this feature should not be over-interpreted because this model imposes a random-effect at the top level only on two parameters. This may lead to considerably uncertain conclusions, and accordingly it should not be generalised to situations where more than two indirect evidence sets are available. Finally, methods that share information only on the heterogeneity rank last, which is not surprising particularly given that direct and indirect evidence is similarly heterogeneous in the base-case scenario.

Rankograms (see Figure D.2.2 in the Appendix, page 248) show that the PED and PrI are broadly in agreement with the exception of constraints. This was discussed in the previous section, and is here also attributed to simulations were the direction of relative effects is reversed between direct and indirect evidence. Compared to FE, under RE, methods seem to fluctuate more between similar ranking positions. This is however not surprising given that more methods can be used under RE leading more easily to ranking changes and to increased ranking uncertainty.

Figure 7.10 shows how the various methods compare to splitting and lumping (only d) for the base-case scenario. Regarding PED, methods can be grouped into two groups: those which yield comparable PED with lumping (top of the graph), and those which lead to much lower PED (bottom of the graph). Within each group, methods are similar in their PED-ratio means but less similar in terms of the PED-ratio credible intervals. For instance, across datasets, the mixture prior that uses the predictive distribution of the indirect evidence may impose a PED that is between 10% and 50% of the PED of lumping. It is worth further noting that in the group that imposes similar PED with lumping, some methods have the potential to yield higher PED that lumping (i.e. their credible interval spans above the dotted line at 1). Further investigation showed that for lumping (d & τ) and power-priors such cases primarily arise when the simulated direct and indirect evidence are very similar in terms of their point estimate (see Figure D.2.3 in Appendix). In these cases, lumping the two sources results in practically the same estimate for the mean and hence the PED of lumping is almost zero. Therefore, when we take the ratio of a method's PED to the PED of lumping (only d), the resulting number is enlarged because the denominator tends to zero. With respect to precision, methods that shared less strongly on PED, also yield, on average, less precise estimates than lumping (i.e. their median Lumping SeR is on the left of the dotted bar). No single method can be distinguished since their credible intervals majorly overlap. However, it needs to be highlighted that the majority of methods have the potential to produce more precise estimates than those under lumping, albeit infrequently. Finally, based on the splitting SeR graph, all methods on average result in precision gains when compared to splitting (i.e. their median Splitting SeR is on the right of the dotted bar), but mixture priors, random-walks, and methods that share only on the heterogeneity parameter, have significant potential to yield precision losses.

Figure 7.10: *RE simulation. Forest plots of PED-ratios, Lumping SeR, and Splitting SeR, across all ISMs used in the base-case scenario.*

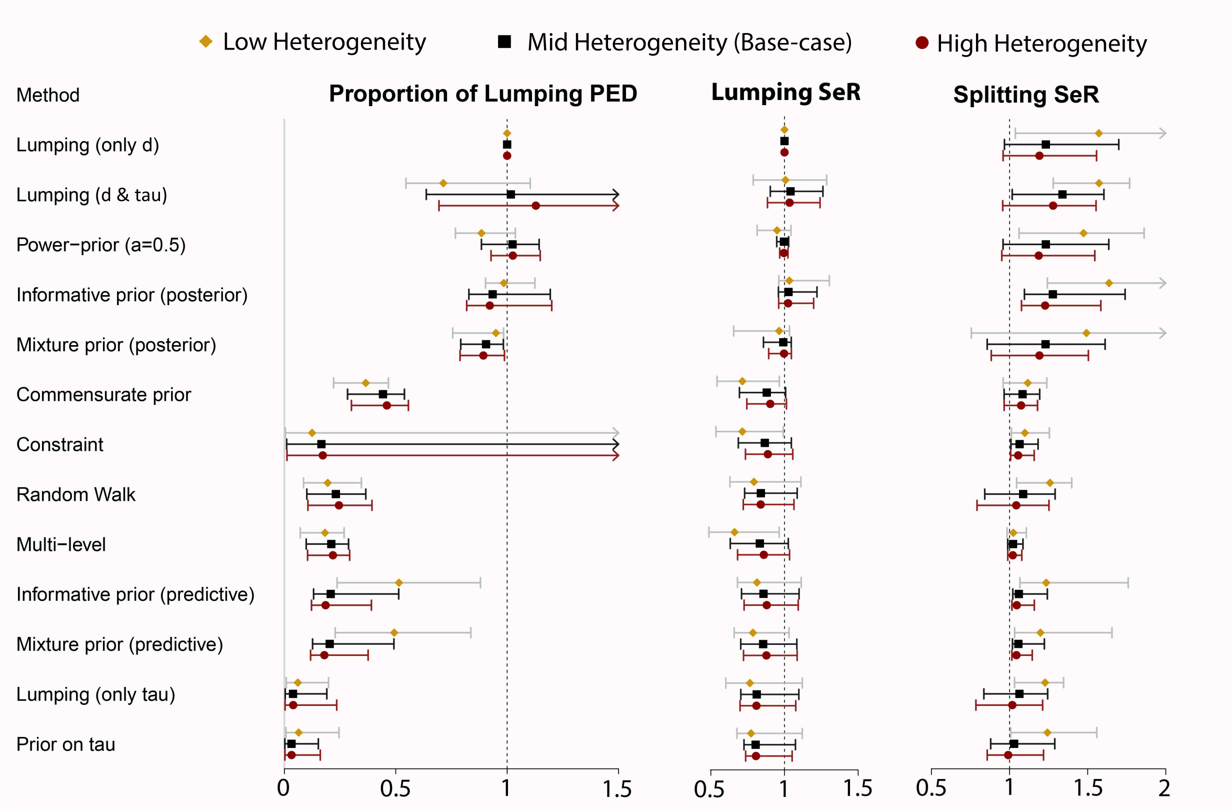


'Proportion of Lumping PED' (i.e. PED-ratio) reflects a methods PED divided by lumping's PED. 'Lumping SeR' is the ratio of a method's standard error divided by lumping's standard error. 'Splitting SeR' is the ratio of a method's standard error divided by splitting's standard error.

7.3.2.2 Alternative scenarios

The sample size of the indirect evidence does not influence how methods compare to lumping (Figure D.2.4 —Appendix). Heterogeneity primarily affects prior-based methods that use the predictive distribution and lumping (d & τ) (Figure 7.11). In particular, informative and mixture priors using the predictive distribution yield lower PED-ratios as heterogeneity increases, because the predictive distribution becomes more vague (i.e. a less informative prior). In contrast, lumping (d & τ) yields higher PED-ratios for increasing heterogeneity. This is because, as shown in Figure D.2.5, lumping (only d) yields a reduced PED for increasing heterogeneity, while lumping (d & τ) remains unaffected; hence, only the denominator of the PED-ratio is reduced and the PED-ratio increases. According to Splitting SeR, heterogeneity impacts how methods compare to splitting, with increasing heterogeneity leading to more moderate precision gains.

Figure 7.11: *RE simulation. Forest plots of PED-ratios, Lumping SeR, and Splitting SeR across ISMs used in the three heterogeneity scenarios.*

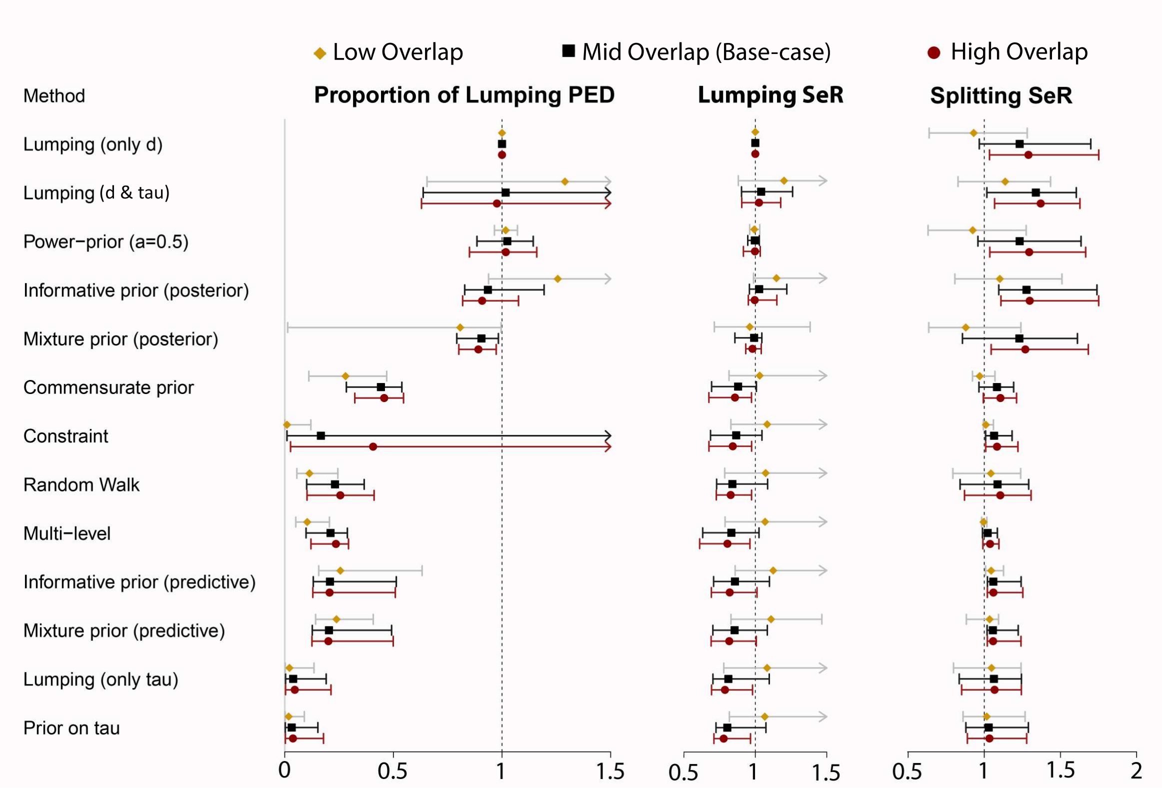


'Base-case scenario' corresponds to a $\tau = 0.56$, 'Low Heterogeneity scenario' to $\tau = 0.24$, and 'High heterogeneity scenario' to $\tau = 0.65$. The 'Proportion of Lumping PED' (i.e. PED-ratio) reflects a methods PED divided by lumping's PED. 'Lumping SeR' is the ratio of a method's standard error divided by lumping's standard error. 'Splitting SeR' is the ratio of a method's standard error divided by splitting's standard error.

Percentage overlap also seems to affect how methods compare to lumping and splitting (Figure 7.12). Regarding PED, for low overlap, lumping (d & τ) and the informative prior that uses the posterior mean of the relative effect of the indirect evidence, on average lead to larger changes in the point estimate than lumping (only d). This is not observed, however, when overlap is higher. In addition, constraints yield proportionately larger PEDs as overlap increases because of the phenomenon described in Figure 7.3. For low overlap, just as in the fixed-effect simulation, mixture priors again show the capacity to share information in the whole range between splitting and lumping; this characteristic was previously attributed to robustness of mixture priors to 'prior data-conflict'. Interestingly, Lumping SeR figures show that for low overlap on average most methods result in more precise estimates than lumping (i.e. the median Lumping SeR is beyond the dotted line), whilst for higher percentage overlap lumping is more precise.

This suggests that when direct and indirect evidence have very different RTE point estimates, Lumping (only d) is less likely to produce precision gains copmpared to other methods. According to Splitting SeR and Figure D.2.6, lumping (only d), power-priors, and mixture priors can lead to estimates more uncertain than splitting when the two evidence sets are far apart.

Figure 7.12: *RE simulation. Forest plots of PED-ratios, Lumping SeR, and Splitting SeR across all ISMs used in the three percentage overlap scenarios.*



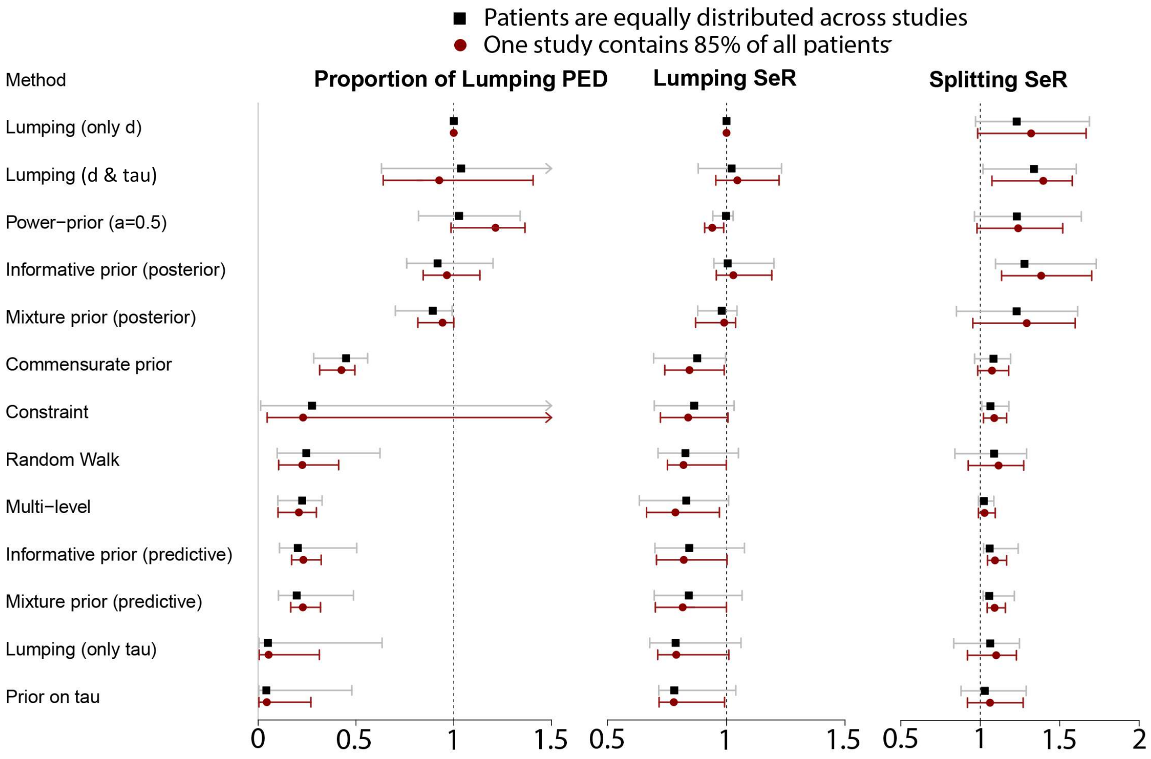
'Base-case scenario' corresponds to an overlapping (OVL) coefficient of 50% (LOR:-0.13, OR:0.87), 'Low overlap scenario' to OVL 5% (LOR: 0.45, OR: 1.57), and 'High overlap scenario' to OVL 75% (LOR: -0.29, OR: 0.74). The 'Proportion of Lumping PED' (i.e. PED-ratio) reflects a methods PED divided by lumping's PED. 'Lumping SeR' is the ratio of a method's standard error divided by lumping's standard error. 'Splitting SeR' is the ratio of a method's standard error divided by splitting's standard error.

In the last scenario, the impact of the distribution of indirect patients across the indirect studies was tested. The reduced variance in all measures that is observed in Figure 7.13 should not be over-interpreted since it is attributed to the way that data was simulated for this scenario[7]. Overall, results do not differ considerably across scenarios for most

[7] Recall that in this scenario the relative effects for the big study are simulated from a small part of the right tail of $N(-0.281, 0.56^2)$; hence draws do not differ much from one another. Also, all the small studies are assumed to exhibit equal relative effects, therefore the simulated indirect datasets are quite similar.

methods. However, this is not the case for power-priors. As speculated in Chapter 5, when power-priors are used to down-weight the effect of an indirect source of data, we need to consider the sizes of the indirect studies. In particular, when there is a study with a much larger size which also suggests a very different relative effect to the other indirect studies, using power-priors with small α values may lead to excessive borrowing from the big study. This is because for low α it is the only indirect study with a non-negligible Bayesian likelihood. As expected, this feature is also apparent in Figure 7.13, suggesting that in contrast to the base-case scenario, the power-prior yields a PED that is on average considerably larger than that of lumping.

Figure 7.13: *RE simulation. Forest plots of PED-ratios, Lumping SeR, and Splitting SeR across all ISMs used in the two scenarios where the indirect patients are differently distributed across studies.*



The 'Proportion of Lumping PED' (i.e. PED-ratio) reflects a methods PED divided by lumping's PED. 'Lumping SeR' is the ratio of a method's standard error divided by lumping's standard error. 'Splitting SeR' is the ratio of a method's standard error divided by splitting's standard error.

7.3.3. Summary of results

The main results of the simulation are listed below:

1. Across all scenarios evaluated, multi-level models, random-walks and commensurate priors impose relatively low degrees of information-sharing leading to relatively small changes in the point estimate and relatively low precision gains. For RE models, prior-based approaches that use the predictive distribution of the indirect evidence, and methods that share information only on heterogeneity fall into the same group. In contrast, power-priors and informative priors that use the posterior mean estimate of the indirect evidence can impose a much stronger degree of information-sharing, comparable to that of lumping.

2. Constraint models most often lead to moderate precision gains but have the potential to result in major shifts of the point estimate. This happens when the direction of direct and indirect effects that is dictated by the constraint (e.g. $d_{dir} \leqslant d_{indir}$) is different from the direction suggested by the available data (e.g. $d_{dir} \geqslant d_{indir}$).

3. Mixture priors can modify the degree of information-sharing that they impose based on the similarity of the direct and indirect evidence; hence, it can be thought of as an 'adaptive' method. When direct and indirect evidence are alike (i.e. with similar point estimates) information-sharing is encouraged and mixture priors essentially become standard informative priors. In contrast, when direct and indirect evidence is considerably different, information-sharing is discouraged with mixture priors becoming more like splitting and sometimes even leading to precision losses.

4. For power-priors, interpretation of α is not straight-forward. In particular, α should not necessarily be viewed as a proxy for the imposed degree of sharing, nor as how the power-prior compares to lumping. Furthermore, the relationship between strength of sharing measures and α may be non-linear. Finally, characteristics of the indirect evidence such as the number of patients in the indirect evidence set can affect how a power-prior model with a given α value compares to lumping (see Figure 7.8).

5. In general, under RE models, methods resulted in wider credible intervals for the various strength of sharing measures. This implies that the way that the methods relate to lumping and splitting (and hence the strength of the assumption that they impose) is more uncertain under RE models. In contrast, under FE, the ratio statistics that compared each method with lumping and splitting resulted in narrower credible

intervals; hence, the degree of information-sharing that each method imposes can be characterised more accurately under FE.

6. Under FE, lumping most often imposes the strongest information-sharing and only the constraints have the capacity to impose a larger degree of sharing under specific circumstances. In contrast, under RE, lumping (d & τ), power-priors, and informative and mixtures priors that use the posterior mean of the indirect evidence can lead to larger PEDs than lumping. Under RE, most methods also have the capacity to lead to higher precision gains than lumping.

7. Under FE, informative priors that use the posterior mean distribution of the RTE are equivalent to Lumping (only d). In contrast, under RE informative priors that use the posterior mean distribution of the RTE can impose a stronger degree of information-sharing than Lumping (only d).

8. Under FE, power-priors with $\alpha = 0.5$ impose strength of sharing slightly above 50% of that of lumping across all metrics, with very narrow credible intervals on sampling uncertainty. As a result, power-priors with $\alpha = 0.5$ never exceed the strength of sharing of lumping. In contrast, under RE, power-priors with $\alpha = 0.5$ yield large credible intervals for all strength of sharing measures implying that they do not relate to lumping in a specific manner across samples. Furthermore, across samples, power-prior models with $\alpha = 0.5$ impose on average a similar degree of sharing with lumping.

9. Under FE, sample size impacts on how methods compare to lumping and splitting with increasing sample size rendering power-priors more similar to lumping in terms of PED. Also, as sample size increases most methods yield proportionately smaller precision gains compared to lumping. In contrast, sample size barely seems to exert any influence on how methods compare to lumping and splitting under RE models.

10. Under FE, the tested levels of heterogeneity do not impact how methods compare to lumping and splitting. However, under RE some methods are affected. These are the methods that share information on both the RTE mean and on the heterogeneity parameter.

7.4. Discussion

In this chapter, a simulation experiment was conducted to investigate how alternative methods share information between direct and indirect evidence. The characteristics of the indirect evidence were varied (in relation to the direct) in a number of scenarios. This work focused on quantifying the degree of information-sharing and explored how this may be influenced by particular characteristics of the indirect evidence. Direct and indirect datasets were generated according to scenario characteristics and these were subsequently synthesised using all the applicable ISMs that were introduced in Chapter 4 (reduced to the pairwise meta-analysis case). The simulation was run twice; once under FE and once under RE models to ensure that potential differences between the two types of models are identified. To my knowledge, this is the first attempt to compare ISMs in terms of the strength of sharing that they impose in a simulation context.

Overall, results were broadly consistent with expectations. Under FE models, three distinct categories are observed. The first is methods that strongly share information across all scenarios explored and includes lumping and informative priors; the second is methods that only minimally shared information and include random-walks, multi-level models, and commensurate priors[8]. The last category comprised of more flexible methods (mixture priors, power-priors, constraints) that can share information in an adaptive manner depending on the characteristics of direct and the indirect evidence sets. Under RE, this classification becomes less clear because the various strength of sharing measures result in large credible intervals. This indicates that, across simulations, it harder to determine how each method compares to lumping and splitting.

In the base-case scenario, we saw that constraints have the potential to be considerably more influential than lumping. This happens when the two evidence sets suggest a direction for the relative effects which is not in accordance with the direction specified by the constraint. This implies that when separate analyses of direct and indirect evidence indicate that the direction of their relative effect means is the opposite from the direction which is assumed by the constraint, we can expect constraints to lead to considerable shifts in the relative effect means and hence to very strong information-sharing. It should be noted that if there is a robust biologic or clinical rationale to support an a priori expected ordering of effects, constraints may be used even if the expected ordering is not reflected in the available evidence. This is because the existing state of the evidence may

[8] Fixed 50% weights in the spike-and-slab hyper-prior were used for the commensurate priors and therefore we did not allow the method to show its potential to borrow strength in an adaptive manner. There is a chance that commensurate priors can also flexibly share information if the aforementioned weight is considered an uncertain parameter and hence estimated within the model. Therefore commensurate priors may also belong in the last category.

be due to extensive sampling error or heterogeneity across studies.

Mixture priors also exhibited interesting features. In particular, in line with previous findings by Roever et al., 2019, mixture priors were robust to 'prior data conflict'. In this work, it is shown that when direct and indirect evidence are indeed similar, mixture priors behave like regular informative priors and borrow strength strongly from the indirect evidence. However, when direct and indirect evidence are not similar (i.e. are in disagreement or in conflict), mixture priors impose minimal information-sharing, if any. This implies that mixture priors offer an adaptive degree of borrowing which may be a desirable property when the appropriate degree of borrowing is unknown. However, it should be highlighted that the way mixture priors determine the degree of borrowing is not necessarily transparent and it may be preferable to retain control of the strength of sharing than allowing it to determined by the model.

Power-priors, which allow the analyst to specify the extent to which the likelihood of the indirect evidence is discounted, also produced some interesting findings. Specifically, the interpretation of α is not directly associated with the extent of information-sharing as that is measured by the metrics used in this chapter. For instance, under FE, the power-prior model with $\alpha = 0.5$ does not share half as much as lumping in terms of any of the metrics used, but more around 60-70%. Furthermore, when the number of patients in the indirect evidence set increases, for the same α, the power-prior imposes larger degree of sharing and is hence more similar to lumping. Under RE, even for $\alpha = 0.5$, on average, the power-prior imposes the same strength of sharing as lumping and often even more. Given the difficulty in relating the α value to the degree of sharing the implication is that it may not be easy to describe α in a straightforward manner that allows it to be obtained from experts in structured elicitation exercises.

The additional scenarios aimed to identify characteristics of the evidence base that may affect how methods compare to lumping. In general, increasing the level of heterogeneity of the indirect evidence can lead to reduced information-sharing for prior-based methods that use the predictive distribution of the indirect evidence and for methods that share information directly on the heterogeneity component. Also, increasing the sample size of the indirect evidence set can result in power-priors, with a given α value, producing results more similar to lumping. Finally, increasing the percentage overlap (i.e. similarity) between point estimates of the direct and indirect evidence sets can considerably increase the degree of information-sharing imposed by constraints and mixture priors.

This work has a number of limitations. First of all, despite that indirect evidence were simulated according to the properties of various scenarios, the same was not the case for the direct evidence which were all simulated based on the same properties. However, what is critical is how the direct and indirect evidence sets compare to each other, and not so much the absolute values of their underlying parameters. This is because the strength of sharing measures (PED, PrI) only consider the relative nature of the two evidence sets; therefore, when the absolute values of the characteristics of the two sources evidence change but their relative difference is preserved, we expect to observe very similar results in terms of strength of sharing. Second, instead of varying dimensions factorially, dimensions were varied here 'one-by-one' to restrict the number of scenarios and the time required in the computing cluster (which already required 20,000 hours). As a result, it cannot be determined how methods strength of sharing is affected by the interaction of dimensions. Finally, only three dimensions were varied (the percentage overlap between direct and indirect evidence, the heterogeneity of the indirect evidence and the sample size of the indirect evidence) and hence there may have been other dimensions could influence methods ranking but were omitted from this study.

Further work could seek to address the limitations of this experiment by potentially exploring more dimensions and varying them factorially to observe the effect of dimension interactions. In addition, researchers could further explore how the characteristics of the evidence make power-priors impose higher or lower degrees of information-sharing and how the value of α should be appropriately interpreted and elicited. Furthermore, in this simulation, only two evidence sets were used. This implies that methods such as the multi-level models cannot, by definition, impose strong degrees of sharing because in the top level they assume a random-effect on just two parameters. Hence, further work could investigate whether the inclusion of additional indirect sources renders multi-level models more comparable to lumping and how they may affect mixture of priors which would require more than two components. Finally, commensurate priors were used only with a fixed 0.5 weights on the spike-and-slab hyper-prior in order to assist parameter identification. This does not allow this method to show if it has the potential to identify differences in the two evidence sets and impose adaptive borrowing accordingly. Consequently, further work may try to compare the 'adaptability' of mixture of priors and commensurate priors.

Chapter 8

Discussion

8.1. Thesis summary

This thesis aimed to address issues that stem from evidence sparsity by using sophisticated evidence synthesis methodologies which facilitate information-sharing between direct and indirect evidence in order to strengthen inference. Chapter 1 provided the necessary background, introducing relevant concepts and highlighting how information-sharing problems have been dealt with to date, and Chapter 2 offered an introduction to the standard methods that are used in HTA for evidence synthesis and decision modelling. Subsequent chapters used this introductory material as a foundation in order to describe approaches that build on those standard methods. What follows in this section is a summary of the scope, main findings, and contributions of each of these chapters.

First, Chapter 3 sought to systematically identify methods that have been used in the biostatistics literature to combine evidence directly and indirectly relating to a research question. To my knowledge, this topic has not been previously reviewed. The identified methods were classified according to the main assumption each method makes to facilitate information-sharing into four distinct 'core' categories: 1. functional relationships, 2. exchangeability-based relationships, 3. prior-based relationships, and 4. multivariate relationships. This classification highlights that there are several alternative methodological options that can facilitate information-sharing and also provides a structured way of thinking around such methods. Papers were categorised according to the main synthesis challenge that they dealt with, allowing researchers faced with specific synthesis issues to get insight into the main contributions in each particular synthesis issue. A further categorisation of papers according to the PICOS level that the external evidence were indirect to was provided in order to reveal specific information-sharing patterns and areas for methods development. This categorisation showed that some types of relationships have not been used to share information on specific PICOS levels. This suggests that there is scope for extending methods that have been used to share information on one PICOS level to other PICOS levels. A non-technical description of the various methods and papers was given, providing a plethora of citations for readers who may want to delve deeper into a particular topic.

Subsequently, in Chapter 4, methods identified in Chapter 3 were adapted to a specific synthesis problem which is common in HTA i.e. the borrowing of strength from aggregate-level evidence pertaining to an indirectly relevant population. Thorough mathematical descriptions and explanations of the different methods were provided along with coding suggestions. Consequently, researchers facing similar synthesis issues can consult this chapter for both a theoretical understanding of the various ISM options and practical suggestions for their implementation. Furthermore, a step-by-step framework that allows the systematic identification of applicable ISMs was introduced. This framework can be used by both researchers who seek to find methods applicable to their own synthesis issues and by appraisers who need to ensure that no applicable method is unjustifiably excluded. The work of this chapter provided the necessary foundation upon which the ISMs used in the subsequent applied chapters were based.

Chapter 5 was an applied case-study where strength was borrowed from paediatric patients to inform inferences in adult patients. In addition to its main contribution —the illustration of the impact of using different ISMs on relative effectiveness estimates —this chapter also suggested methods for exploring heterogeneity in the extended evidence base, and on implementing ISMs when covariate effects need to be considered. Several measures were used to understand how strongly the various methods borrowed strength from the indirect evidence, and how they compared with the two extremes (i.e. lumping and splitting). This work showed that different methods can influence RTE estimates in different ways, such as by modifying point estimates and/or affecting precision, and to different magnitudes. This implies that the use of different methods may lead to different conclusions for HTA, and therefore highlights the need to systematically consider several alternative options. Furthermore, this work illustrates that an information-sharing *'spectrum'* can be defined based on the two extremes (i.e. lumping and splitting) and ISMs can generally be mapped within that spectrum, except for specific situations. Such exceptions include RE power-priors with $\alpha < 1$ which can impose a degree of information-sharing that is higher than that of lumping under models. This finding implies that the characteristics of the indirect studies should be carefully examined before applying power-priors, particularly. The findings of this chapter motivated Chapter 7 in which the methods' properties and imposed strength of sharing were investigated in further detail.

Chapter 6 considered the application of the RTE estimates, which were produced in Chapter 5 using the various ISMs, into an existing decision model. In this way, this chapter revealed the implications of using different ISMs for policy-making, considering both the impact on adoption decisions (approve/reject), and on further research recommendations. The results highlight that the choice of method can result in different policy recommendations, further demonstrating the need for transparency in the choice of ISMs.

Finally, Chapter 7 described a simulation that sought to understand how ISMs compare with lumping and splitting, and which characteristics of the evidence can influence this relationship; all methods explained in Chapter 4 were applied. Given that the results of Chapter 5 suggested that ISMs may perform differently under FE and RE models, two simulations were run (one under FE and one under RE). The results of Chapter 7 can be used by analysts trying to determine the extent of information-sharing that each method imposes in specific evidential scenarios. However, establishing how methods compare to each other —and to lumping and splitting —may be easier under FE than under RE models. Methods could be broadly categorised into three categories: those that generally shared information only weakly resembling splitting, those that generally shared information strongly resembling lumping, and those that could alter their strength of sharing based on the state of the direct and indirect evidence. ISMs sharing information weakly included multi-level models, random-walks, commensurate priors, methods sharing only the heterogeneity parameters, and informative/mixture priors utilising the predictive distribution of the indirect evidence. ISMs sharing information strongly included informative priors using the posterior mean RTE of the indirect evidence and power-prior models with high α values; finally, the only adaptive method was the mixture prior which had the potential of behaving either like lumping or like splitting depending on how similar the direct and indirect evidence were i.e. whether or not there was 'prior-data conflict'. In addition, constraints were found to have the potential to share information more strongly than lumping. This implies that constraints can be very informative and researchers should ensure that there is a robust scientific rationale before imposing such models since they may considerably affect both RTE and cost-effectiveness estimates. Finally, results showed that the interpretation of α in the power-priors is not straightforward as the degree of sharing imposed by a power-prior with a given α might be influenced by the characteristics of the direct and the indirect evidence. Strikingly, under RE, even power-prior models with low α were very often similar to lumping. This implies that the power-prior model should not be viewed as a method that allows the analyst to specify the extent of information-sharing, as the choice of α is not related to the imposed information-sharing in a straightforward manner. Therefore, analysts seeking to elicit α should first determine how α relates to the extent of information-sharing under the conditions of their extended evidence base, and should proceed with the elicitation only if this relationship is clear.

8.2. Recommendations

This section details a set of recommendations for the use of ISMs in HTA practice. These include 1. situations during which the use of indirect evidence may be useful which are based on the general motivation for information-sharing; 2. ways of identifying and selecting ISMs which are based on the general theory behind ISMs; 3. advice for the specification and the implementation of ISMs which are based on the findings of Chapter 7 and could also be useful for the application of ISMs beyond cost-effectiveness analysis and HTA research; 4. suggestions for presentation and reporting of the results of the various ISMs which are based on general principles for good HTA practice and finally, 5. recommendations for further research which are based on remaining uncertainties and realisations made through the thesis.

Situations when indirect evidence may be valuable

In principle, when making a decision all relevant evidence should be considered. This implies that indirect evidence, when available, should always be included to appropriately reflect uncertainty and ensure prioritisation of efficient research. However, this may not be practical in cases when information-sharing is technically challenging, or may be deemed unnecessary when the direct evidence base is sufficiently robust to inform a decision, or when indirect evidence would be unlikely to impact a decision.

Undoubtedly, to even start contemplating using indirect RTE evidence for decision-making three fundamental requirements need to be fulfilled. First, indirect evidence must be relevant to our decision (so that information-sharing is plausible) and must provide information beyond what is provided by the direct evidence (so that information-sharing has the potential to be meaningful). Second, relative effectiveness needs to be a key parameter for cost-effectiveness, so that changes in RTE estimates can affect decisions. Third, indirect evidence should only be considered when we expect that it will have important implications for the RTE estimate of interest. For cases in which the aforementioned requirements are fulfilled, a list of some specific examples where using indirect evidence may be beneficial for decision-making is provided below:

1. When the direct evidence is sparse, leading to imprecise estimates for relative effectiveness, and there is an indirect source of information with a richer evidence base and well characterised RTE estimates.

2. When the direct evidence is of low quality (e.g. there is only observational evidence or RCTs in high risk-of-bias) and its internal validity is questionable —giving rise to bias considerations —, and there is an indirect source of information comprising

studies of better quality e.g. high quality RCTs.

3. When heterogeneity cannot be appropriately explored because the majority of the direct studies do not report the suspected effect modifier, and there is an indirect source of information thought to be influenced by the same effect modifier to a similar extent, and the studies comprising it report complete information.

4. When there is inconsistency between relative effects suggested by direct and indirect evidence even though both sources are thought to comprise of good quality studies and relevance of the indirect evidence is established. This may indicate that some important characteristic of the evidence has been overlooked and/or that heterogeneity has not been appropriately explored.

Identifying and selecting information-sharing methods

1. Analysts should be aware that lumping —despite being the simplest available —is not the only information-sharing option. There are a plethora of methods that use different assumptions, often more moderate, imposing various degrees of information-sharing.

2. The process of identifying applicable ISMs should be transparent and the exclusion of models should be adequately justified based on model assumptions and/or data requirements. The step-by-step framework proposed in Section 4.5 may be useful in this process.

3. ISM choice should not be solely guided by model 'goodness of fit'. This is because methods that impose stronger degrees of information-sharing may naturally yield higher residual deviance and DIC. Also, not all methods yield comparable DICs as some analyse direct and indirect evidence in separate steps, and hence less data are used in the final step.

Specification and implementation of information-sharing methods

1. Analysts should provide all the necessary details regarding the specification of the ISMs used, and where applicable, justify specific parameter choices. For instance, when a lumping approach is adopted it should be clear which parameters are 'lumped' (i.e. RTE mean and/or between-studies heterogeneity). Also, when informative/mixture priors are used for RE base-models, analysts should state whether they used the posterior mean distribution or the predictive distribution of the indirect evidence. In addition, when prior distributions are mixed (e.g. in mixture priors or commensurate priors), it should be clear whether the weights are specified by the analysts or estimated within the synthesis model. Finally, when power-priors are used the value of α should be explicitly stated.

2. Analysts should try where possible to implement several, if not all, applicable ISMs to understand whether decisions are robust to method choice. If the number of methods needs to be restricted, a judgement is required to determine the appropriate degree of information-sharing. If the extended evidence base is similar to one of the scenarios considered in Chapter 7, then the results of that chapter can then be used to identify methods that impose the desirable degree of sharing. Alternatively, if a judgement cannot be made, at least a model that is expected to impose a high degree of sharing, a model that is expected to impose a low degree of sharing, and a model that shares information in an adaptive manner should be implemented.

3. Analysts using power-priors should be careful not to interpret α directly as the degree of information-sharing. Under FE, the relationship between the sample size of the direct and indirect evidence is very important in predicting how the power-prior compares to lumping, and under RE the interpretation of α becomes very hard. Hence, given the nature of the α parameter (i.e. a power weight on a likelihood), α does not seem to always easily map to the degree of sharing in a predictable way.

4. If analysts and policy-makers seek to inform the degree of information-sharing using a structured expert elicitation process, they should be aware that none of the existing flexible models which include parameters controlling borrowing of strength can be easily mapped to the degree of information-sharing spectrum. One option may be to elicit the probability that information is completely transferable between direct and indirect evidence, and model average between lumping and splitting.

Presentation and reporting

1. The assumptions imposed by all ISMs should be interpreted in the context of the synthesis problem at hand, and the limitations of models including uncertainty regarding the imposed assumptions and strength of sharing should be recognised.

2. A summary of all the parameters information was shared on between direct and indirect evidence should be routinely reported along with the models used to facilitate information-sharing on each parameter and justifications for models excluded.

3. Presentation of results from using alternative ISMs should illustrate the impact on all aspects of relative effectiveness estimates (i.e. point estimate and uncertainty surrounding the mean relative effect) using measures such as PED and PrI, as well as on policy relevant quantities such as ICERs, and VoI parameters.

Remaining uncertainties and further research

1. Further research is required to determine the usefulness of ISMs developed in the trial-design field and their capacity to be extended to the MA/NMA field.

2. Further research is required to determine how indirect evidence can be systematically identified, and how information-sharing considerations can be explicitly incorporated in the comprehensive algorithm for approval of health technologies.

3. Further research should try to produce explicit guidance on how ISM choice should be conducted when the degree of information-sharing that each model imposes is unclear.

4. Further research should try to develop ISMs that include parameter(s) which describe the degree of sharing and could be readily elicited from experts.

8.3. Strengths of the thesis

This section provides a general overview of the main strengths of this thesis. More detailed discussions were presented in the final sections of each individual chapter.

First, this is the only work that compares a wide range of ISMs in terms of not only the estimates produced, but also the degree of information-sharing that they impose. In Chapter 5 all the applicable ISMs were used to borrow strength from an indirect evidence set. The resulting estimates were compared using several borrowing-of-strength measures which tried to capture —separately and simultaneously —all policy-relevant aspects of the estimates such as changes in the point estimate and its uncertainty.

Second, in contrast to most existing work on information-sharing which primarily focuses on the biostatistical aspects of information-sharing, this thesis has a clear policy focus. To that end, in Chapter 6, the RTE estimates produced by several ISMs were applied in a decision model developed as part of an HTA, and the implications of methods choice on both adoption and further research recommendation decisions were illustrated. Furthermore, the simulation undertaken in Chapter 7 makes the assumes that we do not know the precise relationship between the true RTE in the direct and the indirect evidence. Hence there is no way of knowing what the appropriate degree of information-sharing is, and classical performance measures —such as bias —cannot be used. In contrast, Chapter 7 compares ISMs with lumping and splitting, and therefore focuses on the imposed degree of information-sharing, as would be the case in policy-making. To date, this is the only simulation that has sought to compare a set of systematically identified ISMs, under several policy-relevant scenarios in terms of measures that retain their relevance for decision-making.

Third, despite the policy focus which is of primary interest here, the methodological developments detailed in this thesis are also useful beyond cost-effectiveness and HTA. For instance, this work is useful for regulatory bodies such as FDA and EMA, which in the process of licensing new interventions for specific patient groups (e.g. children) often face evidence sparsity issues and need to borrow-strength from existing evidence on other relevant populations (e.g. adults) (Food and Drug Administration and Center for Devices and Radiological Health, 2016; European Medicines Agency, 2016). Also, ISMs are relevant to the work undertaken by international research societies, such as the Cochrane collaboration, which is interested in producing guidance on the appropriate use of methods for health care research. Furthermore, the use of indirect evidence can be of interest to funding bodies such as the NIHR Evaluation, Trials and Studies Coordinating Centre (NETSCC) which is faced with the task of prioritising research proposals with minimal or no direct information on the relative effectiveness of the proposed comparison in the population of interest. In addition, appropriate methods for information-sharing can be useful for statisticians working in the trial-design field who may prefer to take account of existing indirect evidence when designing RCTs, and to pharmaceutical companies which could benefit by making use of relevant evidence in the drug-development process. Overall, the work undertaken in this thesis can be useful for anyone, within or outside health research, who seeks to inform decisions using a comprehensive evidence base in an efficient manner; yet, the relevance and the implications of the ISMs should be always be judged in the context of the field/decision-problem that they are applied to.

Fourth, while avoiding excessive focus on specific synthesis problems, this thesis provides a comprehensive list and categorisation of ISMs that are expected to prove useful for analysts. This was achieved by initially adopting a general perspective in Chapter 3, identifying ISMs used across several synthesis problems to enable information-sharing on many model parameters (e.g. relative efficacy, baseline event rate, between-study heterogeneity). This allowed information-sharing patterns to emerge, and enabled a succinct categorisation of ISMs into four 'core' types of relationships each one pertaining to a different type of assumption.

Fifth, in Chapter 4, this thesis 'translated' the methods and code that had been developed for a variety of synthesis problems to a single synthesis problem common in HTA. Mathematical descriptions have been provided alongside thorough discussions of the assumptions used by each method and the technical aspects of their implementation. Despite having been developed for a specific synthesis problem, most methods can also be used for other evidence synthesis problems with only minor modifications. As such, Chapter 4 can be a resource for analysts who seek to obtain a deep understanding of these methods and apply them in their synthesis projects.

Finally, this thesis aids transparency in ISMs choice. To date, there is no systematic process of identifying applicable ISMs and consequently, analysts are usually either unaware of the plethora of options available or prefer to use methods they know better over more complex methods that may potentially be more appropriate. The step-by-step methods identification process developed in Chapter 4 contributes towards the systematisation of ISMs choice. As such, it forces analysts to think through all potentially applicable options, and eliminate methods based on specific reasons (e.g. implausible assumptions). The need for a transparent process was also established in Chapter 5 and Chapter 6 where it was illustrated that different methods can impose varying degrees of information-sharing, and subsequently lead to different adoption and research prioritisation decisions. The current lack of knowledge in the scientific community on the use of appropriate ISMs renders this step-by-step process a necessary tool in assisting transparency in methods choice and consequently in decision-making.

8.4. Limitations and directions for future research

This section discusses the main limitations of this thesis and suggests specific types of research that might be undertaken to address them. The various suggestions are separated in two categories: 1. Methods-related research and 2. Policy-related research.

8.4.1. Methods-related research

First, even though this thesis included a thorough citation-mining review of ISMs in MA and NMA, other fields may also have used similar methods. For instance, in trial-design, relevant methodologies have recently been developed for the analysis of basket trials. These methods combine information from multiple patient subgroups without assuming completely homogenous or unrelated effects (Leon-Novelo et al., 2012; Neuenschwander et al., 2016; Fujikawa et al., 2020). Relevant methods have also been used in ecology (Poole and Raftery, 2000). Although these methods still fall under the existing four 'core' relationships, conducting a review that would seek to identify ISMs developed in non-health research fields may motivate the adaptation of more methods to MA/NMA. Conversely, adapting the ISMs described in this thesis to be suitable for more general use has the potential to provide considerable methodological developments in other fields, beyond HTA, that have historically been interested in the concept information-sharing.

Second, this thesis has only described methods that can be used to borrow strength from indirect studies conducted on a relevant population. This implies that direct and indirect evidence can be distinguished using a study-level covariate to which patients have not been randomised. Hence, the described methods are easily transferable to other cases where indirectness stems from another PICOS level as long as a similar variable can be used to distinguish the evidence sets. However, if a different variable needs to be used, methods may require modifications. For instance, if we wanted to relate the RTEs of different interventions to which patients have been randomised, we may need to adapt the methods to consider that direct and indirect evidence are distinguished with an arm-level variable. Also, when strength is borrowed from indirectly related outcomes, the methods developed in Chapter 4 have limited applicability, and instead multivariate methods that correlate the various outcomes, as well as methods that describe how outcomes are functionally related, are more useful. Finally, this work did not address more complex situations where there may be multiple indirect evidence sets pertaining to different PICOS levels, and therefore the combination of several evidence sets using perhaps different methods under the same synthesis model is a matter for further research.

Third, even though Chapter 7 compared several ISMs under a variety of conditions, further simulation experiments are required to obtain a deeper understanding of some high-level methods features. For instance, Chapter 7 only considered multi-level models in which the top level random-effect was imposed on just two group-specific basic parameters. Given that a potentially unrealistic variance is likely to be estimated for such a random-effect, it is not surprising that multi-level models were consistently ranked at the bottom of the information-sharing spectrum; that is closer to splitting (i.e. no sharing). In contrast, if more indirect evidence sets were included, the random-effect would be applied on more parameters, and perhaps multi-level models would result in stronger information-sharing. Therefore, the relationship between the degree of information-sharing and the number of indirect evidence sets requires further research. Furthermore, in Chapter 7 commensurate priors were not allowed to exhibit their 'adaptive' character because a fixed probability of 0.5 was used in the 'spike-and-slab' hyper-prior. Alternatively, the probability could be considered an uncertain parameter and estimated within the model, based on the similarity between direct and indirect evidence. Hence, simulation experiments can try to assess the commensurate priors' ability to modify the degree of information-sharing they impose under 'prior data conflict' conditions, and perhaps compare it with the mixture priors which were also found to be robust to prior data conflict. Such simulations would reveal whether any of the two adaptive approaches bear any benefits against the other.

Fourth, this work showed that the α likelihood weight in the power-priors does not have a straightforward interpretation, particularly under random-effect models, and does not necessarily link to the imposed degree of sharing in an intuitive manner. Similarly, none of the other flexible methods (i.e. mixture priors, commensurate priors) contain parameters that can intuitively be linked with the imposed degree of sharing. Therefore, developing methods that allow more straightforward mapping between a user-controlled parameter and the degree of information-sharing is warranted.

Finally, further research could describe how methods for structured elicitation could be used to allow experts' beliefs to determine the appropriate degree of information-sharing. Existing guidance (Bojke et al., 2019) suggests that elicitation should focus on simple observable quantities and therefore quantities elicited should at least have some intuitive interpretation. However, given that none of the existing ISMs contain parameters that can be intuitively linked to the extent of information-sharing, the development of elicitation methods is not straightforward and requires further research. In light of this, perhaps a solution would be to elicit the probability that information is completely transferable between the two sources —since this quantity retains some intuitive interpretation ISMs, and subsequently model average lumping and splitting based on that elicited probability. Alternatively, eliciting the probability that RTE parameters pertaining to the indirect evidence are exchangeable with those of the direct evidence may enable the use of partial exchangeability frameworks (e.g. Neuenschwander et al., 2016).

8.4.2. Policy-related research

Whenever evidence is collected with the ultimate purpose of informing a decision-making process, it is vital that all relevant direct and indirect evidence is utilised, so that the evidence base is comprehensive and decision uncertainty is appropriately reflected. However, in practice it may not always be easy to identify all relevant evidence or implement all applicable ISMs. This implies that in these situations decision-makers would need to make *value judgements* regarding the most appropriate/realistic course of action and do so in a transparent manner that makes judgements explicit. In what follows, areas for further policy research are outlined emphasising the required value judgements that would need to be made.

How should the use of indirect evidence be formally considered?

Despite most reimbursement bodies still making dichotomous decisions (approve or reject), existing algorithms suggest more options such as 'Only in Research' (OIR) and 'Approval with Research' (AWR) should be considered. Specifically, the algorithm suggested by McKenna et al., 2015 and Claxton et al., 2016 describes key principles and assessments required to inform reimbursement decisions and formally consider whether additional research is worthwhile, and can be conducted with/without approval.

Carrying out additional research requires valuable resources which could have been devoted to improving patients' care. It is therefore crucial to inform the decision-making process with the best available estimates. This implies that all relevant evidence (including indirect evidence) should be considered when deriving the required estimates because otherwise uncertainty may not be appropriately represented and decisions may be biased.

Furthermore, additional research often requires considerable time during which only a limited number of patients have access to the technology. However, different types of research require considerably different time and resources. For instance, in order to gain more information about the relative effectiveness of a technology, an expensive and time-consuming RCT could be conducted. Alternatively, if an external source of evidence is available and is considered indirect yet relevant to the decision problem, borrowing strength from indirect evidence may provide us with a considerably quicker and cheaper alternative type of research. Consequently, when further research is deemed worthwhile, information-sharing may provide an option that minimises the opportunity cost of research and may be preferred over more expensive and time-consuming types of research. Further research could attempt to formally incorporate information-sharing considerations into the comprehensive algorithm for the approval of technologies, and identify circumstances where information-sharing may be preferable over other types of research that resolve uncertainty in relative effectiveness.

How can indirect evidence be identified?

Even though extending the evidence base to incorporate indirect evidence has the potential to result in several advantages and influence conclusions, there is not yet much work on trying to determine how we can systematically search for indirect evidence. Instead, researchers are usually informed about the existence of such evidence either by clinicians who have expertise on a disease and follow relevant updates, or by systematic reviewers who encounter articles including indirect evidence in the process of screening databases and defining inclusion/exclusion criteria.

Of course the identification of indirect evidence is inevitably going to be easier in some situations than others. For instance, adult evidence is a common indirect source for decisions that consider children, but other sources may exist as well. Decision-makers would therefore often need to judge ‘how far we should go’ to identify indirect evidence. To date, he only work that has specifically developed systematic review strategies to identify indirect evidence was conducted by Hawkins et al., 2009 and only considered the identification of evidence on indirectly related interventions. Further research could try to develop similar strategies in order to systematise the process of identifying indirect evidence on other PICOS levels.

Which information-sharing methods are more appropriate?

Although Chapter 4 provided a step-by-step approach to identify potentially applicable ISMs, ways of choosing amongst the applicable methods were not suggested. This is because the appropriateness of a method relies on the plausibility of its assumptions which cannot be assessed without information on the true relationship between the direct and indirect RTEs. Unfortunately, such information is rarely available in HTA. Therefore, the appropriateness of each method would have to be based on a judgement about the plausibility of each method’s assumption and its expected degree of information-sharing. Importantly, in decision-making such judgements should be in a transparent manner that makes all the necessary assumptions explicit. As noted in the previous section, this information could potentially be elicited from clinical experts who not only have knowledge of differences between patients, treatments and outcomes from their clinical practice, but also often have experience in generating both direct and indirect evidence.

If the preferred degree of sharing can be established, and direct and indirect evidence fall in one of the scenarios explored in Chapter 7, then we may be able to select methods that impose the desired borrowing of strength. However, as illustrated in Chapter 7, ISMs cannot always be mapped to the information-sharing spectrum because the way that they relate to lumping and splitting is in itself uncertain. Therefore, under such circumstances, the point at which each method sits on the information-sharing spectrum is unclear and even obtaining knowledge on the desired or appropriate degree of sharing may not prove helpful in refining the ISMs list.

One way to move forward then is perhaps by applying at least one method from each strength of sharing category i.e. a method expected to impose strong information-sharing, such as an informative prior; a method expected to impose minimal information-sharing, such as a multi-level model; and finally a method that shares information in an adaptive manner, such as a mixture prior. This could be considered as a sensitivity analysis to different ISMs. If all the selected methods lead to similar policy recommendations, decision-makers may confidently infer that the imposed strength of information-sharing is not a key driver of cost-effectiveness. However, if alternative ISMs suggest different conclusions, decision-makers would need to make further *value judgements*. Specifically, they would need to assess the plausibility of the assumptions imposed by each ISM in the context of interest, and derive the appropriate degree of information-sharing between direct and indirect evidence in order to implement only specific ISMs that impose the desired degree of information-sharing.

Overall, although this thesis has provided a detailed description of ISMs, the considerations that analysts must take into account, and the potential implications for decision-making, significant uncertainties still remain. In particular, shedding more light on the issues around the appropriateness of each ISM and the strength of sharing they impose will be challenging. Although simulations may better characterise how some methods compare to one another under specific circumstances, they are unlikely to provide knowledge of how all methods compare to each other in any situation; let alone, which method imposes the most appropriate strength of sharing. Further research could attempt to find alternative ways of incorporating considerations around the appropriate degree of information-sharing into the methods choice process.

8.5. Concluding remarks

In conclusion, this thesis has demonstrated that when indirect evidence is used in HTA to strengthen inference, lumping and splitting are not the only options. Instead, there are a plethora of more sophisticated methods, which bridge the gap between lumping and splitting by imposing more moderate —and perhaps more appropriate —degrees of information-sharing. Researchers working in the field of HTA should therefore consider expanding their toolbox to include more ISMs. To that end, a step-by-step approach to methods identification was suggested. This may be useful for analysts and appraisers who want to ensure that all relevant methodological options were considered. Furthermore, through an application in an existing economic evaluation, it was shown that the use of different ISMs can lead to different adoption and research prioritisation decisions. This highlights that method choice has the potential to considerably influence policy recommendations, and implies that the plausibility of methods' assumptions should be carefully assessed. Finally, by the means of a simulation, it was shown that it may not always be easy to predict the degree of information-sharing that each method imposes. Hence, judgements about the appropriate degree of information-sharing should be associated with particular ISMs only cautiously. Overall, this thesis is to date the first piece of work that has attempted to compile an exhaustive list of ISMs, retaining a clear focus on policy, and with the ultimate purpose of providing a structured way of thinking about information-sharing problems in HTA.

Appendices

A. Appendix to Chapter 3: Classifying information-sharing methods: a citation-mining review

A.1. Methods

Exclusion criteria

Papers were excluded from the search if they fell in any of the categories below:

1. Methods or applications developed outside health research field (e.g. ecology).
2. Applications of standard MA/NMA methods without any extensions or developments to accommodate the inclusion of indirect information.
3. Irrelevant papers. Examples in that category include the following:
   - Papers that developed graphical/presentational methods for MA/NMA.
   - Papers that developed methods intended to assess consistency of the evidence.
   - Papers that aimed to introduce basic concepts and methods of MA/NMA, without any focus on advanced methods that extend the standard models to accommodate the inclusion of indirect evidence.
   - Reviews of the quality of the methods used to conduct MA/NMA.
   - Papers that develop methods to combine sources of information outside the field of MA/NMA. For example, methods that aim to utilise evidence from historical controls in the design of future trials or outside the field of comparative effectiveness research.
4. Protocols for the conduct or analysis of a future study.
5. Articles in which the University of York could not provide access to the full-text.

A.2. Results

Table A.2.1: *A brief summary of the papers included in the citation-mining review.*

	Reference	Summary
1	Achana et al., 2014	They extend network meta-analytic models to the multiple outcomes setting, allowing for strength to be borrowed across different outcomes. They also explain how the constant potency assumption, originally proposed by (Dumouchel and Harris, 1983), can be used to share information across interventions and outcomes simultaneously.
2	Achana et al., 2013	They extend methods that adjust for baseline-risk imbalances from the meta-analysis to the network meta-analytic framework. The models that they impose on the study-specific baselines, effectively share information across studies that enrol different populations (hence the baseline imbalances). They also describe models that can be imposed on the comparison-specific meta-regression slopes thus also sharing information across treatment comparisons.
3	Ades et al., 2010	They describe models that simultaneously analyse multiple mutually exclusive outcomes, specifically accounting for the negative, within-trial, correlations that are induced by this data structure.
4	Ades et al., 2008	This was one of the seminal papers included in the citation-mining review. The authors discuss multi-parameter evidence synthesis and the concept of borrowing strength describing models that have been suggested in the literature such the confidence profile method, hierachical models and multi-variate approaches.
5	Ades et al., 2006	This was one of the seminal papers included in the citation-mining review. The authors discuss the role of Bayesian methods that can accommodate borrowing of strength in cost-effectiveness analysis. They conceptually describe methods that can simultaneously analyse multiple outcomes, that share information across patient subgroups, that incorporate observational evidence and that can be used for bias-adjustment.

6	Ades and Sutton, 2006	They describe approaches for multi-parameter evidence synthesis including the confidence profile method, cross-design synthesis, hierarchical models and functions of parameters.
7	Bujkiewicz et al., 2016	They suggest multivariate models that simultaneously analyse surrogate and final outcomes. The models they suggest can impose structure to the covariance matrix to accommodate both cases where all outcomes are related to each other and cases where some outcomes are conditionally independent.
8	Bujkiewicz et al., 2014	They expand the evidence base by adding evidence on outcomes different than the target outcomes and simultaneously synthesise them using multivariate methods; thus, borrowing strength from related outcomes. In addition, they utilise a different dataset to derive informative prior distributions for the between study correlations.
9	Chaimani and Salanti, 2012	They describe models that estimate and adjust for small-studies bias, thus sharing information across studies of different designs. They also suggest models that can be imposed on the comparison-specific coefficients (i.e. the comparison-specific extend of the small-study effect modification), hence sharing information across treatment comparisons.
10	Cooper et al., 2009	They propose modelling approaches for the comparison-specific effect modification coefficients (i.e. slopes) in meta-regression models.
11	Copas et al., 2018	They explore the combination of primary and secondary outcomes using multivariate RE meta-analytic models. They further show that, usually, the extent of the information gain using multivariate approaches is only modest.
12	da Costa et al., 2017	They describe an application of network meta-analytic models in which they assume a linear (on the modelling scale —log relative dosage) dose-response curve; thus, sharing information across treatment comparisons. They also impose a random-walk across different the relative effects of different follow-ups hence sharing information across different outcomes.

13	Dakin et al., 2011	They describe models to relate outcomes that pertain to measurements taken at different parts during the day (This is here considered sharing information across different endpoints). They further model treatments within classes, allowing strength to be borrowed from interventions that function through similar mechanisms.
14	Daniels and Hughes, 1997	They propose multivariate meta-analytic models that evaluate the association between surrogate markers and final outcomes. Information is shared across the two outcomes by modelling their correlation structure.
15	Del Giovane et al., 2013	They describe network meta-analytic models that relate the relative effects of different dosages of the same treatment. Specifically, they explore lumping all dosages, imposing random-walks, constraints, dose-response curves and class-effects.
16	Dias et al., 2010b	They describe meta-regression type models that simultaneously analyse studies in different levels of risk-of-bias (This is considered here as studies of different design). Their model can estimate and internally adjust for biases that are due to both active *vs* inactive treatment comparisons and active *vs* active treatment comparisons.
17	Dias et al., 2011a	They set out the basic model for NMA in which the between-trial heterogeneities are assumed to be comparison independent; Hence, information is then shared across different treatment comparisons both as part of the consistency equations of the model and as part of the common heterogeneity component.
18	Dias et al., 2011b	They describe models that explore and explain heterogeneity by accounting for specific effect modifiers. They also describe models that can be imposed on the comparison-specific effect modification coefficients to assist their identification.
19	Dias et al., 2011c	They describe models that can imposed across the study-specific baseline parameters such as a simple random-effect across all baselines. This is considered here as sharing information amongst different populations, because the baseline imbalances may be indicative of different types of populations enrolled in different trials.

20	Ding and Fu, 2013	They describe a longitudinal model that combines information from studies that report at multiple/different follow-ups periods without the need for data reconstruction, whilst allowing prediction of relative effects pertaining to follow-ups that have not been observed.
21	Dominici et al., 1999	They describe a meta-analytic model that allows the relative treatment effects of interventions that fall under the same 'class' to shrink towards their class-specific mean. Hence, this assumptions shares information across multiple treatment comparisons.
22	Duarte et al., 2017	Even though they seek to make a decision for a paediatric population, the authors extend the evidence base to include adult evidence and analyse the full evidence set assuming no differences across adult and paediatric patients.
23	Eddy et al., 1990	They describe the confidence profile method which adjusts for known sources of bias by directly modifying the likelihood function. This is categorised here as enabling information-sharing across studies pertaining to different designs.
24	Efthimiou et al., 2014	They describe two multivariate approaches to simultaneously model multiple outcomes in the NMA setting. The first models within- and between- trial correlations separately, and the second expands the alternative model suggested by (Riley et al., 2008), which only models the overall correlation, from MA to NMA.
25	Efthimiou et al., 2017	They describe models to simultaneously synthesise evidence pertaining to several study-designs. The suggested models include hierarchical models, informative prior models and design-adjusted models.
26	Efthimiou et al., 2015	They describe multivariate approaches to simultaneously model multiple outcomes in the NMA.
27	Gamalo-Siebers et al., 2017	They describe prior-based and hierarchical methods (including power-priors) to combine paediatric and adult evidence; thus, sharing information amongst multiple populations.

28	Higgins and Whitehead, 1996	They describe the standard RE NMA model and in addition, suggest a method for using historical information to derive an informative prior for the between-studies heterogeneity. This can be particularly helpful when evidence is sparse and the heterogeneity cannot be appropriately estimated.
29	Hong et al., 2018a	They improve the alternative model suggested by (Riley et al., 2008) by suggesting a robust variance estimator.
30	Hong et al., 2016	They describe contrast-based and arm-based parametrisations of a framework that allows simultaneous synthesis of multiple outcomes. This framework assumes that all studies can contain all treatment arms and hence considers missing arms as missing data and imputes for them.
31	Hong et al., 2018b	Described power and commensurate prior methods to combine aggregate-level and individual-patient level evidence in NMA.
32	Hwang and DeSantis, 2018	They demonstrate that, just as in the MA setting, the use of multivariate methods has the capacity to reduce outcome reporting bias under several outcome missingness scenarios in the NMA setting as well.
33	Jackson et al., 2011	They explained multivariate meta-analysis for multiple outcomes, including within- and between- studies level models and discussed potential benefits and areas of application as well as assumptions and disadvantages.
34	Jackson et al., 2013	They propose a method for multivariate RE meta-analysis that is also able to accommodate the inclusion of covariates through meta-regression.
35	Jackson et al., 2014	They propose a multivariate method to model studies that report survival outcomes at multiple/different follow-up points. Their method models the between-study covariance matrix across different time periods.

36	Jackson and Riley, 2014	They extend a refined method, previously developed by (Hartung and Knapp, 2001) in the univariate setting, to the multivariate setting where multiple outcomes are simultaneously modelled. This method is particularly useful when only few studies are included in the MA causing problems in the estimation of the between-studies covariance matrix.
37	Jackson et al., 2017	They describe multivariate NMA methods and further propose a method for calculating the extent of strength that is borrowed across outcomes. Their method is based on a comparison of the precision of the estimates under the univariate and the multivariate approach.
38	Jackson et al., 2018	They extend univariate NMA methods to the multivariate setting where multiple outcomes are simultaneously synthesised. Their model further allows for two types of variance components. One that is due to between-study heterogeneity and one that is due to inconsistency.
39	Kirkham et al., 2012	They show that multivariate meta-analytic methods have the capacity to reduce outcome reporting bias under several outcome missingness mechanisms.
40	Langford et al., 2018	They developed methods to meta-analyse studies reporting for different/multiple dosages of the same treatment; hence, sharing information across the relative effectiveness of different treatment comparisons. Their method utilises the Emax model that is commonly employed in pharmacology and has several advantages over other, unbounded, approaches such as linear dose-response models.
41	Liu et al., 2018	They develop a multivariate method for simultaneous synthesis of multiple outcomes where within- and between-trials correlations are accounted using copulas.
42	Lu et al., 2007	They extend NMA methods to accommodate cases where the available studies report for multiple/different fixed follow-up periods; hence, their methods borrows strength across different endpoints which is considered here as information-sharing across different outcomes.

43	Lu and Ades, 2009	They model between-trial variance structures that are compatible with consistency assumptions and allow one to incorporate prior information on correlations between treatment arms.
44	Lu et al., 2014	They suggest methods to model the treatment comparison-specific between-trials heterogeneities such as the use of triangle inequalities which stem from second order consistency.
45	Madan et al., 2014	They develop a method to simultaneously analyse multiple outcomes reported at different/several follow-ups of complex interventions. Their model shares information across outcomes and treatment comparisons simultaneously
46	Mak et al., 2009	They use observational evidence to derive an informative prior that is used for the analysis of the available randomised trials. This is a two-step process which results in information-sharing across studies of different designs.
47	Mavridis and Salanti, 2013	They provide a thorough introduction to multivariate meta-analytic methods and a tutorial on how to simultaneously analyse multiple outcomes.
48	Mavridis et al., 2013	They describe an extension of a selection model, previously suggested by (Copas, 1999) that can be used in MA to account for publication bias that at is due to studies' treatment effect size and precision. This is considered here as sharing information across studies that of different designs.
49	Mawdsley et al., 2016	They describe model-based NMA that simultaneously analyses trials that report for multiple dosages of a specific treatment. Their model enables information-sharing across multiple treatment comparisons using the Emax model which is commonly used in pharmacology/pharmacokinetics.

50	McCarron et al., 2010	They describe methods to combine randomised and non-randomised evidence adjusting for imbalances across study arms, within studies. Two approaches are used. One that extends the model previously suggested by (Prevost et al., 2000) and is essentially a three-level hierarchical model, and another that initially meta-analyses the non-randomised evidence and subsequently uses the posterior conclusions as informative priors for the analysis of the randomised evidence.
51	McCarron et al., 2011	They describe a simulation study that compares the methods presented in McCarron et al., 2010. These include multi-level models and prior-based methods to combine randomised and non-randomised evidence, that share information across different designs, accounting for imbalances across study arms.
52	Melendez-Torres et al., 2015	They discuss emergent methods for modelling complex interventions by grouping them into 'clinically meaningful units' or, in other words, according to the components of interventions that they include.
53	Mills et al., 2012	They described methods that model complex interventions by assuming additivity of the relative effects of the various components on the modelling scale. This approach shares information across treatment comparisons and also enables the evaluation of treatment combinations that have not been used in practice.
54	Moreno et al., 2011	They propose a meta-regression method that accounts for publication bias and small-study effects by regressing the treatment effect on its associated variance. The model simultaneously analyses evidence pertaining to 12 interventions, all of which fall into the same 'class' of antidepressants. Their meta-regression model also assumes exchangeability across the treatment comparison-specific meta-regression slopes. Overall, it shares information across different study-designs (small/large studies) and treatment comparisons (class of antidepressants)

55	Musekiwa et al., 2016	They describe a generalised linear mixed model that can simultaneously model studies reporting at multiple pre-determined timepoints (i.e. follow-ups) accounting for within- and between-studies correlations. This model is considered to share information across several outcomes.
56	Nam et al., 2003	They suggest multivariate models that can simultaneously model and share information across multiple outcomes. Two of their models, extend the traditional univariate approach and differ in the assumptions they make at the between-studies level; the third model is a mixed model approach. They compare their approaches using a simulation experiment.
57	Nixon et al., 2007	They suggest methods to model complex interventions. These include meta-regression approaches that assume additive effects among treatment components and also a bivariate approach. They also try a class-effects model where treatments are lumped within classes. All their models share information across parameters pertaining to different treatment comparisons.
58	Owen et al., 2015	They develop a multi-level approach that models interventions within classes of treatments allowing the relative effects of each treatment to shrink toward their class-specific mean. They also impose constraints on the dosages, forcing larger dosages to exhibit larger relative effects. Their models primarily share information across parameters that pertain to different treatment comparisons.
59	Prevost et al., 2000	They suggest a hierarchical, multi-level, approach to model studies pertaining to different study-designs (e.g. randomised and non-randomised studies). This includes initially modelling studies within each design and subsequently modelling allowing all design-specific hyperparameters to shrink towards an overall design-independent hypermean. This approach also allow for separate heterogeneity components to be estimates within each design and across all designs. Their model shares information across studies of different designs.

60	Pullenayegum, 2011	They suggest the use of informative prior distributions for the between-study heterogeneneity when RE meta-analyses analyse sparse evidence. Their priors are derived based on previous meta-analyses and hence this is a meta-epidemiological approach.
61	Ren et al., 2018	They develop a method to elicit informative prior distributions that can be used for the between-trials heterogeneity in RE meta-analyses.
62	Rhodes et al., 2015	They use previous meta-analyses to obtain informative priors that can be used for the between-trials heterogeneity when the number of studies analysed with a random-effect is small and estimation of the between-studies heterogeneity becomes problematic.
63	Rietbergen, 2016	They describe the use of power-priors in many settings. In one of their applications they demonstrate how power-priors can be used to combine randomised and observational evidence by discounting the likelihood of the observational data. Their models share information across multiple study-designs.
64	Riley et al., 2007a	They describe how standard bivariate meta-analysis models can be used and compare them with the univariate approach under a set of scenarios where studies report either complete information on all outcomes or some outcomes are missing at random.
65	Riley et al., 2007b	They describe multivariate RE meta-analytic methods to model simultaneously multiple outcomes and further focus on issues that arise with the estimation of the between covariance matrix, particularly when only few studies are available and the within-study variance is large.
66	Riley et al., 2008	They describe an alternative bivariate random-effect model to analyse multiple outcomes when within-trial correlations are unknown. This model does not distinguish between within-trials and between-trials correlations, and models it as a single correlation, so requires the same data as required for separate univariate meta-analyses. (Hong et al., 2018a) showed that this model may not always appropriately estimate variance and suggested a robust variance estimate that improved on this model.

67	Rodgers et al., 2011	This is an HTA where the authors analysed studies that reported at different follow-up periods without accounting for this difference and effectively lumping across follow-ups. Even though they assumed that all studies reported the same outcome, since different length of follow-ups can be considered essentially different endpoints, here, it is considered that the authors lumped across multiple outcomes.
68	Roever et al., 2019	They demontrate how mixture priors can be used to combine adult and paediatric evidence where the adult evidence are part of the prior. They further show that this approach is robust to 'prior data conflict' (that is cases where direct and indirect evidence are in disagreement) and that therefore mixture priors facilitate adaptive borrowing of strength.
69	Salanti et al., 2010	They develop network meta-regression models to estimate and adjust for novelty bias in which the effectiveness of newer treatments is potentially exaggerated. This is considered here as sharing information across studies pertaining to different designs.
70	Salanti et al., 2009	They develop a network meta-regression model that estimates and adjusts for the effect modification caused by the year of publication. This is considered here a characteristic of the study-design and hence this model shares information across studies of different designs.
71	Schmitz et al., 2013	They suggest modelling approaches to combine randomised and non-randomised studies. These include a simple lumping approach where no differences are considered, using the non-randomised evidence as prior information and analysing both sources with a three-level model that initially models studies within each design and subsequently combines the design-specific hyperparameters.
72	Soares et al., 2014a	They describe modelling approaches that can be used to overcome issues relating to evidence sparsity. Amongst the suggested models there are methods that lump across different population subgroups (patients of different disease severity) and methods that impose a 'class-effect' on intervention functioning through the same mechanism.

73	Spiegelhalter and Best, 2003	They describe a modelling approach that can be used in random-effect meta-analysis to adjust for internal and external biases and therefore combine studies that may pertain to several different study-designs.
74	Tan et al., 2018	They use a bivariate meta-analytic model to obtain estimates required for decision-making that have not been reported and would not be obtainable using standard methods.
75	Thorlund et al., 2013	They conduct a simulation experiment to compare different models, originally suggested by (Lu and Ades, 2009), that can be imposed on the treatment comparison-specific between-trial heterogeneities. These models share information across different treatment comparisons.
76	Trinquart et al., 2012	They describe meta-regression models, similar to those suggested by (Chaimani and Salanti, 2012) that can be used, to estimate and adjust for reporting bias. This is assumed to be linked with the study size and therefore their models share information across studies of different designs.
77	Turner et al., 2015	They utilise meta-epidemiological data from previous meta-analyses in order to obtain 'of-the-shelf' informative priors for the between-trials heterogeneity in RE meta-analyses. These informative priors are particularly useful when there are only few studies in the meta-analysis and the estimation of the between-studies heterogeneity becomes problematic.
78	Turner et al., 2009	They suggest bias-adjustment methods which allow synthesis of studies that differ in rigour (i.e. internal validity) and relevance (i.e. external validity). Their approaches allow for both additive and proportional biases on the modelling scale. These models share information across different study-designs.
79	van Houwelingen et al., 2002a	They describe extensions to the univariate approach (that can only model one outcome at a time). These include bivariate methods that simultaneously model two outcomes and allow information to be shared across outcomes at the within- and the between-studies level.

80	Van Houwelingen et al., 1993	This is one of the seminal papers included in the citation-mining review. The authors set the initial ideas around the use of multivariate meta-analysis to simultaneously model multiple outcomes allowing strength to be borrowed across them through their correlation structure.
81	Warren et al., 2014	They describe how hierarchical, multi-level, methods can be used to model multiple dosages of the same interventions and multiple treatments that fall under the same 'class' (i.e. mechanism of action). Furthermore, they show how dosage constraints can be imposed assuming that larger dosages exhibit larger relative effects.
82	Wei and Higgins, 2013a	They suggest an approach that can be used for multivariate models to approximate within-study covariances when their estimation is problematic because the within-trial correlations are either unknown or cannot be estimated using IPD.
83	Wei and Higgins, 2013b	They set out to extend bivariate meta-analytic methods to cases where more than two outcomes are simultaneously modelled. They further suggest alternatives to the Wishart prior for the variance-covariance matrix and explore simplifying assumptions that can be imposed on the variances and the correlations when their number increases due to additional outcomes included in the analysis.
84	Welton et al., 2009b	They suggest NMA meta-regression approaches that can be used to model complex interventions with multiple treatment components. On top of simple additive -on the modelling scale- relative effects, they also show how synergistic or antagonistic effects can be incorporated in the model.
85	Welton et al., 2009a	They suggest hierarchical models that can be used to simultaneously model studies in high and low risk of bias using a bias-adjustment approach; hence, their models share information across multiple study-designs. They further show how external evidence can be used to derive informative priors for the bias component.

86	Welton et al., 2008	They suggest models that simultaneously synthesize two structurally related time-to-event outcomes. They use constraints to reflect that one outcomes needs to be reached before the other and they also model their between-studies covariance using multivariate methods.
87	Welton et al., 2010	They develop a multi-parameter evidence synthesis framework to model multiple time-to event outcomes. They reflect structural relationships among outcomes by forcing their relative treatment effects to differ by a fixed component term. They also reflect the between-study correlation structure amongst outcomes using multivariate methods.
88	Wolpert and Kerrie, 2004	They suggest models, similar to those developed by (Eddy et al., 1990), to model multiple studies pertaining to several designs by directly modelling sources of bias using adjusted likelihoods.
89	Wu et al., 2018	They describe methods for model-based meta-analysis of biologic products using a linear dose-response relationship where the dosage is proportional to the relative effect -on the modelling scale- and also using the commonly employed in the pharmacokinetics field non-linear Emax model. Their models share information across treatment comparisons (i.e. the relative effects of different dosages of the same treatment).

B. Appendix to Chapter 5: Borrowing strength from paediatric patients to inform relative effectiveness in adults: a case-study

B.1. Previous work

Table B.1.1: *Characteristics of the interventions used in the included studies.*

Study	Control	Treatment	Total dosage (gr/kg)	Volume (ml/kg/day)	Treatment Duration (days)	Jadad Score
Rodriguez 2005	5% HAS + SC	Pentaglobin (Biotest Pharma, Germany)	1.75	7	5	5
Hentrich 2006	HAS + SC	Pentaglobin (Biotest Pharma, Germany)	0.93	6.2	3	3
Karatzsas 2002	PLAC	Pentaglobin (Biotest Pharma, Germany)	0.75	5	3	2
Tugrul 2002	SC	Pentaglobin (Biotest Pharma, Germany)	0.75	5	3	3
Behre 1995	5% HAS + SC	Pentaglobin (Biotest Pharma, Germany)	0.93	6.2	3	1
Shedel 1991	SC	Pentaglobin (Biotest Pharma, Germany)	0.855	5.7	3	3
Wesoly 1990	PLAC	Pentaglobin (Biotest Pharma, Germany)	0.75	5	3	1
Spannbruker 1987	No Treatment	Pentaglobin (Biotest Pharma, Germany)	0.45	3	3	1
Dominioni 1996	5% HAS	Sandoglobulin (Sandoz Pharmaceutical Corp, Italy)	1	4	5	3
Burns 1991	HAS	Sandoglobulin (Sandoz Pharmaceutical Corp, Italy)	1.2	8	3	5
De Simone 1988	SC	Sandoglobulin (Sandoz Pharmaceutical Corp, Italy)	1	3.3	5	1
Werdan 2007	0.1% HAS	Polyglobin N (Bayer Biological Products, Germany)	0.9	9	2	5
Grundmann 1988	No Treatment	Intraglobin F (Biotest Pharma, Germany)	0.5	5	2	2
Darenberg 2003	1% HAS + SC	Endobulin SD (Baxter)	2	13.3	3	5
Lindquist 1981	SC	Pepsin-treated human gamma globulin – Gamma-venin	0.45	3	3	3
Masaoka 200	SC	Not specified	0.21	1.4	3	3
Yakut 1998	20% HAS	Gamimune N 10% (Miles Inc. Pharmaceutical Division, USA)	1.8	5.2	7	3

HAS: Human Albumin Serum, SC: standard of care (i.e. a combination of antibiotics), PLAC: Placebo.

B.2. Systematic review update

Table B.2.1: *Search in Ovid MEDLINE(R).*

#	Searches	Results
1	immunoglobulins/	42452
2	immunoglobulin$.tw.	141513
3	ivig.tw.	6165
4	1 or 2 or 3	165054
5	sepsis/	53622
6	sepsis.tw.	83096
7	septic shock/	20856
8	septic shock.tw.	18886
9	septicemia/	53622
10	septicaemia.tw.	6055
11	septicemia.tw.	12227
12	5 or 6 or 7 or 8 or 9 or 10 or 11	139261
13	4 and 12	1778
14	randomized controlled trial.pt.	464602
15	controlled clinical trial.pt.	92507
16	randomized.ab.	407222
17	placebo.ab.	187477
18	drug therapy.fs.	2031668
19	randomly.ab.	288588
20	trial.ab.	423856
21	groups.ab.	1779563
22	14 or 15 or 16 or 17 or 18 or 19 or 20 or 21	4192834
23	exp animals/ not humans.sh.	4475707
24	22 not 23	3617942
25	13 and 24	553

Ovid MEDLINE(R) 1946 to July Week 3 2018, Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations July 31, 2018.

Table B.2.2: *Search in EMBASE.*

#	Searches	Results
1	immunoglobulins/	115815
2	immunoglobulin$.tw.	163890
3	ivig.tw.	14191
4	1 or 2 or 3	235251
5	sepsis/	137769
6	sepsis.tw.	125463
7	septic shock/	44807
8	septic shock.tw.	29733
9	septicemia/	16197
10	septicaemia.tw.	6613
11	septicemia.tw.	13638
12	5 or 6 or 7 or 8 or 9 or 10 or 11	229315
13	4 and 12	5007
14	random.tw.	268899
15	placebo.mp.	407475
16	double-blind.tw.	170545
17	14 or 15 or 16	716448
18	17 and 13	306
19	animals/ not (animals/ and humans/)	1348142
20	18 not 19	306

EMBASE 1980 to 2018 Week 31.

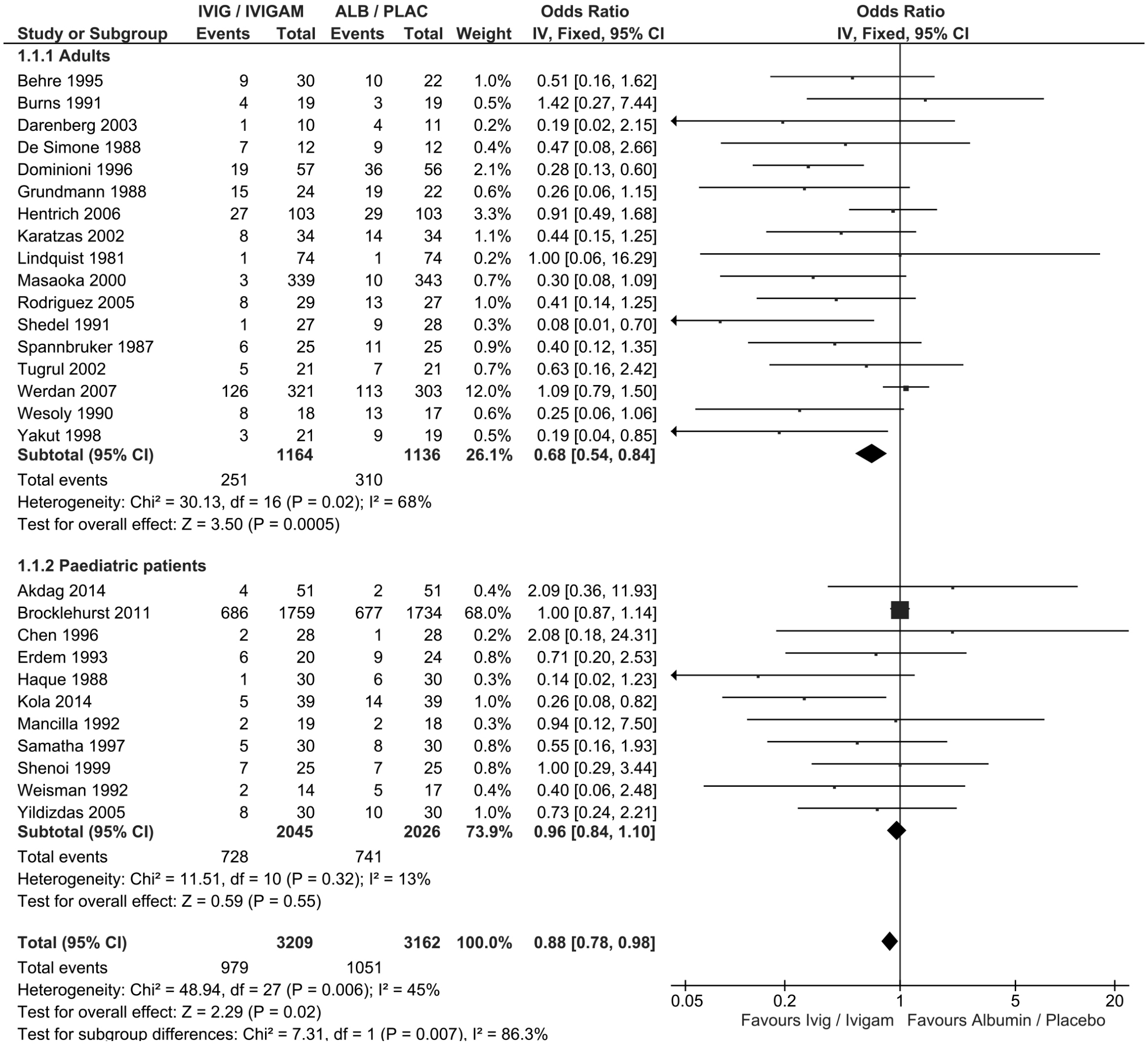
Table B.2.3: *Current evidence totality comprising from RCTs on both adult and paediatric patients.*

#	Reference	Soares et al., 2012	M(-09)	E(-09)	M(10-)	E(10-)	Alejandria and Marissa, 2013	Population
1	Behre et al., 1995	✓	n/a	n/a	n/a	n/a	✓	Adults
2	Burns et al., 1991	✓	n/a	n/a	n/a	n/a	✓	Adults
3	Darenberg et al., 2003	✓	n/a	n/a	n/a	n/a	✓	Adults
4	De Simone et al., 1988	✓	n/a	n/a	n/a	n/a	✓	Adults
5	Dominioni et al., 1996	✓	n/a	n/a	n/a	n/a	✓	Adults
6	Grundmann and Hornung, 1988	✓	n/a	n/a	n/a	n/a	✓	Adults
7	Hentrich et al., 2006	✓	n/a	n/a	n/a	n/a	✓	Adults
8	Karatzas et al., 2002	✓	n/a	n/a	n/a	n/a	✓	Adults
9	Lindquist et al., 1981	✓	n/a	n/a	n/a	n/a	✓	Adults
10	Masaoka et al., 2000	✓	n/a	n/a	n/a	n/a	✓	Adults
11	Rodriguez et al., 2005	✓	n/a	n/a	n/a	n/a	✓	Adults
12	Schedel et al., 1991	✓	n/a	n/a	n/a	n/a	✓	Adults
13	Spannbrucker et al., 1987	✓	n/a	n/a	n/a	n/a	-	Adults
14	Tugrul et al., 2002	✓	n/a	n/a	n/a	n/a	✓	Adults
15	Werdan et al., 2007	✓	n/a	n/a	n/a	n/a	✓	Adults
16	Wesoly et al., 1990	✓	n/a	n/a	n/a	n/a	✓	Adults
17	Yakut et al., 1998	✓	n/a	n/a	n/a	n/a	✓	Adults
18	Chen, 1996	n/a	✓	✓	n/a	n/a	✓	Neonates
19	Erdem et al., 1993	n/a	✓	-	n/a	n/a	✓	Neonates
20	Haque et al., 1988	n/a	✓	-	n/a	n/a	✓	Neonates
21	Mancilla-Ramirez et al., 1992	n/a	✓	-	n/a	n/a	✓	Neonates
22	Samatha et al., 1997	n/a	-	-	n/a	n/a	✓	Neonates
23	Shenoi et al., 1999	n/a	✓	✓	n/a	n/a	✓	Neonates
24	Weisman et al., 1992	n/a	✓	✓	n/a	n/a	✓	Neonates
25	Akdag et al., 2014	n/a	n/a	n/a	✓	✓	n/a	Neonates
26	Brocklehurst et al., 2011	n/a	n/a	n/a	✓	✓	✓	Neonates
27	Kola et al., 2014	n/a	n/a	n/a	-	✓	NA	young children
28	Yildizidas et al., 2005	n/a	-	-	n/a	n/a	✓	young children

The various columns indicate which steps identified which studies. M and E abbreviate Medline and Embase respectively. (-09) indicates that studies where searched for up to December 2009, and (10-) that studies where searched from January 2010 onwards. n/a suggests that a study could not have been identified in this step, while '-' that a study could have been identified but was not.

B.3. Naive analyses

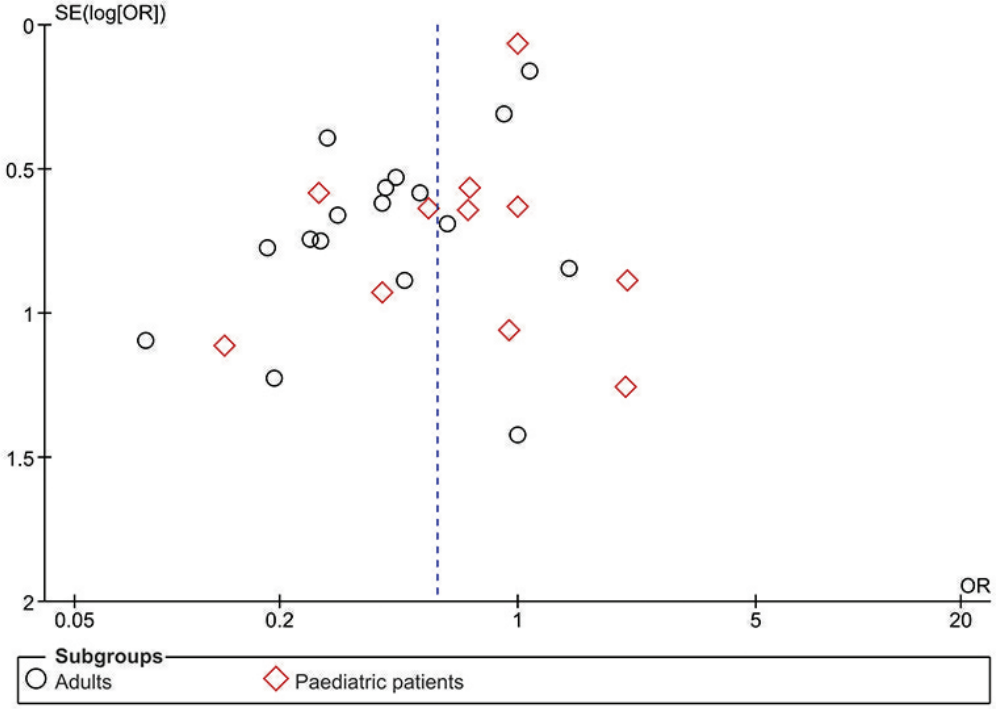
Figure B.3.1: *FE meta-analysis separately within each population and across both populations.*



	IVIG / IVIGAM		ALB / PLAC			Odds Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	IV, Fixed, 95% CI
1.1.1 Adults						
Behre 1995	9	30	10	22	1.0%	0.51 [0.16, 1.62]
Burns 1991	4	19	3	19	0.5%	1.42 [0.27, 7.44]
Darenberg 2003	1	10	4	11	0.2%	0.19 [0.02, 2.15]
De Simone 1988	7	12	9	12	0.4%	0.47 [0.08, 2.66]
Dominioni 1996	19	57	36	56	2.1%	0.28 [0.13, 0.60]
Grundmann 1988	15	24	19	22	0.6%	0.26 [0.06, 1.15]
Hentrich 2006	27	103	29	103	3.3%	0.91 [0.49, 1.68]
Karatzas 2002	8	34	14	34	1.1%	0.44 [0.15, 1.25]
Lindquist 1981	1	74	1	74	0.2%	1.00 [0.06, 16.29]
Masaoka 2000	3	339	10	343	0.7%	0.30 [0.08, 1.09]
Rodriguez 2005	8	29	13	27	1.0%	0.41 [0.14, 1.25]
Shedel 1991	1	27	9	28	0.3%	0.08 [0.01, 0.70]
Spannbruker 1987	6	25	11	25	0.9%	0.40 [0.12, 1.35]
Tugrul 2002	5	21	7	21	0.7%	0.63 [0.16, 2.42]
Werdan 2007	126	321	113	303	12.0%	1.09 [0.79, 1.50]
Wesoly 1990	8	18	13	17	0.6%	0.25 [0.06, 1.06]
Yakut 1998	3	21	9	19	0.5%	0.19 [0.04, 0.85]
Subtotal (95% CI)		**1164**		**1136**	**26.1%**	**0.68 [0.54, 0.84]**
Total events	251		310			
Heterogeneity: Chi² = 30.13, df = 16 (P = 0.02); I² = 68%						
Test for overall effect: Z = 3.50 (P = 0.0005)						
1.1.2 Paediatric patients						
Akdag 2014	4	51	2	51	0.4%	2.09 [0.36, 11.93]
Brocklehurst 2011	686	1759	677	1734	68.0%	1.00 [0.87, 1.14]
Chen 1996	2	28	1	28	0.2%	2.08 [0.18, 24.31]
Erdem 1993	6	20	9	24	0.8%	0.71 [0.20, 2.53]
Haque 1988	1	30	6	30	0.3%	0.14 [0.02, 1.23]
Kola 2014	5	39	14	39	1.0%	0.26 [0.08, 0.82]
Mancilla 1992	2	19	2	18	0.3%	0.94 [0.12, 7.50]
Samatha 1997	5	30	8	30	0.8%	0.55 [0.16, 1.93]
Shenoi 1999	7	25	7	25	0.8%	1.00 [0.29, 3.44]
Weisman 1992	2	14	5	17	0.4%	0.40 [0.06, 2.48]
Yildizdas 2005	8	30	10	30	1.0%	0.73 [0.24, 2.21]
Subtotal (95% CI)		**2045**		**2026**	**73.9%**	**0.96 [0.84, 1.10]**
Total events	728		741			
Heterogeneity: Chi² = 11.51, df = 10 (P = 0.32); I² = 13%						
Test for overall effect: Z = 0.59 (P = 0.55)						
Total (95% CI)		**3209**		**3162**	**100.0%**	**0.88 [0.78, 0.98]**
Total events	979		1051			
Heterogeneity: Chi² = 48.94, df = 27 (P = 0.006); I² = 45%						
Test for overall effect: Z = 2.29 (P = 0.02)						
Test for subgroup differences: Chi² = 7.31, df = 1 (P = 0.007), I² = 86.3%						

The plot was created using Review Manager 5.4 (Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014).

Figure B.3.2: *Funnel-plot including all studies. Adult studies in black circles and paediatric studies in red diamonds.*



The plot was created using Review Manager 5.4 (Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014).

B.4. Heterogeneity re-exploration

Table B.4.1: *Step 2. T2 Network results of meta-regression models on covariates shown in the first column.*

covariate	Model	τ_{AD}	τ_{PE}	β_{AD}	β_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
NULL	FE	n/a	n/a	-	-	77.02	51.43	25.59	303.36
	RE	0.56*	0.47	-	-	51.97	30.82	21.14	289.31
Duration	FE sep	n/a	n/a	−0.38*	0.36	61.39	37.11	24.27	289.70
	FE com	n/a	n/a	−0.27*	−0.27*	69.01	38.12	30.90	296.34
	RE sep	0.37‡	0.46	−0.29‡	0.25	54.01	32.40	21.61	290.33
	RE com	0.43‡	0.57‡	-0.18	-0.18	53.28	31.71	21.57	290.52
Jadad	FE sep	n/a	n/a	0.29*	0.09	64.83	39.24	25.59	293.15
	FE com	n/a	n/a	0.20*	0.20*	66.55	39.87	26.68	293.87
	RE sep	0.44‡	0.55	0.20	0.01	53.78	32.33	21.46	291.51
	RE com	0.47*	0.49	0.12	0.12	53.59	31.53	22.06	290.83
Sample	FE sep	n/a	n/a	−6.64*	−2.98*	58.83	36.55	22.29	287.16
	FE com	n/a	n/a	−4.52*	−4.52*	60.49	37.32	23.17	287.83
	RE sep	0.35	0.44	−5.26*	-2.64	53.61	32.56	21.05	288.94
	RE com	0.37‡	0.39	−4.05*	−4.05*	53.00	31.87	21.13	287.92
Dosage	FE sep	n/a	n/a	-0.18	0.40	77.72	52.22	25.50	306.05
	FE com	n/a	n/a	0.09	0.09	77.89	52.27	25.62	305.23
	RE sep	0.59*	0.51	-0.16	0.22	52.94	31.31	21.63	292.05
	RE com	0.58*	0.50	0.03	0.03	52.51	31.19	21.31	290.86
Year	FE sep	n/a	n/a	0.07*	0.03‡	59.86	35.90	23.96	288.18
	FE com	n/a	n/a	0.05*	0.05*	61.48	36.50	24.98	288.77
	RE sep	0.32	0.47	0.05*	0.02	54.36	32.93	21.43	289.87
	RE com	0.36‡	0.46	0.04*	0.04*	53.55	32.02	21.52	288.97

$*$ indicates a value that is significant at the 5% level, whilst a ‡ at the 10% level. Blue and red shading indicate low and high within-column values respectively. The NULL model refers to the corresponding network without any covariates as that is presented in Table 5.5. Sample size, N, is modelled as $1/\sqrt{N}$.

Table B.4.2: *Step 2. T3a Network results of meta-regression models on covariates.*

covariate	Model	τ_{AD}	τ_{PE}	β_{AD}	β_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
NULL	FE	n/a	n/a	-	-	71.11	50.12	20.99	299.45
	RE	0.60^*	0.35^*	-	-	51.89	31.23	20.67	289.50
Duration	FE sep	n/a	n/a	-0.37^*	0.29	58.24	37.57	20.67	288.55
	FE com	n/a	n/a	-0.27^*	-0.27^*	63.53	38.25	25.28	292.85
	RE sep	$0.42^‡$	0.37	$-0.28^‡$	0.31	53.14	32.63	20.51	290.34
	RE com	$0.49^‡$	0.44	-0.14	-0.14	53.55	31.90	21.65	291.56
Jadad	FE sep	n/a	n/a	0.34^*	-0.03	61.59	39.63	21.95	291.92
	FE com	n/a	n/a	0.17^*	0.17^*	66.80	41.91	24.88	296.11
	RE sep	$0.46^‡$	0.42	0.23	-0.04	53.86	32.70	21.16	291.90
	RE com	0.54^*	0.43	0.08	0.08	53.09	31.63	21.46	291.51
Sample	FE sep	n/a	n/a	-7.17^*	-1.28	58.91	37.36	21.55	289.20
	FE com	n/a	n/a	-4.28^*	-4.28^*	62.38	38.73	23.64	291.70
	RE sep	0.38	0.48	$-5.41^‡$	-1.05	53.97	33.09	20.89	291.22
	RE com	$0.44^‡$	0.45	-3.69	-3.69	52.99	32.06	20.93	290.27
Dosage	FE sep	n/a	n/a	-0.18	-0.04	72.93	50.91	22.03	303.28
	FE com	n/a	n/a	-0.12	-0.12	71.95	50.49	21.46	301.27
	RE sep	0.63^*	0.42	-0.17	-0.03	53.10	31.71	21.39	292.81
	RE com	0.61^*	0.38	-0.10	-0.10	52.50	31.49	21.01	291.06
Year	FE sep	n/a	n/a	0.07^*	0.01	58.14	36.49	21.65	288.45
	FE com	n/a	n/a	0.04^*	0.04^*	61.26	37.61	23.66	290.57
	RE sep	0.36	0.42	$0.05^‡$	0.02	54.06	33.14	20.92	290.66
	RE com	$0.43^‡$	0.44	0.03	0.03	53.12	32.18	20.93	290.29

$*$ indicates a value that is significant at the 5% level, whilst a ‡ at the 10% level. Red shading indicates high within-column values. The NULL model refers to the corresponding network without any covariates as that is presented in Table 5.5. Sample size, N, is modelled as $1/\sqrt{N}$.

Table B.4.3: *Step 2. T4 Network results of meta-regression models on covariates.*

covariate	Model	τ_{AD}	τ_{PE}	β_{AD}	β_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
NULL	FE	n/a	n/a	-	-	65.57	43.58	22.00	295.92
	RE	0.53^{*}	0.46^{*}	-	-	52.99	31.90	21.08	291.76
Duration	FE sep	n/a	n/a	-0.41^{*}	0.54	50.88	28.80	22.08	283.23
	FE com	n/a	n/a	-0.36^{*}	-0.36^{*}	53.29	28.85	24.44	284.61
	RE sep	0.21	0.50	-0.40^{*}	0.54	50.42	29.15	21.27	286.72
	RE com	0.22	0.57	-0.36^{*}	-0.36^{*}	51.10	29.14	21.97	287.43
Jadad	FE sep	n/a	n/a	0.32^{*}	2.99	60.45	38.43	22.01	291.78
	FE com	n/a	n/a	0.33^{*}	0.33^{*}	60.43	38.42	22.01	291.76
	RE sep	$0.46^{\ddagger}$	0.47	0.22	5.57	54.04	32.93	21.10	292.85
	RE com	$0.46^{\ddagger}$	0.47	0.23	0.23	53.96	32.91	21.05	292.67
Sample	FE sep	n/a	n/a	-8.07^{*}	-4.07	55.67	33.67	22.00	287.92
	FE com	n/a	n/a	-6.97^{*}	-6.97^{*}	55.46	33.60	21.85	286.81
	RE sep	0.30	0.49	-7.05^{*}	-4.35	53.57	32.29	21.27	291.08
	RE com	0.30	0.45	-6.31^{*}	-6.31^{*}	52.94	31.92	21.02	289.53
Dosage	FE sep	n/a	n/a	-1.44^{*}	2.04	58.89	36.77	22.11	291.29
	FE com	n/a	n/a	-1.19^{*}	-1.19^{*}	60.74	36.82	23.93	292.09
	RE sep	$0.41^{\ddagger}$	0.50	$-1.26^{\ddagger}$	2.11	52.94	31.72	21.22	291.87
	RE com	$0.43^{\ddagger}$	0.54	-0.98	-0.98	53.27	31.58	21.69	292.27
Year	FE sep	n/a	n/a	0.08^{*}	0.05	57.67	35.56	22.11	290.06
	FE com	n/a	n/a	0.08^{*}	0.08^{*}	56.87	35.38	21.49	288.16
	RE sep	0.35	0.51	0.06	0.06	54.64	33.37	21.27	292.77
	RE com	0.35	0.44	$0.06^{\ddagger}$	$0.06^{\ddagger}$	54.04	33.20	20.84	291.19

$*$ indicates a value that is significant at the 5% level, whilst a ‡ at the 10% level. Blue indicates low and high within-column values. The NULL model refers to the corresponding network without any covariates as that is presented in Table 5.5. Sample size, N, is modelled as $1/\sqrt{N}$.

Table B.4.4: *Results of meta-regression models on sample size for all network parametrisations.*

covariate	Model	τ_{AD}	τ_{PE}	β_{AD}	β_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
T2	FE sep	n/a	n/a	−6.64∗	−2.98∗	58.83	36.55	22.29	287.16
	FE com	n/a	n/a	−4.52∗	−4.52∗	60.49	37.32	23.17	287.83
	RE sep	0.35	0.44	−5.26∗	-2.64	53.61	32.56	21.05	288.94
	RE com	0.37	0.39	−4.05∗	−4.05∗	53.00	31.87	21.13	287.92
T3a	FE sep	n/a	n/a	−7.17∗	-1.28	58.91	37.36	21.55	289.20
	FE com	n/a	n/a	−4.28∗	−4.28∗	62.38	38.73	23.64	291.70
	RE sep	0.38	0.48	-5.41	-1.05	53.97	33.09	20.89	291.22
	RE com	0.44	0.45	-3.69	-3.69	52.99	32.06	20.93	290.27
T3b	FE sep	n/a	n/a	−7.49∗	-4.48	55.98	33.09	22.89	286.28
	FE com	n/a	n/a	−6.70∗	−6.70∗	55.56	32.95	22.61	284.92
	RE sep	0.29	0.50	−6.70∗	-4.51	53.22	31.57	21.65	289.36
	RE com	0.27	0.45	−6.22∗	−6.22∗	52.78	31.47	21.31	287.71
T4	FE sep	n/a	n/a	−8.07∗	-4.07	55.67	33.67	22.00	287.92
	FE com	n/a	n/a	−6.97∗	−6.97∗	55.46	33.60	21.85	286.81
	RE sep	0.30	0.49	−7.05∗	-4.35	53.57	32.29	21.27	291.08
	RE com	0.30	0.45	−6.31∗	−6.31∗	52.94	31.92	21.02	289.53

∗ indicates a value that is significant at the 5% level, whilst a ‡ at the 10% level. Blue shading indicates low within-column values with stronger shading indicating lower values.

Table B.4.5: *Results of Meta-Regression models on Jadad for all network parametrisations.*

covariate	Model	τ_{AD}	τ_{PE}	β_{AD}	β_{PE}	D_{res}	$D_{res_{AD}}$	$D_{res_{PE}}$	DIC
T2	FE sep	n/a	n/a	0.29∗	0.09	64.83	39.24	25.59	293.15
	FE com	n/a	n/a	0.20∗	0.20∗	66.55	39.87	26.68	293.87
	RE sep	0.44	0.55	0.20	0.01	53.78	32.33	21.46	291.51
	RE com	0.47∗	0.49	0.12	0.12	53.59	31.53	22.06	290.83
T3a	FE sep	n/a	n/a	0.34∗	-0.03	61.59	39.63	21.95	291.92
	FE com	n/a	n/a	0.17∗	0.17∗	66.80	41.91	24.88	296.11
	RE sep	0.46	0.42	0.23	-0.04	53.86	32.70	21.16	291.90
	RE com	0.54∗	0.43	0.08	0.08	53.09	31.63	21.46	291.51
T3b	FE sep	n/a	n/a	0.26∗	1.97	61.07	38.23	22.85	290.39
	FE com	n/a	n/a	0.26∗	0.26∗	61.09	38.20	22.88	290.41
	RE sep	0.44	0.45	0.18	-3.18	53.86	32.43	21.43	291.07
	RE com	0.43	0.47	0.19	0.19	53.84	32.46	21.38	291.10
T4	FE sep	n/a	n/a	0.32∗	2.99	60.45	38.43	22.01	291.78
	FE com	n/a	n/a	0.33∗	0.33∗	60.43	38.42	22.01	291.76
	RE sep	0.46	0.47	0.22	5.57	54.04	32.93	21.10	292.85
	RE com	0.46	0.47	0.23	0.23	53.96	32.91	21.05	292.67

∗ indicates a value that is significant at the 5% level, whilst a ‡ at the 10% level. Blue shading indicates low within-column values.

Comparison of the final base-models of this work and of the HTA

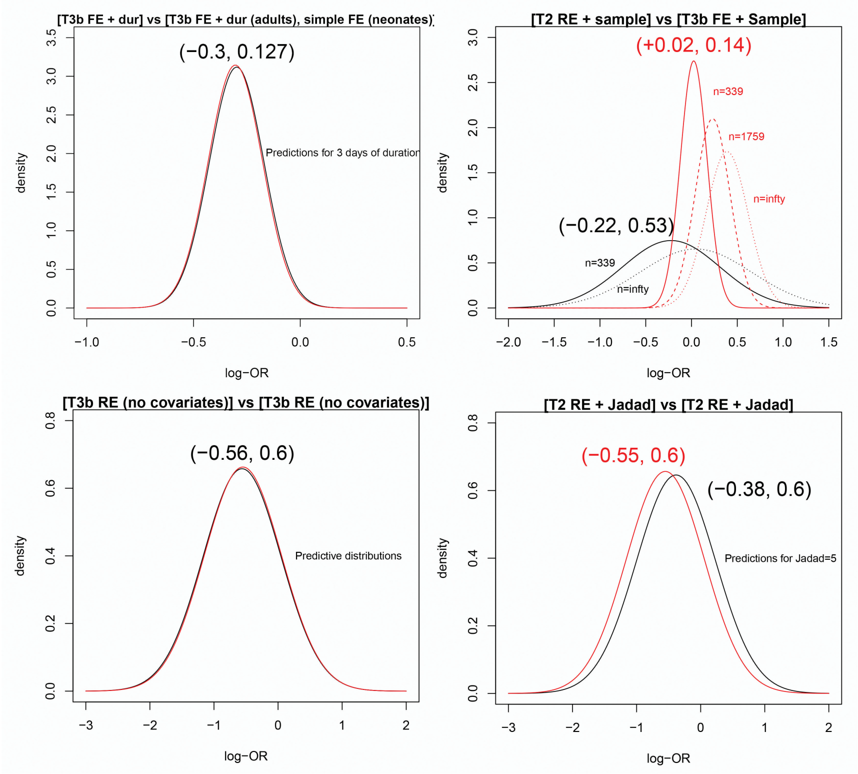
The final list of base-models is illustrated in Table B.4.6 and the two differences from the original list in the HTA are highlighted in yellow. The meta-regression model on sample size ($1/\sqrt{N}$) model is quite different from the one in Soares et al., 2012 because it is a FE T3b model with common effect modification among the two populations. In contrast, the Jadad model is only slightly different and this is due to the common effect modification coefficient which already imposes some degree of information-sharing among adults and paediatric patients. The divergence from the base-models that were used in the original HTA (Soares et al., 2012) is shown in Figure B.4.1.

Table B.4.6: *The final lists of models in Soares et al., 2012 and after the inclusion of the paediatric studies.*

(HTA) Models for CEA	(Adult + paediatric patients) Models for CEA	DIC
T3b : FE + dur	T3b : FE + duration (adults), FE (paediatric patients)	280.1
T2 : RE + sample size	T3b : FE + sample size (common)	284.9
T3b : RE	T3b : RE	290.3
T2 : RE + Jadad	T2 : RE + Jadad (common)	291.1

DICs correspond to the models of this work which are listed in the second column.

Figure B.4.1: *A comparison of the final list of base-models included in Soares et al., 2012 (black) and those included after the heterogeneity re-exploration in this work (red).*



B.5. Sharing information between adults and paediatric patients

A plain language brief description of the ISMs used in Chapter 5[1]

1. *Lumping*: All relative effects are assumed equivalent between adults and paediatric patients. All studies are analysed under a common fixed- or random- effects model; depending on the type of the base-model.

2. *Common Heterogeneity*: The between-studies heterogeneities are assumed equivalent between the adult and paediatric evidence sets.

3. *Multi-level model*: The adult and the paediatric relative treatment effects are assumed to be exchangeable.

4. *Informative prior*: The paediatric evidence is initially analysed. The paediatric posterior relative effect mean estimate is then used as an informative prior for the relative effect in the analysis of the adult evidence. Under RE models the predictive distribution of the paediatric evidence is used as a prior for the analysis of the adult evidence.

5. *Mixture prior* : As the informative prior, only now the prior is a combination of the informative component and a new vague component. A hyper-prior that weights the two components 50-50 is imposed and updated by the model.

6. *Power-prior*: The likelihood of the adults is kept at face value, but the likelihood of the paediatric studies is down-weighted according to a pre-specified weight α. For $\alpha = 0$ paediatric evidence are fully disregarded, while for $\alpha = 1$, the paediatric evidence are considered at face value.

7. *Commensurate prior*: Adult and paediatric evidence are simultaneously analysed. Even though a vague prior is used for the paediatric evidence, a prior that is centered around the paediatric relative effect is used for the adult relative effect. the extent of information-sharing is controlled by imposing a mixture (spike-and-slab) hyperprior on its variance.

8. *Prior on Heterogeneity*: The paediatric evidence is initially analysed with a RE model. The estimated between-studies posterior heterogeneity is then used as an informative prior for the heterogeneity parameter in the analysis of the adult evidence.

[1] For more details see Chapter 4.

Extending Information-sharing method (ISM)s described in Chapter 4 to accommodate information-sharing under meta-regression models

Consider the case where the base-models for both evidence sets account for the same covariate X. Then the synthesis model for the direct evidence would take the following form:

$$\begin{gathered} r_{i,k} \sim Bin(p_{i,k}, n_{i,k}) \\ logit(p_{i,k}) = \theta_{i,k} = \mu_{i_b} + \delta_{i,bk} + \beta^{Dir} \cdot (X_i - \overline{X_{Dir}}) \\ \delta_{i,bk} \sim N(d_{bk}^{Dir}, {\tau^{Dir}}^2) \\ d_{bk}^{Dir} = d_{1k}^{Dir} - d_{1b}^{Dir} \\ d_{11}^{Dir} = 0 \end{gathered}$$

where d_{bk}^{Dir}, d_{1k}^{Dir}, τ^{Dir} are the relative treatment effects, basic parameters, and between-studies heterogeneity specific to the direct evidence. X_i is the study-specific value for covariate X and $\overline{X_{Dir}}$ is the covariate value at which we center. Similarly the synthesis model for the indirect evidence is defined by specifying d_{bk}^{Indir}, d_{1k}^{Indir}, τ^{Indir}, and $\overline{X_{Indir}}$.

The prediction for the relative treatment effect of the direct evidence at a covariate value of, say, $X = 3$ is:

$$m_{1k}^{Dir}[3] = d_{1k}^{Dir} - \beta_{Dir} \cdot \overline{X_{Dir}} + \beta_{Dir} \cdot 3$$

which implies that if we choose to center at $\overline{X_{Dir}} = 3$, then two two last components cancel out and

$$m_{1k}^{Dir}[3] = d_{1k}^{Dir}$$

Similarly, for the indirect evidence the prediction for the relative effect at $X = 3$, if we choose to also center at $\overline{X_{Indir}} = 3$ is

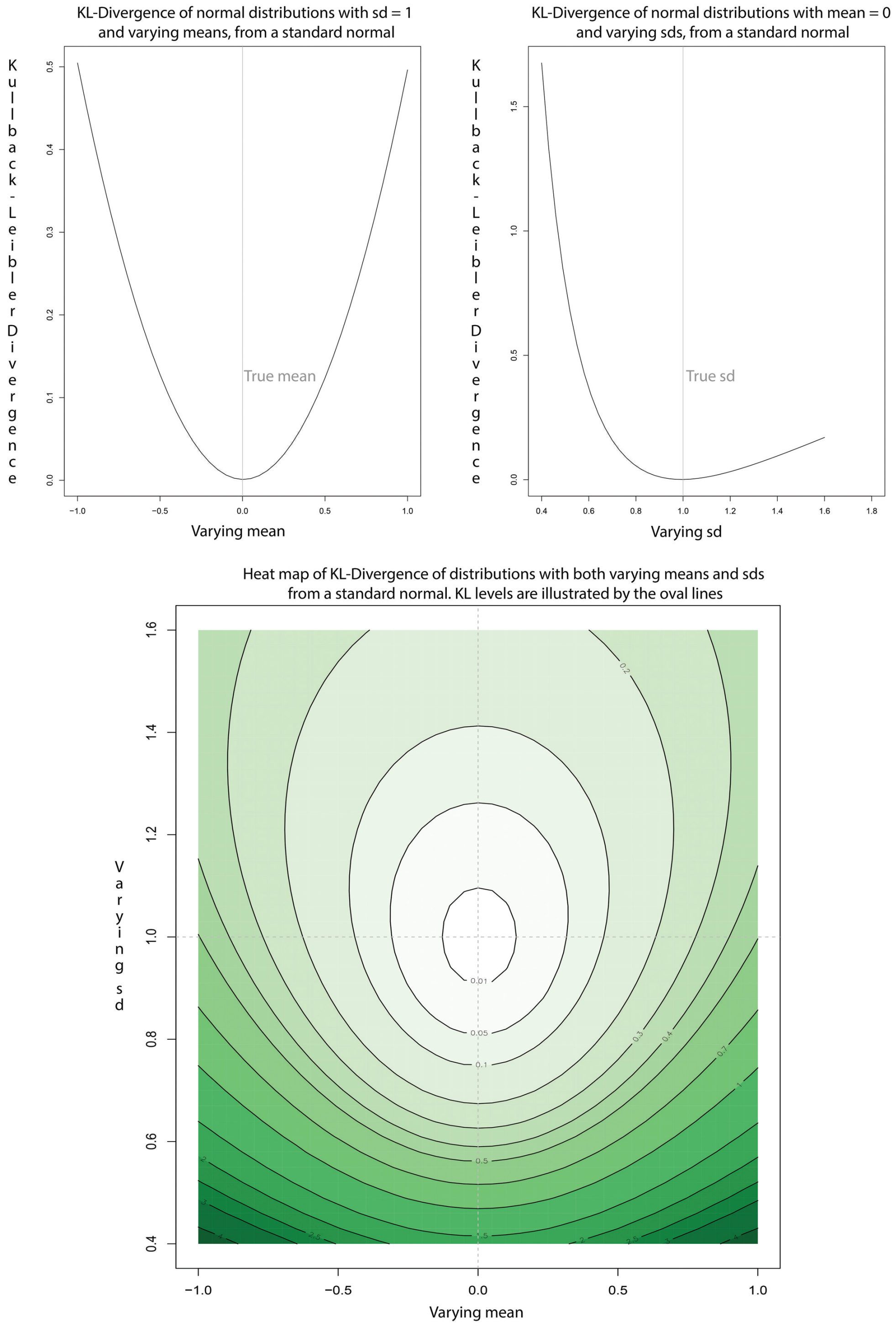
$$m_{1k}^{Indir}[3] = d_{1k}^{Indir}$$

This implies that we can still use the ISMs that were developed in Chapter 4, without any modifications to accommodate the fact that the relative effect now comprises from two components, even if different β coefficients are used in the two evidence sets, as long as we are willing to center both sources' meta-regression models at the covariate value at which we want to relate them.

An illustration of KL divergence

Figure B.5.1 illustrates KL-divergence from a standard normal distribution. The top left figure, shows the KL divergence from a standard normal for distributions with $sd = 1$ and means varying between -1 and 1, while the top right graph for distributions with $mean = 0$ and standard deviations varying between 0.4 and 1.6. It is immediately apparent that KL seems to be symmetrical for divergent means, but more sensitive to lower -than the reference distribution- standard deviations. This feature is also revealed in the bottom graph by the non-circular shapes of the oval KL levels lines. The R code that shows the integration process for the KL calculation is provided in the Appendix and follows the steps that were detailed in Jackson, 2019 for KL calculation of beta distributions.

Figure B.5.1: *An illustration of KL divergence from the standard normal distribution.*



The top left and right graphs show the KL divergence for distribution with varying means and varying standard deviations respectively. The graph in the bottom incorporates both changes simultaneously and reveals the non-symmetrical nature in which KL weights changes in the mean and the standard deviation.

Table B.5.1: *ISMs results for the T3b FE meta-regression on sample size base-model.*

Method	m.ad.339	sd.339	m.ad.infty	sd.infty	B.ad	B.ad.sd	resdev.ad	DIC
Base-case	0.023	0.149	0.384	0.228	-6.654	2.031	32.929	284.847
Lumping	-0.127	0.063	0.148	0.084	-5.069	1.431	34.444	284.847
Multi-level	-0.042	0.139	0.281	0.212	-5.946	1.939	33.353	285.338
Commensurate prior	-0.074	0.130	0.233	0.196	-5.642	1.854	33.679	287.900
Informative prior	-0.108	0.098	0.149	0.150	-5.804	1.837	34.084	
Mixture prior	-0.105	0.100	0.153	0.151	-5.825	1.840	34.085	
Pr-Power (a=0.1)	-0.040	0.116	0.301	0.183	-6.276	1.924	32.906	
Pr-Power (a=0.2)	-0.075	0.099	0.242	0.152	-5.843	1.772	33.057	
Pr-Power (a=0.3)	-0.093	0.089	0.210	0.133	-5.582	1.667	33.236	
Pr-Power (a=0.4)	-0.105	0.082	0.190	0.119	-5.442	1.603	33.442	
Pr-Power (a=0.5)	-0.112	0.076	0.178	0.109	-5.346	1.544	33.581	
Pr-Power (a=0.6)	-0.118	0.073	0.168	0.103	-5.264	1.522	33.806	
Pr-Power (a=0.7)	-0.121	0.069	0.161	0.097	-5.200	1.504	33.982	
Pr-Power (a=0.8)	-0.124	0.067	0.156	0.092	-5.155	1.462	34.096	
Pr-Power (a=0.9)	-0.127	0.065	0.151	0.087	-5.108	1.451	34.301	
Pr-Power (a=1)	-0.128	0.063	0.147	0.084	-5.061	1.433	34.457	

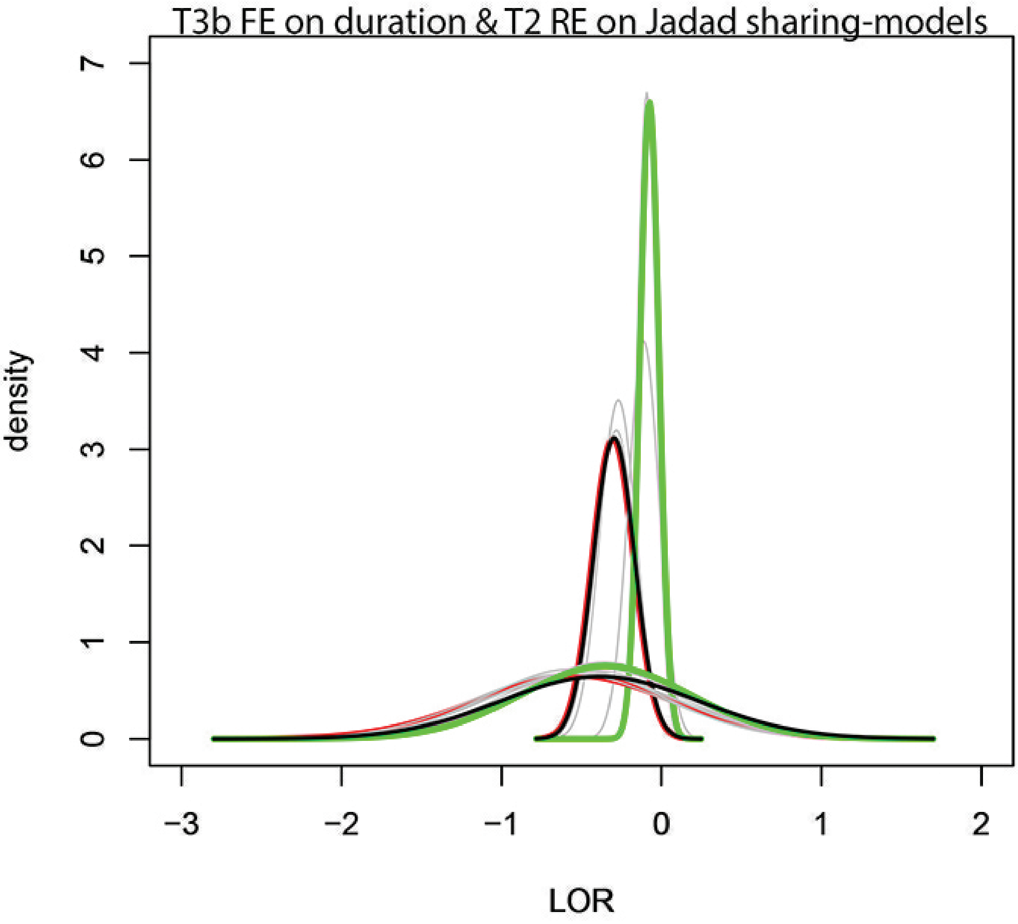
Predictions for treatment arm sizes of $n = 339, n = \infty$. Light red shade indicates the base-case (no sharing/only adults) model for comparison purposes.

Table B.5.2: *ISMs results for T3b RE base-model without any covariates.*

ISM	m.ad	sd	m.ad.pred	sd.pred	tau.ad	resdev.ad	DIC
Base-case (no sharing)	-0.551	0.276	-0.550	0.599	0.486	31.705	290.391
Common Heterogeneity	-0.525	0.256	-0.525	0.528	0.430	31.914	289.229
Lumping	-0.420	0.232	-0.420	0.512	0.424	32.820	289.842
Multi-level	-0.500	0.248	-0.501	0.570	0.471	31.749	289.845
Commensurate prior	-0.473	0.242	-0.473	0.557	0.456	31.934	294.683
Informative prior	-0.516	0.238	-0.517	0.561	0.468	31.617	
Mixture prior	-0.521	0.241	-0.521	0.569	0.471	31.643	
Prior on heterogeneity	-0.550	0.262	-0.549	0.563	0.468	31.402	
Pr-Power (a=0.1)	-0.433	0.244	-0.434	0.560	0.458	32.093	
Pr-Power (a=0.2)	-0.402	0.243	-0.404	0.554	0.452	32.347	
Pr-Power (a=0.3)	-0.395	0.241	-0.397	0.556	0.454	32.406	
Pr-Power (a=0.4)	-0.380	0.244	-0.378	0.561	0.455	32.553	
Pr-Power (a=0.5)	-0.383	0.245	-0.382	0.569	0.464	32.489	
Pr-Power (a=0.6)	-0.380	0.244	-0.379	0.570	0.463	32.545	
Pr-Power (a=0.7)	-0.388	0.242	-0.387	0.562	0.459	32.562	
Pr-Power (a=0.8)	-0.375	0.247	-0.375	0.570	0.465	32.684	
Pr-Power (a=0.9)	-0.391	0.247	-0.390	0.573	0.470	32.631	
Pr-Power (a=1)	-0.415	0.242	-0.415	0.525	0.431	32.790	

Informative and mixture priors use the predictive distribution of the indirect evidence (not the posterior mean). Light red shade indicates the base-case (no sharing/only adults) model for comparison purposes.

Figure B.5.2: *ISMs (except power-prior) of the FE Duration Meta-regression model and the RE Jadad base-models.*



Predictions refer to 3 days of duration and Jadad=5 respectively. The vague distributions correspond to the RE model predictions, whilst the narrower distributions to the FE model predictions. Black estimates correspond the original estimates from Soares et al., 2012, red to the splitting -only adults / no sharing-base-case in this work, green to lumping, and gray to the various remaining ISMs.

Table B.5.3: *ISMs for T2 RE meta-regression on Jadad score.*

ISM	m.ad	sd	m.ad.pred	sd.pred	B.ad	B.ad.sd	tau.ad	resdev.ad	DIC
Base-case (no sharing)	-0.546	0.312	-0.544	0.606	0.117	0.106	0.473	31.490	290.758
Common Heterogeneity	-0.549	0.288	-0.547	0.550	0.110	0.100	0.434	31.506	289.333
Lumping	-0.358	0.242	-0.360	0.504	0.135	0.099	0.398	32.466	289.814
Multi-level	-0.498	0.291	-0.497	0.588	0.121	0.101	0.463	31.609	290.544
Commensurate prior	-0.443	0.277	-0.445	0.561	0.131	0.103	0.438	31.929	293.130
Informative prior -pred-	-0.498	0.291	-0.498	0.580	0.124	0.104	0.456	31.726	
Mixture prior -pred-	-0.510	0.294	-0.511	0.587	0.118	0.104	0.462	31.578	
Pr-Het	-0.580	0.300	-0.578	0.589	0.102	0.104	0.476	31.178	
Pr-Power (a=0.1)	-0.261	0.272	-0.259	0.525	0.231	0.118	0.394	32.820	
Pr-Power (a=0.2)	-0.254	0.261	-0.253	0.513	0.217	0.110	0.387	32.872	
Pr-Power (a=0.3)	-0.257	0.252	-0.255	0.516	0.205	0.107	0.391	32.850	
Pr-Power (a=0.4)	-0.268	0.251	-0.268	0.512	0.196	0.106	0.390	32.814	
Pr-Power (a=0.5)	-0.269	0.245	-0.268	0.512	0.189	0.103	0.390	32.884	
Pr-Power (a=0.6)	-0.285	0.246	-0.285	0.513	0.181	0.103	0.394	32.755	
Pr-Power (a=0.7)	-0.301	0.247	-0.299	0.517	0.168	0.102	0.400	32.737	
Pr-Power (a=0.8)	-0.314	0.244	-0.314	0.516	0.160	0.099	0.401	32.745	
Pr-Power (a=0.9)	-0.326	0.249	-0.328	0.524	0.153	0.102	0.408	32.678	
Pr-Power (a=1)	-0.343	0.249	-0.344	0.528	0.147	0.104	0.414	32.568	

Predictions for Jadad = 5 (i.e. a study of the best quality). Informative and mixture priors use the predictive distribution of the indirect evidence (not the posterior mean). Light red shade indicates the base-case (no sharing/only adults) model for comparison purposes.

B.6. Power-prior

In the 3 treatments network which is found to fit the best to our data (i.e. T3b), the treatment comparison of interest is IVIG/IVIGAM *vs* ALB. The indirect (paediatric) evidence base contains only 2//11 studies that provide information on this comparison. One small study (Weisman et al., 1992) that suggests a strong effect favouring IVIG/IVIGAM, and one very big multi-center study (Brocklehurst et al., 2011) which suggests that IVIG/IVIGAM is no better than ALB.

Using these data, the resulting relative effects of the power-priors that discount the likelihood of the indirect evidence with a range of α weights is shown in Figure B.6.1A. Interestingly, as α increases (i.e. more and more weight is given to the paediatric evidence) the relative effect does not follow a monotonous increasing function. Instead, for low values of α it reaches a maximum which is more extreme than lumping and then follows a gradually decreasing trend to reach a value identical for $\alpha = 1$ that is identical to lumping.

Counter-intuitive as this may seem, it is due to the nature of the indirect evidence and the way the power-prior model works. Initially, for very low values of α, the big study (Brocklehurst et al., 2011) acquires significant weight and thus pulls the overall relative effect closer to its suggested effect (relative effect close to 0). For those low values of α the likelihood of the aforementioned small study is negligible. However, as α increases beyond 0.4 this small study starts influencing the overall estimate and thus pulling towards it suggested strong effect.

This explanation is confirmed by modifying the data and plotting the same graph for two scenarios. In Figure B.6.1B the big study (Brocklehurst et al., 2011) is altered to exhibit a strong effect similar to the small study. The results show the expected shape which is a monotonous decreasing function. Similarly, in Figure B.6.1C, the small study (Weisman et al., 1992) is altered to suggest a similar effect with the big study (i.e. no effect) and the result is a monotonous increasing function. It is hence confirmed that the observed shape in Figure B.6.1A appears because of the disagreement between the two studies informing the relative effect of interest and the fact that they differ significantly in size and therefore start influencing the overall results at different α values.

Figure B.6.1: *Relative treatment effect (IVIG/IVIGAM vs ALB) estimates (y-axis) of re-running power-prior models for different α values (x-axis), for the Jadad RE base-model, under different scenarios for the indirect evidence.*

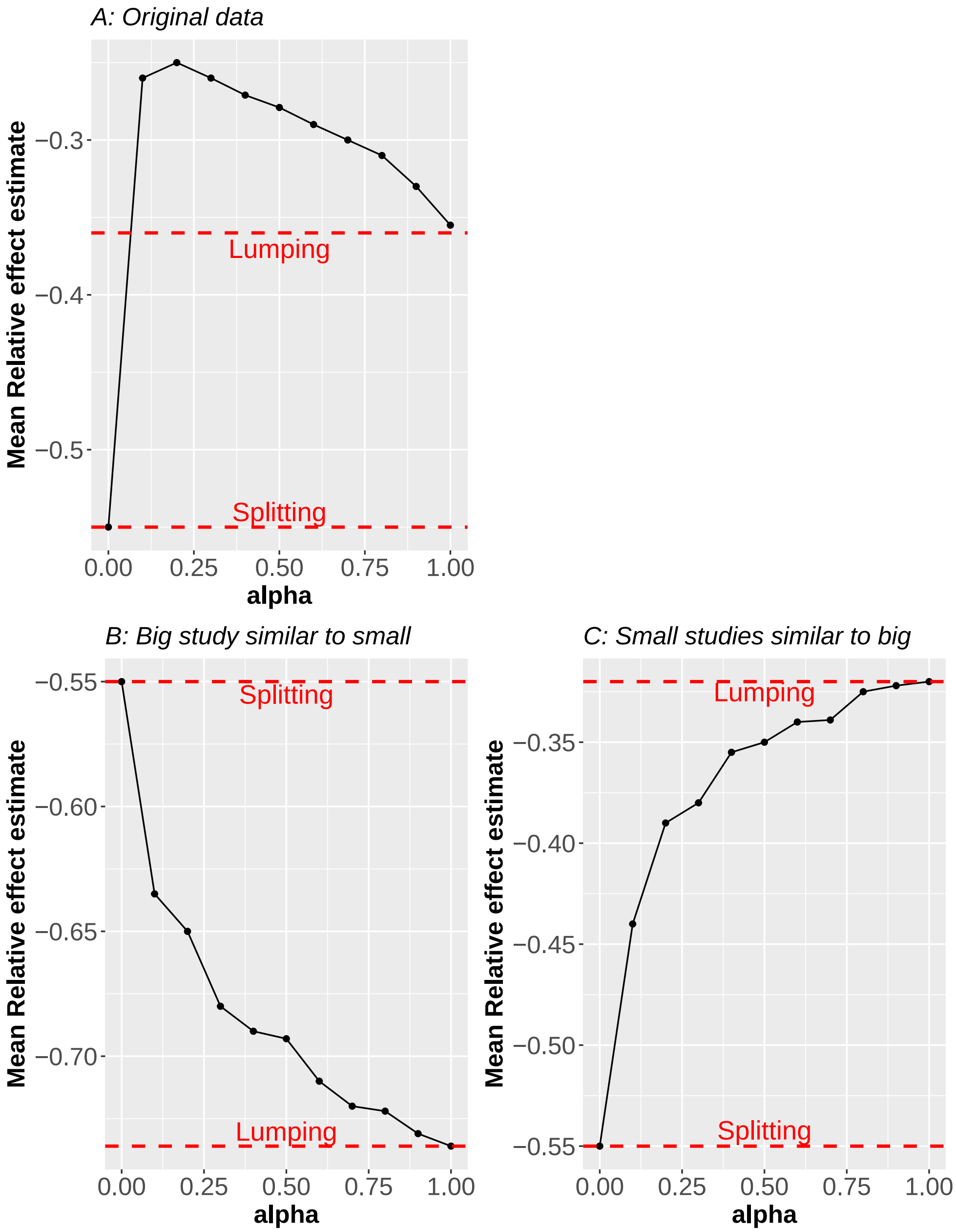


Figure A illustrates the results using the original indirect evidence base. In Figure B the big indirect study (Brocklehurst et al., 2011), which in the original data suggested a RTE of 0, is changed to suggest a similar RTE with the small indirect study (Weisman et al., 1992) (i.e. that IVIG/IVIGAM leads to a considerably lower all-cause mortality than ALB). In Figure C the small indirect study(Weisman et al., 1992) , which in the original data suggested a strong effect, is changed to suggest a similar RTE with the big study (Brocklehurst et al., 2011) (i.e. a RTE close to 0).
These scenarios provide further support of the interpretation given on page 121 which suggested that the heterogeneity of the indirect studies in the study sizes and the RTEs they suggest is causing power-prior models to produce results outside the spectrum defined by splitting and lumping.

C. Appendix to Chapter 6: Policy-implications of information-sharing: cost-effectiveness and value of information analyses

C.1. Methods

We wish to understand how knowledge of the exact relative effectiveness (IVIG/IVIGAM *vs* ALB) reduces the expected cost of uncertainty. However, the relative effect is uncertain in itself. In this process, we are repeatedly fixing the parameter of interest to a deterministic value and calculating the expected value of perfect information of the remaining probabilistic parameters. The steps are detailed below:

1. For the parameter of interest (relative effectiveness here), randomly draw a value from its distribution.

2. Fix the parameter of interest to the deterministic value drawn in Step 1, whilst allowing all the remaining parameters to be probabilistic. Run the PSA for 1000 PSA-iterations.

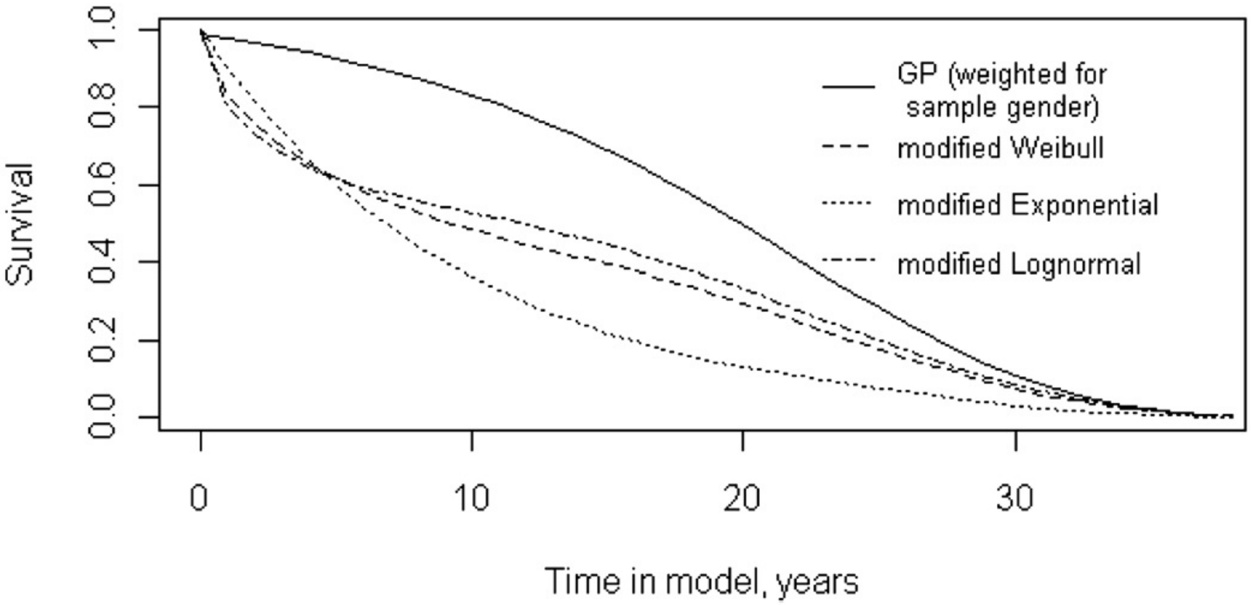
3. Record the average strategy-specific Net Benefits across all 1000 PSA-iterations as well as the maximum average Net Benefit across all strategies.

4. Repeat steps 1-3 for N=1000 times (i.e. for 1000 random draws from the relative effect distribution).

5. Average all quantities recorded in Step 3 across all replications of Step 4

6. Deduct the maximum of the average (across the replications of Step 4) strategy-specific average (across the PSA iterations) Net Benefits from the average (across the replications of Step 4) of the maximum average (across PSA iterations) Net Benefits. Multiply the resulting number with the effective population to get the EVPPI of the parameter of interest.

Figure C.1.1: *Long-term survival of patients who experienced a severe sepsis or septic shock episode. Comparison of different parametric survival curves.*



All models make the additional assumption that the a patient's probability of death cannot drop below that of the general population.

C.2. Results

Table C.2.1: *Policy measures for ISMs under the T2 RE Jadad meta-regression base-model.*

T2 RE meta-regression on Jadad score						
ISM	ICER	pCE	Pop.EVPI	Pop.EVPPI	max.ENBS	Opt.sample
Base-case	16,316	0.29	1,052	918	895	1,170
Common Heterogeneity	16,466	0.29	889	758	734	1,150
Lumping	18,423	0.34	1,113	999	976	1510
Multi-level	17,003	0.31	1,086	958	936	1,180
Informative prio	16,695	0.30	1,014	884	861	1,300
Mixture prior	16,665	0.30	1,029	898	876	1,290
Commensurate prio	17,455	0.32	1,116	993	971	1,470
Prior on heterogeneity	16,224	0.29	942	809	785	1,150
Pr-Power $\alpha = 0.1$	18,295	0.35	1,237	1,122	1,101	1,490
Pr-Power $\alpha = 0.2$	19,048	0.37	1,314	1,206	1,183	1,310
Pr-Power $\alpha = 0.3$	19,229	0.37	1,340	1,234	1,211	1,250
Pr-Power $\alpha = 0.4$	19,831	0.38	1,422	1,321	1,298	1,320
Pr-Power $\alpha = 0.5$	19,757	0.38	1,441	1,339	1,317	1,370
Pr-Power $\alpha = 0.6$	19,867	0.38	1,453	1,353	1,330	1,290
Pr-Power $\alpha = 0.7$	19,554	0.38	1,395	1,291	1,269	1,520
Pr-Power $\alpha = 0.8$	19,995	0.39	1,469	1,369	1,347	1,380
Pr-Power $\alpha = 0.9$	19,538	0.38	1,427	1,323	1,301	1,380
Pr-Power $\alpha = 1$	19,421	0.38	1,426	1,321	1,299	1,130

All measures assume $k = 30,000$ £. Population EVPI and EVPPI further assume that the technology will be relevant for 10 years. All measures have been calculated using the predictive distribution of the relative effect for a study of Jadad = 5 (i.e. of the best possible quality). Light red shade indicates the base-case (no sharing/only adults) model for comparison purposes.

Table C.2.2: *Policy measures for all ISMs used under the T3b Duration FE base-model.*

T3b FE meta-regression on sample size						
ISM	ICER	pCE	Pop.EVPI	Pop.EVPPI	max.ENBS	Opt.sample
Base-case	-117,615	0.89	83	50	20	1,620
Lumping	36,936	0.69	148	52	12	2,500
Multi-level	99,484	0.81	162	100	62	2,300
Informative prior	37,883	0.7	150	52	13	2,600
Mixture prior	39,870	0.72	156	61	16	2,500
Commensurate prior	58,832	0.76	203	123	83	2,550
Pr-Power $\alpha = 0.1$	97,905	0.85	102	54	20	1,850
Pr-Power $\alpha = 0.2$	56,188	0.80	122	59	22	2,200
Pr-Power $\alpha = 0.3$	46,763	0.76	137	63	24	2,300
Pr-Power $\alpha = 0.4$	42,601	0.74	146	64	24	2,600
Pr-Power $\alpha = 0.5$	40,325	0.72	153	64	24	2,600
Pr-Power $\alpha = 0.6$	38,893	0.71	155	63	23	2,600
Pr-Power $\alpha = 0.7$	37,762	0.70	159	63	22	2,600
Pr-Power $\alpha = 0.8$	37,141	0.69	159	61	20	2,500
Pr-Power $\alpha = 0.9$	36,831	0.69	155	57	16	2,600
Pr-Power $\alpha = 1$	36,590	0.69	153	55	14	2,500

All measures assume $k = 30,000$ £. Population EVPI and EVPPI further assume that the technology will be relevant for 10 years. All measures have been calculated using the prediction for the relative effect of a study of treatment arm sample size of 339 patients. Light red shade indicates the base-case (no sharing/only adults) model for comparison purposes.

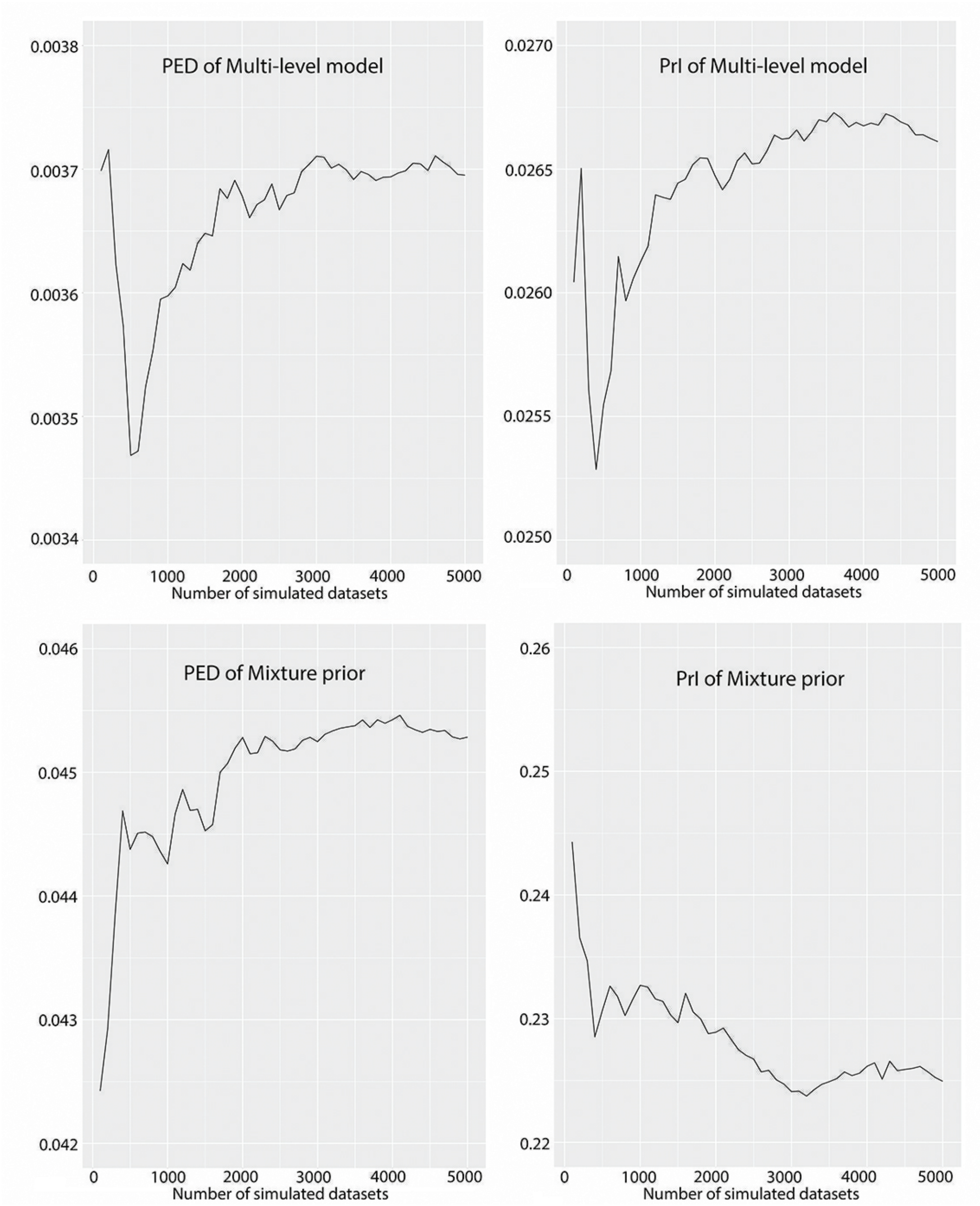
D. Appendix to Chapter 7: Comparing information-sharing methods: a simulation

D.1. Methods

Determining the appropriate number simulations needed

Since the aim is to compare the different methods according to a set of strength of sharing measures, we need to ensure that each method's strength of sharing is accurately estimated. Therefore, to determine the required number of simulations, the two most heavily parametrised and hardest to converge models (i.e. the multi-level model and the mixture of priors) were run for a varying number of simulations to check when the standard deviation of the strength of sharing measures is stabilised. The four graphs that follow show how the standard deviation of PED and PrI change with sample size. It is apparent that standard deviations are stabilised at around 3000 simulations and hence , conservatively, 5000 simulations are used in the experiment.

Figure D.1.1: *Standard deviation of strength of sharing measures (PED, PrI) for two ISMs (Multi-level, Mixture priors) under FE.*



Overlapping coefficient

The overlapping coefficient between two density curves, say $f(x), g(x)$ is defined as the area lying under the density of both curves. Mathematically, if the $f(x), g(x)$ are defined in $(-\infty, +\infty)$ the overlapping coefficient (OVL) is:

$$OVL = \int_{-\infty}^{+\infty} min\{f(x), g(x)\}dx \tag{1}$$

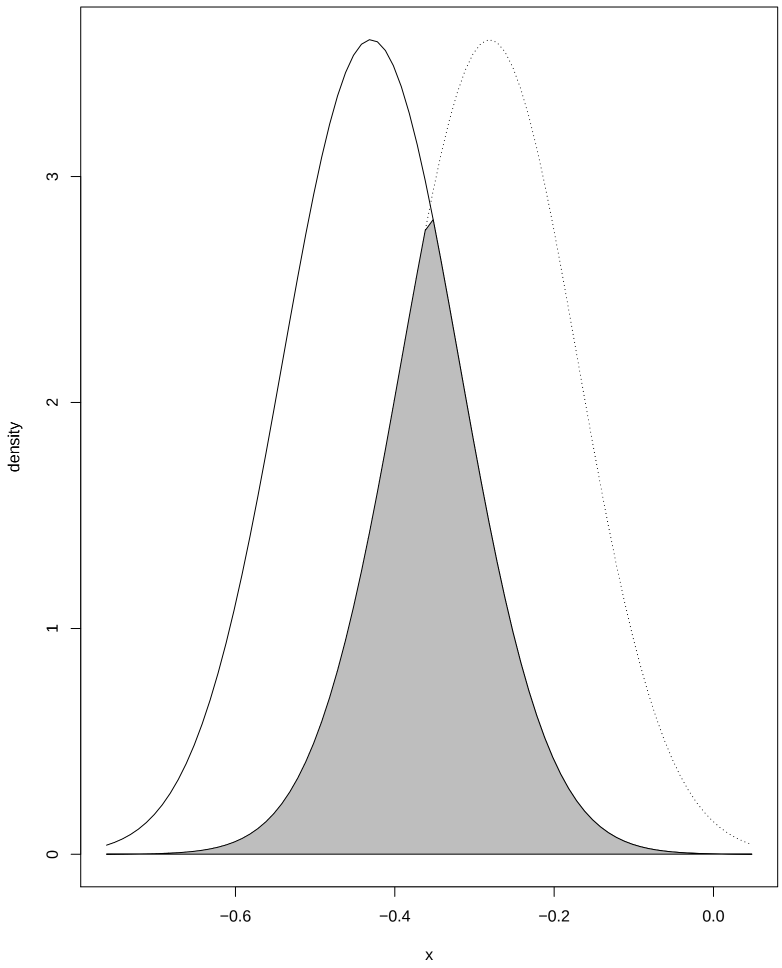
To calculate OVL in R the following code is used:

```
# Define a function that sources the two density curves
# and gives the minimum density of the two for any x
    min.f1f2 <- function(x, mu1, mu2, sd1, sd2) {
  f1 <- dnorm(x, mean=mu1, sd=sd1)
  f2 <- dnorm(x, mean=mu2, sd=sd2)
  pmin(f1, f2)
}

# Define the two densities here
mu1 <- -0.43;    sd1 <- 0.1106100
mu2 <- -0.281;   sd2 <- 0.1106100

# Integrate the function
integrate(min.f1f2, -Inf, Inf, mu1=mu1, mu2=mu2, sd1=sd1, sd2=sd2)
```

Here, to calculate the point estimate for the indirect evidence (i.e. mu2) given the point estimate and standard error of the relative treatment effect of the direct evidence (i.e. mu1 , sd1 respectively), it is assumed that sd2=sd1 i.e. direct and indirect evidence yield relative treatment effect estimates of equal standard errors. Albeit strong, this assumption is not unrealistic, particularly given the fact that in the base-case direct and indirect evidence are assumed to include an equal number of patients. It needs to be highlighted as a limitation that in the scenarios which modify the sample size of the indirect evidence this assumption is less defendable and may lead to slightly lower that 50% overlap between direct and indirect evidence.

Given the assumption of the common standard deviation the only unknown is mu2. Its value is decided here by a trial-and-error approach i.e. attempting multiple different values until the desired overlapping coefficient of 50% is reached. The densities specified above are shown in Figure D.1.2. The solid line corresponds to the direct evidence of the case-study analysed using a simple FE model (mu1, sd1) and the dotted line to the distribution that yields a 50% overlapping coefficient and falls on right i.e. towards the line of no effect where the log-odds ratio is zero.

Figure D.1.2: *An illustration of two distributions with a 50% overlapping coefficient.*



The solid line corresponds to the direct evidence of the case-study analysed using a FE model (mu1, sd1) and the dotted line to the distribution that yields a 50% overlapping coefficient and falls on right i.e. towards the line of no effect.

Back-calculation process

For this scenario we want to simulate a set of 10 indirect studies which, when synthesised all together, produce a given relative treatment effect estimate d_{all} and heterogeneity τ_{all}; these are the overall scenario characteristics of the indirect evidence. Of these 10 studies we want 9 small studies (say studies 1-9) which together include 15% of the indirect patients, and 1 big study (say study 10) which, on its own, includes 85% of the indirect patients, as this was the situation in the case-study. Furthermore, we want the big study to suggest an extreme relative effect estimate. That is because if the big study and the small studies suggest the same relative effect we will not be able to understand whether or not the overall relative effect is disproportionately influenced by the big study. Here we want the big study's effect to fall at the right tail of the predictive distribution that is defined by d_{all} and τ_{all} and, hence, suggest that the new treatment is less effective than the comparator. The challenge now is to create a process by which we can simulate such datasets, preserving the overall characteristics of the indirect evidence d_{all}, τ_{all}.

To simplify the problem we will assume that all small studies (i.e. studies 1-9) have exactly the same relative treatment effect (i.e. $d_1 = d_2 = ... = d_9 = d_{small}$). In the first step, we draw the relative effect (log-odds ratio) for the big study (i.e. $d_{10} = d_{big}$) from the right tail of the predictive distribution (randomly from the 95th-98th percentile) and assuming a probability of an event in the control arm to be $p_{ctl} = 0.366$[2], we calculate the probability of an event in the treatment arm ${p_{big}}^{trt}$. Therefore, since the study sizes (N_{big}, N_{small}) are known, we can calculate the number of events in each arm and the variance of the d_{big} with the following formula:

$$\sigma^2_{big} = \frac{1}{p_{ctl} \cdot \frac{N_{big}}{2}} + \frac{1}{(1 - p_{ctl}) \cdot \frac{N_{big}}{2}} + \frac{1}{{p_{big}}^{trt} \cdot \frac{N_{big}}{2}} + \frac{1}{(1 - {p_{big}}^{trt}) \cdot \frac{N_{big}}{2}} \tag{2}$$

The inverse-variance weights for the big study under FE and RE model can then be calculated as $w_{big} = \frac{1}{\sigma^2_{big}}$ and $w^*_{big} = \frac{1}{\sigma^2_{big} + \tau^2_{all}}$ respectively (Borenstein et al., 2009).

For the small studies, which are all assumed identical the same formula can be used to calculate the variance of d_{small} as follows:

$$\sigma^2_{small} = \frac{1}{p_{ctl} \cdot \frac{N_{small}}{2}} + \frac{1}{(1 - p_{ctl}) \cdot \frac{N_{small}}{2}} + \frac{1}{{p_{small}}^{trt} \cdot \frac{N_{small}}{2}} + \frac{1}{(1 - {p_{small}}^{trt}) \cdot \frac{N_{small}}{2}} \tag{3}$$

However, all parameters are known except for ${p_{small}}^{trt}$ and hence $\sigma_{small}, w_{small}, w^*_{small}$ can be calculated only once ${p_{small}}^{trt}$ is determined.

[2] sourced from the control arm of the direct studies in the case-study

The overall relative effect across all studies d_{all} under a FE model is defined as follows:

$$d_{all} = \frac{\sum_{n=1}^{10} w_i \cdot d_i}{\sum_{n=1}^{10} w_i} \tag{4}$$

which can be decomposed to

$$d_{all} = \frac{\sum_{n=1}^{9} w_i \cdot d_i + w_{10} \cdot d_{10}}{\sum_{n=1}^{9} w_i + w_{10}} \tag{5}$$

and given that all small studies are assumed identical

$$d_{all} = \frac{9 \cdot w_{small} \cdot d_{small} + w_{big} \cdot d_{big}}{9 \cdot w_{small} + w_{big}} \tag{6}$$

where the only unknowns are w_{small} and d_{small}; both being functions of merely the same unknown parameter i.e. ${p_{small}}^{trt}$. Therefore, since d_{all} is known (i.e. the parameter that we want to preserve), ${p_{small}}^{trt}$ can be calculated from this equation and subsequently converted to d_{small} given p_{ctl}.

Under a RE model, the process for calculating ${p_{small}}^{trt}$ is the same as above with the only difference that instead of w_{small}, we need to use w^*_{small}.

As a final step, we need to re-calculate τ^2_{all} to ensure that it has not diverted significantly from its desired value due to the randomly drawn d_{big}. To do that we use its formula as that was described in Borenstein et al., 2009, where all components are known.

$$\tau^2_{all} = \frac{Q - df}{c} = \frac{\sum_1^{10} w_i \cdot d_i^2 - \frac{(\sum_1^{10} w_i \cdot d_i)^2}{\sum_1^{10} w_i} - 9}{\sum_1^{10} w_i - \frac{\sum_1^{10} w_i^2}{\sum_1^{10} w_i}} \tag{7}$$

It is important to note that as d_{big} is drawn from a more extreme part of the predictive distribution (e.g. further at its right tail) there is going to be a point at which the back-calculation process gives us a solution for ${p_{small}}^{trt}$ under a RE model, but not under a FE model. That is because RE models give higher weights to smaller studies (i.e. the 9 non-extreme studies here) and can therefore recover the desired d_{all}. Contrariwise, FE models give lower weights to smaller studies and result easier in no solution. This is exactly the reason why this scenario was explored only under a random-effects model (i.e. to guarantee the stability of the simulation process).

The R code for the back-calculation process is provided in `https://github.com/NikolaidisGFZ/PHD.git`

Calculating τ^2 based on the desired I^2

In a meta-analytic process that uses inverse-variance weights as that is detailed in Borenstein et al., 2009 τ^2 and I^2 are defined as follows:

$$\tau^2 = \frac{Q - df}{C}, I^2 = \frac{Q - df}{Q}$$

Therefore, τ^2 and I^2 can be related through the following formula

$$\tau^2 = \frac{I^2 \cdot Q}{C}$$

where

$$Q = \sum_1^k w_i \cdot d_i^2 - \frac{(\sum_1^k w_i \cdot d_i)^2}{\sum_1^k w_i}, \quad C = \sum_1^k w_i - \frac{\sum_1^k w_i^2}{\sum_1^k w_i}$$

where w_i are the study-specific inverse-variance fixed-effect weights and d_i are the study-specific relative treatment effects.

Here, Iaim to derive τ^2 values based on the desired I^2 values. This can easily be done if we simplify the problem, by assuming that Q/C remains constant as I^2 changes. Even though Q/C is unlikely to remain constant, here we only want to obtain indicative values for τ given particular I^2.

Q/C can be calculated given the estimated $\tau^2 = 0.56^2$ and $I^2 = 68\%$ in the case-study. Then the scenario-specific τ^2 is derived based on the desired I^2.

Step-by-step simulation process

1. Re-analyse the direct evidence of the sepsis case-study, using both FE and RE models to obtain d_{FE}, d_{RE}, τ

2. Draw 5000 x 17 random samples for the study-specific relative treatments effects (log-odds ratios) of the direct evidence. For the FE simulation use the following sampling distribution $d_{dir_i} \sim (d_{FE}, \tau^2)$, whilst for the RE simulations the following $d_{dir_i} \sim (d_{RE}, \tau^2)$.

3. Draw 5000 x 17 random samples for the study-specific baseline (log-odds) nuisance parameters of the direct evidence from the following sampling distribution $\mu_i \sim (0.5, 0.5^2)$. This distribution was estimated using the log-odds of the control arms of the 17 direct studies of the sepsis case-study of Chapter 5).

4. Using the random samples of the two previous parameters, draw study-specific numbers of events for the control and the treatment arms of the direct studies $(r^{ctl}_{dir_i}, r^{trt}_{dir_i})$ as shown in the data-generating mechanism. (eqs. (7.1) to (7.4))

5. Organise the drawn numbers of events along with the fixed studies samples sizes $(n^{ctl}_{dir_i}, n^{trt}_{dir_i})$ into 5000 direct evidence datasets, each one comprising of 17 direct studies.

6. Define the properties of the indirect evidence sampling distribution, according to scenario characteristics. For example, for the base-case scenario that would be $\sim N(-0.281, \tau^2)$ for the FE simulations and $\sim N(-0.511, \tau^2)$ for the RE simulations.

7. Repeat steps 2-5 to obtain $r^{ctl}_{indir_i}, r^{trt}_{indir_i}$ and construct 5000 indirect evidence datasets, each one comprising from 10 indirect studies.

8. Combine the direct datasets obtained in Step 5, with the indirect datasets obtained in Step 7, to construct 5000 hybrid datasets, each one comprising from 27 studies (i.e. 17 direct studies and 10 indirect).

9. For every hybrid dataset, combine the direct and the indirect studies using all the applicable ISMs, including lumping and splitting, and record the method-specific ‘strengthened‘ estimate for the direct relative treatment effect.

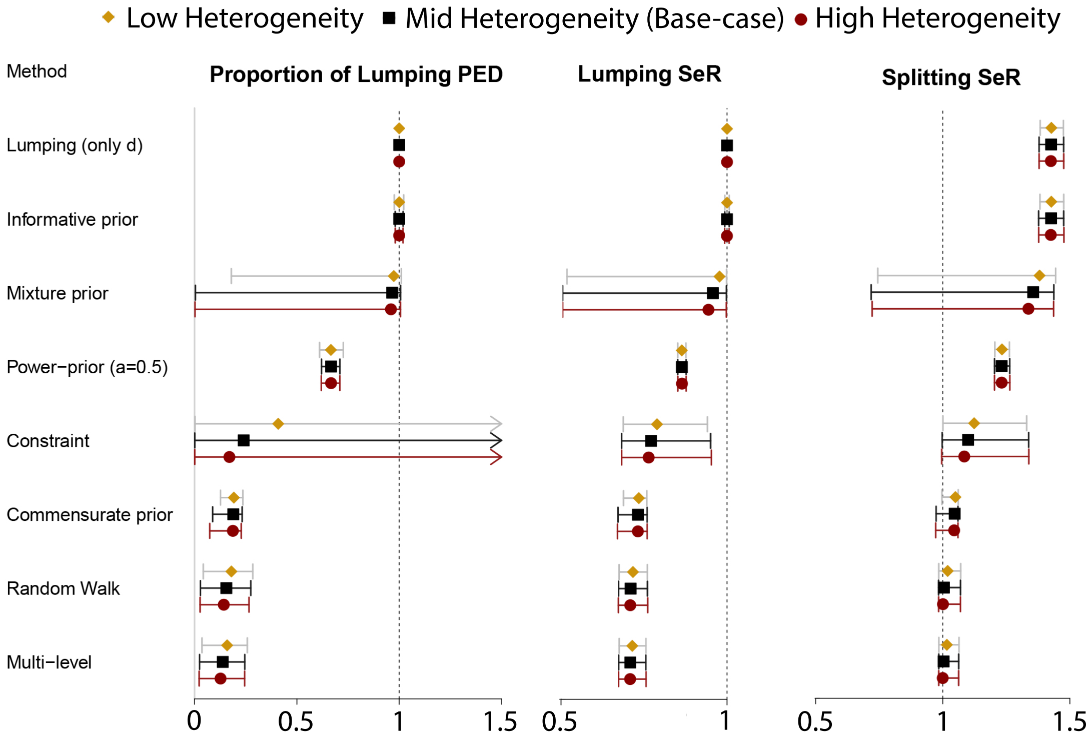
10. For each method applied in each dataset, calculate PED and PrI by comparing the dataset- and method-specific ‘strengthened‘ relative treatment effect with the relative treatment effect that is obtain by splitting the corresponding dataset and only considering the dataset-specific direct evidence.

11. For each method applied in each dataset, rank methods according to the calculated PED and PrI.

12. For each additional scenario repeat Steps 6-11, according to scenario characteristics.

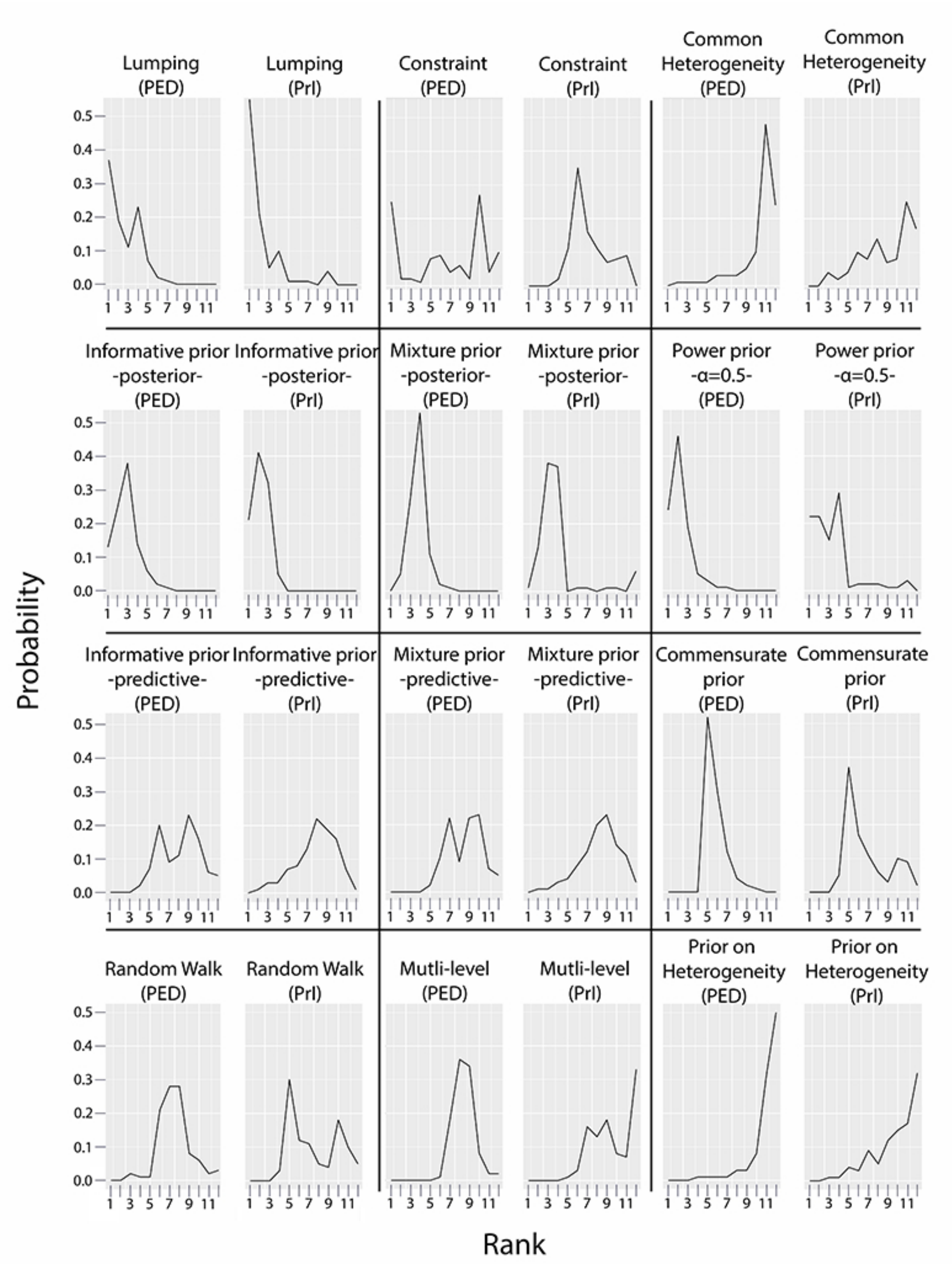
D.2. Results

Figure D.2.1: *FE simulation. PED-ratios, Lumping SeR, and Splitting SeR across all the attempted methods for three scenarios characterised by different heterogeneity of the indirect evidence.*



'Base-case scenario' corresponds to an overlapping (OVL) coefficient of 50% (LOR:-0.281, OR:0.76), 'Low overlap scenario' to OVL 5% (LOR: 0.003, OR: 1.003), and 'High overlap scenario' to OVL 75% (LOR: -0.36, OR: 0.7). The 'Proportion of Lumping PED' (i.e. PED-ratio) reflects a method's PED divided by lumping's PED. Lumping SeR is the ratio of a method's standard error divided by lumping's standard error. Splitting SeR is the ratio of a method's standard error divided by splitting's standard error.

Figure D.2.2: *Rankograms of all ISMs used in the base-case scenario of the RE simulation.*



The graphs are organised in pairs, showing the rankings according to the two strength-of-sharing measures side by side to reveal commonalities and differences. The y-axis depicts the probability of ranking in the position shown in the x-axis.

Figure D.2.3: *PED-ratios for a set of ISMs against the actual point estimate difference in the simulated direct and indirect evidence.*

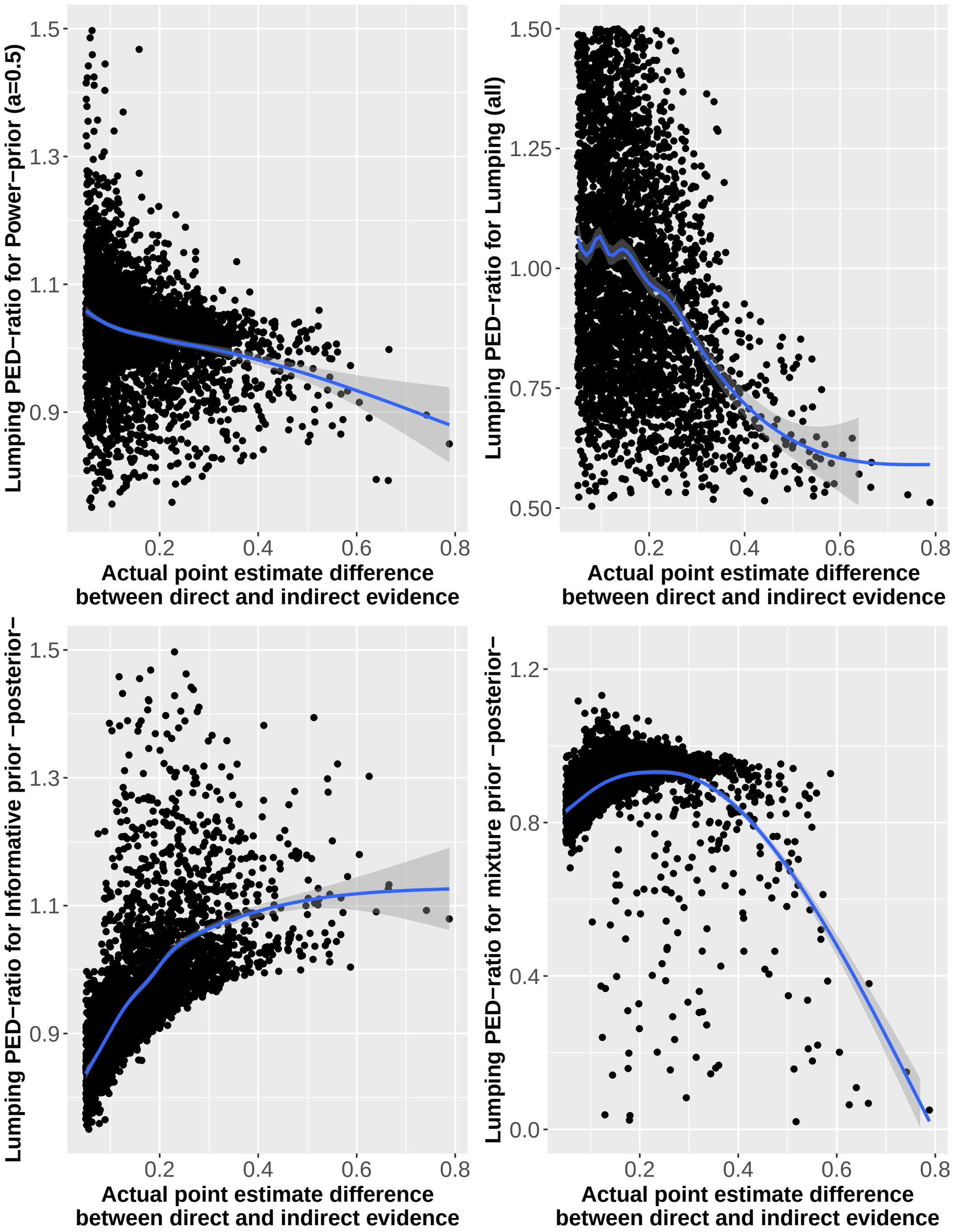


Figure D.2.4: *RE simulation. PED-ratios, Lumping SeR, and Splitting SeR of the various ISMs under three scenarios characterised by a different sample size in the indirect evidence.*

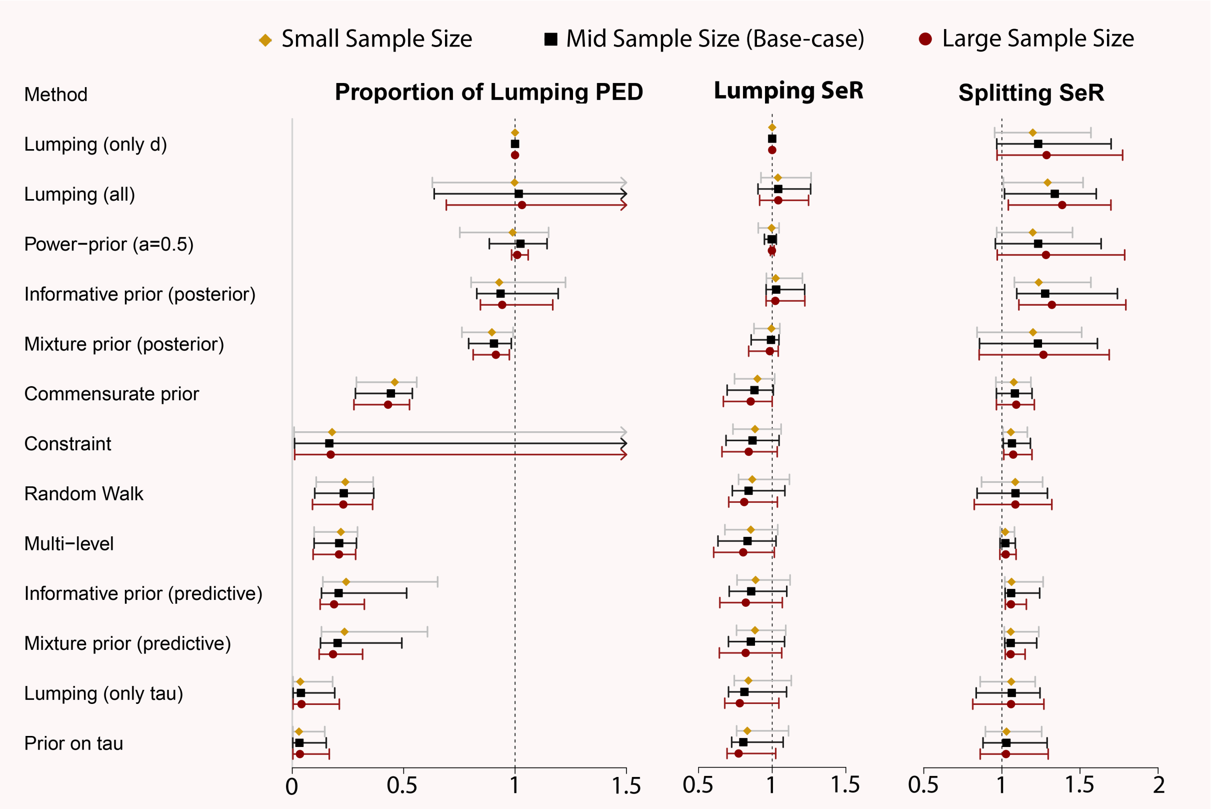


Figure D.2.5: *RE simulation. Absolute values for PED of a subset of methods under three scenarios for the heterogeneity of the indirect evidence.*

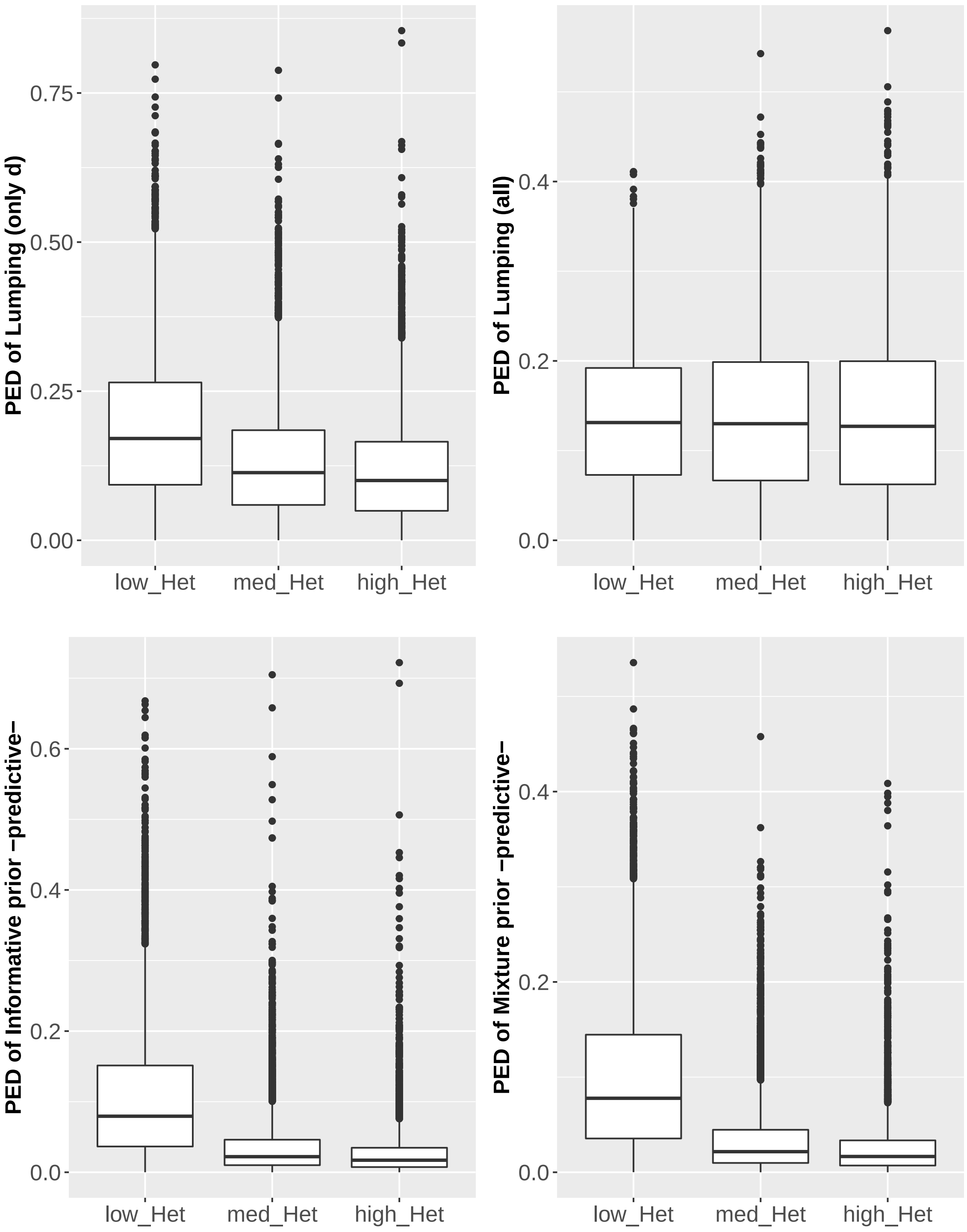
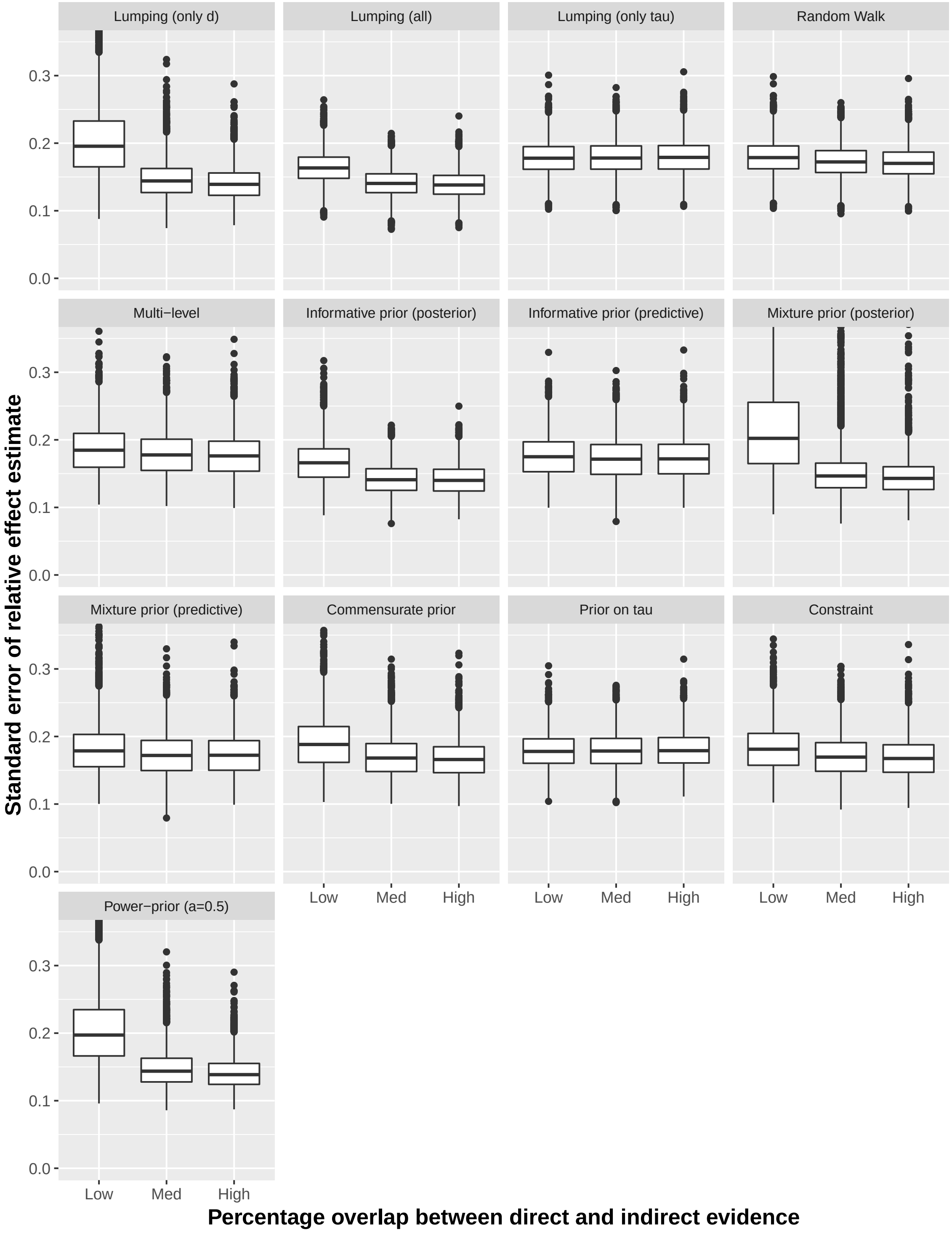


Figure D.2.6: *RE simulation. Standard error of relative effect estimates using different ISMs for three different scenarios. One where direct and indirect evidence exhibit very low percentage overlap (5%), one with medium (50%), and one with high (75%).*



Acronyms

CBM Contrast-based Model. 25, 27

CEAC Cost-Effectiveness Acceptability Curve. 10, 39

DAM Decision Analytic Model. 17

DIC Deviance Information Criterion. 29, 85, 86, 189

DSA Deterministic Sensitivity Analysis. 38

EBM Evidence-based Medicine. 35

EMA European Medicines Agency. 19, 192

ENBS Expected Net Benefit of Sample. 42

EVPI Expected Value of Perfect Information. 40–42

EVPPI Expected Value of Perfect Parameter Information. 40–42

EVSI Expected Value of Sample Information. 40, 42

FDA U.S. Food and Drugs Administration. 19, 192

FE Fixed-Effects. 8–12, 23, 25–27, 30, 60, 62, 64, 68, 74, 75, 81, 87, 101, 104, 106, 108, 111, 112, 114, 118, 123, 124, 126–128, 139, 149, 150, 152, 154–158, 162–165, 167, 170, 175, 180–183, 187, 190, 220, 226, 231, 240, 242–244, 246, 247

HTA Health Technology Assessment. 2, 18, 19, 21–24, 31, 42, 43, 61, 88, 185, 186, 188, 192–194, 197, 199

ICER Incremental Cost-Effectiveness Ratio. 36, 37

IPD Individual-patient data. 30, 60, 87, 214

ISM Information-sharing method. 2, 8–13, 22–24, 43, 44, 47, 57–60, 64, 79, 80, 85–88, 93, 101, 102, 111, 113–120, 122–129, 131, 133, 135–140, 142, 143, 146–151, 156–164, 168, 174, 176–179, 182, 186–199, 227, 228, 231–233, 238–240, 246, 248–250, 252

MA Meta-Analysis. 18, 23, 24, 27–29, 43–45, 47, 56, 191, 194, 201, 205–208

MCMC Markov Chain Monte Carlo. 29

NB Net-Benefit. 37, 38, 40, 42

NETSCC NIHR Evaluation, Trials and Studies Coordinating Centre. 192

NHS National Health Service. 16, 20

NICE National Institute for Health and Care Excellence. 16, 20, 35, 37, 39, 130, 133

NMA Network Meta-Analysis. 10, 18, 19, 22, 23, 27–30, 43–45, 47, 54, 56, 60, 68, 69, 74–77, 157, 161, 191, 194, 201, 204–208, 214

partSA Partition Survival Models. 34

PSA Probabilistic Sensitivity Analysis. 10, 38, 39, 41

QALY Quality-Adjusted Life-Year. 35, 37

QoL Quality of Life. 16–18, 35

RCT Randomised Controlled Trial. 8, 18, 35, 90, 96, 143, 147, 188, 189, 193, 196, 219

RE Random-Effects. 7–13, 23, 25–27, 29, 30, 53, 54, 60, 62, 64, 68, 74, 75, 81, 84, 87, 101–104, 106, 112, 114, 115, 118, 120, 121, 124, 126–128, 135, 139, 147, 149–158, 162, 174–183, 186, 187, 190, 203, 206, 211, 213, 227, 231, 233, 243, 244, 246, 248, 250–252

RTE Relative Treatment Effect. 7, 11, 16–19, 21, 23, 25–30, 35, 42, 52–54, 60, 64–68, 70, 74, 75, 79, 82, 83, 85, 88, 93, 101, 115, 118, 128, 129, 135, 149, 151–154, 156–161, 163, 165, 167, 168, 170, 174, 178, 181, 186–188, 190, 192, 194, 195, 197, 235

UK United Kingdom. 37

WTP Willingness to Pay. 37

Bibliography

Abrams, K., Bujkiewicz, S., Dequen, P., Jenkins, D., and Martina, R. Wp1: Deliverable 1.5 (case study review: Rheumatoid arthritis). Technical report, GetReal - Project No. 115546, 2017.

Achana, F. A., Cooper, N. J., Dias, S., Lu, G., Rice, S. J., Kendrick, D., and Sutton, A. J. Extending methods for investigating the relationship between treatment effect and baseline risk from pairwise meta-analysis to network meta-analysis. *Stat Med*, 32(5): 752–71, 2013.

Achana, F. A., Cooper, N. J., Bujkiewicz, S., Hubbard, S. J., Kendrick, D., Jones, D. R., and Sutton, A. J. Network meta-analysis of multiple outcome measures accounting for borrowing of information across outcomes. *BMC Med Res Methodol*, 14:92, 2014.

Ades, A. E. and Sutton, A. J. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(1):5–35, 2006.

Ades, A. E., Lu, G., and Claxton, K. Expected value of sample information calculations in medical decision modeling. *Medical Decision Making*, 24(2):207–227, 2004. PMID: 15090106.

Ades, A. E., Sculpher, M., Sutton, A., Abrams, K., Cooper, N., Welton, N., and Lu, G. Bayesian methods for evidence synthesis in cost-effectiveness analysis. *PharmacoEconomics*, 24:1–19, 2006.

Ades, A. E., Welton, N. J., Caldwell, D., Price, M., Goubar, A., and Lu, G. Multiparameter evidence synthesis in epidemiology and medical decision-making. *Journal of Health Services Research & Policy*, 13(3):12–22, October 2008.

Ades, A. E., Mavranezouli, I., Dias, S., Welton, N. J., Whittington, C., and Kendall, T. Network meta-analysis with competing risk outcomes. *Value Health*, 13(8):976–83, 2010.

Akdag, A., Dilmen, U., Haque, K., Dilli, D., Erdeve, O., and Goekmen, T. Role of pentoxifylline and/or IgM-enriched intravenous immunoglobulin in the management of neonatal sepsis. *Journal of Perinatology*, 31(10):905–912, November 2014.

Alejandria, M. and Marissa, M. Intravenous immunoglobulin for treating sepsis, severe sepsis and septic shock. *Cochrane Database of Systematic Reviews*, 2013.

Appleby, J., Devlin, N., and Parkin, D. Nice's cost effectiveness threshold. *BMJ*, 335(7616): 358–359, 2007.

Badampudi, D., Wohlin, C., and Petersen, K. Experiences from using snowballing and database searches in systematic literature studies. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, EASE '15, pages 17:1–17:10, New York, NY, USA, 2015. ACM.

Bayes, T. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

Behre, G., Ostermann, H., Schedel, I., Helmerking, M., Schiel, X., Rothenburger, M., Geiger, S., Dedroogh, M., Bockelmann, D., Wormann, B., Kienast, J., and Hiddemann, W. Endotoxin concentrations and therapy with polyclonal igm-enriched immunoglobulins in neutropenic cancer patients with sepsis syndrome: Pilot study and interim analysis of a randomized trial. *Anti Infect Drugs Chemother*, 13:129–134, 01 1995.

Bojke, L., Soares, M., Fox, A., Jankovic, D., Claxton, K., Morton, A., Sharples, L., Jackson, C., Taylor, A., and Colson, A. Developing a reference protocol for expert elicitation in healthcare decision making. *Health Technology Assessment Reports*, 2019.

Borenstein, M., Hedges, V. L., Higgins, P. T. J., and Rothstein, R. H. *Introduction to Meta-Analysis*. Wiley, 2009.

Briggs, A. and Sculpher, M. Sensitivity analysis in economic evaluation: A review of published studies. *Health Economics*, 4(5):355–371, 1995.

Briggs, A., Claxton, K., and Sculpher, M. J. *Decision Modelling for Health Economic Evaluation*. Oxford University Press, 2006.

Briggs, A. H., Weinstein, M. C., Fenwick, E. A. L., Karnon, J., Sculpher, M. J., and Paltiel, A. D. Model parameter estimation and uncertainty analysis: A report of the ispor-smdm modeling good research practices task force working group–6. *Medical Decision Making*, 32(5):722–732, 2012. PMID: 22990087.

Brocklehurst, P., Farrell, B., King, A., Juszczak, E., Darlow, B., Haque, K., Salt, A., Stenson, B., and Tarnow-Mordi, W. Treatment of neonatal sepsis with intravenous immune globulin. *Journal of Medicine*, 365(13):1201–1211, September 2011.

Bucher, H. C., Guyatt, G. H., Griffith, L. E., and Walter, S. D. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology*, 50(6):683 – 691, 1997.

Bujkiewicz, S., Thompson, J. R., Sutton, A. J., Cooper, N. J., Harrison, M. J., Symmons, D. P. M., and Abrams, K. R. Use of bayesian multivariate meta-analysis to estimate the haq for mapping onto the eq-5d questionnaire in rheumatoid arthritis. *Value in Health*, 2014.

Bujkiewicz, S., Thompson, J. R., Sutton, A. J., Cooper, N. J., Harrison, M. J., Symmons, D. P., and Abrams, K. R. Multivariate meta-analysis of mixed outcomes: a bayesian approach. *Statistics in Medicine*, 32(22):3926–3943, 2013.

Bujkiewicz, S., Thompson, J. R., Riley, R. D., and Abrams, K. R. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Statistics In Medicine*, 35(7, SI):1063–1089, March 2016.

Burch, J., Paulden, M., Conti, S., Stock, C., Corbett, M., Welton, N. J., Ades, A., Sutton, A., Cooper, N., Elliot, A., Nicholson, K., Duffy, S., McKenna, C., Stewart, L., Westwood, M., and Palmer, S. Antiviral drugs for the treatment of influenza: A systematic review and economic evaluation. *Health Technology Assessment*, 2008.

Burns, E., Lee, V., and Rubinstein, A. Treatment of septic thrombocytopenia with immune globulin. *Journal of Clinical Immunology*, 1991.

Caldwell, D. M., Ades, A. E., and Higgins, J. P. T. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ*, 331(7521):897–900, 2005.

Capasso, L., Borrelli, A. C., Ferrara, T., Albachiara, R., Coppola, C., and Raimondi, F. Adjuvant therapy in septic neonates with immunoglobulin preparations containing ig isotypes in addition to igg: A critical review of current literature. *Current Pediatric Research*, 21(4), 2017.

Centre for Reviews and Dissemination. Systematic reviews. crd's guidance for undertaking reviews in health care. Technical report, Centre for Reviews and Dissemination, 2006.

Chaimani, A. and Salanti, G. Using network meta-analysis to evaluate the existence of small-study effects in a network of interventions. *Research Synthesis Methods*, 3(2): 161–176, 2012.

Chaimani, A., Higgins, J. P. T., Mavridis, D., Spyridonos, P., and Salanti, G. Graphical tools for network meta-analysis in stata. *PLOS ONE*, 8(10):1–12, 10 2013.

Chen, J. Y. Intravenous immunoglobulin in the treatment of full-term and premature newborns with sepsis. *Journal of the Formosan Medical Association*, 95(11):839–844, November 1996.

Claxton, K. Exploring uncertainty in cost-effectiveness analysis. *PharmacoEconomics*, 26(9): 781–798, Sep 2008.

Claxton, K., Sculpher, M., McCabe, C., Briggs, A., Akehurst, R., Buxton, M., Brazier, J., and O'Hagan, T. Probabilistic sensitivity analysis for nice technology assessment: not an optional extra. *Health Economics*, 14(4):339–347, 2005.

Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P. C., and Sculpher, M. Methods for the estimation of the national institute for health and care excellence cost-effectiveness threshold. *Health technology assessment (Winchester, England)*, 19:1–503, v–vi, Feb 2015a.

Claxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., Devlin, N., Smith, P. C., and Sculpher, M. Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. *Health Technology Assessment (Winchester, England)*, 19(14):1–503, v–vi, February 2015b.

Claxton, K., Palmer, S., Longworth, L., Bojke, L., Griffin, S., Soares, M., Spackman, E., and Rothery, C. A comprehensive algorithm for approval of health technologies with, without, or only in research: The key principles for informing coverage decisions. *Value in Health*, 19(6):885 – 891, 2016.

Cochran, W. G. The comparison of percentages in matched samples. *Biometrika*, 37(3-4): 256–266, 12 1950.

Cooper, N. J., Sutton, A. J., Morris, D., Ades, A. E., and Welton, N. J. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in Medicine*, 28(14):1861–1881, 2009.

Copas, J. What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):95–109, 1999.

Copas, J. B., Jackson, D., White, I. R., and Riley, R. D. The role of secondary outcomes in multivariate meta-analysis. *Journal of the Royal Statistical Society Series C - Applied Statistics*, 67(5):1177–1205, November 2018.

Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration. Review manager (revman) [computer program]. version 5.3, 2014.

Corbett, M., Soares, M., Jhuti, G., Rice, S., Spackman, E., Sideris, E., Moe-Byrne, T., Fox, T., Marzo-Ortega, H., Kay, L., Woolacott, N., and Palmer, S. Tumour necrosis factor-a inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis: a systematic review and economic evaluation. *Health Technology Assessment*, 20(9), February 2016.

Corbett, M., Chehadah, F., Biswas, M., Moe-Byrne, T., Palmer, S., Soares, M., Walton, M., Harden, M., Ho, P., Woolacott, N., and Bojke, L. Certolizumab pegol and secukinumab for treating active psoriatic arthritis following inadequate response to disease-modifying antirheumatic drugs: a systematic review and economic evaluation. *Health Technol Assess*, 21(56):374, October 2017.

Cuthbertson, B. H., Roughton, S., Jenkinson, D., MacLennan, G., and Vale, L. Quality of life in the five years after intensive care: a cohort study. *Critical Care*, 14(1):R6, January 2010.

da Costa, B. R., Reichenbach, S., Keller, N., Nartey, L., Wandel, S., Juni, P., and Trelle, S. Effectiveness of non-steroidal anti-inflammatory drugs for the treatment of pain in knee and hip osteoarthritis: a network meta-analysis. *Lancet*, 390(10090):E21–E33, July 2017.

Dakin, H. A., Welton, N. J., Ades, A. E., Collins, S., Orme, M., and Kelly, S. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Stat Med*, 30(20):2511–35, 2011.

Daniels, M. J. and Hughes, M. D. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med*, 16(17):1965–82, 1997.

Daniels, R. and Nutbeam, T. *The Sepsis Manual*. United Kingdom Sepsis Trust, 2017.

Darenberg, J., Ihendyane, N., Sjolin, J., Aufwerber, E., Haidl, S., Follin, P., Andersson, J., Norrby-Teglund, A., and Group, S. S. Intravenous immunoglobulin G therapy in streptococcal toxic shock syndrome: a European randomized, double-blind, placebo-controlled trial. *Clinical Infectious Diseases*, 37(3):333–340, August 2003.

De Simone, C., Delogu, G., and Corbetta, G. Intravenous immunoglobulins in association with antibiotics: a therapeutic trial in septic intensive care unit patients. *Critical Care Medicine*, 16(1):23–26, January 1988.

Del Giovane, C., Vacchi, L., Mavridis, D., Filippini, G., and Salanti, G. Network meta-analysis models to account for variability in treatment definitions: application to dose effects. *Stat Med*, 32(1):25–39, 2013.

Department of Health. NHS reference costs 2007-2008.

DerSimonian, R. and Laird, N. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7 (3):177 – 188, 1986.

Dias, S., Welton, N. J., Caldwell, D. M., and Ades, A. E. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*, 29(7-8):932–944, March 2010a.

Dias, S., Welton, N. J., Marinho, V. C. C., Salanti, G., Higgins, J. P. T., and Ades, A. E. 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(3):613–629, 2010b.

Dias, S., Welton, N., Sutton, A., and Ades, A. *NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials*. Number TSD2 in Technical Support Document in Evidence Synthesis. National Institute for Health and Clinical Excellence, 8 2011a.

Dias, S., Sutton, A., Welton, N., and Ades, A. *NICE DSU Technical Support Document 3: Heterogeneity: Subgroups, Meta-Regression, Bias and Bias-Adjustment*. National Institute for Health and Clinical Excellence, 9 2011b.

Dias, S., Welton, N., Sutton, A., and Ades, A. *NICE DSU Technical Support Document 5: Evidence Synthesis in the Baseline Natural History Model*. Number TSD5 in NICE DSU Technical Support Document in Evidence Synthesis. National Institute for Health and Clinical Excellence, 8 2011c.

Dias, S., Ades, A., Welton, N. J., Jansen, J. P., and Sutton, A. J. *Network meta-analysis for decision-making*. John Wiley & Sons, 2018.

Ding, Y. and Fu, H. Bayesian indirect and mixed treatment comparisons across longitudinal time points. *Stat Med*, 32(15):2613–28, 2013.

Dominici, F., Parmigiani, G., Wolpert, R. L., and Hasselblad, V. Meta-analysis of migraine headache treatments: Combining information from heterogeneous designs. *Journal of the American Statistical Association*, 94(445):16–28, 1999.

Dominioni, L., Blanchi, V., Imperatori, A., Minoia, G., and Dionigi, R. High-dose intravenous IgG for treatment of severe surgical infections. *Digestive Surgery*, 1996.

Drabinski, A., Williams, G., and Formica, C. Observational evaluation of health state utilities among a cohort of sepsis patients. *Value in Health*, 4(2):128–129, 2001.

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., and Torrance, G. W. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, 2015.

Duarte, A., Mebrahtu, T., and Goncalves, P. Adalimumab, etanercept and ustekinumab for treating plaque psoriasis in children and young people: systematic review and economic evaluation. *Health Technology Assessment*, 2017.

Dumouchel, W. H. and Harris, J. E. Bayes methods for combining the results of cancer studies in humans and other species. *Journal of the American Statistical Association*, 78 (382):293–308, 1983.

Eddy, D. M., Hasselblad, V., and Shachter, R. An introduction to a bayesian method for meta-analysis: The confidence profile method. *Med Decis Making*, 10(1):15–23, 1990.

Efthimiou, O., Mavridis, D., Cipriani, A., Leucht, S., Bagos, P., and Salanti, G. An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. *Statistics in Medicine*, 33(13):2275–2287, 2014.

Efthimiou, O., Mavridis, D., Riley, R. D., Cipriani, A., and Salanti, G. Joint synthesis of multiple correlated outcomes in networks of interventions. *BIOSTATISTICS*, 16(1): 84–97, January 2015.

Efthimiou, O., Debray, T. P. A., van Valkenhoef, G., Trelle, S., Panayidou, K., Moons, K. G. M., Reitsma, J. B., Shang, A., Salanti, G., and on behalf of GetReal Methods Review Group. Getreal in network meta-analysis: a review of the methodology. *Research Synthesis Methods*, 7(3):236–263, 2016.

Efthimiou, O., Mavridis, D., Debray, T. P. A., Samara, M., Belger, M., Siontis, G. C. M., Leucht, S., Salanti, G., and on behalf of GetReal Work, P. Combining randomized and non-randomized evidence in network meta-analysis. *Statistics in Medicine*, 36(8): 1210–1226, 2017.

English, R., Lebovitz, y., and Griffin, R. *Institute of Medicine (US) Forum on Drug Discovery, Development, and Translation: Transforming Clinical Research in the United States: Challenges and Opportunities*. National Academies Press (US), 2010.

Erdem, G., Yurdakok, M., Tekinalp, G., and Ersoy, F. The use of IgM-enriched intravenous immunoglobulin for the treatment of neonatal sepsis in preterm infants. *Journal of Pediatrics*, 35(4):277–281, December 1993.

European Medicines Agency. Extrapolation of efficacy and safety in paediatric medicine development - ema/199678/2016, 2016.

EuroQol Research Foundation. *EQ-5d-5L User Guide*. 2019.

Faria, R., Woods, B., Griffin, S., Palmer, S., Sculpher, M., and Ryder, S. D. Prevention of progression to cirrhosis in hepatitis c with fibrosis: effectiveness and cost effectiveness of sequential therapy with new direct-acting anti-virals. *Alimentary Pharmacology & Therapeutics*, 44(8):866–876, 2016.

Fenwick, E., O'Brien, B. J., and Briggs, A. Cost-effectiveness acceptability curves – facts, fallacies and frequently asked questions. *Health Economics*, 13(5):405–415, 2004.

Food and Drug Administration and Center for Devices and Radiological Health. Leveraging existing clinical data for extrapolation to pediatric uses of medical devices, 2016.

Fujikawa, K., Teramukai, S., Yokota, I., and Daimon, T. A bayesian basket trial design that borrows information across strata based on the similarity between the posterior distributions of the response probability. *Biometrical Journal*, 62(2):330–338, 2020.

Gamalo-Siebers, M., Savic, J., Basu, C., Zhao, X., Gopalakrishnan, M., Gao, A., Song, G., Baygani, S., Thompson, L., Xia, H. A., Price, K., Tiwari, R., and Carlin, B. P. Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. *Pharmaceutical Statistics*, 16(4):232–249, August 2017.

Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal.*, 1(3):515–534, 09 2006a.

Gelman, A. Prior distributions for variance parameters in hierarchical models(Comment on an Article by Browne and Draper). *Bayesian Analysis*, 1(3):515–533, 2006b.

Gelman, A. and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7(4):457–472, 11 1992.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

Goodman, C. *HTA 101: Introduction to Health Technology Assessment*. National Library of Medicine (US), 2014.

Grandage, K., Slawson, D., and Shaughnessy, A. F. Site-ation pearl growing: methods and librarianship history and theory. *J Med Libr Assoc*, 3:298–304, 7 2002.

Gray, A. M., Clarke, P. M., Wolstenholme, J. L., and Wordsworth, S. *Applied Methods of Cost-effectiveness Analysis in Healthcare*. Oxford University Press, 2010.

Grundmann, R. and Hornung, M. Immunoglobulin therapy in patients with endotoxemia and postoperative sepsis–a prospective randomized study. *Progress in Clinical & Biological Research*, 1:339–349, 1988.

Haidich, A. B. Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29–37, December 2010.

Hall, M., Williams, S., DeFrances, C., and Golosinskiy, A. Inpatient care for septicemia or sepsis: a challenge for patients and hospitals. *NCHS Data Brief*, June 2011.

Haque, K. N., Zaidi, M. H., and Bahakim, H. IgM-enriched intravenous immunoglobulin therapy in neonatal sepsis. *Journal of Diseases of Children*, 142(12):1293–1296, December 1988.

Harrison, D. A., Welch, C. A., and Eddleston, J. M. The epidemiology of severe sepsis in England, Wales and Northern Ireland, 1996 to 2004: secondary analysis of a high quality clinical database, the ICNARC Case Mix Programme Database. *Critical Care*, 10 (2):R42, March 2006.

Hartung, J. and Knapp, G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20(24):3875–3889, 2001.

Hawkins, N., Scott, D. A., and Woods, B. How far do you go? efficient searching for indirect evidence. *Medical Decision Making*, 29(3):273–281, 2009.

Hentrich, M., Fehnle, K., Ostermann, H., Kienast, J., Cornely, O., Salat, C., Ubelacker, R., Buchheidt, D., Behre, G., Hiddemann, W., and Schiel, X. IgMA-enriched immunoglobulin in neutropenic patients with sepsis syndrome and septic shock: a randomized, controlled, multiple-center trial. *Critical Care Medicine*, 34(5):1319–1325, May 2006.

Hex, N., Retzler, J., Bartlett, C., and Arber, M. The Cost of Sepsis Care in the UK. *York Health Economics Consortium*, 2017.

Higgins, J. and Green, S. Cochrane handbook for systematic reviews of interventions, 2011.

Higgins, J. P. T. and Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

Higgins, J. P. T. and Whitehead, A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*, 15(24):2733–2749, 1996.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–56, 2011.

Hoch, J. S., Briggs, A. H., and Willan, A. R. Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics*, 11(5):415–430, 2002.

Holmes, E. A. F., Harris, S. D., Hughes, A., Craine, N., and Hughes, D. A. Cost-effectiveness analysis of the use of point-of-care c-reactive protein testing to reduce antibiotic prescribing in primary care. *Antibiotics*, 7(4), 2018.

Hong, C., Riley, R. D., and Chen, Y. An improved method for bivariate meta-analysis when within-study correlations are unknown. *Research Synthesis Methods*, 9(1):73–88, March 2018a.

Hong, H., Chu, H., Zhang, J., and Carlin, B. P. A bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods*, 7(1):6–22, 2016.

Hong, H., Fu, H., and Carlin, B. P. Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2018b.

Hwang, H. and DeSantis, S. M. Multivariate network meta-analysis to mitigate the effects of outcome reporting bias. *Statistics In Medicine*, 37:3254–3266, September 2018.

Ibrahim, J. G. and Chen, M.-H. Power prior distributions for regression models. *Statist. Sci.*, 2000.

IntHout, J., Ioannidis, J. P., Borm, G. F., and Goeman, J. J. Small studies are more heterogeneous than large ones: a meta-meta-analysis. *Journal of Clinical Epidemiology*, 68 (8):860 – 869, 2015.

Jackson, C. Illustration of kullback-leibler divergence calculation. `https://github.com/chjackson`, 2019.

Jackson, D., Riley, R., and White, I. R. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

Jackson, D., White, I. R., and Riley, R. D. A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression. *Biometrical Journal*, 55(2):231–245, March 2013.

Jackson, D., Rollins, K., and Coughlin, P. A multivariate model for the meta-analysis of study level survival data at multiple times. *Research Synthesis Methods*, 5(3):264–272, September 2014.

Jackson, D., White, I. R., Price, M., Copas, J., and Riley, R. D. Borrowing of strength and study weights in multivariate and network meta-analysis. *Statistical Methods in Medical Research*, 26(6):2853–2868, December 2017.

Jackson, D., Bujkiewicz, S., Law, M., Riley, R. D., and White, I. R. A matrix-based method of moments for fitting multivariate network meta-analysis models with multiple outcomes and random inconsistency effects. *Biometrics*, 74(2):548–556, June 2018.

Jackson, D. and Riley, R. D. A refined method for multivariate meta-analysis and meta-regression. *Statistics In Medicine*, 33(4):541–554, February 2014.

Jadad, A. R., Moore, R., Carroll, D., Jenkinson, C., Reynolds, D. M., Gavaghan, D. J., and McQuay, H. J. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, 17(1):1 – 12, 1996.

Karatzas, S., Boutzouka, E., Venetsanou, K., Myrianthefs, P., Fildisis, G., and Baltopoulos, G. The effects of IgM-enriched immunoglobulin preparations in patients with severe sepsis: another point of view. *Critical Care (London, England)*, 6(6):543–544, December 2002.

Karnon, J., Stahl, J., Brennan, A., Caro, J. J., Mar, J., and Möller, J. Modeling using Discrete Event Simulation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. *Value in Health*, 15(6):821 – 827, 2012.

Kirkham, J. J., Riley, R. D., and Williamson, P. R. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics In Medicine*, 31(20):2179–2195, September 2012.

Kola, E., Celaj, E., Bakalli, I., Lluka, R., Sala, D., and Sallabanda, S. Efficacy of an IgM preparation in the treatment of patients with sepsis: A double-blind randomized clinical trial in a pediatric intensive care unit. *Intensive Care Medicine*, September 2014.

Kullback, S. and Leibler, R. A. On information and sufficiency. *Ann. Math. Statist.*, 22(1): 79–86, 03 1951.

Langford, O., Aronson, J. K., van Valkenhoef, G., and Stevens, R. J. Methods for meta-analysis of pharmacodynamic dose-response data with application to multi-arm studies of alogliptin. *Statistical Methods in Medical Research*, 27(2):564–578, February 2018.

Lee, J. Meta-analysis. *J Korean Endocr Soc*, 2008.

Leon-Novelo, L. G., Bekele, B. N., Muller, P., Quintana, F., and Wathen, K. Borrowing strength with nonexchangeable priors over subpopulations. *Biometrics*, 68(2):550–8, 2012.

Lindquist, L., Lundbergh, P., and Maasing, R. Pepsin-treated human gamma globulin in bacterial infections: A randomized study in patients with septicaemia and pneumonia. *Vox Sanguinis*, 40(5):329–337, 1981.

Liu, Y., DeSantis, S. M., and Chen, Y. Bayesian mixed treatment comparisons meta-analysis for correlated outcomes subject to reporting bias. *Journal of the Royal Statistical Society Series C - Applied Statistics*, 67(1):127–144, January 2018.

Lu, G. and Ades, A. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805, 2009.

Lu, G. and Ades, A. E. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, 23(20):3105–3124, 2004.

Lu, G., Ades, A. E., Sutton, A. J., Cooper, N. J., Briggs, A. H., and Caldwell, D. M. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Stat Med*, 26(20):3681–99, 2007.

Lu, G., Kounali, D., and Ades, A. E. Simultaneous Multioutcome Synthesis and Mapping of Treatment Effects to a Common Scale. *Value in Health*, 17(2):280–287, March 2014.

Lumley, T. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, 21(16):2313–2324, 2002.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. *The BUGS Book: A practical introduction to Bayesian Analysis*. Chapman and Hall/CRC, 2013.

Madan, J., Chen, Y.-F., Aveyard, P., Wang, D., Yahaya, I., Munafo, M., Bauld, L., and Welton, N. Synthesis of evidence on heterogeneous interventions with multiple outcomes recorded over multiple follow-up times reported inconsistently: a smoking cessation case-study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(1): 295–314, 2014.

Mak, A., Cheung, M. W. L., Ho, R. C.-M., Cheak, A. A.-C., and Lau, C. S. Bisphosphonates and atrial fibrillation: Bayesian meta-analyses of randomized controlled trials and observational studies. *BMC Musculoskeletal Disorders*, 10, September 2009.

Mancilla-Ramirez, J., Gonzalez-Yunes, R., Castellanos-Cruz, C., Garcia-Roca, P., and Santos-Preciado, J. I. [Intravenous immunoglobulin in the treatment of neonatal septicemia]. [Spanish]. *Boletin Medico del Hospital Infantil de Mexico*, 49(1):4–11, January 1992.

Masaoka, T., Hasegawa, H., Takaku, F., Mizoguchi, H., Asano, S., Ikeda, Y., Urabe, A., Shibata, A., Saito, H., Okuma, M., Horiuchi, A., Saito, Y., Ozawa, K., Usami, M., and Ohashi, Y. The efficacy of intravenous immunoglobulin in combination therapy with antibiotics for severe infections. *Japanese Journal of Chemotherapy*, 48(3):199–217, 2000.

Mavridis, D. and Salanti, G. A practical introduction to multivariate meta-analysis. *Stat Methods Med Res*, 22(2):133–58, 2013.

Mavridis, D., Sutton, A., Cipriani, A., and Salanti, G. A fully bayesian application of the copas selection model for publication bias extended to network meta-analysis. *Stat Med*, 32(1):51–66, 2013.

Mawdsley, D., Bennetts, M., Dias, S., Boucher, M., and Welton, N. J. Model-based network meta-analysis: A framework for evidence synthesis of clinical trial data. *CPT Pharmacometrics Syst Pharmacol*, 5(8):393–401, 2016.

McCabe, C., Claxton, K., and Culyer, A. J. The nice cost-effectiveness threshold. *PharmacoEconomics*, 26(9):733–744, Sep 2008.

McCarron, C. E., Pullenayegum, E. M., Thabane, L., Goeree, R., and Tarride, J.-E. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms. *BMC Medical Research Methodology*, 10, July 2010.

McCarron, C. E., Pullenayegum, E. M., Thabane, L., Goeree, R., and Tarride, J.-E. Bayesian Hierarchical Models Combining Different Study Types and Adjusting for Covariate Imbalances: A Simulation Study to Assess Model Performance. *PLOS ONE*, 6(10), October 2011.

McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. Chapman and Hall/CRC, 1989.

McDaid, C., Griffin, S., Weatherly, H., Duree, K., and van der Burgt, M. Continuous positive airway pressure devices for the treatment of obstructive sleep apnoea-hypopnoea syndrome: a systematic review and economic analysis. *Health Technology Assessment*, 13 (4), February 2009.

McElreath, R. *Statistical Rethinking*. Chapman and Hall/CRC, 2016.

McKenna, C., Soares, M., Claxton, K., Bojke, L., Griffin, S., Palmer, S., and Spackman, E. Unifying research and reimbursement decisions: Case studies demonstrating the sequence of assessment and judgments required. *Value in Health*, 18(6):865 – 875, 2015.

Melendez-Torres, G. J., Bonell, C., and Thomas, J. Emergent approaches to the meta-analysis of multiple heterogeneous complex interventions. *BMC Medical Research Methodology*, 15, June 2015.

Mills, E. J., Thorlund, K., and Ioannidis, J. P. A. Calculating additive treatment effects from multiple randomized trials provides useful estimates of combination therapies. *Journal of Clinical Epidemiology*, 65(12):1282–1288, December 2012.

Moreno, S. G., Sutton, A. J., Ades, A. E., Cooper, N. J., and Abrams, K. R. Adjusting for publication biases across similar interventions performed well when compared with gold standard data. *Journal of Clinical Epidemiology*, 64(11):1230–1241, November 2011.

Morris, T. P., White, I. R., and Crowther, M. J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019.

MRC Biostatistics Unit. The bugs project, 2010.

Musekiwa, A., Manda, S. O. M., Mwambi, H. G., and Chen, D.-G. Meta-Analysis of Effect Sizes Reported at Multiple Time Points Using General Linear Mixed Model. *PLOS ONE*, 11(10), October 2016.

Nam, I., Mengersen, K., and Garthwaite, P. Multivariate meta-analysis. *Statistics In Medicine*, 22(14):2309–2333, July 2003.

National Institute for Health and Care Excellence. *Sepsis: recognition, assessment and early management (NICE Guideline 51)*. July 2016.

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Bailey, S. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, 15 (2):123–134, 2016.

NICE. Guide to the methods of technology appraisal 2013. *NICE process and methods Guides*, 04 2013.

Nixon, R. M., Bansback, N., and Brennan, A. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Stat Med*, 26(6):1237–54, 2007.

Ntzoufras, I. *Bayesian Modeling Using WinBUGS*. John Wiley & Sons, 2008.

Ohlssen, D., Price, K. L., Amy Xia, H., Hong, H., Kerman, J., Fu, H., Quartey, G., Heilmann, C. R., Ma, H., and Carlin, B. P. Guidance on the implementation and reporting of a drug safety bayesian network meta-analysis. *Pharmaceutical Statistics*, 13(1):55–70, 2014.

Owen, R. K., Tincello, D. G., and Keith, R. A. Network meta-analysis: development of a three-level hierarchical modeling approach incorporating dose-related constraints. *Value Health*, 18(1):116–26, 2015.

Petrou, S. and Gray, A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *BMJ*, 342, 2011.

Philips, Z., Ginnelly, L., Sculpher, M., Claxton, K., and Golder, S. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technology Assessment*, 8, 2004.

Poole, D. and Raftery, A. E. Inference for deterministic simulation models: The bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255, 2000.

Prevost, T., Abrams, K., and Jones, D. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics In Medicine*, 19(24):3359–3376, December 2000.

Pullenayegum, E. M. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics In Medicine*, 30(26):3082–3094, November 2011.

R Development Core Team. The r project for statistical computing. *The R foundation for statistical computing*, 2010.

Ren, S., Oakley, J. E., and Stevens, J. W. Incorporating Genuine Prior Information about Between-Study Heterogeneity in Random Effects Pairwise and Network Meta-analyses. *Medical Decision Making*, 38(4):531–542, May 2018.

Rhodes, K. M., Turner, R. M., and Higgins, J. P. T. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology*, 68(1):52–60, January 2015.

Riemsna, R., Lhachimi, S., Armstrong, N., van Asselt, A., Allen, A., Manning, N., Harker, J., Tushabe, D., Severens, J., and Kleijnen, J. *Roflumilast for the management of severe chronic obstructive pulmonary disease: A single technology appraisal*. National Academies Press (US), 2011.

Rietbergen, C. *Quantitative Evidence Synthesis with Power Priors*. PhD thesis, University of Utrecht, 2016.

Rietbergen, C., Groenwold, R. H. H., Hoijtink, H. J. A., Moons, K. G. M., and Klugkist, I. Expert elicitation of study weights for bayesian analysis and meta-analysis. *Journal of Mixed Methods Research*, 10(2):168–181, April 2016.

Riley, R. D., Abrams, K. R., Lambert, P. C., Sutton, A. J., and Thompson, J. R. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics In Medicine*, 26(1):78–97, January 2007a.

Riley, R. D., Thompson, J. R., and Abrams, K. R. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, 9 (1):172–86, 2008.

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., and Thompson, J. R. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, 7, January 2007b.

Riley, R. D., Jackson, D., Salanti, G., Burke, D. L., Price, M., Kirkham, J., and White, I. R. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ*, 358, 2017.

Rodgers, M., Epstein, D., Bojke, L., Yang, H., Craig, D., Fonseca, T., Myers, L., Bruce, I., Chalmers, R., Bujkiewicz, S., Lai, M., Cooper, N., Abrams, K., Spiegelhalter, D., Sutton, A., Sculpher, M., and Woolacott, N. Etanercept, infliximab and adalimumab for the treatment of psoriatic arthritis: a systematic review and economic evaluation. *Health Technology Assessment*, 15(10):1+, February 2011.

Rodriguez, A., Rello, J., Neira, J., Maskin, B., Ceraso, D., Vasta, L., and Palizas, F. Effects of high-dose of intravenous immunoglobulin and antibiotics on survival for severe sepsis undergoing surgery. *Shock*, 23(4):298–304, April 2005.

Roever, C., Wandel, S., and Friede, T. Model averaging for robust extrapolation in evidence synthesis. *Statistics In Medicine*, 38(4, SI):674–694, February 2019.

Salanti, G., Marinho, V., and Higgins, J. P. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *J Clin Epidemiol*, 62(8): 857–64, 2009.

Salanti, G., Dias, S., Welton, N. J., Ades, A. E., Golfinopoulos, V., Kyrgiou, M., Mauri, D., and Ioannidis, J. P. Evaluating novel agent effects in multiple-treatments meta-regression. *Stat Med*, 29(23):2369–83, 2010.

Salanti, G., Ades, A., and Ioannidis, J. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology*, 64(2):163–171, 2011.

Salanti, G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research synthesis methods*, 3:80–97, Jun 2012.

Samatha, S., Jalalu, M., Hegde, R., Vishwanath, D., and PP, M. Role of igm-enriched intravenous immunoglobulin as an adjuvant to antibiotics in neonatal sepsis. *Karnataka Pediatric Journal*, 11:1–6, 1997.

Savitz, L. and Savitz, S. Can delivery systems use cost-effectiveness analysis to reduce healthcare costs and improve value? *F1000Research*, 5(2575), 2016.

Schedel, I., Dreikhausen, U., Nentwig, B., Hockenschnieder, M., Rauthmann, D., Balikcioglu, S., Coldewey, R., and Deicher, H. Treatment of gram-negative septic shock with an immunoglobulin preparation: a prospective, randomized clinical trial. *Critical Care Medicine*, 19(9):1104–1113, September 1991.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

Schmitz, S., Adams, R., and Walsh, C. Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics in Medicine*, 32(17):2935–2949, 2013.

Shenoi, A., Nagesh, N. K., Maiya, P. P., Bhat, S. R., and Subba Rao, S. D. Multicenter randomized placebo controlled trial of therapy with intravenous immunoglobulin in decreasing mortality due to neonatal sepsis. *Indian Pediatrics*, 36(11):1113–1118, November 1999.

Siebert, U., Alagoz, O., Bayoumi, A. M., Jahn, B., Owens, D. K., Cohen, D. J., and Kuntz, K. M. State-Transition Modeling: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. *Value in Health*, 15(6):812 – 820, 2012.

Soares, M. *Making best use of evidence for explicit decisions in health care*. PhD thesis, University of York, 2017.

Soares, M., Welton, N., Harrison, D., Peura, P., and Shankar, H. 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 Technology Assessment*, 16(7), 2012.

Soares, M. O., Dumville, J. C., Ades, A. E., and Welton, N. J. Treatment comparisons for decision making: facing the problems of sparse and few data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(1):259–279, 2014a.

Soares, M., Welton, N., Harrison, D., Peura, P., Shankar, H., Harvey, S., Madan, J., Ades, A., Rowan, K., and Palmer, S. Intravenous immunoglobulin for severe sepsis and septic shock: clinical effectiveness, cost-effectiveness and value of a further randomised controlled trial. *Critical Care*, 2014b.

Spannbrucker, N., Munch, H., Kunze, R., and Vogel, F. Auswirkungen von immunglobulinsubstitution bei sepsis. *Intensivmedizin*, 6:314, 1987.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2002.

Spiegelhalter, D. J. and Best, N. G. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Stat Med*, 22(23):3687–709, 2003.

Spiegelhalter, J. D., Abrams, R., and Myles, P. *Bayesian approaches to clinical trials and health-care evaluation*. Wiley, 2004.

Stinnett, A. A. and Mullahy, J. Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, 18(2_suppl):S68–S80, 1998. PMID: 9566468.

Strong, M., Oakley, J. E., and Brennan, A. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: A non-parametric regression approach. *Medical Decision Making*, 34(3):311–326, 2014. PMID: 24246566.

Sturtz, S., Uwe, L., and Gelman, A. R2openbugs: A package for running openbugs from r.

Sutton, A. Introduction to decision modelling. `https://slideplayer.com/slide/4757593/`, 2016. Online; accessed 3rd October 2019.

Sweeting, M. J., Sutton, A. J., and Lambert, P. C. What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, 23(9):1351–1375, 2004.

Tan, S. H., Abrams, K. R., and Bujkiewicz, S. Bayesian Multiparameter Evidence Synthesis to Inform Decision Making: A Case Study in Metastatic Hormone-Refractory Prostate Cancer. *Medical Decision Making*, 38(7):834–848, October 2018.

Thompson, K. M. and Evans, J. S. The value of improved national exposure information for perchloroethylene (perc): A case study for dry cleaners. *Risk Analysis*, 17(2):253–271, 1997.

Thorlund, K., Thabane, L., and Mills, E. J. Modelling heterogeneity variances in multiple treatment comparison meta-analysis. are informative priors the better solution? *BMC Medical Research Methodology*, 13:2–2, 2013. 1471-2288-13-2[PII] 23311298[pmid] BMC Med Res Methodol.

Trinquart, L., Chatellier, G., and Ravaud, P. Adjustment for reporting bias in network meta-analysis of antidepressant trials. *BMC Medical Research Methodology*, 12:150–150, 2012. 1471-2288-12-150[PII] 23016799[pmid] BMC Med Res Methodol.

Tugrul, S., Ozcan, P. E., Akinci, O., Seyhun, Y., Cagatay, A., Cakar, N., and Esen, F. The effects of IgM-enriched immunoglobulin preparations in patients with severe sepsis [ISRCTN28863830]. *Critical Care (London, England)*, 6(4):357–362, August 2002.

Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S., and Thompson, S. G. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):21–47, 2009.

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., and Higgins, J. P. T. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics In Medicine*, 34(6):984–998, March 2015.

van Houwelingen, H. C., Arends, L. R., and Stijnen, T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in medicine*, 21:589–624, Feb 2002a.

Van Houwelingen, H., Zwinderman, K., and Stijnen, T. A bivariate approach to meta-analysis. *Statistics In Medicine*, 12:2273–2284, December 1993.

van Houwelingen, H., Arends, L., and Stijnen, T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics In Medicine*, 21(4):589–624, February 2002b.

Verde, P. E. and Ohmann, C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods*, 6(1): 45–62, March 2015.

Warner, K. E. Issues in cost effectiveness in health care. *Journal of Public Health Dentistry*, 49(5):272–278, 1989.

Warren, F. C., Abrams, K. R., and Sutton, A. J. Hierarchical network meta-analysis models to address sparsity of events and differing treatment classifications with regard to adverse outcomes. *Statistics In Medicine*, 33(14):2449–2466, 2014.

Wei, Y. and Higgins, J. P. T. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics In Medicine*, 32(7):1191–1205, March 2013a.

Wei, Y. and Higgins, J. P. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(17):2911–2934, 2013b.

Weisman, L. E., Stoll, B. J., Kueser, T. J., Rubio, T. T., Frank, C. G., Heiman, H. S., Subramanian, K. N., Hankins, C. T., Anthony, B. F., and Cruess, D. F. Intravenous immune globulin therapy for early-onset sepsis in premature neonates. *Journal of Pediatrics*, 121(3):434–443, September 1992.

Weitzman, M. S. Measures of overlap of income distributions of white and negro families in the united states, 1970. "A United States Department of Commerce publication.".

Welton, N. J., Cooper, N. J., Ades, A. E., Lu, G., and Sutton, A. J. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza a and b. *Stat Med*, 27(27):5620–39, 2008.

Welton, N. J., Ades, A. E., Carlin, J. B., Altman, D. G., and Sterne, J. A. C. Models for Potentially Biased Evidence in Meta-Analysis Using Empirically Based Priors. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 172(1):119–136, 2009a.

Welton, N. J., Caldwell, D. M., Adamopoulos, E., and Vedhara, K. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *Am J Epidemiol*, 169(9):1158–65, 2009b.

Welton, N. J., Willis, S. R., and Ades, A. E. Synthesis of survival and disease progression outcomes for health technology assessment of cancer therapies. *Research Synthesis Methods*, 1(3-4):239–257, 2010.

Welton, N., Soares, M., Palmer, S., Ades, A., Harrison, D., Shankar, H., and Rowan, K. Accounting for heterogeneity in relative treatment effects for use in cost-effectiveness models and value-of-information analyses. *Medical Decision Making*, 2015.

Werdan, K., Pilz, G., Bujdoso, O., Fraunberger, P., Neeser, G., Schmieder, R. E., Viell, B., Marget, W., Seewald, M., Walger, P., Stuttmann, R., Speichermann, N., Peckelsen, C., Kurowski, V., Osterhues, H.-H., Verner, L., Neumann, R., and Muller-Werdan, U. Score-based immunoglobulin G therapy of patients with sepsis: the SBITS study. *Critical Care Medicine*, 35(12):2693–2701, December 2007.

Wesoly, C., Kipping, N., and Grundmann, R. [Immunoglobulin therapy of postoperative sepsis]. [German]. *Zeitschrift fur Experimentelle Chirurgie, Transplantation, und Kunstliche Organe*, 23(4):213–216, 1990.

Wolpert, R. L. and Kerrie, L. M. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: Effects of environmental tobacco smoke. *Statistical Science*, 19(3):450–471, 2004.

Woods, B., Sideris, E., Palmer, S., Latimer, N., and M, S. Nice dsu technical support document 19. partitioned survival analysis for decision modelling in health care: A critical review, 2017.

Woods, B., Rothery, C., Revill, P., Hallett, T., Phillips, A., and Claxton, K. Che research paper 155: Setting research priorities in global health: Appraising the value of evidence generation activities to support decision-making in health care, 2018.

World Health Organization, 2018.

Wu, J., Banerjee, A., Jin, B., Menon, S. M., Martin, S. W., and Heatherington, A. C. Clinical dose-response for a broad set of biological products: A model-based meta-analysis. *Statistical Methods in Medical Research*, 27(9):2694–2721, September 2018.

Yakut, M., Cetiner, S., Akin, A., Tan, A., Kaymakcioglu, N., Simsek, A., and Sen, D. Effects of immunuglobulin G on surgical sepsis and septic shock. [Turkish]. *Bulletin of Gulhane Military Medical Academy*, 1998.

Yildizidas, D., Yapicioglu, H., Tumgor, G., and Erbey, F. Does polyclonal intravenous immunoglobulin reduce mortality in septic children in pediatric intensive care unit? [cocuk yogun bakim unitesi'nde sepsis nedeni ile izlenen hastalarda poliklonal intravenoz immunglobulin tedavisi mortaliteyi azaltiyor mu? *Cocuk Sagligi Ve Hastaliklari Dergisi*, 48:136–41, 2005.