Calibration of Treatment Effects in Network Meta-Analysis using Individual Patient Data

David Mark Phillippo

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences.

Bristol Medical School

June 2019

Word count: 67 856

Abstract

Health technology assessments require reliable estimates of relative treatment effects for a given patient population, to inform decision making. Standard network meta-analysis (NMA) and indirect comparisons combine aggregate data (AgD) from multiple studies on treatments of interest, assuming that any treatment effect modifiers are balanced across populations. This assumption can be relaxed if individual patient data (IPD) are available from all studies, using an IPD network meta-regression (NMR). However, in many cases IPD are only available from one or a subset of studies.

Recently proposed methods for population-adjusted indirect comparisons aim to adjust for differences between one IPD study and one AgD study. However, the resulting comparison is only valid in the AgD study population without additional assumptions, and the methods cannot be extended to larger treatment networks. Meta-regression approaches can be used in larger networks, but typically incur aggregation bias.

In this thesis, we begin by reviewing the literature on population adjustment and related problems, giving a critique of current methods. We review applications of current methods in the published literature and in National Institute for Health and Care Excellence technology appraisals. Motivated by these reviews we propose a general method, Multilevel Network Meta-Regression (ML-NMR), that overcomes some of the disadvantages of current approaches and reduces to AgD NMA and IPD NMR as special cases. We discuss the computational aspects of implementing ML-NMR, before applying to a real example of plaque psoriasis treatments. The ML-NMR framework is then extended to handle more general likelihoods, illustrated with an artificial example of survival outcomes and a reanalysis of the plaque psoriasis example incorporating multiple outcomes. An extensive simulation study is conducted to assess the performance of ML-NMR and current methods in a range of scenarios and under various failures of assumptions. We conclude with a discussion and suggestions for future research.

Acknowledgements

This work was supported by the United Kingdom Medical Research Council grant MR/P015298/1.

I am hugely thankful to my supervisors, Nicky Welton and Sofia Dias, whose continued advice, support, guidance, and encouragement throughout my PhD has been invaluable. My sincere thanks also to Tony Ades, who has advised and guided me from the beginning. I am very grateful to be part of an excellent research group at the University of Bristol: my thanks to everyone in the Multi-Parameter Evidence Synthesis group.

I would like to thank Mark Belger, Alan Brnabic, Daniel Saure, Alexander Schacht, and Zbigniew Kadziola at Eli Lilly and Company for their valuable collaboration and for providing the plaque psoriasis data.

I am very thankful to my office mates, past and present, for their friendship: Daisy, Grace, Emily, Ben, Hugo, Edna, and Howard.

Many thanks are of course due to my parents Julie and Steve, my brother Matthew, and the rest of my family. I am very grateful to be surrounded in life by close friends, including James, Katy, Matt, Ruth, Pete, Pip, Chris, Emily, Claire, James, Lizzie, Ryan, Rachel, Alex, James, Joel, Mim, and Emilia. I am forever thankful to Jennie, my wife and best friend, for her unwavering love, support, and joy every day. Lastly, my thanks go to Jesus, my strength and purpose: this is for you.

Declaration

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

SIGNED: _____

DATE: 25th June 2019

List of Publications

Parts of this thesis have been published as follows:

Phillippo, D. M., A. E. Ades, S. Dias, S. Palmer, K. R. Abrams, and N. J. Welton (2016). *NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE*. Tech. rep. National Institute for Health and Care Excellence. URL: http://www.nicedsu.org.uk.

Phillippo, D. M., A. E. Ades, S. Dias, S. Palmer, K. R. Abrams, and N. J. Welton (2018). "Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal". In: *Medical Decision Making* 38.2, pp. 200–211. DOI: 10.1177/0272989x17725740.

Phillippo, D. M., S. Dias, A. Elsada, A. E. Ades, and N. J. Welton (2019). "Population adjustment methods for indirect comparisons: A review of National Institute for Health and Care Excellence technology appraisals". In: *International Journal of Technology Assessment in Health Care*. Available online 13/06/2019. DOI: 10.1017/S0266462319000333.

In each case, the candidate conceived and planned the manuscript, was responsible for any data extraction, analysis, and interpretation of results, and drafted and critically revised the manuscript. Co-authors contributed to the conception and planning, and critically revised each manuscript. A. Elsada drafted part of the third manuscript (Phillippo et al. 2019, discussion paragraphs 2–5). Published material is clearly indicated as such throughout this thesis.

We, the first and final authors of the above publications, confirm that this is an accurate description of the candidate's contributions.

SIGNED: _____ *(first author)*

SIGNED: _____ *(final author)*

DATE: 25th June 2019

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Health technology assessments and appraisals require reliable estimates of the relative effectiveness of all relevant treatments or interventions for a given patient population, to inform decision making. In the United Kingdom, the National Institute for Health and Care Excellence (NICE) operates a technology appraisal (TA) process in which a company submits evidence on the clinical and cost effectiveness of their treatment compared to other relevant treatments. However, it is rare that all relevant treatments have been compared head-to-head in a single randomised controlled study; instead, the evidence base is often comprised of several studies, each comparing a subset of the treatments of interest.

When head-to-head evidence is not available, but two treatments of interest have been studied against a common comparator (e.g. placebo or standard care), a standard indirect comparison may be performed using published aggregate data (AgD) from each study (Bucher et al. 1997; Glenny et al. 2005). With larger numbers of treatments and studies, a network meta-analysis (NMA) is the standard approach, of which indirect comparison is a simple special case (Ades 2003; Dias et al. 2011c; Hasselblad 1998; Higgins and Whitehead 1996; Lu and Ades 2004, 2006). NMA combines direct evidence (i.e. head-to-head) and indirect evidence (i.e. via a common comparator) in a coherent manner, and allows any two treatments in the network to be compared regardless of whether they were involved in the same head-to-head study. We introduce NMA in greater detail in Section 1.2.

We make a distinction between prognostic variables, which affect outcomes equally on all treatments, and effect modifiers (EMs), which alter the effect of treatment relative to control on a chosen scale. Some variables may be both prognostic and effect modifying, or purely one or other. Standard methods for indirect comparison and network meta-analysis (and pairwise meta-analysis)

respect randomisation and are thus unaffected by differences in prognostic variables between populations. However, these methods do assume that the distributions of any effect modifying variables do not differ between study populations and the decision target population, so that relative effects are constant across populations. This assumption does not always hold, as evidenced by the frequent use of random effects meta-analysis. Methods which relax this assumption to form *population-adjusted indirect comparisons* are becoming increasingly common for submissions to reimbursement agencies such as NICE. Effect modifying variables may be population characteristics such as age, sex, or disease severity. Other study-level variables can also be seen to modify treatment effect, such as differences in treatment intensity between studies, or differences in healthcare systems by country or region. In a health technology appraisal context the latter are typically tightly-defined; we therefore focus on effect modification by population characteristics.

Ideally, individual patient data (IPD) would be available from all studies to fully adjust for differences between study populations using IPD network meta-regression (NMR) (Berlin et al. 2002; Dias et al. 2011a; Lambert et al. 2002; Riley et al. 2010; Tudur Smith et al. 2005), as aggregate data network meta-regression has low power to detect or adjust for effect modifying covariates and is susceptible to ecological bias (Berlin et al. 2002; Donegan et al. 2013; Rothman et al. 2008; Saramago et al. 2012). However, it is rarely the case that full IPD are available. In particular, a very common scenario in the TA context is when a company has IPD on its own trial but only published AgD on their competitor's trial, typically consisting of average treatment effects and summary patient characteristics (e.g. mean and standard deviation for continuous characteristics, and proportions for discrete). Population adjustment methods aim to use the available IPD to adjust for between-trial imbalances in the distribution of observed covariates. The development and use of these methods is motivated by one of two reasons: either i) there is evidence for effect modification, and these variables are distributed differently in each study population; or ii) there is no common comparator or the relevant studies are single arm, and so adjustment is required for all prognostic and effect modifying variables. We describe the population adjustment scenario in greater detail in Section 1.1.

Two recently proposed methods, *Matching-Adjusted Indirect Comparison* (MAIC) (Ishak et al. 2015; Signorovitch et al. 2012, 2010) and *Simulated Treatment Comparison* (STC) (Caro and Ishak 2010; Ishak et al. 2015), are becoming increasingly popular in the applied literature and in submissions to NICE. These approaches (described in detail in Section 2.2) are based on

reweighting and regression adjustment ideas that date back several decades in the surrounding literatures on standardisation, generalisation, and calibration (reviewed in Section 2.1). MAIC and STC are primarily designed with a simple two-study scenario in mind, where IPD on one study is used to create a population-adjusted indirect comparison in the population of an AgD study. As such, these methods are not easily generalisable to larger treatment networks. Present attempts to extend MAIC to larger networks (e.g. Belger et al. 2015b) match an IPD trial either to the population of a chosen AgD trial in a larger network, or to the overall average population in a larger network, and then perform an NMA on the full network including the reweighted IPD trial. Both of these approaches make different assumptions regarding effect modification in different parts of the network, and are thus difficult to justify: the MAIC is predicated on the fact that EMs are present and imbalanced between the IPD trial and AgD trial (or overall average AgD population), whereas the ensuing NMA assumes that there are no imbalances in any EMs in the AgD trials. MAIC and STC are also limited to providing a comparison adjusted to the population of the AgD trial, which may not match the target population for the decision. If effect modification is present, then relative effect estimates must be provided for the relevant decision target population in order to be useful for decision making.

Separately from the development of MAIC and STC, efforts have also been made to extend the IPD network meta-regression framework to incorporate both IPD and AgD studies (Donegan et al. 2013; Jansen 2012; Saramago et al. 2012; Sutton et al. 2008; Thom et al. 2015). These methods (described in detail in Section 2.2) typically assume common regression coefficients at both the individual and aggregate level, which leads to aggregation bias (a form of ecological bias) when the model is non-linear (Rothman et al. 2008). Two approaches have been proposed to account for this. The first is to split the interaction effect into between-study (or aggregate-level) and within-study (or individual-level) effects (Donegan et al. 2013; Saramago et al. 2012; Sutton et al. 2008; Thom et al. 2015). However, in smaller networks (such as the two-study scenario addressed by MAIC and STC) there are insufficient data to identify the additional parameters. The second approach is to define the aggregate-level model by integrating the individual-level model over the study population. This avoids aggregation bias by properly relating the two levels, and does not introduce additional parameters. So far, however, this approach has only been derived for the simple case of binary outcomes and binary covariates, where the integration reduces to summation (Jansen 2012).

This thesis sets out to propose a new and general method for population-adjusted indirect comparisons and network meta-regression combining IPD and AgD, which can be applied to any connected network of evidence and produce estimates in any specified target population. We generalise and build upon the ideas of Jansen (2012) to extend to other outcomes, likelihoods, and covariate distributions, aided by a general method for numerical integration. We call this new approach *Multilevel Network Meta-Regression* (ML-NMR). ML-NMR has the desirable property that standard IPD NMR and AgD NMA are special cases.

In this chapter, we provide a background on indirect comparisons and NMA, which underlie the rest of this thesis. We start by describing in greater detail the simple two-study scenario, and the methods of standard indirect comparison, MAIC, and STC (Section 1.1, also published as part of Phillippo et al. 2016; Phillippo et al. 2018a). We then provide an overview of NMA and the key concepts, including heterogeneity and inconsistency, and model fit and comparison (Section 1.2). Finally, we summarise this chapter in Section 1.3 before outlining the structure of the remainder of this thesis (Section 1.4).

1.1 Background to population-adjusted indirect comparisons

We begin by describing a simple two-study scenario, in which a comparison is made between two treatments investigated separately in two studies. We distinguish between population adjustment methods that make *anchored* indirect comparisons, where the evidence is connected by a common comparator, and *unanchored* indirect comparisons, where the evidence is disconnected due to a lack of a common comparator or single-arm studies. We make a clear and necessary distinction between *prognostic variables* and *effect modifiers*: prognostic variables are covariates that affect outcome; effect modifiers (also known as predictive variables, Hingorani et al. 2013) are covariates that alter the effect of treatment as measured on a given scale. Effect modifiers are not necessarily also prognostic variables, and may be specific to each treatment. Effect modifier status on one scale does not necessarily imply effect modifier status on another scale (van Valkenhoef and Ades 2013). We assume internal validity of the studies included in the analysis, so that the studies provide unbiased estimates of treatment effects in their respective sample populations.

Consider the scenario in Figure 1.1 with one AB trial for which the analyst has IPD, and one AC trial for which only published AgD are available. We wish to estimate a comparison of the effects of treatments B and C on an

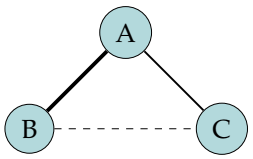


Figure 1.1 An anchored indirect comparison in a simple two-study scenario. IPD are available for the AB study (represented by the thick line); only AgD are available for the AC study (thin line). An anchored indirect comparison (dotted line) compares treatments B and C via the common treatment A arms.

appropriate scale in some target population P, denoted by the parameter $d_{BC(P)}$. Within the AB trial population there are parameters $\theta_{A(AB)}$, $\theta_{B(AB)}$, and $\theta_{C(AB)}$ representing the expected outcome on each treatment (including for treatment C, which was not studied in the AB trial). The AB trial provides estimates $\bar{y}_{A(AB)}$ and $\bar{y}_{B(AB)}$ of $\theta_{A(AB)}$ and $\theta_{B(AB)}$ respectively, which are the summary outcomes (e.g. probability of success, or mean response) on each treatment. The parameter $\theta_{C(AB)}$ is not estimated by the AB trial. There is a parallel system of parameters $\{\theta_{A(AC)}, \theta_{B(AC)}, \theta_{C(AC)}\}$ and estimates $\{\bar{y}_{A(AC)}, \bar{y}_{C(AC)}\}$ in the AC trial.

Having selected a suitable scale, for example a logit, log, risk difference, or mean difference scale, we estimate the population-specific relative treatment effects $d_{AB(AB)}$ and $d_{AC(AC)}$ in each trial using the appropriate link function $g(\cdot)$:

$$\hat{d}_{AB(AB)} = g\left(\bar{y}_{B(AB)}\right) - g\left(\bar{y}_{A(AB)}\right), \tag{1.1a}$$

$$\hat{d}_{AC(AC)} = g\left(\bar{y}_{C(AC)}\right) - g\left(\bar{y}_{A(AC)}\right). \tag{1.1b}$$

Standard methods for indirect comparison make the assumption that there is no difference in the distribution of effect modifying covariates, specific to the chosen scale, between the population in the AB and AC trials or the target population P, so that the population-specific relative treatment effects are equal across populations:

$$d_{AB(AB)} = d_{AB(AC)} = d_{AB(P)}, \tag{1.2a}$$

$$d_{AC(AC)} = d_{AC(AB)} = d_{AC(P)}. \tag{1.2b}$$

Under this assumption, which we call *constancy of relative effects*, a standard indirect comparison estimates the relative effect of C vs. B in population P as

$$\hat{d}_{BC(P)} = \hat{d}_{AC(AC)} - \hat{d}_{AB(AB)}, \tag{1.3}$$

which takes account of the fact that individuals are only randomised *within* trials (Bucher et al. 1997).

For the purposes of technology appraisal, we typically require estimates of absolute effects (for example, as inputs to a cost-effectiveness analysis). The final step is thus to apply these relative effects to a specified target population P in which the summary absolute effect (such as the mean change from baseline, or probability of response) of treatment A is $\bar{y}_{A(P)}$. We can now estimate the summary absolute effects on treatments A, B, and C in the target population, $\theta_{A(P)}$, $\theta_{B(P)}$, $\theta_{C(P)}$, as

$$\bar{y}_{A(P)}, \tag{1.4a}$$

$$\hat{y}_{B(P)} = g^{-1}\Big(g\left(\bar{y}_{A(P)}\right) + \hat{d}_{AB(P)}\Big), \tag{1.4b}$$

$$\hat{y}_{C(P)} = g^{-1}\Big(g\left(\bar{y}_{A(P)}\right) + \hat{d}_{AC(P)}\Big). \tag{1.4c}$$

Suppose that, in each trial, we have information on a common set of covariates $\boldsymbol{x}$. Between-trial differences in the distribution of prognostic variables that are not effect modifiers do not affect inference, because the within-trial randomisation means that they do not impact on the relative treatment effects. Note that effect modifiers $\boldsymbol{x}^{\text{EM}}$, a subset of $\boldsymbol{x}$, are assumed to have an additive effect on the transformed scale. In other words, at any given value of $\boldsymbol{x}^{\text{EM}}$, the *conditional* relative effect is $\gamma_{AB} + \boldsymbol{\beta}^{\intercal}\boldsymbol{x}^{\text{EM}}$, conceptualised as an "intercept" term (the relative effect γ_{AB} at $\boldsymbol{x}^{\text{EM}} = \mathbf{0}$) plus an interaction effect $\boldsymbol{\beta}^{\intercal}\boldsymbol{x}^{\text{EM}}$.

If there are effect modifiers and if these are distributed differently between the populations, then the relative treatment effects $d_{AB(AB)}$ and $d_{AC(AC)}$ that can be estimated directly from each trial are only valid for a population with the distribution of effect modifiers observed in that trial. For example, we would have estimates $\hat{d}_{AB(AB)}$ in the AB population and $\hat{d}_{AC(AC)}$ in the AC population, but it would not be possible to identify a coherent set of estimates, either for the population represented by the AB trial

$$\hat{d}_{AB(AB)}, \quad \hat{d}_{AC(AB)}, \quad \hat{d}_{BC(AB)} = \hat{d}_{AC(AB)} - \hat{d}_{AB(AB)}$$

(since $\hat{d}_{AC(AB)}$ is not available), or for the population represented by the AC trial

$$\hat{d}_{AB(AC)}, \quad \hat{d}_{AC(AC)}, \quad \hat{d}_{BC(AC)} = \hat{d}_{AC(AC)} - \hat{d}_{AB(AC)}$$

(since $\hat{d}_{AB(AC)}$ is not available), or indeed for any other target population.

The premise of population adjustment methods such as MAIC and STC in a connected network is to relax the constancy of relative effects assumption by adjusting for between-trial differences in effect modifying covariates, in order to identify a coherent set of estimates where standard methods of indirect comparison cannot. Both methods use IPD on the AB trial to form predictions

$\hat{y}_{A(AC)}$, $\hat{y}_{B(AC)}$ of the summary outcomes that would be observed on treatments A and B in the AC trial. MAIC and STC, and the process by which these predictions are obtained, are described in greater detail in Section 2.2. The predicted outcomes $\hat{y}_{A(AC)}$ and $\hat{y}_{B(AC)}$ are then used to estimate an anchored indirect comparison:

$$\hat{d}_{BC(AC)} = g\left(\bar{y}_{C(AC)}\right) - g\left(\bar{y}_{A(AC)}\right) - \left(g\left(\hat{y}_{B(AC)}\right) - g\left(\hat{y}_{A(AC)}\right)\right). \tag{1.5}$$

The validity of this anchored comparison assumes that the $d_{AB(AC)}$ relative effect can be reliably predicted using the IPD in the AB study given that all effect modifying covariates are known and adjusted for, which we call *conditional constancy of relative effects*.

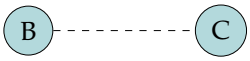


Figure 1.2 An unanchored indirect comparison in a simple two-study scenario. IPD are available for the B study; only AgD are available for the C study. An unanchored indirect comparison (dotted line) compares treatments B and C directly, without a common comparator. The studies involved may also have other treatment arms, or be single-arm as shown here; the key difference between Figure 1.1 is the lack of a common comparator.

In a disconnected network or where single-arm trials are involved (as in Figure 1.2), MAIC and STC instead attempt to improve on a naïve (or "unadjusted") indirect comparison of arms by adjusting for any covariates that influence outcome. In this case, an unanchored indirect comparison is made:

$$\hat{d}_{BC(C)} = g\left(\bar{y}_{C(C)}\right) - g\left(\hat{y}_{B(C)}\right). \tag{1.6}$$

An unanchored comparison assumes that absolute outcomes $\theta_{B(C)}$ can be reliably predicted; this is a very strong assumption, which we call *conditional constancy of absolute effects*. To be valid, all effect modifiers and prognostic variables must be known and adjusted for. An anchored indirect comparison should therefore always be preferred in a connected network, as it respects the randomisation within studies and does not rely on such strong assumptions. However, if the treatment network is disconnected or contains single-arm studies, then there is no common comparator arm through which to make an anchored indirect comparison, and we may be obliged to rely on an unanchored indirect comparison.

Table 1.1 summarises the constancy assumptions made by the different forms of indirect comparison. These assumptions are described in greater detail in Section 2.3.

Table 1.1 All indirect comparisons and network meta-analyses require some form of constancy assumption. Unanchored comparisons require a much stronger assumption, which is widely considered impossible to meet.

	Anchored comparisons		Unanchored comparisons
Method	Standard indirect comparison or network meta-analysis	Anchored population-adjusted indirect comparison	Unanchored population-adjusted indirect comparison
Constancy assumption	Constancy of relative effects	Conditional constancy of relative effects	Conditional constancy of absolute effects
	$\Rightarrow$ Relative effects are the same across populations	$\Rightarrow$ Reliable predictions of relative effects	$\Rightarrow$ Reliable predictions of absolute effects
Valid only if	No effect modifiers in imbalance	All effect modifiers known and adjusted for	All effect modifiers and prognostic variables known and adjusted for
Data requirements	Aggregate data	IPD on at least one trial	IPD on at least one trial

1.2 Background to network meta-analysis

When a larger network of treatments and studies is available, the standard approach is a network meta-analysis (Ades 2003; Dias et al. 2011c; Hasselblad 1998; Higgins and Whitehead 1996; Lu and Ades 2004, 2006). NMA provides a coherent set of relative treatment effects which is essential for decision making (Caldwell et al. 2005), and standard indirect comparison (described above in Section 1.1) is a simple special case. For greatest generality, we label treatments numerically in a NMA (i.e. treatments 1, 2, 3, etc.) instead of alphabetically as in an indirect comparison. An example treatment network is shown in Figure 1.3.

We begin this section by describing the standard parameterisation for writing NMA models, known as the *baseline shift parameterisation*, in which a reference treatment arm is defined for each trial. We then describe an alternative parameterisation, which we term the *reference treatment parameterisation*, that uses a single reference treatment across the entire network. We will show that these two parameterisations are equivalent, and henceforth use the reference treatment parameterisation for the remainder of this thesis. We consider AgD NMA to begin with, but the same arguments apply to IPD NMA (and, by extension, to ML-NMR as developed in Chapter 4).

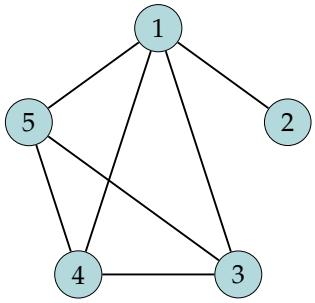


Figure 1.3 An example network of five treatments, connected by study evidence (solid lines). Not all treatments are directly compared in a study, but NMA combines all of the direct and indirect evidence in a coherent manner allowing any two treatments to be compared.

Baseline shift parameterisation 1.2.1

Consider that we have summary outcome data $y_{\bullet jt_k}$ available on treatment t_k in arm k of study j. The subscript $\bullet$ denotes that the data are summaries over individuals; this will help to distinguish from individual-level data y_{ijt_k}, that we introduce later. Using the standard baseline shift parameterisation, we write out the NMA model as

$$y_{\bullet jt_k} \sim \pi(\theta_{\bullet jt_k}) \tag{1.7a}$$

$$\theta_{\bullet jt_k} = g^{-1}(\eta_{\bullet jk}) \tag{1.7b}$$

$$\eta_{\bullet jt_k} = \begin{cases} \mu_j^{(t_1)} & \text{for } k = 1 \\ \mu_j^{(t_1)} + \delta_{jt_1t_k} & \text{for } k > 1 \end{cases} \tag{1.7c}$$

where $\pi(\cdot)$ is a suitable likelihood (e.g. Binomial, Normal, Poisson, etc.), and $g(\cdot)$ is a link function transforming the expected summary outcomes $\theta_{\bullet jt_k}$ onto the linear predictor $\eta_{\bullet jt_k}$. Again, we use a subscript $\bullet$ to help distinguish these parameters, which relate to aggregate-level summaries, from their counterparts introduced later, which relate to individuals. The baseline shift parameterisation dictates the form of the linear predictor $\eta_{\bullet jt_k}$, so that $\mu_j^{(t_1)}$ are study-specific baseline parameters (here we use a superscript referring to the interpretation of the study-specific baselines with respect to the treatment t_1 in arm 1), and $\delta_{jt_1t_k}$ is the study-specific relative effect of treatment t_k in arm k vs. treatment t_1 in arm 1.

In a fixed effect (FE) model, the study-specific relative effects are the same across trials:

$$\delta_{jt_1t_k} = d_{t_1t_k} = d_{1t_k} - d_{1t_1}, \tag{1.8}$$

where d_{ab} is the relative effect of treatment b vs. a, and the consistency

equations specify that

$$d_{ab} = d_{1b} - d_{1a}. \tag{1.9}$$

We often write d_t in place of d_{1t}, and we set $d_1 = 0$. In a Bayesian analysis, prior distributions are placed on the parameters $\mu_j^{(t_1)}$ and d_t.

In a random effects (RE) model, the study-specific relative effects instead have a distribution. For trials with three or more arms the relative effects are correlated, given a multivariate Normal with marginal distributions and correlations given by

$$\delta_{jt_1t_k} \sim \mathrm{N}(d_{t_kt_1}, \tau^2_{t_1t_k}) \quad \forall k > 1 \tag{1.10a}$$

$$\mathrm{cor}(\delta_{jt_1t_{k_1}}, \delta_{jt_1t_{k_2}}) = \psi^{(t_1)}_{t_{k_1}t_{k_2}} \quad \forall k_1, k_2 > 1. \tag{1.10b}$$

The superscript (t_1) on the pairwise correlation parameters by convention denotes the treatment that the correlation between random effects is with respect to: i.e. here the correlation is for random effects of treatments t_{k_1} and t_{k_2} against t_1. For trials with only two arms, the single relative effect on the non-reference arm is univariate Normal with marginal distribution given by (1.10a). As with the fixed effect model, we have the consistency equations (1.9) on the treatment effects, but now we also have a set of second-order consistency equations on the between-study variances (Lu and Ades 2009):

$$\tau^2_{t_{k_1}t_{k_2}} = \tau^2_{t_1t_{k_1}} + \tau^2_{t_1t_{k_2}} - 2\psi^{(t_1)}_{t_{k_1}t_{k_2}}\tau_{t_1t_{k_1}}\tau_{t_1t_{k_2}}. \tag{1.11}$$

Prior distributions on the random effect variances and correlations are non-trivial to specify, as they must jointly satisfy these consistency constraints (Lu and Ades 2009). However, if we assume that the between-study variances are homogeneous and equal to τ^2, this implies that all the correlations are equal to 0.5 (following (1.11), see Higgins and Whitehead 1996).

1.2.2 Reference treatment parameterisation

In the reference treatment parameterisation, the likelihood and link function remain the same as the baseline shift parameterisation (equations (1.7a) and (1.7b)). However, the linear predictor $\eta_{\bullet jt}$ for study j treatment t is now written in terms of the reference treatment, which we set to be treatment 1 without loss of generality, leading to the following NMA model:

$$y_{\bullet jt} \sim \pi(\theta_{\bullet jt}) \tag{1.12a}$$

$$\theta_{\bullet jt} = g^{-1}(\eta_{\bullet jt}) \tag{1.12b}$$

$$\eta_{\bullet jt} = \mu_j^{(1)} + \delta_{jt} \quad \forall t \geq 1, \tag{1.12c}$$

where $\mu_j^{(1)}$ are study-specific baseline parameters for reference treatment 1 (the superscript referring to the interpretation with respect to the reference treatment 1), and $\delta_{jt} = \delta_{j1t}$ is the study-specific relative effect of treatment t vs. treatment 1. The subscript k is dropped from t_k in the reference treatment parameterisation, since it is no longer necessary.

The fixed effect model is written as

$$\delta_{jt} = d_{1t}, \tag{1.13}$$

again often writing d_t for d_{1t} and setting $d_1 = 0$. We still have the set of consistency equations (1.9), but these are not explicitly used in writing down the model. In a Bayesian analysis, prior distributions are placed on the parameters $\mu_j^{(1)}$ and d_t.

For the random effects model, this time every non-treatment 1 arm has a random effect. Therefore we must handle the correlations between the random effects for all trials with two or more non-treatment 1 arms, not just those with three or more arms in total. We specify the random effects to be multivariate Normal, with marginal distributions and correlations given by

$$\delta_{jt} \sim \mathrm{N}(d_t, \tau_t^2) \quad \forall t > 1 \tag{1.14a}$$

$$\mathrm{cor}(\delta_{ja}, \delta_{jb}) = \psi_{ab}^{(1)} \qquad \forall a, b > 1, \tag{1.14b}$$

and $\delta_{j1} = 0$. For two-arm studies that compare a treatment $t > 1$ against treatment 1, there is a single univariate Normal random effect on the non-treatment 1 arm, with distribution given by (1.14a). We still have the second-order consistency equations (1.11), now with respect to the reference treatment (rather than the treatment in the reference arm):

$$\tau_{ab}^2 = \tau_a^2 + \tau_b^2 - 2\psi_{ab}^{(1)}\tau_a\tau_b. \tag{1.15}$$

but these are not explicitly used in writing down the model; prior distributions are therefore straightforward to specify on the τ_t^2 parameters, since the second-order consistency equations are then implicitly satisfied. Again, if the homogeneous variance assumption is used this implies that all the correlations are equal to 0.5.

Equivalence 1.2.3

The standard baseline shift parameterisation (Section 1.2.1) defines a reference arm 1 in each study, in which the treatment is t_1. Study-specific baseline parameters refer to this treatment arm, and the other treatment arms in the trial

are compared to the treatment t_1 as relative effects. The reference treatment parameterisation (Section 1.2.2) instead defines a single reference treatment across the entire network. Without loss of generality, we set treatment 1 to be the reference treatment. Study-specific baseline parameters then always refer to the reference treatment, and relative treatment effects in each trial are always against the reference treatment—even if the trial does not include this treatment arm. These two parameterisations are trivially equivalent when every study has a treatment 1 arm. However, when there are trials without a treatment 1 arm, the baseline shift parameterisation specifies a model on a different set of relative effects to the reference treatment parameterisation. Here, we show that the two parameterisations are indeed equivalent regardless of the network structure (with some caveats, discussed later in Section 1.2.3.1).

Equivalence of the NMA models under the two parameterisations follows by equating

$$\mu_j^{(t_1)} = \mu_j^{(1)} + \delta_{jt_1}. \tag{1.16}$$

Intuitively, we transform the reference treatment parameterisation (where the baselines are with respect to treatment 1) into the baseline shift parameterisation (where the baselines are with respect to treatment t_1 in arm 1) simply by adding the study-specific relative effect of treatment t_1 vs. 1 to the baseline effect of treatment 1, and vice versa.

For both the fixed and random effects models this result follows from the property that relative effects are consistent within a study, requiring the study-specific relative effects to "add up" within a study j (Lu and Ades 2009):

$$\begin{aligned} \delta_{jt_{k_1}t_{k_2}} &= \delta_{jt_1t_{k_2}} - \delta_{jt_1t_{k_1}} \\ &= \delta_{j1t_{k_2}} - \delta_{j1t_{k_1}}, \end{aligned} \tag{1.17}$$

where this relationship can be expressed with respect to the reference arm treatment t_1 or the reference treatment 1 (or indeed any other treatment). (The first and second-order consistency relations are motivated by taking expectations and variances respectively of both sides of (1.17).) We see that (1.16) is the necessary relation to equate the two parameterisations:

$$\begin{aligned} \mu_j^{(t_1)} + \delta_{jt_1t_k} &= \mu_j^{(t_1)} + \delta_{jt_k} - \delta_{jt_1} \\ &= \underbrace{\left(\mu_j^{(t_1)} - \delta_{jt_1}\right)}_{(1.16)} + \delta_{jt_k} \\ &= \mu_j^{(1)} + \delta_{jt_k}. \end{aligned} \tag{1.18}$$

To see that the joint random effects distribution remains the same, consider the joint distribution of the $T-1$ *basic* random effects (i.e. against treatment 1),

where T is the number of treatments:

$$\begin{bmatrix} \delta_{j2} \\ \vdots \\ \delta_{jT} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} d_2 \\ \vdots \\ d_T \end{bmatrix}, \begin{bmatrix} \tau_2^2 & & & \\ \psi^{(1)}_{23}\tau_2\tau_3 & \ddots & & \\ \vdots & \ddots & \ddots & \\ \psi^{(1)}_{2T}\tau_2\tau_T & \dots & \psi^{(1)}_{(T-1)T}\tau_{T-1}\tau_T & \tau_T^2 \end{bmatrix} \right) \tag{1.19}$$

In any given trial, the reference treatment parameterisation selects the necessary subset of these basic random effects. By comparison, the baseline shift parameterisation puts *functional* random effects on non-reference arms, which are linear combinations of the basic random effects (a description used previously by Lu and Ades (2006), with basic/functional terminology from Eddy et al. (1990)). In general, the functional random effects of the baseline shift parameterisation can be written in terms of the basic random effects as

$$\delta_{jt_1t_k} = \delta_{j1t_k} - \delta_{j1t_1} \sim \text{N}(d_{t_k} - d_{t_1}, \tau^2_{t_1} + \tau^2_{t_k} - 2\psi^{(1)}_{t_1t_k}\tau_{t_1}\tau_{t_k}), \tag{1.20}$$

with covariances

$$\begin{aligned} \text{cov}(\delta_{jt_1t_{k_1}}, \delta_{jt_1t_{k_2}}) &= \text{cov}(\delta_{j1t_{k_1}} - \delta_{j1t_1}, \delta_{j1t_{k_2}} - \delta_{j1t_1}) \\ &= \text{cov}(\delta_{j1t_{k_1}}, \delta_{j1t_{k_2}}) - \text{cov}(\delta_{j1t_{k_1}}, \delta_{j1t_1}) \\ &\quad - \text{cov}(\delta_{j1t_1}, \delta_{j1t_{k_2}}) + \text{cov}(\delta_{j1t_1}, \delta_{j1t_1}) \\ &= \psi^{(1)}_{t_{k_1}t_{k_2}}\tau_{t_{k_1}}\tau_{t_{k_2}} - \psi^{(1)}_{t_1t_{k_1}}\tau_{t_1}\tau_{t_{k_1}} - \psi^{(1)}_{t_1t_{k_2}}\tau_{t_1}\tau_{t_{k_2}} + \tau^2_{t_1}. \end{aligned} \tag{1.21}$$

To see that the mean and variance of the functional random effects (1.20) are the same as those specified in the baseline shift parameterisation (1.10) is simply a matter of applying the first (1.9) and second-order (1.15) consistency relations, respectively. To transform correlations with respect to treatment 1 (as used in the reference treatment parameterisation) into correlations with respect to the treatment in the reference arm (as used in the baseline shift parameterisation), we note that the left hand side of (1.21) is equal to $\psi^{(t_1)}_{t_{k_1}t_{k_2}}\tau_{t_{k_1}}\tau_{t_{k_2}}$, and rearrange to find

$$\psi^{(t_1)}_{t_{k_1}t_{k_2}} = \psi^{(1)}_{t_{k_1}t_{k_2}} + \frac{\tau^2_{t_1} - \psi^{(1)}_{t_1t_{k_1}}\tau_{t_1}\tau_{t_{k_1}} - \psi^{(1)}_{t_1t_{k_2}}\tau_{t_1}\tau_{t_{k_2}}}{\tau_{t_{k_1}}\tau_{t_{k_2}}}. \tag{1.22}$$

Thus we have equivalence of random effects NMA under the two parameterisations: the reference treatment parameterisation writes out the model using the basic random effects, whereas the baseline shift parameterisation writes out the model using the functional random effects.

The distribution of these functional random effects can also be computed in matrix form. Writing (1.19) in a compact manner as $\boldsymbol{\delta}_j \sim \text{MVN}(\boldsymbol{d}, \boldsymbol{\Sigma}_\tau)$, we

use the fact that

$$\boldsymbol{D}\boldsymbol{\delta}_j \sim \text{MVN}(\boldsymbol{D}\boldsymbol{d}, \boldsymbol{D}\boldsymbol{\Sigma}_\tau \boldsymbol{D}^\mathsf{T}) \tag{1.23}$$

for any given "design" matrix $\boldsymbol{D}$ selecting the relevant contrasts, to the same result.

Notice that the reference treatment parameterisation aligns closely with the "auxiliary variables" approach used by Lu and Ades (2009) to find prior configurations that satisfy the second-order consistency equations. Lu and Ades (2009) consider expressing the random effects $\delta_{jt_1t_k}$ as the sum of two correlated Normal "auxiliary variables" (and thus the heterogeneity variances $\tau^2_{t_1t_k}$ in a form analogous to (1.15)), leading to simplified prior specification satisfying the second-order consistency equations. In this case, the auxiliary variables are precisely the basic random effects δ_{j1t_1} and δ_{j1t_k}, and we have one-to-one correspondence between the formulations as described above (which is not true for general auxiliary variables).

1.2.3.1 Caveats to equivalence

The equivalence of the two parameterisations is true up to the specification of prior distributions on the study-specific baselines $\mu_j^{(t_1)}$ or $\mu_j^{(1)}$. When $t_1 \neq 1$, the baselines in the two parameterisations refer to absolute responses on different treatments, and thus the same prior distribution has different interpretations under the different parameterisations. It is typical to place non-informative or weakly-informative prior distributions on the study-specific baselines, and in this case there is unlikely to be any noticeable difference to the posterior distribution; however, if stronger prior distributions are to be placed then the different interpretations of the baselines should be noted, and if necessary prior distributions may be transformed from one parameterisation to the other via (1.16).

1.2.4 IPD network meta-regression

When individual patient data are available from every study, an IPD network meta-regression may be performed (Berlin et al. 2002; Dias et al. 2011a; Lambert et al. 2002; Riley et al. 2010; Tudur Smith et al. 2005). Consider that outcomes y_{ijt} and covariates $\boldsymbol{x}_{ijt}$ are available for each individual i in study j on treatment t. Using the reference treatment parameterisation, an IPD NMR model is of

the form

$$y_{ijt} \sim \pi(\theta_{ijt}) \tag{1.24a}$$

$$\theta_{ijt} = g^{-1}(\eta_{jk}(\boldsymbol{x}_{ijt})) \tag{1.24b}$$

$$\eta_{jk}(\boldsymbol{x}_{ijt}) = \mu_j^{(1)} + \boldsymbol{x}_{ijt}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,t}) + \delta_{jt}, \tag{1.24c}$$

where $\pi(\cdot)$ is a suitable likelihood, and $g(\cdot)$ is a link function transforming the expected outcomes θ_{ijt} onto the linear predictor $\eta_{jk}(\boldsymbol{x}_{ijt})$. As with AgD NMA using the reference treatment parameterisation, $\mu_j^{(1)}$ are study-specific baseline parameters for reference treatment 1. Prognostic covariate effects are given by $\boldsymbol{\beta}_1$, and effect modifier interactions with treatment are $\boldsymbol{\beta}_{2,t}$. We set $\boldsymbol{\beta}_{2,1} = \mathbf{0}$. Again, δ_{jt} can either be fixed effect or random effects, except now in terms of individual-level (conditional) relative effects γ_{1t}. We often write γ_t for γ_{1t}.

The fixed effect model is written as

$$\delta_{jt} = \gamma_t, \tag{1.25}$$

with $\gamma_1 = 0$. In a Bayesian analysis, prior distributions are placed on the parameters $\mu_j^{(1)}$ and γ_t.

For the random effects model, we specify the random effects to be multivariate Normal, with marginal distributions and correlations given by

$$\delta_{jt} \sim \mathrm{N}(\gamma_t, \tau_t^2) \quad \forall t > 1 \tag{1.26a}$$

$$\mathrm{cor}(\delta_{ja}, \delta_{jb}) = \psi_{ab}^{(1)} \qquad \forall a, b > 1, \tag{1.26b}$$

and $\delta_{j1} = 0$. For two-arm studies against treatment 1, there is a single univariate Normal random effect on the non-treatment 1 arm, with distribution given by (1.26a). Prior distributions are required for $\mu_j^{(1)}$, γ_t, and τ_t^2. Again, if the homogeneous variance assumption is used this implies that all the correlations are equal to 0.5.

As for the AgD NMA model, we have consistency equations (1.9) on the (now individual-level) relative effects γ_t, and for the random effects model we also have second-order consistency (1.15) on the heterogeneity variances τ_t^2. Furthermore, we now have consistency equations on the effect modifier interaction coefficients $\boldsymbol{\beta}_{2,t}$:

$$\boldsymbol{\beta}_{2,ab} = \boldsymbol{\beta}_{2,1b} - \boldsymbol{\beta}_{2,1a}, \tag{1.27}$$

and we typically write $\boldsymbol{\beta}_{2,t}$ for $\boldsymbol{\beta}_{2,1t}$.

1.2.4.1 Equivalence

Using baseline shift parameterisation, the linear predictor in the IPD NMR model is instead written as

$$\begin{aligned}\eta_{jk}(\boldsymbol{x}_{ijt_k}) &= \mu_j^{(t_1)} + \boldsymbol{x}_{ijt_k}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,t_1t_k}) + \delta_{jt_1t_k} \\ &= \mu_j^{(t_1)} + \boldsymbol{x}_{ijt_k}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,t_k} - \boldsymbol{\beta}_{2,t_1}) + \delta_{jt_k} - \delta_{jt_1}.\end{aligned} \tag{1.28}$$

Equivalence of these two parameterisations relies on a similar relation to before:

$$\mu_j^{(t_1)} = \mu_j^{(1)} + \boldsymbol{x}_{ijt_k}^{\mathsf{T}}\boldsymbol{\beta}_{2,t_1} + \delta_{jt_1}. \tag{1.29}$$

The interpretation of the study-specific baseline parameters under the different parameterisations is now also with respect to the reference level of the covariates (typically centred for continuous covariates). We interpret the baselines under the baseline shift parameterisation (absolute outcomes on treatment t_1) as the baselines under the reference treatment parameterisation (absolute outcomes on treatment 1) plus the study-specific relative effect of treatment t_1 vs. treatment 1, all at the reference level of the covariates. As described in Section 1.2.3.1, the equivalence of the two parameterisations is true up to the specification of prior distributions on the study-specific baselines $\mu_j^{(t_1)}$ or $\mu_j^{(1)}$.

1.2.5 Advantages of different parameterisations

Both the baseline shift and reference treatment parameterisations have their relative advantages and disadvantages.

Firstly, the baseline shift parameterisation is the standard parameterisation for NMA in the literature, and is thus the most familiar form for those applying and developing NMA methods. For RE models, the baseline shift parameterisation specifies only a single random effect for studies with only two arms, thus avoiding the need to account for correlations in networks where only two-arm studies are present. By comparison, the reference treatment parameterisation specifies random effects on every non-reference treatment arm, and so correlations between random effects are present for every network other than a simple "star network" where every study is a single comparison against treatment 1. For both parameterisations, correlations between random effects—where present—are essential, and cannot be ignored. The baseline shift parameterisation is also more obviously *not* a pooling over treatment arms, since the relative effects are explicitly written out using the consistency equations—a fact that may be less obvious when using the reference treatment parameterisation. "Arm-based" NMA models that pool absolute effects over

treatment arms instead of pooling relative effects have been proposed (Hong et al. 2015), but have been criticised for not respecting randomisation (Dias and Ades 2015). Furthermore, since the consistency equations are explicitly written out as part of the baseline shift parameterisation, it is simpler to assess inconsistency through unrelated mean effects or node-splitting models (see Section 1.2.7).

On the other hand, the reference treatment parameterisation allows for simpler implementation in code, as there is no longer a need to keep track of the reference treatment for each study. Moreover, it is much more straightforward to specify prior distributions for non-equal heterogeneity variances τ_t^2 under the reference treatment parameterisation, since the second-order consistency equations hold simply by construction. Lastly, the reference treatment parameterisation allows for the inclusion of single-arm studies through a model on the study-specific baselines $\mu_j^{(1)}$, which is possible in this parameterisation since these all refer to the reference treatment 1, rather than study-specific reference treatments t_1 (Thom et al. 2015). Although this is not a scenario we consider in this thesis, we expect future development of the new methods proposed here to proceed in this direction (see discussion in Section 9.2.7).

For the remainder of this thesis, we adopt the reference treatment parameterisation. We also switch to indexing treatments by $k = 1, \ldots, K$ so that, for example, individual-level data are referenced as individual i in study j on treatment k. This also frees up t to refer to event or survival times when we consider models for time-to-event data later in Chapter 7.

Model fit and comparison 1.2.6

In a Bayesian NMA framework, model fit is often assessed using the residual deviance, and different models are often compared using the Deviance Information Criterion (DIC) (Dias et al. 2011c; Spiegelhalter et al. 2002).

The residual deviance for a data point is defined as the deviance (-2 times the log likelihood) under the current model, minus the deviance under a saturated model where every data point is perfectly predicted (McCullagh and Nelder 1989). The form of the residual deviance is dictated by the form of the likelihood; formulae for some common likelihoods are given by Welton et al. (2012, p. 123), and later in Table 5.1. For example, in an AgD NMA with Binomial likelihood, $y_{\bullet jk} \sim \text{Bin}(N_{jk}, p_{jk})$, the residual deviance contribution $D_{\text{res};\bullet jk}$ for the treatment k arm of study j is

$$D_{\text{res};\bullet jk} = 2\left(y_{\bullet jk}\log\left(\frac{y_{\bullet jk}}{\hat{y}_{\bullet jk}}\right) + (N_{jk} - y_{\bullet jk})\log\left(\frac{N_{jk} - y_{\bullet jk}}{N_{jk} - \hat{y}_{\bullet jk}}\right)\right), \tag{1.30}$$

where $\hat{y}_{\bullet jk} = N_{jk}p_{jk}$ is the predicted number of events under the model. The total residual deviance D_{res} is the sum of the residual deviance contributions from all data points: for example, for an AgD NMA

$$D_{\text{res}} = \sum_j \sum_k D_{\text{res};\bullet jk}. \tag{1.31}$$

For an IPD NMA, the residual deviance contribution for an individual i in study j on treatment k is denoted by $D_{\text{res};ijk}$, and the total residual deviance is

$$D_{\text{res}} = \sum_j \sum_k \sum_i D_{\text{res};ijk}. \tag{1.32}$$

For the purposes of checking absolute model fit, we are interested in the posterior distribution of D_{res}, and of the contribution from each data point (either $D_{\text{res};ijk}$ or $D_{\text{res};\bullet jk}$ in an AgD or IPD NMA, respectively). The posterior mean of the residual deviance, $\mathbb{E}(D_{\text{res}})$, can be compared to the number of independent data points to check if model fit is adequate: the two will be approximately equal under a well-fitting model assuming approximate Normality (which may not hold in practice, e.g. for binary data in small samples or with event probabilities far from 0.5) (Spiegelhalter et al. 2002). Furthermore, the posterior distributions of the residual deviance contributions from each data point can be plotted (e.g. posterior mean and 95% Credible Interval), to identify any poorly-fitting observations (those with contributions much greater than 1).

To compare different models, we need to trade off goodness-of-fit against model complexity. The DIC achieves this by penalising the residual deviance D_{res} by a measure of the effective number of parameters, p_D. Following Welton et al. (2012, p. 126), we calculate p_D as the difference between the posterior mean of D_{res} and the value of D_{res} calculated at the posterior mean of the fitted values (for example, in an AgD NMA we replace $\hat{y}_{\bullet jk}$ by $\mathbb{E}(\hat{y}_{\bullet jk})$):

$$p_D = \mathbb{E}\left(D_{\text{res}}\right) - D_{\text{res}}\big|_{\mathbb{E}(\hat{y})}. \tag{1.33}$$

(Spiegelhalter et al. (2002) originally suggested calculating p_D at the posterior mean of the parameters; the modification of Welton et al. (2012) is more stable in hierarchical models.) The DIC is then calculated as

$$\text{DIC} = \mathbb{E}\left(D_{\text{res}}\right) + p_D. \tag{1.34}$$

When comparing a set of candidate models, lower values of DIC are preferred; typically differences of less than 3 are considered small, and differences of more than 5 are considered substantial (Lunn et al. 2010, pp. 165–167; Dias et al. 2018, p. 69). If differences in DIC are small we would typically prefer the model with the smallest effective number of parameters.

Assessing inconsistency 1.2.7

A key assumption of network meta-analysis is that of *consistency*: that is, treatment effects "add up" across the network through the consistency equations (1.9), $d_{bc} = d_{ac} - d_{ab}$ for all treatments a, b, c (Higgins and Whitehead 1996; Lu and Ades 2006). Consistency is guaranteed to hold within multi-arm trials by design. However, when direct and indirect evidence from different studies are available on the same comparison (e.g. direct evidence on d_{23} and indirect evidence via d_{12} and d_{13}) there is potential for disagreement, or *inconsistency*. Checking consistency is therefore a key part of a network meta-analysis.

Inconsistency and heterogeneity are both caused by differences between studies that alter the treatment effect, such as differences in effect modifying covariates or outcome or treatment definitions. We consider, heterogeneity as affecting studies on the same treatment comparison, and inconsistency as affecting studies on different treatment comparisons. However, in networks containing multi-arm trials this distinction is not uniquely defined, and depends on the choice of baseline treatment. Inconsistency and heterogeneity are therefore often examined together.

Several methods for assessing inconsistency have been proposed, including inconsistency factors (Lu and Ades 2006), design-by-treatment interactions (Higgins et al. 2012), unrelated mean effects (Dias et al. 2011d), and node-splitting (Dias et al. 2010). The inconsistency factors model can be seen as a re-parameterisation of the unrelated mean effects model, with an additional exchangeable structure on the inconsistency parameters; however, the exchangeable inconsistency structure is hard to estimate well and will often be very uncertain (Dias et al. 2011d), and the assumption that inconsistency is exchangeable throughout the network is questionable (Higgins et al. 2012). The inconsistency factors model can also be seen as a restricted form of the design-by-treatment interactions model; however, the design-by-treatment interactions model is over-parameterised unless constraints are placed on the inconsistency parameters (Higgins et al. 2012). We thus focus on the unrelated mean effects and node-splitting models here: the former allows for a simple global assessment of inconsistency, and the latter provides a more powerful and detailed assessment for each potentially inconsistent contrast in turn. Inconsistency is assessed in IPD NMA using the very same techniques, except that the parameters of interest are the individual-level treatment effects γ_k rather than d_k.

1.2.7.1 Unrelated mean effects

Under the assumption of consistency, NMA models define a set of "basic" parameters $d_2, \ldots, d_K$, which relate to relative effects against the reference treatment 1. Prior distributions are placed on these basic parameters, and all other contrasts are derived "functional" parameters using the consistency equations $d_{ab} = d_b - d_a$ for $a < b \in \{2, \ldots, K\}$. The unrelated mean effects (UME) model (Dias et al. 2011d) instead treats all of these parameters, basic and functional, as unrelated parameters with independent prior distributions. In other words, for every observed contrast d_{ab}, we assign a prior distribution such as $d_{ab} \sim \mathrm{N}(0, \sigma_d^2)$ with some appropriate prior variance σ_d^2.

Evidence for inconsistency is then based on comparing the model fit of the unrelated mean effects model to the standard NMA model assuming consistency, for example by comparing residual deviance (Section 1.2.6). Comparing the residual deviance contributions from each data point under each model can indicate where in the network inconsistencies may lie, for example by plotting these against each other and looking for points away from the line of equality, as in Dias et al. (2011d).

1.2.7.2 Node-splitting

Node-splitting, rather than testing for inconsistency globally, examines each potentially inconsistent comparison in turn (Dias et al. 2010). Briefly, the node-splitting approach splits the estimation of a chosen comparison $d_{a'b'}$ into an estimate based only on direct evidence $d^{\mathrm{D}}_{a'b'}$ and an estimate $d^{\mathrm{I}}_{a'b'}$ based on the remainder of the network (i.e. the indirect evidence). Inconsistency is often examined at the same time as heterogeneity, and this is particularly true of node-splitting models where different model parameterisations can shift variation from heterogeneity to inconsistency, or vice versa. The mathematical structure of the RE NMA model is unchanged, except that the random effect for any study with a' and b' arms is now $\delta_{ja'b'} \sim \mathrm{N}(d^{\mathrm{D}}_{a'b'}, \tau^2)$. The estimate of $d^{\mathrm{I}}_{a'b'}$ from the indirect evidence is computed by applying the consistency equations to the remainder of the network in the usual manner. Notice that here we have used the same heterogeneity variance parameter τ^2 for the direct evidence and for the rest of the network. This is not a strict requirement, but in practice there is often not enough data to estimate separate heterogeneity variances for the direct evidence and the rest of the network; moreover, using the same heterogeneity variance parameter allows direct comparison with the heterogeneity variance from the RE NMA.

Again, inconsistency can be assessed by comparing model fit between the standard NMA and the node-split model. Visually comparing the posterior distributions of $d^{\mathrm{D}}_{a'b'}$ and $d^{\mathrm{I}}_{a'b'}$ allows the amount of inconsistency to be judged directly. We can also define the inconsistency parameter $\omega_{a'b'} = d^{\mathrm{D}}_{a'b'} - d^{\mathrm{I}}_{a'b'}$ and examine its posterior distribution to assess whether $\omega_{a'b'} = 0$, for example by computing a Bayesian p-value.

Node-splitting models are more powerful for detecting local inconsistency in a given comparison than the unrelated mean effects model (which assesses inconsistency across the network all at once), however there is a multiple testing issue if several comparisons are to be checked for inconsistency in turn. Node-splitting models also require considerably more effort in determining potentially inconsistent comparisons and then in re-fitting a node-split model for each, although this may be mitigated by automated methods for selecting comparisons to split and then building and fitting the associated models (Valkenhoef et al. 2015).

Inconsistency in meta-regression 1.2.7.3

When meta-regression models are fitted (either to IPD or AgD), the consistency assumption applies to both the treatment effects $d_{ab} = d_b - d_a$ or $\gamma_{ab} = \gamma_b - \gamma_a$, now with respect to the covariates at zero, and to the treatment-covariate interactions $\boldsymbol{\beta}_{2,ab} = \boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}$ (Donegan et al. 2017). This results in four possible scenarios:

1. Consistent treatment effects at zero covariate, consistent interactions;
2. Inconsistent treatment effects at zero covariate, inconsistent interactions;
3. Inconsistent treatment effects at zero covariate, consistent interactions;
4. Consistent treatment effects at zero covariate, inconsistent interactions.

Donegan et al. (2017) suggest using a node-splitting approach to assess consistency of treatment effects and interactions simultaneously. To illustrate, let us consider an AgD meta-regression with a single covariate. For a chosen comparison b' vs. a', we now split estimation of both treatment effects and interactions into direct evidence, $d^{\mathrm{D}}_{a'b'}$ and $\beta^{\mathrm{D}}_{2,a'b'}$ respectively, and indirect evidence from the rest of the network, $d^{\mathrm{I}}_{a'b'}$ and $\beta^{\mathrm{I}}_{2,a'b'}$. The treatment effect of b' vs. a' at a given value x of the covariates, estimated by the direct evidence, is then given by

$$d^{\mathrm{D}}_{a'b'} + x\beta^{\mathrm{D}}_{2,a'b'}, \tag{1.35}$$

and similarly for the indirect evidence by

$$d^{\mathrm{I}}_{a'b'} + x\beta^{\mathrm{I}}_{2,a'b'}. \tag{1.36}$$

These two equations define two lines as a function of x that, when plotted together, should lie parallel (consistent interactions) and have the same intercept (consistent treatment effect at zero covariate). It may be useful to define the inconsistency parameter $\omega_{a'b'}(x)$, in a similar manner to the node-splitting model for standard NMA, except that the inconsistency parameter is now a function of the covariate x:

$$\begin{aligned}\omega_{a'b'}(x) &= d^{\mathrm{D}}_{a'b'} + x\beta^{\mathrm{D}}_{2,a'b'} - d^{\mathrm{I}}_{a'b'} + x\beta^{\mathrm{I}}_{2,a'b'} \\ &= d^{\mathrm{D}}_{a'b'} - d^{\mathrm{I}}_{a'b'} + x\left(\beta^{\mathrm{D}}_{2,a'b'} - \beta^{\mathrm{I}}_{2,a'b'}\right).\end{aligned} \tag{1.37}$$

When plotting the posterior distribution of $\omega_{a'b'}(x)$ against x, this should be a flat horizontal line (consistent interactions) with intercept at zero (consistent treatment effects at zero covariate). Donegan et al. (2017) do not suggest how this approach might extend to interactions with multiple covariates, however there appear to be two options. The first is to consider inconsistency for each covariate interaction separately in turn, following exactly the approach above for each covariate. The second is to node-split all interactions together, replacing $\beta^{\mathrm{D}}_{2,a'b'}$, $\beta^{\mathrm{I}}_{2,a'b'}$, and x above by $\boldsymbol{\beta}^{\mathrm{D}}_{2,a'b'}$, $\boldsymbol{\beta}^{\mathrm{I}}_{2,a'b'}$, and $\boldsymbol{x}$. Visualisations would then be produced for each covariate in turn. It remains to be seen which of these two approaches is preferred.

1.2.8 NMA with contrast-based data

So far, we have considered the NMA scenario where outcome data are available from each treatment arm, such as event counts or mean response. However, it is also possible for outcome data to be reported as summary relative effects on each treatment contrast, such as log odds ratios or mean differences. In this case, the standard approach is to use a Normal likelihood and identity link function for these data (Dias et al. 2011c; Salanti et al. 2007; Woods et al. 2010). Letting $y_{\bullet jab}$ be the observed relative effect of treatment b vs. a in study j with standard error s_{jab}, the NMA model for contrast-based data is then

$$y_{\bullet jab} \sim \mathrm{N}(\delta_{jab}, s^2_{jab}), \tag{1.38}$$

where δ_{jab} is the study-specific relative effect of treatment b vs. a. In studies with three or more arms (and therefore two or more summary relative effects), the corresponding relative effects are correlated. In such cases, a multivariate

Normal likelihood is used to account for these correlations, with marginal distributions given by (1.38) and covariances

$$\operatorname{cov}(y_{\bullet jab}, y_{\bullet jac}) = v^{(a)}_{jbc}. \tag{1.39}$$

If not reported, the covariance $v^{(a)}_{jbc}$ between the relative effects $y_{\bullet jab}$ and $y_{\bullet jac}$ for any two treatments b and c both compared to a may be derived from the standard error s_{jbc} of the c vs. b relative effect using the relation

$$v^{(a)}_{jbc} = \frac{s^2_{jab} + s^2_{jac} - s^2_{jbc}}{2}. \tag{1.40}$$

In cases where the relative effects are simple differences between mean outcomes on each treatment, the covariance $v^{(a)}_{jbc}$ is equal to the variance of the outcome on treatment a (Dias et al. 2011c). When the variance of the outcome on treatment a is not reported, it may be possible to impute from those reported in other trials (Dakin et al. 2011).

Under a fixed effect model, we have

$$\delta_{jab} = d_{ab} = d_b - d_a \tag{1.41}$$

with $d_1 = 0$.

For a random effects model, studies with three or more arms have a multivariate Normal random effects structure, with marginal distributions and correlations given by

$$\delta_{jab} \sim \mathrm{N}(d_b - d_a, \tau^2_{ab}) \tag{1.42a}$$

$$\operatorname{cor}(\delta_{jab}, \delta_{jac}) = \psi^{(a)}_{bc}, \tag{1.42b}$$

and again $d_1 = 0$. For two-arm studies, there is a single summary relative effect and thus a single univariate Normal random effect, with distribution given by (1.42a). Again, we have the second-order consistency equations (1.15) on the heterogeneity variances, and (1.22) on the correlations. If the homogeneous variance assumption is used this implies that all the correlations are equal to 0.5.

Summary 1.3

In this chapter, we began by introducing the problem of population adjustment in a simple two-study scenario, in which an indirect comparison is to be made between two treatments investigated in two different studies (Section 1.1). When the two studies share a common comparator, population adjustment

methods are motivated by the need to adjust for differences in effect modifying variables between the study populations, which would incur bias in a standard indirect comparison. A standard indirect comparison relies on there being no effect modifiers in imbalance between the study populations (the constancy of relative effects assumption), whereas anchored population-adjusted indirect comparisons adjust for effect modifying variables to relax this assumption, assuming that there are no unobserved effect modifiers in imbalance (conditional constancy of relative effects). When there is no common comparator (e.g. if the studies are single-arm), unanchored population adjustment methods aim to improve on a naïve comparison of arms by adjusting for all prognostic and effect modifying variables. However, the assumption that all effect modifiers and prognostic variables are known and have been adjusted for (the conditional constancy of absolute effects assumption) is very strong and difficult to justify.

We then introduced network meta-analysis, which allows larger networks of treatments and studies to be synthesised in a coherent manner (Section 1.2). Standard indirect comparison is a special case of NMA, and (fixed effect) NMA relies on the same constancy of relative effects assumption as standard indirect comparison. (A random effects NMA makes a slightly weaker assumption—constancy of relative effects in expectation—which means that any imbalances in effect modifiers throughout the network are random (we discuss further in Section 2.1.6).) In Section 1.2.1, we outlined the standard parameterisation for writing NMA models, known as the baseline shift parameterisation, in which a reference treatment arm is defined for each trial. The study-specific baselines refer to the reference arm in each trial, and treatments in the other arms of each trial are compared to this as relative effects. We then described an alternative parameterisation in Section 1.2.2, which we term the reference treatment parameterisation. This parameterisation uses the same reference treatment (treatment 1) across the entire network, to which study baseline parameters always refer to, and relative treatment effects in each trial are always against the reference treatment—even if the trial does not include this treatment arm. We showed that the two parameterisations are equivalent (Section 1.2.3) and compared the relative merits of each (Section 1.2.5); we use the reference treatment parameterisation for the remainder of this thesis.

Other parameterisations of NMA models have previously been considered. For example, Lu and Ades (2004) write a model where the study-specific baselines refer to the "average" treatment in the network. This lead to improved computational performance within the MCMC software at the time,

but more recent algorithms perform equally well with the simpler baseline shift parameterisation and as such this parameterisation is seldom used. Previous authors (Hawkins et al. 2015; Tu 2014) have considered models that are also parameterised with respect to a common reference treatment 1 across the network, but that do not account for the correlation structure between random effects. As a result, their models are fundamentally different to the standard baseline shift model (despite similar empirical results in some cases), and have computational difficulties when there are trials in the network with no treatment 1 arm. The `mvmeta` command in Stata has an option to use the reference treatment parameterisation; however, this implementation uses additional "dummy" treatment 1 arms in trials where no treatment 1 arm is present (White 2015).

We considered approaches to assessing model fit and model comparison (Section 1.2.6) and described the inclusion of contrast-based data in NMA (Section 1.2.8). We also introduced IPD network meta-regression (Section 1.2.4) which is the "gold-standard" approach to adjusting for imbalances in effect modifiers throughout the network; however, IPD NMR requires IPD to be available from every study, which is not always possible. Particular issues in all forms of network meta-analysis (both of AgD and of IPD) are heterogeneity and inconsistency, which are both caused by differences in effect modifying variables between studies (either in terms of population characteristics, or other study-level factors such as treatment or outcome definition or study setting). Heterogeneity is assessed using random effects models (Sections 1.2.1, 1.2.2 and 1.2.4), and several approaches exist for assessing inconsistency (Section 1.2.7). Adjusting for effect modifying variables in an IPD NMR is likely to reduce any heterogeneity and inconsistency, however residual heterogeneity or inconsistency indicates that differences in effect modifiers between studies still remain. In the case of a two-study indirect comparison, heterogeneity and inconsistency cannot be checked as there is not enough data.

Thesis overview 1.4

The remainder of this thesis is set out as follows. **Chapter 2** reviews the literature on population adjustment and closely related problems. We describe current methods, including MAIC, STC, and NMR-based approaches, and set out their assumptions and properties. **Chapter 3** contains a systematic review of applications of MAIC and STC in the published literature, and of population

adjustment methods in NICE Technology Appraisals. We aim to determine the ways in which these methods are used and whether the key assumptions are likely to hold, to assess the adequacy of current practice. In **Chapter 4**, we propose a new method for population adjustment called Multilevel Network Meta-Regression, that aims to address the issues with current approaches. **Chapter 5** discusses the computational aspects of implementing ML-NMR models. In **Chapter 6**, we apply ML-NMR to a real example network of plaque psoriasis treatments, and compare with results from MAIC and NMA analyses. **Chapter 7** extends the ML-NMR framework to handle general individual-level likelihoods where the corresponding aggregate-level likelihood does not have a known form, such as survival analysis. We apply the approach to two examples, one with artificial survival outcomes, and another continuing with the plaque psoriasis example to incorporate ordered categorical outcomes. In **Chapter 8**, we perform an extensive simulation study to assess the performance of ML-NMR in a range of scenarios and under various failures of assumptions, comparing with MAIC, STC, and standard indirect comparisons. Finally, in **Chapter 9** we conclude with a discussion, including suggestions for future research. Additional **Appendices** contain listings of Stan code implementing the ML-NMR models, full tabulated results from the simulation study, and details of the computing environments used.

Chapter 2

Review of population adjustment methods

Methods for adjusting treatment effects observed in one population to those that would be observed in another population have been around for many years. In this chapter, we review the literature on methods for population adjustment and other related problems. This chapter was published by the NICE Decision Support Unit as Technical Support Document 18 (Phillippo et al. 2016; also published in abridged form, Phillippo et al. 2018a).

Population adjustment describes a scenario where individual patient data (IPD) in one or more trials are used to adjust for between-trial differences in the distribution of covariates that influence outcome. Recently proposed methods for population adjustment include *Matching Adjusted Indirect Comparison* (MAIC; Ishak et al. 2015; Signorovitch et al. 2010) and *Simulated Treatment Comparison* (STC; Caro and Ishak 2010; Ishak et al. 2015), which are becoming increasingly popular in submissions to regulatory/reimbursement agencies and the wider literature. MAIC and STC are forms of propensity score weighting and outcome regression methods, respectively. These families of methods have previously been used to adjust treatment effects from both randomised and non-randomised studies, under the general headings of "standardisation", "generalisation" or "calibration", and the properties of such methods have been examined in the related literatures. The novelty of population adjustment methods such as MAIC and STC is that they apply the classic propensity score and regression methods to the specific case of indirect comparisons with limited availability of IPD.

Consider the simple two-study scenario introduced in Section 1.1, where IPD are available for an AB trial comparing treatments A and B, and aggregate

data (AgD) are available for an AC trial comparing treatments A and C (see Figure 1.1). The basic premise of MAIC and STC is to effect an indirect comparison between treatments B and C in the AC study population, adjusting for between-trial differences in baseline characteristics. Both methods use IPD from the AB study to form predictions $\hat{y}_{A(AC)}$ and $\hat{y}_{B(AC)}$ of the summary outcomes that would have been observed on treatments A and B in the AC trial. The predicted outcomes are then used to form a population-adjusted indirect comparison in one of two ways, depending on whether or not a common comparator treatment is used. Firstly, the relative effect $d_{BC(AC)}$ may be estimated using an *anchored* comparison, formed as the difference of the relative effects against the common comparator A:

$$\hat{d}_{BC(AC)} = g(\bar{y}_{C(AC)}) - g(\bar{y}_{A(AC)}) - \left(g(\hat{y}_{B(AC)}) - g(\hat{y}_{A(AC)})\right), \tag{2.1}$$

where $\bar{y}_{C(AC)}$ and $\bar{y}_{A(AC)}$ are the mean outcomes observed on each treatment in the AC trial, $\hat{y}_{B(AC)}$ and $\hat{y}_{A(AC)}$ are predicted mean outcomes in the AC population using the population adjustment method, and $g(\cdot)$ is a link function onto some suitable transformed linear scale. The second estimator is an *unanchored* comparison (see Figure 1.2), which does not use a common comparator and is formed as a difference in the mean absolute outcomes on each treatment:

$$\hat{d}_{BC(AC)} = g(\bar{y}_{C(AC)}) - g(\hat{y}_{B(AC)}). \tag{2.2}$$

As we describe in Section 2.3, the anchored indirect comparison respects randomisation and relaxes the assumptions made by standard indirect comparisons, whereas the unanchored indirect comparison requires much stronger assumptions which are very hard to meet.

If IPD are available from every study, an IPD network meta-regression may be performed to adjust for differences in observed effect modifiers between study populations, and is considered the "gold standard" approach (see Section 1.2.4; Berlin et al. 2002; Dias et al. 2011a; Lambert et al. 2002; Riley et al. 2010; Tudur Smith et al. 2005). As a result, there have also been efforts to extend IPD network meta-regression to incorporate aggregate data from published studies (Donegan et al. 2013; Jansen 2012; Saramago et al. 2012; Sutton et al. 2008; Thom et al. 2015). Such developments have thus far proceeded in parallel with the development of MAIC and STC, which are instead largely focused on simple two-study indirect comparisons.

This chapter starts with a review of the earlier literature surrounding population adjustment, including the literature on standardisation, generalisation, and calibration (Section 2.1). We then describe the current methods for population adjustment including MAIC, STC, and network meta-regression based

approaches, noting the similarities and differences with the previous methods (Section 2.2). In Sections 2.3 to 2.5, we detail the particular assumptions that are made by population adjustment methods and discuss specific issues which arise from their practical application, focusing on the scenario of technology appraisal. Finally, in Section 2.6 we conclude with a summary and discussion.

Earlier literature surrounding population adjustment 2.1

We begin with a review of the earlier literature surrounding population adjustment based on propensity score weighting and outcome regression. We start with literature on approaches to standardisation, then look at a related literature on generalisation of treatment effects, and finally at some recent work on calibration.

Model-based standardisation 2.1.1

Standardisation is a method closely related to the kinds of adjustment proposed by MAIC and STC. Here, the mean outcomes or responses to be predicted for a target population P are based on those observed in an unrepresentative sample S from P, taking into account differences in the distributions of characteristics between the sample and full target population. In the following discussion of the methods developed for the standardisation problem, we refer to outcomes under different treatments to remain consistent with our population adjustment scenario; however, in the original context of the standardisation methods, the "treatments" are often exposure classes in an observational context, sampled from a larger target population.

Crude direct standardisation, also known as poststratification, subclassification, or direct adjustment, is a basic method of estimating outcomes in a target population of which the sample is an unrepresentative subpopulation, achieved by stratifying the sample population and reweighting the sample means within each stratum according to the population frequencies (Cochran 1968).

Suppose that the sample population of size $N_{(S)}$ is stratified into Z strata or subclasses based on covariates $\boldsymbol{x}$, with $N_{kz(S)}$ subjects in each stratum z in the sample population receiving treatment k, and $N_{k(S)}$ the total number of subjects receiving treatment k. The target population is of size $N_{(P)}$, with $N_{kz(P)}$ members in each stratum receiving treatment k. Let y_{ikz} denote the response for subject i receiving treatment k in stratum z, which is only observed if the subject is in the sample population (covariate values are however known for

every individual in the full population). The directly standardised estimator of the mean response $\theta_{k(P)} = \mathbb{E}_{(P)}(Y \mid K = k)$ on treatment k in the target population is then

$$\hat{y}_{k(P)} = \frac{1}{N_{k(P)}} \sum_{z=1}^{Z} \sum_{i=1}^{N_{kz(S)}} \frac{N_{kz(P)}}{N_{kz(S)}} y_{ikz}, \tag{2.3}$$

where individuals in strata z receiving treatment k are given weight $\frac{N_{kz(P)}}{N_{kz(S)}}$. Common issues with direct standardisation arise when some strata have small (or zero) membership $N_{kz(S)}$ in the sample population, leading to inflated (or even infinite) weights for these strata; application is further limited by the number of stratification variables, which must also be categorical (or at least discretised in such a manner).

Rosenbaum (1987) proposed a modification of direct standardisation, known as model-based direct standardisation, in which the weights are found using a parametric model rather than observed population frequencies. Individuals in stratum z receiving treatment k are weighted by the inverse of a propensity score estimated using a logistic model, so that the estimator in equation (2.3) becomes

$$\hat{y}_{k(P)} = \frac{1}{N_{k(P)}} \sum_{z=1}^{Z} \sum_{i=1}^{N_{kz(S)}} \frac{y_{ikz}}{p(\boldsymbol{x}_z)}. \tag{2.4}$$

The propensity score (PS) (Rosenbaum and Rubin 1983) is defined in this context as the conditional probability that an individual from the target population is assigned to the sample given the covariates; $p(\boldsymbol{x}_z) = \mathbb{P}\big(\mathbb{I}_{(S)} = 1 \mid \boldsymbol{X} = \boldsymbol{x}_z\big)$, where $\mathbb{I}_{(S)}$ is an indicator of assignment into the trial sample ($\mathbb{I}_{(S)} = 1$) or not ($\mathbb{I}_{(S)} = 0$) from the target population. Rosenbaum (1987) modelled the PS using logistic regression on the observed covariates $\boldsymbol{x}_z$ in each stratum, written in general form as:

$$\text{logit}\big(p(\boldsymbol{x}_z)\big) = h(\boldsymbol{x}_z; \boldsymbol{\alpha}), \tag{2.5}$$

where $\boldsymbol{\alpha}$ is a vector of unknown parameters and $h(\cdot)$ is a known function of the covariates, possibly including interactions or a constant term. A simple linear model for example would be specified with $h(\boldsymbol{x}_z; \boldsymbol{\alpha}) = \alpha_0 + \boldsymbol{x}_z^{\intercal} \boldsymbol{\alpha}_1$.

When correctly specified, the propensity score is a balancing score: conditioning on the PS removes any imbalance in the distribution of $\boldsymbol{x}$ between the sample and target populations. However, incorrect specification of the model in equation (2.5) or the presence of unmeasured effect modifiers or prognostic variables will result in the estimator in equation (2.4) being biased.

Greenland (1991) suggested an alternative form of model-based standardisation where, instead of modelling the propensity score, a generalised linear model (GLM) is fitted to estimate the expected response $\theta_{kz} = \mathbb{E}(Y \mid K = k, \boldsymbol{X} = \boldsymbol{x}_z)$ on treatment k conditional upon the covariates in a stratum z:

$$g(\theta_{kz}) = \beta_{k0} + \boldsymbol{x}_z^\intercal \boldsymbol{\beta}, \tag{2.6}$$

for an appropriate link function $g(\cdot)$ and coefficients β_{k0}, $\boldsymbol{\beta}$. The estimated responses $\hat{\theta}_{kz} = g^{-1}(\hat{\beta}_{k0} + \boldsymbol{x}_z^\intercal \hat{\boldsymbol{\beta}})$ in each stratum z are then weighted by a vector of weights $\boldsymbol{w}$ to match the strata frequencies in the target population, resulting in the estimator

$$\hat{y}_{k(P)} = \boldsymbol{w}^\intercal \left[\hat{\theta}_{k1}, \ldots, \hat{\theta}_{kZ}\right]^\intercal. \tag{2.7}$$

One key difference between this method and that of Rosenbaum (1987), is that the vector of weights $\boldsymbol{w}$ is assumed known—in contrast to Rosenbaum's PS method, which estimates the weights as $1/p(\boldsymbol{x}_z)$ for each stratum z. Again, the estimator is unbiased only if the model is correctly specified, and there are no unmeasured effect modifiers or prognostic variables.

Further propensity score methods 2.1.2

The propensity score, introduced by Rosenbaum and Rubin (1983), has been used in a variety of ways to adjust for imbalances in covariates between a sample population and a target population, of which the sample is an unrepresentative subpopulation. PS adjustment methods in general weight individuals or groups of individuals by the inverse of their PS. Differences between the various methods are found in the coarseness of the weighting applied: at the finest scale to individuals, at the coarsest scale to whole groups or subclasses, or somewhere in between (Stuart et al. 2011). We discuss three common methods here.

Inverse propensity score weighting 2.1.2.1

Inverse propensity score weighting (IPW) applies weights at the finest possible scale; each individual in the trial population is given their own weight. The propensity score is defined for each individual as

$$p(\boldsymbol{x}_i) = \mathbb{P}\big(\mathbb{I}_{(S)i} = 1 \mid \boldsymbol{X} = \boldsymbol{x}_i\big), \tag{2.8}$$

and the logistic regression in equation (2.5) is therefore modified so that $\text{logit}(p(\boldsymbol{x}_i)) = h(\boldsymbol{x}_i; \boldsymbol{\alpha})$. The individual propensity scores (2.8) are then used in the estimator (2.4). However, IPW can result in unstable estimates if extreme

weights are estimated—a problem not evident in coarser weighting schemes which stabilise the weights. Furthermore, simulation studies have shown that IPW is heavily reliant on correct specification of the PS model, and that bias and imprecision are increased by misspecification (Kang and Schafer 2007).

2.1.2.2 Subclassification

Subclassification is the coarsest weighting method, where individuals with similar propensity scores are grouped and the subclass average responses are weighted. The model-based standardisation method of Rosenbaum (1987) sits in this category (see Section 2.1.1), as individuals in the same stratum are assigned the same weight—equivalent to weighting their average responses. Kang and Schafer (2007) showed that the amount of bias reduction of subclassification methods is not as great as an IPW method with a correct PS model; however, the estimator is more efficient and the amount of bias reduction increases with the number of subclasses. Usually only five or six subclasses are used (Rubin 2001), for example by the quintiles of the PS distribution, which may not be enough for sufficient bias reduction (Stuart et al. 2011). Subclassification is also more robust to misspecification of the PS model than IPW.

2.1.2.3 Full matching

Full matching is a weighting method that lies between the two extremes of IPW and subclassification in terms of coarseness of the weights. Subclasses are formed so that each contains at least one individual from both the trial and target populations. Rosenbaum (1991) first proposed the full matching method, and also showed that it is the weighting method which minimises differences in propensity scores between trial and target populations within the subclasses.

2.1.3 Outcome regression

An alternative to PS weighting is known as outcome regression. In this method, instead of modelling the propensity score and applying a weighting scheme to the sample subjects, a model for the mean response (or outcome) given treatment and observed covariates $\theta_k(\boldsymbol{x}) = \mathbb{E}(Y \mid K = k, \boldsymbol{X} = \boldsymbol{x})$ is fitted:

$$g(\theta_k(\boldsymbol{x})) = m\left(k, \boldsymbol{x}; \boldsymbol{\beta}\right) \tag{2.9}$$

where $\boldsymbol{\beta}$ is an unknown parameter, $m(\cdot)$ is a known function, possibly including interactions or a constant term, and $g(\cdot)$ is an appropriate link function.

Estimators of expected outcomes in treatment group k in the target population P are then formed using prediction based on the observed covariates in the target population:

$$\hat{y}_{k(P)} = \frac{1}{N_{k(P)}} \sum_{i=1}^{N_{k(P)}} g^{-1}\left(m\left(k, \boldsymbol{x}_{ik}; \hat{\boldsymbol{\beta}}\right)\right) \tag{2.10}$$

These are unbiased if the outcome model is correctly specified and there are no unmeasured covariates. Simulation studies (Kang and Schafer 2007) have shown that estimators based on a misspecified outcome regression model are less biased and more efficient than estimators based on a misspecified PS model, however the associated precisions are overestimated.

Outcome regression is similar to the standardisation method proposed by Greenland (1991), in the sense that a model for the outcome conditional on covariates is fitted, however the methods differ in some key ways. Firstly, the regression model in equation (2.9) is more general than that originally proposed by Greenland in equation (2.6), which is a special case with $m(k, \boldsymbol{x}; \boldsymbol{\beta}) = \beta_k + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}$. Greenland's model-based standardisation method fits the outcome regression at the level of mean response within a subclass, however outcome regression fits a model at the individual response level. Furthermore, Greenland's method estimates the mean response within each subclass in the trial population and then weights the subclass responses by the relative frequencies of the subclasses in the target population (equation (2.7)); outcome regression predicts the response for each individual in a target population based on their covariate values, and then takes the average (equation (2.10)).

Doubly robust estimation 2.1.4

Both propensity score weighting/matching and outcome regression provide methods to estimate outcomes in a target population from a sample subpopulation that differs in covariate balance. However, the estimators are only unbiased if their respective models (for propensity score or outcome) are correctly specified, and if there are no unmeasured effect modifiers or prognostic variables.

Doubly robust (DR) estimators aim to reduce the impact of model misspecification by incorporating both outcome regression and propensity score models into one estimator, which is consistent (and for some estimators unbiased) if at least one of the constituent models is correct. Doubly robust estimators can be constructed in various ways (Funk et al. 2011; Kang and Schafer 2007; Robins et al. 2007). Robins et al. (2007) refer to DR estimators of

the general form

$$\hat{y}_{k(P)} = \frac{1}{N_{k(P)}} \sum_{i=1}^{N_{k(P)}} \hat{\theta}_k(\boldsymbol{x}_{ik}) + \frac{1}{N_{k(P)}} \sum_{i=1}^{N_{k(S)}} \frac{1}{\hat{p}(\boldsymbol{x}_{ik})} \left(y_{ik} - \hat{\theta}_k \left(\boldsymbol{x}_{ik} \right) \right) \tag{2.11}$$

or, normalising the second term by the sum of the weights, the form

$$\hat{y}_{k(P)} = \frac{1}{N_{k(P)}} \sum_{i=1}^{N_{k(P)}} \hat{\theta}_k(\boldsymbol{x}_{ik}) + \frac{\sum_{i=1}^{N_{k(S)}} \frac{1}{\hat{p}(\boldsymbol{x}_{ik})} \left(y_{ik} - \hat{\theta}_k \left(\boldsymbol{x}_{ik} \right) \right)}{\sum_{i=1}^{N_{k(S)}} \frac{1}{\hat{p}(\boldsymbol{x}_{ik})}} \tag{2.12}$$

where $\hat{\theta}_k(\boldsymbol{x}) = g^{-1}(m(k, \boldsymbol{x}; \hat{\boldsymbol{\beta}}))$ is the outcome regression estimator of the mean response at covariate value $\boldsymbol{x}$, and $\hat{p}(\boldsymbol{x})$ is the estimated propensity score. Note that, in the DR estimators (2.11) and (2.12) above, the first term is a summation over every individual in the full target population (of which the sample population is a subpopulation), and the second term is a summation over only those individuals in the sample. (Recall that, in this standardisation scenario, individuals not in the sample are missing outcomes y_{ik}, but covariate values are known for every individual in the full population.) The estimator proposed by Funk et al. (2011) is also of this form.

These estimators are consistent (they converge to the true value as sample size increases) and unbiased when at least one of the PS or outcome models is correct. Funk et al. (2011) show that these estimators can be rearranged into an unbiased estimator plus an augmentation term—the product of the biases in the PS and outcome regression models—which vanishes when either model is correct. Robins et al. (2007) however note that DR estimators in this form do not have the property of *boundedness*: when the sample population is of finite size, the estimates do not lie in the parameter space with probability 1. This is an issue when the parameter space is not the entire real line, for example when fitting logistic models for the probability of an event. Robins et al. suggest that it is primarily desirable for an estimator to display boundedness, possibly at the expense of unbiasedness, and describe the construction of three DR estimators, known as regression doubly-robust estimators, which have the boundedness property. Each of these three regression DR estimators is a modification of the outcome regression estimator in equation (2.10):

1. Modify the outcome regression model to include the inverse of the propensity score $\hat{p}(\boldsymbol{x}_{ik})^{-1}$ as a covariate;

2. Modify the outcome regression model to use weighted least squares (WLS), using the inverse propensity scores as weights;

3. Modify the outcome regression model to include the propensity score $\hat{p}(\boldsymbol{x}_{ik})$ as a covariate.

The first estimator (also suggested by Bang and Robins 2005), including the inverse PS as a covariate in the outcome regression, can perform very poorly when the inverse PS are highly variable, due to large extrapolations being made from the sample population to the target population. The second estimator, using the inverse PS as weights for WLS outcome regression, does not suffer from this issue, and is expected in general to outperform the first. The third estimator, including the propensity score as a covariate in the outcome regression, should also fare better than the first in the case of highly variable weights, perhaps as well as the second. Doubly robust methods "give the analyst two chances" to specify correct models (Bang and Robins 2005), and demonstrate little loss of efficiency in practice. If neither model is correctly specified then the resulting estimator will be biased and inconsistent.

Generalising treatment effects to a target population 2.1.5

There is substantial literature developed to generalise estimates of relative treatment effects obtained from a RCT into a target population. The methods used are broadly similar to the standardisation literature discussed so far, including propensity score methods and outcome regression. Here, the quantity of interest is the average relative treatment effect in the target population (the population average treatment effect, or PATE)—in contrast to the standardisation literature, which in general is interested in standardising expected outcomes (or absolute effects) to a target population P. Mathematically, the PATE (on a suitable scale with transformation $g(\cdot)$) is

$$d_{AB(P)} = \frac{1}{N_{(P)}} \sum_{i=1}^{N_{(P)}} \big(g(Y_i(B)) - g(Y_i(A)) \big) , \tag{2.13}$$

where $N_{(P)}$ is the number of individuals in the target population, and $Y_i(A)$ and $Y_i(B)$ are the potential outcomes for individual i on treatments A and B respectively (one of which is unobserved); we could equivalently write the RHS of equation (2.13) in (finite sample) expectation notation as $\mathbb{E}_{(P)}\big(g(Y(B)) - g(Y(A))\big)$, where the expectation is with respect to the joint covariate distribution in the target population P. The PATE may be estimated by generalising the sample average treatment effect (SATE), given by $d_{AB} = N_{(S)}^{-1} \sum_{i=1}^{N_{(S)}} \big(g(Y_i(B)) - g(Y_i(A)) \big)$ where the sum is over the $N_{(S)}$ individuals in the sample (trial) population S, which is the usual quantity estimated

by a RCT. Some authors (e.g. Hartman et al. 2015) focus on estimating a related quantity for the treated population only—the population average treatment effect on the treated (PATT), estimated from the sample average treatment effect on the treated (SATT)—which is pertinent in some policy decisions. The PATT and SATT are analogous to the PATE and SATE, except expectation of treatment effect is taken only over the individuals actually assigned treatment. In a RCT, SATE and SATT are asymptotically equal due to randomisation (assuming sufficiently large sample size and proper randomisation).

Of particular significance in this literature are the introduction of a rigorous decomposition of the biases in estimating PATE (Imai et al. 2008), and tests for generalisability which provide means to verify the assumptions required.

The underlying assumptions required for generalisability and valid estimation of PATE are given by several authors (Hartman et al. 2015; Stuart et al. 2011):

1. **Homogeneity of outcomes on each treatment.** Outcomes on treatment and control are the same whether the individual is assigned to the trial or not.

2. **Stable unit treatment value.** The outcomes of one individual are not affected by any other individuals.

3. **Strongly ignorable treatment assignment.** Treatment assignment is random and independent of sample selection from the target population given the observed covariates. This means that there are no prognostic factors or effect modifiers in imbalance between arms of a study.

4. **Strongly ignorable sample assignment.** There are no unmeasured variables related to both sample selection and outcome and, given observed covariates, each individual in the target population has a non-trivial probability (i.e. not zero or one) of being selected into the trial sample.

Assumption 1 may be violated by, for example, protocol differences in inclusion/exclusion criteria (Hartman et al. 2015). Assumption 2 is met by appropriate study design, and is necessary for causal inference. Assumption 3 is met in RCTs by proper randomisation, where treatment assignment only depends on the observed covariates. Assumption 4 is violated if there are unmeasured effect modifiers or prognostic variables in imbalance between the populations.

In order to assess the assumptions required for generalisability, several authors have proposed what are known as placebo tests, proposed by Stuart et al. (2011) in the context of PS models and more generally by Hartman et al. (2015). Outcomes on the placebo (or control) arm of the trial are generalised to the target population, and then compared with the outcomes observed in the target population. The null hypothesis is that there is no difference in the average outcome between populations; however, tests of this null hypothesis can have low power, particularly if conditional outcomes by subgroup are investigated or if the outcome measure has a large variance (Hartman et al. 2015). An alternative proposition is to use the reverse null hypothesis that there is a difference in average control outcome between populations, a test of which will then only support generalisability if there is sufficient evidence and sufficient power to reject the null (Hartman and Hidalgo 2011).

Placebo tests can demonstrate failures of assumptions 1, 2, and 4 above, however they cannot ascertain which assumption or assumptions are violated, nor can they detect multiple violations whose resulting biases cancel each other out (Hartman et al. 2015). Furthermore, a placebo test only has capacity to check for unobserved prognostic variables in imbalance in assumption 4 but not for unobserved effect modifiers in imbalance. A "placebo" test comparing observed and predicted outcomes on a common active comparator (if available) would additionally be able to detect unobserved effect modifiers in imbalance (but would not be able to discern whether the unobserved covariate was an effect modifier or prognostic variable).

When PS methods are used to generalise relative treatment effects, Stuart et al. (2011) suggest examining the difference in average propensity scores between the trial population and target population; a difference in the mean PS greater than 0.25 standard deviations indicates that the generalisation is largely based on extrapolation, and will be heavily dependent on the PS model used.

Calibration of treatment effects 2.1.6

The literature reviewed thus far seeks to generalise either the outcomes (Sections 2.1.1 to 2.1.4) or the relative treatment effects (Section 2.1.5) observed in a sample sub-population, under some strict assumptions, to those that would be observed in a target population. There has however been no attempt to perform treatment comparisons in the target population, which is our problem of interest. Additionally, we now wish to consider the sample and target populations as distinct and independent (e.g. from two non-overlapping clinical

trials), whereas previously the sample was considered an unrepresentative subpopulation of the target population. Several authors have framed this as a calibration problem, where information on treatment effects and covariates in one population is used to estimate treatment effects in another population with different known covariate values (Nie and Soon 2010; Nie et al. 2013; Zhang 2009; Zhang et al. 2015), and note that it is similar to the generalisation problem (Section 2.1.5).

The work on calibration assumes that IPD is available on both the AB and AC trials, so the methods proposed are not strictly relevant to the problem that MAIC and STC set out to address. However, we review this literature here because it contains some clear statements of the assumptions made by MAIC and STC.

Recently there has been a specific interest from the US Food and Drug Administration (FDA) in calibration methods for the analysis of non-inferiority studies, which compare a treatment A with an active comparator B and thus lack a placebo arm C. When the quantity of interest is treatment effect relative to placebo, a historical placebo-controlled trial with the active comparator may be used to calibrate the treatment effect by estimating the placebo effect that would be observed in the AB trial (known as a *putative placebo analysis*).

Calibration methods are interested in estimating the average relative treatment effect of B vs. C in the AB population on an appropriate scale, which can be done in one of several ways depending on the assumptions one is willing to make. (This is in contrast with MAIC/STC, where the target of inference is the B vs. C effect that would be observed in the AC trial if the B arm was included.)

The first possibility is an approach based on the assumption of *constancy of absolute effects*

$$\mathbb{E}_{(AB)}\big(g(Y(C))\big) = \mathbb{E}_{(AC)}\big(g(Y(C))\big), \tag{2.14}$$

which requires that there are no prognostic variables or effect modifiers in imbalance between the two populations. This is of course absurd, as there is no randomisation between trials—only within. No accepted methods for evidence synthesis or indirect comparison, whether population-adjusted or not, make this impossibly strong assumption (see Table 2.1).

Another possibility is an approach based on the assumption of *conditional constancy of absolute effects* (also known as *treatment-specific conditional constancy* in the calibration literature):

$$\mathbb{E}_{(AB)}\big(g(Y(C)) \mid \boldsymbol{X}_{(AB)}\big) = \mathbb{E}_{(AC)}\big(g(Y(C)) \mid \boldsymbol{X}_{(AC)}\big). \tag{2.15}$$

This means that the expected absolute outcomes under treatment C are identical between the two trial populations at any given set of covariate values. This assumption is very strong (if not implausibly so), as it requires all effect modifiers and prognostic variables to be available (Zhang et al. 2015). Estimation of the indirect comparison under this assumption proceeds via one of the previously discussed methods (e.g. propensity score weighting, outcome regression), which is used to predict mean absolute outcomes $\hat{y}_{C(AB)}$ in the AB population. $\mathbb{E}_{(AB)}\big(g(Y(B))\big)$ is estimated as $\bar{y}_{B(AB)}$ in standard fashion directly from the AB trial. The result is an unanchored comparison in the AB population,

$$\hat{d}_{BC(AB)} = g(\hat{y}_{C(AB)}) - g(\bar{y}_{B(AB)}). \tag{2.16}$$

(This is the same idea as the unanchored comparison in equation (2.2), except that here the comparison is made in the AB population.) We note that treatment-specific conditional constancy is equivalent to ignorable sample assignment as described in the generalisation literature (assumption 4, Section 2.1.5).

To avoid making such a strong assumption about prognostic variables, inferences could be made instead using an assumption of *constancy of relative effects* (sometimes referred to simply as *constancy*), meaning that the relative C vs. A effect observed in the AC trial is identical to that which would be observed in the AB trial:

$$d_{AC(AB)} = d_{AC(AC)}. \tag{2.17}$$

However, this is often questionable, as constancy of relative effects requires that all effect modifiers (whether measured or unmeasured) are balanced between the two trial populations. This is akin to the consistency assumption (on the transformed scale) that is standard in NMA (Lu and Ades 2006): consistency is assumed to hold exactly for a fixed effect analysis, and is relaxed in a random effects analysis where consistency is only assumed to hold in expectation. The consistency assumption in random effects models is reasonable when contrasts are informed by many trials, allowing the impact of effect modifiers to "balance out", but less so in sparse networks. Development of population adjustment for the very sparse networks of comparisons often seen in submissions to NICE is therefore well motivated.

Instead of making any of the three strong assumptions above, calibration methods rely on an assumption of *conditional constancy of relative effects* (sometimes referred to simply as *conditional constancy*):

$$\gamma_{AC(AB)}(\boldsymbol{x}) = \gamma_{AC(AC)}(\boldsymbol{x}), \tag{2.18}$$

where $\gamma_{AC(AB)}(\boldsymbol{x}) = \mathbb{E}_{(AB)}\big(g(Y(C)) - g(Y(A)) \mid \boldsymbol{X} = \boldsymbol{x}\big)$ is the relative conditional treatment effect. This means that the relative C vs. A effect observed in the AC trial at a given covariate value (e.g. the effect at age 55) is equal to the C vs. A effect which would be observed in the AB trial at that same covariate value. This assumption may be more valid, as only effect modifiers are required to be adjusted for; estimators based on the conditional constancy of relative effects assumption respect randomisation which balances prognostic variables within studies. $d_{AC(AB)}$ is identified using the relation $d_{AC(AB)} = \mathbb{E}_{(AB)}\big(\gamma_{AC(AB)}(\boldsymbol{x})\big)$. $d_{AB(AB)}$ is estimated in standard fashion from the AB trial. The result is an anchored comparison in the AB population,

$$\hat{d}_{BC(AB)} = \hat{d}_{AC(AB)} - \bar{d}_{AB(AB)}. \tag{2.19}$$

(This is the same idea as the anchored comparison in equation (2.1), except that here the comparison is made in the AB population.)

Calibration methods have been proposed in various forms: covariate adjustment, which is a form of outcome regression (Zhang 2009); likelihood reweighting, which is a form of PS weighting (Nie et al. 2013); and doubly robust methods (Zhang et al. 2015). Another estimator recently proposed by Zhang et al. (2015) is known as a conditional effect estimator, which models the conditional relative effect $\gamma_{AC(AB)}(\boldsymbol{x})$ directly rather than modelling (transformed) outcomes, and may also be combined into doubly robust estimators. We are yet to see any published applications of conditional effect models. In practice, all of the above methods require IPD on the historical AC trial in order to infer comparisons in the AB population; this differs from the calibration scenarios into which MAIC and STC have been proposed, where IPD on the AC trial are unavailable and comparisons are inferred in the AC population. Zhang (2009) notes that covariate adjustment may be performed using aggregate data from the AC trial if the coefficients in the outcome regression and their covariance matrix are published, although this seems unlikely.

2.2 Population adjustment combining IPD and AgD

The core principles of MAIC and STC remain the same as in the general calibration literature (Section 2.1.6), however the problem scenario is modified slightly: rather than IPD being available in all study populations, IPD are only available in the AB trial, with AgD in the AC trial along with information on the covariate distribution. Ideally the full joint distribution of $\boldsymbol{x}$ is known, but frequently in practice only marginal summary statistics for each covariate

are known (e.g. mean and standard deviation for continuous covariates, proportions for discrete covariates). Due to the lack of IPD from the AC trial, standard approaches to fitting both propensity score and outcome models may not be used. We outline both MAIC and STC below (Sections 2.2.1 and 2.2.2), along with alternative approaches arising from the literature on network meta-regression (Section 2.2.3).

Matching-Adjusted Indirect Comparison (MAIC) 2.2.1

MAIC is a form of non-parametric likelihood reweighting (see Section 2.1.6), which allows the propensity score logistic regression model to be estimated without IPD in the AC population. The mean outcome on treatment $k = A, B$ in the AC target population is estimated by taking a weighted average of the outcomes $y_{ik(AB)}$ of the $N_{k(AB)}$ individuals on treatment k in the AB population

$$\hat{y}_{k(AC)} = \frac{\sum_{i=1}^{N_{k(AB)}} y_{ik(AB)} w_{ik}}{\sum_{i=1}^{N_{k(AB)}} w_{ik}}, \tag{2.20}$$

where the weight w_{ik} assigned to the i-th individual receiving treatment k is equal to the odds of being enrolled in the AC trial vs. the AB trial. Conceptually, this is very similar to the inverse propensity weighting method discussed in the standardisation literature (Section 2.1.2.1). As with likelihood reweighting (from which MAIC is derived), the weights themselves are estimated using logistic regression as $\log(w_{ik}) = \alpha_0 + \boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\alpha}_1$, where $\boldsymbol{x}_{ik}$ is the covariate vector for the i-th individual receiving treatment k; however, the regression parameters are not estimable using standard methods due to the lack of IPD in the AC trial, in particular a lack of information on the joint distribution of covariates. If the joint covariate distribution was available in the AC trial, then the likelihood reweighting approach of Nie et al. (2013) would be feasible, with the possibility of the sufficient statistics replacing the full IPD. Because only marginal information is available, Signorovitch et al. (2010) propose using a method of moments to estimate $\boldsymbol{\alpha}_1$ so that the weights exactly balance the mean covariate values (and any included higher order terms, for example squared covariate values to balance the variance) between the weighted AB population and the AC population. When $\bar{\boldsymbol{x}}_{(AC)} = \boldsymbol{0}$, Signorovitch et al. show that this is equivalent to minimising $\sum_{k=A,B}\sum_{i=1}^{N_{k(AB)}} \exp(\boldsymbol{x}_{ik}^{\intercal}\boldsymbol{\alpha}_1)$. The estimator in equation (2.20) is then equal to

$$\hat{y}_{k(AC)} = \frac{\sum_{i=1}^{N_{k(AB)}} y_{ik(AB)} \exp\left(\boldsymbol{x}_{ik}^{\intercal}\hat{\boldsymbol{\alpha}}_1\right)}{\sum_{i=1}^{N_{k(AB)}} \exp\left(\boldsymbol{x}_{ik}^{\intercal}\hat{\boldsymbol{\alpha}}_1\right)},$$

noting that $\exp \hat{\alpha}_0$ cancels from the top and bottom of the fraction. Anchored and unanchored indirect comparisons are then formed using equations (2.1) and (2.2) respectively. Although MAIC can be used to facilitate indirect comparisons on any scale, the MAIC literature almost exclusively performs comparisons on the natural outcome scale (i.e. with $g(\cdot)$ the identity function). Typically, standard errors for MAIC estimates are calculated using a robust sandwich estimator (White 1980) (see the appendix of Signorovitch et al. (2010)). Sandwich estimators are derived empirically from the data rather than making overly strong assumptions about the weights, to account for the fact that the weights are estimated rather than fixed and known. Signorovitch et al. (2010) suggest that the effective sample size (ESS) of the pseudo-population formed by weighting the AB population is approximated by

$$\text{ESS} = \frac{\left(\sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} \hat{w}_{ik}\right)^2}{\sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} \hat{w}_{ik}^2} \tag{2.21}$$

This approximate ESS is only accurate if the weights are fixed and known, or if they are uncorrelated with outcome—neither of which is true here; as such, this approximation is likely to be an underestimate of the true ESS (Vartivarian and Little 2004). However, small effective sample sizes are an indication that the weights are highly variable due to a lack of population overlap, and that the estimate may be unstable. The distribution of weights themselves should also be examined directly, to diagnose population overlap and to highlight any overly influential individuals. It is not possible to apply traditional propensity score tools for "balance checking" here, as propensity scores are only estimated for the AB trial, and the method of moments by definition ensures covariate balance (at least in the means, and up to the level of information published in the AC trial).

2.2.2 Simulated Treatment Comparison (STC)

STC is a modification of covariate adjustment (see Section 2.1.6). Firstly, an outcome model is fitted using the IPD in the AB trial:

$$g\left(\theta_{k(AB)}(\boldsymbol{x})\right) = \mu_{(AB)} + \boldsymbol{\beta_1}^{\mathsf{T}}\boldsymbol{x} + \left(\boldsymbol{\beta_2}^{\mathsf{T}}\boldsymbol{x}^{\text{EM}} + \gamma_B\right) \mathbb{I}(k = B), \tag{2.22}$$

where $\mu_{(AB)}$ is an intercept term, $\boldsymbol{\beta}_1$ is a vector of coefficients for prognostic variables, $\boldsymbol{\beta}_2$ is a vector of coefficients for effect modifiers $\boldsymbol{x}^{\text{EM}}$ (a subvector of the full covariate vector $\boldsymbol{x}$), γ_B is the individual-level relative effect of treatment B compared to A at $\boldsymbol{x} = \boldsymbol{0}$, and $\theta_{k(AB)}(\boldsymbol{x})$ is the expected outcome of an

individual assigned treatment k with covariate values $\boldsymbol{x}$ which is transformed onto a chosen linear predictor scale with link function $g(\cdot)$.

The model in equation (2.22) is a more general form of that given by Ishak et al. (2015), which does not include any effect modifier terms. Ishak et al. then form (on the natural outcome scale) either an unanchored indirect comparison $\hat{d}_{BC(AC)} = \bar{y}_{C(AC)} - \hat{y}_{B(AC)}$, or an anchored indirect comparison $\hat{d}_{BC(AC)} = \bar{y}_{C(AC)} - \bar{y}_{A(AC)} - \left(\hat{y}_{B(AC)} - \hat{y}_{A(AC)}\right)$, where $\hat{y}_{A(AC)}$ and $\hat{y}_{B(AC)}$ are predicted from the outcome regression by substituting in mean covariate values to obtain

$$\hat{y}_{A(AC)} = g^{-1}\left(\hat{\mu}_{(AB)} + \hat{\boldsymbol{\beta}}_1^{\intercal}\bar{\boldsymbol{x}}_{(AC)}\right)$$

and

$$\hat{y}_{B(AC)} = g^{-1}\left(\hat{\mu}_{(AB)} + \hat{\boldsymbol{\beta}}_1^{\intercal}\bar{\boldsymbol{x}}_{(AC)} + \hat{\boldsymbol{\beta}}_2^{\intercal}\bar{\boldsymbol{x}}_{(AC)}^{\text{EM}} + \hat{\gamma}_B\right).$$

These estimators (and hence any indirect comparison based on them) are systematically biased whenever $g(\cdot)$ is not the identity function, because the mean outcome depends on the full distribution of the covariates and not just their mean (Ishak et al. 2015). Instead of substituting in mean covariate values, Ishak et al. suggest that estimates are obtained by first drawing samples from the joint covariate distribution in the AC trial and then averaging over the predicted outcomes based on the regression model. This simulation approach however introduces additional variation as, rather than computing an average over the distribution of covariates in the AC population, the estimated quantity is now the expected effect for a randomly selected individual from the AC population (i.e. the predictive distribution), leading to an underestimate of the precision of the final indirect comparison estimate.

Forming indirect comparisons directly on the natural outcome scale, as advocated by the STC literature and described above, causes several problems due to a conflict between the scale of the linear predictor and the scale of the indirect comparison (see Section 2.3.1.2). To avoid these, we strongly recommend that anchored and unanchored indirect comparisons are formed on the linear predictor scale using equations (2.1) and (2.2) respectively. Standard tools for model checking (such as AIC/DIC, examining residuals, etc.) may be used when constructing the outcome model in the AB trial; however (as with MAIC), additional assumptions are required to predict outcomes in the AC population, which are difficult to test when there is little data available.

Whilst the above formulation of STC is seen in Ishak et al. (2015) and in all the published applications of STC to date, an earlier paper (Ishak 2014)

suggests that an indirect comparison may be performed in the AB population via extension to the above steps. We have not identified any applications employing this method.

2.2.3 Network Meta-Regression

If individual patient data are available on both the AB and AC studies, a network meta-regression using IPD is the gold standard approach (Berlin et al. 2002; Dias et al. 2011a; Lambert et al. 2002; Riley et al. 2010; Tudur Smith et al. 2005). There has, understandably, been interest in generalising network meta-regression to situations where combinations of IPD and AgD are available in a network of treatment comparisons; the scenario with one AB IPD study and one AC aggregate study is then a special case. Currently, there are two main forms of network meta-regression which combine both IPD and aggregate data, which primarily differ in how the regression model is defined at the individual level and at the aggregate level. We discuss both approaches here, in the context of the two-study scenario.

The first approach (Donegan et al. 2013; Jansen 2012; Saramago et al. 2012; Thom et al. 2015) builds upon models previously proposed for pairwise meta-regression (Riley et al. 2008; Riley and Steyerberg 2010; Sutton et al. 2008). Two regression models are fitted simultaneously, one describing individual level outcomes in the AB trial, and another describing the aggregate outcome in the AC trial:

Individual:

$$g\left(\theta_{k(AB)}(\boldsymbol{x})\right) = \mu_{(AB)} + \boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{x} + \left(\boldsymbol{\beta}_2^{\mathsf{T}}\boldsymbol{x}^{\mathrm{EM}} + \gamma_B\right)\mathbb{I}(k = B) \tag{2.23a}$$

Aggregate:

$$g\left(\theta_{\bullet k(AC)}\right) = \mu_{(AC)} + \left(\boldsymbol{\beta}_2^{\mathsf{T}}\bar{\boldsymbol{x}}_{(AC)}^{\mathrm{EM}} + \gamma_C\right)\mathbb{I}(k = C) \tag{2.23b}$$

Due to the lack of data, there are some restrictions on the more general models which have been proposed for larger networks (Donegan et al. 2013; Saramago et al. 2012) and for pairwise meta-regression of multiple studies (Riley et al. 2008; Riley and Steyerberg 2010): a fixed effects model must be used, and the treatment by effect modifier interaction coefficient $\boldsymbol{\beta}_2$ must be shared between treatments B and C and between the individual and aggregate level. This second restriction is at first glance akin to the shared effect modifier assumption discussed later (Section 2.5), although on further inspection it is far stronger—the effect modifier is required to act in the same manner on

both the aggregate level and on the individual level. This assumption is only valid if the identity link is used and all effect modifiers are accounted for (and proper randomisation has occurred); imposing this assumption when it does not hold results in aggregation bias (a form of ecological bias) (Berlin et al. 2002; Donegan et al. 2013; Rothman et al. 2008; Saramago et al. 2012).

The second approach derives from a type of model proposed by Jackson et al. (2006, 2008) known as hierarchical related regression. This model avoids the pitfalls of the first by correctly relating the individual and aggregate levels so that aggregation bias does not occur. The basic idea is a natural one; the aggregate data arise from averaging over a population of individuals, so the aggregate level model arises from averaging (i.e. integrating) the individual model over a population. The resulting model may be written in most general form as

Individual:

$$g\left(\theta_{k(AB)}(\boldsymbol{x})\right) = \mu_{(AB)} + \boldsymbol{\beta}_1^{\intercal}\boldsymbol{x} + \left(\boldsymbol{\beta}_2^{\intercal}\boldsymbol{x}^{\mathrm{EM}} + \gamma_B\right) \mathbb{I}(k = B) \tag{2.24a}$$

Aggregate:

$$\theta_{\bullet k(AC)} = \int_{\mathfrak{X}} g^{-1}\left(\mu_{(AC)} + \boldsymbol{\beta}_1^{\intercal}\boldsymbol{x} + \left(\boldsymbol{\beta}_2^{\intercal}\boldsymbol{x}^{\mathrm{EM}} + \gamma_C\right) \mathbb{I}(k = C)\right) f_{(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \tag{2.24b}$$

where $f_{(AC)}(\boldsymbol{x})$ is the joint distribution of $\boldsymbol{x}$ in the AC trial population, and $\mathfrak{X}$ is the support of $\boldsymbol{x}$. If the full joint distribution is not available for the AC trial (as is likely with published data), an approximation may be used—for example by assuming a Normal distribution (or another appropriate distribution, such as log Normal) for continuous covariates with the reported mean and standard deviation, and either imputing correlations between covariates from the AB trial or assuming that they are zero. Note that this model reduces to the gold-standard IPD network meta-regression when IPD are available for all studies, and is equally applicable for analysing larger networks of treatments with a mixture of IPD and aggregate data available. When used in the simple two-study scenario, model (2.24) does require the shared effect modifier assumption (see Section 2.5) in order to estimate the parameters due to lack of data; however, this assumption may not be required when a larger network of studies is available, or perhaps if external information on the effect modifiers of treatment C is available. Model (2.24) is equivalent to model (2.23) if an identity link is used and all effect modifiers are accounted for.

The individual level model (2.24a) here is of the same form as above in model (2.23a). The aggregate level model (2.24b) however is found by integration of this individual level model, and therefore may not be straightforward to

explicitly write down. Jansen (2012) describes a special case of model (2.24) for the simple case of a binary outcome and binary covariates. When all covariates are binary (or categorical), it is simple to rewrite the integration as a sum over each level of the covariates, so that the aggregate level model (2.24b) becomes

$$\theta_{\bullet k(AC)} = \sum_{z} g^{-1}\left(\mu_{(AC)} + \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{x}_z + \left(\boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{x}_z^{\mathrm{EM}} + \gamma_C\right)\mathbb{I}(k = C)\right) f_{(AC)}(\boldsymbol{x}_z) \qquad (2.25)$$

where $\boldsymbol{x}_z$ is a discrete level of the covariates, and $f_{(AC)}(\boldsymbol{x}_z)$ is simply the proportion of AC trial individuals in the category $\boldsymbol{x}_z$. We are not currently aware of any more general applications of model (2.24) in the literature; in the absence of a more sophisticated approach, model (2.25) may be used to incorporate continuous covariates by splitting them into discrete categories (e.g. splitting ages into 5 year bands), at the expense of loss of information.

An alternative approach to avoiding aggregation bias, proposed by Yamaguchi et al. (2014) for pairwise meta-regression combining IPD and AgD studies, is to use multiple imputation to generate simulated IPD for the AgD studies. Each set of simulated IPD from the AgD studies may then be analysed in an IPD meta-regression along with the IPD, before combining the results using Rubin's rules (Rubin 1987). This approach should avoid aggregation bias, since an individual-level model is used for all studies. However, the imputation algorithm described by Yamaguchi et al. (2014) is not immediately applicable to indirect comparisons or larger networks, assumes covariates are Normally distributed, and is limited to fixed effect models.

The hierarchical network meta-regression approach in model (2.24) represents an alternative class of methods to those such as MAIC and STC. The hierarchical approach models individual-level relationships and is able to provide internally consistent inferences at both the individual level and at an aggregate level like a standard indirect comparison. Methods such as MAIC and STC use IPD to predict average outcomes on study arms, and then effect the indirect comparison at the aggregate study level. We could therefore refer to MAIC and STC as forms of *population-adjusted study-level indirect comparisons*, and the hierarchical approach as a form of *population-adjusted individual-level indirect comparison*. Despite the apparent benefits of the hierarchical approach, MAIC and (to a lesser extent) STC are the most widely used approaches in the applied literature and in submissions to NICE (Chapter 3). We therefore focus largely on the properties and assumptions of MAIC and STC for the remainder of this chapter. However, we expect that these properties and assumptions will apply to population adjustment methods in general, and in particular those for STC will broadly apply to network meta-regression based

approaches (including those that we go on to develop in later chapters).

Other forms of population reweighting 2.2.4

The application of weights to individuals in the IPD population in order to balance the covariate distributions between trials is a general technique which we shall refer to as *population reweighting*. MAIC as described in Section 2.2.1 is currently the most widely used form of population reweighting when IPD are only available for the AC trial.

Belger et al. (2015a,b) suggest another form of population reweighting based on entropy balancing (Hainmueller 2012). Rather than seeking to estimate a propensity score with which to create weights, entropy balancing methods are designed to estimate weights by directly matching moments of the covariate distributions (such as the mean and standard deviation). As MAIC uses the method of moments to estimate weights, the methods up to this point are effectively identical. However, entropy balancing methods apply an additional constraint when estimating the weights; the optimal entropy balancing weights are those which are as close as possible to uniform weights (that is, as close as possible to no weighting at all). This additional constraint means that entropy balancing methods should have equal or reduced standard error compared to MAIC, whilst achieving the same reduction in bias. However, as we now show, estimation of weights via entropy balancing and standard MAIC (using the method of moments) are in fact entirely equivalent.

To see this, let us consider the objective functions that are minimised for MAIC and for entropy balancing. As we described in Section 2.2.1, after centring the IPD covariates around the means in the AC study (i.e. so that $\bar{\boldsymbol{x}}_{AC} = \mathbf{0}$), MAIC minimises the objective function

$$H_{\text{MAIC}}(\boldsymbol{\alpha}) = \sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} \exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right). \tag{2.26a}$$

The (normalised) weights w_{ik} are then given by

$$w_{ik} = \frac{\exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right)}{\sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} \exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right)}. \tag{2.26b}$$

(We use the normalised weights here to better show the equivalence to entropy balancing; a set of weights can be rescaled arbitrarily without affecting the estimate (2.20).) Entropy balancing also seeks weights that match the moments of covariates between studies, but that further minimise the entropy distance from uniform weights, $\sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} w_{ik} \log(N_{(AB)} w_{ik})$. Hainmueller (2012)

uses Lagrange multipliers to find an unconstrained dual optimisation problem, which (again after setting $\bar{x}_{AC} = \mathbf{0}$) gives the objective function

$$H_{\text{EB}}(\boldsymbol{\alpha}) = \log\left(\frac{1}{N_{(AB)}} \sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} \exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right)\right). \tag{2.27a}$$

Again, the weights are given by

$$w_{ik} = \frac{\exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right)}{\sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} \exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right)}. \tag{2.27b}$$

Comparing the objective functions (2.26) and (2.27), we see that

$$H_{\text{EB}}(\boldsymbol{\alpha}) = \log(H_{\text{MAIC}}(\boldsymbol{\alpha})) - \log(N_{(AB)}). \tag{2.28}$$

Therefore, since the logarithm is a monotonic function and $\log(N_{(AB)})$ is constant, the solutions of these two minimisation problems are identical.

As a result, we have shown that the MAIC weights (being identical to entropy balancing weights) have the additional desirable property that they are as close as possible to uniform weights (no weighting at all), in an entropy sense. Entropy balancing performs the minimisation on the log scale which may perform better computationally, but the estimated weights will be identical for MAIC and entropy balancing, up to optimisation error. Whilst MAIC and entropy balancing are mathematically identical under this formulation, alternative loss functions could be used in the entropy balancing scheme which may change the performance of the method. Hainmueller (2012) also notes that other "base weights" to minimise the distance from could be used instead of uniform weights, and this would affect the equivalence to standard MAIC. For example, other base weights could be used to perform non-parametric covariate adjustment (Williamson et al. 2013), or to adjust for treatment crossover (Robins and Finkelstein 2000), prior to reweighting to match the AC population. With non-uniform base weights $w_{ik}^{(0)}$, the entropy balancing objective function in (2.27) becomes

$$H_{\text{EB}}(\boldsymbol{\alpha}) = \log\left(w_{ik}^{(0)} \sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} \exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right)\right), \tag{2.29a}$$

and the weights are then given by

$$w_{ik} = \frac{w_{ik}^{(0)} \exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right)}{\sum_{k=A,B} \sum_{i=1}^{N_{k(AB)}} w_{ik}^{(0)} \exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{\alpha}\right)}. \tag{2.29b}$$

Setting uniform base weights $w_{ik}^{(0)} = 1/N_{(AB)}$ in (2.29) recovers the formulae in (2.27) above.

Different schemes for applying weights have also been proposed. MAIC as described in Section 2.2.1 estimates weights for the entire AB population at once to balance covariate distributions with the entire AC population. Belger et al. (2015a,b) compare anchored and unanchored MAIC with other possible approaches, which involve splitting apart trial arms and balancing covariate distributions separately between the control arms (A) and between the treatment arms (B and C) in the IPD and aggregate populations. The properties of such "splitting" approaches in comparison with a more typical population reweighting are largely unknown and require further investigation; however, some initial simulation studies have reported performance benefits over standard MAIC (Petto et al. 2019).

2.3 Assumptions and properties of MAIC and STC in anchored and unanchored comparisons

We now examine in detail the assumptions made by MAIC and STC which are required to achieve a valid indirect comparison in the target population. If these assumptions are violated, the resulting estimate may be biased. It is critical to observe that the necessary assumptions differ between the anchored and unanchored forms of indirect comparison (equations (2.1) and (2.2) respectively), with the unanchored indirect comparison requiring stronger assumptions. We do not discuss the first three core assumptions specified in the generalisation literature (homogeneity of effects, stable unit treatment value, and ignorable treatment assignment), as they must generally be assumed to hold for any form of indirect comparison or meta-analysis.

2.3.1 Anchored comparisons

The MAIC and STC literature typically advocates performing indirect comparisons directly on the outcome scale, with $g(\cdot)$ the identity function in equation (2.1) for an anchored comparison, so that

$$\hat{d}_{BC(AC)} = \bar{y}_{C(AC)} - \bar{y}_{A(AC)} - \left(\hat{y}_{B(AC)} - \hat{y}_{A(AC)}\right) \tag{2.30}$$

2.3.1.1 MAIC, and STC with an identity link

When making an anchored indirect comparison in the AC population on the outcome scale as in equation (2.30), both MAIC and STC (using a linear

outcome model with identity link) rely on an assumption of conditional constancy of relative effects on the outcome scale—that the differences in the relative effects that would be observed between studies are entirely accounted for by an imbalance in the effect modifier variables $\boldsymbol{x}^{\text{EM}}$ (see Section 2.1.6). The implication of this assumption is that $\boldsymbol{x}^{\text{EM}}$ must contain every effect modifier that is in imbalance between the two studies, otherwise the indirect comparison is still biased. Note that both effect modifiers and conditional constancy of relative effects here are defined on the outcome scale due to the indirect comparison being made on this scale.

STC requires the correct specification of the form of the outcome model in order to provide unbiased estimates. When an anchored comparison is made, an unbiased estimate is still obtained even if some or all prognostic variables (that are not also effect modifiers) are omitted from or misspecified in the model (and an intercept term is included). However, inclusion of prognostic variables in the outcome model should in theory lead to more precise estimation of the treatment effect and effect modifier parameters within the model and the resulting indirect comparison, as a portion of the variability is accounted for by the prognostic variables.

In the present MAIC literature (Ishak et al. 2015; Signorovitch et al. 2012, 2010), there is no discussion of which variables (prognostic and/or effect modifying) should be included in the weighting model; the prevailing choice in applications of MAIC to date appears to be to include as many variables as possible, regardless of effect modifier status or level of imbalance (see Chapter 3). However, the choice of variables to be matched/weighted on should be carefully considered: including too many variables will reduce the effective sample size, negatively affecting the precision of the estimate; conversely, failure to include relevant variables will result in a biased estimate. Therefore, for an anchored indirect comparison, the weighting model must include all effect modifiers (both those in balance and imbalance between the studies), but no prognostic variables. Including effect modifiers that are already balanced in the weighting model ensures that they remain balanced after the weighting, and there will be negligible impact on the standard error due to their inclusion. Imbalances in prognostic variables are taken care of by the randomisation within studies (and the subsequent "adjustment" to the comparison with the control arms), and their inclusion in the matching model only reduces the effective sample size.

STC with a non-identity link 2.3.1.2

In the case that STC is carried out with a non-identity link function, there arises a conflict of scale when equation (2.30) is used to form an indirect comparison on the natural outcome scale: the outcome model defines a specific transformed linear predictor scale, upon which additivity is assumed and effect modifiers and prognostic variables are defined, whereas the indirect comparison is formed on the natural outcome scale. Effect modifier status is mathematically demonstrable to be scale-specific (e.g. Brumback and Berg 2008), and the status of a variable as an effect modifier on one scale does not imply (either positively or negatively) the effect modifier status on any other scale. Therefore, performing the indirect comparison on one scale whilst fitting the outcome model on another raises questions about the interpretation of the model and of the indirect comparison.

The advantage of an anchored indirect comparison over an unanchored indirect comparison is also in doubt in this case, as the aim of cancelling out prognostic variables on the outcome scale in the anchored indirect comparison is in contradiction with their definition on the linear predictor scale in the outcome model. It is unclear at present whether the anchored comparison leads to a reduction in bias and reliance on model specification or an increase, compared to the unanchored comparison. However, it is clear that, as prognostic variables (defined on the linear predictor scale) will not cancel in the anchored indirect comparison (defined on the outcome scale), any misspecification or omission of prognostic variables in the outcome model will lead to a biased estimate. Therefore, an indirect comparison made using STC with a non-identity link makes the assumption that $\boldsymbol{x}$ contains both all effect modifiers and all prognostic variables (i.e. conditional constancy of absolute effects) with respect to the linear predictor scale, and that the outcome model is correctly specified.

Performing the indirect comparison on the transformed linear predictor scale as in equation (2.1) (instead of the outcome scale) would eliminate these concerns, and once again lead to reliance upon the weaker assumption of conditional constancy of relative effects. This is the usual method employed in standard indirect comparisons (Bucher et al. 1997; Dias et al. 2013a). We discuss the choice of scale further in Section 2.3.3.

Unanchored comparisons 2.3.2

Regulators are, increasingly, approving new products on the basis of single-arm studies, especially in oncology (50% of all FDA accelerated oncology approvals

in 2015 were based on single-arm trials; FDA 2016), and reimbursement authorities are increasingly asked to assess treatments where only single-arm studies or disconnected networks are available. In this case unanchored MAIC or STC can be used to improve on "unadjusted" or naïve indirect comparisons by taking into account the different distributions of prognostic factors and effect modifiers in the two studies. (In the same way that MAIC and STC may improve upon standard "adjusted" indirect comparison by taking account of the distribution of effect modifiers.) However, it is essential that decision makers understand the different sources of error that attach to standard ("adjusted") indirect comparisons, naïve "unadjusted" indirect comparisons, and MAIC/STC in their anchored and unanchored forms.

If an unanchored comparison is made (2.2), whether on the outcome scale or transformed scale, then both MAIC and STC rely on the conditional constancy of absolute effects assumption; the differences between absolute outcomes that would be observed in each trial are entirely explained by imbalances in prognostic variables and effect modifiers $\boldsymbol{x}$ with respect to the chosen scale. Under this assumption, $\boldsymbol{x}$ must contain both every prognostic variable and every effect modifier that is in imbalance between the two studies—an assumption that is largely deemed unreasonable (if it were, there would be no reason to undertake randomised controlled trials). Conditional constancy of absolute effects may be partially assessed in a connected scenario through the use of placebo tests using the common comparator (see Section 2.1.5). If the conditional constancy of absolute effects assumption fails then the unanchored estimator is invalid and an anchored estimator making use of the conditional constancy of relative effects assumption should be used. However, such tests cannot be used to justify an unanchored comparison for two reasons: (i) lack of statistical power; and (ii) conditional constancy of absolute effects is only partially assessed if the common comparator is placebo, as residual imbalances in observed or unobserved effect modifiers cannot be evaluated. It should also be noted that, whilst the traditional approach is to adjust for all available variables, these may nevertheless be limited (especially in the published aggregate data), and therefore such an approach alone is not sufficient justification for the conditional constancy of absolute effects assumption.

STC furthermore assumes that the outcome model is correctly specified in both prognostic variables and effect modifiers; it is thus more burdensome to specify an outcome model for an unanchored comparison than for an anchored comparison, as the prognostic variables and their model specification become

critical in the unanchored case. The impact of performing an unanchored indirect comparison on a different scale to that of the linear predictor is currently unknown, although the concerns over interpretability raised for the anchored case in Section 2.3.1.2 still stand.

If a MAIC is to be performed, the weighting model must include every effect modifier and prognostic variable—compared to the anchored case, where only effect modifiers are required. An immediate consequence of this is that an unanchored indirect comparison performed using MAIC will always have less precision than an anchored indirect comparison using MAIC in the presence of an imbalance of prognostic variables, and—more importantly—is more likely to be biased given that all prognostic variables in imbalance must be included in the weighting model as well as effect modifiers.

Choice of scale for indirect comparisons 2.3.3

The standard practice for indirect comparison and network meta-analysis is that they are made on a pre-specified transformed linear predictor scale (e.g. on the logit scale for proportions or the log scale for rate outcomes), rather than on the natural outcome scale (Bucher et al. 1997; Dias et al. 2013a); for the purposes of a CEA, the resulting estimates may be back-transformed onto the (possibly more interpretable) natural scale. The reasons for this choice include approximate normality and the stabilisation of variance, however the most critical reason with regards to indirect comparisons is that effects are assumed to be linear and additive on the transformed scale. Therefore the apparently pervasive choice amongst present applications of MAIC and STC (see Section 3.1) to perform comparisons directly on the natural outcome scale in the face of a more usual transformed scale is disconcerting, and somewhat a contradiction of assumptions. Furthermore, as effect modification is defined with respect to the scale of the comparison, variables that are effect modifiers in standard indirect comparison might not be in MAIC/STC, and variables which are effect modifiers in MAIC/STC may not be effect modifiers in a standard indirect comparison analysis.

This is made most clear by STC when an outcome model is (quite correctly) specified with a non-identity link function (see Section 2.3.1.2): the outcome model defines effects linearly and additively on the transformed linear predictor scale, which is in direct contradiction with the subsequent assumption of linearity and additivity on the outcome scale used by the indirect comparison. Furthermore, the definition and interpretation of effect modifiers and prognostic variables is entirely scale-specific, and results in

conflicts and contradictions when the outcome model and indirect comparison are on differing scales. We cannot be certain of the impact of such conflicts of scale without comprehensive simulation studies.

Although the identification of the "correct" scale for any specific outcome is debatable, there is a considerable literature (e.g. Deeks 2002) that shows that relative treatment effects for binary or rate outcomes are more stable across trials when they are expressed on logit or log scales, compared to absolute scales such as the risk difference, meaning there are fewer effect modifiers or that effect modification is weaker. Another concern of scale choice in the context of indirect comparisons is that different scales can lead to reverse conclusions, particularly for binary and rate outcomes when baseline event rates are diverse (Norton et al. 2012). This reversal is due to the additivity assumption not being valid on all scales (indeed, it is impossible for additivity to hold on all scales; van Valkenhoef and Ades 2013). The choice of an appropriate scale is therefore critical, and should be made using biological and clinical knowledge (Caldwell et al. 2012); moreover, where a standard scale exists for a given outcome upon which additivity is commonly accepted, the use of an alternative scale is hard to justify.

In a decision-making context, the possibility of effect modification has to be handled thoughtfully. The NICE Guide to the Methods of Technology Appraisal (NICE 2013) is explicit that effect modifiers must be pre-specified and clinically plausible, and that supporting evidence must be provided from a thorough review of the subject area or from expert clinical opinion (see Section 5.2.7 of the NICE Methods Guide). Moreover, although in the present context controlling for effect modifiers is undertaken to generate less biased population-average relative effects, the existence of an effect modifier can change the nature of the decision problem: for example if age is considered to be an effect modifier, it raises the possibility that a treatment that is effective at one age might not be effective at another.

A potential and oft-cited advantage of MAIC is that it is perceived to be "scale-free", in the sense that the definition of the weighting model does not require any fixed outcome scale to be chosen (Ishak et al. 2015; Signorovitch et al. 2010). We however express caution at this notion: it is true that no outcome model need be assumed to create the weighting model, but the subsequent indirect comparison does assume additivity on a specific scale, and therefore neither MAIC nor STC are "scale-free" in this important sense.

Impact of having access to only marginal covariate information 2.3.4

Thus far we have considered MAIC and STC in the scenario where, despite not having access to IPD on the AC trial, sufficient information on the joint covariate distribution is available. In practice even this level of detail is unlikely, as published trials frequently report only details of the marginal covariate distributions (e.g. mean/median and standard deviation for continuous covariates, or proportion of individuals with a binary/categorical trait). This leads to an additional assumption being required for both MAIC and STC: either that (i) the joint distribution of covariates in the AC trial is the product of the (published) marginal distributions, or (ii) the correlations between covariates in the AC trial may be imputed using those observed in the AB trial.

This assumption is most explicit when STC is used. Ishak et al. (2015) propose that, in order to create predictions into the AC population, missing correlations between covariates in the AC population are assumed to be the same as those observed in the AB population.

MAIC does not explicitly specify any form of outcome model, however there is an implicit outcome model which is inferred when the indirect comparison is formed. Specifically, effects are assumed to be additive on the scale of the indirect comparison, as are the actions of effect modifiers and prognostic variables. When covariate correlations are not available from the AC population (and therefore cannot be balanced by inclusion in the weighting model), they are assumed to be equal to the correlations amongst covariates in the pseudo-population formed by weighting the AB population.

However, if an anchored indirect comparison is made (from either MAIC or STC), then, due to the cancellation of prognostic variables, only correlations amongst effect modifiers will affect the indirect comparison, and the assumption of identical correlations amongst prognostic variables between the two trial populations can be dropped. Furthermore, if there are no multi-way treatment by effect modifier interactions in the (for MAIC, implicit) outcome model (or any interactions at all, for an unanchored comparison), then the estimated indirect comparison will remain unbiased even if the correlations between covariates differ between the two trial populations.

Choice of target population 2.3.5

The premise of both MAIC and STC is that the treatment effect depends on the population. It is therefore not sufficient to use MAIC or STC to generate an "unbiased" comparison in just any population; they only achieve this purpose if they can produce a fair comparison in the target population for the decision.

In general, the target population should be a UK cohort or registry study population relevant to the clinical decision, which is unlikely to match the population of the AC trial. However, MAIC and STC as currently proposed are unable to achieve estimates in any population other than that of the AC study. We present an extension in Section 2.5 which enables indirect comparisons to be made in any target population, given an additional assumption.

The population-specific nature of MAIC and STC analyses can lead to apparently contradictory conclusions being drawn from the same pair of trials, simply by taking the alternate company's perspective and swapping the roles of the AB and AC studies, having instead IPD on the AC trial and aggregate data on the AB trial. This problem has already arisen in analyses from competing companies: Novartis and AbbVie presented MAIC analyses of the same two trials comparing secukinumab and adalimumab to placebo as treatments for ankylosing spondylitis (Betts et al. 2016; Maksymowych et al. 2016). Each company had IPD on their own trial, but not on their competitor's trial. The results from each company's MAIC appear to be in conflict, with one company claiming significant differences in efficacy in favour of secukinumab, and the other claiming comparable efficacy but improvements in cost effectiveness for adalimumab. Importantly we note that, as MAIC (and STC) attempts to produce estimates in the AC population, the two MAIC analyses are aiming to provide estimates in two different target populations—the population of the competitor's trial in each case. Furthermore, the Novartis trial population included both treatment experienced and treatment naïve patients, whereas the AbbVie trial population included only treatment naïve patients. Due to the lack of population overlap concerning treatment experienced patients, it is impossible for a MAIC from AbbVie's perspective to generate estimates for the full Novartis trial population. However, even if both trial populations overlapped perfectly, we would still expect there to be differing estimates depending on which company's perspective is taken—precisely because the two study populations have been deemed incomparable directly due to an imbalance in effect modifiers; if there were no such imbalance, then there would be no need to conduct an anchored indirect comparison instead of the usual indirect comparison. The real conflict, therefore, lies not in the results produced by the two MAICs, but in deciding which of the two study populations better represents the true target population. Ironically, each company is left in the position of implicitly assuming that their competitor's trial is more representative than their own.

This prospect of conflicting estimates from different companies becomes

exponentially worse as MAIC/STC is extended to multiple trials and multiple treatments. For example in a star-like structure of AB, AC, AD, AE studies, if each company performed a MAIC/STC using IPD available on their own trial, and effect modification was present, they would generate among them four incoherent sets of three pair-wise indirect comparisons, none of which could be compared to each other.

Sampling variation in the target population 2.3.6

MAIC and STC, as currently portrayed, produce estimates of mean outcomes on each treatment in the AC study sample, rather than in the AC population. In other words, the sampling uncertainty of the AC trial sample is ignored.

There is substantial literature on super-population average treatment effects (SPATE), which addresses precisely this issue (for an introduction, see Imbens and Rubin 2015, chapter 6). In the context of our calibration scenario, the AB and AC trials are seen as samples from a larger super-population (the true target population), and the estimates in the AC trial can be turned into estimates in the target population by accounting for the additional sampling variation. A notable special case occurs when the inclusion/exclusion criteria of the AC trial match exactly the true target population and the individuals enrolled in the AC trial are randomly sampled from the true target population; then the point estimates provided by MAIC or STC in the sample population are exactly carried over to the true target population, with an increase in standard error reflecting the sampling uncertainty.

Uncertainty propagation 2.4

We break down the uncertainty in the estimates resulting from MAIC and STC into three sources: sampling variation within the studies, uncertainty due to the imbalance in covariate distributions, and uncertainty due to estimation of the weighting/outcome model. (We do not consider additional methodological or structural uncertainty here, such as that arising from the choice of covariates to adjust for, and these are not reflected in the uncertainty of the estimates.) Both MAIC and STC fully account for the sampling variation within the studies, and propagate this through to the final estimate.

MAIC inherently accounts for the uncertainty due to the imbalance in covariate distributions: greater differences between the covariate distributions lead to an increase in the variation of weights (some become larger, some become smaller) and hence a reduction in effective sample size. Standard errors

for MAIC estimates are typically obtained using robust sandwich estimators (Signorovitch et al. 2010), which account for the fact that the weights are estimated rather than fixed and known. Alternative methods for incorporating all sources of uncertainty in MAIC include bootstrapping techniques (Efron 1979).

Whether or not STC takes into account the latter two sources of variation depends upon how the predicted outcomes into the AC study are treated. If the predicted outcomes are treated as fixed and known (as if they had actually been observed), then the estimates resulting from STC will not take into account either the uncertainty due to covariate imbalance (which may lead to extrapolation if there is insufficient overlap between the two populations), or due to the estimation of the outcome model parameters. However, if the predicted outcomes are correctly considered along with their associated prediction error, then the resulting estimates will account for all three sources of variation.

2.5 Calibrating population-adjusted estimates to the correct target population

In Section 2.3.5 it was pointed out that MAIC and STC as presently used, although based on the idea that the size of a relative treatment effect depends on the population, do not in general succeed in generating comparisons calibrated to the target population for the decision (unless the target population matches the AC trial population, which is unlikely). We propose that an additional assumption is made, which we call the *shared effect modifier assumption*, which will allow relative treatment effects to be projected into any population. One of the results of this assumption is that active-active treatment comparisons (e.g. B vs. C) may be transported into any target population, as any effect modifiers cancel out; indeed, the shared effect modifier assumption is required in order for this to be possible.

2.5.1 Shared effect modifier assumption

The shared effect modifier assumption applies to a set of active treatments $\mathcal{T}$, and states that (i) the effect modifiers of all treatments in $\mathcal{T}$ are the same, and (ii) the change in treatment effect caused by each effect modifier is the same for all treatments in $\mathcal{T}$.

This assumption is not required for MAIC or STC as currently used. However, if this assumption is deemed reasonable, then it may be leveraged

to produce indirect comparisons in any given target population (see below). The shared effect modifier assumption is evaluated on a clinical and biological basis; treatments in the same class (i.e. sharing biological properties or mode of action) are likely to satisfy the shared effect modifier assumption, and those from different classes are not. In some circumstances, where effect modification is an artefact of the scale of measurement (possibly indicating a poor choice of scale), it will be valid for all active treatments. This assumption is, in fact, commonly made when meta-regression is used (as noted by Dias et al. 2011a). One of the reasons for assuming that treatments in the same class have the same effect modifiers, in the absence of overwhelming evidence to the contrary, is that relaxing this assumption could lead to seemingly perverse decisions. For example, it is not uncommon to switch from recommending no treatment to recommending a given treatment past a certain age, but it would be most unusual to switch among several treatments within the same class at various ages (say treatment B is most effective at age 50, treatment C at age 60, and treatment D at age 70, and so on). In the present "anchored" scenario, it is common that A is placebo or a standard treatment, and we might make the shared effect modifier assumption for the set of treatments $\mathcal{T} = \{B, C\}$.

Using the shared effect modifier assumption 2.5.2

The shared effect modifier assumption allows us to transpose indirect comparisons from any population where a relative effect has been observed, such as an AC trial, to any other population of interest P, and recreate a full set of relative or absolute effects given an observed relative or absolute effect in the P population. In general, we make use of the following relation concerning the marginal relative effects for a set of treatments $\mathcal{T}$ for which the shared effect modifier assumption holds:

$$d_{ab(P)} = d_{ab} \quad \forall a, b \in \mathcal{T} \text{ and } \forall P. \tag{2.31}$$

That is, the b vs. a relative effects are constant across populations for any two active treatments a and b in $\mathcal{T}$.

Proof of shared effect modifier relationship 2.5.3

To see that the relation in (2.31) holds, assume additivity on an appropriate linear predictor scale and write the transformed conditional absolute treatment effects $\eta_k(\boldsymbol{x}, \boldsymbol{u})$ as

$$\eta_k(\boldsymbol{x}, \boldsymbol{u}) = \beta_0 + \boldsymbol{\beta}_1^{\intercal}\boldsymbol{x} + \boldsymbol{\phi}_1^{\intercal}\boldsymbol{u} + \left(\boldsymbol{\beta}_{2,k}^{\intercal}\boldsymbol{x}^{\text{EM}} + \boldsymbol{\phi}_{2,k}^{\intercal}\boldsymbol{u}^{\text{EM}} + \gamma_k\right)\mathbb{I}(k \neq A), \tag{2.32}$$

where $\boldsymbol{x}$ and $\boldsymbol{u}$ are vectors of observed and unobserved covariates respectively (possibly including interactions or higher order terms), with corresponding subvectors of effect modifiers $\boldsymbol{x}^{\text{EM}}$ and $\boldsymbol{u}^{\text{EM}}$. Equation (2.32) represents the underlying (transformed) outcome model, which cannot be estimated directly as $\boldsymbol{u}$ are unobserved.

Using the shared effect modifier assumption on the set of treatments $\mathcal{T}$, which means that $\boldsymbol{\beta}_{2,k} = \boldsymbol{\beta}_2$ and $\boldsymbol{\phi}_{2,k} = \boldsymbol{\phi}_2$ $\forall k \in \mathcal{T}$, we rewrite the outcome model (2.32) for $k \in \mathcal{T}$ as

$$\eta_k(\boldsymbol{x}, \boldsymbol{u}) = \beta_0 + \boldsymbol{\beta}_1^{\intercal}\boldsymbol{x} + \boldsymbol{\phi}_1^{\intercal}\boldsymbol{u} + \left(\boldsymbol{\beta}_2^{\intercal}\boldsymbol{x}^{\text{EM}} + \boldsymbol{\phi}_2^{\intercal}\boldsymbol{u}^{\text{EM}} + \gamma_k\right) \mathbb{I}(k \neq A). \tag{2.33}$$

We are now ready to proceed in proving (2.31).

Notice that, for any two treatments $a, b \in \mathcal{T}$, we can write the marginal relative effects in a population P in terms of the conditional absolute effects by using equation (2.33) and taking expectation over the population P :

$$\begin{aligned} d_{ab(P)} &= \mathbb{E}_{(P)}\big(\eta_b(\boldsymbol{x}, \boldsymbol{u}) - \eta_a(\boldsymbol{x}, \boldsymbol{u})\big) \\ &= \gamma_b - \gamma_a + \mathbb{E}_{(P)}\left(\boldsymbol{\beta}_2^{\intercal}\boldsymbol{x}^{\text{EM}} + \boldsymbol{\phi}_2^{\intercal}\boldsymbol{u}^{\text{EM}}\right) - \mathbb{E}_{(P)}\big(\boldsymbol{\beta}_2^{\intercal}\boldsymbol{x}^{\text{EM}} + \boldsymbol{\phi}_2^{\intercal}\boldsymbol{u}^{\text{EM}}\big) \\ &= \gamma_b - \gamma_a. \end{aligned}$$

The right-hand side of this equation does not depend on the population P. Therefore $d_{ab(P)}$ is constant across populations for all $a, b \in \mathcal{T}$.

2.5.4 Example of applying the shared effect modifier assumption

Hence, if all relative effects are known in one population (say, the AC population) and for another population (say P) we are given an estimate of any single relative effect $d_{Ak(P)}$, where k is in $\mathcal{T}$, then immediately we can calculate estimates of all other relative effects $d_{Ab(P)}$, where b is in $\mathcal{T}$, in the new population via equation (2.31). Similarly, if we are given an estimate of a single absolute effect $\theta_{k(P)}$, where k is in $\mathcal{T}$, in the P population, then we can calculate estimates of all absolute effects $\theta_{b(P)}$ and relative effects $d_{ab(P)}$ for all a, b in $\mathcal{T}$ via equation (2.31).

For example, suppose the log odds ratios in the AC population have been estimated to be

$$\hat{d}_{AB(AC)} = 1.3, \qquad \hat{d}_{AC(AC)} = 0.8.$$

Furthermore, in a population P the log odds ratio for treatment B compared to A is estimated to be $\hat{d}_{AB(P)} = 0.7$. We make the shared effect modifier assumption for treatments $\mathcal{T} = \{B, C\}$. From the AC trial we have that $\hat{d}_{BC(AC)} = \hat{d}_{AC(AC)} - \hat{d}_{AB(AC)} = -0.5$. Now using (2.31), we have that $\hat{d}_{BC(AC)} =$

$\hat{d}_{BC(P)}$, and the log odds ratio for treatment C compared to A is inferred to be $\hat{d}_{AC(P)} = \hat{d}_{AB(P)} + \hat{d}_{BC(P)} = 0.2$.

In practical terms for our two-study scenario, if the shared effect modifier assumption holds for treatments B and C, then the estimated d_{BC} marginal relative treatment effect (whether obtained using anchored or unanchored MAIC/STC) will be applicable to *any* population. Notice also that if $\{A, B\} \subseteq \mathcal{T}$ then $\hat{d}_{AB(AB)}$ is valid for any population and no population adjustment is necessary for this comparison, and similarly if $\{A, C\} \subseteq \mathcal{T}$ then $\hat{d}_{AC(AC)}$ is valid for any population.

Summary 2.6

Population adjustment methods such as MAIC and STC are based upon well-known methods dating back several decades. In this chapter we began with a review of the earlier literature surrounding the population adjustment scenario and other closely-related problems.

We started by considering a related problem known as standardisation (Section 2.1.1), in which outcomes are to be predicted in (or standardised to) a target population based on a sample population which has a different covariate distribution. Two standardisation methods were discussed, namely propensity score weighting (of which there are several variants; Section 2.1.2), and outcome regression (Section 2.1.3); these methods are common across the broader literature and within various problem settings, and may also be combined into doubly robust methods (Section 2.1.4). Propensity score methods seek to apply weights (the inverse propensity score) to the sample population in order that its covariate distribution matches that of the target population; the propensity scores are usually found using logistic regression for inclusion into the target population based on the set of covariates. Outcome regression fits a model to the outcomes on a suitable scale in the sample population based on the set of covariates, and uses this model to predict the outcomes that would be seen in the target population.

We then discussed more recent literature concerning another related problem known as generalisation (Section 2.1.5), where relative treatment effects observed in a sample population are to be generalised to a target population. The techniques involved in such scenarios are broadly similar to those described above in the standardisation literature. An important contribution of the generalisation literature is the discussion of the assumptions required for valid inference in the target population, along with methods to

(at least partially) test these assumptions (placebo tests), and an exposition of the biases involved in generalising treatment effects across populations.

Finally, we explored the calibration literature (Section 2.1.6), which seeks to perform indirect comparisons between treatments studied in two different populations, relaxing the assumptions typically made in the generalisation literature. Treatment-specific conditional constancy (that is, conditional constancy of absolute effects) assumes that, on a suitable scale, treatment effects are constant across populations given a set of covariates. In practice this is highly implausible, as the set of covariates is required to contain every prognostic variable and effect modifier that is in imbalance between the two populations; arguably if all of these were known then there would be no need for RCTs. Calibration methods instead rely upon a less stringent assumption of conditional constancy of relative effects, assuming that (again on a suitable scale) relative treatment effects are constant across populations given a set of effect modifiers, allowing the randomisation within studies to cancel out the effects of any prognostic variables. The calibration literature again makes use of the standard propensity score and outcome regression ideas. The key difference between calibration methods and population adjustment method such as MAIC and STC is that the standard calibration methods require the availability of IPD in both studies. This inhibits their immediate application in our specific problem setting where IPD is available only on the manufacturer's AB trial, although the assumptions, advantages, and disadvantages of the different calibration methods carry over to MAIC and STC, since they share the same basis in reweighting or regression adjustment.

In light of the surrounding literature on standardisation, generalisation, and calibration, we then proceeded to describe the current methods for population adjustment (Section 2.2) and to set out their properties and assumptions (Section 2.3). All methods for indirect comparison—population-adjusted or otherwise—require some form of constancy assumption with respect to the chosen comparison scale, as described earlier by the calibration literature. The strength of this assumption depends on whether an anchored comparison (comparing relative effects via a common comparator treatment) or an unanchored comparison (comparing absolute effects with no commmon comparator) is made. A standard indirect comparison or network meta-analysis assumes constancy of relative effects, so that there are no imbalances any effect modifiers between study populations (for random effects NMA, this is relaxed to no imbalance on average). Anchored forms of population-adjusted indirect comparisons rely on conditional constancy of relative effects. This means

that the relative treatment effects are assumed constant between studies at any given level of the effect modifiers. No assumptions are needed regarding between-study differences in the distribution of prognostic variables, because these are accounted for by randomisation. Unanchored population-adjusted indirect comparisons make the much stronger assumption of conditional constancy of absolute effects (called treatment-specific conditional constancy by Zhang et al. (2015) in the calibration literature). This means that the absolute treatment effects are assumed constant at any given level of the effect modifiers and prognostic variables, and all effect modifiers and prognostic variables are required to be known. This is a far more demanding assumption, and it is widely accepted that it is very hard to meet or even verify. Unanchored comparisons based on disconnected networks and/or involving single-arm studies are therefore problematic. The required constancy assumption for each population adjustment method are summarised in Table 2.1.

In Section 2.5, we described how the shared effect modifier assumption could be used to produce estimates for a target population other than that of the AgD AC study. This is especially important in a decision making context since estimates are only useful if they can be produced for the relevant decision population, which is unlikely to be represented by the AC study. MAIC and STC do not require the shared EM assumption to produce estimates for the AC population, but do require the shared EM assumption to generalise estimates to other populations. Current network meta-regression based approaches do require the shared EM assumption in order to identify the model in a scenario with one IPD and one AgD study. However, it is possible to relax the shared EM assumption if IPD are available from both studies, or if enough AgD AC studies are available. Later in Section 4.6.1, we describe more generally how the shared EM assumption may be relaxed and assessed under our new method (also based on network meta-regression). Table 2.1 also summarises the use of the shared EM assumption for each population adjustment method.

In Section 2.2.3 we discussed a further class of methods based upon network meta-regression, with regression models defined at both the individual and aggregate levels. Of particular interest is a method derived from the hierarchical related regression introduced by Jackson et al. (2006, 2008), where the aggregate-level model is an integration of the individual-level model over the aggregate study population, although at present these models have only been derived for the special case of binary covariates (Jansen 2012). This method requires the same assumptions as MAIC and STC, namely that all effect modifiers in imbalance are accounted for (conditional constancy of relative

effects), and (in the two-study scenario) the shared effect modifier assumption. This approach differs conceptually from MAIC and STC, in that it models individual-level relationships and is able to provide internally consistent inferences at both the individual level and at an aggregate level like a standard indirect comparison. Methods such as MAIC and STC use IPD to predict average outcomes on study arms, and then produce an indirect comparison at the aggregate study level. We regard this as a promising approach with some attractive properties. Most importantly: (i) it reduces to the gold-standard IPD network meta-regression if IPD are available for all trials, and to standard AgD NMA if no adjustment is required; (ii) it generalises naturally to connected networks of any size, unlike MAIC and STC; and (iii) aggregation bias is avoided by correctly relating the individual and aggregate levels of the model through integration, unlike simpler meta-regression approaches that "plug in" mean covariate values (Berlin et al. 2002; Rothman et al. 2008). Later in Chapter 4, we build upon this approach to develop a general framework for population-adjusted indirect comparisons and network meta-regression combining IPD and AgD, which is then compared with MAIC and STC through an extensive simulation study in Chapter 8.

Table 2.1 Key assumptions required by different methods for indirect comparisons.

Assumption	Method			
	Standard indirect comparison and NMA	Network meta-regression*	Anchored MAIC and STC	Unanchored MAIC and STC
Constancy	N	N	N	N
Constancy of absolute effects				
Conditional constancy of absolute effects	N	N	N	Y
Constancy of relative effects	Y For RE NMA relaxed to constancy in expectation	N	N	N
Conditional constancy of relative effects	N	Y	Y	N
Shared effect modifiers	N/A	Y Not required if IPD are available in both studies	N†	N†

* The assumptions set out here are applicable to all forms of network meta-regression with varying combinations of IPD and AgD (both studies IPD, both studies AgD, one IPD and one AgD), with the exception of the shared effect modifier assumption which is not required if IPD are available on both studies and may also potentially be relaxed in larger networks.

† The shared effect modifier assumption is not required, but may be additionally assumed in order to present estimates for another target population.

Chapter 3

Review of applications of population adjustment methods

In this chapter, we undertake two reviews of the applications of population adjustment methods. As well as investigating the uptake of population adjustment methods, we are interested in the ways in which these methods are used and whether the key assumptions are likely to hold in order to assess the adequacy of current practice, particularly for decision making. The first (Section 3.1) is a review of applications of MAIC and STC in the published literature, which has been published as part of NICE Decision Support Unit Technical Support Document 18 (Phillippo et al. 2016). We focus on MAIC and STC in this review, as they are the most commonly used population adjustment methods at present, although other approaches based on network meta-regression have also been proposed (see Section 2.2.3). This review was carried out at the beginning of the PhD project. Later on in the PhD project we wished to perform an updated review; however, to keep this manageable we focused on applications of population adjustment methods in NICE Technology Appraisals (TAs). Thus, in the second review (Section 3.2), we review all NICE TAs published since 2010—when MAIC and STC were first suggested in the literature (Caro and Ishak 2010; Signorovitch et al. 2010)—for uses of any population adjustment methods. This review has been published as Phillippo et al. (2019a). We conclude with a discussion and suggest several key improvements to current practice, towards providing better evidence for decision makers and greater impact for those performing such analyses.

3.1 Published applications of MAIC and STC in the literature

In this section we review the published applications of MAIC and STC in the literature, to examine how these new methods are being used in practice, and how well the methodology and assumptions underlying them are understood. Applied papers were found using a simple search amongst titles, abstracts, and keywords for "matching-adjusted indirect comparison" and "simulated treatment comparison" in Scopus and PubMed on 07/07/2016, by checking citing articles of the methodological papers (Caro and Ishak 2010; Ishak et al. 2015; Signorovitch et al. 2010), and examining papers identified in a published scoping review (Veroniki et al. 2016).

3.1.1 Applications of MAIC in the literature

In the short time since the first papers on MAIC (Signorovitch et al. 2010) and STC (Caro and Ishak 2010) were published, the use of these methods—particularly MAIC—has increased dramatically. In Table 3.1 we list the ten published applications of MAIC that our search identified in the literature up until 07/07/2016, along with particular features and properties of the analyses, which we now discuss.

3.1.1.1 Anchored and unanchored comparisons

The majority (60%) of the analyses involved randomised controlled trials with a common comparator. Of these, four out of six performed anchored indirect comparisons. Three out of six analyses involved an unanchored indirect comparison (one performed both anchored and unanchored indirect comparisons on different outcomes). In two of these, the unanchored approach was due to the outcome of interest being overall survival (OS) in a trial subject to treatment switches, where the placebo arm is contaminated by individuals crossing-over to active treatment after disease progression. Focusing on progression free survival (PFS) rather than OS as the primary outcome avoids this issue, although estimates of OS are typically required for economic modelling. One analysis by Signorovitch et al. (2013b) performed an anchored indirect comparison for PFS and an unanchored indirect comparison for OS. Several methods to account for treatment switching in the analysis of OS are available, including rank-preserving structural failure time models (Robins and Tsiatis 1991), iterative parameter estimation (Branson and Whitehead 2002), and inverse probability of censoring weighting (Robins and Finkelstein 2000), yet none of the analyses made use of these approaches. (An overview

of treatment switching methods and guidance on their use in NICE appraisals is given by Latimer and Abrams (2014).)

An analysis by Sikirica et al. (2013) had common placebo arms between the two trials, yet made an unanchored indirect comparison. The authors' justification was that, in the matching procedure, weights were additionally constrained to exactly balance placebo outcomes across trials. This method has yet to be evaluated either formally or through simulation studies, and its properties and performance in comparison with anchored methods are uncertain; in particular it is unlikely that balancing placebo outcomes is equivalent to relying on randomisation to remove residual differences due to unobserved prognostic variables.

A sizeable proportion (40%) of analyses applied MAIC to single-arm trials, or in situations with no common comparator. The only choice in such a scenario is to perform an unanchored indirect comparison. As in all cases where unanchored indirect comparisons are performed, a strong assumption is made that all prognostic variables and all effect modifiers are accounted for and correctly specified—an assumption largely considered to be implausibly strong. The published applications of unanchored MAIC acknowledge the possibility of residual bias due to unobserved prognostic variables and effect modifiers; however, it is not made clear that the accuracy of the resulting estimates is entirely unknown, because there is no analysis of the potential magnitude of residual bias, and hence no idea of the degree of error in unanchored MAIC estimates. Moreover, the inclusion of single-arm studies in an analysis is subject to the additional assumptions and biases incurred by these study designs (Deeks et al. 2003).

Availability of multiple studies 3.1.1.2

In half of the published analyses, issues arose with multiple IPD or aggregate populations for the same treatments. In both cases where multiple populations with IPD were available, the populations were simply pooled and treated as one large population. There was seemingly no attempt to account for the clustering of individuals within the component trials, which has been seen to incur bias and reduce power in the closely related context of IPD meta-analysis (Abo-Zaid et al. 2013). A better option in this scenario, in the absence of MAIC methodology which accounts for clustering, is to perform identical MAICs based on each IPD population, and then pool the relative effect estimates (on the linear predictor scale) with standard (network) meta-analysis methods (e.g. Ades 2003; Dias et al. 2011c; Hasselblad 1998; Higgins and Whitehead

1996; Lu and Ades 2004).

Multiple aggregate populations were pooled in two out of three cases, and analysed separately in one other. When aggregate populations are pooled, this should always be done with relative effects on the linear predictor scale to avoid complications such as conflicts of scale (see Section 2.3.3). There are two equivalent ways in which such an analysis may be done: (i) perform identical MAICs into each aggregate population, and then pool the $\hat{d}_{BC}$ relative effect estimates; or (ii) pool the aggregate populations and the $\hat{d}_{AC}$ relative effect estimates, and then perform a single MAIC into the pooled population. In either case, the pooling of relative effect estimates should take place on the linear predictor scale using standard methods (Ades 2003; Dias et al. 2011c; Hasselblad 1998; Higgins and Whitehead 1996; Lu and Ades 2004), and the resulting target population will be the (appropriately weighted) combination of the aggregate populations—which may or may not match the true target population for the decision. If separate analyses are performed for each aggregate population, it should be noted that the resulting estimates are each valid for a different target population (i.e. for each aggregate population) and are not comparable unless the target populations have balanced distributions of effect modifiers.

3.1.1.3 Larger treatment networks

Two papers presented analyses involving more than three treatments. One by Signorovitch et al. (2011b) had four treatments arranged in a square network (Figure 3.1a), essentially giving two possible common comparators (placebo and another active treatment) between the treatments of interest B and C. The other by Kirson et al. (2013) had four treatments in a star network (Figure 3.1b), in this case having two competitor treatments C and D to make indirect comparisons with B.

Signorovitch et al. (2011b) had access to IPD on the AB and BD studies, with aggregate data on the AC and CD studies; therefore two possible MAIC analyses could be performed, one via treatment A, and another via treatment D. The two resulting indirect comparison estimates are valid for different target populations—one for AC and one for CD—which were then pooled. The target population of the MAIC in this case is therefore a weighted combination of the AC and CD populations, which is unlikely to match the true target population for the decision.

Kirson et al. (2013) faced a similar scenario, where there were two competitor treatments C and D with aggregate AC and AD trial data with which to

form an indirect comparison. Again, two MAICs were performed, this time giving an estimate of $d_{BC(AC)}$ and of $d_{BD(AD)}$. These relative effect estimates are not comparable as they are both valid for different target populations (AC and AD respectively), unless the two target populations have balanced distributions of effect modifiers. There is no way with current MAIC methods to achieve a coherent comparison of all four treatments in this case when the AC and AD trial populations differ in terms of effect modifiers.

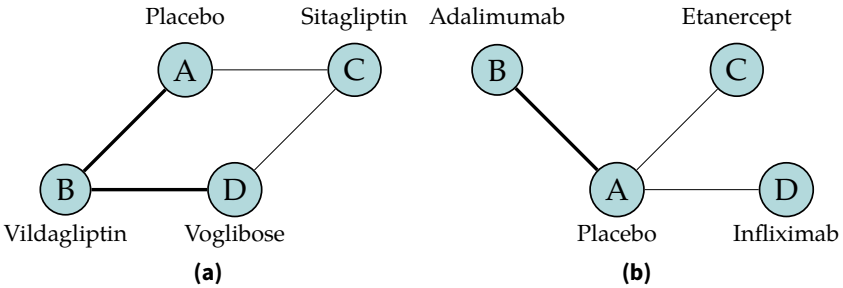


Figure 3.1 (a) Signorovitch et al. (2011b) perform two MAICs via alternate common comparators; (b) Kirson et al. (2013) perform two MAICs for two different competitor treatments. Thick edges indicate availability of IPD, thin edges indicate only AgD being available.

Effective sample size and weight distributions 3.1.1.4

Only 40% of the published MAIC analyses made any mention of either effective sample size or the distribution of weights: three included an ESS, and one other included a summary of the distribution of weights. The reporting of at least one of these is fundamental to understanding and diagnosing poor overlap between the IPD and aggregate populations. When the ESS is markedly reduced, or equivalently the weights are highly variable, estimates become unstable and inferences depend heavily on just a small number of individuals. The three papers reporting ESS saw an 80% average reduction from the original sample size (range: 57–98%). Absolute values for ESS ranged from 14 to 591.

Choice of matching variables 3.1.1.5

The number of matching variables used in the published MAIC analyses varied between 2 and 17. Most analyses balanced the standard deviations of covariates as well as means or other summary statistics between the populations, but only one (Sikirica et al. 2013) included any interactions or higher order terms in the weighting model. The majority of published

MAIC analyses therefore are subject to the additional assumptions set out in Section 2.3.4 due to the use of marginal covariate distributions instead of the joint distribution; in particular, an assumption must be made either regarding the balance of covariate correlations between populations, or regarding the lack of interaction terms in the implicit outcome model induced on the scale of the indirect comparison.

In no anchored analysis was there any attempt to justify the effect modifier status of the variables included in the weighting model, either with clinical expertise or with prior empirical evidence. The NICE Methods Guide (NICE 2013) is explicit that effect modifier status should be justified prior to analysis. For unanchored comparisons, every prognostic variable as well as effect modifier should be included; only three analyses justified the included variables as being prognostic or effect modifying in any manner.

In general, published anchored MAIC analyses reported comparative estimates before and after the weighting adjustment, and noted any difference. However, the observation of a difference in relative effects after an analysis has been done should not be used to justify that an anchored MAIC should be preferred over a standard indirect comparison; such arguments amount to post hoc reasoning, whereas in the context of NICE technology appraisals all analyses should be clearly pre-specified (NICE 2013). No attempts were made prior to any analysis to assess the magnitude of impact of effect modifier imbalance on the indirect comparison.

In some cases where common placebo arms were present, placebo tests were performed as an attempt to justify the validity of the MAIC. However, such tests can only detect imbalance in observed or unobserved prognostic variables, and are completely unable to detect imbalances in observed or unobserved effect modifiers. It is arguable whether placebo tests in this context add any value at all: anchored indirect comparisons by design account for differences in prognostic variables between the two populations, so any imbalanced prognostic variables will not lead to bias in the indirect comparison but will cause a placebo test to "fail"; placebo tests should not be used to "justify" unanchored indirect comparisons due to their low power.

3.1.1.6 Choice of scale

The choice of scale for an indirect comparison is important, as assumptions are implied on the indirect comparison scale regarding additivity of effects, definition of prognostic and effect modifying variables, and distributional properties (Section 2.3.3). Almost all published MAICs carried out the indirect

comparison on the natural outcome scale. In many cases this led to indirect comparisons being made on scales not commonly used for meta-analyses, such as probability differences rather than log odds ratios. As in meta-analysis, the appropriate scale should be considered on a case-by-case basis, in light of the biological and clinical knowledge, with the default scale determined by existing literature.

Table 3.1 Applications of MAIC in the literature.

Paper	Trials and treatments	IPD sample size	AgD sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of outcome in indirect comparison
Signorovitch et al. (2010)	Company: Adalimumab (B) vs. Placebo (A) Pooled two AB populations Competitor: Etancercept (C) vs. Placebo (A)	Original: 1359 After excl. criteria: 1025 ESS: 591	330	10 (5 with SD)	2 Based on clinical reasoning	4 (statistically significant)	Anchored	Response probability, percent change in PASI
Chang et al. (2011)	Company: Bevacizumab + cisplatin (B) Competitor: Pemetrexed + cisplatin (C) Two single-arm trials	Original: 2172 After excl. criteria: 72 ESS: 46	67	2 (0 with SD)	0 Variables described as "potentially prognostic"	2 (numerically different)	Unanchored	Median PFS

Table 3.1 (continued)

Paper	Trials and treatments	IPD sample size	AgD sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of indirect comparison
Signorovitch et al. (2011a)	Company: Nilotinib (*B*) vs. Imatinib (*A*) Competitor: Dasatinib (*C*) vs. Imatinib (*A*)	Original: *A*: 282 *B*: 283 After excl. criteria: *A*: 280 *B*: 273 ESS: Not reported	*A*: 260 *C*: 259	10 (0 with SD)	0	3 (numerically different)	Unanchored	Proportion of MMR, PFS, and OS at 1 year

Table 3.1 (continued)

Paper	Trials and treatments	IPD sample size	AgD sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of indirect comparison
Signorovitch et al. (2011b)	Company: Vildagliptin (B) vs. Placebo (A) Vildagliptin (B) vs. Voglibose (D) Competitors: Sitagliptin (C) vs. Placebo (A) Sitagliptin (C) vs. Voglibose (D) Two AC populations pooled for one analysis at one dose level	Original: AB: 148 BD: 380 After excl. criteria: AB: 148 BD: 363 ESS: Not reported	AC: 145 CD: 319	6 (5 with SD)	0 Noted large heterogeneity in previous meta-analyses	3 (statistically significant)	Anchored Two MAICs performed with Placebo and Voglibose as common comparators, results then pooled	Mean HbA1c

Table 3.1 (continued)

Paper	Trials and treatments	IPD sample size	AgD sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of indirect comparison
Kirson et al. (2013)	Company: Adalimumab (B) vs. Placebo (A) Competitors: Etancercept (C) vs. Placebo (A) Infliximab (D) vs. Placebo (A)	Original: 313 After excl. criteria: 296 (for AC) 234 (for AD) ESS: Not reported	AC: 205 AD: 200	For AC: 12 (6 with SD) For AD: 17 (11 with SD)	0	2 for AC and 4 for AD (statistically significant)	Anchored	Response rates, percent change
Signorovitch et al. (2013b)	Company: Everolimus (B) vs. Placebo (A) Competitor: Sunitinib (C) vs. Placebo (A)	Original: 410 After excl. criteria: 394 ESS: Not reported	171	9 (0 with SD)	0	3 (statistically significant)	Anchored for PFS Unanchored for OS	log hazard ratios

Table 3.1 (continued)

Paper	Trials and treatments	IPD sample size	AgD sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of indirect comparison
Sikirica et al. (2013)	Company: Guanfacine (B) vs. Placebo (A) Pooled two AB populations Competitor: Atomoxetine (C) vs. Placebo (A)	Original: 631 After excl. criteria: A: 136 B: 82 ESS: Not reported	A: 83 C: 84	4 (with SDs, pairwise interactions, quadratic and cubic terms)	0	1 (statistically significant)	Unanchored Weights are constrained such that placebo arms match exactly	Mean ADHD scores
Sherman et al. (2015)	Company: Everolimus (B) Competitor: Axitinib (C) No common comparator, other arms ignored	Original: 277 After excl. criteria: 43 ESS: Not reported	194	3	0 Variables found using latent class model as being influential on PFS	3 (numerically different)	Unanchored	Median PFS

Table 3.1 (continued)

Paper	Trials and treatments	IPD sample size	AgD sample size	Number of matching variables	Variables where evidence for effect modifier status is presented	Variables where evidence of imbalance is presented	Anchored or unanchored indirect comparison	Scale of indirect comparison
Van Sanden et al. (2016)	Company: Simeprevir + peginterferon alfa 2a + ribavirin (*B*) Competitor: Peginterferon alfa 2a + ribavirin (*C*1–5) Single arms, multiple *C* populations	Original: 107 After excl. criteria (ESS): For *C*1: 35 (29) For *C*2: 35 (15) For *C*3: 57 (14) For *C*4: 35 (26) For *C*5: 19 (17)	*C*1: 30 *C*2: 18 *C*3: 95 *C*4: 40 *C*5: 109	5–6	0 Consulted two experienced hepatologists for variables "relevant to treatment response"	Some numerical differences	Unanchored	Proportion achieving sustained virologic response
Swallow et al. (2016)	Company: Daclatasvir + sofosbuvir (*B*) Competitor: Sofosbuvir + ribavirin (*C*) Pooled two *C* populations All open label, single-arm	Original: 153 After excl. criteria: 91 ESS: Not reported	455	14 (3 with SD)	0	4 (statistically significant)	Unanchored	Proportion achieving sustained virologic response

3.1.2 Applications of STC in the literature

Our literature search returned only one published application of STC to date. Nixon et al. (2014) present an analysis of oral therapies for the treatment of relapsing-remitting multiple sclerosis. A network diagram is shown in Figure 3.2. The AB population consisted of 1556 patients randomised to either fingolimod (B) or placebo (A) across two original trials with IPD. Unlike any MAIC analyses using pooled IPD, Nixon et al. correctly accounted for the clustering of the IPD by including a study-level baseline risk term in the outcome model (i.e. a separate intercept for each AB study). There were three trials with aggregate data: two comparing dimethyl fumarate (C) to placebo in a total of 2301 patients, and another comparing teriflunomide (D) to placebo in 1088 patients. Risk ratios and covariate distributions of the two AC trials were pooled simply using inverse variance weighting (essentially a fixed-effect meta-analysis of the two trials). Differences in covariate and outcome definitions between the AC and AD studies led Nixon et al. to produce two STC models, one using the AC definitions for prediction into the AC population, and the other using the AD definitions for prediction into the AD population.

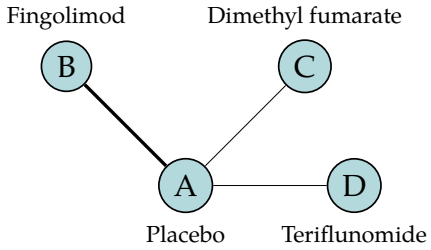


Figure 3.2 Network diagram for the STC analyses performed by Nixon et al. (2014). Thick edges indicate availability of IPD, thin edges indicate only AgD being available.

Of all published applications across MAIC and STC, Nixon et al. (2014) are the only authors to attempt to justify effect modifier status of any variables; both expert clinical opinion and the results of previous subgroup analyses were used in evidence. There was no analysis of the imbalance in any covariates between the three populations beyond simple numerical differences, however the use of an AIC-based backwards selection algorithm to choose the final model suggests that the remaining covariates were significantly predictive of outcome. The outcome model itself was a linear probability model, using an identity link function to regress the probability of response against the covariates. As noted earlier this is an uncommon modelling choice, not least because such models can lead to predicted probabilities that lie outside the

range 0 to 1. Similarly, this choice of model scale in this case also leads to problems with the anchored indirect comparison, which is constructed naturally on the (log) relative risk scale. It therefore breaks the "anchoring" which is taken advantage of by the anchored indirect comparison. In the outcome regression, prognostic variables (and effect modifiers) are defined with respect to the linear probability scale, however the use of the log RR for the anchored indirect comparison means that prognostic variables will not cancel (see Section 2.2.2).

3.2 Applications of population adjustment in NICE Technology Appraisals

The use of population adjustment methodology is becoming increasingly widespread in technology appraisals, in which it is typical for only limited IPD to be available. A company submitting to a regulatory or reimbursement agency will have IPD available for their own trials, but likely only published AgD from their competitors'. In this section, we undertake a review of TAs published by NICE (NICE 2018d), aiming to characterise the use of population adjustment methods. As well as investigating the uptake of population adjustment in different clinical areas, we are interested in the ways in which these methods are used and whether the key assumptions are likely to hold, to assess the adequacy of current practice for decision making. We discuss how these methods have been received by appraisal committees and how they have impacted decision making.

3.2.1 Review outline

We reviewed all NICE TAs published between 1st January 2010 and 20th April 2018 for the use of population adjustment methods. We excluded appraisals that had access to IPD from all included studies, and focused on those with only partial availability of IPD. From those appraisals using one or more forms of population adjustment, we extracted the following information from company submissions:

- Population adjustment method used;
- Whether the comparison was anchored or unanchored;
- Outcome type;
- Clinical area;
- Number of covariates adjusted for;
- How the covariates were chosen;

- For appraisals using MAIC, effective sample sizes after weighting;
- Whether a larger network structure was present (e.g. multiple comparators and/or aggregate studies), and how this was dealt with.

3.2.2 Results

A total of 268 technology appraisals have been published by NICE since 2010—when MAIC and STC were first suggested in the literature (Caro and Ishak 2010; Signorovitch et al. 2010)—up until 20th April 2018. Of these, 21 appraisals used a form of population adjustment; three of these had IPD available from all included studies, so we focus on the remaining 18 appraisals with only partial IPD. Figure 3.3 shows the selection process. The included appraisals are tabulated in Table 3.2.

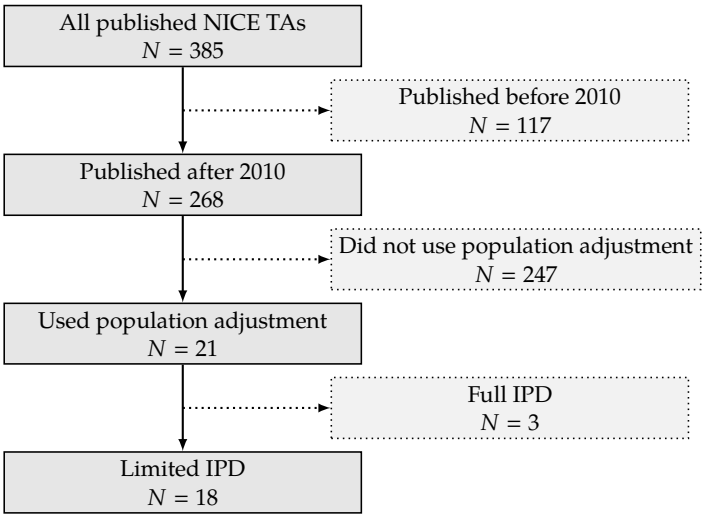


Figure 3.3 Flow chart showing the process of selecting technology appraisals, and the numbers excluded and remaining at each stage.

The first use of population adjustment in a TA was TA311 in 2014. Since then, the use of population adjustment in TAs has increased rapidly, in terms of both the absolute number and the relative proportion of appraisals using population adjustment methods Figure 3.4. In 2017, a total of nine appraisals used population adjustment, accounting for 14.5% of all appraisals that year.

3.2.2.1 Usage by clinical area

Since 2010, almost half of all published TAs have been in oncology (127 of 268, 47.4%). Of these, 15 (11.8%) used population adjustment, accounting for over 80 percent of all applications of population adjustment in appraisals to date. Only two other clinical areas saw any applications of population adjustment:

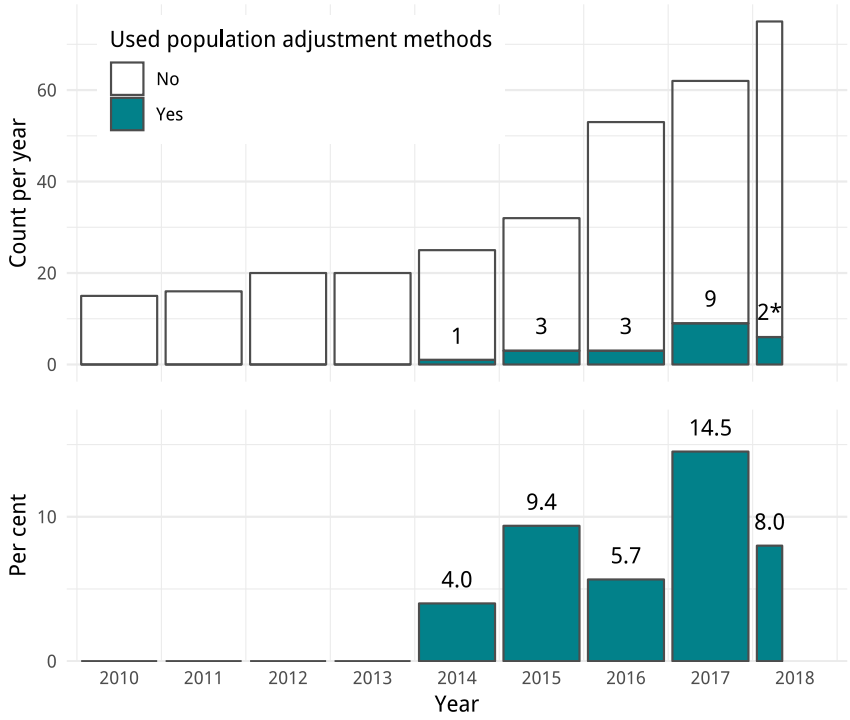


Figure 3.4 The number and percentage of NICE technology appraisals using population adjustment methodology has increased greatly since the introduction of these methods in the literature in 2010.
* Two TAs used population adjustment out of 25 up to 20th April 2018.

two out of 12 (16.7%) appraisals in hepatology (both for hepatitis C), and one out of 28 (3.6%) appraisals in rheumatology. The usage of population adjustment methods in oncology TAs has increased since 2010, both in terms of the number and proportion of TAs using these methods. In 2017, a total of 9 appraisals in oncology (25.7%) used population adjustment methods, up from one appraisal (9.1%) in 2014 (Figure 3.5). The increasing use of population adjustment in oncology appraisals, which themselves make up the largest proportion of all appraisals, is the main driver behind the overall results in Figure 3.4.

Outcome types 3.2.2.2

Unsurprisingly, due to the majority of applications of population adjustment being in oncology appraisals, survival outcomes (e.g. progression-free survival, overall survival) were the most common outcome type used in population adjustment—13 of 18 appraisals (72.2%) included a population-adjusted survival outcome. Rate outcomes such as response rates were used in 5 appraisals, and duration and change from baseline outcomes in one appraisal

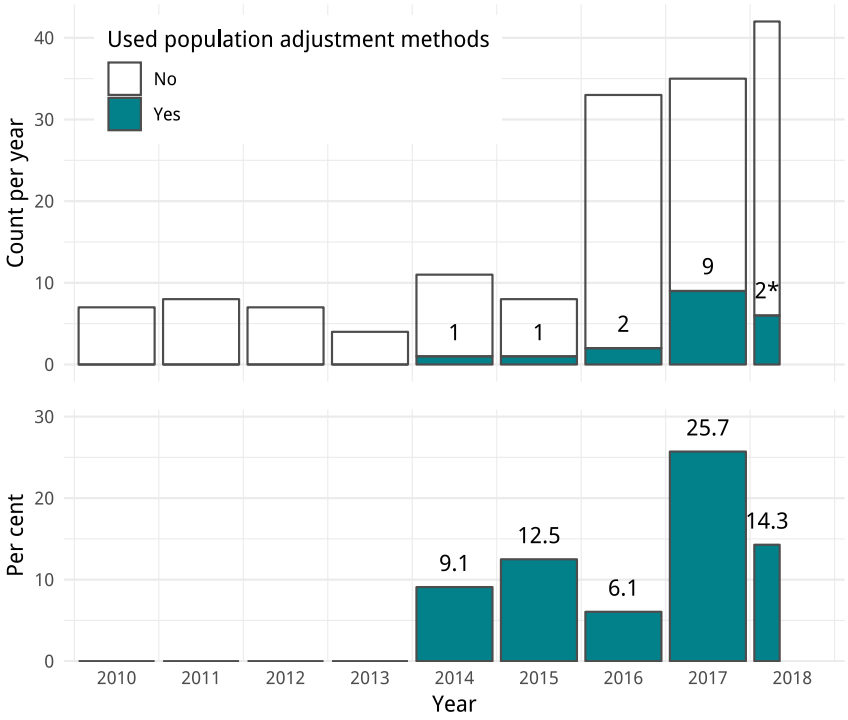


Figure 3.5 For technology appraisals in oncology, the number and percentage using population adjustment methods has increased greatly since the introduction of these methods in the literature in 2010.
* Two TAs used population adjustment out of 12 up to 20th April 2018.

each. Two appraisals (TA462, TA451) used population adjustment for more than one type of outcome (survival and response rate, and response rate and duration respectively).

3.2.2.3 Population adjustment method

The large majority of appraisals using some form of population adjustment used MAIC (16 out of 18, 88.9%). STC was less popular, used in only 3 appraisals (16.7%). Two appraisals used both MAIC and STC and compared the results, which were reported to be similar in each case (TA383, TA492).

One appraisal (TA410) used neither MAIC nor STC. In this appraisal, a published prediction model (developed for a previous appraisal (NICE 2014b)) was used to adjust the survival curves from the AgD trials to the population of the IPD trial.

Of the 16 appraisals performing MAIC, only 9 (56.3%) reported an effective sample size (ESS) (see equation (2.21) in Section 2.2.1). Of these, the median effective sample size was 80.0 (range: 4.0 to 335.5, IQR: 15.4 to 52.0), with a median reduction in effective sample size from the original sample size of

74.2% (range: 7.9% to 94.1%, IQR: 48.0% to 84.6%). Such large reductions in ESS indicate that in many cases there may be poor overlap between the IPD and AgD studies. A substantial proportion of TAs reported small absolute ESS, and the resulting comparisons are therefore dependent on a small number of individuals in the IPD study and may be unstable.

Anchored and unanchored comparisons 3.2.2.4

Only 2/18 appraisals (11.1%) formed anchored comparisons (TA383, TA449). The remaining 16 appraisals (88.9%) instead formed unanchored comparisons without a common comparator, relying on strong assumptions that are very difficult to justify and are thus subject to unknown amounts of residual bias. No appraisals attempted to quantify residual bias, although this is challenging to achieve (Phillippo et al. 2016). Appraisal committees and review groups treated estimates from unanchored comparisons with strong caution.

Covariates adjusted for 3.2.2.5

For appraisals reporting unanchored comparisons, the median number of covariates adjusted for was 6, and ranged from 1 to 13 covariates. Only one of the two appraisals reporting anchored comparisons presented any information on the choice of covariates; in this appraisal (TA383) 10 covariates were adjusted for.

Common covariates adjusted for in oncology appraisals were age, Eastern Cooperative Oncology Group (ECOG) performance status, gender, and the number and/or type of previous therapies. Many appraisals also adjusted for other clinical factors such as biomarker levels or disease subtypes.

Both hepatitis C appraisals (TA364, TA331) adjusted for age, body mass index (BMI), gender, fibrosis staging, and viral load. One appraisal (TA364) further adjusted for race, genotype, and several biomarker levels in two MAIC analyses for different genotypes and comparator treatments, but in a third MAIC analysis only had sufficient sample size to adjust for viral load.

The single rheumatology appraisal (TA383) adjusted for 10 covariates including age, gender, race, concomitant treatments, two biomarkers and three functional/activity scores.

The most common justification for covariate selection amongst appraisals reporting unanchored comparisons was simply to adjust for all baseline characteristics reported in both studies. This was also true for appraisal TA383 which used an anchored comparison, despite the fact that adjustment is only required for covariates which were effect modifiers in anchored comparisons.

(The other appraisal with an anchored comparison, TA449, did not report any information on variable selection.) Unnecessary adjustment will not introduce bias but may increase uncertainty, particularly with MAIC (see Section 2.3.1 and Phillippo et al. (2016), although we note that TA38 took place before the advice in Phillippo et al. (2016) was published). Two appraisals (TA429, TA457) justified the selection of covariates using expert clinical opinion. One appraisal using MAIC (TA510) asked experts to rank covariates by importance, then added covariates into the model one-by-one in decreasing order of importance; the final model choice was determined by consideration of effective sample size. Unanchored MAICs in particular have to make trade-offs between effective sample size and the number of adjustment variables, since the number of potential prognostic factors is likely to be large. However, unless all prognostic factors and effect modifiers are included in the adjustment, the estimates will remain biased (see Section 2.3.2, and Phillippo et al. 2016). Moreover, the covariates for which the effective sample size reduction is greatest are those which are most imbalanced between populations, and are therefore more important to adjust for amongst the covariates with similar prognostic or effect modifying strength. Two appraisals using STC used statistical techniques to choose covariates. One (TA333) selected covariates that were "significant" in the regression model, which is again likely to incur residual bias—particularly in small samples (Steyerberg et al. 1999). Another (TA492) selected covariates to maximise cross-validated predictive performance, which is more appropriate than selection based on "significance" given that STC relies on accurate predictions into the aggregate population, but is still subject to the limitations of in-sample validation (Phillippo et al. 2016).

3.2.2.6 Larger networks

As originally proposed, MAIC and STC cannot be extended to larger network structures with multiple comparators of interest and/or multiple aggregate studies. However, these scenarios frequently arise in practice: a total of 10 out of 18 TAs (55.6%) involved larger networks of treatments and studies.

In five of these (71.4%; TA331, TA383, TA429, TA500, TA510), multiple population adjusted indirect comparisons were performed and then simply left as stand-alone estimates. Each of these estimates will be valid for different target populations, and so cannot be interpreted together coherently unless additional assumptions are met—namely that all the target populations are in fact identical (in terms of effect modifiers for anchored comparisons, and also in terms of prognostic variables for unanchored comparisons).

One appraisal (TA492) used STC (and MAIC as a sensitivity analysis) to predict active treatment arms for each single-arm study in an unconnected network, and then analysed this newly-connected network using network meta-analysis (NMA). This results in a coherent set of relative effect estimates. However, aside from the very strong assumptions required for the unanchored comparisons, this analysis must also assume that there are no imbalances in effect modifiers between the single-arm studies included in the NMA. Another serious concern is the repeated use of the predicted active treatment arms, which are all based on the same data set and so are not independent.

Two appraisals (TA311, TA380) had wider networks of treatments and studies including the two treatments of primary interest, but that were not fully connected. These networks were analysed using NMA (without any population adjustment) using an equivalency assumption for two treatments (TA311) and a matched pairs analysis (TA380) to connect the networks. Separate unanchored MAICs were then used to create population-adjusted comparisons as sensitivity analyses.

One appraisal (TA427) had additional single-arm IPD sources which were used to provide additional stand-alone comparisons (in this case using Cox regression for survival outcomes). Guidance on single-arm comparisons with full IPD in NICE TAs has previously been published by Faria et al. (2015).

Lastly, the method of analysis was unclear for one appraisal (TA364) which had multiple comparators of interest, some with several AgD studies available. However, given that unanchored MAIC was used, this analysis is susceptible to the same sets of pitfalls described above depending on whether the estimates were left as stand-alone estimates or synthesised as a network.

Table 3.2 Applications of population adjustment in NICE Technology Appraisals. Superscripts in the covariates column indicate the subset of covariates selected using the approach detailed in the adjacent column.

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA510: Daratumumab monotherapy for treating relapsed and refractory multiple myeloma (NICE 2018c)	March 2018	MAIC	Unanchored	Oncology	Survival	refractory status*†, ECOG*†, prior treatments*, creatinine clearance*, time since diagnosis, myeloma subtype, race, bone lesions, prior ASCT, age	Using literature review and expert opinion, covariates ranked by importance then forward selected. MAICs performed into two target trials, the first adjusting for 4 covariates (*), the second for 2 covariates (†)	84 (57.8), 80 (54.1)	Stand-alone comparisons

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA500: Ceritinib for untreated ALK-positive non-small-cell lung cancer (NICE 2018b)	January 2018	MAIC	Unanchored	Oncology	Survival	age, gender, race, smoking status, adenocarcinoma, ECOG, metastatic disease, brain metastases	All baseline characteristics available in both trials	171 (90.4), 174 (92.1)	Stand-alone comparisons
TA492: Atezolizumab for untreated locally advanced or metastatic urothelial cancer when cisplatin is unsuitable (NICE 2017a)	December 2017	STC, MAIC	Unanchored	Oncology	Survival	age*, gender*, ECOG*, liver metastases*, number of prior therapies	STC selected covariates (*) to maximise cross-validated predictive performance. MAIC used all available covariates.	-	Active arms predicted for each study to connect the network, then analysed using NMA

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA462: Nivolumab for treating relapsed or refractory classical Hodgkin lymphoma (NICE 2017h)	July 2017	MAIC	Unanchored	Oncology	Survival, Response rate	age, gender, disease stage, B symptoms, haemoglobin, lymphocytes, white cell count, albumin, extranodal site, ECOG, tumour diameter, number of prior therapies	-	81 (42)	-
TA478: Brentuximab vedotin for treating relapsed or refractory systemic anaplastic large cell lymphoma (NICE 2017b)	October 2017	MAIC	Unanchored	Oncology	Survival	age, gender, elevated lactate dehydrogenase, disease stage, ECOG, response to primary therapy	All baseline characteristics available in both trials	4.8 (8.3)	-

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA457: Carfilzomib for previously treated multiple myeloma (NICE 2017c)	July 2017	MAIC	Unanchored	Oncology	Survival	age, ISS stage, time since diagnosis, creatinine clearance, number of prior therapies, prior SCT, prior bortezomib, prior IMiD, refractory to last therapy	Covariates identified as prognostic factors by UK clincal experts reported in both studies	335.5 (52)	-
TA449: Everolimus and sunitinib for treating unresectable or metastatic neuroendocrine tumours in people with progressive disease (NICE 2017d)	June 2017	MAIC	Anchored	Oncology	Survival	-	-	-	-

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA451: Ponatinib for treating chronic myeloid leukaemia and acute lymphoblastic leukaemia (NICE 2017j)	June 2017	MAIC	Unanchored	Oncology	Response rate, Duration	age, gender, T315I mutation, race, duration of disease, ECOG	All baseline characteristics available in both trials	69 (25.8)	-
TA432: Everolimus for advanced renal cell carcinoma after previous treatment (NICE 2017e)	February 2017	MAIC	Unanchored	Oncology	Survival	-	-	-	-

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA429: Ibrutinib for previously treated chronic lymphocytic leukaemia and untreated chronic lymphocytic leukaemia with 17p deletion or TP53 mutation (NICE 2017f)	January 2017	MAIC	Unanchored	Oncology	Survival	17p deletion status, number of prior therapies, purine refactory status, age, binet/RAI, IGVH status, beta2-microglobulin, del 11q, creatinine clearance, platelets, gender, haemoglobin, lymphocytes	All clinically relevant baseline characteristics available in both trials, reviewed by clinical experts	30 (15.4)	Stand-alone comparisons
TA427: Pomalidomide for multiple myeloma previously treated with lenalidomide and bortezomib (NICE 2017i)	January 2017	MAIC	Unanchored	Oncology	Survival	age, ECOG, number of prior therapies, prior thalidomide	-	-	Other single-arm IPD sources used for additonal stand-alone comparisons

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA410: Talimogene laherparepvec for treating unresectable metastatic melanoma (NICE 2016c)	September 2016	Prediction model	Unanchored	Oncology	Survival	gender, ECOG, visceral status, brain metastases, LDH	Modification of published prediction model (Korn model)	-	-
TA383: TNF-alpha inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis (NICE 2016d)	February 2016	MAIC, STC	Anchored	Rheumatology	Change from baseline	gender, race, age, concomitant DMARD, concomitant NSAID, HLA-B27, BASDAI, BASFI, ASDAS, CRP	All baseline characteristics available in both trials	-	Stand-alone comparisons

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA380: Panobinostat for treating multiple myeloma after at least 2 previous treatments (NICE 2016b)	January 2016	MAIC	Unanchored	Oncology	Survival	age, gender, time since diagnosis, ECOG, number of prior therapies, prior thalidomide, prior bortezomib, prior stem cell transplant, beta2-microglobulin	All baseline characteristics available in both trials	137 (35.4), 23 (5.9)	Network of studies available, MAIC used to target single comparison
TA364: Daclatasvir for treating chronic hepatitis C (NICE 2015b)	November 2015	MAIC	Unanchored	Hepatology	Rate	age, BMI, race, gender, HCV genotype, plasma HCV RNA, fibrosis staging, IL28B genotype, platets, ALT, previous treatment	All baseline characteristics available in both trials One MAIC only had sufficient sample size to adjust for one covariate (HCV RNA)	-	Unclear

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA331: Simeprevir in combination with peginterferon alfa and ribavirin for treating genotypes 1 and 4 chronic hepatitis C (NICE 2015c)	February 2015	MAIC	Unanchored	Hepatology	Rate	fibrosis score, viral load, BMI, age, gender	-	15 (14)	Stand-alone comparisons
TA333: Axitinib for treating advanced renal cell carcinoma after failure of prior systemic treatment (NICE 2015a)	February 2015	STC	Unanchored	Oncology	Survival	gender, age*, nephrectomy status, previous radiotherapy, previous cytokine therapy, MSKCC*†, clear cell carcinoma, ECOG, time on subitinib†	Significant predictors of outcome in regression model ($p < 0.1$) for PFS (*) and OS (†)	-	-

Table 3.2 (continued)

Appraisal	Date published	Population adjustment method	Anchored or unanchored comparison	Clinical area	Outcome type	Covariates	How were covariates chosen?	MAIC effective sample size (%)	If available, how was a larger network dealt with?
TA311: Bortezomib for induction therapy in multiple myeloma before high-dose chemotherapy and autologous stem cell transplantation (NICE 2014a)	April 2014	MAIC	Unanchored	Oncology	Response rate	ISS stage, beta2-microglobulin, cytogenetic abnormality t4, age, gender, light chain myeloma, IG-A, IG-D, IG-G	-	-	Network of studies available, MAIC used to target single comparison

3.3 Discussion

In these two reviews, we have seen that population adjustment methods are becoming ever more prevalent in the published literature and in NICE Technology Appraisals. Different practices may be found in submissions to other reimbursement agencies, who may also receive and interpret such analyses differently. Both reviews span a limited time period since these methods were first published and practice is likely to continue to evolve, for example as methodological guidance is published. A further limitation of these reviews is that the data extraction was carried out by a single reviewer only. Since its publication in late 2016, the NICE Decision Support Unit Technical Support Document 18 (Phillippo et al. 2016) has been cited in every following TA that used population adjustment methods, with committees and review groups using TSD 18 to inform and justify their conclusions regarding population-adjusted analyses.

In NICE TAs, decisions are not based solely on clinical effectiveness; cost considerations are also taken into account in a cost-effectiveness analysis. The impact of population adjustment methods on appraisals is therefore understood within this context. Through discussions with colleagues at NICE (Phillippo et al. 2019a), it has become apparent that committees often had concerns regarding the quality of evidence produced by such methods: the data used in the analyses were often weak (for example, immature follow-up data or small single-arm studies), and there was additional uncertainty over the covariates that were adjusted for (which covariates were selected and how, and whether and to what extent any unobserved covariates introduced bias). As a result, committees typically looked for greater cost-effectiveness before making a positive recommendation to offset the perceived risks from lower-quality, uncertain evidence. Appraisal committees were typically more likely to use the results of population-adjusted analyses for decision-making when they were presented alongside an additional confirmatory analysis, and when the uncertainty in the method was acknowledged and explored (for example using sensitivity analyses).

In both the published literature and in NICE TAs, a large number of analyses were unanchored with no common comparator, and hence rely on very strong assumptions as outlined in Section 2.3.2. The proliferation of unanchored analyses is likely to escalate, in large part due to the rise of single-arm studies for accelerated or conditional approval with regulators such as the US Food and Drug Administration or the European Medicines Agency (Hatswell et al. 2016). However, the evidential requirements for

demonstrating clinical efficacy (to obtain licensing) can be less stringent than those for demonstrating cost effectiveness (to obtain reimbursement). NICE appraisal committees and evidence review groups have been justifiably wary of the use of unanchored population adjustment methods to bridge this evidence gap, with many commenting that the results should be interpreted with caution and may contain an unknown amount of bias. Increased dialogue between regulators and reimbursement agencies may help bridge this gap in evidence requirements.

All current population adjustment methods assume that there are no unmeasured effect modifiers when making anchored comparisons. For unanchored comparisons, it is further assumed that there are no unmeasured prognostic factors. This latter assumption is particularly strong and difficult to justify. Quantifying residual bias due to unmeasured confounding is an area for future work, which we discuss in Section 9.2.8.

Several applications in the literature and in technology appraisals had multiple comparators and/or AgD study populations for which comparisons were required (see Sections 3.1.1.3, 3.1.2 and 3.2.2.6). Current MAIC and STC methodology cannot handle larger network structures: multiple analyses were performed in each case, and then either left as stand-alone comparisons or themselves synthesised using network meta-analysis, requiring further assumptions in the process. Furthermore, current MAIC and STC methods produce estimates which are valid only for the aggregate study population (typically that of a competitor) without additional assumptions, which may not match the target population for the decision (Phillippo et al. 2016). This fact has been largely overlooked in appraisals and in the published literature to date, although one appraisal (TA451) did note that the MAIC analysis that was performed took the results of an IPD trial deemed to be relevant to the decision population and adjusted them into a non-representative aggregate trial population. Clearly if effect modification is present then it is not enough to simply produce "unbiased" estimates: the estimates produced must be specific to the decision population, otherwise they are of little use to decision-makers. This motivates the development of new methods which can extend naturally to larger networks of treatments and can produce estimates for a given target decision population, in the following chapter. Furthermore, if all trials are a subset of the decision target population with respect to one or more effect modifiers, then any adjustment must rely on extrapolation; if these effect modifiers are discrete, adjustment may be impossible.

The large majority of published analyses and technology appraisals used

MAIC to obtain population-adjusted indirect comparisons. Effective sample sizes were typically small and often substantially reduced compared to the original sample sizes, indicating potential lack of overlap between the IPD and AgD populations. Lack of overlap is of particular concern with re-weighting methods such as MAIC since they cannot extrapolate to account for covariate values beyond those observed in the IPD, and thus may produce estimates that remain biased even when all necessary covariates are included in the model. This motivates the need for simulation studies to explore the robustness of MAIC (and other population adjustment methods) in scenarios where there is a lack of overlap between populations (see Section 8.2.5).

Three appraisals were excluded from our review of TAs, as IPD were available from all included studies (NICE 2016a, 2017g, 2018a). These appraisals were all unanchored comparisons of survival outcomes in oncology, and used a selection of propensity score, covariate matching, and regression methods. Having IPD available from all studies is the gold-standard and is preferable if at all possible. This is because IPD allows for analyses that have more statistical power and may rely on less stringent assumptions, and allows assumptions to be tested. Methodological guidance is available for analyses with full IPD in NICE TAs (Faria et al. 2015).

For population-adjusted analyses to have the desired impact on decision making, including in technology appraisals, several key improvements are needed to current practice (Phillippo et al. 2016). Firstly, a target population relevant to decision makers should be defined, and estimates must be produced for this population in order to be relevant. Current population adjustment methods can only produce estimates valid for the population represented by the aggregate study unless further assumptions are made, which may not represent the decision at hand; this has been largely overlooked in appraisals and applications in the literature to date (although note that several of the TAs and all of the applications in the literature that we identified pre-date published guidance). For anchored comparisons there should be clear prior justification for effect modification, based on empirical evidence from previous studies and/or clinical expertise. To date, only one published application of anchored population adjustment made any attempt to justify effect modification (Nixon et al. 2014, see Section 3.1.2); no technology appraisals or any other applications in the literature that reported anchored population-adjusted analyses provided any such justification. Unanchored comparisons require reliable predictions of absolute effects via adjustment for both prognostic and effect modifying covariates, and are highly susceptible to unobserved confounding due to a

lack of randomisation. Simply adjusting for all available covariates, as is currently common practice, is not sufficient. For unanchored comparisons to be impactful, covariates should be selected with predictive performance in mind and estimates of the potential range of residual bias are required; otherwise, the amount of bias in the estimates is unknown and may even be larger than for an unadjusted comparison. This is not easy to achieve (some suggestions are made in the discussion, see Section 9.2.8), but without such reassurance decision makers are likely to remain justifiably wary of unanchored analyses. Many of the above issues can be mitigated—at least in part—by the availability of IPD from all studies in an analysis, and thus the increased sharing of IPD is greatly encouraged.

Chapter 4

A new approach: Multilevel Network Meta-Regression

Current methods for population-adjusted indirect comparisons combining IPD and AgD include MAIC and STC, and network meta-regression (Section 2.2). MAIC and STC (Caro and Ishak 2010; Ishak et al. 2015; Signorovitch et al. 2010) are based on reweighting and regression adjustment ideas seen in the standardisation literature, and represent a departure from the standard framework for indirect comparisons and network meta-analysis (Sections 2.2.1 and 2.2.2). MAIC and STC are designed with a simple two-study scenario in mind where there is IPD available from an AB study and only published AgD from an AC study, and are not readily generalisable to larger treatment networks. Moreover, they are also limited to providing a comparison adjusted to the population of the AgD trial without further assumptions, which may not match the relevant target population for decision makers—rendering estimates irrelevant (Section 2.5). Current network meta-regression approaches (Section 2.2.3) are consistent with the standard NMA framework, can be applied to networks of all sizes, and can produce estimates for relevant target populations. These methods combine available IPD with AgD in a NMA framework and estimate treatment-covariate interaction terms for effect modifiers (Donegan et al. 2013; Jansen 2012; Saramago et al. 2012; Sutton et al. 2008; Thom et al. 2015). Typically, the same model is applied at both the individual and aggregate level, which leads to aggregation bias (a form of ecological bias) when the model is non-linear (Rothman et al. 2008). Two approaches have been proposed to account for this. The first is to split the interaction effect into between-study (or aggregate-level) and within-study (or individual-level) effects (Donegan et al. 2013; Saramago et al. 2012; Sutton

et al. 2008; Thom et al. 2015). However, in the two-study scenario there is insufficient data to identify the additional between-study parameter alongside the treatment effect; this approach is therefore not applicable in this scenario. The second approach is to define the aggregate-level model by integrating the individual-level model over the study population, avoiding aggregation bias by properly relating the two levels. So far, this approach has only been derived for the simple case of discrete covariates, where the integration reduces to summation (Jansen 2012), limiting its usefulness. There is therefore a clear need for new methods which address these issues.

As we found in our review of the literature, there have been promising attempts to combine information from several levels in an ecological context, where data are available in aggregate form on, say, geographical areas, but only limited individual survey data are available (Jackson et al. 2006, 2008; see Section 2.2.3). These methods motivated the approach of Jansen (2012) mentioned above to network meta-regression combining IPD and AgD. In this chapter we build on the ideas of Jansen (2012) and the ecological inference literature, further generalising these methods to incorporate both continuous and discrete covariates for use in population-adjusted indirect comparisons and network meta-regression combining IPD and AgD. We call this approach *Multilevel Network Meta-Regression* (ML-NMR). ML-NMR extends the standard NMA framework to synthesise evidence from both IPD and AgD simultaneously whilst avoiding aggregation bias, and is highly flexible. When IPD are available on all studies, ML-NMR reduces to IPD network meta-regression—the "gold standard" approach.

We begin by outlining the general framework for ML-NMR, based on that set out by Jackson et al. (2006, 2008) but contextualised for indirect comparisons in a simple two-study scenario (Section 4.1). Derivation of the model depends upon two things: i) the likelihood for the individual-level data, and ii) the link function used for the individual-level linear predictor. Firstly, we consider deriving the correct form of the aggregate-level likelihood from the individual-level likelihood (Section 4.2). We then describe the derivation of ML-NMR models when all covariates are discrete (Section 4.3.1), as considered by Jansen (2012). ML-NMR models are then derived for continuous covariates (Section 4.3.2), however analytic approaches are not always tractable. This motivates discussion of general numerical methods in Section 4.3.3, which are widely applicable and easily implementable. In Section 4.3.4, we demonstrate how models combining discrete and continuous covariates may be obtained from the previous results. We discuss some practical considerations in

Section 4.5, including identifiability and data availability. In Section 4.6, we show the natural extension of ML-NMR to larger treatment networks. Finally, we conclude with a discussion (Section 4.7).

General framework for ML-NMR 4.1

To develop the general framework for ML-NMR, we consider an individual-level regression model, which we can fit directly to the IPD in the AB study. The individual-level model also underlies the AgD AC study, however we do not have the IPD to fit this directly. Instead, we integrate the individual-level model over the AC study population to form an aggregate-level model with which to fit to the summary outcomes. Mathematically, we write

Individual:

$$y_{ik(AB)} \sim \pi_{\text{Ind}}(\theta_{ik(AB)}) \tag{4.1a}$$

$$g(\theta_{ik(AB)}) = \mu_{(AB)} + \boldsymbol{x}_{ik(AB)}^{\intercal}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k \tag{4.1b}$$

Aggregate:

$$y_{\bullet k(AC)} \sim \pi_{\text{Agg}}(\theta_{\bullet k(AC)}) \tag{4.1c}$$

$$\theta_{\bullet k(AC)} = \int_{\mathfrak{X}} g^{-1}\big(\mu_{(AC)} + \boldsymbol{x}^{\intercal}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k\big) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \tag{4.1d}$$

The individual- and aggregate-level data are given appropriate likelihood distributions $\pi_{\text{Ind}}(\cdot)$ and $\pi_{\text{Agg}}(\cdot)$, where the choice of individual-level likelihood will determine the corresponding aggregate-level likelihood (see Section 4.2). $\theta_{ik(AB)}$ is the conditional mean outcome for an individual i on treatment k in the AB trial with covariates $\boldsymbol{x}_{ik(AB)}$. $\theta_{\bullet k(AC)}$ is the marginal mean outcome on treatment k in the AC trial, and $f_{k(AC)}(\boldsymbol{x})$ is the distribution of $\boldsymbol{x}$ in those assigned to treatment k the AC trial. $g(\cdot)$ is a suitable link function, and $\mathfrak{X}$ denotes the support of $\boldsymbol{x}$. The coefficients μ are study-specific baselines, $\boldsymbol{\beta}_1$ are coefficients for prognostic variables, and $\boldsymbol{\beta}_{2,k}$ are coefficients for effect modifiers specific to each treatment k. The effect of the k-th treatment (at the individual level), γ_k, is defined with respect to the reference treatment A, and we set $\gamma_A = 0$ and $\boldsymbol{\beta}_{2,A} = \mathbf{0}$. Some coefficients of $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_{2,k}$ may be set to zero, if it is known that a particular covariate is not prognostic or effect modifying respectively. In a Bayesian analysis, prior distributions are placed over each of the parameters $\mu_{(AB)}$, $\mu_{(AC)}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_{2,B}$, $\boldsymbol{\beta}_{2,C}$, γ_B, γ_C. For brevity, we will frequently refer to the individual-level linear predictor for individuals on

treatment k in a population P with covariates $\boldsymbol{x}$ using the notation

$$\eta_{k(P)}(\boldsymbol{x}) = \mu_{(P)} + \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k,$$

so that the model (4.1) may be more succinctly written as

Individual:

$$\begin{aligned} y_{ik(AB)} &\sim \pi_{\text{Ind}}(\theta_{ik(AB)}) \\ g(\theta_{ik(AB)}) &= \eta_{k(AB)}(\boldsymbol{x}_{ik(AB)}) \end{aligned}$$

Aggregate:

$$\begin{aligned} y_{\bullet k(AC)} &\sim \pi_{\text{Agg}}(\theta_{\bullet k(AC)}) \\ \theta_{\bullet k(AC)} &= \int_{\mathfrak{X}} g^{-1}\big(\eta_{k(AC)}(\boldsymbol{x})\big) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \end{aligned}$$

We must address some key issues before the ML-NMR model (4.1) can be implemented. Firstly, we must derive an aggregate level likelihood (4.1c) from the individual level likelihood (4.1a), ensuring that the relationship between levels is maintained. For several individual level likelihoods the corresponding aggregate likelihood is intuitive, but this is not always the case. We discuss the proper derivation of aggregate level likelihoods in Section 4.2. Secondly, the integral in the aggregate level model (4.1d) must be evaluated; the manner of this evaluation depends on the type of covariates (discrete, continuous, or a mixture of both) and the link function $g(\cdot)$. Algebraic approaches are developed in Section 4.3.2, and numerical integration examined in Section 4.3.3.

4.2 From individual to aggregate likelihoods

The choice of an individual level likelihood distribution is natural, however care must be taken in subsequently defining the appropriate aggregate level likelihood distribution as a summary (e.g. a mean or summation) over individual likelihoods.

For example, if a continuous outcome is of interest with a Normal individual likelihood, summarised in the AgD by mean outcome, then the respective aggregate likelihood—the mean of the (independent) individual likelihoods—will also be Normal:

$$\text{Individual:} \qquad y_{ik(AB)} \sim \mathrm{N}\Big(\theta_{ik(AB)}, \sigma^2_{k(AB)}\Big) \tag{4.2a}$$

$$\text{Aggregate:} \qquad \bar{y}_{k(AC)} = N^{-1}_{k(AC)} \sum_i y_{ik(AC)} \sim \mathrm{N}\Big(\theta_{\bullet k(AC)}, s^2_{k(AC)}\Big), \tag{4.2b}$$

where $y_{ik(AB)}$ is the outcome for an individual i on treatment k in the AB trial, and $\bar{y}_{k(AC)}$ is the mean outcome on treatment k in the AC trial. The variance parameters $\sigma^2_{k(AB)}$ would be given prior distributions and estimated from the data in a Bayesian analysis, and the standard errors of the means $s_{k(AC)}$ would be given as data.

If the data are in the form of counts or numbers of events, then an individual Poisson likelihood may be used, the sum over which gives a Poisson aggregate likelihood:

$$\text{Individual:} \qquad y_{ik(AB)} \sim \text{Pois}\big(\lambda_{ik(AB)} E_{ik(AB)}\big) \tag{4.3a}$$

$$\text{Aggregate:} \qquad y_{\bullet k(AC)} = \sum_i y_{ik(AC)} \sim \text{Pois}\big(\bar{\lambda}_{k(AC)} E_{\bullet k(AC)}\big), \tag{4.3b}$$

where $y_{ik(AB)}$ is the number of events occurring in exposure time $E_{ik(AB)}$ for an individual i on treatment k in the AB trial, and $E_{\bullet k(AC)}$ is the total exposure time on treatment k in the AC trial. Note that the aggregate summary for the Poisson model is the total number of events $y_{\bullet k(AC)}$, as opposed to the mean outcome in the Normal model above. The parameters $\lambda_{ik(AB)}$ and $\bar{\lambda}_{k(AC)}$ are conditional and marginal event rates. These are modelled by $\theta_{ik(AB)}$ and $\theta_{\bullet k(AC)}$ respectively, typically with a log link function (see Section 4.3.2.2).

However, the situation is not so straightforward when binary data are observed. With a Bernoulli individual likelihood the respective aggregate likelihood is *not* Binomial, as in general each individual has a different event probability due to differences in prognostic factors. In this case the true aggregate likelihood, the sum of independent (but not identically distributed) Bernoulli likelihoods, is Poisson Binomial:

$$\text{Individual:} \qquad y_{ik(AB)} \sim \text{Bern}\big(p_{ik(AB)}\big) \tag{4.4a}$$

$$\text{Aggregate:} \qquad y_{\bullet k(AC)} = \sum_i y_{ik(AC)} \sim \text{PoBin}\big(\boldsymbol{p}_{k(AC)}\big), \tag{4.4b}$$

where $y_{ik(AB)}$ is a binary event indicator for individual i on treatment k in the AB study, and $y_{\bullet k(AC)}$ is the total number of events on treatment k in the AC study. Here, $p_{ik(AB)}$ is the probability of an event for an individual i on treatment k in the AB study, which is modelled by $\theta_{ik(AB)}$, typically using a logit link function. The parameter vector of the Poisson Binomial $\boldsymbol{p}_{k(AC)} = \big(p_{1k(AC)}, \ldots, p_{N_{k(AC)}(AC)}\big)$ is a vector of event probabilities for the $N_{k(AC)}$ individuals on treatment k in the AC study.

Aside from the lack of IPD in the AC trial to estimate the vector $\boldsymbol{p}_{k(AC)}$, there is a further issue which prevents us from using (4.4b) directly: evaluation of the Poisson Binomial likelihood is not straightforward. Exact evaluation involves

a recursive summation of factorially many terms, which is very slow and can be numerically unstable; the individual event probabilities are also required, which are not available due to the lack of IPD. More efficient approaches using the Discrete Fourier Transform have been proposed (Fernandez and Williams 2010; Hong 2013) which are fast and avoid numerical instabilities; however these still rely upon the availability of individual event probabilities. To circumvent these issues, we can use an approximation of the Poisson Binomial likelihood, which will not only lead to efficient computation, but will be tractable when only aggregate data are available.

4.2.1 Binomial approximations to the Poisson Binomial likelihood

The most common approximation (also used in standard aggregate data NMA) is to consider the number of events $y_{\bullet k(AC)}$ to be Binomially distributed with average probability $\bar{p}_{k(AC)}$,

$$y_{\bullet k(AC)} \sim \text{Bin}(N_{k(AC)}, \bar{p}_{k(AC)}). \tag{4.5}$$

In this case, the average probability is modelled using $\theta_{\bullet k(AC)}$, typically with a logit link function (see Section 4.3.2.4).

This approximation only works well when the individual event probabilities $p_{ik(AC)}$ are all approximately equal for a given treatment k (Ehm 1991). When this is not the case, the Binomial and Poisson Binomial share the same mean:

$$\begin{aligned}\text{Binomial mean} &= N_{k(AC)}\bar{p}_{k(AC)} \\ \text{Poisson Binomial mean} &= \sum_i p_{ik(AC)}\end{aligned}$$

and we have

$$\sum_i p_{ik(AC)} = N_{k(AC)}\frac{1}{N_{k(AC)}}\sum_i p_{ik(AC)} = N_{k(AC)}\bar{p}_{k(AC)},$$

but differ in their variance:

$$\begin{aligned}\text{Binomial variance} &= N_{k(AC)}\bar{p}_{k(AC)}(1-\bar{p}_{k(AC)}) \\ \text{Poisson Binomial variance} &= \sum_i p_{ik(AC)}(1-p_{ik(AC)})\end{aligned}$$

and it can be shown that

$$\sum_i p_{ik(AC)}(1-p_{ik(AC)}) \leq N_{k(AC)}\bar{p}_{k(AC)}(1-\bar{p}_{k(AC)}),$$

with equality if and only if $p_{ik(AC)} = \bar{p}_{k(AC)}\ \forall i$.

Le Cam (1960) considered an alternative Binomial approximation, whereby both $N_{k(AC)}$ and $\bar{p}_{k(AC)}$ are adjusted in order to match both the variance and the mean of the Poisson Binomial:

$$y_{\bullet k(AC)} \sim \text{Bin}(N'_{k(AC)}, \bar{p}'_{k(AC)}), \tag{4.6}$$

where

$$N'_{k(AC)} = \frac{\sum_i p_{ik(AC)}}{\bar{p}'_{k(AC)}} = N_{k(AC)} \frac{\bar{p}^2_{k(AC)}}{\bar{p^2}_{k(AC)}}$$
$$\bar{p}'_{k(AC)} = \frac{\sum_i p^2_{ik(AC)}}{\sum_i p_{ik(AC)}} = \frac{\bar{p^2}_{k(AC)}}{\bar{p}_{k(AC)}}$$

defining $\bar{p^2}_{k(AC)} = N^{-1}_{k(AC)} \sum_i p^2_{ik(AC)}$. The mean probability parameter $\bar{p}_{k(AC)}$ is again modelled with $\theta_{\bullet k(AC)}$ and a logit link. The second parameter $\bar{p^2}_{k(AC)}$ (which is related to the variance of the individual probabilities) is modelled in a similar manner, where the integration in (4.1d) is over the *squared* linear predictor. We discuss this in detail in Section 4.3.2.4.

Peköz et al. (2009) derive a shifted-Binomial approximation, which further matches the skewness of the Poisson Binomial by adjusting three parameters (the third being the shift parameter); however Peköz et al. (2010) find that in practice the largest improvements arise from the use of the two-parameter approximation over the simple approximation, with the three-parameter approximation giving little additional improvement at the expense of greater complexity and computational cost. We therefore do not explore any higher-order approximations beyond the two-parameter Binomial approximation.

Deriving the aggregate level model by integration 4.3

Implementing ML-NMR involves an integration step in the aggregate-level model (4.1d). In a Bayesian framework, ML-NMR is likely to be implemented using Markov Chain Monte Carlo (MCMC), for example in WinBUGS, JAGS, or Stan. At each iteration of the MCMC algorithm, we must evaluate the integral (4.1d) at a given set of values for the parameters $(\mu_{(AC)}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{2,k}, \gamma_k)$. There are two possible approaches: (i) analytical, where the exact form of the aggregate-level model is derived and can be explicitly written into the MCMC algorithm; or (ii) numerical, where a numerical integration method is enlisted. Analytical approaches are fast and offer insight into the nature of multilevel models and aggregation bias; however, different approaches are required for different link functions and covariate distributions, meaning that

a general implementation is not possible, and the resulting expressions for the aggregate-level model can grow increasingly complex, or may even be intractable. This motivates the exploration of numerical approaches, which sacrifice some speed and mathematical insight but may be broadly applied to fit ML-NMR models in myriad scenarios.

4.3.1 ML-NMR with discrete covariates

If the covariates $\boldsymbol{x}$ are all discrete, then the integration in the aggregate level model (4.1d) becomes a summation which is straightforward to perform:

$$\theta_{\bullet k(AC)} = \sum_{\mathfrak{x}} g^{-1}\big(\eta_{k(AC)}(\boldsymbol{x})\big) f_{k(AC)}(\boldsymbol{x}) \tag{4.7}$$

where now $\mathfrak{x}$ is the set of discrete levels of $\boldsymbol{x}$, and $f_{k(AC)}(\boldsymbol{x})$ is the proportion of individuals in each level in the AC trial. Jansen (2012) describes how a network meta-analysis combining IPD and AgD may be performed using (4.7) in a logistic regression scenario with only discrete covariates, and using the simple one-parameter Binomial approximation to the Poisson Binomial likelihood (4.5).

4.3.2 ML-NMR with continuous covariates: analytic approaches

When all covariates are continuous, the manner in which the aggregate level model is evaluated depends on the choice of link function $g(\cdot)$.

4.3.2.1 Integration with an identity link

When $g(y) = y$, the identity link, the integration is trivially the expectation over the linear predictor. By linearity of expectation, we have:

$$\begin{aligned}\theta_{\bullet k(AC)} &= \int_{\mathfrak{x}} \eta_{k(AC)}(\boldsymbol{x}) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \eta_{k(AC)}(\bar{\boldsymbol{x}}_{(AC)}).\end{aligned} \tag{4.8}$$

With any other non-linear link function, simply "plugging in" the mean covariate values in this manner will lead to aggregation bias since $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$.

4.3.2.2 Integration with a log link

When $g(\cdot)$ is the log link function, for example with a continuous outcome taking positive real values, the aggregate level model (4.1d) becomes

$$\theta_{\bullet k(AC)} = \int_{\mathfrak{x}} \exp(\eta_{k(AC)}(\boldsymbol{x})) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x}.$$

In this case, the integration may be performed explicitly by noting the definition of the *moment generating function* (MGF) for a (multivariate) random variable $\boldsymbol{x}$ (see Severini 2005, p. 109):

$$\begin{aligned} M_{\boldsymbol{x}(AC)}(\boldsymbol{s}) &= \mathbb{E}_{(AC)}\big(\exp(\boldsymbol{s}^\intercal \boldsymbol{x})\big) \\ &= \int_{\mathfrak{X}} \exp(\boldsymbol{s}^\intercal \boldsymbol{x}) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned} \tag{4.9}$$

Therefore, we rewrite the aggregate level model in terms of the joint MGF of $\boldsymbol{x}$ in the aggregate AC population:

$$\begin{aligned} \theta_{\bullet k(AC)} &= \int_{\mathfrak{X}} \exp\big(\mu_{(AC)} + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k\big) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \exp(\mu_{(AC)} + \gamma_k) \int_{\mathfrak{X}} \exp\big(\boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\big) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \exp(\mu_{(AC)} + \gamma_k) M_{\boldsymbol{x}(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}). \end{aligned} \tag{4.10}$$

Table 4.1 lists the MGFs for some common distributions likely to be encountered in practice.

Table 4.1 Moment generating functions for common distributions.

Distribution	Moment generating function	Sufficient statistics required
Normal $x \sim \mathrm{N}(m, \sigma^2)$	$M_{x(AC)}(s) = \exp(ms + \frac{1}{2}\sigma^2 s^2)$	Mean m Variance σ^2
Multivariate Normal $\boldsymbol{x} \sim \mathrm{MVN}(\boldsymbol{m}, \Sigma)$	$M_{\boldsymbol{x}(AC)}(\boldsymbol{s}) = \exp(\boldsymbol{s}^\intercal \boldsymbol{m} + \frac{1}{2}\boldsymbol{s}^\intercal \Sigma \boldsymbol{s})$	Mean vector $\boldsymbol{m}$ Covariance matrix Σ
Gamma $x \sim \mathrm{Gam}(a, b)$	$M_{x(AC)}(s) = (1 - s/b)^{-a}$	Mean a/b Variance a/b^2
Bernoulli $x \sim \mathrm{Bern}(p)$	$M_{x(AC)}(s) = 1 - p + p\exp(s)$	Proportion p

If the aggregate data are instead given as relative effect estimates—the difference in transformed mean outcomes on each treatment—we can simply derive the aggregate level model for relative effects data from (4.10):

$$\begin{aligned} d_{AC(AC)} &= \log(\theta_{\bullet C(AC)}) - \log(\theta_{\bullet A(AC)}) \\ &= \gamma_C + \log\left(\frac{M_{\boldsymbol{x}(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,C})}{M_{\boldsymbol{x}(AC)}(\boldsymbol{\beta}_1)}\right) \end{aligned} \tag{4.11}$$

The joint MGF $M_{\boldsymbol{x}(AC)}(\cdot)$ could be derived empirically if IPD were available on the covariates in the AC study using the estimator $\hat{M}_{\boldsymbol{x}(AC)}(\boldsymbol{s}) = N_{(AC)}^{-1} \sum_{i=1}^{N_{(AC)}} \exp(\boldsymbol{s}^\intercal \boldsymbol{x}_i)$. However, such level of detail is unlikely to be provided in published reports. Instead, if a specific covariate distribution is either

known or may be assumed, sufficient statistics can be used to construct the MGF. One practical consideration must be noted with this approach: not every distribution has a MGF, including most notably the log-Normal distribution. For skew covariates defined to be positive, which might otherwise be assumed to follow a log-Normal distribution, the Gamma distribution could be used instead (later in Section 8.2.6, simulations suggest that the assumed form of the covariate distribution may not greatly impact the results).

For example, with multivariate Normal covariates $\boldsymbol{x} \sim \text{MVN}(\boldsymbol{m}, \Sigma)$ with mean vector $\boldsymbol{m}$ and covariance matrix Σ, the aggregate level model (4.10) becomes

$$\theta_{\bullet k(AC)} = \exp\Big(\underbrace{\mu_{(AC)} + \gamma_k + \boldsymbol{m}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})}_{\text{(a)}} + \underbrace{\tfrac{1}{2}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^\intercal \Sigma (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})}_{\text{(b)}} \Big). \quad (4.12)$$

In this case, the aggregate level model for relative effects data (4.11) becomes

$$\begin{aligned} d_{AC(AC)} &= \log(\theta_{\bullet C(AC)}) - \log(\theta_{\bullet A(AC)}) \\ &= \mu_{(AC)} + \gamma_C + \boldsymbol{m}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,C}) + \tfrac{1}{2}\boldsymbol{\beta}_1^\intercal \Sigma \boldsymbol{\beta}_1 + \left(\boldsymbol{\beta}_1 + \tfrac{1}{2}\boldsymbol{\beta}_{2,C}\right)^\intercal \Sigma \boldsymbol{\beta}_{2,C} \\ &\quad - \left(\mu_{(AC)} + \boldsymbol{m}^\intercal \boldsymbol{\beta}_1 + \tfrac{1}{2}\boldsymbol{\beta}_1^\intercal \Sigma \boldsymbol{\beta}_1\right) \\ &= \underbrace{\gamma_C + \boldsymbol{m}^\intercal \boldsymbol{\beta}_{2,C}}_{\text{(a)}} + \underbrace{\left(\boldsymbol{\beta}_1 + \tfrac{1}{2}\boldsymbol{\beta}_{2,C}\right)^\intercal \Sigma \boldsymbol{\beta}_{2,C}}_{\text{(b)}}. \end{aligned} \quad (4.13)$$

The multivariate Normal scenario offers some insight into the relationship between ML-NMR and the naïve approach of "plugging in" mean covariate values. Both (4.12) and (4.13) can be viewed as an adjustment to the naïve model (4.12a, 4.13a), with the adjustment terms (4.12b, 4.13b) involving the covariate covariance matrix Σ accounting for the integration over the population. Examining the adjustment term (4.13b), we notice that this depends upon

- The covariance of prognostic variables and effect modifiers, and the strength of each, through $\boldsymbol{\beta}_1^\intercal \Sigma \boldsymbol{\beta}_{2,C}$. This is large when highly prognostic variables are correlated with strong effect modifiers, or when strong effect modifiers which are also highly prognostic take a wide range of values in the AC population (the population variance is large).
- The covariance of effect modifiers and their strength, through $\frac{1}{2}\boldsymbol{\beta}_{2,C}^\intercal \Sigma \boldsymbol{\beta}_{2,C}$. This is large when strong effect modifiers are highly correlated, or when strong effect modifiers take on a wide range of values in the AC population (the population variance is large).

This lends theoretical justification to intuitive thinking for when effect modification may lead to aggregation bias. In particular, aggregation bias is expected to be small if either effect modifiers are weak, or the within-study variance of effect modifiers is small. (The covariance inequality, a form of the Cauchy-Schwarz inequality, bounds the covariance between two variables as $\left|\text{cov}(X_1, X_2)\right| \leq \sqrt{\text{var}(X_1)\,\text{var}(X_2)}$. Small within-study variance in the effect modifiers is therefore sufficient to also limit the aggregation bias due to correlations.)

Integration with a probit link 4.3.2.3

Suppose that a binary outcome is of interest, and the probit link function $g(p) = \Phi^{-1}(p)$ (the standard Normal inverse cumulative distribution function) is used. We saw in Section 4.2.1 that, for a Bernoulli individual-level likelihood, the corresponding aggregate-level likelihood is Poisson Binomial, to which we make either a one- or two-parameter Binomial approximation.

One-parameter Binomial approximation To recap, the one-parameter approximation (4.5) is to consider the number of events $y_{\bullet k(AC)}$ to be Binomially distributed with some mean probability $\bar{p}_{k(AC)}$,

$$y_{\bullet k(AC)} \sim \text{Bin}\big(N_{k(AC)}, \bar{p}_{k(AC)}\big).$$

The mean probability $\bar{p}_{k(AC)}$ is modelled by $\theta_{\bullet k(AC)}$, and using the probit link the aggregate-level model becomes

$$\theta_{\bullet k(AC)} = \int_{\mathfrak{X}} \Phi\big(\eta_{k(AC)}(\boldsymbol{x})\big) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{4.14}$$

This integral does not always have a closed form solution; however, we can obtain a solution if the covariates are multivariate-Normal, $\boldsymbol{x}_{k(AC)} \sim \text{MVN}(\boldsymbol{m}_{k(AC)}, \Sigma_{k(AC)})$. Using the following result for the expectation of the probit of a Normal random variable $Z \sim \text{N}(m, \sigma^2)$ (see Theorem 2, Corollary 1 of Ellison 1964):

$$\mathbb{E}(\Phi(Z)) = \Phi\left(\frac{m}{\sqrt{1+\sigma^2}}\right), \tag{4.15}$$

and noting that the linear predictor $\eta_{k(AC)}(\boldsymbol{x}_{k(AC)})$ is a linear transformation of Normal random variables, so is itself Normally distributed:

$$\begin{aligned}
\eta(\boldsymbol{x}_{k(AC)}) &= \mu_{(AC)} + \boldsymbol{x}_{k(AC)}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k \\
&\sim \text{N}\Big(\mu_{(AC)} + \boldsymbol{m}_{k(AC)}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k, (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^{\mathsf{T}}\Sigma_{k(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\Big) \\
&\sim \text{N}\big(\eta(\boldsymbol{m}_{k(AC)}), (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^{\mathsf{T}}\Sigma_{k(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\big),
\end{aligned} \tag{4.16}$$

we have that (4.14) can be written as

$$\theta_{\bullet k(AC)} = \Phi\left(\eta_{k(AC)}(\boldsymbol{m}_{k(AC)})\left(1 + (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^{\intercal}\Sigma_{k(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\right). \quad (4.17)$$

As in Section 4.3.2.2, the ML-NMR aggregate-level model can be viewed as an adjustment to the naïve approach of "plugging in" mean covariate values to $\theta_{\bullet k(AC)} = \Phi(\eta_{k(AC)}(\boldsymbol{m}_{k(AC)}))$, and the adjustment factor again involves the strength of the prognostic and effect modifying covariates and their covariance matrix through $(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^{\intercal}\Sigma_{k(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})$.

Two-parameter Binomial approximation As noted in Section 4.2.1, the one-parameter Binomial approximation to the Poisson Binomial likelihood only works well when the individual event probabilities are approximately equal. The improved two-parameter Binomial approximation (4.6) is

$$y_{\bullet k(AC)} \sim \text{Bin}\left(N'_{k(AC)}, \bar{p}'_{k(AC)}\right),$$

where

$$N'_{k(AC)} = \frac{\sum_i p_{ik(AC)}}{\bar{p}'_{k(AC)}} = N_{k(AC)}\frac{\bar{p}^2_{k(AC)}}{\bar{p^2}_{k(AC)}}$$
$$\bar{p}'_{k(AC)} = \frac{\sum_i p^2_{ik(AC)}}{\sum_i p_{ik(AC)}} = \frac{\bar{p^2}_{k(AC)}}{\bar{p}_{k(AC)}}.$$

To derive the AgD model using the two-parameter Binomial approximation, we therefore need to model $\bar{p}_{k(AC)}$ and $\bar{p^2}_{k(AC)} = N^{-1}_{k(AC)}\sum_i p^2_{ik(AC)}$. We model $\bar{p}_{k(AC)}$ by $\theta_{\bullet k(AC)}$ as in the one-parameter case using (4.14), and we model $\bar{p^2}_{k(AC)}$ in an analogous manner as

$$\theta_{2\bullet k(AC)} = \int_{\mathfrak{x}} \Phi\left(\eta_{k(AC)}(\boldsymbol{x})\right)^2 f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x}. \quad (4.18)$$

Again, the necessary integrals are not generally analytically tractable, but we can obtain analytic results in the special case that $\boldsymbol{x}_{k(AC)} \sim \text{MVN}(\boldsymbol{m}_{k(AC)}, \Sigma_{k(AC)})$. We use the same result for $\theta_{\bullet k(AC)}$ obtained in (4.17). For $\theta_{2\bullet k(AC)}$, we again note the Normal distribution of $\eta_{k(AC)}(\boldsymbol{x}_{k(AC)})$ (4.16), and use the result of Owen (1980) for $Z \sim \text{N}(m, \sigma^2)$ that

$$\mathbb{E}(\Phi(Z)^2) = \Phi\left(\frac{m}{\sqrt{1+\sigma^2}}\right) - 2\,\text{T}\left(\frac{m}{\sqrt{1+\sigma^2}}, \frac{1}{\sqrt{1+2\sigma^2}}\right), \quad (4.19)$$

where $\mathrm{T}(\cdot,\cdot)$ is *Owen's T function* (Owen 1956), to obtain

$$\begin{aligned}\theta_{2\bullet k(AC)} = \Phi\Big(&\eta_{k(AC)}(\boldsymbol{m}_{(AC)})\left(1+(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})^{\mathsf{T}}\Sigma_{(AC)}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\Big)\\ &-2\,\mathrm{T}\Bigg(\eta_{k(AC)}(\boldsymbol{m}_{(AC)})\left(1+(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})^{\mathsf{T}}\Sigma_{(AC)}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}},\\ &\qquad\left(1+2(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})^{\mathsf{T}}\Sigma_{(AC)}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\Bigg)\end{aligned} \tag{4.20}$$

Owen's T may be evaluated using fast numerical methods: an efficient hybrid approach that selects between six alternate evaluation methods is given by Patefield and Tandy (2000), and an adapted version of this algorithm is implemented in the Stan function `owens_t`.

Integration with a logit link 4.3.2.4

Suppose again that a binary outcome is of interest, but that this time the logit link function is used. The corresponding aggregate-level likelihood is Poisson Binomial, to which we make either a one- or two-parameter Binomial approximation (Section 4.2.1).

One-parameter Binomial approximation The one-parameter approximation (4.5) is to consider the number of events $y_{\bullet k(AC)}$ to be Binomially distributed with some mean probability $\bar{p}_{k(AC)}$,

$$y_{\bullet k(AC)} \sim \mathrm{Bin}\big(N_{k(AC)}, \bar{p}_{k(AC)}\big).$$

The mean probability $\bar{p}_{k(AC)}$ is modelled by $\theta_{\bullet k(AC)}$, and using the logit link the aggregate-level model becomes

$$\theta_{\bullet k(AC)} = \int_{\mathfrak{X}} \mathrm{logit}^{-1}\big(\eta_{k(AC)}(\boldsymbol{x})\big)\, f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \tag{4.21}$$

Unfortunately, this integral has no closed form solution.

Previous authors have suggested approximating the logit link by a probit (Demidenko 2004; Salway and Wakefield 2005), for example

$$\text{One-probit approx.:}\;\; \mathrm{logit}^{-1}(\eta) \approx \Phi\Big(\tfrac{16\sqrt{3}}{15\pi}\eta\Big)$$

$$\text{Two-probit approx.:}\;\; \mathrm{logit}^{-1}(\eta) \approx 0.4353\,\Phi\Big(\frac{\eta}{2.2967}\Big) + 0.5647\,\Phi\Big(\frac{\eta}{1.3017}\Big)$$

The two-probit approximation is particularly robust, and still results in a proper likelihood distribution as the coefficients sum to one. Higher order

approximations are possible and can be even more accurate, but do not necessarily result in proper likelihood distributions.

Continuing with the one-probit approximation first, the average event probability on treatment k in the AC study is then approximately

$$\theta_{\bullet k(AC)} \approx \int_{\mathfrak{X}} \Phi\left(\tfrac{16\sqrt{3}}{15\pi}\eta_{k(AC)}(\boldsymbol{x})\right) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x}.$$

If the covariates are multivariate-Normal, $\boldsymbol{x}_{k(AC)} \sim \text{MVN}(\boldsymbol{m}_{k(AC)}, \Sigma_{k(AC)})$ then, following the same approach as Section 4.3.2.3 by noting the Normal distribution of $\eta_{k(AC)}(\boldsymbol{x}_{k(AC)})$ and using the result (4.15), we have

$$\theta_{\bullet k(AC)} \approx \text{logit}^{-1}\left(\eta_{k(AC)}(\boldsymbol{m}_{(AC)})\left(1 + \left(\tfrac{16\sqrt{3}}{15\pi}\right)^2 (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^{\intercal}\Sigma_{(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\right). \tag{4.22}$$

This result is given by Salway and Wakefield (2005) for bivariate Normal covariates. Simlarly to Section 4.3.2.2 and Section 4.3.2.3, this aggregate-level model can be viewed as an adjustment to the naïve approach of "plugging in" mean covariate values to $\theta_{\bullet k(AC)} = \text{logit}^{-1}(\eta_{k(AC)}(\boldsymbol{m}_{k(AC)}))$, and the adjustment factor again involves the strength of the prognostic and effect modifying covariates and their covariance matrix through $(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^{\intercal}\Sigma_{k(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})$.

With the two-probit approximation and multivariate-Normal $\boldsymbol{x}_{k(AC)}$, again noting the Normal distribution of $\eta_{k(AC)}(\boldsymbol{x}_{k(AC)})$ and using the result (4.15), we obtain

$$\begin{aligned}\theta_{\bullet k(AC)} \approx\ & 0.4353\,\Phi\left(\frac{\eta_{k(AC)}(\boldsymbol{m}_{k(AC)})}{2.2967}\left(1 + 2.2967^{-2}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^{\intercal}\Sigma_{k(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\right) \\ & + 0.5647\,\Phi\left(\frac{\eta_{k(AC)}(\boldsymbol{m}_{k(AC)})}{1.3017}\left(1 + 1.3017^{-2}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^{\intercal}\Sigma_{k(AC)}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\right)\end{aligned} \tag{4.23}$$

For skew covariates, as for models with a log link or probit link (see Section 4.3.2.2 and Section 4.3.2.3), evaluation of the marginalisation integral is troublesome: analytic integration for both log-Normal and Gamma distributed covariates is not straightforward. In an ecological context, Salway and Wakefield (2005) noted that estimates remained biased when skew covariates were treated as Normal (although less so than the naïve approach), and uncertainty was underestimated. For discrete covariates, the aggregation integral is simply a sum over the levels of the covariates, and no approximation is required (see Section 4.3.4).

Two-parameter Binomial approximation As noted previously (Section 4.2.1), the one-parameter Binomial approximation to the Poisson Binomial likelihood

only works well when the individual event probabilities are approximately equal. The improved two-parameter Binomial approximation (4.6) is

$$y_{\bullet k(AC)} \sim \text{Bin}\Big(N'_{k(AC)}, \bar{p}'_{k(AC)}\Big),$$

where

$$N'_{k(AC)} = \frac{\sum_i p_{ik(AC)}}{\bar{p}'_{k(AC)}} = N_{k(AC)} \frac{\bar{p}^2_{k(AC)}}{\bar{p^2}_{k(AC)}}$$
$$\bar{p}'_{k(AC)} = \frac{\sum_i p^2_{ik(AC)}}{\sum_i p_{ik(AC)}} = \frac{\bar{p^2}_{k(AC)}}{\bar{p}_{k(AC)}}.$$

To derive the AgD model using the two-parameter Binomial approximation, we therefore need to model $\bar{p}_{k(AC)}$ and $\bar{p^2}_{k(AC)} = N^{-1}_{k(AC)} \sum_i p^2_{ik(AC)}$. We model $\bar{p}_{k(AC)}$ by $\theta_{\bullet k(AC)}$ as in the one-parameter case using (4.21), and we model $\bar{p^2}_{k(AC)}$ in an analogous manner as

$$\theta_{2\bullet k(AC)} = \int_{\mathfrak{X}} \text{logit}^{-1}\big(\eta_{k(AC)}(\boldsymbol{x})\big)^2 f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{4.24}$$

Again, the necessary integrals are not generally analytically tractable, but we can obtain analytic results in the special case that $\boldsymbol{x}_{k(AC)} \sim \text{MVN}(\boldsymbol{m}_{k(AC)}, \Sigma_{k(AC)})$ using the one-probit approximation to the logit. We use the same result for $\theta_{\bullet k(AC)}$ obtained in (4.22). For $\theta_{2\bullet k(AC)}$, we again note the Normal distribution of $\eta_{k(AC)}(\boldsymbol{x}_{k(AC)})$ (4.16), and use the result (4.19) to obtain

$$\begin{aligned}
\theta_{2\bullet k(AC)} &= \int_{\mathfrak{X}} \text{logit}^{-1}(\eta_{k(AC)}(\boldsymbol{x}))^2 f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \\
&\approx \int_{\mathfrak{X}} \Phi\Big(\tfrac{16\sqrt{3}}{15\pi}\eta_{k(AC)}(\boldsymbol{x})\Big)^2 f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \Phi\left(\tfrac{16\sqrt{3}}{15\pi}\eta_{k(AC)}(\boldsymbol{m}_{(AC)})\left(1 + \left(\tfrac{16\sqrt{3}}{15\pi}\right)^2 (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^\intercal \Sigma_{(AC)} (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\right) \\
&\quad - 2\,\text{T}\left(\tfrac{16\sqrt{3}}{15\pi}\eta_{k(AC)}(\boldsymbol{m}_{(AC)})\left(1 + \left(\tfrac{16\sqrt{3}}{15\pi}\right)^2 (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^\intercal \Sigma_{(AC)} (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}},\right. \\
&\qquad \left.\left(1 + 2\left(\tfrac{16\sqrt{3}}{15\pi}\right)^2 (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})^\intercal \Sigma_{(AC)} (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\right)
\end{aligned}$$

$$\approx \text{logit}^{-1}\left(\eta_{k(AC)}(\boldsymbol{m}_{(AC)})\left(1+\left(\tfrac{16\sqrt{3}}{15\pi}\right)^2(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})^{\mathsf{T}}\Sigma_{(AC)}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\right)$$
$$-2\,\mathrm{T}\left(\tfrac{16\sqrt{3}}{15\pi}\eta_{k(AC)}(\boldsymbol{m}_{(AC)})\left(1+\left(\tfrac{16\sqrt{3}}{15\pi}\right)^2(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})^{\mathsf{T}}\Sigma_{(AC)}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}},$$
$$\left(1+2\left(\tfrac{16\sqrt{3}}{15\pi}\right)^2(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})^{\mathsf{T}}\Sigma_{(AC)}(\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2,k})\right)^{-\frac{1}{2}}\Bigg) \tag{4.25}$$

Derivation of the aggregate-level model for multivariate-Normal covariates using both the two-probit and two-parameter Binomial approximations together is not straightforward: the integral for $\theta_{\bullet k(AC)}$ may be evaluated in a similar fashion to the one-probit approximation described here, but the integral for the average squared probabilities modelled by $\theta_{2\bullet k(AC)}$ becomes intractable.

4.3.3 ML-NMR with continuous covariates: generalised numerical approaches

Thus far we have considered analytic approaches to performing the marginalisation integral (4.1d). The resulting analytic forms of the aggregate level model enable efficient computation, and lead to greater insight into the mathematical nature of the multilevel model—specifically that the ML-NMR aggregate level model may be viewed as an adjustment to the naïve aggregate level model. However, such approaches are context-specific, requiring a different approach for each link function, and are of little use when the marginalisation integral becomes intractable (most notably when covariates are skewed). There is, therefore, the need for a general numerical approach which is both flexible and robust enough to be widely applicable.

Practically, the ML-NMR model will be implemented using Markov Chain Monte Carlo (MCMC) for flexibility and simplicity (such as WinBUGS, JAGS, or Stan). The role of numerical integration is therefore to evaluate the marginalisation integral (4.1d) for a given value of the parameters $(\mu_{(AC)}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{2,k}, \gamma_k)$ at each iteration of the MCMC chain. We discuss several possible numerical approaches, along with their merits and disadvantages, below.

4.3.3.1 Quadrature

One potential approach is to use a quadrature rule. For example, if the covariate distribution is Normal, Gauss-Hermite quadrature may be used to efficiently and accurately evaluate $\theta_{\bullet k(AC)}$ for a given set of parameter values

at each MCMC iteration. Alternatively, more general quadrature methods such as the trapezium rule, Simpson's rule, or Romberg integration (see Press et al. 2007, Chapter 4, for an overview), may be utilised. Whilst these methods are fast and accurate, they cannot handle higher dimensional integrals (i.e. with larger numbers of covariates) well: the number of quadrature points required increases to the power of the number of dimensions. The appropriate quadrature rule to use will also depend on the specific form of $f_{k(AC)}(\cdot)$ or $g(\cdot)$.

Monte Carlo integration 4.3.3.2

In contrast to quadrature rules which are *deterministic* algorithms, we might instead consider forms of *stochastic* numerical integration, which involve random sampling from the joint covariate distribution $f_{k(AC)}(\cdot)$. Here, we consider Monte Carlo integration. A large number $\tilde{N}$ of integration points $\tilde{\boldsymbol{x}}_{k(AC)}$ are simulated from $f_{k(AC)}(\cdot)$ before the MCMC run, and treated as additional data for the MCMC algorithm. At each MCMC iteration, the marginalisation integral is then estimated by a summation:

$$\begin{aligned}\theta_{\bullet k(AC)} &= \int_{\mathfrak{X}} g^{-1}(\eta_{k(AC)}(\boldsymbol{x})) f_{k(AC)}(\boldsymbol{x})\, d\boldsymbol{x} \\ &\approx \frac{1}{\tilde{N}} \sum g^{-1}(\eta_{k(AC)}(\tilde{\boldsymbol{x}})) \end{aligned} \tag{4.26}$$

For standard Monte Carlo integration, $\tilde{\boldsymbol{x}}_{k(AC)}$ are sampled (pseudo-) randomly from the covariate distribution $f_{k(AC)}(\cdot)$; however, this can require large numbers of integration points to obtain sufficiently accurate estimates. The approximation can be made more efficient by using a quasi-random sample of points (Quasi Monte Carlo, QMC), which are chosen sequentially to cover the covariate space $\mathfrak{X}$ more uniformly than pseudo-random samples, resulting in fewer integration points needed to achieve the same accuracy. QMC integration commonly achieves integration error rates of the order $\tilde{N}^{-1}$, compared to standard Monte Carlo which only achieves $\tilde{N}^{-\frac{1}{2}}$ (Caflisch 1998; Niederreiter 1978). In other words, when using 10,000 integration points drawn from a quasi-random sequence we would expect accuracy down to around 4 decimal places, compared with an accuracy of 2 decimal places using standard Monte Carlo integration.

A numerical integration approach based on QMC is more widely applicable than one based on quadrature (Section 4.3.3.1), since it can be applied regardless of the number of covariates or their distributions, the type of link function, or the complexity of the model. Moreover, QMC integration accounts for correlations between covariates by design since $\tilde{\boldsymbol{x}}_{k(AC)}$ are drawn from the

joint covariate distribution; accounting for correlations between covariates is not straightforward when using quadrature. We discuss the practical implementation of QMC integration in greater detail later in section Section 5.1.

4.3.4 Combining discrete and continuous covariates

In Section 4.3.1, we saw that if all covariates are discrete, the integration in the aggregate level model (4.1d) becomes a straightforward summation. Sections 4.3.2 and 4.3.3 presented methods for deriving the aggregate level model when all covariates are continuous, using either analytic or numerical approaches.

When both continuous and discrete covariates are to be included in the model, we may write the aggregate level model as an expansion over the levels of the discrete covariates using the Laws of Total Probability and Expectation. Dichotomising for a moment the covariates into discrete z and continuous $\boldsymbol{x}$, with support $\mathfrak{Z}$ and $\mathfrak{X}$ respectively, we have

$$\begin{aligned}
\theta_{\bullet k(AC)} &= \int_{\mathfrak{Z}} \int_{\mathfrak{X}} g^{-1}\left(\eta_{k(AC)}(\boldsymbol{x}, z)\right) f_{k(AC)}(\boldsymbol{x}, z)\, d\boldsymbol{x}\, dz \\
&= \int_{\mathfrak{Z}} \int_{\mathfrak{X}} g^{-1}\left(\eta_{k(AC)}(\boldsymbol{x}, z)\right) f_{k(AC)}(\boldsymbol{x}|z) f_{k(AC)}(z)\, d\boldsymbol{x}\, dz \\
&= \sum_{\mathfrak{Z}} \left[\int_{\mathfrak{X}} g^{-1}\left(\eta_{k(AC)}(\boldsymbol{x}, z)\right) f_{k(AC)}(\boldsymbol{x}|z)\, d\boldsymbol{x} \right] f_{k(AC)}(z). \qquad (4.27)
\end{aligned}$$

The inner integration of (4.27) is then performed in the manner described in the preceding sections.

Alternatively, when using QMC integration the discrete and continuous covariates may be jointly simulated from $f_{k(AC)}(\cdot)$. (This is made possible by the copula approach described in Section 4.3.3.2, and later in greater detail in Section 5.1.) The numerical integration (4.26) is then a summation over both the continuous and discrete covariates. Although this has the potential to increase the number of integration points required, particularly when there are high correlations between covariates and many discrete covariates, in practice we have not observed this to be an issue. Furthermore, implementation in code is greatly simplified and made more general, since only a single summation rather than the full expansion (4.27) is performed.

4.4 Producing estimates for a specific target population

Once the ML-NMR model has been specified and fitted to the data, we can produce estimates of different quantities of interest. Typically in a technology

appraisal context the quantities of interest are population-level quantities such as average relative treatment effects (contrasts) or the predicted proportion of individuals with a binary outcome, rather than individual-level treatment effects and interactions. To be relevant for decision making, population-adjusted estimates must be produced for the relevant target population. ML-NMR can produce population-adjusted estimates for any target population for which covariate information is available.

To produce estimates of population-average relative effects $d_{ab(P)}$ between any two treatments a and b in a population P, we can simply "plug in" the mean covariate values $\bar{\boldsymbol{x}}_{(P)}$ in population P to the linear predictor for each treatment like so:

$$\begin{aligned} d_{ab(P)} &= \eta_{b(P)}\big(\bar{\boldsymbol{x}}_{(P)}\big) - \eta_{a(P)}\big(\bar{\boldsymbol{x}}_{(P)}\big) \\ &= \bar{\boldsymbol{x}}_{(P)}^{\mathsf{T}}(\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}) + \gamma_b - \gamma_a. \end{aligned} \tag{4.28}$$

In a Bayesian framework using MCMC, we evaluate this formula at each of the posterior samples of the parameters $\boldsymbol{\beta}_{2,a}$, $\boldsymbol{\beta}_{2,b}$, γ_a, and γ_b to produce posterior samples for $d_{ab(P)}$.

In general, we consider estimating the average $\bar{h}_{(P)}$ of a quantity $h(\boldsymbol{x}, \boldsymbol{\xi})$ in a target population P, which is some function of the covariates $\boldsymbol{x}$ and a set $\boldsymbol{\xi}$ of the parameters $\mu_{(P)}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_{2,k}$, γ_k. Unless $h(\boldsymbol{x}, \boldsymbol{\xi})$ is linear in $\boldsymbol{x}$, like the population-average relative effects $d_{ab(P)}$ in (4.28) above, we cannot simply "plug in" mean covariate values to produce estimates. Doing so incurs aggregation bias since $\mathbb{E}_{(P)}\big(h(\boldsymbol{x}, \boldsymbol{\xi})\big) \neq h\big(\mathbb{E}_{(P)}(\boldsymbol{x}), \boldsymbol{\xi}\big)$.

Instead, given the joint covariate distribution $f_{(P)}(\cdot)$ in the target population, $\bar{h}_{(P)}$ is obtained by integrating over the joint covariate distribution (as a generalisation of equation (4.1d)):

$$\bar{h}_{(P)} = \int_{\mathfrak{X}} h(\boldsymbol{x}, \boldsymbol{\xi}) f_{(P)}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{4.29}$$

Alternatively, if the target population is represented by a study or registry with individual covariate information available, $\bar{h}_{(P)}$ may be obtained by taking the average of $h(\boldsymbol{x}, \boldsymbol{\xi})$ applied to each individual in P:

$$\bar{h}_{(P)} = \frac{1}{N_P} \sum_{i=1}^{N_P} h(\boldsymbol{x}_{i(P)}, \boldsymbol{\xi}), \tag{4.30}$$

where N_P is the number of individuals in the sample, each with covariate information $\boldsymbol{x}_{i(P)}$. Another option in this case is to estimate $f_{(P)}(\cdot)$ using the individual covariate information, and use this in (4.29). This may be a more robust approach, particularly if the sample size in the target population is

small; comparing these approaches, perhaps through a simulation study, is an area for further research.

In a Bayesian framework using MCMC, equation (4.29) or (4.30) is evaluated at each posterior sample to produce posterior samples for $\bar{h}_{(P)}$.

For example, to calculate the average absolute response on each treatment k on the natural outcome scale (or the predicted proportion of events for a binary outcome), we use

$$\begin{aligned} h(\boldsymbol{x}, \mu_{(P)}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{2,k}, \gamma_k) &= g^{-1}\big(\eta_{k(P)}(\boldsymbol{x})\big) \\ &= g^{-1}\big(\mu_{(P)} + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k\big). \end{aligned} \tag{4.31}$$

The baseline response $\mu_{(P)}$ in population P may be equal to $\mu_{(AB)}$ or $\mu_{(AC)}$ if P is represented by the AB or AC study respectively, or may be estimated from external data on population P.

Equation (4.28) for calculating the population-average relative effects $d_{ab(P)}$ between any two treatments b and a is obtained using

$$h(\boldsymbol{x}, \boldsymbol{\beta}_{2,a}, \boldsymbol{\beta}_{2,b}, \gamma_a, \gamma_b) = \boldsymbol{x}^\intercal(\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}) + \gamma_b - \gamma_a. \tag{4.32}$$

In this case, since $h(\boldsymbol{x}, \boldsymbol{\xi})$ is linear in $\boldsymbol{x}$, using (4.32) in either (4.29) or (4.30) reduces to (4.28), which is simply $h(\bar{\boldsymbol{x}}_{(P)}, \boldsymbol{\xi})$.

4.5 Practical considerations

4.5.1 Using published marginal covariate information

When using ML-NMR in practice, we typically only have access to limited covariate information from the AgD study. This is often in the form of published marginal covariate summaries, and we have no information on the correlation structure between the covariates or on the true distributional form of the marginal distributions. Where the true marginal distributions of the covariates in the AgD study are unknown, we can instead choose distributional forms for these covariates based on their theoretical properties and/or to approximately match the observed distributional forms in the IPD. For example, a covariate may be well-known to be approximately log-Normal and/or observed as such in the IPD study, and we can then assume a marginal distribution for this covariate in the AgD study to match the published summary statistics. To account for the missing correlation structure, rather than assuming that all correlations are zero (which may be unreasonable), we can utilise the correlation structure observed amongst the covariates in the IPD study and assume that this holds for the covariates in the AgD study.

It seems reasonable to assume that, whilst the marginal summaries for each covariate may change from study to study (e.g. proportion of males, or mean and standard deviation of weight), the relationships between covariates are likely to remain similar (e.g. duration of disease is positively correlated with the number of previous lines of treatment).

We have specified a model where the covariate distribution $f_{k(AC)}$ in the AgD AC study may be different in each arm. Whilst randomisation ensures that covariates are balanced across arms within a study on average, there are likely to be chance imbalances in the covariate distributions between arms. Allowing for distinct covariate distributions $f_{k(AC)}$ in each arm of the AC trial therefore allows us to account for these chance imbalances. Reporting baseline covariate distributions by arm is not uncommon for published trials—indeed, it has been an explicit requirement of the CONSORT statement since its first revision in 2001 (Moher et al. 2001; Schulz et al. 2010); however, if only the overall baseline covariate distribution $f_{(AC)}$ is reported this may be used instead, possibly at the expense of some bias.

The Monte Carlo integration approach outlined in Section 4.3.3.2 allows the correlation structure for the AC study to be specified in a straightforward manner. When producing the integration points $\tilde{x}_{k(AC)}$, we can impose the assumed joint covariate distribution by using a Gaussian copula (Nelsen 2006), so that the correlation matrix of the covariates in the AC study matches that of the IPD AB study, whilst retaining the given marginal distributions. We describe QMC integration with copulæ in detail in Section 5.1.

Identifiability in small networks 4.5.2

In the small two-study scenario with an AB and an AC trial, model (4.1) is not identifiable due to the lack of data: it is not possible to estimate both $\boldsymbol{\beta}_{2,C}$ and γ_C from a single data point (the C arm of the AC trial, as we do not have IPD for the AC trial). There are several possible solutions, which we discuss briefly.

Firstly, the shared effect modifier assumption (Section 2.5) may be used if justifiable (typically on biological/clinical grounds, for example if B and C are from the same class of treatments), so that $\boldsymbol{\beta}_{2,B} = \boldsymbol{\beta}_{2,C}$.

Alternatively, if the shared EM assumption is not justifiable, the mean outcomes on treatments A and B may be predicted in the AC trial and compared directly with the observed A and C outcomes without constructing an aggregate level model for the AC trial as follows:

Individual model in AB trial:

$$\begin{aligned} y_{ik(AB)} &\sim \pi_{\text{Ind}}(\theta_{ik(AB)}) \\ g(\theta_{ik(AB)}) &= \eta_{k(AB)}(\boldsymbol{x}_{ik(AB)}) \end{aligned} \tag{4.33a}$$

Predicted relative effect in AC trial:

$$\hat{d}_{AB(AC)} = \bar{\boldsymbol{x}}_{(AC)}^{\intercal}(\hat{\boldsymbol{\beta}}_{2,B} - \hat{\boldsymbol{\beta}}_{2,A}) + \hat{\gamma}_B - \hat{\gamma}_A \tag{4.33b}$$

Targeted comparison in AC trial:

$$\hat{d}_{BC(AC)} = \hat{d}_{AC(AC)} - \hat{d}_{AB(AC)} \tag{4.33c}$$

where (4.33b) is a result of equation (4.28), and $\hat{d}_{AC(AC)}$ is estimated by the AC trial. This is conceptually similar to STC (except undertaken in a Bayesian framework and with aggregation bias explicitly accounted for rather than using simulation), and should be considered as a "targeted comparison" rather than a true evidence synthesis, as information from the AC trial is not propagated through to the parameter estimates (although in practice this is unlikely to make any meaningful difference, as the aggregate summaries from a single trial do not contain much information about the individual level parameters). The main disadvantage of this approach is that the indirect comparison must be made in the AC population, which is unlikely to match the decision target population.

Other approaches are possible if more data are available. For example, if external data are available on the strength of effect modification, it may be possible to construct an informative prior distribution for $\boldsymbol{\beta}_{2,C}$; however note that the prior information is unlikely to be updated—the posterior distribution for $\boldsymbol{\beta}_{2,C}$ will be the same as the prior distribution—due to a lack of information in the data. In larger treatment networks, greater flexibility in estimation is possible (see Section 4.6).

4.6 Extension to larger treatment networks

Thus far we have described ML-NMR in a simple two study setting, motivated by comparison with methods such as MAIC and STC. However, application to larger networks of treatments is straightforward, and is simply a matter of notation. Where we have used subscript parentheses to denote studies and populations, we now refer to studies indexed numerically with index j. We write the model using the reference treatment parameterisation (Section 1.2.2), so that individual-level treatment effects γ_{1k} and effect modifier coefficients

$\boldsymbol{\beta}_{2,1k}$ refer to the relative effect of treatment k against treatment 1, and by convention we drop the subscript 1 for brevity: $\gamma_k = \gamma_{1k}$ and $\boldsymbol{\beta}_{2,k} = \boldsymbol{\beta}_{2,1k}$. This generalises (4.1) as follows:

Individual:

$$y_{ijk} \sim \pi_{\text{Ind}}(\theta_{ijk}) \tag{4.34a}$$

$$g(\theta_{ijk}) = \eta_{jk}(\boldsymbol{x}_{ijk}) = \mu_j + \boldsymbol{x}_{ijk}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k \tag{4.34b}$$

Aggregate:

$$y_{\bullet jk} \sim \pi_{\text{Agg}}(\theta_{\bullet jk}) \tag{4.34c}$$

$$\theta_{\bullet jk} = \int_{\mathfrak{X}} g^{-1}\big(\eta_{jk}(\boldsymbol{x})\big) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \tag{4.34d}$$

Again we set $\gamma_1 = 0$, $\boldsymbol{\beta}_{2,1} = \mathbf{0}$.

Here we have written a fixed effect (FE) model with no heterogeneity in treatment effects γ_k between studies, due to the conditional constancy of relative effects assumption which supposes that there are no unobserved effect modifiers. We consider a random effects model in Section 4.6.2. Using this notation allows ML-NMR to be applied to networks of any number of trials and treatments. This notation also highlights how ML-NMR is a generalisation of standard network meta-analysis models. When all studies have IPD available, equations (4.34a) and (4.34b) describe a typical FE IPD NMR. When no covariates are included, ML-NMR reduces to a standard FE AgD NMA: (4.34d) becomes $\theta_{\bullet jk} = g^{-1}\big(\mu_j + \gamma_k\big)$, and the model is fitted to the summary outcomes from each trial:

$$\begin{aligned} y_{\bullet jk} &\sim \pi_{\text{Agg}}(\theta_{\bullet jk}) \\ g(\theta_{\bullet jk}) &= \mu_j + \gamma_k \end{aligned} \tag{4.35}$$

and γ_k are equal to d_k, the aggregate-level relative effects.

The AgD NMA framework defines the consistency equations $d_{ab} = d_b - d_a$ for any two treatments a and b (see Section 1.2.7; Higgins and Whitehead 1996; Lu and Ades 2006). These relate the (population-level) relative effects between each pair of treatments in the network, and ensure that the resulting set of relative effect estimates is consistent. However, in the ML-NMR framework the population-level relative effects are not assumed constant between populations as the constancy of relative effects assumption is relaxed. Instead, ML-NMR defines consistency equations at the *individual* level on the individual-level treatment effects and effect modifier coefficients:

$$\begin{aligned} \gamma_{ab} &= \gamma_b - \gamma_a \\ \boldsymbol{\beta}_{2,ab} &= \boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a} \end{aligned} \tag{4.36}$$

Notably, this implies consistency of the population-average relative effects within each population, $d_{ab(P)} = d_{b(P)} - d_{a(P)}$:

$$\begin{aligned} d_{ab(P)} &= \bar{\boldsymbol{x}}_{(P)}^{\mathsf{T}} \boldsymbol{\beta}_{2,ab} + \gamma_{ab} \\ &= \bar{\boldsymbol{x}}_{(P)}^{\mathsf{T}} (\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}) + \gamma_b - \gamma_a \\ &= (\bar{\boldsymbol{x}}_{(P)}^{\mathsf{T}} \boldsymbol{\beta}_{2,b} + \gamma_b) - (\bar{\boldsymbol{x}}_{(P)}^{\mathsf{T}} \boldsymbol{\beta}_{2,a} + \gamma_a) \\ &= d_{b(P)} - d_{a(P)} \end{aligned} \tag{4.37}$$

following the same reasoning in Section 4.4, where integration over the covariate distribution in population P reduces to plugging in mean covariate values $\bar{\boldsymbol{x}}_{(P)}$ when working on the linear predictor scale. These consistency relationships do *not* in general hold across populations. Standard AgD NMA is a special case where it is assumed that there is no effect modification (or that effect modifiers are balanced between all studies), so the population-average relative effects are identical between studies and the usual NMA consistency equations $d_{ab} = d_b - d_a$ are recovered.

Larger networks also offer the opportunity to assess and, if necessary, relax modelling assumptions given sufficient data. In the two study scenario we are forced to make the shared effect modifier assumption in order to identify the model, or otherwise we are limited to a targeted comparison in the AgD population (see Section 4.5.2). With a larger network it may be possible to relax the shared effect modifier assumption (Section 4.6.1), or to assess residual heterogeneity or inconsistency (Sections 4.6.2 and 4.6.3).

4.6.1 Relaxing the shared effect modifier assumption

Within the ML-NMR framework, the shared effect modifier assumption (Section 2.5) asserts that the regression coefficients $\boldsymbol{\beta}_{2,k}$ for a set of treatments $k \in \mathcal{T}$ are all equal. As discussed in Section 4.5.2, this assumption is necessary to identify the model in the two study scenario (one IPD AB study and one AgD AC study), where we assume that $\{B, C\} = \mathcal{T}$.

In a larger treatment network, the data requirements for estimating a treatment effect and independent EM interaction terms for a given treatment k are either i) IPD from one or more trials including treatment k, or ii) sufficiently many AgD studies including treatment k, with enough variation in covariate values (equivalent to the requirements of a standard AgD meta-regression). In the case that neither of these requirements can be met, either informative prior distributions or additional assumptions are required to estimate the model.

As in the two study scenario, the shared EM assumption can make the model estimable by assuming that the effect modifier coefficients $\boldsymbol{\beta}_{2,k}$ are

identical amongst a set or class of treatments $\mathcal{T}$. The data requirements described above then apply to the set $\mathcal{T}$ as a whole, rather than to each individual treatment.

To relax the shared EM assumption for a set of treatments $\mathcal{T}$, we can instead assume that the EM interactions are exchangeable. A hierarchical model is placed on the EM interaction coefficients

$$\beta_{2,k;l} \sim \mathrm{N}\left(m_{\beta_2,l}, \sigma^2_{\beta_2,l}\right) \tag{4.38}$$

for $k \in \mathcal{T}$ and each coefficient $\beta_{2,k;l}$ in the vector $\boldsymbol{\beta}_{2,k}$, where l indexes the different covariates. The hyperparameters $m_{\beta_2,l}$ and $\sigma^2_{\beta_2,l}$ are themselves given prior distributions in a Bayesian framework.

There may be multiple (mutually exclusive) treatment sets $\mathcal{T}_1, \mathcal{T}_2, \ldots$ within the treatment network, and some of these may only contain a single treatment on its own. Different assumptions may be made within each set if required. For example, the shared EM assumption may be made in some sets, exchangeable EM interactions may be estimated in others (with separate variance components $\sigma^2_{\beta_2,l,\mathcal{T}_1}$, $\sigma^2_{\beta_2,l,\mathcal{T}_2}$, etc. in each set), and other treatments and their EM interactions may be estimated independently.

In practice, the data requirements for estimating exchangeable EM interactions (4.38) may also be beyond the data available. In particular, the estimation of the variance components $\boldsymbol{\sigma}^2_{\beta_2}$ is challenging, and only improves with increasing numbers of treatments in $\mathcal{T}$ and/or informative prior distributions. If there are multiple treatment sets within the network for which the shared EM assumption is made, then the variance components could be assumed equal between the treatment sets to further aid estimation, $\sigma^2_{\beta_2,l,\mathcal{T}_1} = \sigma^2_{\beta_2,l,\mathcal{T}_2} = \sigma^2_{\beta_2,l}$ for each covariate l. Another simplifying assumption which may aid estimation is to assume that the variance components are equal between covariates, $\sigma^2_{\beta_2,l} = \sigma^2_{\beta_2}$. However, this latter assumption is much harder to justify, and additionally relies on covariates being suitably scaled relative to each other so that the magnitude of their effect modification is similar. Estimation of the treatment-specific interactions $\boldsymbol{\beta}_{2,k}$ also requires either IPD or a sufficient number of AgD studies on each treatment in $\mathcal{T}$, particularly when $\boldsymbol{\sigma}^2_{\beta_2}$ is large or imprecisely estimated and little information is shared between treatments in $\mathcal{T}$.

Assessing residual heterogeneity 4.6.2

The key assumption of population adjustment methods is that relative effects are constant given the effect modifiers adjusted for (conditional constancy

of relative effects). Within the ML-NMR framework (4.34), this assumption corresponds to the individual-level relative effects γ_k being constant across populations—in other words, we fit a fixed effect model. In the two study scenario this is an untestable assumption. However, with a larger network of studies and treatments it is possible to assess the conditional constancy of relative effects assumption by fitting a random effects model. This is achieved by modifying the linear predictor $\eta_{jk}(\boldsymbol{x})$ in equation (4.34) as follows:

Fixed Effect:

$$\eta_{jk}(\boldsymbol{x}) = \mu_j + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \delta_{jk}, \qquad \delta_{jk} = \gamma_k \tag{4.39}$$

Random Effects:

$$\eta_{jk}(\boldsymbol{x}) = \mu_j + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \delta_{jk} \tag{4.40a}$$

$$\delta_{jk} \sim \mathrm{N}(\gamma_k, \tau^2) \tag{4.40b}$$

$$\mathrm{cor}(\delta_{ja}, \delta_{jb}) = 0.5 \tag{4.40c}$$

where $\delta_{j1} = 0$. Using the reference treatment parameterisation (Section 1.2.2), the random effects are multivariate Normal with marginal distributions given by (4.40b) and correlations equal to 0.5 (4.40c), under the assumption of common heterogeneity variance τ^2 (Higgins and Whitehead 1996). (Given sufficient data, separate τ_k^2 for each treatment and correlations $\mathrm{cor}(\delta_{ja}, \delta_{jb}) = \psi_{ab}^{(1)}$ may be fitted with an appropriate prior distribution on the covariance structure (Lu and Ades 2009), though this is less common in practice.) For two-arm studies against treatment 1, there is a single univariate Normal random effect on the non-treatment 1 arm, with distribution given by (4.40b). The appropriateness of the conditional constancy of relative effects assumption can then be assessed by comparing model fit (e.g. using the Deviance Information Criterion) between the fixed and random effects models and by examining the posterior distribution of the residual heterogeneity variance τ^2.

The presence of residual heterogeneity has several potential causes. For example, there may be effect modifiers that have not been included in the model or other model misspecification, the assumed joint covariate distributions used to adjust the results from aggregate studies may be incorrect, or the shared effect modifier assumption (if it was used) may be invalid. Attempts may be made to rectify these issues in a revised model—if data permits—and the residual heterogeneity of the revised model may then be checked.

The random effects model (4.40) is likely to be more widely applicable in practice than the exchangeable interactions model (4.38) as a relaxation of the standard ML-NMR model (4.34b), since the data requirements are lesser. The random effects model only requires the total number of studies to be large enough to estimate the common heterogeneity variance τ^2, whereas the exchangeable interactions model requires the number of studies and treatments in $\mathcal{T}$ to be large enough to estimate $\boldsymbol{\beta}_{2,k}$, $\boldsymbol{m}_{\beta_2}$, and $\boldsymbol{\sigma}^2_{\beta_2}$.

When producing estimates for a specific target population under the random effects model, we again follow the approach described in Section 4.4. However, estimates can be produced in two different ways, depending on whether the additional uncertainty due to residual heterogeneity is accounted for. For clarity, let us consider producing estimates of population-average treatment effects, as given by (4.28):

$$d_{ab(P)} = \bar{\boldsymbol{x}}^{\mathsf{T}}_{(P)}(\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}) + \gamma_b - \gamma_a.$$

Under the random effects model, the parameters γ_k are interpreted as mean individual-level treatment effects, and so using these γ_k in (4.28) estimates the mean of the population-average treatment effects one would expect to see in studies representative of population P (*super*-population average treatment effects, see Section 2.3.6). However, these estimates do not account for the additional uncertainty due to residual heterogeneity: in any one study performed in population P, the observed individual-level treatment effects differ from the mean treatment effect due to residual heterogeneity described by the random effects distribution $\mathrm{N}(\gamma_k, \tau^2)$. This additional uncertainty can be accounted for by using the posterior predictive distribution for the individual-level treatment effects (Dias et al. 2011a; Higgins et al. 2009; Smith et al. 1995), which is multivariate Normal with marginal distributions and correlations given by

$$\begin{aligned} \delta_{k;\text{new}} &\sim \mathrm{N}(\gamma_k, \tau^2) \\ \mathrm{cor}(\delta_{a;\text{new}}, \delta_{b;\text{new}}) &= 0.5 \end{aligned} \tag{4.41}$$

and $\delta_{1;\text{new}} = 0$. The $\delta_{k;\text{new}}$ describe the predicted individual-level treatment effects one would expect to see in a new study in population P, and are then substituted for γ_k in (4.28) to estimate the population-average treatment effects in a new study in population P:

$$d_{ab(P);\text{new}} = \bar{\boldsymbol{x}}^{\mathsf{T}}_{(P)}(\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}) + \delta_{b;\text{new}} - \delta_{a;\text{new}}. \tag{4.42}$$

The posterior distribution of $d_{ab(P);\text{new}}$ will display more uncertainty than that of $d_{ab(P)}$ (e.g. credible intervals will be wider), reflecting the amount of residual heterogeneity present.

As with the fixed effect model, the ML-NMR model with random effects reduces to IPD RE network meta-regression with full IPD from every study:

$$y_{ijk} \sim \pi_{\text{Ind}}(\theta_{ijk}) \tag{4.43a}$$

$$g(\theta_{ijk}) = \eta_{jk}(\boldsymbol{x}_{ijk}) = \mu_j + \boldsymbol{x}_{ijk}^{\intercal}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \delta_{jk} \tag{4.43b}$$

$$\delta_{jk} \sim \text{N}(\gamma_k, \tau^2) \tag{4.43c}$$

$$\text{cor}(\delta_{ja}, \delta_{jb}) = 0.5 \tag{4.43d}$$

Similarly, the RE ML-NMR model reduces to AgD RE NMA when there are no covariates included in the model: (4.34d) becomes $\theta_{\bullet jk} = g^{-1}\big(\mu_j + \delta_{jk}\big)$, and the model is fitted to the summary outcomes from each trial:

$$y_{\bullet jk} \sim \pi_{\text{Agg}}(\theta_{\bullet jk}) \tag{4.44a}$$

$$g(\theta_{\bullet jk}) = \mu_j + \delta_{jk} \tag{4.44b}$$

$$\delta_{jk} \sim \text{N}(\gamma_k, \tau^2) \tag{4.44c}$$

$$\text{cor}(\delta_{ja}, \delta_{jb}) = 0.5 \tag{4.44d}$$

and in this case γ_k are equal to d_k, the aggregate-level relative effects.

4.6.3 Assessing inconsistency

ML-NMR, like standard IPD and AgD NMA, makes an assumption of consistency that is enforced through a set of consistency equations (Section 1.2.7). For ML-NMR, these apply to both the individual-level treatment effects and the effect modifier interactions, following the equations in (4.36). The causes of inconsistency in ML-NMR are the same as the causes of heterogeneity described previously in Section 4.6.2. For example, there may be effect modifiers that have not been included in the model or other model misspecification, the assumed joint covariate distributions used to adjust the results from aggregate studies may be incorrect, or the shared effect modifier assumption (if it was used) may be invalid. Attempts may be made to rectify these issues in a revised model—if data permits—and the revised model may then be assessed for inconsistency. To assess inconsistency, we can use the same approaches described in Section 1.2.7—in particular the unrelated mean effects model, and node-splitting models.

4.6.3.1 Unrelated mean effects

As originally described for NMA, the unrelated mean effects (UME) model (Dias et al. 2011d) treats all contrasts—both basic (i.e. against treatment 1)

and functional—as independent parameters, without imposing consistency Section 1.2.7.1. The UME model needs to be written with the study-specific baselines referring to a reference *arm* in each trial (as in the baseline shift parameterisation, Section 1.2.1), rather than the reference treatment 1, since the reference treatment parameterisation imposes consistency implicitly. For random effects ML-NMR (4.40), we must also consider the EM interaction terms, and whether or not we allow these to be inconsistent too. The linear predictor and random effects structure of the UME model for RE ML-NMR are written as

$$\eta_{jk}(\boldsymbol{x}) = \mu_j^{(t_1)} + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,t_1k}) + \delta_{jt_1k} \tag{4.45a}$$

$$\delta_{jt_1k} \sim \mathrm{N}(\gamma_{t_1k}, \tau^2) \tag{4.45b}$$

$$\mathrm{cor}(\delta_{jt_1a}, \delta_{jt_1b}) = 0.5 \tag{4.45c}$$

for a study j with treatment t_1 in arm 1, where $\mu_j^{(t_1)}$ is the study-specific baseline with respect to t_1.

If we apply the consistency equations to the EM interactions, $\boldsymbol{\beta}_{2,ab} = \boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}$, then (4.45) only relaxes consistency in the treatment effects. There can also be inconsistency in the EM interactions (Donegan et al. 2017); to assess this as well, we instead place independent prior distributions on $\boldsymbol{\beta}_{2,ab}$, such as $\mathrm{N}(0, \sigma^2_{\beta_2})$ for a suitable prior variance $\sigma^2_{\beta_2}$. However, this requires sufficient data on each contrast to estimate independent interactions. This may not be possible, for example if there are contrasts which are only informed by a small number of AgD studies.

An intermediate approach is possible when the shared EM assumption (Section 2.5) is used to fit the ML-NMR model, so that the regression coefficients $\boldsymbol{\beta}_{2,k}$ for a set of treatments $k \in \mathcal{T}$ are all equal. In this case, we can use the shared EM assumption—which implies that certain interactions are zero or equal to each other—and allow the remaining interactions to be inconsistent. To achieve this, consider (without loss of generality) that every treatment is assigned to a mutually exclusive set $\mathcal{T}_1, \mathcal{T}_2, \ldots$ (some treatments may be in a set by themselves). Then, using the shared EM assumption, EM interactions for contrasts between any two treatments within a given set $\mathcal{T}$ are equal to zero, $\boldsymbol{\beta}_{2,ab} = \mathbf{0}$ for any two treatments $a, b \in \mathcal{T}$. EM interactions for contrasts between treatments in any two different sets $\mathcal{T}_1, \mathcal{T}_2$ are equal, $\boldsymbol{\beta}_{2,a_1a_2} = \boldsymbol{\beta}_{2,b_1b_2}$ for any treatments $a_1, b_1 \in \mathcal{T}_1$ and $a_2, b_2 \in \mathcal{T}_2$, and are assigned a prior distribution such as $\mathrm{N}(0, \sigma^2_{\beta_2})$. This allows us to assess inconsistency of the shared EM interactions, and such a model should always be identifiable when

the corresponding standard (consistency) ML-NMR model with shared EM interactions is identifiable.

In any case, evidence for inconsistency is then based on comparing the model fit (e.g. using residual deviance and DIC) between the ML-NMR model assuming consistency and the UME model without consistency (see Section 1.2.7.1).

4.6.3.2 Node-splitting

The node-splitting approach for network meta-regression models (Donegan et al. 2017) described in Section 1.2.7.3 is naturally applicable to ML-NMR in the same manner as IPD network meta-regression. For a given contrast b' vs. a', the node-splitting model splits the estimation of the relative effect $\gamma_{a'b'}$ and effect modifier interactions $\boldsymbol{\beta}_{2,a'b'}$ into parameters estimated by direct evidence only, $\gamma^{\mathrm{D}}_{a'b'}$ and $\boldsymbol{\beta}^{\mathrm{D}}_{2,a'b'}$, and parameters estimated by the indirect evidence from the rest of the network, $\gamma^{\mathrm{I}}_{a'b'}$ and $\boldsymbol{\beta}^{\mathrm{I}}_{2,a'b'}$. To achieve this, the random effects ML-NMR model (4.40) remains the same for studies not including both a' and b' treatment arms. For studies including both a' and b' treatment arms, we choose to re-parameterise the model with a' as the reference treatment within these studies. Mathematically, we write out the linear predictor and random effects for this node-splitting model as

For studies without both a' and b' treatment arms:

$$\eta_{jk}(\boldsymbol{x}) = \mu_j^{(1)} + \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \delta_{jk} \tag{4.46a}$$

$$\delta_{jk} \sim \mathrm{N}(\gamma_k, \tau^2) \tag{4.46b}$$

$$\mathrm{cor}(\delta_{ja}, \delta_{jb}) = 0.5 \tag{4.46c}$$

For studies with both a' and b' treatment arms:

$$\eta_{ja'}(\boldsymbol{x}) = \mu_j^{(a')} + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}_1 \tag{4.46d}$$

$$\eta_{jb'}(\boldsymbol{x}) = \mu_j^{(a')} + \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}^{\mathrm{D}}_{2,a'b'}) + \delta_{ja'b'} \tag{4.46e}$$

$$\eta_{jk}(\boldsymbol{x}) = \mu_j^{(a')} + \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k} - \boldsymbol{\beta}_{2,a'}) + \delta_{ja'k} \quad \text{for } k \neq a', b' \tag{4.46f}$$

$$\delta_{ja'b'} \sim \mathrm{N}(\gamma^{\mathrm{D}}_{a'b'}, \tau^2) \tag{4.46g}$$

$$\delta_{ja'k} \sim \mathrm{N}(\gamma_k - \gamma_{a'}, \tau^2) \quad \text{for } k \neq a', b' \tag{4.46h}$$

$$\mathrm{cor}(\delta_{ja'a}, \delta_{ja'b}) = 0.5 \quad \text{for } a, b \neq a', b' \tag{4.46i}$$

$$\mathrm{cor}(\delta_{ja'k}, \delta_{ja'b'}) = 0 \quad \text{for } k \neq a', b' \tag{4.46j}$$

where the re-parameterised study-specific baselines with respect to treatment a' are denoted by $\mu_j^{(a')}$, and we write the study-specific baselines with respect

to treatment 1 as $\mu_j^{(1)} = \mu_j$ for additional clarity. As usual we set $\gamma_1 = \delta_{j1} = 0$ and $\boldsymbol{\beta}_{2,1} = \mathbf{0}$. If there are multi-arm studies with both a' and b' treatment arms, then the other random effects $\delta_{ja'k}$ with $k \neq b'$ are uncorrelated with $\delta_{ja'b'}$, but are still correlated between themselves with $\text{cor}(\delta_{ja'a}, \delta_{ja'b}) = 0.5$ for $a, b \neq a', b'$ (assuming homogeneous τ^2). The indirect estimates $\gamma^{\text{I}}_{a'b'}$ and $\boldsymbol{\beta}^{\text{I}}_{2,a'b'}$ are obtained from the consistency equations

$$\begin{aligned} \gamma^{\text{I}}_{a'b'} &= \gamma_{b'} - \gamma_{a'}, \\ \boldsymbol{\beta}^{\text{I}}_{2,a'b'} &= \boldsymbol{\beta}_{2,b'} - \boldsymbol{\beta}_{2,a'}. \end{aligned} \tag{4.47}$$

The node-splitting model as written in (4.46) splits the EM interaction terms for all covariates at once. Alternatively, a separate node-splitting model could be fitted for each covariate in turn, where $\boldsymbol{\beta}^{\text{D}}_{2,a'b'}$ is broken down into a split interaction for one covariate, $\beta^{\text{D}}_{2,a'b';l}$, and the consistency equations are applied for the remaining covariates $\beta_{2,a'b';l} = \beta_{2,b';l} - \beta_{2,a';l}$. The latter approach may be more tractable in scenarios with smaller amounts of data on the b' vs. a' contrast and/or large numbers of effect modifying covariates, since there are only 2 more parameters than the standard RE ML-NMR model ($\gamma^{\text{D}}_{a'b'}$ and $\beta^{\text{D}}_{2,a'b';l}$), as opposed to $L + 1$ more when splitting all EM interactions at once ($\gamma^{\text{D}}_{a'b'}$ and $\boldsymbol{\beta}^{\text{D}}_{2,a'b'}$), where L is the number of covariates.

Furthermore, there may be insufficient data on the b' vs. a' contrast even to node-split the EM interaction terms one covariate at a time, for example if the direct evidence consists of only a small number of AgD studies. In this case, we may be able to assess inconsistency in the treatment contrast $\gamma_{a'b'}$ by node-splitting into $\gamma^{\text{D}}_{a'b'}$ and $\gamma^{\text{I}}_{a'b'}$, but not in the EM interactions, keeping the consistency equations $\boldsymbol{\beta}_{2,a'b'} = \boldsymbol{\beta}_{2,b'} - \boldsymbol{\beta}_{2,a'}$.

Section 1.2.7.2 describes how to interpret the results of node-splitting models to assess inconsistency. As with the unrelated mean effects model (Section 4.6.3.1), one check for inconsistency is to compare the model fit (e.g. using residual deviance and DIC) between the ML-NMR model with and without node-splitting. For each of the L covariates x_l in the vector $\boldsymbol{x}$, the posterior distributions of $\gamma^{\text{D}}_{a'b'} + x_l\beta^{\text{D}}_{2,a'b';l}$ and $\gamma^{\text{I}}_{a'b'} + x_l\beta^{\text{I}}_{2,a'b';l}$ can be plotted as functions of x_l and compared, as can the posterior distribution of the inconsistency parameter $\omega_{a'b'}(x_l) = (\gamma^{\text{D}}_{a'b'} - \gamma^{\text{I}}_{a'b'}) + x_l(\beta^{\text{D}}_{2,a'b';l} - \beta^{\text{I}}_{2,a'b';l})$.

Using published marginal covariate information 4.6.4

To implement ML-NMR, we need to describe the joint covariate distribution $f_{jk}(\cdot)$ in each treatment arm of each AgD trial. In practice, joint covariate information is unlikely to be available from the AgD studies, so instead

we infer the forms of the marginal distributions and correlation structure from other information such as clinical knowledge or from the IPD studies. The approaches described and discussed in Section 4.5.1 for the two-study scenario apply equally in larger networks, although there are some additional considerations when multiple IPD studies are available from which to infer information on the joint covariate distribution.

If multiple IPD studies are available, it is possible that the observed marginal covariate distributions may differ in form between the IPD studies (perhaps due to differences in inclusion criteria or other aspects of study design). In this case, marginal distributions for each AgD study may be inferred from the IPD study or studies deemed most representative.

Similarly, when inferring a correlation matrix for the AgD studies from multiple IPD studies, one option is to choose the correlation matrix from an IPD study deemed most representative for each AgD study. Another option is to use a weighted average of the covariance matrices from the IPD studies (or a suitable representative subset), for example by applying the approach described by Hedges and Olkin (1985, Chapter 11) to each pairwise correlation. The correlations $\rho_{j;l_1l_2}$ between covariates x_{l_1} and x_{l_2} in each study j are z-transformed as

$$\zeta_{j;l_1l_2} = \frac{1}{2}\log\left(\frac{1+\rho_{j;l_1l_2}}{1-\rho_{j;l_1l_2}}\right), \tag{4.48}$$

before taking a weighted average

$$\bar{\zeta}_{l_1l_2} = \frac{\sum_{\text{IPD } j} w_j \zeta_{j;l_1l_2}}{\sum_{\text{IPD } j} w_j}, \tag{4.49}$$

where the weights $w_j = N_j - 3$ are the inverse of the asymptotic variance of $\zeta_{j;l_1l_2}$, and N_j is the number of individuals in study j. The weighted average correlations are then obtained by back-transformation of $\bar{\zeta}_{l_1l_2}$ using

$$\bar{\rho}_{l_1l_2} = \frac{\exp(2\bar{\zeta}_{l_1l_2}) - 1}{\exp(2\bar{\zeta}_{l_1l_2}) + 1}. \tag{4.50}$$

4.7 Discussion

In this chapter, we have proposed a new method for population-adjusted indirect comparisons and network meta-regression. ML-NMR derives from a method presented in the ecological inference literature, where the aggregate-level model is obtained by integrating the individual-level model over the covariate distribution (Jackson et al. 2006, 2008) (discussed in Section 2.2.3). The methods in the ecological inference literature have been applied in the

context of NMA previously, but were only derived for the simple case of a binary outcome and binary covariates (Jansen 2012). There are several key advantages to this approach, particularly in comparison with methods such as MAIC (Ishak et al. 2015; Signorovitch et al. 2010) (see Section 2.2.1), STC (Caro and Ishak 2010; Ishak et al. 2015) (see Section 2.2.2), or network meta-regression based approaches (Donegan et al. 2013; Saramago et al. 2012; Sutton et al. 2008; Thom et al. 2015) (see Section 2.2.3).

MAIC and STC are designed with a simple two-study scenario in mind (one IPD AB study, one AgD AC study). By comparison, ML-NMR is applicable to treatment networks of any size, allowing use of all available information. Both MAIC and STC can be considered as "targeted comparisons" rather than true evidence syntheses, since information from the AC trial is not propagated through to the parameter estimates. As we showed in Section 4.5.2, ML-NMR can also be used in this manner, fitting a model in the IPD AB trial and predicting outcomes on treatments A and B the AC trial population. However, as with MAIC and STC, the results of this targeted comparison are then limited in validity to the AC trial population, which may not reflect the target population of interest.

We have considered an individual-level model where there is a single treatment-covariate interaction term per effect modifying covariate. In the context of IPD network meta-regression, Hua et al. (2016) suggest splitting the EM interaction term into a within-study interaction $\boldsymbol{\beta}_{2,k}^{(w)}$ and a between-study interaction $\boldsymbol{\beta}_{2,k}^{(b)}$ like so:

$$\eta_{jk}(\boldsymbol{x}) = \mu_j + (\boldsymbol{x}_{ijk} - \bar{\boldsymbol{x}}_j)^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}^{(w)}) + \bar{\boldsymbol{x}}_j^\intercal \boldsymbol{\beta}_{2,k}^{(b)} + \delta_{jk}. \tag{4.51}$$

Their reasoning is that combining information on interactions from within and between studies can result in ecological bias due to unobserved effect modifiers, and splitting the interaction term in this manner means that $\boldsymbol{\beta}_{2,k}^{(w)}$ should be free from this bias. Hua et al. (2016) suggest drawing conclusions from only the within-study interactions, and interpret differences between the within- and between-study interaction estimates as evidence for ecological bias. This same "split-interactions" model has been used by several authors to incorporate AgD studies (Donegan et al. 2013; Riley et al. 2008; Riley and Steyerberg 2010; Saramago et al. 2012) (others choose not to split interactions (Sutton et al. 2008), see further discussion in Section 2.2.3). Using the notation of ML-NMR, these models are written as

Individual:

$$y_{ijk} \sim \pi_{\text{Ind}}(\theta_{ijk}) \tag{4.52a}$$

$$g(\theta_{ijk}) = \mu_j + (\boldsymbol{x}_{ijk} - \bar{\boldsymbol{x}}_j)^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}^{(w)}) + \bar{\boldsymbol{x}}_j^\intercal \boldsymbol{\beta}_{2,k}^{(b)} + \delta_{jk} \tag{4.52b}$$

Aggregate:

$$y_{\bullet jk} \sim \pi_{\text{Agg}}(\theta_{\bullet jk}) \tag{4.52c}$$

$$g(\theta_{\bullet jk}) = \mu_j + \bar{\boldsymbol{x}}_j^\intercal \boldsymbol{\beta}_{2,k}^{(b)} + \delta_{jk} \tag{4.52d}$$

where δ_{jk} are either fixed or random effects as before. Notably, for the AgD studies only the between-study interactions are used and the mean covariate values $\bar{\boldsymbol{x}}_j$ are "plugged in" to the linear predictor. As a result, differences between the within- and between-study interactions are now due to two possible sources of ecological bias: one from unobserved effect modifiers, and an additional aggregation bias when the model is non-linear due to "plugging in" means (see Greenland 1992, who refers to the latter as "pure specification" bias). As well as population characteristics, effect modification may be due to other study-level factors at such as differences in treatment intensity or study setting; in practice, study-level factors may be difficult to determine from the available data, and adjustment is more difficult than for population characteristics since they do not vary between the individuals within each study. Clearly, ML-NMR avoids the aggregation bias, since the individual model is appropriately integrated over the covariate distribution in the AgD studies. The other possible source of ecological bias—unobserved effect modifiers—is assumed not to be present when fitting the standard ML-NMR model due to the conditional constancy of relative effects assumption, which is required to hold in order to produce unbiased estimates of population-adjusted relative effects. As such, we have not split the interaction terms in the ML-NMR model. ML-NMR could be modified to include split interaction terms, however this results in a non-identifiable model in the two-study scenario, and likely requires substantial amounts of data to estimate well. By comparison, the RE ML-NMR model may be used to investigate residual heterogeneity due to unobserved effect modifiers (or from other sources, such as an invalid shared EM assumption), and is less data-intensive (Section 4.6.2). In practice, we thus propose to use the RE ML-NMR model, although further research to compare the different approaches is warranted (Section 9.2.2).

In Section 4.6, we conceptualised ML-NMR as an extension of the standard NMA framework, which is an established and accepted method with a broad literature. Standard IPD and AgD NMA are special cases of ML-NMR: ML-NMR reduces to standard AgD NMA when there are no covariates included in

the model, and to IPD network meta-regression when IPD are available from every study (see Section 4.6). When implemented in a Bayesian framework ML-NMR retains the flexibility and extensibility of Bayesian NMA, so that, for example, prior information could be utilised, several data types with differing likelihoods could be included, or the analysis embedded in a probabilistic cost-effectiveness framework as widely used by decision makers (Claxton et al. 2005; Critchfield and Willard 1986; Dias et al. 2013b; Doubilet et al. 1985). Inconsistency checks between the direct and indirect information are frequently performed for NMA (Dias et al. 2010, 2013d; Higgins et al. 2012; Lu and Ades 2006; see Section 1.2.7), and are equally applicable in a ML-NMR analysis where a larger network of studies is available, as we described in Section 4.6.3. ML-NMR aims to avoid heterogeneity and inconsistency by adjusting for differences in effect modifying covariates, however it is important to investigate whether there is any residual heterogeneity or inconsistency after adjustment, and attempt to rectify this if data permits (Sections 4.6.2 and 4.6.3).

In Section 4.3.2, we presented algebraic derivations of the aggregate-level model for the probit link function, logit link function (via approximation to the probit), and log link function (via moment generating functions). However, these algebraic approaches quickly become complex, are context-specific, and are not always tractable (e.g. with skew covariates). Using QMC integration instead provides a general numerical approach that is straightforward and broadly applicable (Section 4.3.3.2), and, in practical terms, much quicker to implement than an algebraic approach. The use of quasi-random sequences for numerical integration improves upon the convergence rates of standard pseudo-random Monte Carlo integration, and retains this performance in high-dimensions. We describe the implementation of QMC integration in detail in Section 5.1.

As is crucial for relevance in decision making, comparisons may be provided in any target population given sufficient information on the covariate distribution (Section 4.4), without the need for additional assumptions required by methods such as MAIC or STC. However, extrapolation may be necessary if the covariate distribution in the target population does not overlap with the covariate distributions in the observed data. The target population need not be a clinical trial, and could be taken from other data sources such as registries.

All population adjustment methods rely on the availability of covariate information in all included studies. In a connected network, adjustment is

only required for effect modifiers (Phillippo et al. 2016; Phillippo et al. 2018a) (Section 2.3.1) which, being of high clinical relevance, are more likely to be widely reported in publications. Aside from unmeasured or unreported covariates, other forms of missing information are also an issue.

Firstly, it is unlikely for publications to report the correlation structure between covariates, although this may be available on request. Methods such as MAIC ignore the correlations between covariates. However, as we have shown algebraically for multivariate Normal covariates (Section 4.3.2), correlations involving effect modifiers are implicated in aggregation bias along with within-study variation. Assuming that correlation structures are the same within the AgD and IPD, which may be more reasonable than assuming all correlations are zero, we can impute the missing correlations when generating the AgD integration points (Section 4.5.1). Correlation structures could differ between the populations if, for example, different characteristics coexist, or the sampling and randomisation methods differ between the trials. However, later in the simulation study we show that varying the assumed correlation structure of the AgD trial has negligible impact on the results (Section 8.2.6). Furthermore, the true forms of the marginal distributions of the aggregate covariates are likely unknown. Instead, we choose distributional forms for the aggregate covariates based on theoretical properties and to approximately match the observed distributional forms in the IPD (Section 4.5.1). However, the true marginal distributions could differ between the populations if, for example, the sampling and randomisation methods differ between the trials. Later in the simulation study, we see that altering the assumed marginal distributions for the AgD trial may have little impact on the results Section 8.2.6. We propose a very flexible approach based on copulæ, allowing any set of desired marginal distributions to be combined under a given joint correlation structure (Section 4.3.3.2, and greater detail in Section 5.1). A limitation of this approach is that uncertainty in selecting marginal distributions and correlation structures is not accounted for.

Missing values within the IPD are also likely to be encountered. One solution is to simply remove those individuals with missing covariate values from the analysis, which may be reasonable if the proportion of individuals with missing values is very small. However, in general this is not recommended, as complete case analyses can incur bias and loss of precision. Multiple imputation is a widely-used method of dealing with missing data (Kenward and Carpenter 2007; Little and Rubin 2002). The Bayesian framework is well-suited to implementing multiple imputation as covariates can be imputed

at every iteration of the MCMC sampler, incorporating the uncertainty arising from the missing values into the posterior distribution (Jackson et al. 2009; Mason et al. 2012). Such approaches have previously been described for IPD NMA (Quartagno and Carpenter 2016), and apply similarly to ML-NMR by extension.

Derivation of the aggregate likelihood is not always straightforward (Section 4.2), and may even be intractable. Most notably this is the case for survival analysis—which represents the large majority of applications of population adjustment methodology to date (Chapter 3)—where the aggregate likelihood cannot be derived analytically. In Chapter 7 we extend ML-NMR to general likelihoods, including for survival data, further increasing the applicability of this new method. In Chapter 8, we perform an extensive simulation study to assess the performance of ML-NMR in a wide variety of scenarios and its robustness to failure in assumptions, in comparison with other methods. In the following chapter we discuss computation of ML-NMR models, before applying the methods described here to a real example in Chapter 6.

Chapter 5

Computation of ML-NMR models

In this chapter, we discuss the computational aspects of implementing ML-NMR. As we introduced in Section 4.1, one of the key ideas behind ML-NMR is that the individual-level model $\theta_{ijk} = g^{-1}(\eta_{jk}(\boldsymbol{x}_{ijk}))$ is integrated over the covariate distribution $f_{jk}(\cdot)$ in an AgD trial arm to obtain the appropriate aggregate-level model for that arm:

$$\theta_{\bullet jk} = \int_{\mathfrak{X}} g^{-1}(\eta_{jk}(\boldsymbol{x})) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{5.1}$$

In Section 4.3.2, we considered the algebraic solutions to (5.1) in some special cases; however, in general such algebraic solutions quickly become complex, and may even be intractable. Instead, in Section 4.3.3 we proposed to use numerical integration techniques to evaluate the integral. The most attractive numerical approach for our purposes is Quasi-Monte Carlo integration (Section 4.3.3.2), since it is more widely applicable than quadrature (Section 4.3.3.1) and can be applied regardless of the number of covariates or the type or complexity of the model.

Models involving IPD, including ML-NMR, can be large and slow to analyse in both Bayesian and frequentist frameworks. It is therefore particularly desirable to implemented such models in an efficient manner, saving both time and computational effort. Efficiency may be dichotomised into two concepts: *computational* efficiency, and *statistical* efficiency. In a Bayesian MCMC framework, computational efficiency refers to minimising the amount of time[1] required in performing the calculations for a given sample iteration, for example by "vectorising" code (removing `for` loops in favour of more efficient

[1] Other measures of computational effort may also be of interest, such as the number of computation operations or the amount of memory required, but often we are simply concerned with speeding up model run-times.

vector/matrix statements). Statistical efficiency corresponds to obtaining the greatest effective number of samples from the posterior distribution for a given number of iterations, and is achieved by choosing a suitable model parameterisation.

We begin by describing how Quasi-Monte Carlo integration is used to evaluate the marginalisation integral (5.1), and suggest visual checks for the magnitude of integration error. We then discuss the efficient implementation of ML-NMR in Stan, focusing on issues of statistical efficiency (efficient parameterisation) as opposed to computational efficiency. We also review techniques for checking convergence and other sampler diagnostics in Stan. Finally we explore assessment of model fit and model comparison in the ML-NMR framework, before concluding with a discussion.

5.1 QMC integration with copulæ

In this section, we describe in greater detail the implementation of QMC integration. We start by considering only continuous covariates; however, in Section 5.1.1 we describe a simple extension to incorporating discrete covariates also.

To begin, we need to describe the joint covariate distribution $f_{jk}(\cdot)$ in each AgD study j on treatment k. Let us summarise $f_{jk}(\cdot)$ by the L constituent marginal distributions $f_{jk;l}(\cdot)$, $l = 1, \ldots, L$, where L is the number of covariates, and information on the relationship structure between covariates. Sklar's theorem shows that any joint distribution can be fully described in this manner, by decomposition into marginal distributions and a relationship structure (Sklar 1959, see also Nelsen 2006):

$$f_{jk}(\boldsymbol{x}) = C(f_{jk;1}(x_1), \ldots, f_{jk;L}(x_L)). \tag{5.2}$$

The function $C(\cdots)$ is called a *copula*, and encodes the relationships between the marginal distributions to "couple" them into the joint distribution. A range of copula functions are available, each encoding different relationships between covariates, whilst allowing arbitrary marginal distributions to be specified (Nelsen 2006). We assume that the marginal distributions and relationship structure are known—or at least may be inferred, since in practice it is likely that only marginal covariate summaries (e.g. means and standard deviations for continuous covariates, or proportions for discrete covariates) will be reported from the AgD studies. As described in Section 4.5.1, when the true marginal distributions in the AgD studies are unknown, we can choose

distributional forms for these covariates based on their theoretical properties and/or to approximately match the observed distributional forms in the IPD, and then match these distributions to the reported summary statistics. Similarly, we can utilise the relationship structure observed between covariates in the IPD studies, and assume that this holds also in the AgD studies. Here we will use a Gaussian copula, encoding the relationship structure between covariates with a correlation matrix $\boldsymbol{\Omega}_{jk}$; however, the following process applies in the same manner when other copulæ are chosen. Note that "Gaussian" here refers only to the form of the copula: no restrictions are placed on the form of the marginal distributions, which may take any arbitrary distribution (even discrete marginal distributions are allowed, which we utilise in Section 5.1.1). In practice, we find that this is a very flexible approach and can account for a wide range of relationships between covariates. Using a Gaussian copula, the necessary summaries from each AgD study are thus a description of the marginal distributions $f_{jk;l}(\cdot)$ (for example summarised by means and standard deviations and a given distributional form) and the correlation matrix $\boldsymbol{\Omega}_{jk}$.

The aim is to obtain a sample of integration points $\tilde{\boldsymbol{x}}_{ijk}$, $i = 1, \ldots, \tilde{N}$ from the joint covariate distribution, with which we can evaluate the integral of any given function over $f_{jk}(\cdot)$. Rather than a (pseudo-)random sample from $f_{jk}(\cdot)$, which we could use for Monte Carlo integration over the covariate distribution, we will instead seek a *quasi-random* sample of points that cover the covariate space $\mathfrak{X}$ more uniformly than a random sample, resulting in a Quasi-Monte Carlo integration scheme. As outlined in Section 4.3.3.2, QMC integration can achieve much faster convergence of integration errors than standard Monte Carlo integration. Standard Monte Carlo integration has an expected error rate of order $\tilde{N}^{-\frac{1}{2}}$, whereas QMC integration—whilst having a worst-case error rate of the order $\tilde{N}^{-1}(\log \tilde{N})^L$—often achieves an error rate of the order $\tilde{N}^{-1}$, even in high dimensions (Caflisch 1998; Niederreiter 1978). (See Section 5.1.2 for suggestions on checking the rate of convergence.)

We start with an L-dimensional sequence of quasi-random points $\tilde{\boldsymbol{u}}_{ijk}$, $i = 1, \ldots, \tilde{N}$ in the unit hypercube $[0, 1]^L$, where $L = |\mathfrak{X}|$ is the dimension of the covariate space $\mathfrak{X}$ (i.e. the number of covariates). Several possible quasi-random sequences have been proposed (see Caflisch 1998). Here, we use a Sobol' sequence (Sobol' 1967) to generate $\tilde{\boldsymbol{u}}_{ijk}$, using the function `sobol` from the R package `randtoolbox` (Christophe and Petr 2018). Figure 5.1 shows a sample of 2048 Sobol' points in two dimensions, alongside the same number of pseudo-random points (generated using the R function `runif`) for comparison.

We use a Gaussian copula (Nelsen 2006) to describe the relationship

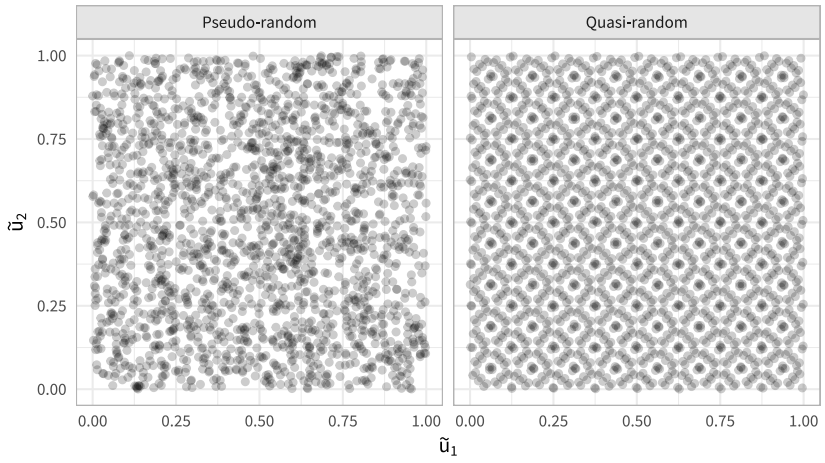


Figure 5.1 Samples of 2048 points $\tilde{\boldsymbol{u}}$ in two dimensions, generated pseudo-randomly (left) and quasi-randomly using a two-dimensional Sobol' sequence (right).

structure between covariates, which imposes the correlation matrix $\boldsymbol{\Omega}_{jk}$ on the Sobol' points $\tilde{\boldsymbol{u}}_{ijk}$. This is equivalent to applying the inverse cumulative distribution function (CDF) $\Phi_{\Omega}^{-1}(\cdot)$ of the multivariate Normal with correlation matrix $\boldsymbol{\Omega}_{jk}$, and then the standard multivariate Normal CDF $\Phi(\cdot)$,

$$\tilde{\boldsymbol{u}}^*_{ijk} = \Phi(\Phi_{\Omega}^{-1}(\tilde{\boldsymbol{u}}_{ijk})), \tag{5.3}$$

to obtain correlated Sobol' points $\tilde{\boldsymbol{u}}^*_{ijk}$. (In practice this is computed component-wise as conditional univariate Normal distributions; we use the implementation in the R package `copula` (Yan 2007).) Figure 5.2 shows this transformation applied to the uncorrelated points from Figure 5.1, with a correlation of 0.4.

The inverse CDF method (also called the inverse transform method) is a widely-used technique for generating samples from any distribution for which the inverse CDF $F^{-1}(\cdot)$ is known (see Devroye 1986, Chapter 2). The process is simple: given a sample of uniformly distributed points u, the inverse CDF is applied to obtain a sample $x = F^{-1}(u)$ from the desired target distribution. We use this technique here to transform the correlated Sobol' points $\tilde{u}^*_{jk;l}$ to match the marginal covariate distributions reported in the AgD trials:

$$\tilde{x}_{jk;l} = F^{-1}_{jk;l}(\tilde{u}^*_{jk;l}) \quad \text{for } l = 1, \ldots, L, \tag{5.4}$$

where $F^{-1}_{jk;l}(\cdot)$ is the inverse CDF of the marginal distribution of the l-th covariate in study j on treatment k. The resulting integration points $\tilde{\boldsymbol{x}}_{ijk}$ capture the correlations between the covariates (e.g. longer duration of psoriasis is correlated with having previous systemic treatment) whilst preserving

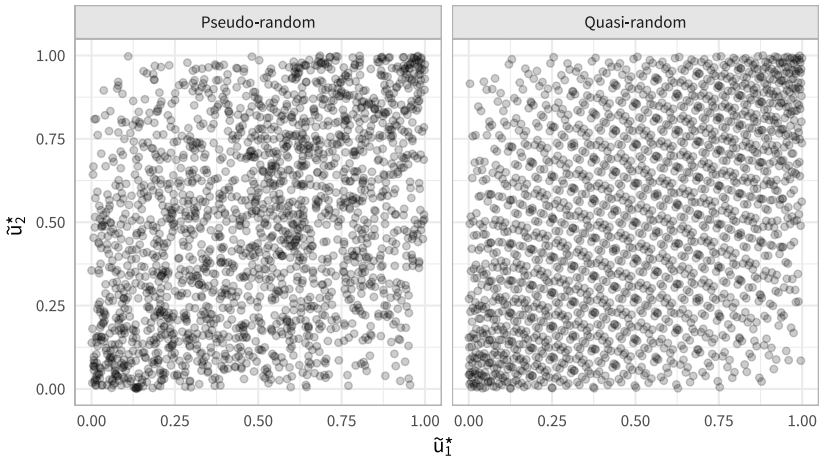


Figure 5.2 The correlated points $\tilde{\boldsymbol{u}}^*$ after applying a correlation of 0.4 to the points $\tilde{\boldsymbol{u}}$, for pseudo-random samples (left) and quasi-random (Sobol') samples (right).

the marginal distribution for each covariate. As an example, we consider transforming the correlated Sobol' points in two dimensions from Figure 5.2 to have marginal distributions with means 4 and 2, standard deviations 1.5 and 0.8, and distributed as Normal and Gamma distributions respectively. Figure 5.3 shows marginal histograms of the integration points against the true marginal distributions (solid line), and Figure 5.4 shows the integration points jointly in two dimensions along with empirical density contours. Notice how the marginal histograms for the quasi-random Sobol' points follow the true marginal densities much more closely for the same number of integration points. Similarly, in two-dimensions the points cover the joint distribution more uniformly and have smoother empirical density contours.

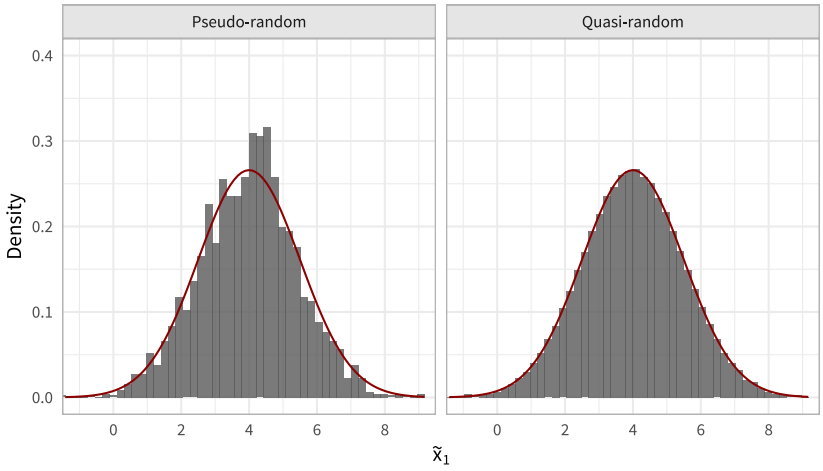
The integral (5.1) is then evaluated using the integration points $\tilde{\boldsymbol{x}}_{ijk}$ as

$$\int_{\mathfrak{X}} g^{-1}(\eta_{jk}(\boldsymbol{x}))f_{jk}(\boldsymbol{x})\,d\boldsymbol{x} \approx \frac{1}{\tilde{N}}\sum_{i=1}^{\tilde{N}} g^{-1}(\eta_{jk}(\tilde{\boldsymbol{x}}_{ijk})). \tag{5.5}$$

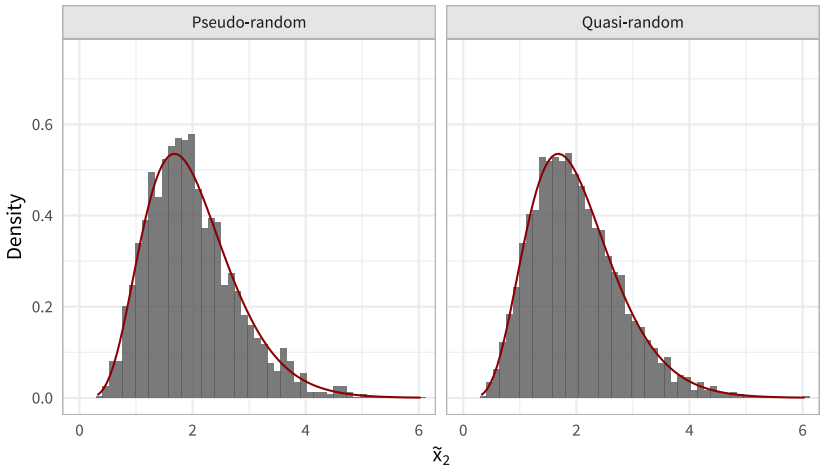
Indeed, the integral of any general function $h(\boldsymbol{x})$ of the covariates can be evaluated using the integration points in this manner:

$$\int_{\mathfrak{X}} h(\boldsymbol{x})f_{jk}(\boldsymbol{x})\,d\boldsymbol{x} \approx \frac{1}{\tilde{N}}\sum_{i=1}^{\tilde{N}} h(\tilde{\boldsymbol{x}}_{ijk}), \tag{5.6}$$

for example to produce estimates of population-average quantities in a given target population, as described in Section 4.4.



(a) First marginal is Normally distributed with mean 4 and standard deviation 1.5.



(b) Second marginal is Gamma distributed with mean 2 and standard deviation 0.8.

Figure 5.3 Histograms of the integration points $\tilde{\boldsymbol{x}}$ in each dimension, after applying the inverse CDFs to $\tilde{\boldsymbol{u}}^*$. The overlaid line shows the density of the true marginal distribution from which the points are sampled. The same transformation is applied to pseudo-random points (left) and quasi-random (Sobol') points (right).

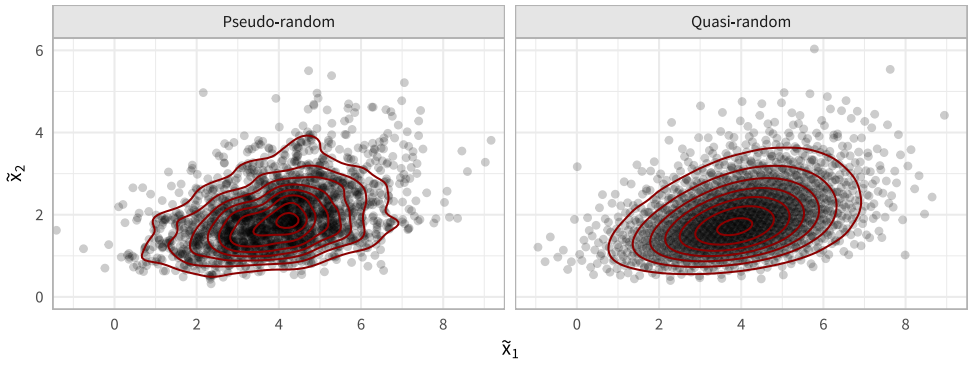


Figure 5.4 The integration points $\tilde{x}$ in two-dimensions, after applying the inverse CDFs to $\tilde{u}^*$. Empirical contours of the integration points are overlaid. The same transformation is applied to pseudo-random points (left) and quasi-random (Sobol') points (right).

Incorporating discrete covariates 5.1.1

Earlier in Section 4.3.4, we suggested that one approach to integration over discrete and continuous covariates simultaneously was to nest the integration over continuous covariates in a summation over the levels of discrete covariates (equation (4.27)). However, this quickly becomes an unwieldy approach as the number of discrete covariates increases, and furthermore does not account for correlations between discrete and continuous covariates unless summary statistics are available within each discrete subgroup.

Instead, we propose to integrate over both discrete and continuous covariates together using QMC integration. The process described previously is unchanged by the inclusion of discrete covariates. If the correlation matrix $\boldsymbol{\Omega}_{jk}$ is to be computed from the IPD and applied to the AgD studies, we use Spearman's rank correlation to allow for discrete covariates. This approach allows correlations between discrete and continuous covariates to be accounted for, and computation of the integral (5.1) via equation (5.5) is unchanged—regardless of the number or type of covariates—avoiding unwieldy nested summations. One possible drawback of this approach is that it has the potential to increase the number of integration points required, particularly when there are large correlations and many discrete covariates; however, in practice we have not observed this to be an issue.

Checking integration error 5.1.2

The convergence of Monte Carlo integration schemes as $\tilde{N} \to \infty$, evaluating the integral (5.1) using (5.5), is guaranteed (under some mild regularity

conditions which are unlikely to be of practical concern here) (Niederreiter 1978). However, in practice we would like to examine integration error for finite $\tilde{N}$, with the aim of determining a suitable value for $\tilde{N}$. We propose plotting empirical integration error against the number of integration points, where empirical integration error is estimated by the relative difference from the final estimate, at every posterior sample (since we are fitting ML-NMR models using MCMC). Letting $I_{n,s}$ be the estimated value of the integral (5.5) at posterior sample s using n integration points, mathematically the empirical integration error at posterior sample s is written

$$I_{n,s} - I_{\tilde{N},s}, \quad \text{for } n = 1, \ldots, \tilde{N} - 1 \text{ and } \forall s. \tag{5.7}$$

For practical visualisation purposes, we suggest plotting the empirical integration error at suitable steps of n (e.g. in steps of 100), and summarising the empirical integration error at each step of n over the entire posterior distribution using a "violin" plot (a smooth density estimate) or box plot. An example plot is shown in Figure 5.5; in this example we see that the integration error over the entire posterior distribution decreases at the expected rate of $\tilde{N}^{-1}$ (given by the dashed lines). Based on these plots, a judgement can be made over suitable $\tilde{N}$, weighing up decreasing integration error against increasing computational cost. We use this technique in the applied examples throughout this thesis, for example Figure 6.4 in Chapter 6.

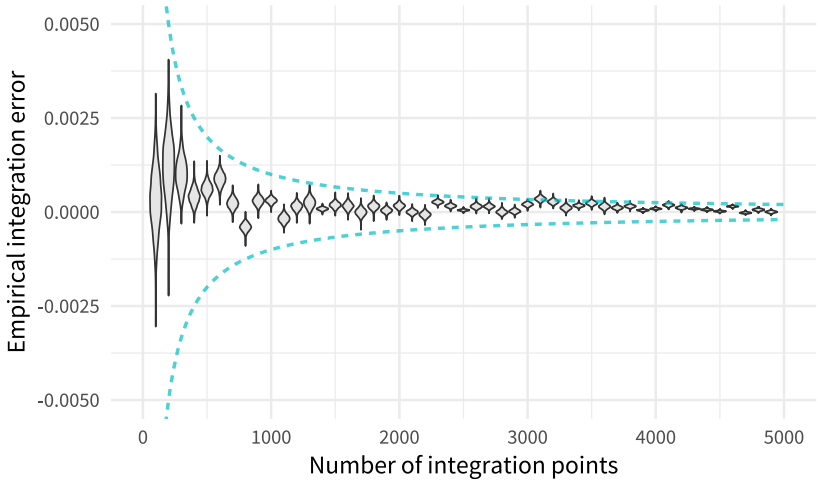


Figure 5.5 An example plot of empirical integration error against the number of (quasi-random Sobol') integration points. Steps of 100 are used, at each of which the empirical integration error is summarised over all posterior samples using a violin plot. The dashed line is $\pm\tilde{N}^{-1}$, showing that the integration error is of this order.

Efficient implementation in Stan 5.2

We now consider several issues regarding the efficient implementation of ML-NMR models in Stan. Here, we focus specifically on statistical efficiency of ML-NMR models through efficient parameterisation, although our Stan code (Appendix A) is written with computational efficiency in mind too. Techniques for achieving both statistical and computational efficiency in Stan are discussed in Section 23 of the Stan User's Guide (Stan Development Team 2018).

In this section it will be helpful to consider the ML-NMR model written in a general matrix form, with QMC integration (5.5) substituted in for the aggregation integral (5.1), which also aligns with the implementation of the model in Stan. Considering the fixed effect model for now (the random effects model is discussed in Section 5.2.3), we write:

Individual:

$$y_{ijk} \sim \pi_{\text{Ind}}(\theta_{ijk}) \tag{5.8a}$$

$$g(\boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\xi} \tag{5.8b}$$

Aggregate:

$$y_{\bullet jk} \sim \pi_{\text{Agg}}(\theta_{\bullet jk}) \tag{5.8c}$$

$$\theta_{\bullet jk} \approx \frac{1}{\tilde{N}} \sum g^{-1}\left(\tilde{\boldsymbol{X}}_{jk}\boldsymbol{\xi}\right) \tag{5.8d}$$

where $\boldsymbol{X}$ is a design matrix with a row for each individual in the IPD, picking out the appropriate components from parameter vector

$$\boldsymbol{\xi} = \left(\mu_1, \ldots, \mu_J \mid \boldsymbol{\beta}_1^{\intercal} \mid \boldsymbol{\beta}_{2,2}^{\intercal}, \ldots, \boldsymbol{\beta}_{2,K}^{\intercal} \mid \gamma_2, \ldots, \gamma_K\right)^{\intercal},$$

and $\boldsymbol{\theta}$ is the corresponding vector of individual-level predictors θ_{ijk}. Vertical bars emphasise the partitioning of $\boldsymbol{\xi}$ corresponding to study intercepts μ_j, prognostic effects $\boldsymbol{\beta}_1$, EM interactions $\boldsymbol{\beta}_{2,k}$, and treatment effects γ_k. For example, an individual in study 2 on treatment 4 with covariates $\boldsymbol{x}_{i24}$ would have corresponding row of $\boldsymbol{X}$ equal to

$$\left(0, 1, 0 \ldots, 0 \mid \boldsymbol{x}_{i24}^{\intercal} \mid 0, 0, \boldsymbol{x}_{i24}^{\intercal}, 0, \ldots, 0 \mid 0, 0, \boldsymbol{x}_{i24}^{\intercal}, 0, \ldots, 0\right).$$

At the aggregate level, $\tilde{\boldsymbol{X}}_{jk}$ is the "design matrix" of integration points for study j treatment k, formed in the same way as the individual-level $\boldsymbol{X}$ but with $\tilde{N}$ rows, one for each integration point. The summation in (5.8d) is over the $\tilde{N}$ components of the vector $g^{-1}(\tilde{\boldsymbol{X}}_{jk}\boldsymbol{\xi})$, containing the values of the back-transformed linear predictor evaluated at every integration point.

5.2.1 Transforming covariates

Due to the way in which Stan works, sampling is most efficient if the posterior distribution of the parameters is approximately uncorrelated and unit scaled (standard deviation of 1). The following techniques aim to achieve this by transforming the covariates in some manner.

5.2.1.1 Centring and scaling

The simplest change we can make to the model is to centre and scale the (continuous) covariates. These are common practice in both Bayesian and frequentist settings; the Stan User's Guide covers these topics in Section 23.12 (Stan Development Team 2018).

Centring involves subtracting an overall mean value from each covariate. This effectively unlinks the corresponding regression slopes from the "intercepts" (here a generalised notion of an intercept, relating to any discrete parameters), removing posterior correlation between these two sets of parameters. In this context, the slopes are the components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{2,k}$ corresponding to continuous covariates, and the "intercepts" are the study intercepts μ_j, individual-level treatment effects γ_k, and components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{2,k}$ for any categorical covariates. The interpretation of the slope parameters is unchanged by centring, but the interpretation of the "intercept" parameters is now with respect to the mean values (instead of zero). (As a byproduct of centring, these "intercept" parameters may become more interpretable: for example, we are rarely interested in an individual with zero weight or age.)

Scaling involves dividing each of the covariates by a fixed value, either chosen given observed covariate ranges or clinical knowledge, or set to the observed standard deviation (in which case this is called *standardising*). For ML-NMR, since we have only AgD available from some studies and an overall standard deviation is not straightforward to calculate, the simplest approach is the former. We are not concerned with choosing scaling values that result in precisely unit-scaled parameters: a sensible approximate scaling is often good enough.

Models fitted with and without centring and scaling of covariates are equivalent: posterior estimates from a model fit with centred and scaled covariates can be transformed to estimates from a model fit without these transformations, simply by adding and multiplying by the centring and scaling values respectively. However, this equivalence only holds if the different interpretation of the model parameters is accounted for when specifying prior distributions (particularly when the prior distributions are informative).

An additional advantage of using QMC integration is that it allows us to simply transform the integration points $\tilde{\boldsymbol{X}}_{jk}$ *after* they have been obtained, rather than attempting to appropriately transform the AgD marginal distributions and summary statistics before generating the integration points.

QR decomposition 5.2.1.2

Centring and scaling covariates is enough to ensure efficient computation in Stan when the covariates themselves are uncorrelated (or any correlations are minimal). However, when covariates are correlated this induces correlations between the regression parameters in the posterior distribution, which hinders efficient sampling. Instead, we can reparameterise the model using a QR decomposition of the design matrix, resulting in a transformed posterior parameter space that is much more efficient to sample from (see Section 1.2 of the Stan User's Guide, Stan Development Team 2018).

The QR decomposition of a general $r \times c$ matrix $\boldsymbol{X}$ (with $r \geq c$) is $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ is an $r \times c$ orthogonal matrix and $\boldsymbol{R}$ is a $c \times c$ upper-triangular matrix (Golub and Van Loan 1996, Section 5.2).[2] Since we are using QMC integration to evaluate the integral for the aggregate-level model, we apply the QR decomposition to the augmented design matrix $\boldsymbol{X}^*$ formed by joining the numerical integration "design" matrices $\tilde{\boldsymbol{X}}_{jk}$ for the AgD below the IPD design matrix like so:

$$\boldsymbol{X}^* = \begin{pmatrix} \boldsymbol{X} \\ \tilde{\boldsymbol{X}}_{11} \\ \vdots \\ \tilde{\boldsymbol{X}}_{JK} \end{pmatrix}. \tag{5.9}$$

The QR decomposition of this augmented design matrix is then

$$\boldsymbol{X}^* = \boldsymbol{Q}\boldsymbol{R},$$

which we further rescale as

$$\begin{aligned} &= \left(\sqrt{N^* - 1}\boldsymbol{Q}\right)\left(\frac{1}{\sqrt{N^* - 1}}\boldsymbol{R}\right) \\ &= \boldsymbol{Q}^*\boldsymbol{R}^*, \end{aligned} \tag{5.10}$$

where N^* is the number of rows of $\boldsymbol{X}^*$. This ensures that the columns of $\boldsymbol{Q}^*$ are unit scaled (have standard deviation 1). We partition $\boldsymbol{Q}^*$ into sub-matrices

[2] This is the *thin* QR decomposition. There is a corresponding *fat* QR decomposition where $\boldsymbol{Q}_{\text{fat}}$ is $r \times r$ and $\boldsymbol{R}_{\text{fat}}$ is $r \times c$. The two are related by $\boldsymbol{Q}_{\text{fat}} = [\boldsymbol{Q}, \boldsymbol{Q}_2]$, with $r - c$ additional columns $\boldsymbol{Q}_2$, and $\boldsymbol{R}_{\text{fat}} = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix}$. The fat QR decomposition is not useful here.

consisting of the corresponding rows from the IPD, and from the numerical integration points for each AgD arm, echoing the construction of $\boldsymbol{X}^*$ in (5.9):

$$\boldsymbol{Q}^* = \begin{pmatrix} \boldsymbol{Q}^*_{\text{Ind}} \\ \tilde{\boldsymbol{Q}}^*_{11} \\ \vdots \\ \tilde{\boldsymbol{Q}}^*_{JK} \end{pmatrix}. \tag{5.11}$$

We then reparameterise the FE ML-NMR model (5.8) as

Individual:

$$y_{ijk} \sim \pi_{\text{Ind}}(\theta_{ijk}) \tag{5.12a}$$
$$g(\boldsymbol{\theta}) = \boldsymbol{Q}^*_{\text{Ind}}\boldsymbol{\xi}^* \tag{5.12b}$$

Aggregate:

$$y_{\bullet jk} \sim \pi_{\text{Agg}}(\theta_{\bullet jk}) \tag{5.12c}$$
$$\theta_{\bullet jk} \approx \frac{1}{\tilde{N}} \sum g^{-1}\left(\tilde{\boldsymbol{Q}}^*_{jk}\boldsymbol{\xi}^*\right) \tag{5.12d}$$

where $\boldsymbol{\xi}^* = \boldsymbol{R}^*\boldsymbol{\xi}$. The parameters on the original scale can be recovered by the back-transformation $\boldsymbol{\xi} = (\boldsymbol{R}^*)^{-1}\boldsymbol{\xi}^*$.

The use of QMC integration allows us to easily implement a QR parameterisation, by considering the augmented design matrix $\boldsymbol{X}^*$ as a whole. However, there are some caveats to the QR parameterisation. Firstly, since the QR decomposition involves the numerical integration points $\tilde{\boldsymbol{X}}_{jk}$, the decomposition and resulting transformed parameter space may vary depending on the number of numerical integration points $\tilde{N}$ used. Whilst this will not introduce bias (the underlying model remains identical and can always be recovered), this may mean that the resulting transformed parameter space is not as optimal as possible. Centring and scaling the covariates before the QR decomposition is applied may help mitigate this issue. Similarly, we should specify prior distributions on the original parameters $\boldsymbol{\xi}$ rather than on the transformed parameters $\boldsymbol{\xi}^*$, since the interpretation of the latter is not only unnatural but is liable to change with different choices of $\tilde{N}$. Furthermore, the QR parameterisation only works well when non- or weakly-informative prior distributions are used; informative prior distributions can induce strong correlations in the posterior distribution of the transformed parameters $\boldsymbol{\xi}^*$, hindering sampling, in which case the original parameterisation may perform better.

Exchangeable effect modifier coefficients 5.2.2

The ML-NMR model with exchangeable effect modifier coefficients (Section 4.6.1) specifies that, for a set of treatments $k \in \mathcal{T}$, the effect modifier interaction coefficients $\beta_{2,k;l}$ have a distribution

$$\beta_{2,k;l} \sim \text{N}(m_{\beta_2,l}, \sigma^2_{\beta_2,l}), \tag{5.13}$$

for each covariate l. Fitting hierarchical structures such as these can be problematic for MCMC sampling algorithms, leading to slow sampling and biased posterior estimates, as the strong dependence between the hierarchical mean and standard deviation induces a phenomenon known as *Neal's funnel* (after Neal 2003). (In Stan, this manifests as divergent transition errors clustered around small values of the hierarchical standard deviation; see Section 5.2.5) The parameterisation in (5.13) is commonly referred to as *centred*; an alternative is the *non-centred* parameterisation, which instead samples over a parameter $\beta^*_{2,k;l}$ with a standard Normal distribution and then back-transforms to the required distribution (Betancourt and Girolami 2013; Papaspiliopoulos et al. 2007):

$$\begin{aligned} \beta^*_{2,k;l} &\sim \text{N}(0,1), \\ \beta_{2,k;l} &= m_{\beta_2,l} + \beta^*_{2,k;l}\sigma^2_{\beta_2,l}. \end{aligned} \tag{5.14}$$

Neal's funnel and its remedy through the non-centred parameterisation are discussed in Section 21.7 of the Stan User's Guide (Stan Development Team 2018). The non-centred parameterisation can greatly improve sampling efficiency, particularly when the number of treatments in the set $\mathcal{T}$ is small (as is common); we use this parameterisation when coding the ML-NMR model in Stan (Appendix A). However, using both the non-centred parameterisation and QR decomposition (Section 5.2.1) together is not straightforward due to the different transformations applied to the same set of EM interaction parameters. In practice, we find that the non-centred parameterisation is often necessary for well-behaved sampling (i.e. without divergent transition errors, see Section 5.2.5): we therefore use the non-centred parameterisation and simply centre and scale the covariates instead of using the QR decomposition when fitting models with exchangeable EM interactions.

Random effects 5.2.3

In Section 4.6.2, we modified the fixed effect ML-NMR model (given in matrix form in equation (5.8)) to include random treatment effects accounting for any residual heterogeneity. As shown in (4.40), the linear predictor $\eta_{jk}(\boldsymbol{x})$ becomes

$\eta_{jk}(\boldsymbol{x}) = \mu_j + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \delta_{jk}$, where the δ_{jk} are study-specific random effects. Using the reference-treatment notation (Section 1.2.2), we place a random effect δ_{jk} on every non-treatment 1 arm in each study (i.e. setting $\delta_{j1} = 0$). These have a multivariate Normal distribution, with marginal distributions $\delta_{jk} \sim \mathrm{N}(\gamma_k, \tau^2)$ and correlations $\mathrm{cor}(\delta_{ja}, \delta_{jb}) = 0.5$ between random effects on non-treatment 1 arms in the same study (under the assumption of common heterogeneity variance τ^2) (Higgins and Whitehead 1996).

To write these random effects in a vector-matrix form for efficient implementation, let $\boldsymbol{\delta}_j$ be the vector of random effects in study j (again, with no random effect on $k = 1$ arms), let $\boldsymbol{D}_j$ be a design matrix selecting the corresponding treatment effects from the vector $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_K)^\intercal$, and let $\boldsymbol{\Sigma}_{\tau;j}$ be the covariance matrix between the random effects in study j, with τ^2 on the diagonal and $\tau^2/2$ off the diagonal. The random effects distribution for each study is then written as $\boldsymbol{\delta}_j \sim \mathrm{MVN}(\boldsymbol{D}_j\boldsymbol{\gamma}, \boldsymbol{\Sigma}_{\tau;j})$. Joining the vectors of random effects from each study together, the overall random effects vector has distribution

$$\boldsymbol{\delta} \sim \mathrm{MVN}\big(\boldsymbol{D}\boldsymbol{\gamma}, \boldsymbol{\Sigma}_\tau\big), \tag{5.15}$$

where

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \vdots \\ \boldsymbol{\delta}_J \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 \\ \vdots \\ \boldsymbol{D}_J \end{pmatrix}, \quad \boldsymbol{\Sigma}_\tau = \begin{pmatrix} \boldsymbol{\Sigma}_{\tau;1} & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_{\tau;J} \end{pmatrix}.$$

This parameterisation of the random effects is a centred parameterisation, and hence will suffer similar issues to those discussed in Section 5.2.2, particularly when the number of studies is small. Instead, we implement a non-centred parameterisation of (5.15), as discussed in Section 21.7 of the Stan User's Guide (Stan Development Team 2018). In Section 5.2.2, we described the non-centred parameterisation of exchangeable EM interactions, where the hierarchical distribution was univariate Normal (5.13); here, deriving the non-centred parameterisation of the random effects follows a similar process, except now the hierarchical distribution is multivariate Normal (5.15).

To derive the non-centred parameterisation, we need to find a suitable linear transformation relating the desired random effects $\boldsymbol{\delta}$ to independent standard Normal parameters $\boldsymbol{\delta}^*$ (over which we can efficiently sample). One way of achieving this is to use Cholesky decomposition (Golub and Van Loan 1996, Section 4.2). Firstly, we write the block-diagonal covariance matrix $\boldsymbol{\Sigma}_\tau$ in terms of a block-diagonal correlation matrix $\boldsymbol{\Psi}$ (where, under the assumption of common heterogeneity variance, each block has 1s on the diagonal and 0.5s

elsewhere) and the common heterogeneity variance τ^2:

$$\boldsymbol{\Sigma}_\tau = \tau^2 \boldsymbol{\Psi}. \tag{5.16}$$

We take the Cholesky decomposition $\boldsymbol{L}$ of the correlation matrix,

$$\boldsymbol{\Psi} = \boldsymbol{L}\boldsymbol{L}^\intercal, \tag{5.17}$$

and thus decompose the covariance matrix as

$$\boldsymbol{\Sigma}_\tau = \tau \boldsymbol{L}\boldsymbol{L}^\intercal \tau. \tag{5.18}$$

Using (5.18), we then write the non-centred parameterisation as

$$\begin{aligned} \boldsymbol{\delta}^* &\sim \text{MVN}(\mathbf{0}, \boldsymbol{I}) \\ \boldsymbol{\delta} &= \boldsymbol{D}\boldsymbol{\gamma} + \tau \boldsymbol{L}\boldsymbol{\delta}^*, \end{aligned} \tag{5.19}$$

where $\mathbf{0}$ is a vector of zeros and $\boldsymbol{I}$ is the identity matrix.

Finally, we use this non-centred random effects parameterisation to write the random effects ML-NMR model in matrix form. Notationally (and computationally), it is convenient to define a vector $\boldsymbol{\delta}_0$ of *zero mean* random effects augmented with fixed zeros for arms with $k = 1$, where the elements of $\boldsymbol{\delta}_0$ are

$$\delta_{0;jk} = \begin{cases} 0 & \text{if } k = 1 \\ \tau \left[\boldsymbol{L}\boldsymbol{\delta}^*\right]_{jk} & \text{if } k > 1 \end{cases} \tag{5.20}$$

and we have $\delta_{jk} = \gamma_k + \delta_{0;jk}$. The full RE ML-NMR model parameterised with non-centred random effects is then

Individual:

$$y_{ijk} \sim \pi_{\text{Ind}}(\theta_{ijk}) \tag{5.21a}$$

$$g(\boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\delta}_{0;\text{Ind}} \tag{5.21b}$$

$$\boldsymbol{\xi} = \left(\mu_1, \ldots, \mu_J, \boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_{2,2}^\intercal, \ldots, \boldsymbol{\beta}_{2,K}^\intercal, \gamma_2, \ldots, \gamma_K\right)^\intercal \tag{5.21c}$$

Random Effects:

$$\boldsymbol{\delta}^* \sim \text{MVN}(\mathbf{0}, \boldsymbol{I}) \tag{5.21d}$$

$$\delta_{0;jk} = \begin{cases} 0 & \text{if } k = 1 \\ \tau \left[\boldsymbol{L}\boldsymbol{\delta}^*\right]_{jk} & \text{if } k > 1 \end{cases} \tag{5.21e}$$

Aggregate:

$$y_{\bullet jk} \sim \pi_{\text{Agg}}(\theta_{\bullet jk}) \tag{5.21f}$$

$$\theta_{\bullet jk} \approx \frac{1}{\tilde{N}} \sum g^{-1}\left(\tilde{\boldsymbol{X}}_{jk}\boldsymbol{\xi} + \delta_{0;jk}\right) \tag{5.21g}$$

We set $\delta_{j1} = \gamma_1 = 0$ and $\boldsymbol{\beta}_{2,1} = \mathbf{0}$, and let $\boldsymbol{\delta}_{0;\text{Ind}}$ be the elements of $\boldsymbol{\delta}_0$ corresponding to IPD study arms. The IPD design matrix $\boldsymbol{X}$ and the numerical integration "design" matrices $\boldsymbol{X}_{jk}$ are the same as in the FE model (5.8). Prior distributions are specified for the parameters in $\boldsymbol{\xi}$ and for τ.

We also apply the covariate transformation techniques described in Section 5.2.1 here, either centring and scaling the covariates or using the QR decomposition. Unlike with the exchangeable EM interactions model (Section 5.2.2), the QR decomposition is straightforward to apply here: the random effects terms are separated out from the design matrix in equations (5.21b) and (5.21g), which we rewrite as

$$g(\boldsymbol{\theta}) = \boldsymbol{Q}^*_{\text{Ind}}\boldsymbol{\xi}^* + \boldsymbol{\delta}_0 \tag{5.22a}$$

and

$$\theta_{\bullet jk} \approx \frac{1}{\tilde{N}} \sum g^{-1}\left(\tilde{\boldsymbol{Q}}^*_{jk}\boldsymbol{\xi}^* + \delta_{0;jk}\right) \tag{5.22b}$$

respectively.

5.2.4 Specifying initial values

MCMC algorithms require the user to specify initial values for the sampling algorithm. Software packages such as WinBUGS offer to randomly generate initial values; however, these are generated from the given prior distribution and may be extreme (particularly with non-informative prior distributions) which can cause numerical issues and/or slow convergence. By comparison, Stan offers to generate initial values within a given interval on an unconstrained scale (e.g. the log scale for parameters constrained to be positive). This removes the possibility of extreme values causing numerical issues, whilst hopefully still providing initial values that are disparate enough across multiple chains to detect issues such as multiple posterior modes. The interval in which Stan generates initial values is set to $[-2, 2]$ by default, but whether this is appropriate depends on the scaling of outcomes and covariates.

When using the two-parameter Binomial approximation to the Poisson Binomial aggregate likelihood (Section 4.2.1), we must ensure that Stan does not choose invalid initial values—i.e. those that correspond to an adjusted number of individuals N'_{jk} less than the observed number of events $y_{\bullet jk}$. To implement this restriction, we use a `reject` statement to reject the generated initial values (and force Stan to try again) when $N'_{jk} < y_{\bullet jk}$. In this way, we can still rely on Stan's inbuilt mechanism for generating initial values. This condition should always be met once Stan is sampling from the posterior

distribution: any rejections during sampling indicate serious conflict between the model and the data, and any results should not be trusted in such case.

Checking convergence and other diagnostics 5.2.5

The MCMC sampling algorithm used by Stan is different to that used by other Bayesian software packages such as WinBUGS or JAGS. Stan implements a form of Hamiltonian Monte Carlo (HMC), specifically the No U-Turn Sampler (NUTS; Hoffman and Gelman 2011). Instead of producing samples in a random walk, with acceptance probabilities determined by the ratio of the posterior density at successive points, HMC produces samples by simulating an analogous physical system under Hamiltonian dynamics, where the sampler has potential energy (determined by the posterior density) and momentum (which is tuned for efficiency) as it moves through the sample space (for an introduction, see Neal 2012). Whilst the computational cost for obtaining each posterior sample is greater, the resulting samples themselves show much lower autocorrelation than those from WinBUGS or JAGS. Far fewer iterations are therefore needed to obtain a suitable effective sample size (or corresponding Monte Carlo error); typically only a few thousand iterations in Stan are required to give the same effective sample size as tens or even hundreds of thousands from WinBUGS or JAGS. (It is even possible for Stan to produce anticorrelated samples, in which case the effective sample size may be larger than the number of iterations.)

General MCMC diagnostics are applicable when fitting Stan models. For example, convergence may be assessed by running multiple parallel chains from disparate initial values, and then calculating the value of $\hat{R}$, also known as the potential scale reduction factor or Gelman-Rubin statistic (Gelman and Rubin 1992; Gelman et al. 2013a, pp. 284–285; see also Section 15.3 of the Stan Language Reference Manual, Stan Development Team 2018). $\hat{R}$ assesses convergence by comparing the between- and within-chain variances to determine whether the different chains are sampling from the same posterior distribution. ($\hat{R}$ can be interpreted as an estimate of the overdispersion of the samples from the true posterior distribution, hence it is sometimes referred to as the potential scale reduction factor.) We wish $\hat{R}$ to be close to 1; typically values greater than 1.1 are considered too large and indicate non-convergence, and a larger number of iterations may be required (or there may be other issues, such as chains sampling from different posterior modes).

Stan also provides a number of advanced diagnostics to help indicate problems with sampling. One key diagnostic is the presence of *divergent*

transitions; these mean that the NUTS algorithm used by Stan has detected that it was not able to suitably sample from a region of the posterior distribution (Betancourt and Girolami 2013; see also Section 14.5 of the Stan Language Reference Manual, Stan Development Team 2018). For example, if random effects are implemented using the centred parameterisation instead of the non-centred parameterisation (see Section 5.2.3), then divergent transitions are likely to occur at small values of τ (and the posterior estimate of τ will be biased upward as a result). Stan will raise a warning if divergent transitions occur, and their presence means that the posterior estimates may be biased. It may be possible to remedy divergent transitions by reducing the step size used in the NUTS algorithm, forcing the sampler to move more slowly;[3] otherwise a more suitable model parameterisation may be required, such as those discussed in the previous sections.

These diagnostics are all provided in the `rstan` R package which is used to run Stan, and printed as part of the default output. Alternatively, the `shinystan` package provides a graphical user interface for interactive convergence checking and sampling diagnostics.

5.3 Model fit and comparison

As we described in Section 1.2.6 for standard NMA, in a Bayesian framework model fit is often assessed using the residual deviance D_{res}, and different models are often compared using the Deviance Information Criterion (DIC) (Dias et al. 2011c; Spiegelhalter et al. 2002). We can follow this approach for ML-NMR models also, as long as the form of the aggregate likelihood is known. (We discuss model comparison for models with general likelihoods where the aggregate likelihood may not have an explicit form later in Section 7.2.) The form of the residual deviance (and thus p_D and DIC) is dictated by the form of the likelihood, and so we use one form for the IPD corresponding to the individual-level likelihood, and one form for the AgD corresponding to the aggregate-level likelihood.

The residual deviance for a data point is defined as the deviance (-2 times the log likelihood) under the current model, minus the deviance under a saturated model where every data point is perfectly predicted (McCullagh and Nelder 1989). For the IPD, we denote the residual deviance contributions by $D_{\text{res};ijk}$, and for the AgD we denote the contributions by $D_{\text{res};\bullet jk}$. The total

[3] This is achieved by increasing target acceptance rate, which is set via the Stan control argument `adapt_delta`. Increasing the target acceptance rate from its default value of 0.8, to a higher value such as 0.95 or 0.99, results in a smaller step size.

residual deviance, summing over the contributions from each data point, is thus

$$D_{\text{res}} = \sum_{\text{IPD}\ j}\sum_{k}\sum_{i} D_{\text{res};ijk} + \sum_{\text{AgD}\ j}\sum_{k} D_{\text{res};\bullet jk}. \tag{5.23}$$

Table 5.1 lists residual deviance formulae for some common individual and aggregate likelihoods.

For the purposes of checking absolute model fit, we are interested in the posterior distribution of D_{res}, and of $D_{\text{res};ijk}$ and $D_{\text{res};\bullet jk}$ for each data point (individual or aggregate, respectively). The posterior mean of the residual deviance, $\mathbb{E}(D_{\text{res}})$, can be compared to the number of independent data points to check if model fit might be improved: the two will be approximately equal under a well-fitting model assuming approximate Normality (which may not hold in practice, e.g. for binary data in small samples or with event probabilities far from 0.5) (Spiegelhalter et al. 2002). The posterior distributions of the residual deviance contributions $D_{\text{res};ijk}$ or $D_{\text{res};\bullet jk}$ from each data point can be plotted (e.g. posterior mean and 95% Credible Interval), to identify any poorly-fitting observations (those with contributions much greater than 1). It may be useful to check model fit for the IPD and AgD separately, as well as overall: for example by splitting the total residual deviance (5.23) into totals for the IPD and AgD, or by distinguishing between IPD and AgD in plots of residual deviance contributions.

The DIC is calculated using equation (1.34) as

$$\text{DIC} = \mathbb{E}\left(D_{\text{res}}\right) + p_D, \tag{5.24}$$

penalising the residual deviance D_{res} by a measure of the effective number of parameters, p_D. Following Welton et al. (2012, p. 126), we calculate p_D as the difference between the posterior mean of D_{res} and the value of D_{res} calculated at the posterior mean of the fitted values (i.e. replacing $\hat{y}_{ijk}$ and $\hat{y}_{\bullet jk}$ by $\mathbb{E}(\hat{y}_{ijk})$ and $\mathbb{E}(\hat{y}_{\bullet jk})$ respectively in the formulae in Table 5.1):

$$p_D = \mathbb{E}\left(D_{\text{res}}\right) - D_{\text{res}}\big|_{\mathbb{E}(\hat{y})}. \tag{5.25}$$

When comparing a set of candidate models, lower values of DIC are preferred; typically differences of less than 3 are considered small, and differences of more than 5 are considered substantial (Lunn et al. 2010, pp. 165–167; Dias et al. 2018, p. 69). If differences in DIC are small we would typically prefer the model with the smallest effective number of parameters.

Table 5.1 Residual deviance contributions for some common individual and aggregate-level likelihoods. For full details on each likelihood and its notation, see the definitions in Section 4.2.

Likelihood	Model prediction	Residual deviance contribution
Normal individual, Normal aggregate *		
$y_{ijk} \sim \mathrm{N}(\theta_{ijk}, \sigma^2_{jk})$	$\hat{y}_{ijk} = \theta_{ijk}$	$D_{\mathrm{res};ijk} = \frac{(y_{ijk} - \hat{y}_{ijk})^2}{\sigma^2_{jk}}$
$y_{\bullet jk} \sim \mathrm{N}(\theta_{\bullet jk}, s^2_{jk})$	$\hat{y}_{\bullet jk} = \theta_{\bullet jk}$	$D_{\mathrm{res};\bullet jk} = \frac{(y_{\bullet jk} - \hat{y}_{\bullet jk})^2}{s^2_{jk}}$
Bernoulli individual, one-parameter Binomial aggregate		
$y_{ijk} \sim \mathrm{Bern}(p_{ijk})$	$\hat{y}_{ijk} = p_{ijk}$	$D_{\mathrm{res};ijk} = -2\left(y_{ijk}\log\hat{y}_{ijk} + (1-y_{ijk})\log(1-\hat{y}_{ijk})\right)$
$y_{\bullet jk} \sim \mathrm{Bin}(N_{jk}, \bar{p}_{jk})$	$\hat{y}_{\bullet jk} = N_{jk}\bar{p}_{jk}$	$D_{\mathrm{res};\bullet jk} = 2\left(y_{\bullet jk}\log\left(\frac{y_{\bullet jk}}{\hat{y}_{\bullet jk}}\right) + (N_{jk} - y_{\bullet jk})\log\left(\frac{N_{jk}-y_{\bullet jk}}{N_{jk}-\hat{y}_{\bullet jk}}\right)\right)$
Bernoulli individual, two-parameter Binomial aggregate †		
$y_{ijk} \sim \mathrm{Bern}(p_{ijk})$	$\hat{y}_{ijk} = p_{ijk}$	$D_{\mathrm{res};ijk} = -2\left(y_{ijk}\log\hat{y}_{ijk} + (1-y_{ijk})\log(1-\hat{y}_{ijk})\right)$
$y_{\bullet jk} \sim \mathrm{Bin}(N'_{jk}, \bar{p}'_{jk})$	$\hat{y}_{\bullet jk} = N'_{jk}\bar{p}'_{jk}$	$D_{\mathrm{res};\bullet jk} = 2\left(y_{\bullet jk}\log\left(\frac{y_{\bullet jk}}{\hat{y}_{\bullet jk}}\right) + (N'_{jk} - y_{\bullet jk})\log\left(\frac{N'_{jk}-y_{\bullet jk}}{N'_{jk}-\hat{y}_{\bullet jk}}\right)\right)$
Poisson individidual, Poisson aggregate		
$y_{ijk} \sim \mathrm{Pois}(\lambda_{ijk}E_{ijk})$	$\hat{y}_{ijk} = \lambda_{ijk}E_{ijk}$	$D_{\mathrm{res};ijk} = 2\left((\hat{y}_{ijk} - y_{ijk}) + y_{ijk}\log\left(\frac{y_{ijk}}{\hat{y}_{ijk}}\right)\right)$
$y_{\bullet jk} \sim \mathrm{Pois}(\lambda_{\bullet jk}E_{\bullet jk})$	$\hat{y}_{\bullet jk} = \lambda_{\bullet jk}E_{\bullet jk}$	$D_{\mathrm{res};\bullet jk} = 2\left((\hat{y}_{\bullet jk} - y_{\bullet jk}) + y_{\bullet jk}\log\left(\frac{y_{\bullet jk}}{\hat{y}_{\bullet jk}}\right)\right)$

* When calculating p_D, set σ^2_{jk} to posterior median (see Section 5.3.1.1); s^2_{jk} assumed known.

† When calculating p_D, set N'_{jk} to posterior median (see Section 5.3.1.2).

Considerations for likelihoods with more than one parameter 5.3.1

Calculation of the residual deviance, p_D, and DIC is straightforward in most cases when the aggregate likelihood has a known form. However, when the likelihood has more than one unknown parameter the calculations become more complicated. In this case, calculation of p_D requires choosing suitable plug-in values for each parameter, and if multiple parameters are modelled the saturated deviance is also not uniquely defined. We now consider two common cases where these issues arise.

Considerations for a Normal likelihood with unknown variance 5.3.1.1

With a Normal individual-level likelihood, $y_{ijk} \sim \mathrm{N}(\theta_{ijk}, \sigma^2_{jk})$, a model is placed on the mean parameter θ_{ijk} and the variance σ^2_{jk} is unknown. (At the aggregate level we take s^2_{jk} as data, assumed known, so we do not have the same issue.) When calculating p_D, we evaluate the residual deviance at the posterior mean of the fitted values; we therefore need to also plug in a suitable value for σ^2_{jk}. A typical choice is the posterior median, $\mathrm{med}(\sigma^2_{jk})$, resulting in

$$D_{\mathrm{res};ijk}\big|_{\mathbb{E}(\hat{y}_{ijk}),\mathrm{med}(\sigma^2_{jk})} = \frac{\left(y_{ijk} - \mathbb{E}(\hat{y}_{ijk})\right)^2}{\mathrm{med}(\sigma^2_{jk})}. \tag{5.26}$$

In general, when calculating p_D for a likelihood with both location and scale parameters unknown, the location parameter (which is modelled) is determined by the posterior mean of the fitted values, and the scale parameter is fixed at a suitable value such as the posterior median.

Considerations for the two-parameter Binomial likelihood 5.3.1.2

When fitting a model using the two-parameter Binomial approximation to the aggregate Poisson Binomial likelihood (Section 4.2.1), there are two issues that must be considered. Firstly, when calculating the residual deviance the saturated model is not uniquely defined. This is because both likelihood parameters $\bar{p}'_{jk}$ and N'_{jk} are modelled, and there are an infinite number of ways in which $\bar{p}'_{jk}$ and N'_{jk} can be chosen together to perfectly predict the observed number of events $y_{\bullet jk}$. Secondly, calculating p_D requires choosing a suitable plug-in value, in this case for N'_{jk}.

To alleviate the first issue, one possibility is to avoid calculating the saturated deviance altogether. The deviance may be used to calculate p_D and DIC by itself, without subtracting the saturated deviance (Spiegelhalter et al. 2002). This results in the same values for p_D, since the saturated deviance

cancels in equation (5.25), although suitable plug-in values will still need to be chosen (i.e. the second issue still stands). Comparisons of deviance and DIC between candidate models will also be unaffected, since the saturated deviance is the same under all models and cancels out. However, the saturated deviance acts as a standardising term, allowing the residual deviance to be interpreted as an absolute measure of model fit (Spiegelhalter et al. 2002). The need to calculate this standardising term is arguably more important in ML-NMR models than in NMA with only IPD or AgD: since the individual-level and aggregate-level likelihoods are different, the deviance contributions require standardising in order to compare model fit between the IPD and AgD. Therefore we now consider the calculation of the residual deviance for the two-parameter Binomial model.

We choose to calculate the saturated deviance under a model where the average event probability $\bar{p}'_{jk}$ perfectly predicts the observed number of events, for a given posterior sample of N'_{jk}, since in our experience there is less posterior variation in N'_{jk} than $\bar{p}'_{jk}$. This is the formulation given in Table 5.1:

$$D_{\text{res};\bullet jk} = 2\left(y_{\bullet jk}\log\left(\frac{y_{\bullet jk}}{\hat{y}_{\bullet jk}}\right) + (N'_{jk} - y_{\bullet jk})\log\left(\frac{N'_{jk} - y_{\bullet jk}}{N'_{jk} - \hat{y}_{\bullet jk}}\right)\right), \tag{5.27}$$

where $\hat{y}_{\bullet jk} = N'_{jk}\bar{p}'_{jk}$. An alternative choice would be to calculate the saturated deviance under a model with N'_{jk} fixed to equal the actual sample size N_{jk} (i.e. the one-parameter Binomial model), resulting in

$$\begin{aligned} D_{\text{res};\bullet jk} = 2\Bigg(&y_{\bullet jk}\log\left(\frac{y_{\bullet jk}}{\hat{y}_{\bullet jk}}\right) + (N_{jk} - y_{\bullet jk})\log\big(N_{jk} - y_{\bullet jk}\big) \\ &- (N'_{jk} - y_{\bullet jk})\log\Big(N'_{jk} - \hat{y}_{\bullet jk}\Big)\Bigg). \end{aligned} \tag{5.28}$$

Under this formulation the additional flexibility of the two-parameter Binomial is not reflected in the saturated deviance, but when comparing DIC between candidate models the calculated saturated deviances are all the same so cancel out. This cancellation is not guaranteed under (5.27): the calculated saturated deviances are not necessarily the same for all candidate models, since N'_{jk} may be estimated differently under each fitted model.

Then, when calculating the effective number of parameters p_D, we evaluate the residual deviance at the posterior mean of the fitted values. In this case the residual deviance also depends on N'_{jk}, for which we also need to plug in a suitable value. (This is also true of the deviance, so when calculating p_D with the deviance instead of the residual deviance the same choices apply.) We

choose to use the posterior median of the adjusted sample size, $\text{med}(N'_{jk})$, so that (along with the formulation of residual deviance given in (5.27)), we have

$$D_{\text{res};\bullet jk}\big|_{\mathbb{E}(\hat{y}_{\bullet jk}),\text{med}(N'_{jk})} = 2\left(\mathbb{E}(\hat{y}_{\bullet jk})\log\left(\frac{y_{\bullet jk}}{\mathbb{E}(\hat{y}_{\bullet jk})}\right) + (\text{med}(N'_{jk}) - y_{\bullet jk})\log\left(\frac{\text{med}(N'_{jk}) - y_{\bullet jk}}{\text{med}(N'_{jk}) - \mathbb{E}(\hat{y}_{\bullet jk})}\right)\right). \tag{5.29}$$

Another choice would be the unadjusted sample size N_{jk}, which gives

$$D_{\text{res};\bullet jk}\big|_{\mathbb{E}(\hat{y}_{\bullet jk}),\text{med}(N'_{jk})} = 2\left(\mathbb{E}(\hat{y}_{\bullet jk})\log\left(\frac{y_{\bullet jk}}{\mathbb{E}(\hat{y}_{\bullet jk})}\right) + (N_{jk} - y_{\bullet jk})\log\left(\frac{N_{jk} - y_{\bullet jk}}{N_{jk} - \mathbb{E}(\hat{y}_{\bullet jk})}\right)\right). \tag{5.30}$$

It remains to be seen which formulation of D_{resdev}, p_D, and DIC is theoretically most desirable. However, these choices only relate to the aggregate part of the model; since the values of D_{resdev}, p_D, and DIC will be dominated by the individual-level model and its fit to the IPD, the choice of formulation is likely to make very little practical difference.

5.4 Discussion

In this chapter, we have discussed several computational aspects of implementing ML-NMR models in Stan. Obtaining the aggregate-level model using QMC integration with copulæ is a flexible and widely applicable approach, which can be used regardless of the number or form of covariates or the complexity of the model. Using quasi-random sequences for numerical integration improves upon the convergence rates of standard pseudo-random Monte Carlo integration, and retains this performance in high-dimensions. We have used a Gaussian copula here to impart the given correlations on the integration points (Section 5.1). This accounts for linear correlation in the underlying relationships between covariates, but no restrictions are placed on the form of the marginal distributions. In practice, we find that this is a very flexible approach and can account for a wide range of observed relationships between covariates; however, other copulæ could be used to assert covariance structures under different assumptions (Nelsen 2006).

In Section 5.2, we discussed several techniques for improving statistical efficiency, by transforming covariates (Section 5.2.1) and choosing an efficient parameterisation of hierarchical structures (Sections 5.2.2 and 5.2.3). In the

extreme case, we must choose a suitable model parameterisation to avoid pathological sampling behaviour and biased posterior estimates. Whilst we discussed these concepts in the context of implementing ML-NMR in Stan, they are equally applicable to implementations of AgD NMA and IPD network meta-regression, and also to implementations of this family of models in other Bayesian software. A key advantage of Stan in this regard is that pathological behaviour is automatically identified through divergent transitions, and the user can take suitable remedial actions (see Section 5.2.5). In contrast, software such as WinBUGS or JAGS in which these models are more traditionally implemented will not explicitly warn of such pathological behaviour, instead relying on more subtle signs such as high autocorrelation or "sticking" chains.

A significant advantage of Stan over other Bayesian software such as WinBUGS or JAGS is that it is much more statistically efficient: much higher effective sample sizes can be obtained for a set number of iterations, due to the use of Hamiltonian Monte Carlo (Section 5.2.5). Not only can this lead to substantial time savings, but there are further benefits for probabilistic cost-effectiveness analysis where posterior samples are used to evaluate a cost-effectiveness model (Claxton et al. 2005; Critchfield and Willard 1986; Dias et al. 2011b; Doubilet et al. 1985), including avoiding the need for "thinning" and reducing Monte Carlo error in the results.

Residual deviance and DIC are widely-used for checking model fit and for model selection of NMA and meta-regression models in a Bayesian framework (Dias et al. 2011c; Hoaglin et al. 2011; Welton et al. 2012, Chapter 6; Dias et al. 2018, Chapter 3). Other model comparison criteria besides DIC have been proposed in the literature, each with different relative merits (for an overview, see Gelman et al. 2013b), but have yet to be widely used in the NMA literature. In Section 5.3, we extended the residual deviance and DIC ideas to the ML-NMR framework by considering the contributions from the individual and aggregate parts of the model. When the aggregate-level likelihood has a known form this approach is straightforward, although there are complications when the likelihood has more than one unknown parameter (such as the two-parameter Binomial, or Normal with unknown variance). In such cases, calculation of p_D requires choosing suitable plug-in values for each parameter, and if multiple parameters are modelled the saturated deviance is not simple to determine (see Section 5.3.1). We discuss alternative approaches for when the aggregate-level likelihood has an unknown form later in Section 7.2—namely the use of approximate leave-one-out (LOO) cross validation and the resulting information criterion LOOIC (analogous to DIC), as proposed by (Vehtari

et al. 2016). Calculating the LOOIC requires only posterior samples of the log likelihood contribution for each data point (over the posterior distribution of the model parameters)—which are always well-defined—and thus avoids the need to specify the form of the likelihood. Indeed, LOOIC also avoids the need to specify the saturated deviance or to plug in posterior estimates to calculate p_D, and so appears an attractive alternative to DIC when faced with multi-parameter likelihoods. However, the approximations used in calculating LOOIC break down when some data points are highly influential, or effectively saturated in the model (i.e. a parameter informed by only a single data point). Unfortunately, these scenarios are common in ML-NMR (and AgD NMA), since often there are treatment comparisons in a network only informed by a small number of studies or only one study. In such cases, these AgD studies will be highly influential on the posterior distribution of the corresponding treatment effect, and the approximate LOO approach breaks down. Some solutions are discussed in Section 7.2, but these are non-trivial to implement. For this reason we have continued to focus on DIC in this chapter, and we use DIC for the examples in the following Chapter 6. However, the use of approximate LOO and LOOIC for ML-NMR and AgD NMA is an interesting area for further research.

Stan code implementing ML-NMR models in an efficient manner using the techniques described in this chapter is given in Appendix A. Whilst the provided Stan code is general for networks of any size and can be readily adapted to other likelihoods, implementation still requires familiarity with Stan. Future work is to develop an R package that provides a user-friendly interface to fitting ML-NMR models (and, by extension, IPD and AgD NMA) in Stan. This package would also simplify the generation of integration points for QMC integration, produce plots of empirical integration error (as described in Section 5.1.2), provide model fit and comparison statistics such as residual deviance and DIC, and produce graphical and tabular summaries of results. Such a package would streamline the process of implementing ML-NMR models, and make the methods accessible to a wider range of users.

Chapter 6

Applied example of ML-NMR: Plaque psoriasis

In this chapter, we apply Multilevel Network Meta-Regression to a real example, and compare with current methods. Three treatments for moderate-to-severe plaque psoriasis were compared with placebo over 12 weeks in four phase 3 trials. In UNCOVER-1, patients were randomised to receive placebo (PBO), ixekizumab every 2 weeks (IXE Q2W), or ixekizumab every 4 weeks (IXE Q4W) (Gordon et al. 2016). In UNCOVER-2 and UNCOVER-3, patients were randomised to receive placebo, etanercept (ETN), ixekizumab Q2W, or ixekizumab Q4W (Gordon et al. 2016; Griffiths et al. 2015). In FIXTURE, patients were randomised to receive placebo, secukinumab 150 mg (SEC 150), secukinumab 300 mg (SEC 300), or etanercept (Langley et al. 2014). Figure 6.1 displays the resulting treatment network formed by the four studies. IPD were available for the three UNCOVER trials. Outcomes of interest include success/failure to achieve 75%, 90%, and 100% improvement on the Psoriasis Area and Severity Index (PASI) scale (denoted PASI 75, PASI 90, and PASI 100 respectively) at twelve weeks. Information on five clinically-relevant covariates thought to be potential effect modifiers is available on individuals in the UNCOVER trials, and summary statistics on the same covariates were extracted from the FIXTURE trial (Langley et al. 2014). Table 6.1 summarises the distribution of these at baseline in each trial.

We begin by performing a simple "targeted comparison" (see Section 4.5.2) between ixekizumab Q2W and secukinumab 300 mg (the licensed dosages) via the etanercept common comparator, comparing the results with a reanalysis of a previously published MAIC (Strober et al. 2016). However, this analysis ignores information from multiple other treatment arms and common comparators.

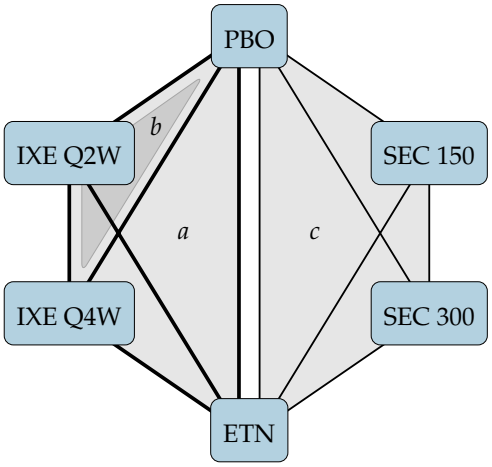


Figure 6.1 The UNCOVER (Gordon et al. 2016; Griffiths et al. 2015) and FIXTURE (Langley et al. 2014) trials form a network of six treatments. Shading indicates comparisons made in: (*a*) UNCOVER-2 and UNCOVER-3; (*b*) UNCOVER-1; (*c*) FIXTURE. The thick and thin lines represent availability of IPD and AgD on a comparison respectively. PBO = placebo, IXE = ixekizumab, SEC = secukinumab, ETN = etanercept. IXE and SEC were each investigated with two different dosing regimens.

Using ML-NMR, we then extend the analysis to incorporate evidence from all trial arms. Finally, we extend the analysis using the additional IPD and AgD studies, allowing us to assess the shared EM assumption and investigate residual heterogeneity. All analyses were performed using R (R Core Team 2018) and Stan (Carpenter et al. 2017).

Table 6.1 Baseline covariate summaries from the UNCOVER and FIXTURE trials, over all trial arms. Reported sample size for UNCOVER-2 and 3 after removing two individuals from each study with missing weight. Statistics are mean (SD) unless otherwise specified.

	UNCOVER-1 (N = 1296)	UNCOVER-2 (N = 1219)	UNCOVER-3 (N = 1339)	FIXTURE (N = 1306)
Age, years	45.7 (12.9)	45.0 (13.0)	45.7 (13.1)	44.5 (12.9)
* Body surface area, per cent	27.7 (17.3)	26.0 (16.5)	28.3 (17.1)	34.4 (18.9)
* Duration of psoriasis, years	19.6 (11.9)	18.7 (12.5)	18.2 (12.2)	16.5 (12.0)
Baseline PASI score	20.1 (8.0)	19.6 (7.2)	20.9 (8.2)	23.7 (10.2)
* Previous systemic treatment (%)	71.3	64.2	57.1	64.0
* Psoriatic arthritis (%)	26.3	23.6	20.5	14.7
Male (%)	68.1	67.0	68.2	71.1
* Weight, kg	92.2 (23.8)	91.6 (22.2)	91.2 (23.5)	83.3 (20.8)

* Covariate considered a potential effect modifier, to be included in population adjustment.

A simple population-adjusted indirect comparison 6.1

A previous MAIC sought to create a population-adjusted indirect comparison between ixekizumab Q2W and secukinumab 300 mg via etanercept, adjusting for the baseline covariates in Table 6.1, using the data from UNCOVER-2 and 3 and FIXTURE (Strober et al. 2016). Data from UNCOVER-1 could not be used, as this study did not include an etanercept arm. We recreate the MAIC analysis of Strober et al., and compare with a "targeted comparison" performed using ML-NMR (as in Section 4.5.2). Figure 6.2 shows this comparison in the context of the full network formed by the UNCOVER and FIXTURE trials (which was shown in Figure 6.1). We focus on the PASI 75 outcome for the analyses in this chapter, since the more demanding PASI 90 and PASI 100 outcomes present difficulties for estimation with small numbers of observed events. (This issue is resolved later in Chapter 7 using a joint multinomial model for the three PASI outcomes.) In this section, we will refer to etanercept as treatment A, ixekizumab Q2W as B, and secukinumab 300 mg as C.

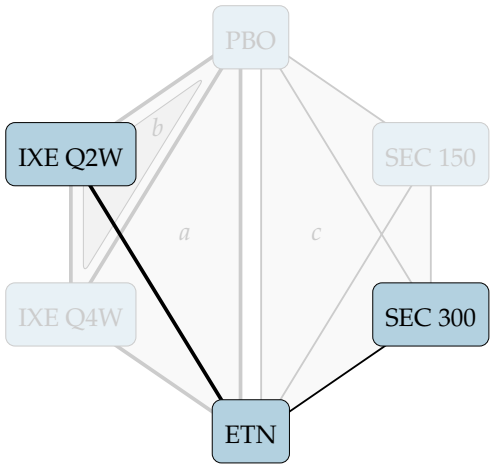


Figure 6.2 The comparison between ixekizumab Q2W and secukinumab 300 mg targeted by a previous MAIC, using etanercept as a common comparator, highlighted as a subset of the full network formed by the UNCOVER and FIXTURE trials (Figure 6.1). The thick and thin lines represent availability of IPD and AgD on a comparison respectively. IXE = ixekizumab, SEC = secukinumab, ETN = etanercept.

Methods 6.1.1

A population-adjusted indirect comparison in the FIXTURE study population is given by $d_{BC(AC)} = d_{AC(AC)} - d_{AB(AC)}$, where the relative effects are defined as standardised mean differences (SMD) on the PASI scale. The relative effect $d_{AC(AC)}$ of secukinumab 300 mg vs. etanercept is estimated by the FIXTURE

study. We will use MAIC and ML-NMR to estimate $d_{AB(AC)}$, the relative effect of ixekizumab Q2W vs. etanercept in the FIXTURE study population.

6.1.1.1 MAIC

The MAIC analysis (Section 2.2.1) matches the mean and standard deviation of the continuous covariates (duration of psoriasis, weight, and body surface area) and the proportion of binary covariates (previous systemic treatment, psoriatic arthritis) in the IPD studies UNCOVER-2 and 3 to the AgD study FIXTURE. Since there is more than one IPD study available, we calculate weights for each study separately to respect randomisation. Following Section 2.2.1, we obtain the weights for UNCOVER-2 and UNCOVER-3 in turn by minimising

$$H_j(\boldsymbol{\alpha_j}) = \sum_k \sum_{i=1}^{N_{jk}} \exp\Big((\boldsymbol{x}^*_{ijk})^\intercal \boldsymbol{\alpha}_j\Big), \tag{6.1}$$

where N_{jk} is the number of individuals in study j on treatment k, and $\boldsymbol{x}^*_{ijk}$ is a vector of covariate moments for individual i in study j on treatment k, centred around the covariate moments in the FIXTURE trial. That is, for a continuous covariate with mean m and standard deviation s in the FIXTURE trial, we include both $x_{ijk} - m$ and $x^2_{ijk} - m^2 - s^2$ in the vector $\boldsymbol{x}^*_{ijk}$; for a binary covariate with proportion m in FIXTURE, we include $x_{ijk} - m$. The weights are then given by

$$w_{ijk} = \exp\Big((\boldsymbol{x}^*_{ijk})^\intercal \boldsymbol{\alpha}_j\Big). \tag{6.2}$$

We then estimate $d_{AB(AC)}$, the relative effect of ixekizumab Q2W vs. etanercept in the FIXTURE study population, using the weights. Following Section 3.1.1.2 (also Phillippo et al. 2016), we do this using a weighted one-stage IPD meta-analysis over the UNCOVER trials. This is implemented as a weighted Binomial GLM with a probit link function, fitted using the weights w_{ijk}, with a model including only a treatment indicator and study-specific intercept to account for clustering:

$$\mathbb{E}(y_{ijk}) = \Phi\big(\mu_j + d_{AB(AC)}\mathbb{I}(k = B)\big). \tag{6.3}$$

The standard error of the estimate $\hat{d}_{AB(AC)}$ is calculated using bootstrapping.

The MAIC population-adjusted indirect comparison in the FIXTURE population is then $\hat{d}_{BC(AC)} = \hat{d}_{AC(AC)} - \hat{d}_{AB(AC)}$, where $\hat{d}_{AC(AC)}$ is the relative effect estimate from the FIXTURE trial.

In the interests of a fair comparison, the MAIC implemented here differs slightly from the original analysis of Strober et al. (2016): in the original

analysis the indirect comparison was effected on the risk difference scale, and it was unclear whether randomisation was fully respected in the weighting process. However, the results of our updated analysis are very similar to those reported originally (Strober et al. 2016).

ML-NMR 6.1.1.2

A simple population-adjusted indirect comparison using ML-NMR is conceptually very similar to STC (which is also a regression adjustment approach, see Section 2.2.2). The main differences are that we perform ML-NMR in a Bayesian framework and use numerical integration to produce estimates on transformed scales rather than simulation. To perform a targeted comparison (see Section 4.5.2), we fit a probit regression in the IPD UNCOVER trials and then make predictions into the AgD FIXTURE population. The same set of potentially effect modifying covariates from Table 6.1 are included in the model, each with a main (prognostic) effect and an interaction effect with treatment. (Note that baseline PASI score and body surface area are highly correlated since the PASI score is based on cutpoints of body surface area, so we only include body surface area in the adjustment in line with the previous MAIC.) For each individual in the UNCOVER studies, the binary outcome of success/failure to achieve PASI 75 follows a Bernoulli distribution, with some individual success probability p_{ijk} which is modelled with a probit link function, and the "targeted comparison" (4.33) is written as:

Individual model in UNCOVER trials:

$$y_{ijk} \sim \text{Bern}(p_{ijk}) \tag{6.4a}$$

$$\Phi^{-1}(p_{ijk}) = \eta_{jk}(\boldsymbol{x}_{ijk}) = \mu_j + \boldsymbol{x}_{ijk}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k \tag{6.4b}$$

Predicted relative effect in FIXTURE trial:

$$\hat{d}_{AB(AC)} = \bar{\boldsymbol{x}}_{(AC)}^{\mathsf{T}}(\hat{\boldsymbol{\beta}}_{2,B} - \hat{\boldsymbol{\beta}}_{2,A}) + \hat{\gamma}_B - \hat{\gamma}_A \tag{6.4c}$$

Targeted comparison in FIXTURE trial:

$$\hat{d}_{BC(AC)} = \hat{d}_{AC(AC)} - \hat{d}_{AB(AC)} \tag{6.4d}$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function (CDF), and $\hat{d}_{AC(AC)}$ is the relative effect estimate from the FIXTURE trial. $\eta_{jk}(\boldsymbol{x}_{ijk})$ is the linear predictor for an individual on treatment k in trial j with covariate vector $\boldsymbol{x}_{ijk}$. The coefficients μ_j are study-specific baselines, $\boldsymbol{\beta}_1$ are coefficients

for prognostic variables, and $\boldsymbol{\beta}_{2,k}$ are coefficients for effect modifiers specific to each treatment k. The effect of the k-th treatment (at the individual level), γ_k, is defined with respect to the reference treatment A, and we set $\gamma_A = 0$ and $\boldsymbol{\beta}_{2,A} = \mathbf{0}$. We implement this model in Stan using the `stan_glm` function from the `rstanarm` R package, placing weakly-informative $\text{N}(0, 10^2)$ prior distributions on the QR-transformed parameters (see Section 5.2.1.2). We assess convergence using $\hat{R}$ for each parameter, and check that there are no divergent transitions (see Section 5.2.5).

6.1.2 Results

The results of the MAIC and ML-NMR analyses are very similar (Table 6.2). The MAIC estimate of the relative effect of secukinumab 300 mg vs. ixekizumab Q2W in the FIXTURE population is −0.28 SMD (95% Confidence Interval: −0.56, −0.00). The ML-NMR estimate is −0.26 SMD (95% Credible Interval: −0.53, 0.01), and is slightly more precise than the MAIC estimate. MAIC effective sample sizes were 418 (59.1% of the original 707) and 558 (72.8% of the original 766) in UNCOVER-2 and 3 respectively.

Table 6.2 Results of the MAIC and ML-NMR population-adjusted indirect comparisons in the FIXTURE study population. The uncertainty intervals are 95% Credible Intervals for ML-NMR, and 95% Confidence Intervals for MAIC and the FIXTURE study estimate.

	Method		
Contrast	ML-NMR	MAIC	FIXTURE study
IXE Q2W vs. ETN $\hat{d}_{AB(AC)}$	1.15 (0.99, 1.33)	1.18 (0.99, 1.37)	
SEC 300 vs. ETN $\hat{d}_{AC(AC)}$			0.89 (0.69, 1.10)
SEC 300 vs. IXE Q2W $\hat{d}_{BC(AC)}$	−0.26 (−0.53 0.01)	−0.28 (−0.56, −0.00)	

6.1.3 Limitations

This analysis has several important limitations. Firstly, etanercept was chosen as the common comparator by Strober et al., but there is also a common placebo comparator in the network (Figure 6.1). Etanercept was likely chosen over placebo to increase precision of the indirect comparison, as there are few events in the placebo arms. However, the placebo arms still contain information that we can incorporate in our analysis. This is particularly desirable since excluding

the placebo arms results in ignoring the IPD from an entire trial (UNCOVER-1). Secondly, the network contains other doses of ixekizumab and secukinumab, and we may be interested in comparisons between all treatments at all doses for clinical or regulatory purposes. However, performing multiple population-adjusted indirect comparisons between each pair of treatments does not result in a consistent set of relative effect estimates—in exactly the same way that multiple pairwise meta-analyses do not provide a consistent set of estimates (Caldwell et al. 2005). Finally, the results of these population-adjusted indirect comparisons are valid only for the FIXTURE study population, which may not represent the decision target population (and hence leads the sponsor of the UNCOVER trials to arguing that their competitor's FIXTURE trial is more representative). The results may be generalised to any target population if the shared effect modifier assumption is made for ixekizumab Q2W and secukinumab 300 mg (Section 2.5). However, this assumption is untestable with the data available for this analysis.

Incorporating evidence from all trial arms 6.2

Using ML-NMR it is straightforward to synthesise all of the available data on all treatments from each of the studies (Figure 6.1), and to produce population-adjusted indirect comparisons between any pair of treatments in any chosen target population. We illustrate this by producing estimates for both the FIXTURE and UNCOVER study populations.

Methods 6.2.1

The full ML-NMR model (4.34) in this scenario is as follows. At the individual level, the binary outcome of success/failure to achieve PASI 75 follows a Bernoulli distribution, with some individual success probability p_{ijk} that is modelled with a probit link function in the same manner as the targeted comparison (equation 6.4):

$$y_{ijk} \sim \text{Bern}(p_{ijk}) \tag{6.5a}$$

$$\Phi^{-1}(p_{ijk}) = \Phi^{-1}(\theta_{ijk}) = \eta_{jk}(\boldsymbol{x}_{ijk}) = \mu_j + \boldsymbol{x}_{ijk}^{\intercal}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k \tag{6.5b}$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function (CDF). θ_{ijk} and $\eta_{jk}(\boldsymbol{x}_{ijk})$ are the conditional mean outcome and linear predictor for an individual on treatment k in trial j with covariate vector $\boldsymbol{x}_{ijk}$. The coefficients μ_j are study-specific baselines, $\boldsymbol{\beta}_1$ are coefficients for prognostic variables, and $\boldsymbol{\beta}_{2,k}$ are coefficients for effect modifiers specific to each treatment k. The effect

of the k-th treatment (at the individual level), γ_k, is defined with respect to the reference treatment 1 (here chosen to be placebo), and we set $\gamma_1 = 0$ and $\boldsymbol{\beta}_{2,1} = \mathbf{0}$. The individual-level model (equations 6.5a and 6.5b) is identical to that used in the targeted comparison in the previous section (equations 6.4a and 6.4b). We use the two-parameter Binomial approximation (4.6) to define the aggregate likelihood

$$y_{\bullet jk} \sim \text{Bin}(N'_{jk}, \bar{p}'_{jk}), \tag{6.5c}$$

$$N'_{jk} = \frac{\sum_i p_{ijk}}{\bar{p}'_{jk}} = N_{jk} \frac{\bar{p}^2_{jk}}{\bar{p^2}_{jk}} \tag{6.5d}$$

$$\bar{p}'_{jk} = \frac{\sum_i p^2_{ijk}}{\sum_i p_{ijk}} = \frac{\bar{p^2}_{jk}}{\bar{p}_{jk}} \tag{6.5e}$$

with parameters $\bar{p}_{jk}$ and $\bar{p^2}_{jk}$ modelled by integrating the individual-level model over the covariate distribution $f_{jk}(\cdot)$:

$$\bar{p}_{jk} = \theta_{\bullet jk} = \int_{\mathfrak{x}} \Phi\big(\eta_{jk}(\boldsymbol{x})\big) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \tag{6.5f}$$

$$\bar{p^2}_{jk} = \int_{\mathfrak{x}} \Phi\big(\eta_{jk}(\boldsymbol{x})\big)^2 f_{jk}(\boldsymbol{x})\, d\boldsymbol{x}, \tag{6.5g}$$

where $\theta_{\bullet jk}$ is the marginal mean outcome on treatment k in trial j, and $\mathfrak{x}$ denotes the support of $\boldsymbol{x}$.

To implement the aggregate-level model, we use the QMC integration approach described in Section 4.3.3.2 to integrate the individual-level model over the covariate distribution in the FIXTURE trial (Table 6.1). First, 10,000 points are taken from a five-dimensional Sobol' sequence, one dimension for each covariate: body surface area, duration of psoriasis, previous systemic treatment, psoriatic arthritis, and weight. Marginal distributions for each covariate in the FIXTURE trial are chosen to match the reported summary statistics, with specific form based on theoretical properties and the observed distributions in the UNCOVER trials: weight and duration are given a Gamma distribution to account for skewness, and body surface area as a percentage is given a scaled logit-Normal distribution (see Figure 6.3). Previous systemic treatment and psoriatic arthritis are binary covariates. Since no information on the correlations between covariates is available in FIXTURE, these are assumed to match those observed in the UNCOVER trials. To account for this, we compute a correlation matrix from the IPD UNCOVER trials (as described

in Section 4.6.4)

$$\begin{bmatrix} 1 & 0.19 & 0.05 & -0.00 & 0.08 \\ 0.19 & 1 & 0.04 & -0.05 & 0.14 \\ 0.05 & 0.04 & 1 & 0.04 & 0.05 \\ -0.00 & -0.05 & 0.04 & 1 & -0.00 \\ 0.08 & 0.14 & 0.05 & -0.00 & 1 \end{bmatrix} \begin{matrix} \text{Duration of psoriasis} \\ \text{Previous systemic treatment} \\ \text{Body surface area} \\ \text{Weight} \\ \text{Psoriatic arthritis} \end{matrix} \tag{6.6}$$

and impose this upon the Sobol' points using a Gaussian copula, before transforming to the required marginal distributions using the inverse CDF method (see Section 5.1). The resulting integration points capture the correlations between the covariates (e.g. longer duration of psoriasis is correlated with having previous systemic treatment) whilst preserving the marginal distribution for each covariate. Figures 6.4 and 6.5 demonstrate that empirical integration error rates of the order $\tilde{N}^{-1}$ are indeed achieved over the entire posterior distribution of the parameters, for both $\bar{p}$ and $\bar{p^2}$ in each arm of the FIXTURE study.

The ML-NMR model (6.5) is fitted to the PASI 75 outcomes, including interaction terms for the five potential effect modifiers in Table 6.1. We take a Bayesian approach implemented in Stan (Carpenter et al. 2017), placing a non-informative $\text{N}(0, 100^2)$ prior distribution on each parameter. We assess convergence using $\hat{R}$ for each parameter, and check that there are no divergent transitions (see Section 5.2.5). With only two AgD secukinumab treatment arms available, it is not possible to identify a model with five distinct effect modifier interactions and a treatment effect for each secukinumab dose. However, since secukinumab and ixekizumab share modes of action as interleukin-17A blockers, we assume that the effect modifier interaction parameters are common between these treatments across all doses (the shared effect modifier assumption) to identify the model (Section 2.5).

Finally, we produce estimates of contrasts between each pair of treatments, and of the proportion of individuals achieving PASI 75, in both the UNCOVER and FIXTURE populations. The SMD contrasts between each pair of treatments are produced by "plugging in" mean covariate values from the population of interest into equation (4.28), as described in Section 4.4. To calculate the predicted proportion of individuals achieving PASI 75 response on treatment k in population P, we define

$$h(\boldsymbol{x}, \mu_P, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{2,k}, \gamma_k) = \Phi(\mu_P + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k), \tag{6.7}$$

where μ_P is the individual-level reference effect in population P, which may be equal to μ_j if P is study j in the analysis, or may be estimated from external

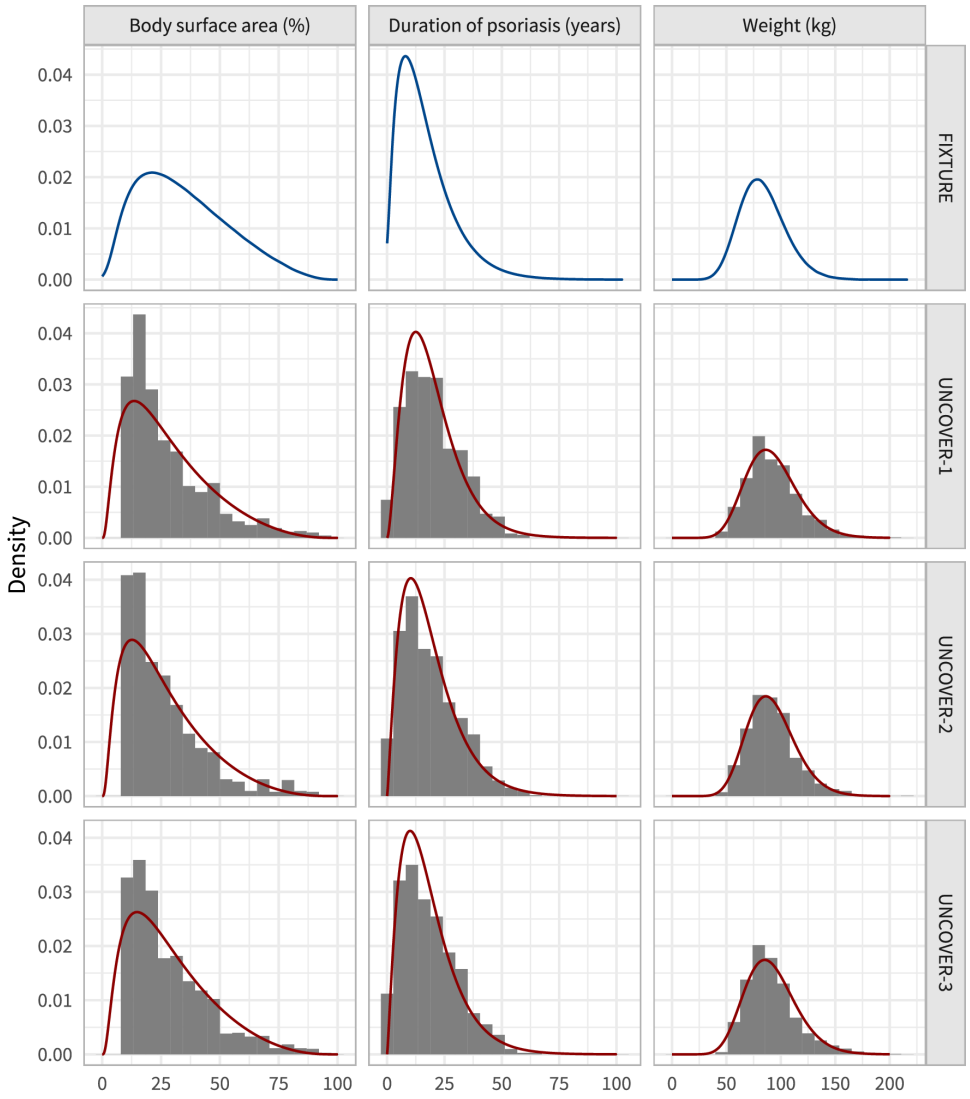


Figure 6.3 The forms of the marginal distributions for the covariates in the FIXTURE study are assumed to match those in the UNCOVER studies. Histograms show the observed marginal distributions in the UNCOVER studies, which are overlaid with the assumed distribution. The assumed marginal distributions are shown for the FIXTURE study. Body surface area is assumed a scaled logit-Normal distribution; weight and duration of psoriasis are assumed a Gamma distribution.

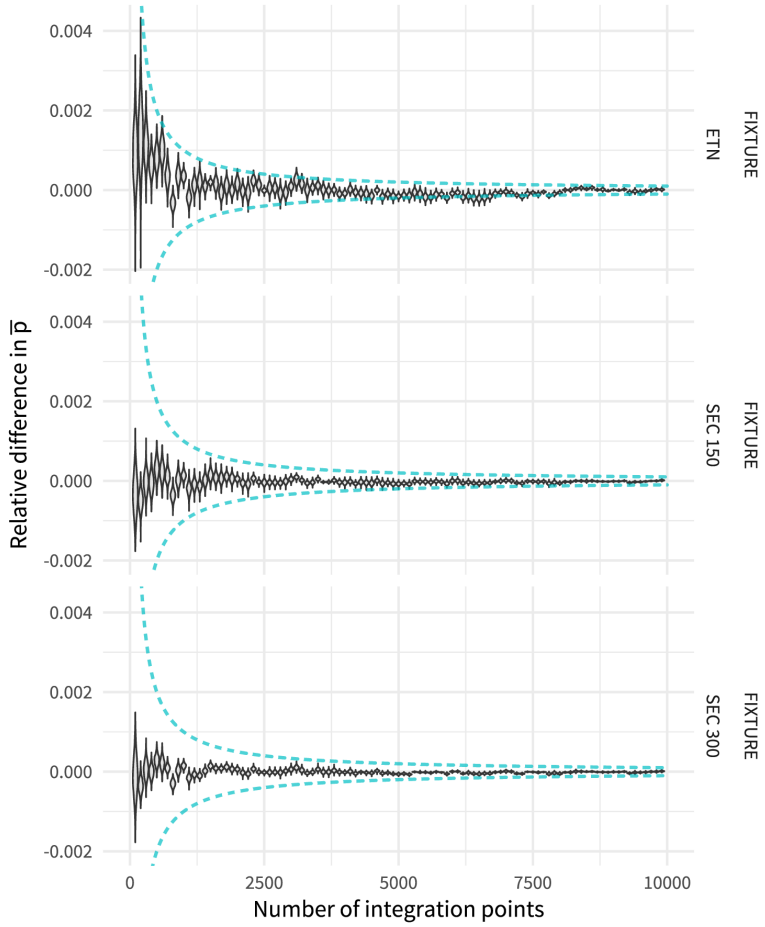


Figure 6.4 Empirical integration error for $\bar{p}$ over the entire posterior distribution of the model parameters, estimated as a relative difference from the final estimate with 10,000 integration points (at each posterior sample). The dashed line is $\pm\tilde{N}^{-1}$, showing that the integration error rate is of this order.

data on P. We then produce estimates as described in Section 4.4. To obtain the predicted proportion achieving PASI 75 in the IPD UNCOVER populations, (6.7) is summarised over every individual using equation (4.30). For the FIXTURE population, the predicted proportion is produced by integrating (6.7) over the joint covariate distribution using numerical integration as in equation (4.29).

Results 6.2.2

The resulting contrast estimates are shown in Table 6.3 and Figure 6.6. There are small differences in the estimated average treatment effects in each population, for example, etanercept appears slightly more effective relative to placebo in the FIXTURE study population than in the UNCOVER study populations.

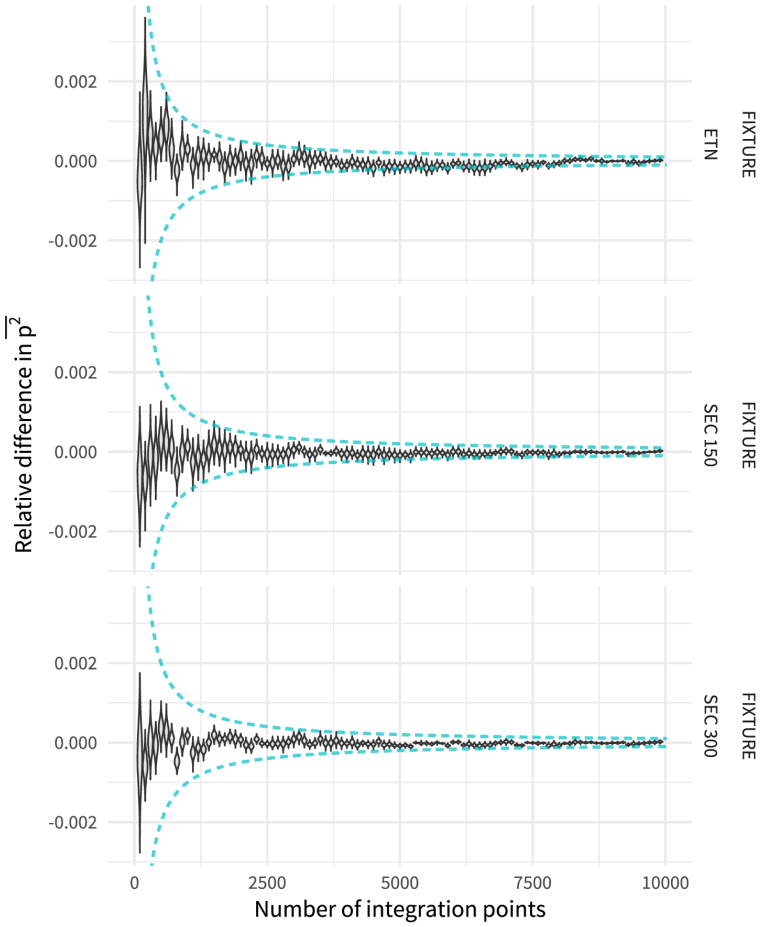


Figure 6.5 Empirical integration error for $\bar{p}^2$ over the entire posterior distribution of the model parameters, estimated as a relative difference from the final estimate with 10,000 integration points (at each posterior sample). The dashed line is $\pm\tilde{N}^{-1}$, showing that the integration error rate is of this order.

Examining the estimated effect modifier interactions in Table 6.4 alongside the covariate summaries in Table 6.1, we see that this is to be expected: the differences in mean covariate values between study populations are small when combined with the size of the interaction terms. Furthermore, a random effects (RE) NMA of the studies estimates the between-study heterogeneity standard deviation to be 0.17 (0.02, 0.46), which is moderate compared to the size of the average treatment effects. Using MAIC, the comparison between ixekizumab Q2W and secukinumab 300 mg in the FIXTURE population is estimated as a SMD of 0.28 (0.00, 0.56) in favour of ixekizumab Q2W; with ML-NMR, we estimate 0.34 (0.10, 0.58). (A standard indirect comparison estimates 0.37 (0.12, 0.63).) The point estimates are similar between the two population adjustment approaches, as we would expect, but ML-NMR has

reduced uncertainty compared to MAIC due to incorporating all available information. The RE NMA estimate for this contrast is 0.45 (−0.02, 0.92), which assumes that any imbalance in effect modifiers is random (Section 2.1.6). Since the differences in effect modifiers between trials are small, the possible bias in the RE NMA estimates are likely to also be small, and indeed all of the ML-NMR estimates are close to the corresponding RE NMA estimates (Figure 6.6); however, ML-NMR increases the precision of the estimates by explaining the within-trial variation due to effect modification.

Figure 6.7 and Table 6.5 show the estimated proportion of individuals achieving PASI 75 in each population, using MAIC and ML-NMR. Again, ML-NMR has reduced uncertainty compared to MAIC, and we are able to produce estimates for any target population—not just the FIXTURE trial population.

Every active treatment is effective compared to placebo, with the class of interleukin-17A blockers more effective than anti-TNFα treatment. Ixekizumab Q2W displays the highest estimated proportion of individuals achieving PASI 75, with posterior mean estimates ranging from 85.9% to 90.3% across the UNCOVER and FIXTURE studies. The 95% Credible Intervals for comparisons of ixekizumab Q2W against every other treatment exclude zero (on the probit SMD scale), in all study populations assessed. In a decision making context, estimates could be produced for the decision target population with a defined covariate distribution, which need not match any of the FIXTURE or UNCOVER studies.

The total residual deviance was 3146.26 on 3858 data points, of which the contribution from 3854 individual data points was 3141.81, and 4.45 from 4 aggregate data points, demonstrating that the model is a good fit to both the IPD and AgD. Standard fixed and random effects NMA models have total residual deviance of 3216.01 and 3210.44 respectively. Comparing the deviance information criterion (DIC) (Spiegelhalter et al. 2002) between the different approaches, ML-NMR has a lower DIC (3170) than either fixed (3225) or random effects (3223) NMA. Comparing the FE and RE NMA models shows little evidence of between-study heterogeneity, although there are only four studies in this analysis. Despite this, the ML-NMR model achieves better fit than both the FE and RE NMAs and is more interpretable and informative, as both between- and within-study variation is explained rather than averaged over.

This analysis relies upon the shared effect modifier assumption to identify the model parameters due to the small number of trials and treatments in the network (Figure 6.1). We also assume that there are no unobserved effect

Table 6.3 Estimated SMD contrasts and 95% Credible Intervals for each pair of treatments in each study population (for ML-NMR) and from a random effects NMA. Note that the ML-NMR contrast estimates between ixekizumab and secukinumab treatments are the same in every population due to the shared effect modifier assumption for these treatments.

	ML-NMR study population				RE NMA
Contrast	FIXTURE	UNCOVER-1	UNCOVER-2	UNCOVER-3	Weighted overall
IXE Q2W vs. PBO	2.94 (2.74, 3.14)	2.98 (2.80, 3.17)	2.95 (2.77, 3.13)	2.93 (2.76, 3.11)	2.91 (2.64, 3.19)
IXE Q4W vs. PBO	2.65 (2.45, 2.84)	2.69 (2.51, 2.89)	2.66 (2.47, 2.84)	2.64 (2.46, 2.82)	2.61 (2.32, 2.88)
ETN vs. PBO	1.74 (1.55, 1.93)	1.65 (1.47, 1.83)	1.64 (1.46, 1.81)	1.65 (1.47, 1.81)	1.61 (1.34, 1.91)
SEC 150 vs. PBO	2.29 (2.07, 2.53)	2.33 (2.10, 2.58)	2.30 (2.07, 2.54)	2.28 (2.05, 2.52)	2.16 (1.71, 2.61)
SEC 300 vs. PBO	2.60 (2.36, 2.83)	2.64 (2.40, 2.90)	2.61 (2.36, 2.86)	2.59 (2.35, 2.83)	2.46 (2.02, 2.89)
IXE Q4W vs. IXE Q2W	−0.30 (−0.42, −0.17)	−0.30 (−0.42, −0.17)	−0.30 (−0.42, −0.17)	−0.30 (−0.42, −0.17)	−0.31 (−0.57, −0.05)
ETN vs. IXE Q2W	−1.20* (−1.35, −1.06)	−1.33 (−1.47, −1.19)	−1.31 (−1.45, −1.18)	−1.29 (−1.42, −1.15)	−1.30 (−1.58, −1.01)
SEC 150 vs. IXE Q2W	−0.65 (−0.89, −0.42)	−0.65 (−0.89, −0.42)	−0.65 (−0.89, −0.42)	−0.65 (−0.89, −0.42)	−0.75 (−1.24, −0.25)
SEC 300 vs. IXE Q2W	−0.34† (−0.58, −0.10)	−0.34 (−0.58, −0.10)	−0.34 (−0.58, −0.10)	−0.34 (−0.58, −0.10)	−0.45 (−0.92, 0.02)
ETN vs. IXE Q4W	−0.91 (−1.05, −0.75)	−1.04 (−1.17, −0.90)	−1.02 (−1.15, −0.89)	−0.99 (−1.12, −0.85)	−0.99 (−1.27, −0.68)
SEC 150 vs. IXE Q4W	−0.36 (−0.59, −0.12)	−0.36 (−0.59, −0.12)	−0.36 (−0.59, −0.12)	−0.36 (−0.59, −0.12)	−0.45 (−0.93, 0.03)
SEC 300 vs. IXE Q4W	−0.05 (−0.27, 0.19)	−0.05 (−0.27, 0.19)	−0.05 (−0.27, 0.19)	−0.05 (−0.27, 0.19)	−0.14 (−0.62, 0.31)
SEC 150 vs. ETN	0.55 (0.36, 0.74)	0.68 (0.48, 0.88)	0.66 (0.46, 0.87)	0.63 (0.43, 0.84)	0.54 (0.11, 0.97)
SEC 300 vs. ETN	0.86 (0.65, 1.06)	0.99 (0.77, 1.20)	0.97 (0.75, 1.18)	0.94 (0.72, 1.14)	0.85 (0.41, 1.26)
SEC 300 vs. SEC 150	0.31 (0.11, 0.53)	0.31 (0.11, 0.53)	0.31 (0.11, 0.53)	0.31 (0.11, 0.53)	0.31 (−0.16, 0.77)

* MAIC estimate is −1.18 (−1.37, −0.99). Standard indirect comparison uses the pooled study estimate −1.27 (−1.42, −1.12) from UNCOVER-2 and 3.
† MAIC estimate is −0.28 (−0.56, −0.00). Standard indirect comparison estimate is −0.37 (−0.63, 0.12).

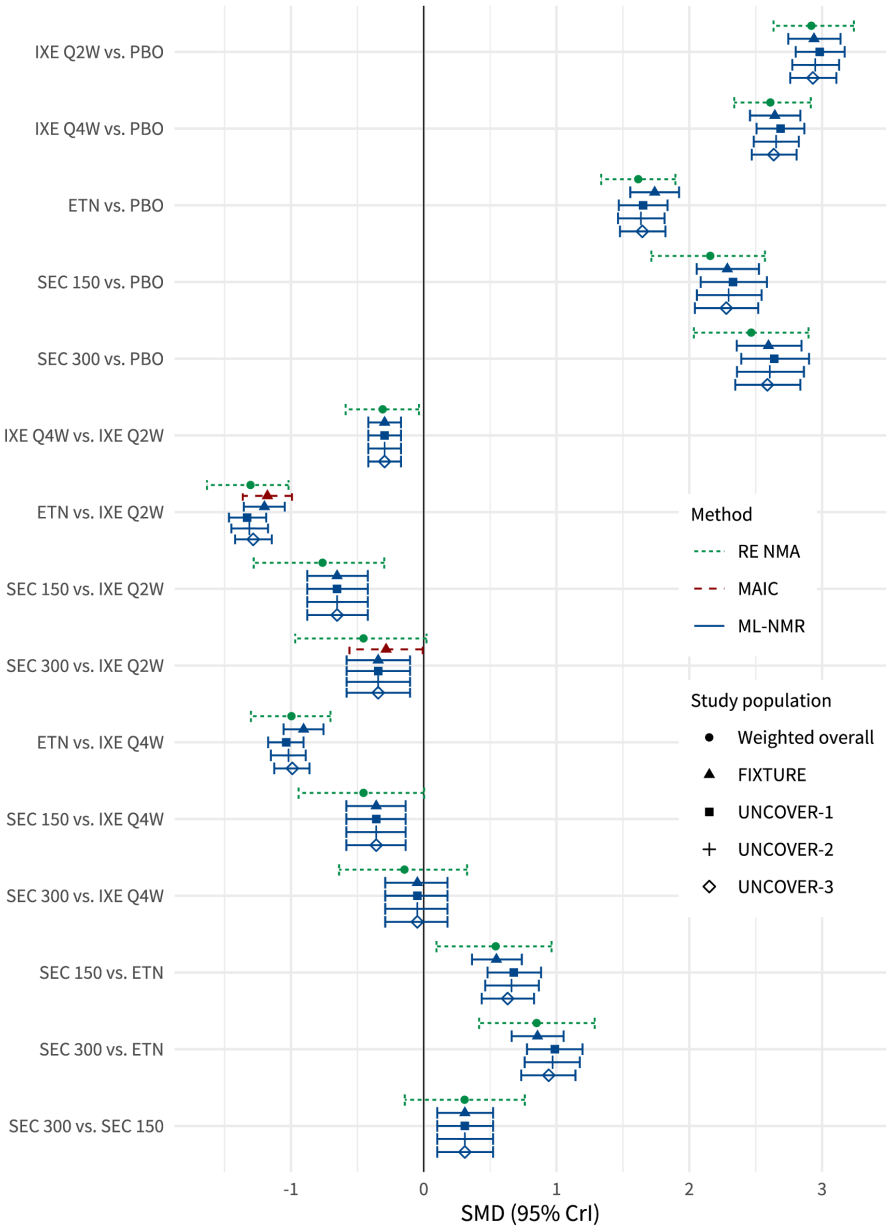


Figure 6.6 Estimated contrasts at the population level, for each pair of treatments in each study population. Note that the interval for MAIC is a 95% Confidence Interval, as MAIC is a frequentist method.

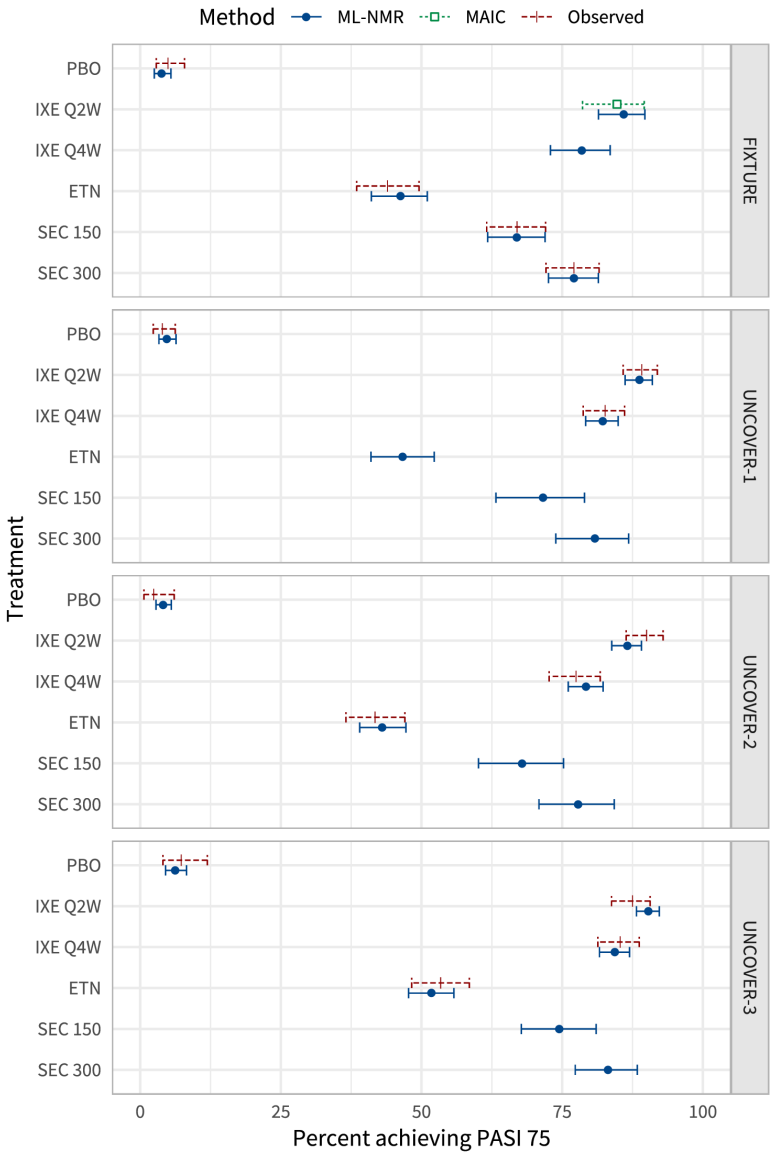


Figure 6.7 Estimated proportion of individuals achieving PASI 75 on each treatment, in each study population. Note that the MAIC estimate is produced in the FIXTURE study population, and the corresponding interval is a 95% Confidence Interval as MAIC is a frequentist method.

Table 6.4 Estimated interactions for each treatment class and potential effect modifier, and estimated individual-level treatment effects. All estimates are standardised mean differences versus placebo, with 95% Credible Intervals.

	Treatment class	
	Anti-TNFα	IL-17A blocker
Effect modifier interaction		
Previous systemic use	−0.00 (−0.38, 0.35)	0.12 (−0.21, 0.46)
Duration of psoriasis, per 10 years	0.14 (−0.03, 0.30)	0.17 (0.02, 0.33)
Body surface area, per 10%	0.06 (−0.05, 0.17)	0.02 (−0.09, 0.13)
Weight, per 10 kg	−0.10 (−0.18, −0.02)	−0.04 (−0.11, 0.04)
Psoriatic arthritis	0.01 (−0.43, 0.48)	0.25 (−0.17, 0.71)
Reference individual treatment effect		
IXE Q2W		2.82 (2.56, 3.10)
IXE Q4W		2.52 (2.25, 2.80)
ETN	1.67 (1.38, 1.96)	
SEC 150		2.16 (1.86, 2.49)
SEC 300		2.47 (2.17, 2.79)

modifiers so that the conditional constancy of relative effects assumption holds. In a larger network with more data available, we can attempt to relax and assess both of these assumptions.

Table 6.5 Estimated proportion of individuals achieving PASI 75 on each treatment in each study population, along with 95% Credible Intervals, for each method. The observed proportions are accompanied by 95% Confidence Intervals, calculated on the probit scale.

		Treatment					
Study population	Method	Placebo	Ixekizumab Q2W	Ixekizumab Q4W	Etanercept	Secukinumab 150 mg	Secukinumab 300 mg
FIXTURE	Observed	4.94 (2.85, 7.90)	-	-	43.96 (38.47, 49.56)	66.97 (61.59, 72.05)	77.09 (72.11, 81.56)
	ML-NMR	3.83 (2.58, 5.38)	85.93* (81.58, 89.64)	78.46 (72.84, 83.27)	46.10 (40.93, 51.11)	66.96 (61.82, 71.72)	77.07 (72.72, 81.34)
UNCOVER-1	Observed	3.94 (2.31, 6.24)	89.15 (85.83, 91.91)	82.64 (78.73, 86.09)	-	-	-
	ML-NMR	4.76 (3.37, 6.47)	88.72 (86.48, 90.79)	82.19 (79.10, 85.11)	46.57 (40.92, 52.22)	71.59 (63.69, 78.78)	80.79 (73.72, 86.93)
UNCOVER-2	Observed	2.41 (0.66, 6.05)	90.00 (86.37, 92.94)	77.46 (72.68, 81.75)	41.74 (36.57, 47.04)	-	-
	ML-NMR	4.06 (2.80, 5.66)	86.59 (84.20, 88.88)	79.31 (76.35, 82.11)	43.05 (39.23, 46.97)	67.92 (60.01, 75.19)	77.83 (70.69, 84.00)
UNCOVER-3	Observed	7.29 (4.04, 11.93)	87.50 (83.77, 90.64)	85.30 (81.34, 88.70)	53.40 (48.26, 58.49)	-	-
	ML-NMR	6.23 (4.48, 8.35)	90.33 (88.35, 92.20)	84.41 (81.74, 86.94)	51.69 (47.55, 55.94)	74.52 (67.41, 80.93)	83.13 (76.49, 88.31)

* MAIC estimate is 84.74 (78.54, 89.62).

Extending the network further 6.3

The ML-NMR analysis described in Section 6.2 makes full use of the data available in the UNCOVER and FIXTURE trials (Figure 6.1). However, IPD from an additional study IXORA-S is also available, comparing ixekizumab with another treatment of interest, ustekinumab (UST). Furthermore, a literature search carried out by our industry collaborators found an additional three AgD studies which compare secukinumab doses with placebo (ERASURE, FEATURE, and JUNCTURE) and one which compares secukinumab with ustekinumab (CLEAR). The full treatment network is shown in Figure 6.8 and Table 6.6.

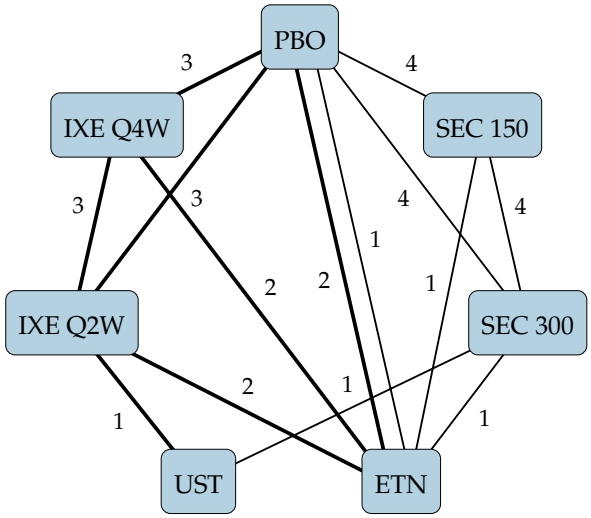


Figure 6.8 The full plaque psoriasis treatment network. The thick and thin lines represent availability of IPD and AgD on a comparison respectively. The number next to each line shows the number of studies making this comparison. PBO = placebo, IXE = ixekizumab, SEC = secukinumab, ETN = etanercept, UST = ustekinumab. IXE and SEC were each investigated with two different dosing regimens.

Methods 6.3.1

We implement the same ML-NMR model (6.5) described in Section 6.2—a Bernoulli individual likelihood with probit link function, and the two-parameter Binomial approximation for the aggregate likelihood. We again use a five-dimensional Sobol' sequence for the QMC integration over the five covariates, except now we have a set of integration points for five AgD populations not just one. We halve the number of integration points to 5,000 for the full network analysis, since the convergence plots Figures 6.4 and 6.5 indicate that the integration error with this number is already very small over

Table 6.6 Treatment comparisons made by studies in the full plaque psoriasis treatment network; ● indicates that the study included this treatment arm.

Study	Placebo	Ixekizumab Q2W	Ixekizumab Q4W	Etanercept	Secukinumab 150 mg	Secukinumab 300 mg	Ustekinumab
IPD studies							
UNCOVER-1	●	●	●	-	-	-	-
UNCOVER-2	●	●	●	●	-	-	-
UNCOVER-3	●	●	●	●	-	-	-
IXORA-S	-	●	-	-	-	-	●
AgD studies							
CLEAR	-	-	-	-	-	●	●
ERASURE	●	-	-	-	●	●	-
FEATURE	●	-	-	-	●	●	-
FIXTURE	●	-	-	●	●	●	-
JUNCTURE	●	-	-	-	●	●	-

the entire posterior distribution. We again place a non-informative $N(0, 100^2)$ prior distribution on each parameter. We assess convergence using $\hat{R}$ for each parameter, and check that there are no divergent transitions (see Section 5.2.5). The larger network also presents opportunities to explore the shared EM assumption and any residual heterogeneity or inconsistency, using the models described in Section 4.6.1, Section 4.6.2, and Section 4.6.3 respectively.

Results 6.3.2

The resulting contrast estimates are shown in Table 6.7 for each study population. Again, the differences in estimated average treatment effects between each study population are small since the differences in covariate distributions between each population are small when compared with the size of the estimated EM interaction terms (Table 6.8). The point estimates for each contrast are very similar to those in the smaller network (compare with Table 6.3). However, as we would expect, the contrasts are more precisely estimated in this analysis which draws upon more information than before: this is most apparent for the etanercept and secukinumab treatments, where the additional AgD studies provide the greatest amount of additional information. The gain in precision over the RE NMA is smaller in this analysis compared to the smaller network. The estimate of the heterogeneity standard deviation from the RE NMA is 0.10 (0.01, 0.26), which is both smaller and more precise than before. Indeed, the DIC of fixed and random effects NMA are nearly identical (3497.0 and 3497.6 respectively), so we would choose the more parsimonious FE model based on DIC alone. Fitting an unrelated mean effects (UME) NMA (as described in Section 1.2.7.1) gives a DIC of 3498.8 and residual deviance of 3478.6, which are not substantially different to either the FE or RE NMAs (with residual deviance of 3482.4 and 3478.4, respectively). There is also no change in the heterogeneity standard deviation under the UME NMA model, which is estimated as 0.10 (0.01, 0.28). Figure 6.9 compares the residual deviance contributions under each NMA model (FE, RE, and UME). The points lie largely on the line of equality, indicating little difference in fit between the models. There is some suggestion that two IPD study arms (ixekizumab Q2W arms in UNCOVER-2 and 3) fit better under the RE NMA model than the FE NMA model (lying below the line of equality in Figure 6.9a), although the 95% Credible Intervals are very wide and include the line of equality. Overall, there is little evidence for substantial heterogeneity or inconsistency in the NMA, yet the ML-NMR model has a much lower DIC of 3451.4 (and residual deviance of 3415.8). The ML-NMR model allows us to explain both between

and within study variation, resulting in better fit.

The estimated proportion of individuals achieving the PASI 75 outcome in each study population are given in Table 6.9 and Figure 6.10. As for the contrast estimates, the estimated proportions have changed little from the smaller network analysis (Table 6.5), but there is an increase in precision—particularly for the etanercept and secukinumab treatments.

As before, for decision making purposes estimates could be provided in a decision target population with a defined covariate distribution, which need not match any of the included studies. For absolute estimates of the proportion of individuals achieving the PASI 75 outcome (which may be required for economic modelling and cost-effectiveness analysis, for example), an estimate of the reference probability of achieving PASI 75 on placebo is also required. This could be taken from the literature, or estimated from a registry or other study in the target population.

The model presented above again makes the shared EM assumption between the interleukin-17A blockers secukinumab and ixekizumab. In the smaller network (Section 6.2), this assumption was required for identifiability. With the addition of three further AgD trials involving secukinumab, we hoped to be able to relax this assumption, either by using the exchangeable interactions model (see Section 4.6.1) or by fitting independent interactions. However this proved difficult with the amount of data available. Firstly, we attempted to fit exchangeable interactions within the class of interleukin-17A blockers for all five effect modifiers at once. To do this, we modify the ML-NMR model given in (6.5) to include a hierarchical structure on the respective EM interaction parameters. Letting $\mathcal{T}$ = {IXE Q2W, IXE Q4W, SEC 150, SEC 300}, we write this as

$$\beta_{2,k;l} \sim \mathrm{N}\left(m_{\beta_2,l}, \sigma^2_{\beta_2,l}\right) \quad \forall k \in \mathcal{T} \text{ and } l \in \{1,\dots,5\}, \tag{6.8}$$

and the other $\boldsymbol{\beta}_{2,k}$ corresponding to etanercept and ustekinumab are given independent prior distributions. Even with strong prior distributions on the hierarchical means $m_{\beta_2,l}$ and standard deviations $\sigma_{\beta_2,l}$, this model was too weakly-identified by the available data to achieve a reliable fit. We then attempted to fit exchangeable interactions for each effect modifier in turn, keeping the shared EM assumption for the other four effect modifiers. For each effect modifying covariate $l^* \in \{1,\dots,5\}$ in turn, we write this model as

$$\begin{aligned} \beta_{2,k;l^*} &\sim \mathrm{N}\left(m_{\beta_2,l^*}, \sigma^2_{\beta_2,l^*}\right) \quad \forall k \in \mathcal{T} \\ \beta_{2,a;l} &= \beta_{2,b;l} \quad \forall a, b \in \mathcal{T} \text{ and } l \in \{1,\dots,5\} \setminus l^*, \end{aligned} \tag{6.9}$$

and the other $\boldsymbol{\beta}_{2,k}$ corresponding to etanercept and ustekinumab again given independent prior distributions. These models were also too weakly-identified. There was also not enough data to fit a model with independent EM interactions for all treatments (every $\beta_{2,k;l}$ given an independent prior), or to fit models where the EM interactions are only shared between the two doses of secukinumab (i.e. with $\mathcal{T}$ = {SEC 150, SEC 300} in (6.8)).

Whilst there were insufficient data to explicitly relax the shared EM assumption using the models described above (see Section 4.6.1 for a discussion of these models and their data requirements), there are enough data to assess whether there is residual heterogeneity with a random effects model (see Section 4.6.2). If residual heterogeneity is present, this could either be due to unobserved EMs or an inappropriate shared EM assumption (see Section 4.6.2). Thus we can still assess the shared EM assumption in this network, even though we cannot explicitly relax it. We also fit an unrelated mean effects ML-NMR model, as described in Section 4.6.3.1, to check for inconsistency. There are insufficient data to fully relax consistency of the EM interactions (i.e. with no shared EM assumption), so we keep the shared EM assumption for the class of interleukin-17A blockers and instead check inconsistency of the EM interactions at the treatment class level. The results of the RE ML-NMR model are very similar to the standard fixed effect (FE) model. The posterior median of the heterogeneity standard deviation is estimated as 0.09 (0.01, 0.25), which is small compared to the magnitude of the relative effects (Table 6.8). This is only slightly smaller than the heterogeneity estimate from the RE NMA, suggesting that the improved fit of ML-NMR is largely due to explaining within-study variation rather than any between-study heterogeneity. Indeed, the DIC of the RE ML-NMR model is 3449.2 (residual deviance 3412.2), which does not suggest any substantial improvement over the FE model with DIC 3451.4 (residual deviance 3415.8). The UME ML-NMR model has DIC 3448.8 and residual deviance 3410.7, which again does not suggest any substantial improvement over either the FE or RE ML-NMR models with consistency (Table 6.10). The estimated heterogeneity standard deviation is also unchanged between the RE ML-NMR and UME ML-NMR. Table 6.10 shows that the slight reductions in residual deviance against the FE ML-NMR model (to be expected, since the RE and UME models are more flexible) are balanced against the increased number of effective parameters (the FE consistency model is more parsimonious). When examining Figure 6.9 which compares the residual deviance contributions under each ML-NMR model (FE, RE, and UME), we see that the points lie on the line of equality, indicating little difference in fit

between the models. There is some suggestion that one AgD point (the placebo arm of the FIXTURE trial) fits better under the RE ML-NMR model than the FE ML-NMR model (lying below the line of equality in Figure 6.11a), although the 95% Credible Intervals are very wide and include the line of equality. Overall, there is therefore little evidence for substantial residual heterogeneity, and no evidence for inconsistency in this analysis: we have not detected any failings in either the conditional constancy of relative effects assumption, the shared effect modifier assumption, or the consistency assumption.

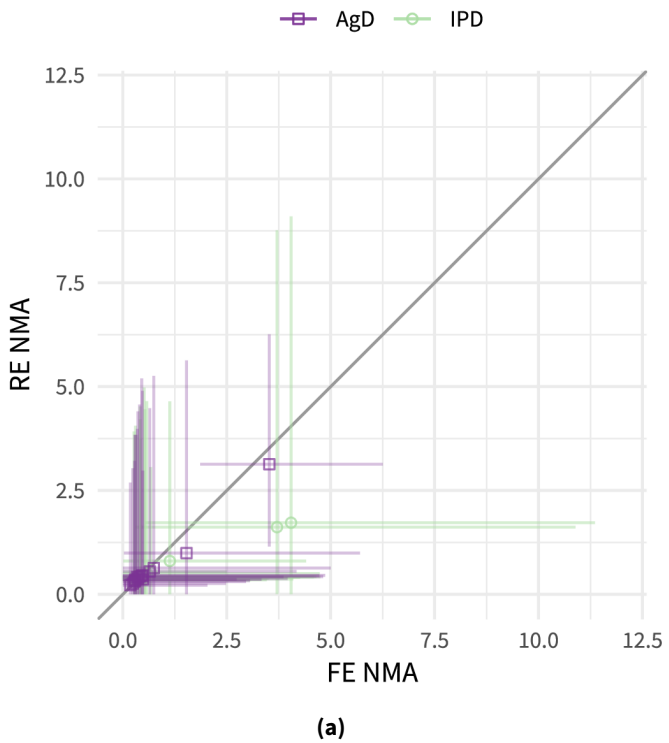
Table 6.7 Estimated basic SMD contrasts and 95% Credible Intervals for each treatment compared to placebo, plus the comparison targeted by the previous MAIC, in each study population using the ML-NMR model and for the RE NMA.

	ML-NMR study population									RE NMA
Contrast	CLEAR	ERASURE	FEATURE	FIXTURE	IXORA	JUNCTURE	UNCOVER-1	UNCOVER-2	UNCOVER-3	Weighted overall
IXE Q2W vs. PBO	2.96 (2.78, 3.15)	2.96 (2.78, 3.14)	2.97 (2.79, 3.16)	2.94 (2.75, 3.14)	2.97 (2.77, 3.19)	2.99 (2.80, 3.17)	2.99 (2.81, 3.18)	2.96 (2.79, 3.14)	2.94 (2.77, 3.12)	2.92 (2.71, 3.13)
IXE Q4W vs. PBO	2.67 (2.50, 2.85)	2.67 (2.50, 2.84)	2.68 (2.52, 2.87)	2.65 (2.48, 2.84)	2.69 (2.49, 2.90)	2.70 (2.53, 2.88)	2.71 (2.53, 2.89)	2.67 (2.51, 2.84)	2.65 (2.49, 2.82)	2.62 (2.42, 2.83)
ETN vs. PBO	1.74 (1.58, 1.90)	1.72 (1.56, 1.88)	1.72 (1.55, 1.89)	1.77 (1.60, 1.95)	1.72 (1.52, 1.93)	1.70 (1.53, 1.88)	1.69 (1.51, 1.87)	1.67 (1.50, 1.84)	1.68 (1.52, 1.84)	1.65 (1.46, 1.85)
SEC 150 vs. PBO	2.38 (2.19, 2.58)	2.37 (2.19, 2.57)	2.39 (2.19, 2.60)	2.36 (2.16, 2.55)	2.39 (2.17, 2.62)	2.40 (2.20, 2.61)	2.41 (2.21, 2.62)	2.38 (2.18, 2.58)	2.36 (2.17, 2.55)	2.28 (2.06, 2.54)
SEC 300 vs. PBO	2.71 (2.53, 2.90)	2.71 (2.52, 2.90)	2.72 (2.53, 2.93)	2.69 (2.50, 2.88)	2.72 (2.51, 2.95)	2.73 (2.54, 2.95)	2.74 (2.55, 2.95)	2.71 (2.52, 2.91)	2.69 (2.51, 2.89)	2.61 (2.39, 2.86)
UST vs. PBO	2.26 (1.98, 2.58)	2.27 (1.97, 2.60)	2.27 (1.98, 2.59)	2.26 (1.96, 2.59)	2.23 (1.92, 2.54)	2.30 (1.95, 2.71)	2.28 (1.98, 2.60)	2.27 (1.96, 2.62)	2.26 (1.94, 2.64)	2.12 (1.80, 2.46)
SEC 300 vs. IXE Q2W	−0.25 (−0.45, −0.06)	−0.25 (−0.45, −0.06)	−0.25 (−0.45, −0.06)	−0.25* (−0.45, −0.06)	−0.25 (−0.45, −0.06)	−0.25 (−0.45, −0.06)	−0.25 (−0.45, −0.06)	−0.25 (−0.45, −0.06)	−0.25 (−0.45, −0.06)	−0.31 (−0.55, −0.04)

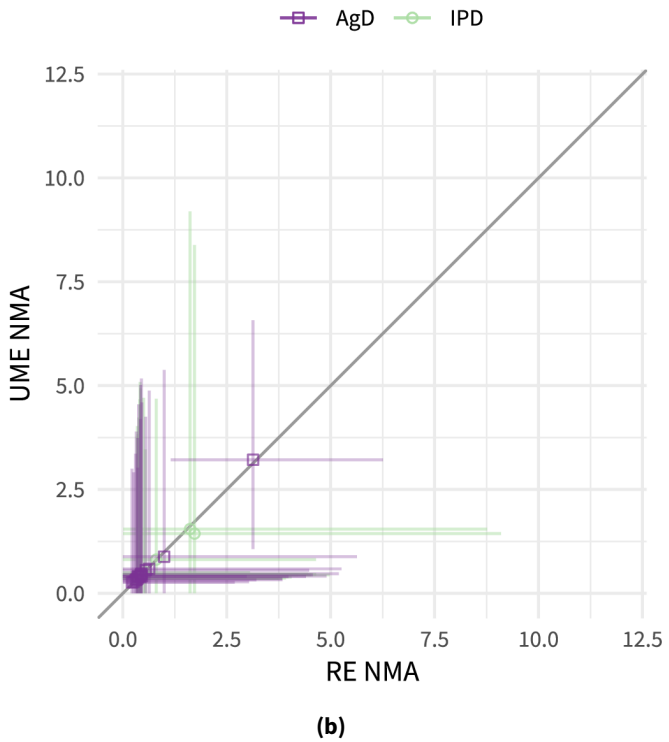
* MAIC estimate is −0.28 (−0.56, −0.00). Standard indirect comparison estimate is −0.37 (−0.63, 0.12).

Table 6.8 Estimated interactions for each treatment class and potential effect modifier, and estimated individual-level treatment effects, for the ML-NMR model in the full network. All estimates are standardised mean differences versus placebo, with 95% Credible Intervals.

	Treatment class		
	Anti-TNFα	IL-12 and IL-23 blocker	IL-17A blocker
Effect modifier interaction			
Previous systemic use	0.00 (−0.39, 0.39)	−0.11 (−1.17, 0.81)	0.11 (−0.25, 0.47)
Duration of psoriasis, per 10 years	0.14 (−0.01, 0.31)	0.15 (−0.08, 0.37)	0.18 (0.03, 0.33)
Body surface area, per 10%	0.06 (−0.05, 0.17)	0.01 (−0.14, 0.18)	0.02 (−0.08, 0.13)
Weight, per 10 kg	−0.10 (−0.18, −0.02)	−0.04 (−0.16, 0.08)	−0.03 (−0.10, 0.05)
Psoriatic arthritis	−0.01 (−0.46, 0.49)	0.21 (−0.53, 0.98)	0.26 (−0.15, 0.71)
Reference individual treatment effect			
IXE Q2W			2.83 (2.56, 3.14)
IXE Q4W			2.54 (2.26, 2.84)
ETN	1.71 (1.43, 2.00)		
SEC 150			2.25 (1.97, 2.55)
SEC 300			2.58 (2.31, 2.88)
UST		2.29 (1.55, 3.19)	



(a)



(b)

Figure 6.9 Residual deviance contributions (posterior mean and 95% Credible Interval) under (a) RE vs. FE NMA models, and (b) UME vs. RE NMA models.

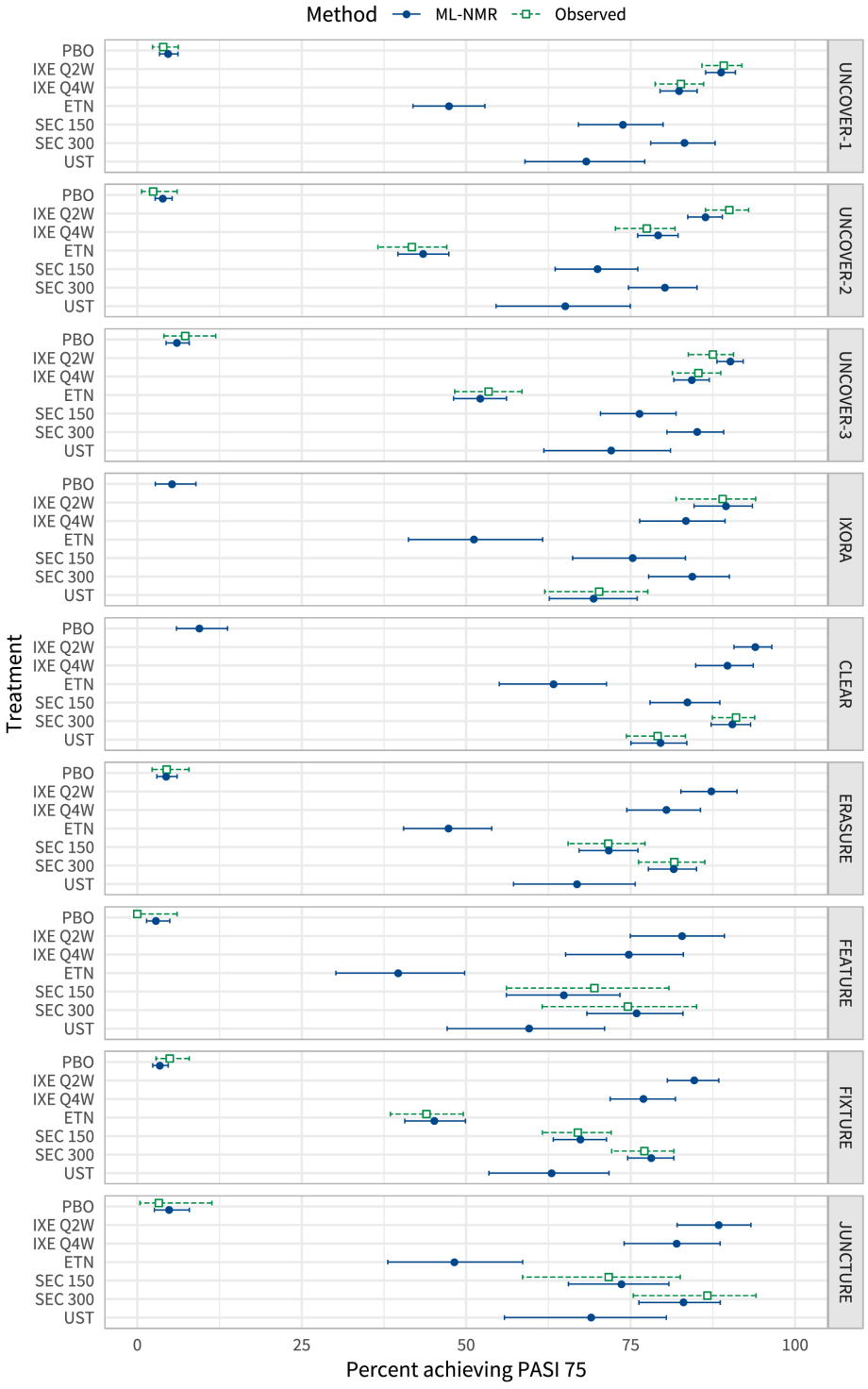


Figure 6.10 Estimated proportion of individuals achieving PASI 75 on each treatment, in each study population, using the ML-NMR model.

Table 6.9 Estimated proportion of individuals achieving PASI 75 on each treatment in each study population, along with 95% Credible Intervals, using ML-NMR.

	Treatment						
Study population	Placebo	Ixekizumab Q2W	Ixekizumab Q4W	Etanercept	Secukinumab 150 mg	Secukinumab 300 mg	Ustekinumab
CLEAR	9.42 (5.96, 13.72)	93.92 (90.71, 96.46)	89.71 (84.89, 93.63)	63.30 (55.04, 71.32)	83.62 (77.95, 88.56)	90.45 (87.24, 93.23)	79.52 (75.05, 83.53)
ERASURE	4.37 (2.98, 6.05)	87.31 (82.64, 91.16)	80.46 (74.41, 85.61)	47.32 (40.47, 53.88)	71.65 (67.17, 76.11)	81.55 (77.70, 85.02)	66.88 (57.20, 75.69)
FEATURE	2.87 (1.42, 4.95)	82.80 (74.95, 89.25)	74.71 (65.12, 83.02)	39.70 (30.19, 49.77)	64.82 (56.10, 73.38)	75.95 (68.34, 82.95)	59.57 (47.10, 71.04)
FIXTURE	3.42 (2.35, 4.70)	84.67* (80.58, 88.40)	76.99 (71.89, 81.81)	45.20 (40.67, 49.89)	67.37 (63.24, 71.33)	78.15 (74.54, 81.56)	62.98 (53.46, 71.71)
IXORA	5.30 (2.73, 8.92)	89.53 (84.64, 93.51)	83.43 (76.36, 89.35)	51.16 (41.22, 61.62)	75.30 (66.18, 83.33)	84.35 (77.74, 90.02)	69.40 (62.64, 75.99)
JUNCTURE	4.84 (2.59, 7.91)	88.40 (82.07, 93.28)	81.98 (74.01, 88.60)	48.19 (38.09, 58.59)	73.61 (65.54, 80.81)	83.01 (76.26, 88.61)	69.01 (55.82, 80.42)
UNCOVER-1	4.67 (3.37, 6.19)	88.73 (86.39, 90.92)	82.35 (79.47, 85.11)	47.36 (41.92, 52.84)	73.83 (67.06, 79.93)	83.23 (78.05, 87.84)	68.29 (58.93, 77.15)
UNCOVER-2	3.91 (2.72, 5.29)	86.39 (83.68, 88.93)	79.21 (76.07, 82.21)	43.48 (39.63, 47.36)	69.96 (63.52, 76.08)	80.20 (74.68, 85.08)	65.04 (54.54, 74.95)
UNCOVER-3	6.01 (4.38, 7.89)	90.19 (88.12, 92.12)	84.34 (81.57, 86.96)	52.15 (48.08, 56.13)	76.36 (70.42, 81.91)	85.16 (80.51, 89.13)	72.05 (61.81, 81.05)

* MAIC estimate is 84.74 (78.54, 89.62).

Table 6.10 Model fit statistics for the RE and UME ML-NMR models compared with the standard FE model. p_D is a measure of the effective number of parameters. We also present estimates and 95% Credible Intervals for the heterogeneity standard deviation τ, for the RE and UME models.

	NMA			ML-NMR		
	FE	RE	UME	FE	RE	UME
Residual deviance	3482.4	3478.4	3478.6	3415.8	3412.2	3410.7
p_D	14.6	19.3	20.2	35.6	37.1	38.0
DIC	3497.0	3497.6	3498.8	3451.4	3449.2	3448.8
τ	-	0.10 (0.01, 0.26)	0.10 (0.01, 0.28)	-	0.09 (0.01, 0.25)	0.09 (0.00, 0.27)

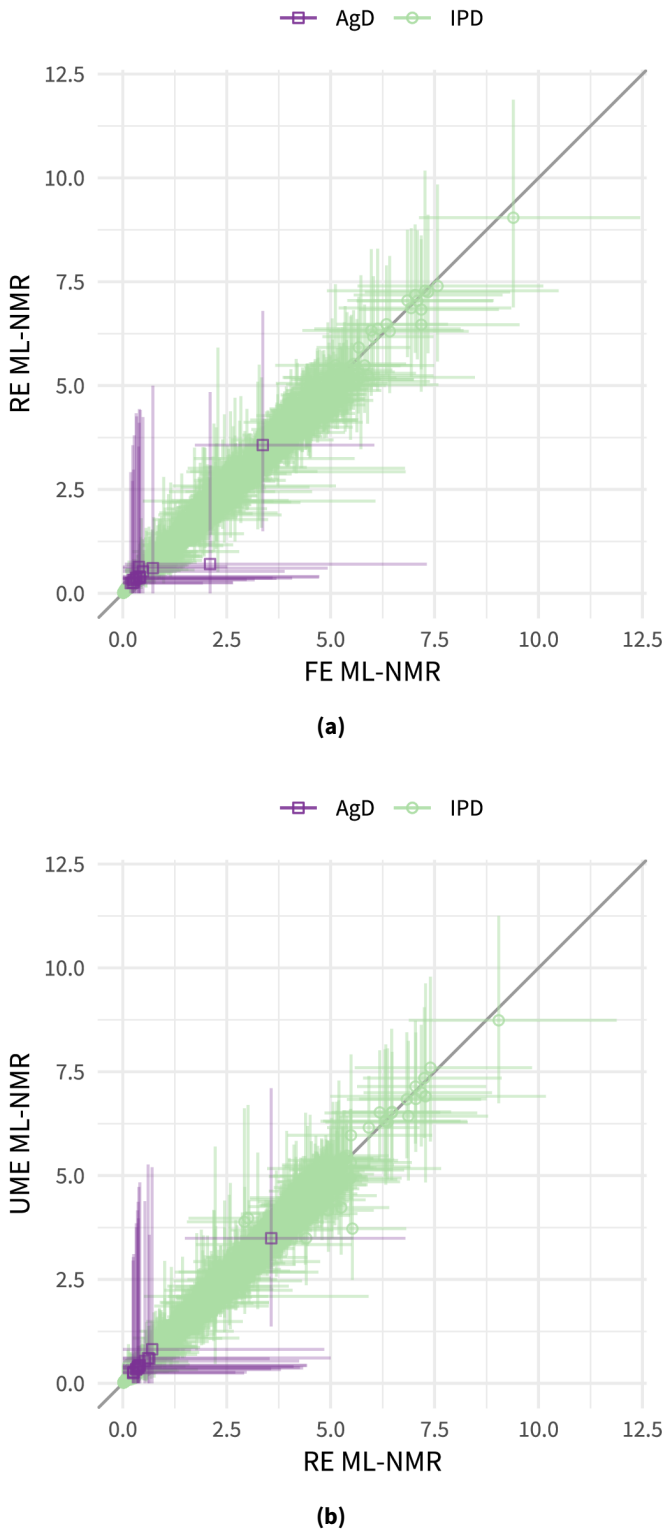


Figure 6.11 Residual deviance contributions (posterior mean and 95% Credible Interval) under (a) RE vs. FE ML-NMR models, and (b) UME vs. RE ML-NMR models.

6.4 Discussion

In this chapter, we began by comparing a targeted comparison performed using ML-NMR to a previous MAIC analysis. We then used the full ML-NMR model to incorporate information from all the available treatment arms and studies in the network formed by the UNCOVER and FIXTURE studies. Using all the available information in the analysis resulted in more precise estimates than the MAIC and targeted comparison, and the ML-NMR model out-performed standard fixed and random effects NMA (in terms of DIC) as between- and within-study heterogeneity was explained rather than averaged over. Finally, we extended the network further, and attempted to assess the shared EM and conditional constancy of relative effects assumptions. Whilst the random effects ML-NMR models found no evidence of either assumption being broken, we could not fit exchangeable EM interaction models to assess or relax the shared EM assumption explicitly. ML-NMR models with exchangeable EM interactions are likely to be data-intensive, and without informative prior distributions for the hierarchical variance components they may be just as data-intensive as a fully independent EM interaction model. Future work could use simulation studies to determine the data requirements for such models.

All of the studies in the extended network reported (at least) three PASI endpoints: 75, 90, and 100% improvement from baseline. We focused on the PASI 75 outcome in this chapter for computational reasons, since the small numbers of observed events on the more demanding PASI 90 and PASI 100 outcomes posed difficulties for estimation. However, the higher PASI outcomes are also of interest for decision-making: from a patient perspective the higher outcomes are more desirable, and the different outcomes are likely to have different utilities in a cost-effectiveness analysis for a reimbursement decision. This motivates consideration of a joint model, where all PASI outcomes are simultaneously modelled, aiming to alleviate the issues with estimating the higher PASI outcomes separately. In the following chapter we extend the ML-NMR framework to handle general likelihoods, and subsequently fit a model to all three PASI outcomes simultaneously using a multinomial likelihood.

Chapter 7

Extension to general likelihoods

Previously in Chapter 4, we presented the general ML-NMR model in equation (4.34). Importantly, derivation of the aggregate-level model is split into two steps: i) deriving the aggregate likelihood from the individual likelihood, and ii) integrating the individual-level model over the covariate distribution in the aggregate population to form the aggregate-level model. As we described in Section 4.3.3.1, the integration step can be performed numerically using Quasi Monte Carlo (QMC) integration, regardless of the number or distribution of covariates, or the form of the link function. However, derivation of the aggregate likelihood is not always straightforward (Section 4.2), and may even be intractable. Most notably this is the case for the analysis of survival outcomes—which represent the large majority of applications of population adjustment methodology to date (Chapter 3)—where the aggregate likelihood cannot be derived analytically. This is also the case for the plaque psoriasis example in the previous chapter (Chapter 6) if we wish to jointly analyse the ordered categorical outcomes (PASI 75, 90, and 100 cutpoints), and the appropriate likelihood for the aggregate summary outcomes is not readily apparent.

In this chapter, we begin by setting out the ML-NMR framework in a more general form, based directly on the likelihood contributions from different sources of data (Section 7.1). Likelihood contributions may be either *individual* or *aggregate*, depending on whether they refer to an individual in a study or an aggregate summary, respectively. Individual likelihood contributions may also be either *conditional* or *marginal*, depending on whether they depend on given covariate values or are averaged over the covariate distribution in a population, respectively. When we specify an individual-level model (with a likelihood, link function, and linear predictor), we are also specifying an

individual conditional likelihood function. We can directly integrate the individual conditional likelihood function over the joint covariate distribution in a study to obtain an individual marginal likelihood function, describing the likelihood where individual outcomes are known but individual covariates are not. For example, this is the case when analysing survival outcomes using time-to-event data reconstructed from published Kaplan-Meier curves but with only published summary covariate information at baseline. Using this approach, we describe the application of ML-NMR to censored survival outcomes with general survival and hazard functions in Section 7.1.1. We can perform the necessary integration numerically using the same approaches described in Section 4.3.3 (here, we use QMC integration). We also consider obtaining an aggregate marginal likelihood function by multiplying together the individual marginal likelihood functions, which gives simple results when outcomes are discrete. We apply this approach to obtain analytic results for binary outcomes (which we compare with Section 4.2.1) and ordered categorical outcomes in Sections 7.1.2 and 7.1.3, respectively. In Section 7.2, we then consider the issue of model comparison under the generalised framework, where the lack of an explicit aggregate likelihood complicates evaluation of the usual quantities of interest (e.g. residual deviance, DIC). Finally we apply these ideas to two examples, one of simulated survival data (Section 7.3) and the other continuing with the plaque psoriasis example from Chapter 6 to jointly model the outcomes (Section 7.4).

7.1 Deriving the aggregate likelihood using integration

We begin with the same individual-level model as before in equations (4.34a) and (4.34b):

$$\begin{aligned} y_{ijk} &\sim \pi_{\text{Ind}}(\theta_{ijk}) \\ g(\theta_{ijk}) = \eta_{jk}(\boldsymbol{x}_{ijk}) &= \mu_j + \boldsymbol{x}_{ijk}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k \end{aligned}$$

with IPD outcomes y_{ijk} for individuals i in study j receiving treatment k given the likelihood distribution $\pi_{\text{Ind}}(\theta_{ijk})$. The link function $g(\cdot)$ links the conditional mean outcomes θ_{ijk} to the linear predictor $\eta_{jk}(\boldsymbol{x}_{ijk})$, with covariates $\boldsymbol{x}_{ijk}$. The parameters μ_j are study-specific intercepts, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{2,k}$ are regression coefficients for prognostic and effect modifying covariates respectively, and γ_k are individual-level treatment effects. We set $\boldsymbol{\beta}_{2,1} = \gamma_1 = 0$ for the reference treatment 1.

Rather than proceeding to determine the appropriate aggregate likelihood (for example using known statistical properties, such as the Normality of the sum of Normally distributed outcomes), and then using numerical integration to evaluate the marginal mean outcome in each aggregate population, we instead consider the likelihood contributions from each level of the model. Let $\boldsymbol{\xi}$ denote the set of all model parameters $\{\mu_j, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{2,k}, \gamma_k : \forall j, k\}$, and denote the individual conditional likelihood contributions (conditional on the covariates) by $L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}_{ijk})$. The form of the individual conditional likelihood function follows from the chosen individual-level model, in particular the individual-level likelihood $\pi_{\text{Ind}}(\cdot)$, link function $g(\cdot)$, and linear predictor $\eta_{jk}(\cdot)$. (To be strictly precise, we refer to the likelihood *contribution* as the value of the likelihood *function* evaluated at a given $\boldsymbol{\xi}$, y_{ijk}, and $\boldsymbol{x}_{ijk}$.) We then integrate the individual conditional likelihood function over the covariate distribution $f_{jk}(\cdot)$ on treatment k in study j to obtain the individual marginal likelihood contributions

$$L_{ijk}(\boldsymbol{\xi}; y_{ijk}) = \int_{\mathfrak{X}} L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x}, \tag{7.1}$$

which no longer depend on $\boldsymbol{x}$. In other words, for an individual on treatment k in study j with outcome y_{ijk}, if we do not know their individual covariate vector $\boldsymbol{x}_{ijk}$ but only the distribution $f_{jk}(\cdot)$, their likelihood contribution is given by (7.1). This integration may be performed using the QMC numerical integration technique described previously (Section 4.3.3.2), or by hand if analytically tractable. If we have summary outcomes $y_{\bullet jk}$ on a given treatment k in study j, we can attempt to derive a corresponding aggregate marginal likelihood contribution as the product of the individual marginal likelihood contributions (7.1), up to a normalising constant:

$$L_{\bullet jk}(\boldsymbol{\xi}; y_{\bullet jk}) \propto \prod_{i=1}^{N_{jk}} L_{ijk}(\boldsymbol{\xi}; y_{ijk}). \tag{7.2}$$

If the result can be rearranged in terms of $y_{\bullet jk}$, we can then use $L_{\bullet jk}(\boldsymbol{\xi}; y_{\bullet jk})$ to evaluate the aggregate marginal likelihood contributions. This is straightforward when outcomes are discrete (as we see in Sections 7.1.2 and 7.1.3), but may not be in general.

By working directly with the likelihood contributions from each level of the model, we avoid having to explicitly derive the form of the aggregate likelihood. The full ML-NMR model for general likelihoods may be written using (7.1) and (7.2) as

Individual:

$$L_{ijk|\boldsymbol{x}}(\xi; y_{ijk}, \boldsymbol{x}_{ijk}) = \pi_{\text{Ind}}(y_{ijk}|\theta_{ijk}) \tag{7.3a}$$

$$g(\theta_{ijk}) = \eta_{jk}(\boldsymbol{x}_{ijk}) = \mu_j + \boldsymbol{x}_{ijk}^{\intercal}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k \tag{7.3b}$$

Aggregate:

$$L_{ijk}(\xi; y_{ijk}) = \int_{\mathfrak{X}} L_{ijk|\boldsymbol{x}}(\xi; y_{ijk}, \boldsymbol{x}) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \tag{7.3c}$$

$$L_{\bullet jk}(\xi; y_{\bullet jk}) \propto \prod_{i=1}^{N_{jk}} L_{ijk}(\xi; y_{ijk}) \tag{7.3d}$$

where in a Bayesian analysis, prior distributions are placed over each of the parameters μ_j, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_{2,k}$, γ_k. Once the ML-NMR model has been fitted, estimates may be produced for a specific target population following exactly the approach in Section 4.4.

Computationally, we can fit these models in WinBUGS/OpenBUGS/JAGS using the "zeros trick" to provide the correct (log) likelihood contributions via a Poisson distribution with dummy zero observations (see Lunn et al. 2010, Section 9.5.1, for further information). In Stan, we can directly code the (log) likelihood contributions using a `target +=` statement.

To clarify ideas, let us apply these ideas to derive ML-NMR models for three examples.

7.1.1 Survival outcome data

The large majority of applications of population adjustment methodology to date have been for survival analysis (Chapter 3). We are therefore interested in extending the ML-NMR framework to handle time-to-event outcomes such as survival, as this will greatly increase the applicability of the method.

We consider a scenario where every study provides outcome times t_{ijk} for each individual i in study j receiving treatment k, along with an indicator y_{ijk} of whether t_{ijk} is an event ($y_{ijk} = 1$) or censoring ($y_{ijk} = 0$) time. For the AgD studies, this data could be obtained by digitising published Kaplan-Meier curves and reconstructing the event and censoring times using an algorithm such as that described by Guyot et al. (2012). Individual covariate information $\boldsymbol{x}_{ijk}$ is available for every individual in the IPD studies, but for the AgD studies only the joint distribution of the covariates at baseline $f_{jk}(\cdot)$ is available. Indeed, it is likely that only marginal covariate summaries are available from the AgD studies, but we can reconstruct the full joint distribution under the assumption that the form of the marginal covariate distributions and the

pairwise correlations are the same as those observed in the IPD studies (see Section 4.5.1).

In general, the individual-level conditional likelihood is defined by two quantities: a survival function giving the probability of surviving to time t on treatment k in study j given covariates $\boldsymbol{x}$:

$$S_{jk}(t|\boldsymbol{x}) = \mathbb{P}(T \geq t|\boldsymbol{x}, j, k), \tag{7.4}$$

and a hazard function describing the instantaneous hazard rate at time t:

$$\lambda_{jk}(t|\boldsymbol{x}) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt|T \geq t, \boldsymbol{x}, j, k)}{dt}. \tag{7.5}$$

The individual conditional likelihood contributions from each time t_{ijk} in the IPD are given by

$$L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; t_{ijk}, y_{ijk}, \boldsymbol{x}_{ijk}) = S_{jk}(t_{ijk}|\boldsymbol{x}_{ijk})\lambda_{jk}(t_{ijk}|\boldsymbol{x}_{ijk})^{y_{ijk}}, \tag{7.6}$$

so that both observed and censored events get a contribution for surviving up until time t_{ijk}, but only observed events get the additional contribution from the instantaneous hazard rate at time t_{ijk}. The forms of the survival and hazard functions depend on the specific model chosen, but the framework described here may be applied in any case, as long as both the survival and hazard functions are specified. This includes parametric proportional hazards models (Exponential, Weibull, Gompertz, etc.) (see Cox 1984, Chapter 2; Collett 2003, Chapter 5), parametric accelerated failure time models (Weibull, log-Logistic, Gamma, log-Normal, etc.) (see Collett 2003, Chapter 6), and flexible parametric baseline hazard models (using splines as in the Royston-Parmar model (Freeman and Carpenter 2017; Royston and Parmar 2002), or fractional polynomials (Jansen 2011), etc.). Table 7.1 gives the survival and hazard functions for some of these commonly-used survival models.

Using equation (7.1), the individual marginal likelihood contributions are

$$L_{ijk}(\boldsymbol{\xi}; t_{ijk}, y_{ijk}) = \int_{\mathfrak{X}} L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; t_{ijk}, y_{ijk}, \boldsymbol{x}) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \tag{7.7}$$

$$= \int_{\mathfrak{X}} S_{jk}(t_{ijk}|\boldsymbol{x})\lambda_{jk}(t_{ijk}|\boldsymbol{x})^{y_{ijk}} f_{jk}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{7.8}$$

Since we have the (reconstructed) outcome times for each individual in the AgD studies, we use the individual marginal likelihood contributions directly (rather than taking their product to obtain the aggregate likelihood contribution for a single summary outcome). Unlike the previous examples, this integral cannot be simplified analytically; instead, we employ the QMC integration

Table 7.1 Survival and hazard functions for some common parametric survival models.

Survival model	Parameters	Survival and hazard functions
Proportional hazards		
Exponential	Hazard rate θ_{jk}, modelled with $\exp(\eta_{jk}(\boldsymbol{x}))$	$S(t) = \exp(-\theta t)$ $\lambda(t) = \theta$
Weibull	Hazard rate θ_{jk}, modelled with $\exp(\eta_{jk}(\boldsymbol{x}))$; shape ν_j	$S(t) = \exp(-\theta t^{\nu})$ $\lambda(t) = \nu\theta t^{\nu-1}$
Gompertz	Hazard rate θ_{jk}, modelled with $\exp(\eta_{jk}(\boldsymbol{x}))$; shape ν_j	$S(t) = \exp\left(-\frac{\theta}{\nu}\left(\exp(t\nu) - 1\right)\right)$ $\lambda(t) = \theta\exp(t\nu)$
Accelerated failure time		
Weibull	Acceleration factor θ_{jk}, modelled with $\exp(-\eta_{jk}(\boldsymbol{x}))$; shape ν_j	$S(t) = \exp(-\theta^{\nu} t^{\nu})$ $\lambda(t) = \nu\theta^{\nu} t^{\nu-1}$
log-Normal	Acceleration factor θ_{jk}, modelled with $\exp(-\eta_{jk}(\boldsymbol{x}))$; variance σ_j^2; denote the Normal probability density function $\phi(\cdot)$ and cumulative density function $\Phi(\cdot)$	$S(t) = 1 - \Phi\left(\frac{\log(t)-\log(\theta)}{\sigma}\right)$ $\lambda(t) = \frac{\phi\left(\frac{\log(t)-\log(\theta)}{\sigma}\right)}{t\sigma\,\Phi\left(\frac{\log(t)-\log(\theta)}{\sigma}\right)}$
log-Logistic	Acceleration factor θ_{jk}, modelled with $\exp(-\eta_{jk}(\boldsymbol{x}))$; shape ν_j	$S(t) = \frac{1}{1+(t/\theta)^{\nu}}$ $\lambda(t) = \frac{(\nu/\theta)(t/\theta)^{\nu-1}}{1+(t/\theta)^{\nu}}$
Flexible baseline hazards		
Royston-Parmar (splines)	Spline function $\zeta_j(\log(t))$, with derivative $d\zeta_j(\log(t))$ with respect to $\log(t)$; hazard rate θ_{jk} modelled with $\exp(\eta_{jk}(\boldsymbol{x}))$	$S(t) = \exp(-\exp(\zeta(\log(t)))\theta)$ $\lambda(t) = d\zeta(\log(t))\exp(\zeta(\log(t)))\theta$

approach detailed in Section 4.3.3.2 to evaluate the integral. With a set of $\tilde{N}$ integration points $\tilde{\boldsymbol{x}}_{jk}$ drawn from $f_{jk}(\cdot)$, the individual marginal likelihood contributions are evaluated as

$$L_{ijk}(\boldsymbol{\xi}; t_{ijk}, y_{ijk}) = \tilde{N}^{-1} \sum_{\tilde{\boldsymbol{x}}} S_{jk}(t_{ijk}|\tilde{\boldsymbol{x}}) \lambda_{jk}(t_{ijk}|\tilde{\boldsymbol{x}})^{y_{ijk}} \tag{7.9}$$

We apply this model later in Section 7.3 to an example of simulated survival outcomes.

7.1.2 Binary outcome data

Suppose that we have binary outcomes $y_{ijk} \sim \text{Bern}(p_{ijk})$. In this case, the individual conditional likelihood contributions are

$$L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}_{ijk}) = p_{ijk}^{y_{ijk}} (1 - p_{ijk})^{(1-y_{ijk})},$$

where the individual event probabilities p_{ijk} are modelled using $\theta_{ijk} = g^{-1}(\eta_{jk}(\boldsymbol{x}_{ijk}))$ with a suitable link function $g(\cdot)$ (e.g. a logit or probit link

function). Using equation (7.1), the individual marginal likelihood contributions are

$$\begin{aligned} L_{ijk}(\xi; y_{ijk}) &= \int_{\mathfrak{X}} L_{ijk|\boldsymbol{x}}(\xi; y_{ijk}, \boldsymbol{x}) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\mathfrak{X}} g^{-1}(\eta_{jk}(\boldsymbol{x}))^{y_{ijk}} \left(1 - g^{-1}(\eta_{jk}(\boldsymbol{x}))\right)^{(1-y_{ijk})} f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \bar{p}_{jk}^{\,y_{ijk}} (1 - \bar{p}_{jk})^{(1-y_{ijk})} \end{aligned}$$

where $\bar{p}_{jk} = \int_{\mathfrak{X}} g^{-1}(\eta_{jk}(\boldsymbol{x})) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x}$ is the mean event probability on treatment k in study j (as in Sections 4.3.2.3 and 4.3.2.4), since $y_{ijk} \in \{0, 1\}$. The aggregate likelihood contribution for $y_{\bullet jk}$ events out of N_{jk} individuals on treatment k in study j is then proportional to the product of $y_{\bullet jk}$ many $L_{ijk}(\xi; 1)$ terms and $N_{jk} - y_{\bullet jk}$ many $L_{ijk}(\xi; 0)$ terms:

$$\begin{aligned} L_{\bullet jk}(\xi; y_{\bullet jk}) &\propto L_{ijk}(\xi; 1)^{y_{\bullet jk}} L_{ijk}(\xi; 0)^{(N_{jk} - y_{\bullet jk})} \\ &= \bar{p}_{jk}^{\,y_{\bullet jk}} (1 - \bar{p}_{jk})^{(N_{jk} - y_{\bullet jk})} \end{aligned}$$

which we recognise as a $\text{Bin}(N_{jk}, \bar{p}_{jk})$ likelihood. In other words, we recover the one-parameter Binomial likelihood described in Section 4.2.1.

In Section 4.2.1, we improved upon the one-parameter Binomial likelihood with a two-parameter Binomial likelihood in which both $\bar{p}_{jk}$ and N_{jk} were adjusted, aiming to obtain a likelihood closer to the "true" Poisson Binomial aggregate likelihood. The Poisson Binomial likelihood describes the total number of events given a vector of individual probabilities, where the exact individuals experiencing an event are unknown; however, we cannot use this likelihood directly as the parameter vector is not identifiable given the aggregate data. Instead, the one-parameter Binomial likelihood assigns the same event probability $\bar{p}_{jk}$ to each individual on treament k in study j; however, this is not the most efficient model since we know that the individual event probabilities differ. The two-parameter Binomial likelihood acknowledges this, and as a result has a smaller variance (matching that of the Poisson Binomial). Given full IPD, the individual-level Bernoulli likelihood would additionally make use of the information on precisely which individuals experienced events. Intuitively then, the two-parameter Binomial likelihood lies in between the one-parameter Binomial likelihood and the full IPD individual-level Bernoulli likelihood in terms of efficiency. The marginal likelihood approach is not "wrong" here, it is just not the most efficient. In this case, we can improve on the one-parameter Binomial likelihood obtained through the marginal likelihood approach since we know the "true" form of the aggregate likelihood, although in practice we find that this makes little difference to the results.

However, this will not be possible in general, as we cannot always derive (or even approximate) the appropriate likelihood distribution for the aggregate data.

7.1.3 Ordered categorical outcome data

In the plaque psoriasis example introduced in Chapter 6, the outcomes of interest are success/failure to achieve 75%, 90%, and 100% improvement in PASI score from baseline, denoted PASI 75, 90, and 100 respectively. These are ordered categorical outcomes, where obtaining an outcome (e.g. PASI 90) necessarily means that the preceding outcomes have also been obtained (e.g. PASI 75). Synthesis of such outcomes using aggregate data NMA has been described previously (Dias et al. 2011c; Woolacott et al. 2006), using a multinomial likelihood with ordered latent cutoffs on the underlying transformed scale (e.g. with a probit link function).

In general, let us consider outcomes $y_{ijk} \in \{1, \ldots, M\}$ in M ordered categories. At the individual level, these have a categorical likelihood

$$L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}_{ijk}) = \begin{cases} p_{ijk;1} & \text{if} \quad y_{ijk} = 1 \\ \vdots & \qquad \vdots \\ p_{ijk;M} & \text{if} \quad y_{ijk} = M \end{cases}, \quad \text{where} \sum_{m=1}^{M} p_{ijk,m} = 1 \quad \forall i, j, k \tag{7.10}$$

The individual category probabilities $p_{ijk;m}$ are modelled by

$$p_{ijk,m} = g^{-1}(\eta_{jk}(\boldsymbol{x}_{ijk}) - c_{m-1}) - g^{-1}(\eta_{jk}(\boldsymbol{x}_{ijk}) - c_m), \tag{7.11}$$

with a suitable link function $g(\cdot)$, such as a probit or logit link function. The latent cutpoints c_m are subject to the ordering constraints

$$c_0 < c_1 < \cdots < c_{M-1} < c_M,$$

with $c_0 = -\infty$ and $c_M = +\infty$ so that

$$p_{ijk,1} = 1 - g^{-1}(\eta_{jk}(\boldsymbol{x}_{ijk}) - c_1)$$
$$p_{ijk,M} = g^{-1}(\eta_{jk}(\boldsymbol{x}_{ijk}) - c_{M-1})$$

We further set $c_1 = 0$, since c_1 is equivalent to a global intercept and cannot be identified at the same time as the study-specific intercepts μ_j. This model reduces to that for binary data with a Bernoulli likelihood when $M = 2$. As such, we proceed with the derivations in a very similar manner to before (Section 7.1.2).

Using equation (7.1), the individual marginal likelihood contributions are

$$\begin{aligned} L_{ijk}(\boldsymbol{\xi}; y_{ijk}) &= \int_{\mathfrak{X}} L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\mathfrak{X}} \left(g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_{m-1}) - g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_m)\right) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \quad \text{if } y_{ijk} = m \\ &= \begin{cases} \bar{p}_{jk,1} & \text{if} \quad y_{ijk} = 1 \\ \vdots & \qquad\vdots \\ \bar{p}_{jk,M} & \text{if} \quad y_{ijk} = M \end{cases} \end{aligned}$$

where

$$\bar{p}_{jk,m} = \int_{\mathfrak{X}} \left(g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_{m-1}) - g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_m)\right) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \tag{7.12}$$

is the mean event probability for outcome m on treatment k in study j.

The aggregate data are now vectors of (mutually exclusive) outcome counts $\boldsymbol{y}_{\bullet jk} = (y_{\bullet jk,1}, \ldots, y_{\bullet jk,M})^{\intercal}$, where $\sum_{m=1}^{M} y_{\bullet jk,m} = N_{jk}$. The model is parameterised so that the count data are mutually exclusive: individuals are each assigned to one of M mutually exclusive intervals defined by the M categories (so for example, an individual achieving the PASI 90 outcome is only counted in the PASI 90 category, and *not* counted for PASI 75 also). The aggregate likelihood contributions from each $\boldsymbol{y}_{\bullet jk}$ are thus proportional to the product of $y_{\bullet jk,1}$ many $L_{ijk}(\boldsymbol{\xi}; 1)$ terms, $y_{\bullet jk,2}$ many $L_{ijk}(\boldsymbol{\xi}; 2)$ terms, and so on, up to $y_{\bullet jk,M}$ many $L_{ijk}(\boldsymbol{\xi}; M)$ terms:

$$\begin{aligned} L_{\bullet jk}(\boldsymbol{\xi}; \boldsymbol{y}_{\bullet jk}) &\propto \prod_{m=1}^{M} L_{ijk}(\boldsymbol{\xi}; m)^{y_{\bullet jk,m}} \\ &= \prod_{m=1}^{M} \bar{p}_{jk,m}^{\,y_{\bullet jk}} \end{aligned} \tag{7.13}$$

where proportionality is up to a normalising constant. We recognise this as a multinomial distribution $\text{Multi}(\bar{p}_{jk,1}, \ldots, \bar{p}_{jk,M})$, with average event probability $\bar{p}_{jk,m}$ in each category m. Notice that the characterisation as a multinomial distribution is well-defined, since we have for each j and k that

$$\begin{aligned} \sum_{m=1}^{M} \bar{p}_{jk,m} &= \sum_{m=1}^{M} \int_{\mathfrak{X}} \left(g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_{m-1}) - g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_m)\right) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\mathfrak{X}} \sum_{m=1}^{M} \left(g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_{m-1}) - g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_m)\right) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\mathfrak{X}} \left(g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_0) - g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_M)\right) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \end{aligned}$$

and since $c_0 = -\infty$ and $c_M = +\infty$

$$= \int_{\mathfrak{X}} 1 \cdot f_{jk}(\boldsymbol{x})\, d\boldsymbol{x}$$
$$= 1.$$

Fitting the aggregate part of the model using the aggregate marginal likelihood contributions in (7.13) requires the calculation of the average event probabilities $\bar{p}_{jk,m}$ at every iteration. In practice, we perform the necessary integration in equation (7.12) numerically (Section 4.3.3), which carries a non-trivial computational cost. To reduce the computational cost, we can avoid unnecessary repeated calculations by rewriting (7.12) as

$$\begin{aligned}
\bar{p}_{jk,m} &= \int_{\mathfrak{X}} \left(g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_{m-1}) - g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_m) \right) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{\mathfrak{X}} g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_{m-1}) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} - \int_{\mathfrak{X}} g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_m) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \bar{q}_{jk,m-1} - \bar{q}_{jk,m},
\end{aligned}$$

where

$$\bar{q}_{jk,m} = \int_{\mathfrak{X}} g^{-1}(\eta_{jk}(\boldsymbol{x}) - c_m) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x}, \tag{7.14}$$

with $\bar{q}_{jk,0} = 1$ and $\bar{q}_{jk,M} = 0$. We compute each $\bar{q}_{jk,m}$ only once, and re-use the necessary values when calculating $\bar{p}_{jk,m}$. Further cost savings can be made by computing $\eta_{jk}(\boldsymbol{x})$ only once, and using the stored value in the calculation of each $\bar{q}_{jk,m}$.

We apply this model later in Section 7.4 to the plaque psoriasis example.

7.2 Model comparison

There are three broad scenarios in the general framework given by equation (7.3) depending on how the aggregate likelihood is evaluated, which each entail different approaches to model comparison:

1. The integrand in (7.3c) may be simplified, so that the aggregate-level likelihood is available in closed form (as in (4.34));

2. There is no closed-form aggregate likelihood and individual-level outcomes are available, so the aggregate-level model is fitted using the individual marginal likelihood contributions $L_{ijk}(\xi, y_{ijk})$;

3. There is no closed-form aggregate likelihood and only aggregate-level outcomes are available, so the aggregate-level model is fitted using the aggregate marginal likelihood contributions $L_{\bullet jk}(\xi, y_{\bullet jk})$.

(In all scenarios, there is only aggregate covariate information available from the AgD studies.) Scenario 1 is exactly that discussed in Chapter 4; the form of the aggregate likelihood is known and we can calculate the residual deviance D_{res} and the effective number of parameters p_D, and hence the DIC, for model comparison purposes (see Section 5.3).

For scenarios 2 and 3 the aggregate likelihood has no closed form, which complicates calculation of the DIC. Instead, we propose to use approximate leave-one-out (LOO) cross validation for model comparison.

Cross validation is a widely-used method for assessing the predictive ability of regression models (Picard and Cook 1984). Performance in a future data set is approximated by repeatedly re-fitting the model in question to the sample data set, each time leaving out a different "hold-out" point (a single point for LOO cross validation, or a partition of points for k-fold cross validation) against which the predictive performance is evaluated. A general measure of predictive performance for a given data point is the *log predictive density*, which is equal to the expectation of the log likelihood of the data point over the posterior distribution of the model parameters (Gelman et al. 2013b). Cross validation then provides an approximation to the *expected log predictive density* (ELPD) in a new data set. The ELPD is a measure of the expected predictive performance of the model, and given a set of candidate models we would choose the model with the greatest ELPD. Vehtari et al. (2016) also define a LOO information criterion (LOOIC) as $-2 \cdot \text{ELPD}$, transforming ELPD onto the deviance scale to be used in an analogous manner to DIC (i.e. lower values are better). The DIC (Section 1.2.6) can also be seen as an approximation to the ELPD, up to the scaling by -2, where plug-in posterior values are used for the log predictive density rather than an expectation over the full posterior distribution (see Gelman et al. 2013b).

Exact LOO cross validation is computationally expensive: for each data point in turn, the model must be re-fit without that data point, and the log predictive density of the hold-out point evaluated. Instead, it is possible to estimate the ELPD using an importance sampling approach, where the posterior samples of the log predictive density (i.e. the log likelihood for each data point at each posterior sample of the model parameters) are weighted to approximate the LOO cross validation ELPD (Gelfand et al. 1992). However, the importance weights are often very unstable. Vehtari et al. (2016) propose an improved approximate LOO approach, called Pareto-smoothed importance sampling LOO (PSIS-LOO), where the importance weights are stabilised (smoothed) by fitting a Pareto distribution to the upper tail of the distribution

of importance weights. The R package `loo` implements PSIS-LOO, and provides diagnostics for determining the adequacy of the PSIS approximation. Vehtari et al. (2016) also describe how to calculate an estimate p_{LOO} of the effective number of parameters, and standard errors for ELPD, LOOIC, and p_{LOO}. Standard errors facilitate the judgement of substantial differences in these quantities when comparing models. By comparison, when using DIC the common approach is to compare differences with a χ^2 distribution, which is theoretically justified for Normal linear models or asymptotically but may not always be appropriate in practice (Lunn et al. 2010, pp. 165–167; Dias et al. 2018, p. 69). Importantly, PSIS-LOO does not require the model to be re-run, and the required inputs are simply the posterior samples of the log likelihood contribution for each data point.

Survival analysis, as discussed in Section 7.1.1, falls under scenario 2: there is no closed-form aggregate likelihood, and it is assumed that we have event/censoring times for every individual (e.g. reconstructed from Kaplan-Meier curves). In this scenario in general, to calculate the ELPD using PSIS-LOO we use the posterior samples of the individual conditional log likelihood contributions $\log L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}_{ijk})$ for the individuals in the IPD studies (equation (7.3a)), and the posterior samples of the individual marginal log likelihood contributions $\log L_{ijk}(\boldsymbol{\xi}; y_{ijk})$ for the individuals in the AgD studies (equation (7.3c)).

In scenario 3, we have only aggregate outcomes in the AgD trials. This poses problems for model comparison with PSIS-LOO, as the PSIS approximation breaks down for highly influential data points. If we only have a small number of AgD studies on a given treatment k, then each of these is influential in the estimation of the the treatment effect γ_k (the remaining regression coefficients for the prognostic and effect modifying variables will be almost entirely informed by the IPD). In the extreme case there is only a single AgD study informing a given treatment k, and LOO cross validation for this data point is undefined—there is no way to predict this data point from the other studies.

To avoid these issues, we must modify the PSIS-LOO approach described above. The simplest option is to perform model comparison based on the IPD only, ignoring the AgD studies. This may be a justifiable strategy when the number of AgD studies is small, given that the AgD will be contributing very little to the estimation of the regression model. With only a single AgD study informing a treatment k, the model is essentially saturated for that treatment—the treatment k arm informs exactly one parameter, γ_k, which

provides perfect fit to that data point regardless of the model—and this AgD study arm can be ignored for the purpose of model comparison. A more rigorous solution, at the expense of increased computational effort, is to use the approach referred to by Vehtari et al. (2016) as "PSIS-LOO+". PSIS-LOO is used as a first pass, followed by exact LOO cross validation for the problematic data points where the approximation is inadequate, re-fitting the model with these data points omitted one by one and computing their ELPD contributions directly. Again, any treatments where the model is saturated (only a single AgD study arm providing information) can be safely ignored.

Example: survival analysis 7.3

In this section we consider an artificial example of survival (time-to-event) outcomes. Since the data are simulated, we can compare the results and performance of ML-NMR using only partial IPD to that of a full IPD NMA, and to the known true values.

Artificial scenario 7.3.1

The artificial scenario involves two studies, one comparing treatments A and B, and the other comparing treatments A and C. The AB study randomised 500 individuals 1:1 to each treatment, and the AC study randomised 400 individuals 1:1. For each individual, we generate three covariates according to the following (independent) distributions in each study:

AB study:

$$X_{1(AB)} \sim \mathrm{N}\left(0, 0.5^2\right), \qquad X_{2(AB)} \sim \mathrm{Gam}(4, 2), \qquad X_{3(AB)} \sim \mathrm{Bern}(0.2)$$

AC study:

$$X_{1(AC)} \sim \mathrm{N}\left(1, 0.4^2\right), \qquad X_{2(AC)} \sim \mathrm{Gam}(6, 2), \qquad X_{3(AC)} \sim \mathrm{Bern}(0.7)$$

Outcomes are simulated from a Weibull model in each study, with scales $\alpha_{(AB)} = 6.2$, $\alpha_{(AC)} = 5.8$ and shapes $\nu_{(AB)} = 0.8$, $\nu_{(AC)} = 1.2$. Under a proportional hazards model, the hazard function at time t for an individual i on treatment k in study j is then given by

$$\lambda_{jk}(t|\boldsymbol{x}_{ijk}) = \lambda_{j;0}(t) \exp\left(\boldsymbol{x}_{ijk}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k\right),$$

where $\boldsymbol{x}_{ijk} = (x_{ijk;1}, x_{ijk;2}, x_{ijk;3})^{\mathsf{T}}$ is a vector of simulated covariates, $\lambda_{j;0}(t)$ is a baseline hazard function for trial j, $\boldsymbol{\beta}_1$ are prognostic coefficients, $\boldsymbol{\beta}_{2,k}$ are

effect modifying coefficients, and γ_k are individual-level treatment effects. For this simulation we set these coefficients equal to

$$\begin{aligned}\boldsymbol{\beta_1} &= (0.1, 0.05, -0.25)^\intercal \\ \boldsymbol{\beta}_{2,A} &= \mathbf{0}, \quad \boldsymbol{\beta}_{2,B} = \boldsymbol{\beta}_{2,C} = (-0.2, -0.2, -0.1)^\intercal \\ \gamma_A &= 0, \quad \gamma_B = -1.2, \quad \gamma_C = -0.5\end{aligned}$$

Notice that all covariates are both prognostic and effect modifying, and we have $\boldsymbol{\beta}_{2,B} = \boldsymbol{\beta}_{2,C}$ so that the shared EM assumption holds. For the Weibull model, the baseline hazard function is

$$\lambda_{j;0}(t) = \nu_j \alpha_j t^{\nu_j - 1}.$$

Survival times are simulated using the Cumulative Distribution Function inversion method described by Bender et al. (2005), implemented in the R package `simsurv` (Brilleman 2018). We censor all surviving individuals at time $t = 1$ for both studies, and further uniformly censor 10% of individuals within each study. The resulting Kaplan-Meier survival curves are shown in Figure 7.1. For the ML-NMR analysis, we provide only summary covariate information for the AC trial (means and standard deviations for the continuous covariates, and a proportion for the discrete covariate).

7.3.2 Methods

We implement the ML-NMR model for general likelihoods in equation (7.3), described for survival outcomes in Section 7.1.1. We will fit Exponential, Weibull, and Gompertz proportional hazards models, and use the LOOIC (Section 7.2) to select the most appropriate model.

For each model, the linear predictor is

$$\eta_{jk}(\boldsymbol{x}_{ijk}) = \beta_{0,j} + \boldsymbol{x}_{ijk}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k, \tag{7.15}$$

and we set $\boldsymbol{\beta}_{2,A} = \mathbf{0}$, $\gamma_A = 0$. Note that the study-specific intercept $\beta_{0,j}$ is the log scale parameter for each study, so that $\alpha_j = \exp(\beta_{0,j})$. We also define $\theta_{jk}(\boldsymbol{x}) = \exp(\eta_{jk}(\boldsymbol{x}))$, which are interpreted as hazard rates modelled with a log link function.

The Weibull model[1] is specified by the survival and hazard functions

$$S_{jk}(t|\boldsymbol{x}) = \exp\left(-\theta_{jk}(\boldsymbol{x})t^{\nu_j}\right) \tag{7.16a}$$

$$\lambda_{jk}(t|\boldsymbol{x}) = \nu_j \theta_{jk}(\boldsymbol{x}) t^{\nu_j - 1}. \tag{7.16b}$$

[1]Note that Stan uses an alternative Weibull parameterisation, based on a scale (inverse rate) parameter instead of a hazard rate. Model (7.16) is equivalent to a model on the scale parameters using $\exp\left(-\frac{\eta_{jk}(\boldsymbol{x})}{\nu_j}\right)$.

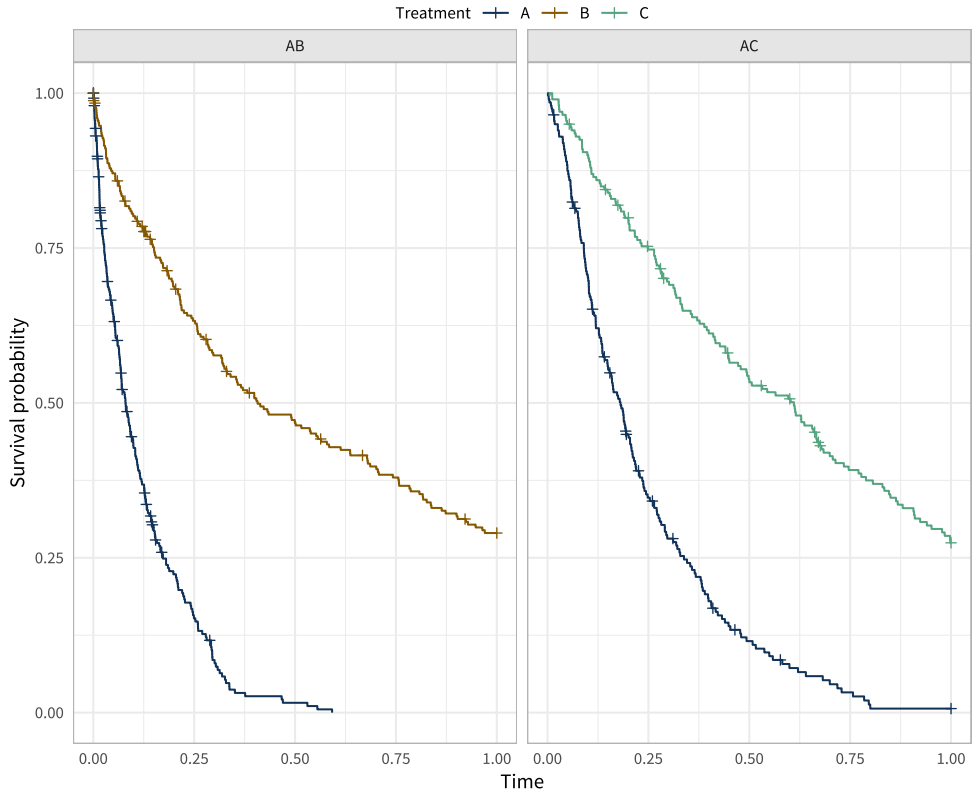


Figure 7.1 Simulated Kaplan-Meier survival curves for each treatment in each study. Censored events are marked with a cross (+).

The Exponential model is a special case of the Weibull model where $\nu_j = 1 \; \forall j$, so the survival and hazard functions are

$$S_{jk}(t|\boldsymbol{x}) = \exp\big(-t\theta_{jk}(\boldsymbol{x})\big) \tag{7.17a}$$

$$\lambda_{jk}(t|\boldsymbol{x}) = \theta_{jk}(\boldsymbol{x}). \tag{7.17b}$$

Finally, the Gompertz model has survival and hazard functions

$$S_{jk}(t|\boldsymbol{x}) = \exp\left(-\frac{\theta_{jk}(\boldsymbol{x})}{\nu_j}\left(\exp(t\nu_j) - 1\right)\right) \tag{7.18a}$$

$$\lambda_{jk}(t|\boldsymbol{x}) = \theta_{jk}(\boldsymbol{x})\exp(t\nu_j). \tag{7.18b}$$

For an individual in the IPD AB study with event/censoring time t_{ijk} and covariates $\boldsymbol{x}_{ijk}$, their (individual conditional) likelihood contribution is given by substituting in $S_{jk}(t_{ijk}|\boldsymbol{x}_{ijk})$ and $\lambda_{jk}(t_{ijk}|\boldsymbol{x}_{ijk})$ from (7.16), (7.17), or (7.18) into equation (7.6). For each event/censoring time in the AgD AC study, the (individual marginal) likelihood contributions are given by equation (7.7). To evaluate the integral, we employ the numerical integration approach

described in Section 4.3.3.2, with $\tilde{N} = 100$ integration points $\tilde{\boldsymbol{x}}_{jk}$ drawn from joint distribution $f_{jk}(\cdot)$ of the covariates on each treatment k in study j. The likelihood contributions are then given by substituting $S_{jk}(t_{ijk}|\tilde{\boldsymbol{x}})$ and $\lambda_{jk}(t_{ijk}|\tilde{\boldsymbol{x}})$ into equation (7.9). The regression parameters $\beta_{0,j}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_{2,k}$, and γ_k are all given vague $\text{N}(0, 100^2)$ prior distributions, and (for the Weibull and Gompertz models) the shape parameters ν_j are given improper uniform prior distributions $\text{U}(0, +\infty)$. We assess convergence using $\hat{R}$ for each parameter, and check that there are no divergent transitions (see Section 5.2.5).

We also fit the Weibull, Exponential, and Gopmertz models (equations (7.16), (7.17), and (7.18)) in an IPD NMA with full IPD (i.e. individual outcomes and covariates) available from both studies. The likelihood contributions for every individual in both studies are given by (7.6). We also perform a standard (unadjusted) indirect comparison: log hazard ratios are estimated in each study separately using a Weibull model without adjustment for effect modifiers, d_{AB} for the AB study and d_{AC} for the AC study, and then the indirect comparison between B and C is formed as $d_{BC} = d_{AC} - d_{AB}$.

7.3.3 Results

We begin by comparing the expected predictive performance of each model. Table 7.2 shows the LOO model comparison statistics (see Section 7.2) for each of the three models, fit using ML-NMR with only aggregate covariate information in the AC study, and also fit using IPD NMA with full IPD. The LOOIC is lowest (and, equivalently, ELPD highest) for the Weibull model, for both ML-NMR and IPD NMA. The difference in ELPD, when compared with the standard error of the difference, suggests that the Weibull model is a substantially better fit than either the Exponential or Gompertz models, in both the ML-NMR and IPD NMA scenarios. We therefore present the results for the Weibull model in the remainder of this section, comparing between ML-NMR, full IPD NMA, and the true values used for simulation. (Note that, since $\text{LOOIC} = -2 \cdot \text{ELPD}$, the standard errors for the differences in LOOIC are simply double the standard errors for the differences in ELPD—and we reach exactly the same conclusion on either scale.) Notably, the LOO model comparison statistics are very similar when the same model is fit with ML-NMR or with IPD NMA. Figure 7.2 compares the LOOIC contributions from each individual event/censoring time between the Weibull model fit using ML-NMR and using full IPD NMA. The LOOIC contributions follow the straight line of equality well, showing that the same observations are fitted similarly well whether the full IPD was used or aggregate AC data. The

Table 7.2 Model comparison results, using full IPD NMA and ML-NMR. The leave-one-out information criterion (LOOIC) is equal to $-2 \cdot$ ELPD, where ELPD is the expected log pointwise predictive density, and lower LOOIC values indicate better expected predictive performance. p_{loo} is the effective number of parameters. The ELPD differences are in comparison with the respective Weibull models, with positive values favouring the Weibull model. Standard errors for each statistic are given alongside in small brackets.

	IPD NMA			ML-NMR		
	Exponential	Gompertz	Weibull	Exponential	Gompertz	Weibull
LOOIC	−231.5 (63.5)	−232.9 (63.4)	−251.7 (64.5)	−214.5 (63.8)	−214.8 (63.7)	−234.4 (64.6)
ELPD	115.8 (31.8)	116.4 (31.7)	125.9 (32.2)	107.3 (31.9)	107.4 (31.9)	117.2 (32.3)
p_{LOO}	8.4 (0.6)	9.9 (0.7)	9.8 (0.7)	8.8 (0.9)	10.3 (1.0)	9.5 (0.8)
ELPD difference	10.1 (4.7)	9.4 (4.3)		9.9 (4.6)	9.8 (4.2)	

only noticable exception to this is a horizontal cluster of LOOIC contributions for a set of censoring times in the C treatment arm of the AgD AC trial, which all have LOOIC contributions around 2.5 under the ML-NMR model. These correspond to individuals all censored at the end of the trial ($t = 1$), which under the ML-NMR model are all given the same marginal likelihood contribution.

The estimated population-average survival curves on each treatment in each study population under the Weibull model fitted using ML-NMR are shown in Figure 7.3, overlaid on the observed Kaplan-Meier curves. Visually, the estimated survival curves are a good fit to the observed data. Table 7.3 presents the estimated log Hazard Ratios (HRs) for each pairwise comparison in each population, along with the true values from the simulation. The ML-NMR estimates agree well with both the IPD NMA and the true values, and the B vs. A and C vs. A estimates within the AB and AC study populations respectively are unchanged in point estimate or standard error. Standard errors for comparisons not observed in the data are slightly increased (by 2–6%) using ML-NMR compared to full IPD NMA, which is expected due to the reduced information available. The standard (unadjusted) indirect comparison, also presented in Table 7.3, produces estimates that are clearly biased in this scenario. The B vs. A and C vs. A log HRs are assumed to be constant across both study populations, when in fact these are altered by effect modifying covariates. The B vs. A estimate is therefore incorrectly applied to the AC study population, resulting in an incorrect estimate of C vs. B in this population; similarly, the C vs. A estimate is invalid in the AB study, and the C vs. B estimate is invalid here too.

Examining the parameters from the ML-NMR and IPD NMA models in

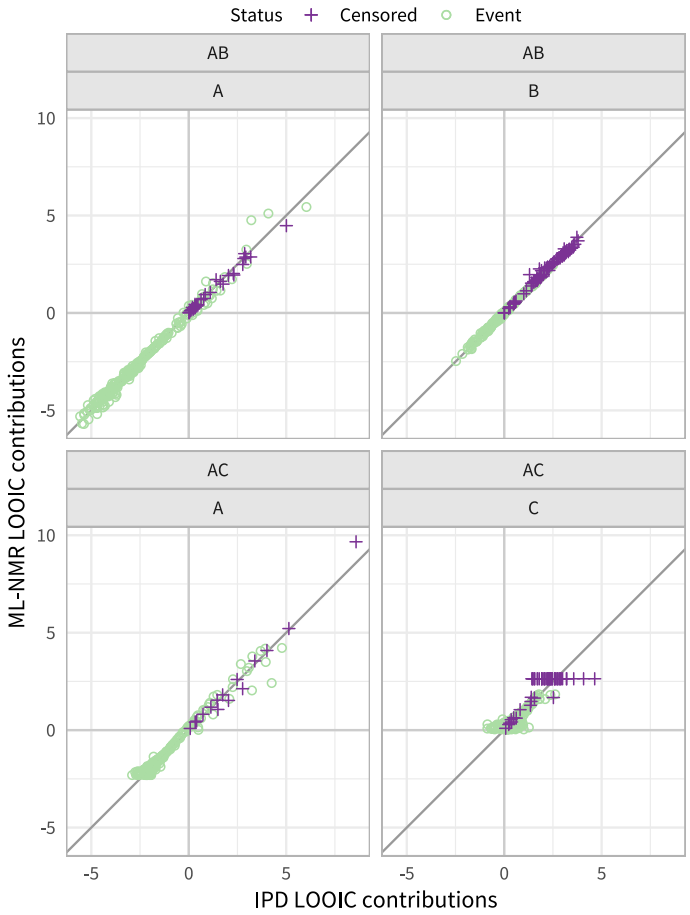


Figure 7.2 Contributions to the LOOIC from each event and censoring time in the Weibull model, plotted for ML-NMR using only summary covariate information in the AC study against an IPD NMA with full information from every study.

Table 7.4, we see that these agree closely with each other and recover the true parameter values well, with the true values lying within the majority of the Credible Intervals.

7.3.4 Conclusions

In this artificial example, we have demonstrated how the ML-NMR framework extended to general likelihoods (Section 7.1) can be used to fit survival models. Whilst we focused on parametric proportional hazards models here (Weibull, Exponetial, and Gompertz models), the derivations in Section 7.1.1 can be applied in the same manner to fit survival models with any given hazard and survival functions.

The results of the models fit with ML-NMR (using only aggregate covariate

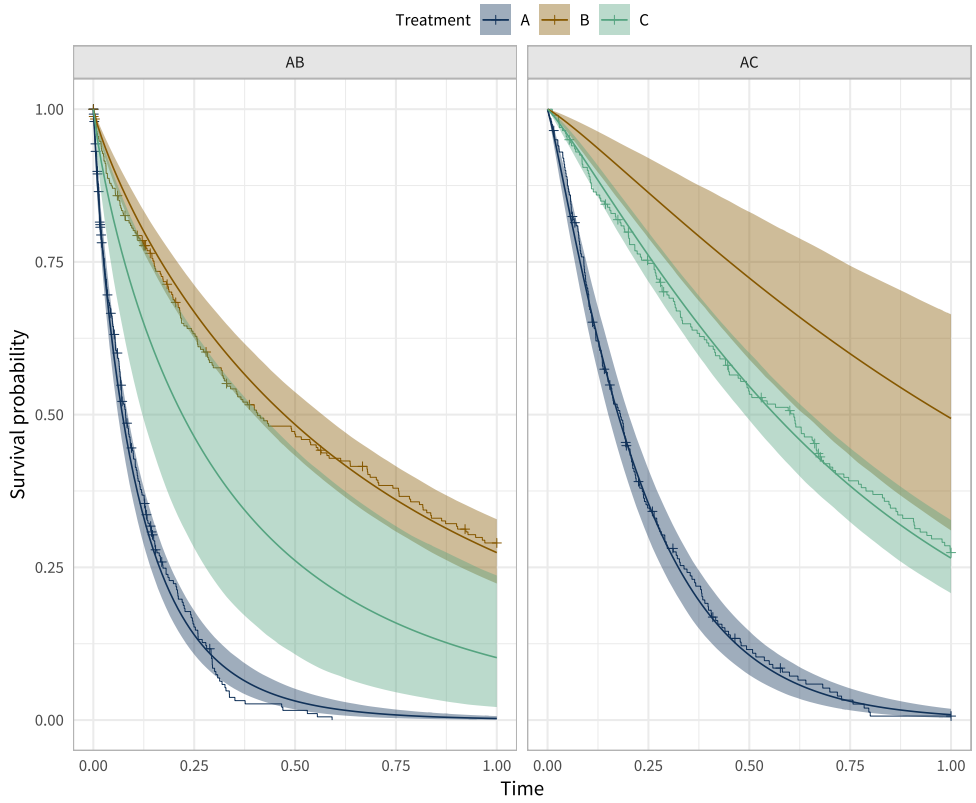


Figure 7.3 ML-NMR estimated survival curves on each treatment in each study population, under a Weibull model. Shaded bands indicate the 95% Credible Intervals for the survival curves (thick lines), overlaid on the observed Kaplan-Meier curves from the treatments in each study (thin lines).

information from the AC study) agree closely with the results from full IPD NMA. Furthermore, the lack of IPD in the AC study did not greatly reduce precision for ML-NMR compare to IPD NMA; the standard errors of population-average log Hazard Ratios were the same for comparisons observed within each study population, and only slightly increased for the comparisons not observed. As described in Section 7.2, we used the LOOIC to select the appropriate parametric model form. The conclusions of the model selection process were identical for ML-NMR and IPD NMA, in both cases correctly identifying the Weibull model as the most appropriate model from those fitted.

For the QMC numerical integration, we used $\tilde{N} = 100$ integration points. This number is much smaller than the 10,000 points used earlier in Chapter 6 for probit regression. The trade-off between the precision of the numerical integration and the computation time is much more severe for survival models (and indeed any ML-NMR models fit using the individual marginal likelihood),

Table 7.3 Table of estimated log Hazard Ratios and 95% Credible intervals from the ML-NMR model, the full IPD NMA, and the standard indirect comparison, alongside the true log Hazard Ratios, in the AB and AC study populations.

Study	Model	B vs. A	C vs. A	C vs. B
AB	Truth	−1.62	−0.92	0.70
	ML-NMR	−1.63 (−1.87, −1.41)	−0.98 (−1.49, −0.50)	0.65 (0.12, 1.21)
	IPD NMA	−1.62 (−1.85, −1.38)	−0.87 (−1.32, −0.40)	0.75 (0.23, 1.29)
	Standard IC	−1.60 (−1.84, −1.37)	−1.37 (−1.61, −1.14)	0.23 (−0.24, 0.70)
AC	Truth	−2.07	−1.37	0.70
	ML-NMR	−2.00 (−2.53, −1.47)	−1.35 (−1.60, −1.10)	0.65 (0.12, 1.21)
	IPD NMA	−2.12 (−2.61, −1.64)	−1.36 (−1.62, −1.11)	0.75 (0.23, 1.29)
	Standard IC	−1.60 (−1.84, −1.37)	−1.37 (−1.61, −1.14)	0.23 (−0.24, 0.70)

since the numerical integration is performed for every individual in the aggregate data, rather than only once per AgD treatment arm. Computation time for the Weibull ML-NMR model with $\tilde{N} = 100$ took around 5 minutes on a modern laptop (see Appendix C for details of the computing environment). Visual inspection of the 400 cumulative numerical integration plots for the individuals in the AC study confirmed that the expected convergence rate of $\tilde{N}^{-1}$ was achieved, and the numerical integration was accurate down to approximately 3 decimal places. Figure 7.4 shows an example of one of these plots, for an individual on treatment C in the AC study. The empirical integration error is estimated as a relative difference from the final estimate with $\tilde{N} = 100$ integration points, for each posterior sample; these are summarised using a smooth density estimate over the entire posterior distribution of the model parameters, in increasing steps of 10 integration points. We see that the empirical integration error decreases as $\tilde{N}^{-1}$ over the entire posterior distribution, as the density estimates lie within the dashed line as $\tilde{N}$ increases. Section 5.1.2 describes these plots, their construction, and interpretation in greater detail. Whilst $\tilde{N}$ could be increased for more precise integration (perhaps leveraging the computing power of a supercomputer), $\tilde{N} = 100$ seems to be adequate in this scenario. In other scenarios with greater numbers of covariates and/or different hazard and survival functions a larger value of

Table 7.4 Table of estimated model parameters and 95% Credible Intervals from the ML-NMR model and the full IPD NMA, alongside the true values used in the simulation.

Parameter		Truth	IPD NMA	ML-NMR
Treatment Effect	γ_B	−1.20	−1.28 (−1.61, −0.95)	−1.08 (−1.53, −0.62)
	γ_C	−0.50	−0.52 (−1.12, 0.10)	−0.43 (−1.11, 0.26)
Prognostic Effect	$\beta_{1;1}$	0.10	0.17 (−0.05, 0.39)	0.12 (−0.14, 0.37)
	$\beta_{1;2}$	0.05	0.02 (−0.07, 0.11)	0.14 (−0.01, 0.28)
	$\beta_{1;3}$	−0.25	−0.32 (−0.57, −0.08)	−0.12 (−0.45, 0.21)
EM Interaction	$\beta_{2;1}$	−0.20	−0.23 (−0.54, 0.10)	−0.20 (−0.63, 0.20)
	$\beta_{2;2}$	−0.20	−0.15 (−0.28, −0.03)	−0.24 (−0.44, −0.05)
	$\beta_{2;3}$	−0.10	−0.22 (−0.59, 0.14)	−0.41 (−0.94, 0.10)
Shape	ν_{AB}	0.80	0.86 (0.79, 0.93)	0.86 (0.79, 0.94)
	ν_{AC}	1.20	1.16 (1.06, 1.26)	1.15 (1.05, 1.26)
Scale	α_{AB}	6.20	6.82 (5.28, 8.62)	5.24 (3.72, 7.27)
	α_{AC}	5.80	5.17 (3.23, 7.68)	3.07 (1.91, 4.89)

$\tilde{N}$ may be necessary, and the cumulative integration plots should always be checked for convergence and suitable precision.

Whilst ML-NMR and IPD NMA were seen to perform very similarly in this scenario, the additional IPD available to IPD NMA does offer additional possibilities for analysis. For example, ML-NMR makes the shared EM assumption (Section 2.5) in this scenario for identifiability. In the interests of a fair comparison between ML-NMR and IPD NMA, both methods made use of this assumption in this analysis—which is known to hold due to the simulation setup. However, IPD NMA could relax this assumption and estimate separate EM interaction coefficients $\boldsymbol{\beta}_{2,B}$ and $\boldsymbol{\beta}_{2,C}$, rather than assuming equality. (In this scenario, since we know that $\boldsymbol{\beta}_{2,B} = \boldsymbol{\beta}_{2,C}$, the standard errors for IPD NMA would be inflated by the unnecessarily more flexible model.) In larger treatment networks it is possible to relax the shared EM assumption for

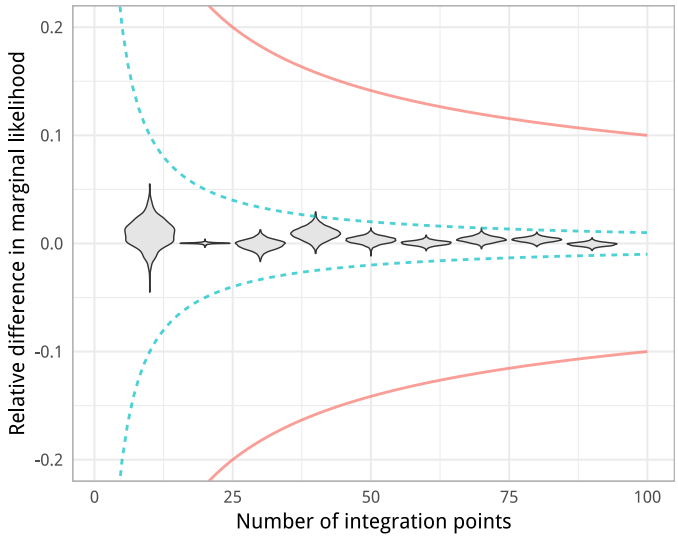


Figure 7.4 An example plot of empirical integration error using QMC integration to evaluate the marginal likelihood, for one individual on treatment C in the AC trial. The integration error is estimated as a relative difference from the final estimate (with $\tilde{N} = 100$ integration points), over the entire posterior distribution of the model parameters (i.e. at each posterior sample). The dashed line is $\pm\tilde{N}^{-1}$, showing the the integration error is of this order; for comparison, the solid line is $\pm\tilde{N}^{-\frac{1}{2}}$, which is the expected error rate for standard (pseudo-random) Monte Carlo integration.

ML-NMR, as described in Section 4.6.

Finally, although ML-NMR was seen to perform well in comparison to IPD NMA and the known truth, this scenario is only a single instance. Further simulations in a simulation study could validate the performance of ML-NMR for survival analysis, and investigate the impact of invalid assumptions. However, we expect the results and conclusions of the simulation study on binary outcomes in Chapter 8 to apply broadly to ML-NMR models of general forms—including for survival analysis.

7.4 Example: plaque psoriasis with ordered categorical PASI outcomes

In Chapter 6, we introduced an example comparing plaque psoriasis treatments. Every trial in the network reported three PASI outcomes at different cutoffs (75, 90, and 100% improvement in PASI score from baseline). The full treatment network (Figure 6.8) was analysed using ML-NMR in Section 6.3, but the analysis focused on only the PASI 75 outcome for computational reasons. The numbers of observed events for the more demanding PASI 90 and PASI 100

outcomes were small, and posed difficulties for estimation. However, from a decision-making perspective the higher PASI outcomes are of greater interest. In this section, we use the ML-NMR model developed in Section 7.1.3 to synthesise all three PASI outcomes simultaneously. By sharing information between the outcomes in a coherent model, we aim to avoid the computational problems associated with modelling the higher PASI outcomes separately, and obtain more precise estimates across all three outcomes.

7.4.1 Methods

We implement the ML-NMR model for ordered categorical outcomes described in Section 7.1.3. There are $M = 4$ categories, and we let $y_{ijk} \in \{1, \ldots, 4\}$ correspond to less than 75% reduction in PASI score (i.e. failure to achieve PASI 75), $\geq$ 75% and $<$ 90% reduction (achieving PASI 75 but not PASI 90), $\geq$ 90% and $<$ 100% reduction (achieving PASI 90 but not PASI 100), and 100% reduction (achieving PASI 100), respectively, for an individual i in study j receiving treatment k. The individual categorical likelihood given in equation (7.10) is

$$L_{ijk|\boldsymbol{x}}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}_{ijk}) = \begin{cases} p_{ijk;1} & \text{if} \quad y_{ijk} = 1 \\ \vdots & \qquad \vdots \\ p_{ijk;M} & \text{if} \quad y_{ijk} = 4 \end{cases}, \quad \text{where} \sum_{m=1}^{4} p_{ijk,m} = 1 \quad \forall i, j, k$$

and we use the probit link function $\Phi(\cdot)$ in equation (7.11) to model the individual category probabilities as

$$p_{ijk,m} = \Phi(\eta_{jk}(\boldsymbol{x}_{ijk}) - c_{m-1}) - \Phi(\eta_{jk}(\boldsymbol{x}_{ijk}) - c_m),$$

for $m = 1, \ldots, 4$, given individual covariate vectors $\boldsymbol{x}_{ijk}$ and linear predictor $\eta_{jk}(\boldsymbol{x}) = \mu_j + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k$ as defined in Section 4.1, where we set $\boldsymbol{\beta}_{2,A} = \mathbf{0}$ and $\gamma_A = 0$. The latent cutpoints c_m are subject to the ordering constraints

$$c_0 < c_1 < c_2 < c_3 < c_4, \tag{7.19}$$

and, we set $c_0 = -\infty$, $c_1 = 0$, and $c_4 = +\infty$. The latent cutpoint c_1 corresponds to achieving PASI 75, c_2 corresponds to achieving PASI 90, and c_3 corresponds to achieving PASI 100.

Priors for the latent cutpoints are most straightforward to specify on the differences between adjacent cutpoints, for example $c_m - c_{m-1} \sim \text{U}(0, u_c)$ for $m = 2, 3$ with an appropriate upper bound u_c (as used by Dias et al. (2011c) with $u_c = 5$), so that the ordering constraints (7.19) are satisfied. When fitting the model in Stan, the ordering constraints (7.19) are guaranteed by

declaring the c_m to be an `ordered` vector, so prior distributions can be placed directly on the cutpoints if desired. In this analysis, we place improper uniform prior distributions $\mathrm{U}(-\infty, +\infty)$ on c_2 and c_3, which are automatically truncated to satisfy the ordering constraints (7.19). We also place $\mathrm{N}(0, 10^2)$ prior distributions on each of the parameters μ_j, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_{2,j}$.

The aggregate-level likelihood for the summary outcome vector $\boldsymbol{y}_{\bullet jk} = (y_{\bullet jk,1}, \ldots, y_{\bullet jk,4})^\mathsf{T}$ from study j on treatment k is given by (7.13):

$$L_{\bullet jk}(\boldsymbol{\xi}; \boldsymbol{y}_{\bullet jk}) \propto \prod_{m=1}^{4} \bar{p}_{jk,m}^{\,y_{\bullet jk}},$$

which is a Multinomial likelihood with average event probability $\bar{p}_{jk,m}$ in each category $m = 1, 2, 3, 4$. The average event probabilities are calculated from the intermediate quantities $\bar{q}_{jk,m}$ as $\bar{p}_{jk,m} = \bar{q}_{jk,m-1} - \bar{q}_{jk,m}$, where $\bar{q}_{jk,m}$ is given in equation (7.14):

$$\bar{q}_{jk,m} = \int_{\mathfrak{X}} \Phi(\eta_{jk}(\boldsymbol{x}) - c_m) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x},$$

avoiding unnecessary repeated integration. We compute these integrals using QMC integration, as described in Section 4.3.3.2, with $\tilde{N} = 5000$ integration points $\tilde{\boldsymbol{x}}_{jk}$ drawn from joint distribution $f_{jk}(\cdot)$ of the covariates on each treatment k in study j, so that

$$\bar{q}_{jk,m} \simeq \tilde{N}^{-1} \sum \Phi(\eta_{jk}(\tilde{\boldsymbol{x}}_{jk}) - c_m).$$

We calculate population-average treatment effects in each study population following the approach in Section 4.4. On the SMD scale (the linear predictor scale), these are equivalent to "plugging-in" mean covariate values from the population of interest, so the population-average treatment effect between treatments a and b in population P is estimated using

$$d_{ab(P)} = \bar{\boldsymbol{x}}_{(P)}^\mathsf{T}(\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}) + \gamma_b - \gamma_a, \tag{7.20}$$

where $\bar{\boldsymbol{x}}_{(P)}$ is the vector of mean covariate values in population P.

To estimate the proportion of individuals achieving each PASI endpoint in a given population, we again use the approach detailed in Section 4.4. We define the quantity

$$h_{km(P)}(\boldsymbol{x}) = \Phi\big(\mu_{(P)} + \boldsymbol{x}^\mathsf{T}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k - c_m\big), \tag{7.21}$$

which represents the probability of an individual in population P with covariate values $\boldsymbol{x}$ achieving the PASI endpoint corresponding to the latent cutpoint c_m (i.e. $m = 1$ corresponds to PASI 75, $m = 2$ to PASI 90, and $m = 3$ to PASI 100).

The population base rate $\mu_{(P)}$ is required, which may be estimated from external data on the population P; here, we produce estimates for each of the observed study populations, and simply use the intercepts μ_j estimated in the model. Since equation (7.21) is non-linear in terms of the model parameters, we must use (4.29) (in AgD studies) or (4.30) (in IPD studies) to evaluate the population average of (7.21). We will use QMC integration to evaluate (4.29) in the AgD studies, so the population-average probabilities are given generally by

$$\bar{h}_{km(P)} = N^{-1} \sum_{i=1}^{N} h_{km(P)}(\boldsymbol{x}_{i(P)}), \tag{7.22}$$

where $\boldsymbol{x}_{i(P)}$ and N are either the N_j covariate values of individuals in the IPD study j, or the $\tilde{N}$ integration points $\tilde{\boldsymbol{x}} \sim f_j(\cdot)$ for AgD study j.

For comparison, we also fit standard AgD fixed and random effects NMA models with no covariate adjustment, using a multinomial likelihood. For all models, we assess convergence using $\hat{R}$ for each parameter, and check that there are no divergent transitions (see Section 5.2.5).

Results 7.4.2

Figure 7.5 and Table 7.5 show the estimated population-average SMD contrasts for each treatment compared to placebo in each study population, calculated using (7.20). Also shown in Figure 7.5 are the contrasts calculated previously in Section 6.3 (see Table 6.7), using only the PASI 75 data. The point estimates from both models agree closely, however uncertainty is slightly reduced for all contrasts when utilising data on all three PASI outcomes compared to just PASI 75—as we would expect. There is little variation in the population-average treatment effects between populations; as before, this is due to the differences in effect modifier distributions between study populations being small when combined with the strength of the interaction. The estimated heterogeneity standard deviation from the RE NMA without covariate adjustment was 0.09 (0.01, 0.26), which is small compared to the magnitude of the relative effects (Table 7.5), and is unchanged from the RE NMA analysis of the PASI 75 outcome only (Section 6.3.2). The DIC values for the FE and RE NMA models without covariate adjustment were 8958.4 and 8956.6 respectively; there is little difference between these models, and we would choose the more parsimonious fixed effect model based on DIC alone. However, despite a lack of evidence for substantial between-study heterogeneity, the ML-NMR model has a much lower DIC of 8815.1. The ML-NMR model allows us to explain both between

and within study variation, resulting in better fit and reduced uncertainty in contrast estimates across the study populations (Table 7.5).

Examining the individual-level treatment effect and EM interaction parameters in Table 7.6 and comparing with the same parameters in the model using only PASI 75 data (Table 6.8), we see that that these are again very similar, with some gains in precision due to using all PASI outcomes rather than only PASI 75. The greatest gains in precision are seen for ustekinumab, which has the least amount of data (both IPD and AgD) in the network amongst all the treatments, and is the sole IL-12/23 blocker so is not sharing information on EM interactions with any other treatments.

The estimated proportion of individuals in each study population achieving each PASI outcome are shown in Figure 7.6, and listed in Tables 7.7 to 7.9. For interpretability and comparability with the previous results from the PASI 75 only model, these are non-exclusive probabilities (e.g. the probability of achieving 75% reduction or greater in PASI score), as opposed to exclusive category probabilities as used to parameterise the model (e.g. the probability of achieving 75% or greater but less than 90% reduction in PASI score). For the predicted proportion of individuals achieving PASI 75, the point estimates are very similar to those based on only the PASI 75 data (Figure 6.10 and Table 6.9), but with reduced uncertainty due to incorporating data from all three PASI outcomes.

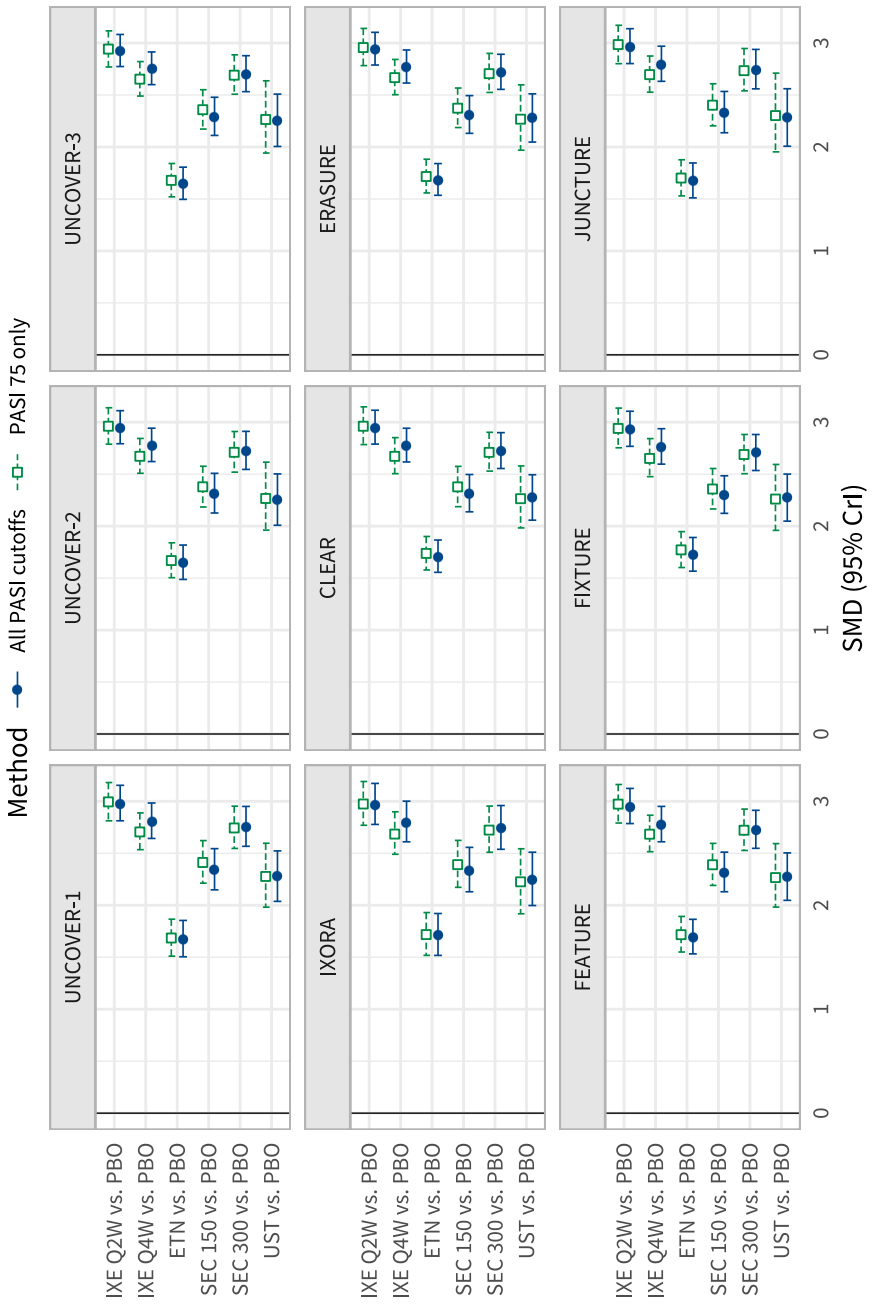


Figure 7.5 Estimated contrasts (standardised mean differences) at the population level for each pair of treatments in each study population, from the ML-NMR model combining information from all PASI endpoints and that using only PASI 75.

Table 7.5 Estimated basic SMD contrasts and 95% Credible Intervals for each treatment compared to placebo, plus the comparison targeted by the previous MAIC, in each study population using the ML-NMR model and for the RE NMA.

	ML-NMR study population									RE NMA
Contrast	CLEAR	ERASURE	FEATURE	FIXTURE	IXORA	JUNCTURE	UNCOVER-1	UNCOVER-2	UNCOVER-3	Weighted overall
IXE Q2W vs. PBO	2.95 (2.79, 3.11)	2.94 (2.79, 3.10)	2.95 (2.79, 3.12)	2.93 (2.77, 3.11)	2.97 (2.78, 3.17)	2.96 (2.80, 3.14)	2.98 (2.81, 3.15)	2.95 (2.79, 3.11)	2.92 (2.77, 3.08)	2.87 (2.69, 3.07)
IXE Q4W vs. PBO	2.78 (2.62, 2.94)	2.77 (2.62, 2.93)	2.78 (2.61, 2.95)	2.76 (2.60, 2.94)	2.80 (2.61, 3.00)	2.79 (2.63, 2.97)	2.81 (2.64, 2.98)	2.78 (2.62, 2.94)	2.75 (2.60, 2.91)	2.69 (2.49, 2.89)
ETN vs. PBO	1.70 (1.56, 1.87)	1.68 (1.53, 1.84)	1.69 (1.53, 1.86)	1.73 (1.57, 1.89)	1.71 (1.52, 1.92)	1.68 (1.51, 1.85)	1.67 (1.50, 1.85)	1.65 (1.49, 1.82)	1.65 (1.50, 1.81)	1.61 (1.43, 1.81)
SEC 150 vs. PBO	2.31 (2.14, 2.50)	2.31 (2.13, 2.49)	2.31 (2.13, 2.51)	2.30 (2.12, 2.48)	2.33 (2.13, 2.56)	2.33 (2.14, 2.53)	2.34 (2.15, 2.54)	2.31 (2.13, 2.51)	2.29 (2.11, 2.48)	2.19 (2.00, 2.42)
SEC 300 vs. PBO	2.72 (2.55, 2.90)	2.72 (2.55, 2.89)	2.73 (2.55, 2.91)	2.71 (2.54, 2.88)	2.74 (2.54, 2.96)	2.74 (2.56, 2.94)	2.75 (2.57, 2.95)	2.72 (2.55, 2.91)	2.70 (2.53, 2.88)	2.60 (2.42, 2.82)
UST vs. PBO	2.28 (2.06, 2.49)	2.28 (2.05, 2.51)	2.28 (2.05, 2.50)	2.28 (2.05, 2.50)	2.25 (2.00, 2.51)	2.29 (2.01, 2.56)	2.28 (2.04, 2.52)	2.26 (2.01, 2.50)	2.26 (2.01, 2.51)	2.16 (1.92, 2.44)
SEC 300 vs. IXE Q2W	−0.22 (−0.37, −0.06)	−0.22 (−0.37, −0.06)	−0.22 (−0.37, −0.06)	−0.22* (−0.37, −0.06)	−0.22 (−0.37, −0.06)	−0.22 (−0.37, −0.06)	−0.22 (−0.37, −0.06)	−0.22 (−0.37, −0.06)	−0.22 (−0.37, −0.06)	−0.27 (−0.49, −0.03)

* Based on the PASI 75 data only, the MAIC estimate was −0.28 (−0.56, −0.00), and the standard indirect comparison estimate was -0.37 (−0.63, 0.12).

Table 7.6 Estimated interactions for each treatment class and potential effect modifier, and estimated individual-level treatment effects, for the ML-NMR model combining information from all PASI endpoints. All estimates are standardised mean differences versus placebo, with 95% Credible Intervals.

	Treatment class		
	Anti-TNFα	IL-12 and IL-23 blocker	IL-17A blocker
Effect modifier interaction			
Previous systemic use	0.10 (−0.28, 0.48)	−0.01 (−0.69, 0.69)	0.13 (−0.22, 0.48)
Duration of psoriasis, per 10 years	0.17 (0.02, 0.32)	0.12 (−0.08, 0.32)	0.17 (0.03, 0.31)
Body surface area, per 10%	0.04 (−0.06, 0.15)	0.05 (−0.08, 0.20)	0.00 (−0.09, 0.11)
Weight, per 10 kg	−0.09 (−0.17, −0.01)	−0.04 (−0.14, 0.07)	−0.05 (−0.12, 0.02)
Psoriatic arthritis	0.01 (−0.45, 0.47)	0.31 (−0.36, 1.00)	0.27 (−0.14, 0.70)
Reference individual treatment effect			
IXE Q2W			2.81 (2.56, 3.07)
IXE Q4W			2.63 (2.38, 2.90)
ETN	1.61 (1.35, 1.88)		
SEC 150			2.17 (1.91, 2.45)
SEC 300			2.58 (2.33, 2.85)
UST		2.21 (1.63, 2.75)	

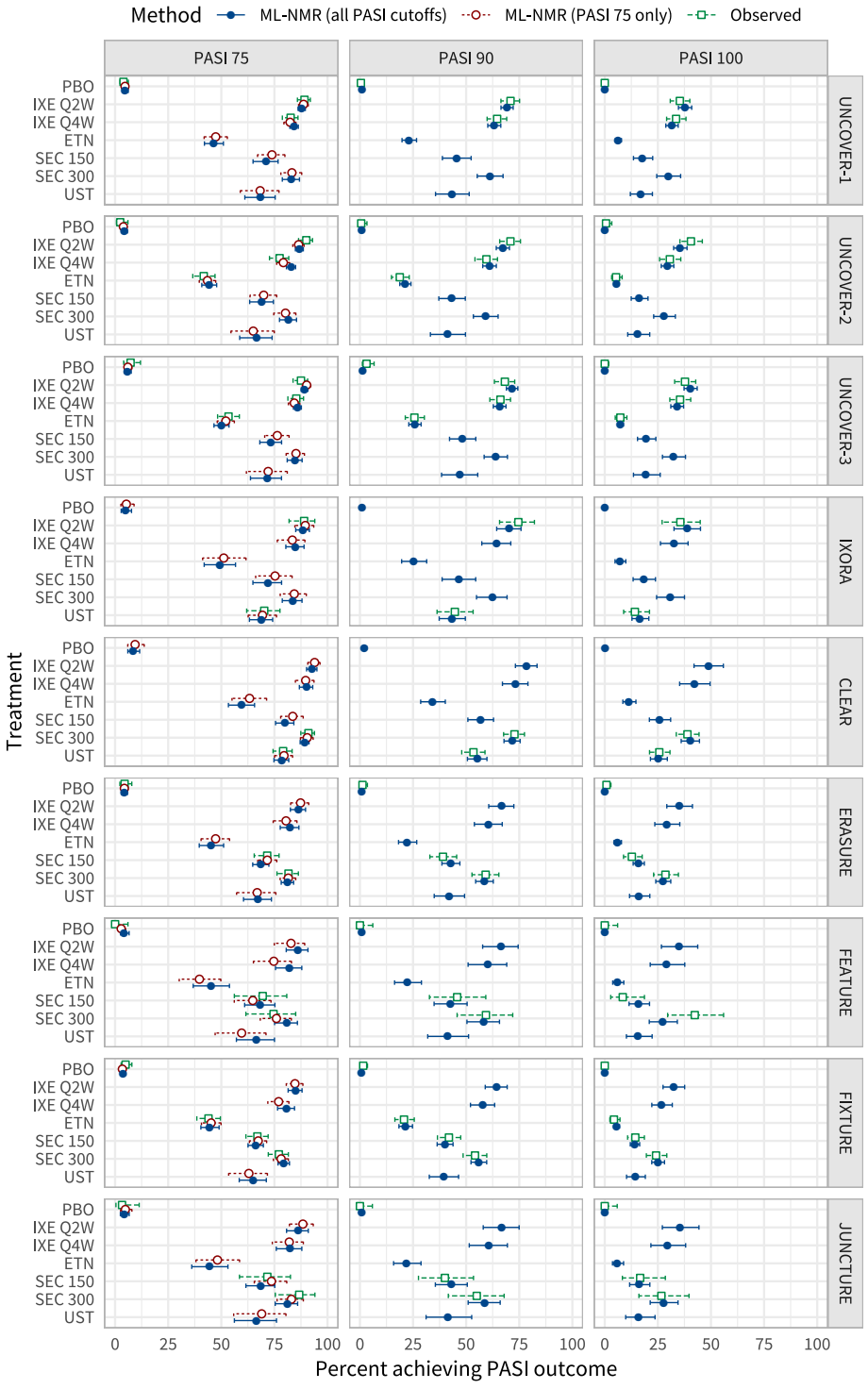


Figure 7.6 Estimated proportion of individuals achieving each PASI endpoint on each treatment, in each study population, using the ML-NMR model.

Table 7.7 Estimated proportion of individuals achieving PASI 75 on each treatment in each study population, along with 95% Credible Intervals, using ML-NMR combining information from all PASI endpoints.

	Treatment						
Study population	Placebo	Ixekizumab Q2W	Ixekizumab Q4W	Etanercept	Secukinumab 150 mg	Secukinumab 300 mg	Ustekinumab
CLEAR	8.44 (5.99, 11.56)	92.75 (90.13, 95.01)	90.16 (86.74, 93.04)	59.57 (53.35, 65.65)	80.01 (75.56, 84.15)	89.33 (87.12, 91.30)	78.47 (74.79, 81.74)
ERASURE	4.32 (2.99, 5.83)	86.37 (82.52, 89.65)	82.42 (77.73, 86.52)	45.26 (39.62, 51.09)	68.61 (64.79, 72.38)	81.16 (78.20, 84.02)	67.23 (60.46, 73.59)
FEATURE	4.27 (2.54, 6.60)	86.10 (80.56, 90.78)	82.12 (75.53, 87.88)	45.22 (36.77, 53.83)	68.29 (60.99, 75.22)	80.86 (75.10, 85.82)	66.51 (57.10, 75.04)
FIXTURE	3.74 (2.64, 5.07)	85.00 (81.46, 88.00)	80.78 (76.49, 84.46)	44.55 (40.39, 48.95)	66.25 (62.58, 69.78)	79.42 (76.70, 82.08)	65.05 (58.44, 71.18)
IXORA	4.94 (2.86, 7.73)	88.47 (85.04, 91.45)	84.92 (80.38, 88.95)	49.40 (41.94, 56.69)	72.03 (64.87, 78.37)	83.71 (78.70, 87.99)	68.87 (63.23, 74.17)
JUNCTURE	4.30 (2.49, 6.65)	86.28 (80.71, 90.95)	82.36 (75.89, 87.90)	44.51 (36.09, 53.13)	68.66 (61.58, 75.17)	81.11 (75.37, 85.93)	66.56 (56.14, 75.99)
UNCOVER-1	4.71 (3.48, 6.15)	87.91 (86.19, 89.55)	84.25 (82.27, 86.22)	46.46 (42.05, 50.99)	71.04 (64.94, 76.73)	82.97 (78.72, 86.79)	68.49 (61.08, 75.36)
UNCOVER-2	4.41 (3.21, 5.79)	86.80 (84.87, 88.60)	82.91 (80.66, 84.96)	44.34 (40.81, 47.82)	69.10 (63.35, 74.52)	81.56 (77.39, 85.41)	66.66 (58.66, 73.91)
UNCOVER-3	5.90 (4.44, 7.57)	89.25 (87.65, 90.75)	85.87 (83.95, 87.68)	50.07 (46.55, 53.62)	73.37 (68.13, 78.37)	84.67 (81.02, 88.03)	71.65 (63.72, 78.33)

Table 7.8 Estimated proportion of individuals achieving PASI 90 on each treatment in each study population, along with 95% Credible Intervals, using ML-NMR combining information from all PASI endpoints.

	Treatment						
Study population	Placebo	Ixekizumab Q2W	Ixekizumab Q4W	Etanercept	Secukinumab 150 mg	Secukinumab 300 mg	Ustekinumab
CLEAR	2.11 (1.32, 3.19)	78.50 (73.20, 83.38)	73.33 (67.08, 79.08)	34.12 (28.60, 40.19)	56.85 (50.81, 62.89)	71.71 (67.83, 75.41)	55.34 (50.61, 59.79)
ERASURE	0.89 (0.55, 1.31)	66.72 (60.64, 72.45)	60.51 (53.88, 66.99)	22.22 (18.18, 26.74)	42.77 (38.60, 47.05)	58.60 (54.48, 62.77)	42.01 (34.97, 49.26)
FEATURE	0.88 (0.45, 1.53)	66.40 (57.78, 74.49)	60.21 (50.93, 69.11)	22.28 (16.32, 28.95)	42.54 (34.91, 50.45)	58.30 (50.38, 65.69)	41.32 (31.92, 51.13)
FIXTURE	0.74 (0.47, 1.10)	64.31 (58.99, 69.32)	57.93 (52.04, 63.42)	21.44 (18.37, 24.74)	40.06 (36.36, 43.83)	55.97 (52.29, 59.67)	39.53 (32.66, 46.44)
IXORA	1.05 (0.51, 1.86)	70.30 (64.37, 75.83)	64.31 (57.37, 71.01)	25.32 (19.64, 31.43)	46.69 (38.70, 54.50)	62.43 (54.89, 69.26)	43.40 (37.31, 49.61)
JUNCTURE	0.89 (0.44, 1.58)	66.78 (58.04, 74.94)	60.63 (51.47, 69.41)	21.83 (15.86, 28.75)	43.02 (35.51, 50.53)	58.74 (50.96, 66.02)	41.51 (31.15, 52.61)
UNCOVER-1	1.00 (0.67, 1.42)	69.30 (66.40, 72.15)	63.26 (60.19, 66.31)	23.14 (19.82, 26.65)	45.59 (38.83, 52.35)	61.34 (55.22, 67.38)	43.38 (35.59, 51.43)
UNCOVER-2	0.92 (0.59, 1.32)	67.29 (64.17, 70.38)	61.08 (57.81, 64.20)	21.33 (18.76, 24.01)	43.26 (37.15, 49.60)	59.13 (53.40, 65.04)	41.30 (33.18, 49.69)
UNCOVER-3	1.33 (0.89, 1.87)	71.69 (68.95, 74.34)	65.83 (62.80, 68.75)	25.84 (23.02, 28.81)	48.32 (42.18, 54.55)	63.95 (58.47, 69.43)	47.00 (38.43, 55.46)

Table 7.9 Estimated proportion of individuals achieving PASI 100 on each treatment in each study population, along with 95% Credible Intervals, using ML-NMR combining information from all PASI endpoints.

	Treatment						
Study population	Placebo	Ixekizumab Q2W	Ixekizumab Q4W	Etanercept	Secukinumab 150 mg	Secukinumab 300 mg	Ustekinumab
CLEAR	0.23 (0.12, 0.39)	48.84 (42.01, 55.83)	42.26 (35.25, 49.59)	11.35 (8.53, 14.67)	25.88 (21.01, 31.01)	40.28 (35.97, 44.57)	25.29 (21.51, 29.45)
ERASURE	0.08 (0.04, 0.13)	35.11 (29.21, 41.27)	29.19 (23.56, 35.32)	5.97 (4.40, 7.86)	15.88 (13.37, 18.62)	27.45 (24.02, 31.07)	16.02 (11.67, 21.21)
FEATURE	0.08 (0.03, 0.15)	34.94 (26.68, 43.76)	29.06 (21.40, 37.66)	6.03 (3.75, 8.95)	15.84 (11.43, 21.07)	27.33 (20.95, 34.16)	15.64 (10.20, 22.29)
FIXTURE	0.06 (0.03, 0.10)	32.51 (27.44, 37.61)	26.76 (22.15, 31.78)	5.55 (4.37, 6.94)	14.15 (12.02, 16.43)	25.10 (22.13, 28.12)	14.41 (10.26, 19.09)
IXORA	0.09 (0.04, 0.19)	38.80 (32.60, 45.06)	32.61 (26.23, 39.22)	7.17 (4.84, 9.91)	18.38 (13.35, 23.87)	30.83 (24.41, 37.53)	16.56 (12.80, 20.75)
JUNCTURE	0.08 (0.03, 0.16)	35.40 (27.11, 44.34)	29.50 (21.72, 38.08)	5.89 (3.64, 8.85)	16.17 (11.71, 21.22)	27.76 (21.40, 34.48)	15.85 (9.86, 23.66)
UNCOVER-1	0.09 (0.05, 0.14)	37.75 (34.69, 40.93)	31.62 (28.74, 34.56)	6.33 (4.98, 7.85)	17.71 (13.52, 22.57)	29.91 (24.48, 35.71)	16.87 (12.04, 22.43)
UNCOVER-2	0.08 (0.04, 0.13)	35.50 (32.43, 38.73)	29.51 (26.60, 32.49)	5.53 (4.50, 6.63)	16.15 (12.46, 20.35)	27.84 (23.03, 33.23)	15.51 (10.82, 21.16)
UNCOVER-3	0.13 (0.07, 0.20)	40.38 (37.33, 43.51)	34.10 (31.10, 37.10)	7.39 (6.12, 8.79)	19.52 (15.42, 24.04)	32.31 (27.17, 38.00)	19.27 (13.52, 26.11)

7.4.3 Conclusions

In this section, we used the ML-NMR model derived in Section 7.1.3 to synthesise ordered categorical outcomes, namely 75%, 90%, and 100% cutpoints in PASI score improvements from baseline. Previously in Chapter 6, we synthesised only the PASI 75 data: whilst the higher PASI outcomes are more interesting from a decision-making perspective, the low numbers of observed events posed difficulties for estimation in stand-alone analyses. By considering the PASI 75, 90, and 100 outcomes together as an ordered categorical outcome, we share information across the three cutpoints which alleviates the estimation issues with the higher PASI outcomes and leads to more precise estimates. Estimated average treatment effects and the proportion of individuals achieving each PASI endpoint are available in any of the included study populations or in an external target population for decision-making (given a covariate distribution and, for absolute proportions, a baseline event rate).

Synthesising the three PASI outcomes together in this manner does require additional assumptions: the outcomes are assumed to be defined by cutpoints in some underlying continuous variable, which further implies that the treatments and covariates affect each outcome via this underlying continuous variable and that their effects are constant across cutpoints. These assumptions are likely justified, given that the outcomes are indeed specified as cutpoints in percent improvement in PASI score. Additionally, the probit model assumes that the treatments and covariates act only on the location of the distribution of the underlying continuous variable, and not on its scale (i.e. the standard deviation used in standardising outcomes to SMDs is the same across treatments and covariate values). Furthermore, in this analysis we have used fixed latent cutpoints c_m, which are assumed to be the same in every trial. Variations in outcome definition between trials could affect this assumption, and instead "random" cutpoints c_{mj} which are exchangeable at level m between trials j could be fitted (Dias et al. 2011c). However, we have not attempted to fit a random cutpoint model here.

7.5 Discussion

In this chapter, we have extended the ML-NMR framework to handle general likelihoods and expand the range of models which can be fitted. As in Chapter 4, we began with a fully-specified individual-level model, and considered how to aggregate this model to apply to summary data. In Chapter 4, we proceeded by determining explicitly the form of the aggregate likelihood, rely-

ing on well-known results such as the sum of Normally distributed outcomes being Normally distributed. However, in Section 7.1 we instead proceeded by considering the likelihood contributions from each level of the model. Individuals in studies with IPD have individual conditional likelihood contributions defined by the individual-level model, conditional on their covariate values. Integrating the individual conditional likelihood function over the covariate distribution in a study results in an individual marginal likelihood function, which is then used in one of two ways, depending on the data available, with different levels of generality.

Firstly, in settings where the aggregate data consist of individual outcomes but only summary covariate information (such as survival data reconstructed from Kaplan-Meier curves), the aggregate part of the model is fitted directly using the individual marginal likelihood contributions. In this case, the method is fully general: individual conditional likelihood functions of any form can be integrated numerically to evaluate the individual marginal likelihood function using one of the approaches described in Section 4.3.3.

Secondly, we have settings where the aggregate data consist of summary outcomes and summary covariate information. In this case, the individual marginal likelihood contributions are multiplied together to obtain the aggregate marginal likelihood contributions for the summary outcomes. Evaluation of the aggregate marginal likelihood contributions requires that these can be expressed in terms of the summary outcomes, which is only straightforward for discrete outcomes. This would appear to limit the generality of the approach for continuous outcomes; however, the aggregate-level likelihood has a known closed form for many continuous individual-level likelihoods common in practice (Section 4.2).

The survival analysis example in Section 7.3 focused on a scenario where event/censoring times were available from each individual in the aggregate studies, for example reconstructed from Kaplan-Meier plots (Guyot et al. 2012). If individual event/censoring times are not available, but instead only log hazard ratios are reported (or can be recovered, see Parmar et al. 1998), the simplest solution is to synthesise the log HRs using a Normal likelihood, and if the individual-level linear predictor $\eta_{jk}(\cdot)$ is on the log HR scale no numerical integration is needed and mean covariate values can simply be "plugged in". For example, for the log HR of treatment b vs. treatment a in study j the likelihood would be

$$\mathrm{N}\Big(\eta_{jb}(\bar{\boldsymbol{x}}_j) - \eta_{ja}(\bar{\boldsymbol{x}}_j), s^2_{jab}\Big),$$

where s^2_{jab} is the variance of the log HR (given as data) and $\bar{\boldsymbol{x}}_j$ is the vector

of mean covariate values in study j. (Studies with three or more arms would require the correlations between log HRs to be accounted for in the likelihood, as in Section 1.2.8.) It may also be possible to instead synthesise reported summary outcomes such as hazard ratios or median survival times by considering the aggregate marginal likelihood contributions for these data (Section 7.1). This remains an area for further research (see Section 9.2.3).

We have only considered adjusting for covariates measured at baseline: time-varying covariates were not considered since it is likely that, in the aggregate studies, summary covariate information is available only available at baseline and not throughout follow-up. The inclusion of time-varying covariates in a survival model is often an attempt to correct for observed non-proportionality (e.g. failure of the proportional hazards assumption). However, as noted by Therneau and Grambsch (2000, Section 6.6), such problems may by symptomatic of other issues such as omitted covariates, an incorrect functional form for a covariate, or using an inappropriate model form (e.g. a proportional hazards model when an accelerated failure time model would be more appropriate). Notably, the solutions for these issues can be dealt with within the ML-NMR framework we have described, without requiring further information on time-varying covariates.

Although our survival analysis example focused on parametric proportional hazards models, it is straightforward to fit other survival models such as accelerated failure time models, or more flexible models on the baseline hazard function such as the Royston-Parmar model (Freeman and Carpenter 2017; Royston and Parmar 2002) or fractional polynomials (Jansen 2011). Piecewise exponential models with a Poisson likelihood have been shown to be equivalent to the partial likelihood for a semi-parametric Cox model (Cox 1972) when the number of intervals (within which the hazard is modelled as constant) is equal to the number of events (Crowther et al. 2012). This formulation admits a Bayesian approach to the semi-parametric Cox model, and it should be possible to apply the marginal likelihood approach to fit piecewise exponential models in the ML-NMR framework. However, implementation of the Cox model in a Bayesian framework in this manner is often very computationally intensive (Crowther et al. 2012), and the addition of numerical integration will only exacerbate this. A wide range of other semi- and non-parametric prior distributions for the baseline hazard function have been proposed (for an overview, see Ibrahim et al. 2001, Chapter 3; Müller et al. 2015, Chapter 6), which could also be investigated. The Stan code (see Appendix A.3.2) that we have developed is modular, and all that is required to fit a range of alternative

models in the ML-NMR framework is to specify the form of the survival and hazard functions for the individual conditional likelihood (i.e. conditional on the covariates). Once these have been specified, the numerical integration step to obtain the individual marginal likelihood remains the same, and is automatically implemented in the Stan code.

Fitting models where part of the model has no explicit closed form presents new challenges, in particular for model comparison. Whilst the relative fit of data points under a given model can always be investigated using the deviance (-2 times the log likelihood), computation of standard model comparison statistics such as the DIC (Spiegelhalter et al. 2002) is complicated when the aggregate likelihood has no closed form. In Section 7.2, we proposed instead to use approximate leave-one-out cross validation for model comparison, as described by Vehtari et al. (2016). This approach requires only the posterior samples of the log likelihood contributions, and so can be computed without a closed form aggregate likelihood. We applied this approach successfully in Section 7.3 to select an appropriate survival model.

Chapter 8

Simulation study

Despite their increasing popularity, population adjustment methods such as MAIC and STC have yet to be subjected to extensive simulation studies. In this chapter, we undertake a thorough simulation study designed to test the performance of ML-NMR alongside MAIC and STC in a wide range of scenarios, in particular under failure of assumptions. Section 8.1 sets out the simulation study plan, and Section 8.2 presents and discusses the results.

Simulation study plan 8.1

This simulation study plan follows the ADEMP framework (Morris et al. 2019), which breaks down the simulation study into five key elements:

- **A**ims
- **D**ata-generating mechanisms
- **E**stimands
- **M**ethods
- **P**erformance measures

The following sections are devoted to describing each of these elements in turn.

Aims 8.1.1

The simulation study aims to assess the performance of ML-NMR against current population adjustment methods (MAIC and STC), in a range of ideal and non-ideal scenarios under various failures of assumptions. The primary

concerns are bias and efficiency of the estimators. We shall also consider other frequentist properties such as coverage.

8.1.2 Data-generating mechanisms

For this simulation study, we will consider a binary outcome, generated under a logit (log odds ratio) model. The basic data structure involves three treatments (A, B, C) investigated in two studies, AB and AC. We will simulate two continuous covariates which modify the effect of treatments B and C. Individual patient data (IPD) are available for the AB study as a binary outcome and covariate information for each individual. Only aggregate data (AgD) are available for the AC study, given as an overall event count and summary covariate information. For the AC study, individual outcomes will be simulated and then aggregated.

The underlying model is of the form

$$\begin{aligned} y_{ijk} &\sim \text{Bern}(\theta_{ijk}) \\ \text{logit}(\theta_{ijk}) &= \mu_j + q\left(\boldsymbol{x}_{ijk}^{\mathsf{T}} - \boldsymbol{m}_j^{\mathsf{T}}\right)\boldsymbol{\beta}_k + \gamma_k \\ \boldsymbol{x}_{ijk} &\sim \phi_j \end{aligned}$$

where outcomes y_{ijk} for individual i in study j receiving treatment k are generated from a Bernoulli distribution with individual event probability θ_{ijk}. A logit link function is used to define an outcome model on the linear predictor scale, with study intercept μ_j, covariate vector $\boldsymbol{x}_{ijk}$ and coefficients $\boldsymbol{\beta}_k$, and treatment effects γ_k. The covariates are generated following the joint distribution ϕ_j. The function $q(\cdot)$ defines a (potentially non-linear) relationship between the covariates and outcome. Covariates are centred in the regression model against the mean in each study $\boldsymbol{m}_j$, so that the model coefficients are more easily interpreted as log odds and log odds ratios at the mean.

We set $\gamma_A = 0$, $\gamma_B = -2$, $\gamma_C = -1.5$, $\mu_{AB} = 1$, $\mu_{AC} = 1.5$. The values of $\boldsymbol{\beta}_k$ will be set depending on the scenarios set out below, but always $\boldsymbol{\beta}_A = \mathbf{0}$ because A is the reference treatment.

In the AB trial, the two covariates will be generated with means $m_{X_1(AB)} = 1$ and $m_{X_2(AB)} = 0.5$, and standard deviations $\sigma_{X_1(AB)} = 0.5$ and $\sigma_{X_2(AB)} = 0.1$ (as either a Normal or Gamma covariate, see g and h below). The distributions of the covariates in the AC trial are set to achieve the required overlap (see e below).

We shall consider varying the following parameters (where the reference levels for each parameter are **bold and underlined** below):

a. Sample size N in AB and AC trials (100, **500**, 1000). We will use 1:1 randomisation within each study.

b. Strength of effect modification. Modifying the treatment effect log odds ratio by **0.1**, 0.5 per AB covariate SD, i.e. $\boldsymbol{\beta}_k = 0.1\sigma_{X(AB)}$ or $0.5\sigma_{X(AB)}$. This corresponds to roughly ±10% and ±50% modification of treatment effects within ±2 SD of the mean covariate value $\boldsymbol{m}_{X(AB)}$ in the AB study.

c. Shared effect modification. Effect modifier coefficients are **shared** ($\boldsymbol{\beta}_B = \boldsymbol{\beta}_C$) or not between treatments B and C.

d. Strength of correlation between covariates in each study ($\rho_{(AB)} = \rho_{(AC)} =$ 0, **0.25**, 0.5).

e. Between-study overlap. Full overlap (AC contained entirely within AB), and **50%**, 100% of AC population outside of AB. Set using a proxy parameter κ (see below).

f. Covariate-outcome relationship. **Linear**, non-linear beyond the range of the AB study.

g. Distribution of covariates in AB. Consider **Normal covariates** and Gamma covariates.

h. Distribution of covariates in AC. Consider **Normal covariates** and Gamma covariates.

i. Correlation structures. **Same in both studies** ($\rho_{(AB)} = \rho_{(AC)}$), different in each study ($\rho_{(AB)} \neq \rho_{(AC)}$).

For the purposes of computation time and reporting, we will not consider a full factorial design of all possible data-generating mechanisms. Instead, each parameter will be varied independently. However, we will consider a factorial examination of scenarios e and f (investigating extrapolation when it is not valid), and of g, h, and i (investigating assuming the same form of covariate distribution and correlation structure between studies when not valid).

For scenario e, we wish to vary the overlap between the study populations, which we define as the proportion of the AC joint density $\phi_{(AC)}$ contained within the 95% highest density region (HDR) of the AB joint density $\phi_{(AB)}$. Thus the overlap is given by the integral

$$\int_{\mathfrak{X}} \mathbb{I}\big(\boldsymbol{x} \in R_{(AB),0.05}\big)\phi_{(AC)}(\boldsymbol{x})\, d\boldsymbol{x}, \tag{8.1}$$

where the $(1-\alpha)$% HDR is the region $R_{(AB),\alpha} = \{\boldsymbol{x} : \phi_{(AB)}(\boldsymbol{x}) \geq c_{(AB),\alpha}\}$, with $c_{(AB),\alpha}$ the largest constant satisfying $\mathbb{P}(\boldsymbol{x} \in R_{(AB),\alpha}) \geq 1-\alpha$. Clearly the exact relationship between the parameters of the joint distribution in each study and the true overlap is highly complex, and so instead we use a proxy parameter κ. Approximately, $\kappa = 0$ corresponds to no overlap, $\kappa = 0.5$ to 50% overlap, and $\kappa = 1$ to full overlap. Using κ, we then define the mean and standard deviation of each covariate in the AC population as

$$\begin{aligned} m_{X(AC)} &= \left(1.1 + (1-\kappa)^2\right) m_{X(AB)} \\ \sigma_{X(AC)} &= 0.75\sigma_{X(AB)}. \end{aligned}$$

We later calculate the true overlap in several scenarios using numerical integration to evaluate the integral (8.1). The HDR is calculated numerically using the density quantile method, outlined by Hyndman (1996); for numerical integration over two dimensions we use the R package `cubature` (available from `https://cran.r-project.org/package=cubature`).

For scenario f, the non-linear relationship will be specified with a sigmoid function, parametrised so that $y \approx x$ within ± 2 SD of the mean covariate value in the AB study, but then attenuating outside this range (Figure 8.1):

$$q(x) = 8\sigma_{X(AB)} \left(\frac{1}{1 + \exp(-(x - m_{X(AB)})/2\sigma_{X(AB)})} - \frac{1}{2} \right) + 1. \tag{8.2}$$

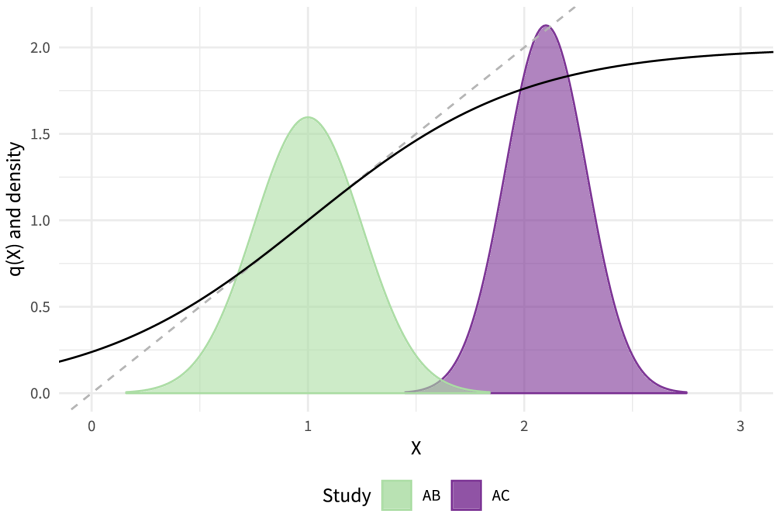


Figure 8.1 The function $q(\cdot)$ is approximately linear within the range of the AB study, but attenuates outside this range.

Estimands 8.1.3

The estimands of interest will be population-average relative effects d_{AB}, d_{AC}, d_{BC} with target population:

1. Represented by the AC trial population,
2. Represented by the AB trial population.

Methods 8.1.4

The following methods will be compared:

1. ML-NMR,
2. STC,
3. MAIC,
4. Standard indirect comparison.

For each of methods 1–3, we will consider models with a full set of effect modifiers or missing one effect modifier. ML-NMR will be carried out using Quasi-Monte Carlo numerical integration as described in Chapter 4, with a Gaussian copula to account for the correlations between covariates. MAIC and STC are described in Section 2.2.1 and Section 2.2.2 respectively. Standard errors for MAIC will be estimating using bootstrapping. The standard indirect comparison (without population adjustment) follows the Bucher method (Bucher et al. 1997).

Performance measures 8.1.5

The following performance measures will be computed:

- Bias = $\frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} (\hat{d}_i - d)$,
- Empirical standard error = $\sqrt{\frac{1}{N_{\text{sim}}-1} \sum_{i=1}^{N_{\text{sim}}} (\hat{d}_i - \bar{d})^2}$,
- Model standard error = $\sqrt{\frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \widehat{\text{var}}(\hat{d}_i)}$,
- Coverage probability = $\frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} \mathbb{I}\left(\hat{d}_{\text{lower},i} \leq d \leq \hat{d}_{\text{upper},i}\right)$.

where $\hat{d}_i$ are N_{sim} repeated estimates of some truth d, with sample mean $\bar{d} = \sum_{i=1}^{N_{\text{sim}}} \hat{d}_i$, model estimated variance $\widehat{\text{var}}(\hat{d}_i)$, and lower and upper confidence/credible interval limits $\hat{d}_{\text{lower},i}$ and $\hat{d}_{\text{upper},i}$. Bias is the difference

between the expected value of an estimator and the truth, estimated as the average difference between the repeated estimates and the truth. An unbiased estimator (bias of zero) is desirable. The empirical standard error is the true variability of the estimator, estimated as the observed standard error of the repeated estimates. The model standard error is the average standard error reported by a method over all repetitions (taken on the variance scale). We thus desire both that the empirical standard error is small (the estimator is precise) and that the empirical standard error is well-estimated by the model standard error (so that uncertainty is appropriately quantified). The coverage probability is the probability that the confidence or credible intervals contain the true value, estimated as the proportion of repetitions with intervals that include the true value, which we wish to be at the nominal level (here 95%).

Since the primary concerns are bias and efficiency, we will evaluate sufficient sample size based on bias and empirical standard error. The Monte Carlo standard errors (MCSE) for these quantities are

$$\text{MCSE}_{\text{Bias}} = \sqrt{\frac{\text{var}(\hat{d})}{N_{\text{sim}}}}$$

$$\text{MCSE}_{\text{EmpSE}} = \sqrt{\frac{\text{var}(\hat{d})}{2(N_{\text{sim}} - 1)}}$$

We will begin by running $N_{\text{sim}} = 1000$ simulations. At this point, we will evaluate the Monte Carlo standard errors to determine whether these are sufficiently low with respect to the magnitude of the treatment effects, and consider running further simulations.

8.2 Simulation results

The simulation study was run on BlueCrystal Phase 3, the University of Bristol supercomputer (see Appendix C for details of the computing environment). The R package `simsalapar` was utilised to simplify the mechanics of the study, including parallelisation, error handling, and saving of results (Hofert and Mächler 2016). Each scenario was replicated 2000 times to achieve Monte Carlo standard errors (on the log odds ratio scale) below 0.035 for bias and 0.025 for empirical standard error, for all methods (other than a small number of scenarios for which MAIC was highly unstable). By running the simulations in parallel, the total 4-core slave time of over 1800 hours (11 weeks) was reduced to an actual run time of around 10 days.

Due to the size of simulation study, tabulated summaries of results can be found in Appendix B.

Scenario a: sample size 8.2.1

In scenario a, the sample sizes of the two studies were varied between 100, 500, and 1000. Figure 8.2a shows the bias in the population-average contrast estimates for each method, adjusting for all effect modifiers, along with 95% Monte Carlo confidence intervals. To aid comparison between methods, the points are coloured by contrast, with lighter shades for the AB population and darker for the AC population. Figure 8.2b shows the corresponding empirical and model standard errors. Coverage zip plots (Morris et al. 2019) for the $d_{BC(AC)}$ contrast estimate are shown in Figure 8.3, centred around zero bias. These display the 95% confidence or credible intervals for each repetition, centile-ranked on the vertical axis by

$$\left| \frac{(\hat{d}_i - d)}{\sqrt{\widehat{\text{var}}(\hat{d}_i)}} \right|, \tag{8.3}$$

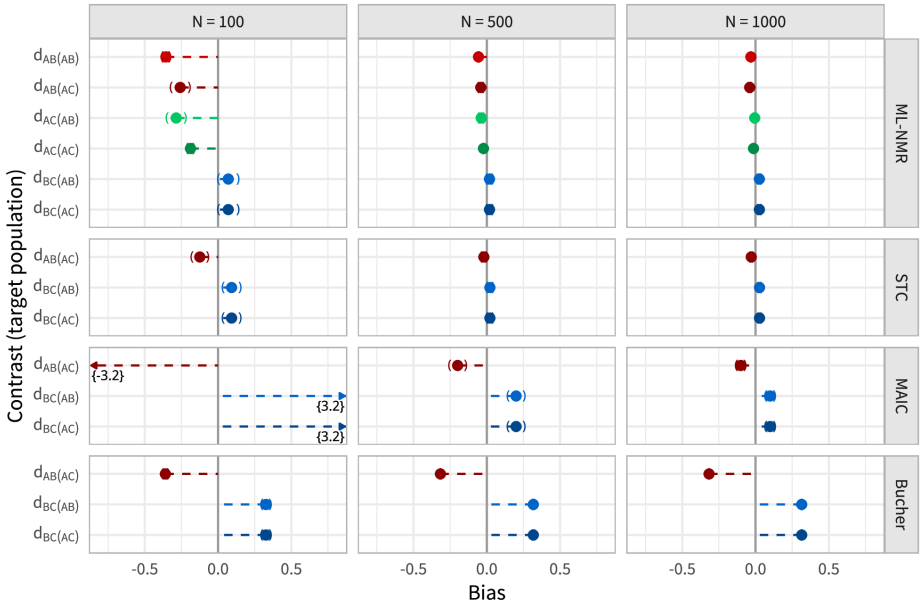
and coloured according to whether the interval includes the truth ("coverers", in green, at the bottom) or not ("non-coverers", in purple, at the top). The location of the colour change on the vertical axis is thus the estimated coverage (i.e. the proportion of coverers). If the intervals have nominal coverage, the colour change will occur at the 95th centile (i.e. at the nominal level). The horizontal dashed lines give the 95% Monte Carlo confidence interval for the coverage. The zip plot will appear symmetrical around zero (with a Y or zip shape) if the estimates are unbiased, otherwise the zip plot will be skewed to one side. Table B.1 provides the results in tabular format.

We see that ML-NMR and STC are comparable in terms of both bias and standard error, largely eliminating the bias that the Bucher method (standard indirect comparison) incurs, with similar empirical standard errors that are well-estimated by the model standard errors. For the smallest sample size (100), the model standard errors slightly underestimate the empirical standard errors. Some small bias remains in each of the estimates for ML-NMR and STC, and is most pronounced at the smallest sample size; this is likely due to the small sample bias inherent to logistic regression and the low average number of events on treatment B (13.6 from 50 individuals) (Vittinghoff and McCulloch 2007). As expected, ML-NMR produces more precise estimates (i.e. with lower standard error) for contrasts between treatments where there is direct evidence than those based on indirect evidence ($d_{AB(AB)}$ and $d_{AC(AC)}$ are more precisely estimated than the remaining contrasts). ML-NMR and STC also perform similarly in terms of coverage, achieving nominal coverage for sample sizes 500 and 1000. Both methods display slight under-coverage

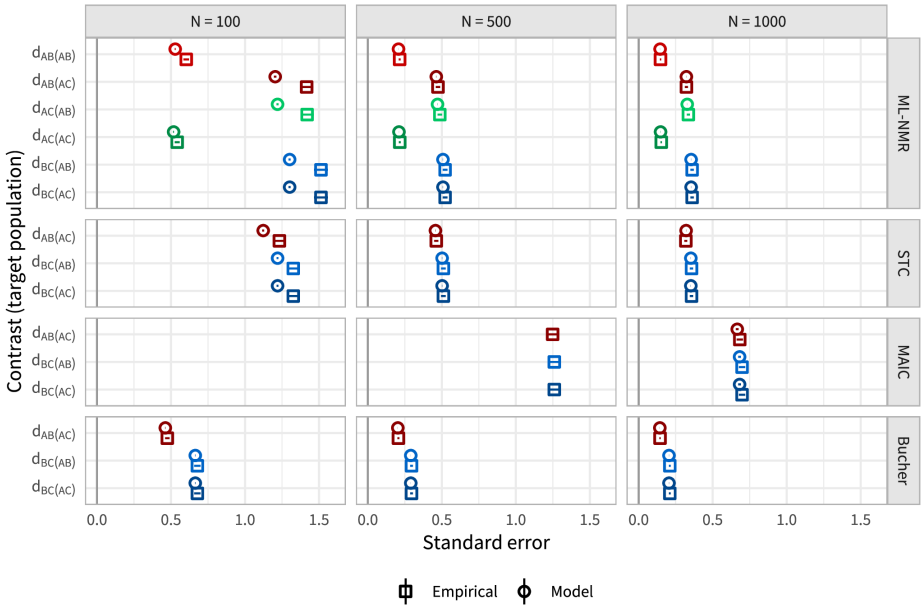
at the smallest sample size (100), 90.6% (89.3, 91.8) for ML-NMR and 93.6% (92.6, 94.7) for STC, perhaps due to increased small-sample bias. The standard indirect comparison is always below nominal coverage, and coverage drops off severely with increasing sample size as a biased estimator is more precisely estimated (note the rightward skew to the zip, which agrees with Figure 8.2a).

MAIC does not perform as well as ML-NMR or STC in this scenario. The reference level of between-study overlap is set at 50%, but this means that any reweighting scheme such as MAIC cannot hope to eliminate the bias since extrapolation is not possible. As a result, MAIC provides a lesser bias reduction than ML-NMR and STC for sample sizes 500 and 100, and for the smallest sample size actually substantially increases the bias compared to a standard indirect comparison. Empirical standard errors for MAIC are larger than for ML-NMR and STC, and the MAIC model standard error is only a good estimate at the largest sample size. For sample size 500, the MAIC model standard errors (derived through bootstrapping) are extremely unstable, as the weights are highly dependent on a small number of individuals, and were only successfully obtained 47% of the time. For sample size 100, estimation failed entirely 23% of the time, and no attempts to obtain model standard errors were successful. Coverage for sample size 500 is at the nominal value, despite remaining bias and unstable standard errors. For sample size 1000, coverage of 90.7% (89.4, 91.9) is below the nominal value, as the standard errors have stabilised but the estimator remains biased.

When one of the two effect modifiers is not adjusted for, none of the methods are able to eliminate bias from the estimates (Figure 8.4a, also Table B.2). ML-NMR and STC show slightly reduced standard errors compared to the models with the full set of effect modifiers (Figure 8.4b). Again, the empirical standard errors are well-estimated by the model standard errors for the larger sample sizes, with a slight underestimation for the smallest sample size. MAIC has markedly reduced standard errors compared to adjustment for both effect modifiers, and the bootstrap model standard errors are now stable for the larger sample sizes. This is because the overlap between studies is necessarily lower as the number of covariates increases. However, MAIC is still unstable for the smallest sample size. Coverage of all three population adjustment methods begins to drop from the nominal level as sample size increases, down to around 92% for sample size 1000 (Figure 8.5). This is due to the bias remaining in the estimates, as further evidenced by the rightward skew to the zip plots. However, coverage is still well above that of the standard indirect comparison, as bias is reduced.



(a)



(b)

Figure 8.2 Bias (a) and standard errors (b) for the population-average contrast estimates for scenario a, along with 95% Monte Carlo confidence intervals. Each method (other than Bucher) adjusts for the full set of effect modifiers. Sample size is varied between 100, 500, and 1000. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

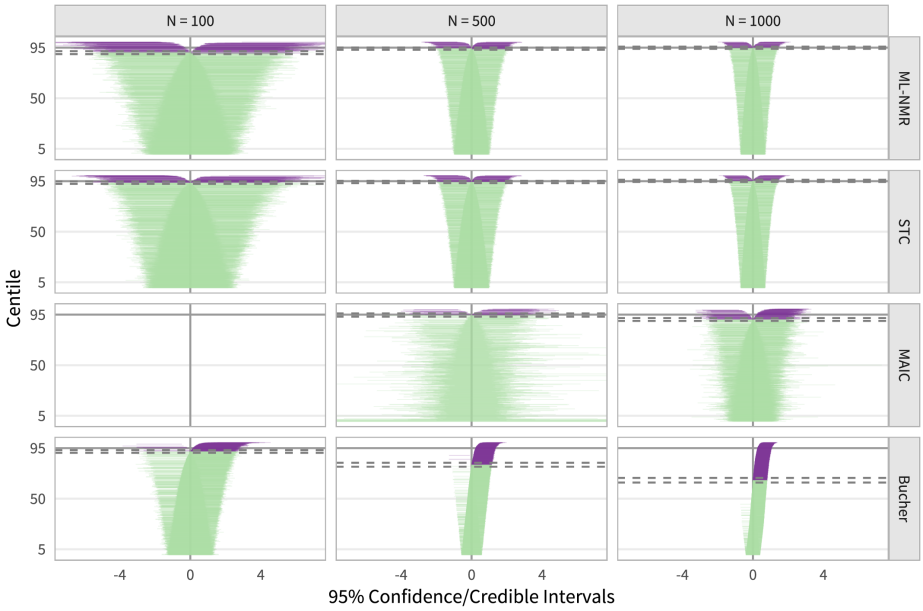
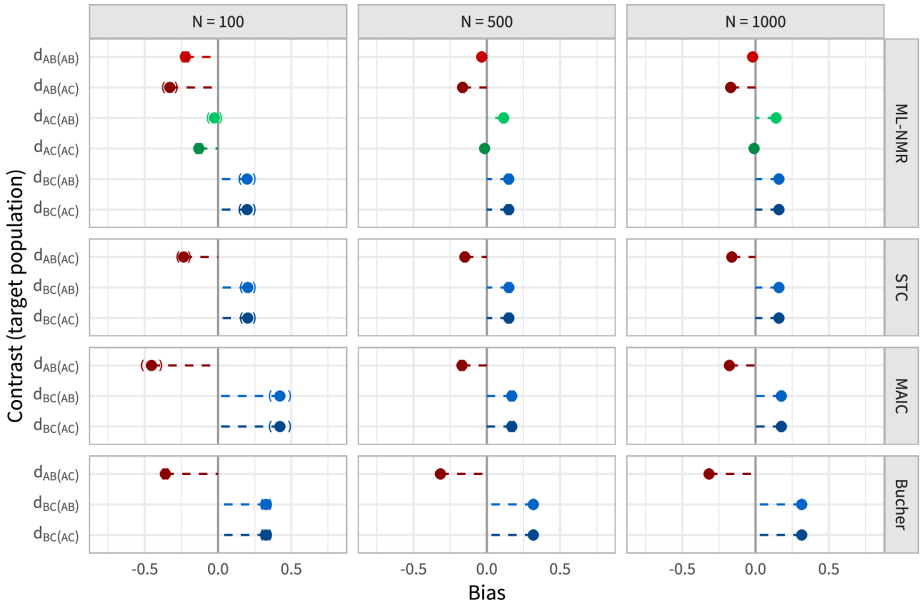
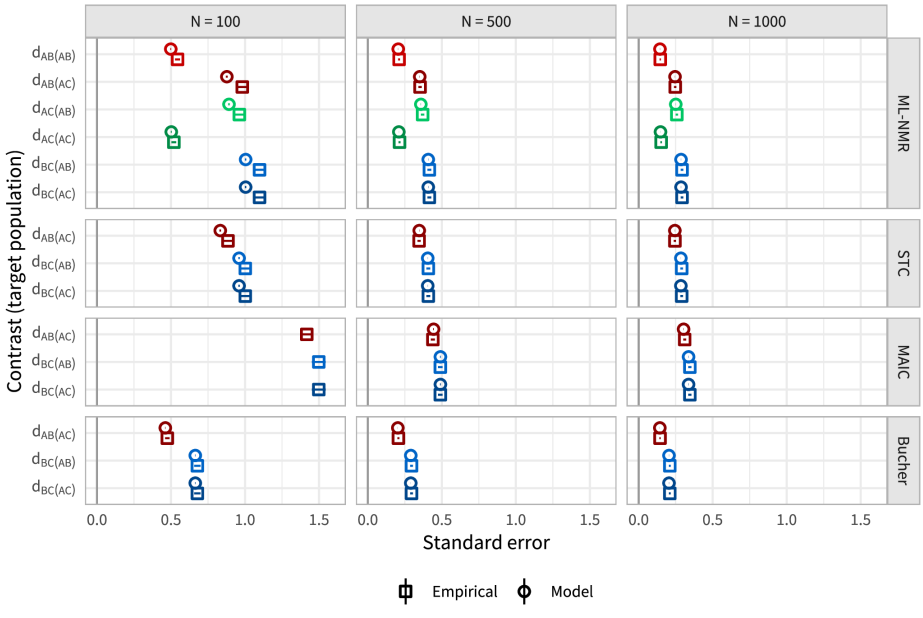


Figure 8.3 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenario a. Each method (other than Bucher) adjusts for the full set of effect modifiers. Sample size is varied between 100, 500, and 1000. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.



(a)



(b)

Figure 8.4 Bias (a) and standard errors (b) for the population-average contrast estimates for scenario a, along with 95% Monte Carlo confidence intervals. One of the two effect modifiers was not adjusted for. Sample size is varied between 100, 500, and 1000. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

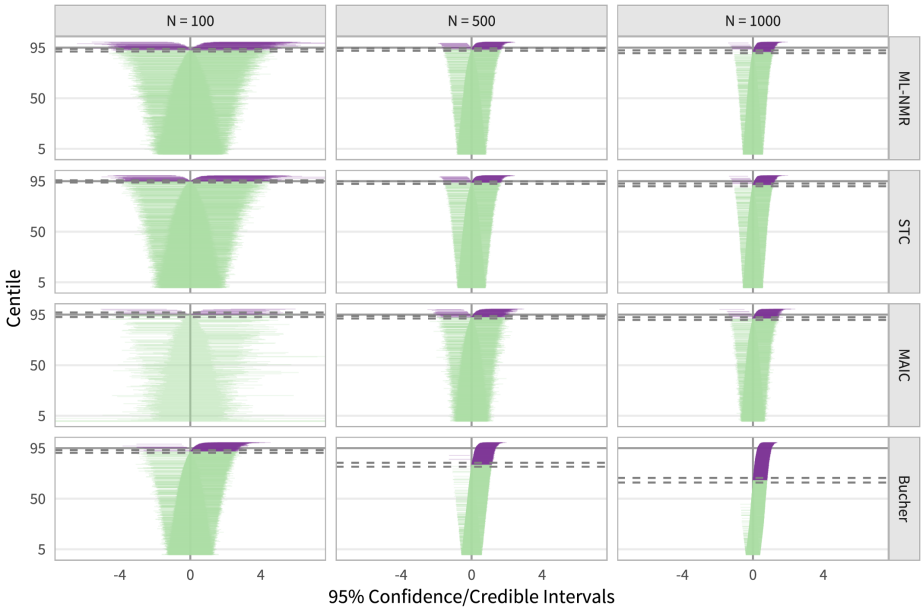
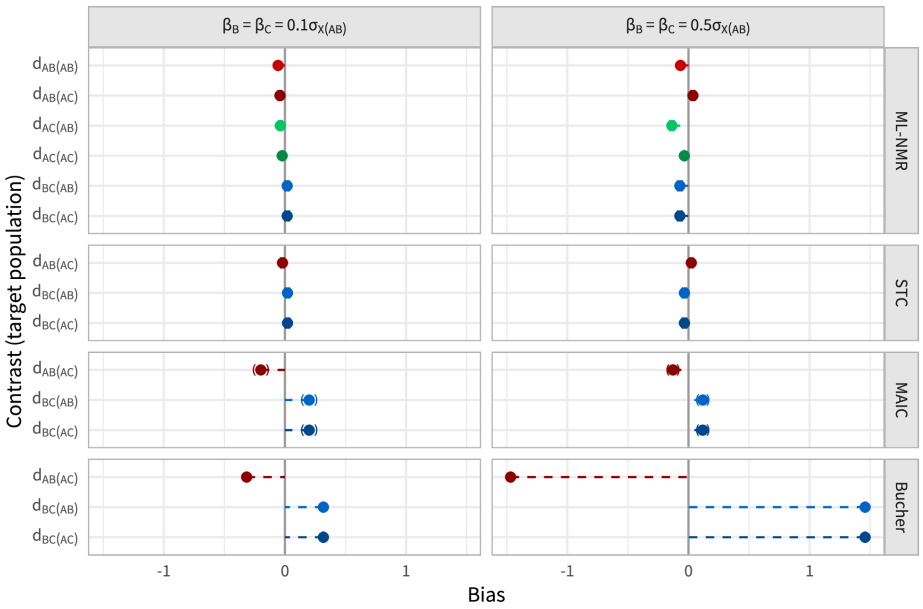


Figure 8.5 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenario a. One of the two effect modifiers was not adjusted for. Sample size is varied between 100, 500, and 1000. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.

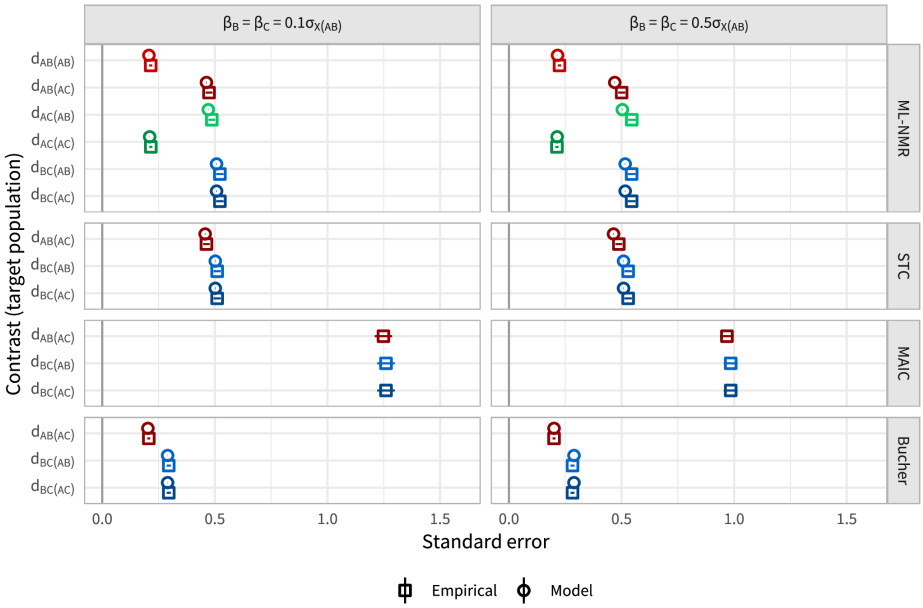
Scenario b: strength of effect modification 8.2.2

In scenario b, the strength of effect modification was varied from weak (0.1 change in log odds ratio per covariate standard deviation in the AB study) to strong (0.5 change in log odds ratio per covariate standard deviation in the AB study), and was the same for both treament B and C (so the shared effect modifier assumption holds). Figure 8.6a shows the bias in the population-average contrast estimates for each method, adjusting for all effect modifiers, along with 95% Monte Carlo confidence intervals. Figure 8.6b shows the corresponding empirical and model standard errors. (See also Table B.3.) Both ML-NMR and STC have successfully removed the bias from the standard indirect comparison, and display very similar standard errors. The standard errors are largely unaffected by the strength of the effect modification, although the model standard errors for ML-NMR and STC slightly underestimate the empirical standard errors when the effect modification is strong. MAIC has also removed most of the bias, though some remains due to lack of overlap between the two studies, and appears to do better when the effect modification is stronger. Empirical standard errors for MAIC are higher than for ML-NMR and STC, and again the bootstrap model standard errors are extremely unstable. Coverage for all three population adjustment methods is at the nominal level (Figure 8.7), although there is a small drop for ML-NMR and STC to 93.0% (91.9, 94.2) and 93.9% (92.9, 94.9) when the effect modification is stronger due to the slight underestimation of the standard error. Coverage for the standard indirect comparison drops to zero when the effect modification is strong, as the incurred bias is so large.

When one of the two effect modifiers is not adjusted for, all population adjustment methods fail to remove the bias (Figure 8.8a, also Table B.4). As expected, the amount of bias remaining is larger when the missing effect modifier is stronger. The standard errors of all population adjustment methods are reduced compared to adjustment for all effect modifiers, and again do not differ by the strength of effect modification (Figure 8.8b). The coverage for all population adjustment methods drops further when the missing effect modifier is stronger (Figure 8.9), due to the increased residual bias.



(a)



(b)

Figure 8.6 Bias (a) and standard errors (b) for the population-average contrast estimates for scenario b, along with 95% Monte Carlo confidence intervals. Strength of effect modification is varied from weak (0.1 change in log odds ratio per covariate standard deviation in the *AB* study) to strong (0.5 change in log odds ratio per covariate standard deviation in the *AB* study). Each method (other than Bucher) adjusts for the full set of effect modifiers. The points are coloured by contrast, with lighter shades for the *AB* population and darker for the *AC* population.

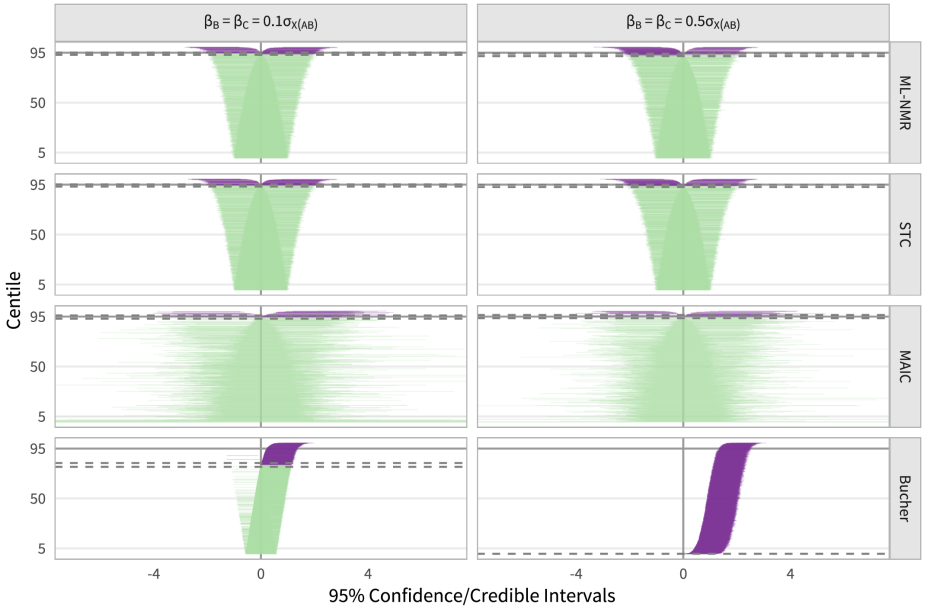
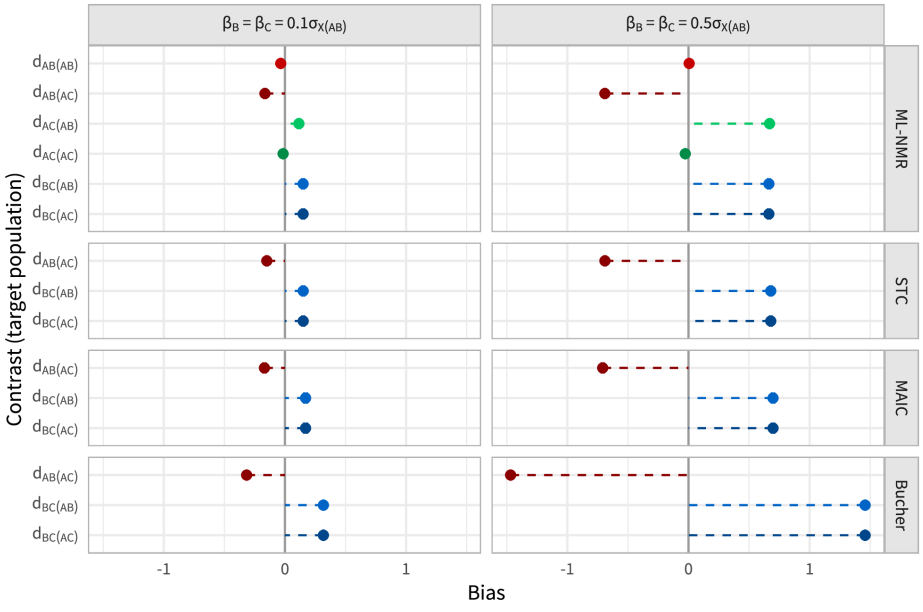
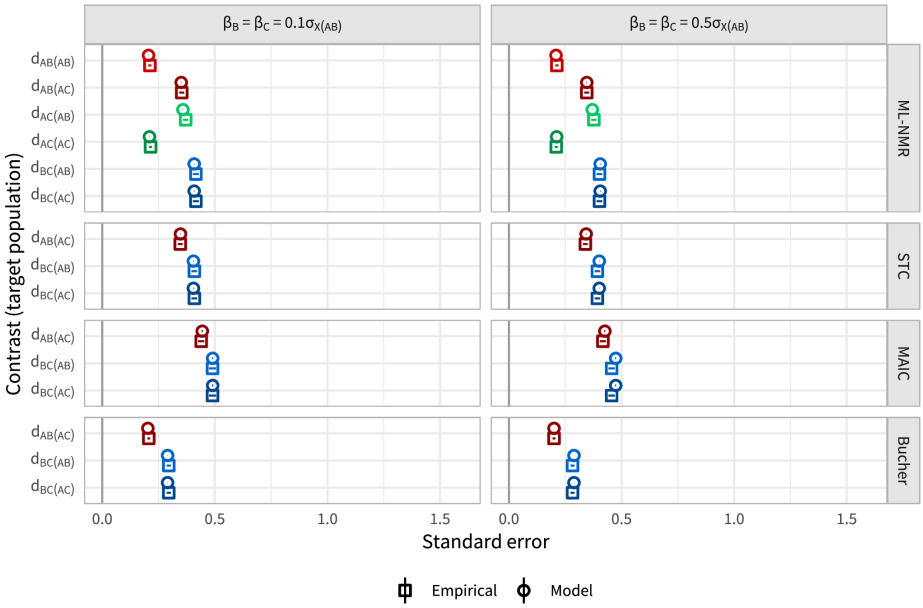


Figure 8.7 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenario b. Strength of effect modification is varied from weak (0.1 change in log odds ratio per covariate standard deviation in the AB study) to strong (0.5 change in log odds ratio per covariate standard deviation in the AB study). Each method (other than Bucher) adjusts for the full set of effect modifiers. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.



(a)



(b)

Figure 8.8 Bias (a) and standard errors (b) for the population-average contrast estimates for scenario b, along with 95% Monte Carlo confidence intervals. Strength of effect modification is varied from weak (0.1 change in log odds ratio per covariate standard deviation in the AB study) to strong (0.5 change in log odds ratio per covariate standard deviation in the AB study). One of the two effect modifiers was not adjusted for. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

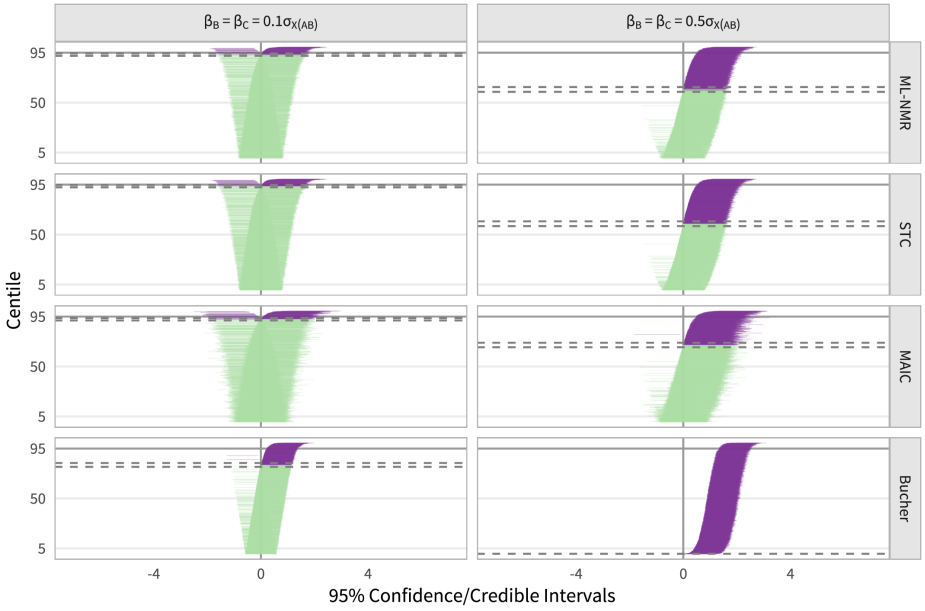


Figure 8.9 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenario b. Strength of effect modification is varied from weak (0.1 change in log odds ratio per covariate standard deviation in the AB study) to strong (0.5 change in log odds ratio per covariate standard deviation in the AB study). One of the two effect modifiers was not adjusted for. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.

8.2.3 Scenario c: shared EM assumption

In scenario c the shared effect modifier assumption is broken, so that treatment B is subject to weak effect modification whilst treatment C is subject to strong effect modification and vice versa. We expect to see that all population adjustment methods are capable of producing unbiased estimates in the AC population, but that extrapolation into other populations is biased, and indeed this is what occurs (Figure 8.10a). Standard errors (Figure 8.10b) and coverage (Figure 8.11) are largely unchanged from scenario b (see also Table B.5). Again, ML-NMR and STC show slight underestimation of the empirical standard error when there is strong effect modification, leading to a small drop in coverage down to around 93%.

The increase in bias for a standard indirect comparison when treatment B is strongly modified compared to weakly modified (Figure 8.10a) is due to the focus on the AC population. Since a standard indirect comparison does not adjust for population differences, the estimates of d_{AB}, d_{AC}, and d_{BC} are the same regardless of population. However, when the target of inference is a comparison in the AC population, the estimate of $d_{AC(AC)}$ is unbiased and bias in the estimate of $d_{AB(AC)}$ (and thus $d_{BC(AC)}$) is driven by the strength of effect modification of treatment B. If the target of inference was instead a comparison in the AB population, this pattern would be reversed (bias now being driven by the strength of effect modification of treatment C). A similar pattern, driven by the same mechanism, is observed in scenario b (see Figure 8.2a).

When one of the two effect modifiers is not adjusted for, further bias is introduced into the estimates, in all target populations (Figure 8.12a, see also Table B.6). This bias is generally larger for estimates involving extrapolation of treatment effects affected by strong unobserved effect modifiers, e.g. $d_{AC(AB)}$ and $d_{BC(AB)}$ when C is strongly modified by a missing effect modifier. Again, standard error is reduced when fewer effect modifiers are adjusted for (Figure 8.12b). Coverage for $d_{BC(AC)}$ is close to nominal level for all population adjustment methods when the missing effect modifier for B is weak as the incurred bias is small, but drops severely when the missing effect modifier for B is strong as the incurred bias is large (Figure 8.13).

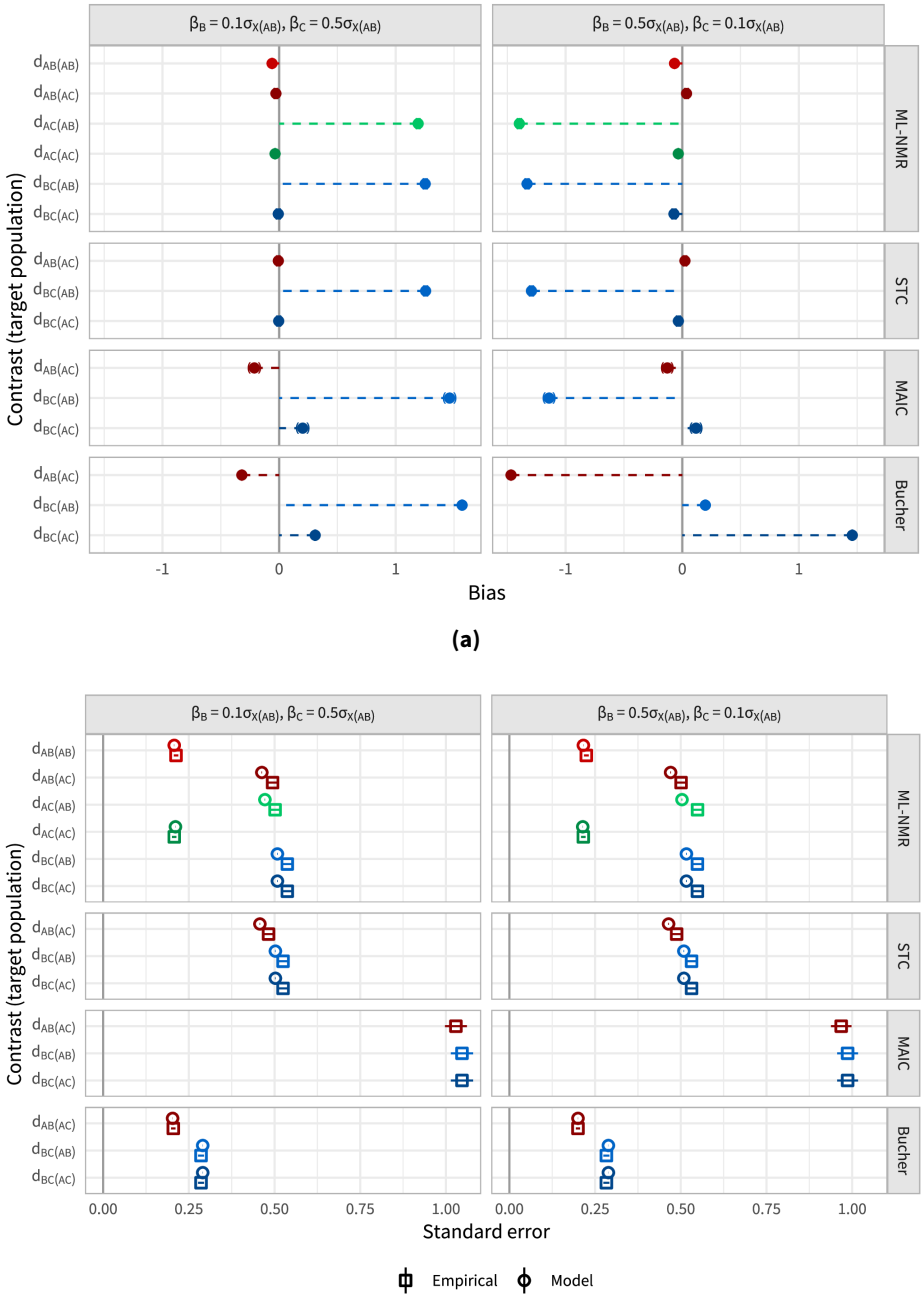


Figure 8.10 Bias (a) and standard error (b) for the population-average contrast estimates for scenario c, along with 95% Monte Carlo confidence intervals. Each method (other than Bucher) adjusts for the full set of effect modifiers. The shared effect modifier assumption is broken, so that treatment B is subject to weak effect modification whilst treatment C is subject to strong effect modification and vice versa. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

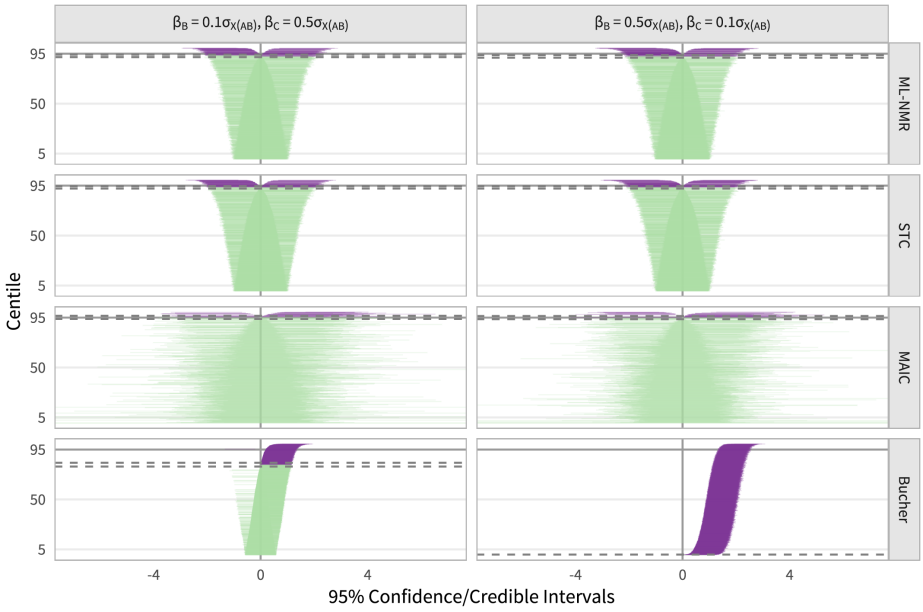
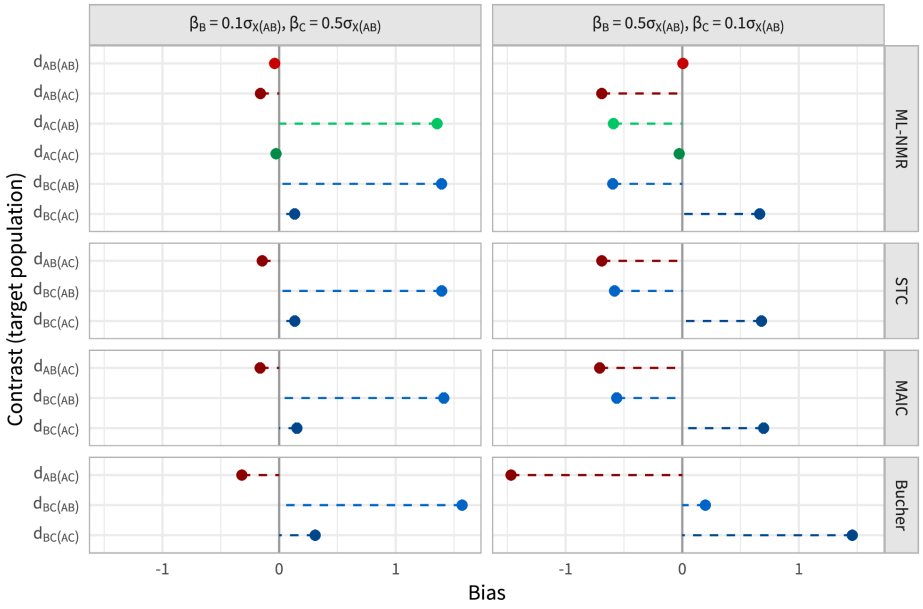
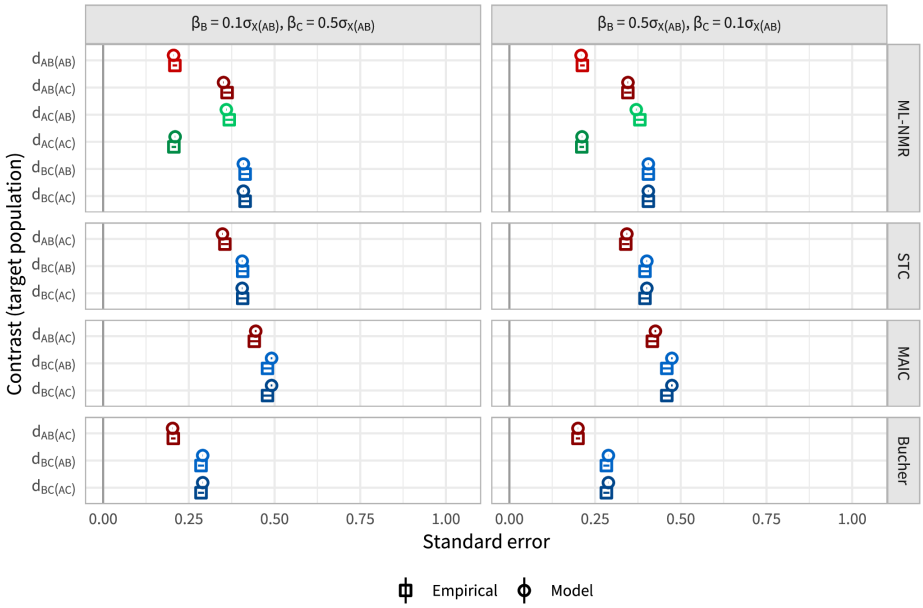


Figure 8.11 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenario c. Each method (other than Bucher) adjusts for the full set of effect modifiers. The shared effect modifier assumption is broken, so that treatment B is subject to weak effect modification whilst treatment C is subject to strong effect modification and vice versa. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.



(a)



(b)

Figure 8.12 Bias (a) and standard errors (b) for the population-average contrast estimates for scenario c, along with 95% Monte Carlo confidence intervals. One of the two effect modifiers was not adjusted for. The shared effect modifier assumption is broken, so that treatment B is subject to weak effect modification whilst treatment C is subject to strong effect modification and vice versa. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

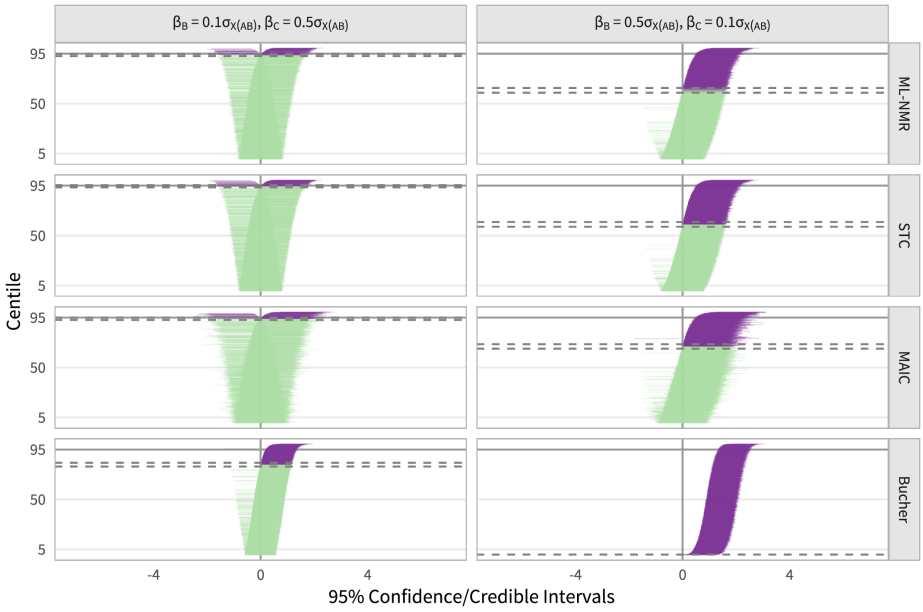
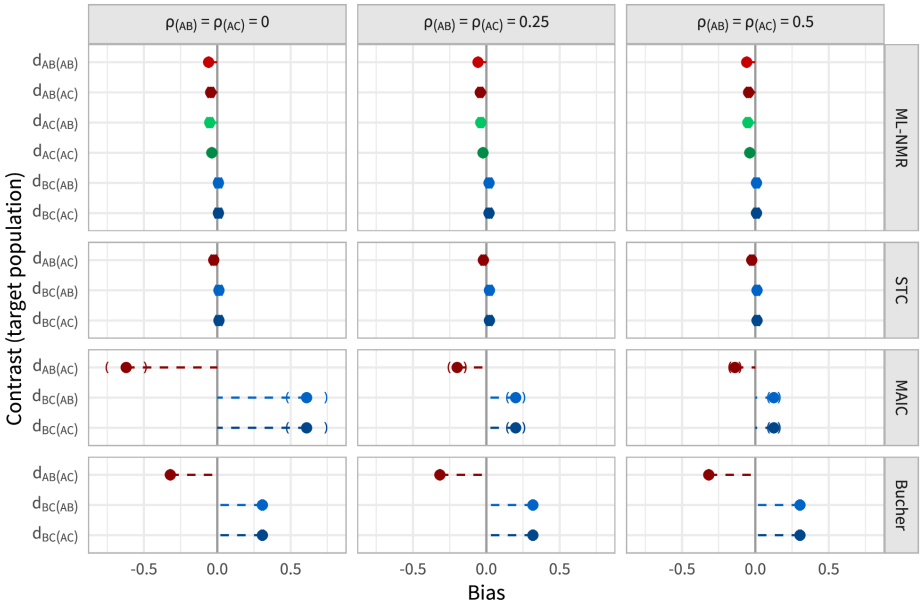


Figure 8.13 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenario c. One of the two effect modifiers was not adjusted for. The shared effect modifier assumption is broken, so that treatment B is subject to weak effect modification whilst treatment C is subject to strong effect modification and vice versa. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.

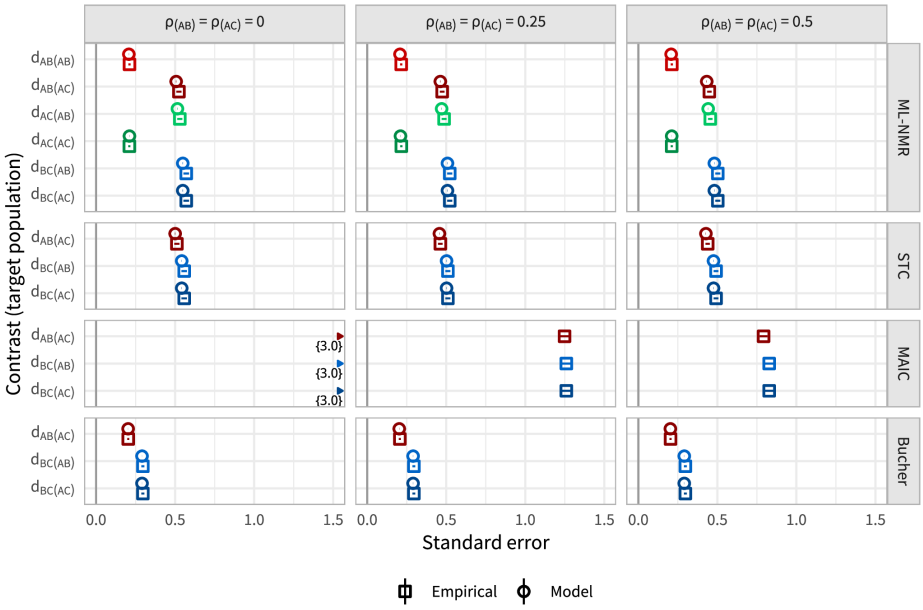
Scenario d: correlation between covariates 8.2.4

In scenario d, the correlation between covariates is varied between 0, 0.25, and 0.5, and the correlation is the same in both study populations. Figure 8.14a shows that ML-NMR and STC are unaffected by the correlation between covariates, achieving bias removal regardless of the correlation. Standard errors are similarly unaffected (Figure 8.14b), and coverage is at the nominal level (Figure 8.15). Table B.7 shows the results in tabular format. For MAIC, both bias and standard error are reduced as the correlation between covariates increases. This is because, as the correlation increases, the effective number of covariates decreases and the overlap between study populations increases (Figure 8.16). The simulation parameter κ is only a proxy for overlap and does not account for correlation between covariates, so the true overlap changes with the correlation despite holding κ constant at 0.5. Since MAIC cannot eliminate the bias, coverage drops as the standard error decreases (as the correlation increases), down to 93.4% (92.3, 94.6) when the correlation is 0.5.

When one of the two effect modifiers is not adjusted for, all population adjustment methods produce biased estimates (Figure 8.17a, see also Table B.8). However, as we expect, the amount of bias reduces as the correlation between the observed and missing effect modifiers increases. If we were to continue simulations with correlations tending closer to 1, the bias due to a missing effect modifier would disappear entirely. Again, standard error is reduced when fewer effect modifiers are adjusted for (Figure 8.17b). Coverage for all population adjustment methods was slightly below the nominal level (around 93%) due to the remaining bias (Figure 8.18).



(a)



(b)

Figure 8.14 Bias (a) and standard errors (b) for the population-average contrast estimates for scenario d, along with 95% Monte Carlo confidence intervals. The correlation between covariates is varied between 0, 0.25, and 0.5. Each method (other than Bucher) adjusts for the full set of effect modifiers. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

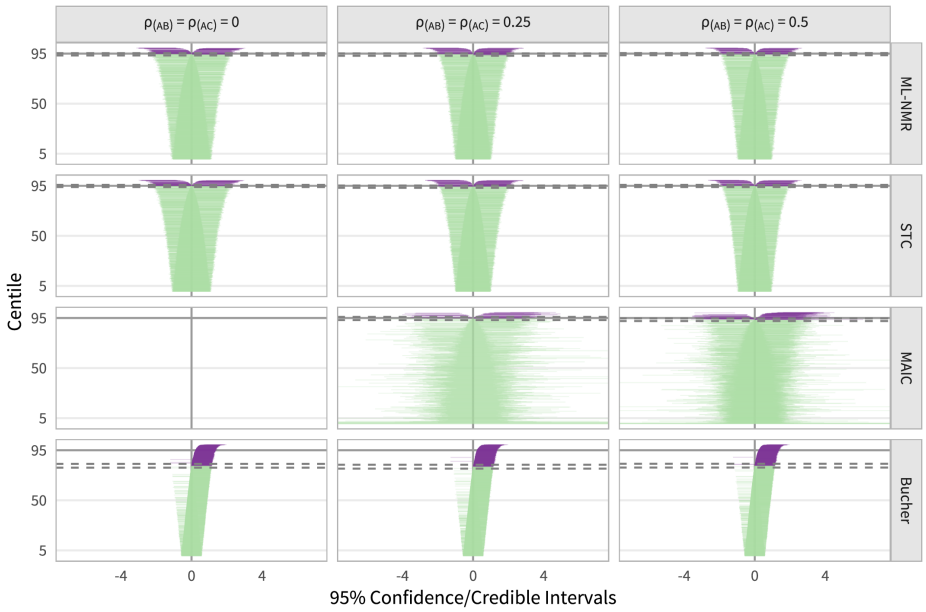


Figure 8.15 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenario d. The correlation between covariates is varied between 0, 0.25, and 0.5. Each method (other than Bucher) adjusts for the full set of effect modifiers. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.

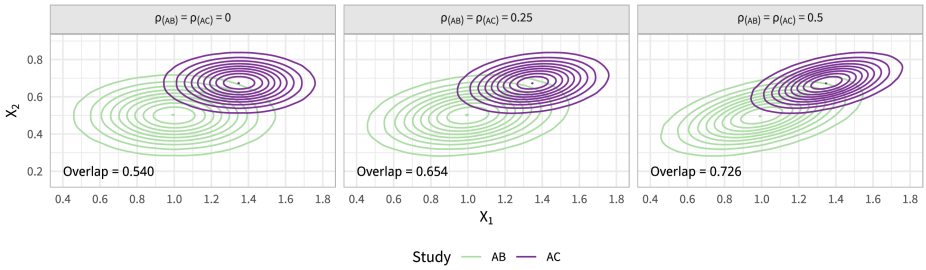
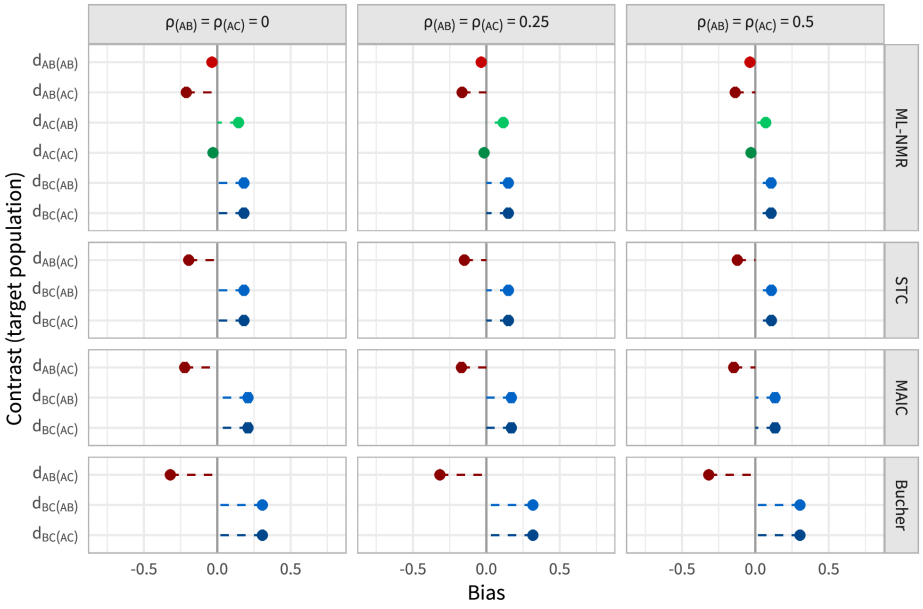
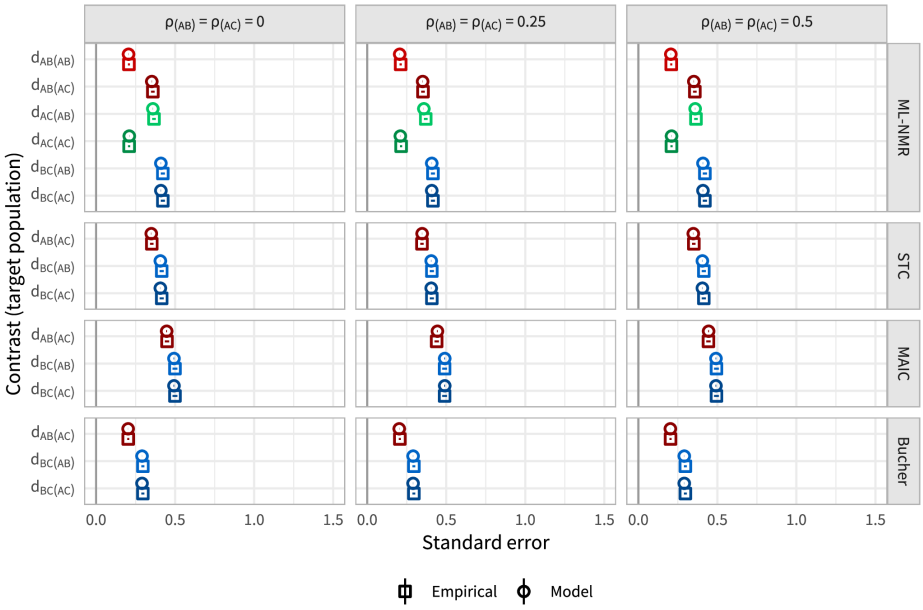


Figure 8.16 Joint covariate distributions in the AB and AC study, as the correlation between covariates is varied. The true overlap is defined as the proportion of the AC joint density contained within the 95% HDR of the AB joint density, calculated using numerical integration.



(a)



(b)

Figure 8.17 Bias (a) and standard errors (b) for the population-average contrast estimates for scenario d, along with 95% Monte Carlo confidence intervals. The correlation between covariates is varied between 0, 0.25, and 0.5. One of the two effect modifiers was not adjusted for. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

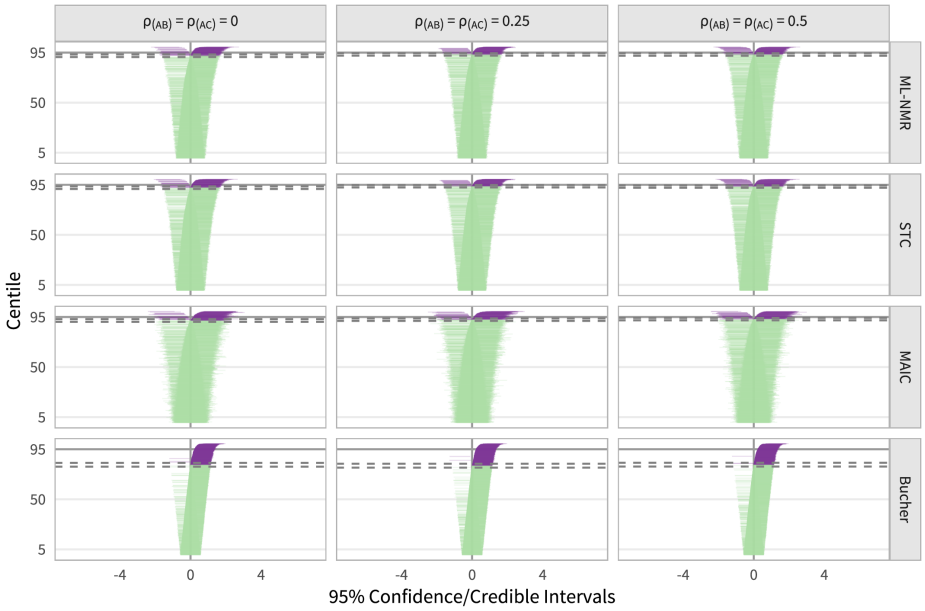


Figure 8.18 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenario d. The correlation between covariates is varied between 0, 0.25, and 0.5. One of the two effect modifiers was not adjusted for. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.

8.2.5 Scenarios e and f: between-study overlap and covariate-outcome relationship

In scenarios e and f, the between-study overlap and covariate-outcome relationship are varied jointly. The between-study overlap is varied between 0 (no overlap, all of AC population outside of the AB population), 0.5 (approximately 50% of AC outside of AB), and 1 (full overlap, all of AC within AB), and the covariate-outcome relationship is either linear or non-linear (equation 8.2).

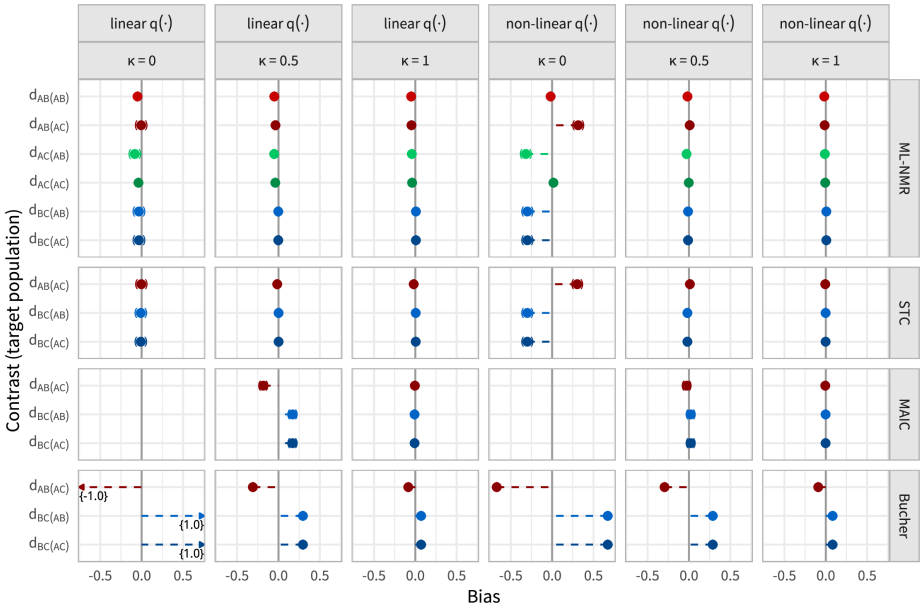
When the covariate-outcome relationship is linear, both ML-NMR and STC produce unbiased estimates (Figure 8.19a, Table B.9). MAIC is unable to produce any estimates when there is no overlap between study populations, and remains biased when the overlap is 0.5. Only when the study populations overlap completely does MAIC produce unbiased estimates.

When the covariate-outcome relationship is non-linear, ML-NMR and STC (set to fit a linear relationship) are both biased when there is no overlap between the study populations. There is no discernible bias when the overlap is 0.5 or 1, as equation (8.2) behaves linearly within the range of the AB population. However, prediction into another target population with a greater difference in covariate values from the AB and AC populations would still result in bias. Again, MAIC cannot produce estimates when there is no overlap between study populations.

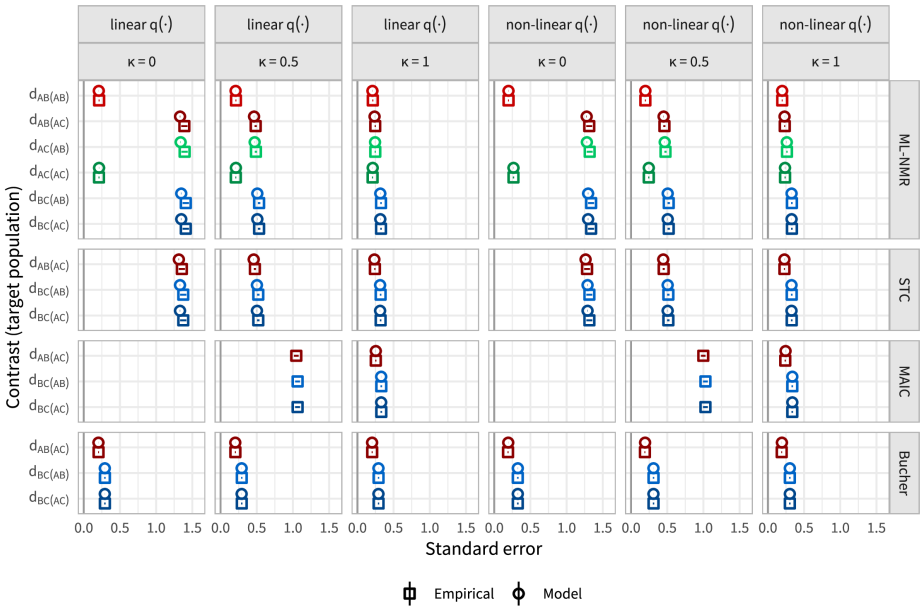
Standard errors for all population adjustment methods decrease as the level of overlap increases (Figure 8.19b). Empirical standard errors are well-estimated by model standard errors for ML-NMR and STC, however MAIC only produced stable bootstrap model standard errors when the overlap was 1.

For the population adjustment methods, nominal coverage is achieved when the covariate-outcome relationship is linear (Figure 8.20). With a non-linear relationship, coverage drops slightly to 93.0% (91.9, 94.2) for ML-NMR and to 93.4% (92.4, 94.5) when there is no overlap, but is unaffected at 0.5 and full overlap as there is no discernable bias. MAIC achieves nominal coverage for 0.5 and full overlap.

When one of the two effect modifiers is not adjusted for, all population adjustment methods produce biased estimates (Figure 8.21a, Table B.10). As expected, the bias due to the missing effect modifier reduces as the overlap between studies increases, since the difference in the missing effect modifier between the study populations becomes smaller. Due to this, coverage shows the opposite relationship with overlap, dropping as the overlap decreases (Figure 8.22). Again, standard error is reduced when fewer effect modifiers are adjusted for (Figure 8.21b).



(a)



(b)

Figure 8.19 Bias (a) and standard errors (b) the population-average contrast estimates for scenarios e and f, along with 95% Monte Carlo confidence intervals. Each method (other than Bucher) adjusts for the full set of effect modifiers. The between-study overlap and covariate-outcome relationship are varied jointly. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

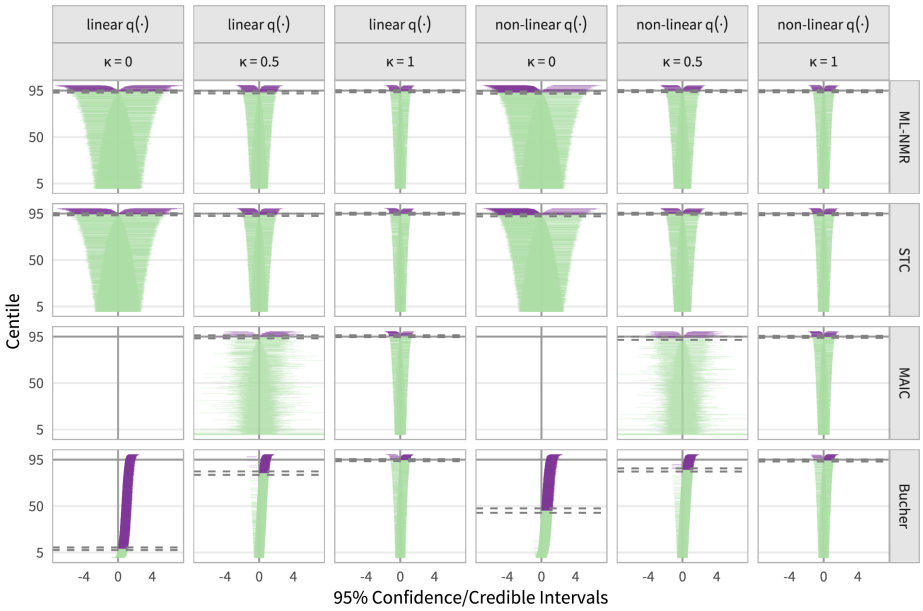
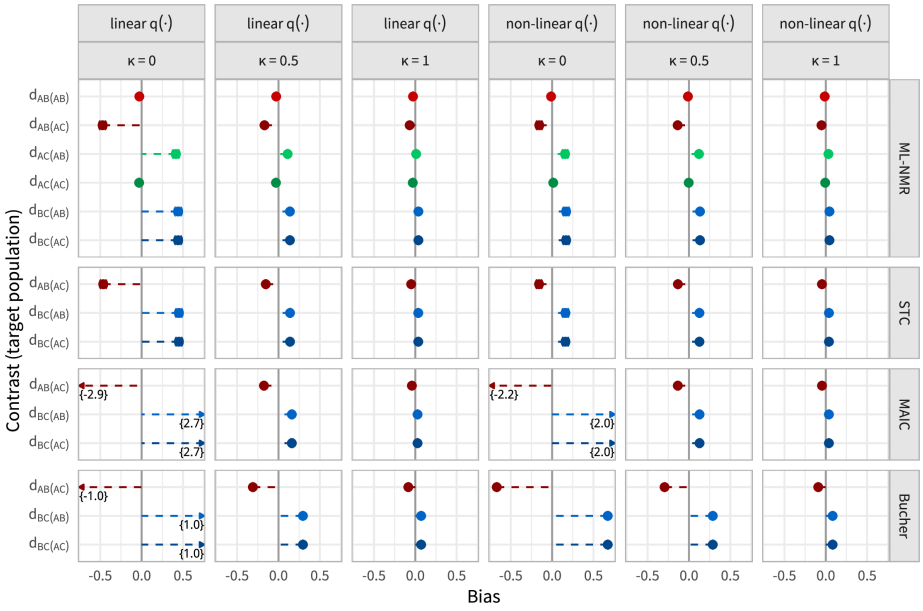
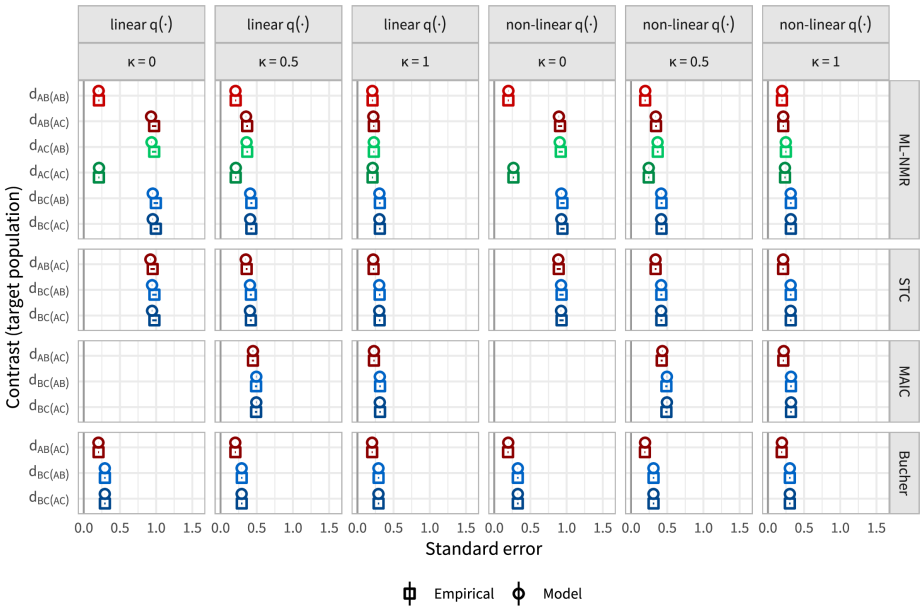


Figure 8.20 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenarios e and f. Each method (other than Bucher) adjusts for the full set of effect modifiers. The between-study overlap and covariate-outcome relationship are varied jointly. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.



(a)



(b)

Figure 8.21 Bias (a) and standard errors (b) for the population-average contrast estimates for scenarios e and f, along with 95% Monte Carlo confidence intervals. One of the two effect modifiers was not adjusted for. The between-study overlap and covariate-outcome relationship are varied jointly. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

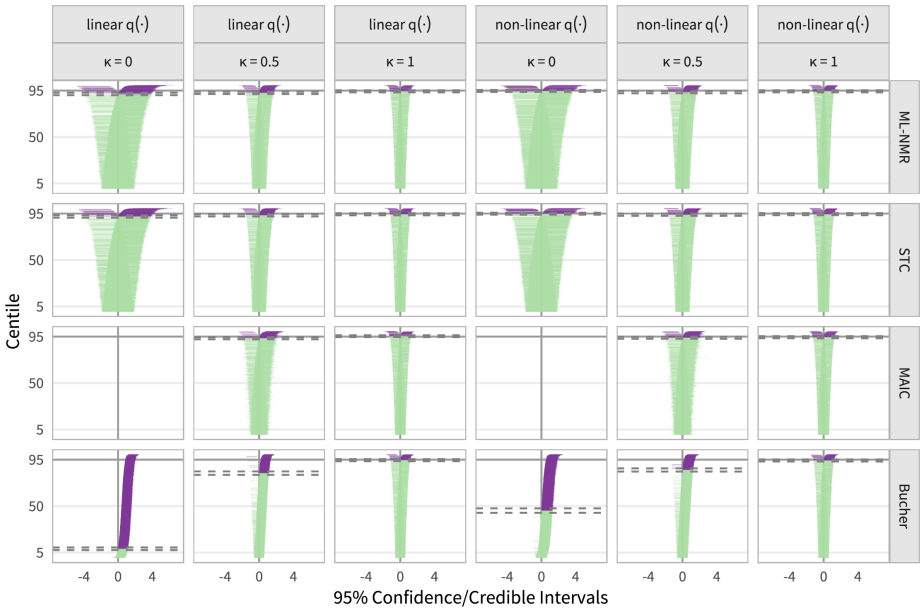


Figure 8.22 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenarios e and f. One of the two effect modifiers was not adjusted for. The between-study overlap and covariate-outcome relationship are varied jointly. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.

Scenarios g, h, and i: covariate distributions and correlation structures 8.2.6

In scenarios g, h, and i, the covariate distributions and correlation structures in each study population are varied jointly. The covariate distributions are set to either Normal or Gamma in each of the study populations, and the correlation between covariates in the AC population was varied between 0, 0.25, and 0.5. The correlation between covariates in the AB population was held constant at 0.25.

We are particularly interested in the performance of ML-NMR in this scenario, since ML-NMR makes the assumption that distributional form and correlation structure are the same in the AgD studies as in the IPD. Scenarios g, h, and i break this assumption. Despite this, ML-NMR is not seen to incur any bias when the distributional form and/or correlation structure are different in each population (Figure 8.23). The empirical standard errors, and their estimation by the model standard errors, are also unaffected (Figure 8.24). The bias and standard error of MAIC are slightly smaller when the covariates in the AB population are Gamma distributed compared to Normally distributed; however, this is due to the skew of the Gamma distribution slightly increasing the true overlap between the study populations (compare row 1 with 3, and 2 with 4, in Figure 8.26). Nominal coverage is achieved by all population adjustment methods, regardless of the covariate distributions or correlation structure (Figure 8.25). See Table B.11 for a table of these results.

Once again, when one of the two effect modifiers is not adjusted for, all population adjustment methods produce biased estimates (Figure 8.27) but standard error is reduced (Figure 8.28). Coverage is slightly reduced for all population adjustment methods, to around 94%, as a result of the bias (Figure 8.25). The results are tabulated in Table B.12.

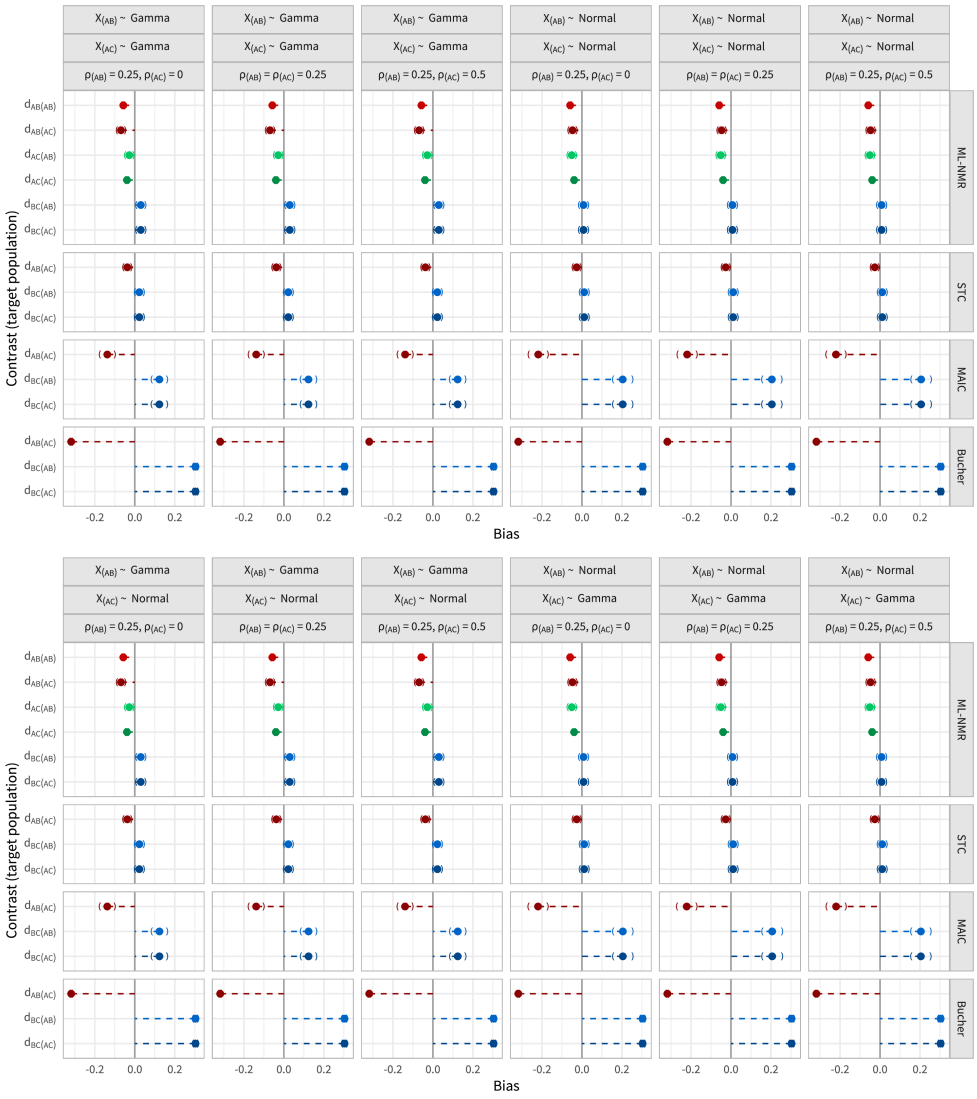


Figure 8.23 Bias in the population-average contrast estimates for scenarios g, h, and i, along with 95% Monte Carlo confidence intervals. Each method (other than Bucher) adjusts for the full set of effect modifiers. The covariate distributions and correlation structures in each study population are varied jointly. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

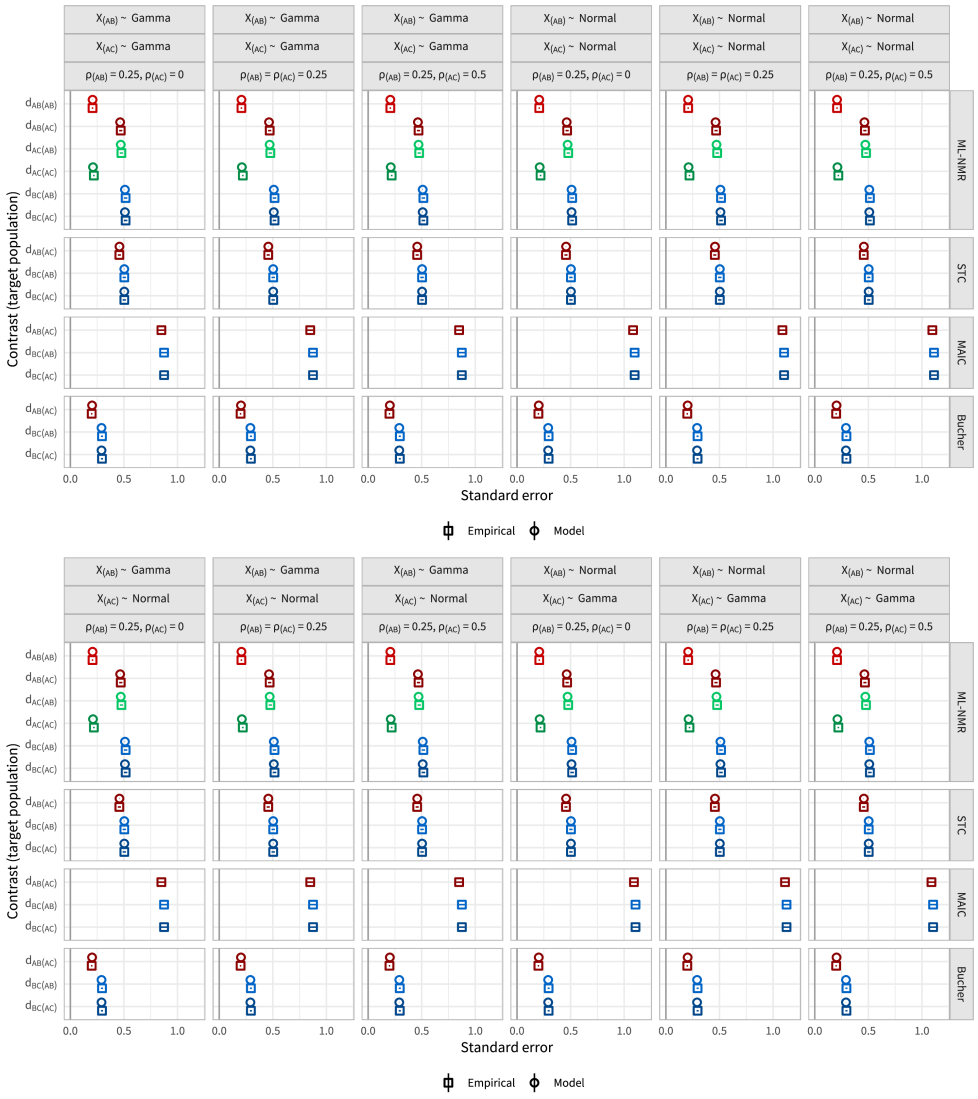


Figure 8.24 Empirical and model standard errors for scenarios g, h, and i, along with 95% Monte Carlo confidence intervals. Each method (other than Bucher) adjusts for the full set of effect modifiers. The covariate distributions and correlation structures in each study population are varied jointly. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

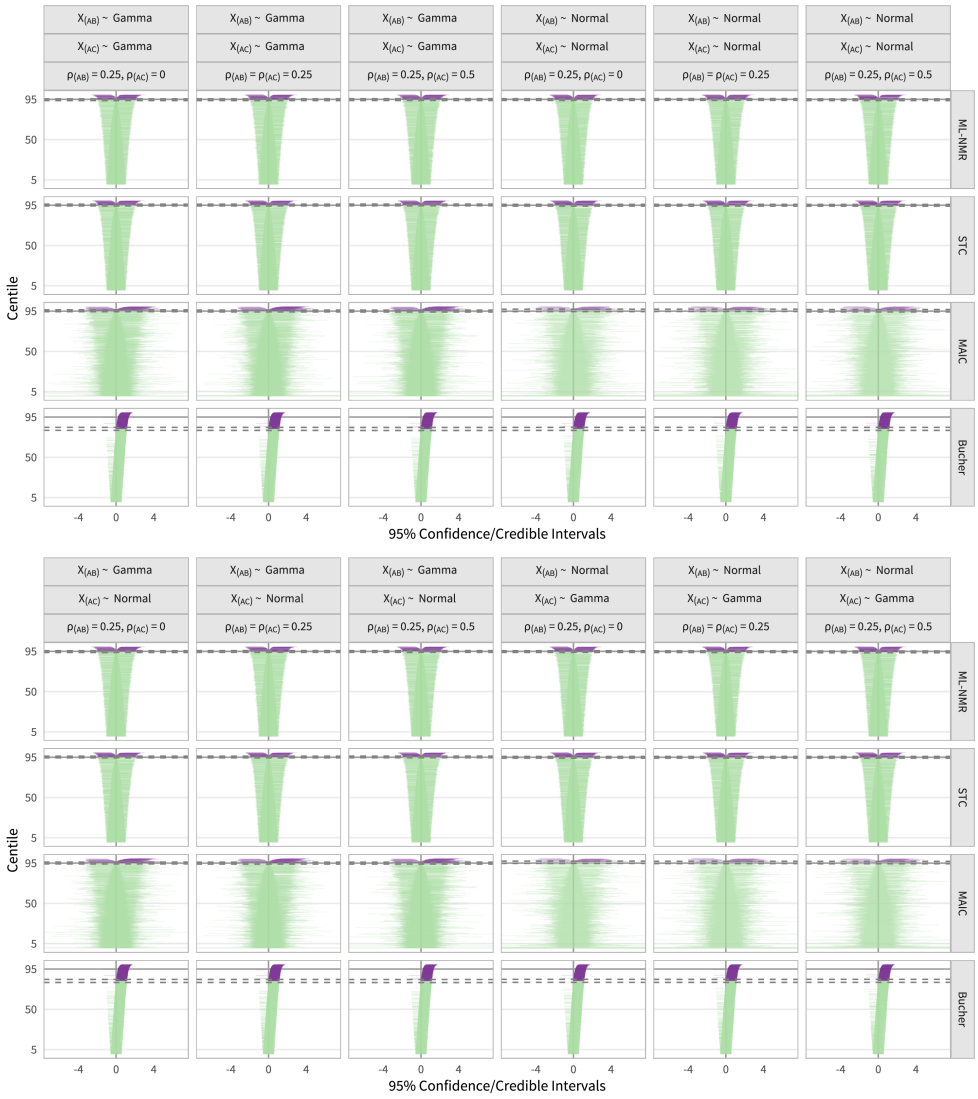


Figure 8.25 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenarios g, h, and i. Each method (other than Bucher) adjusts for the full set of effect modifiers. The covariate distributions and correlation structures in each study population are varied jointly. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.

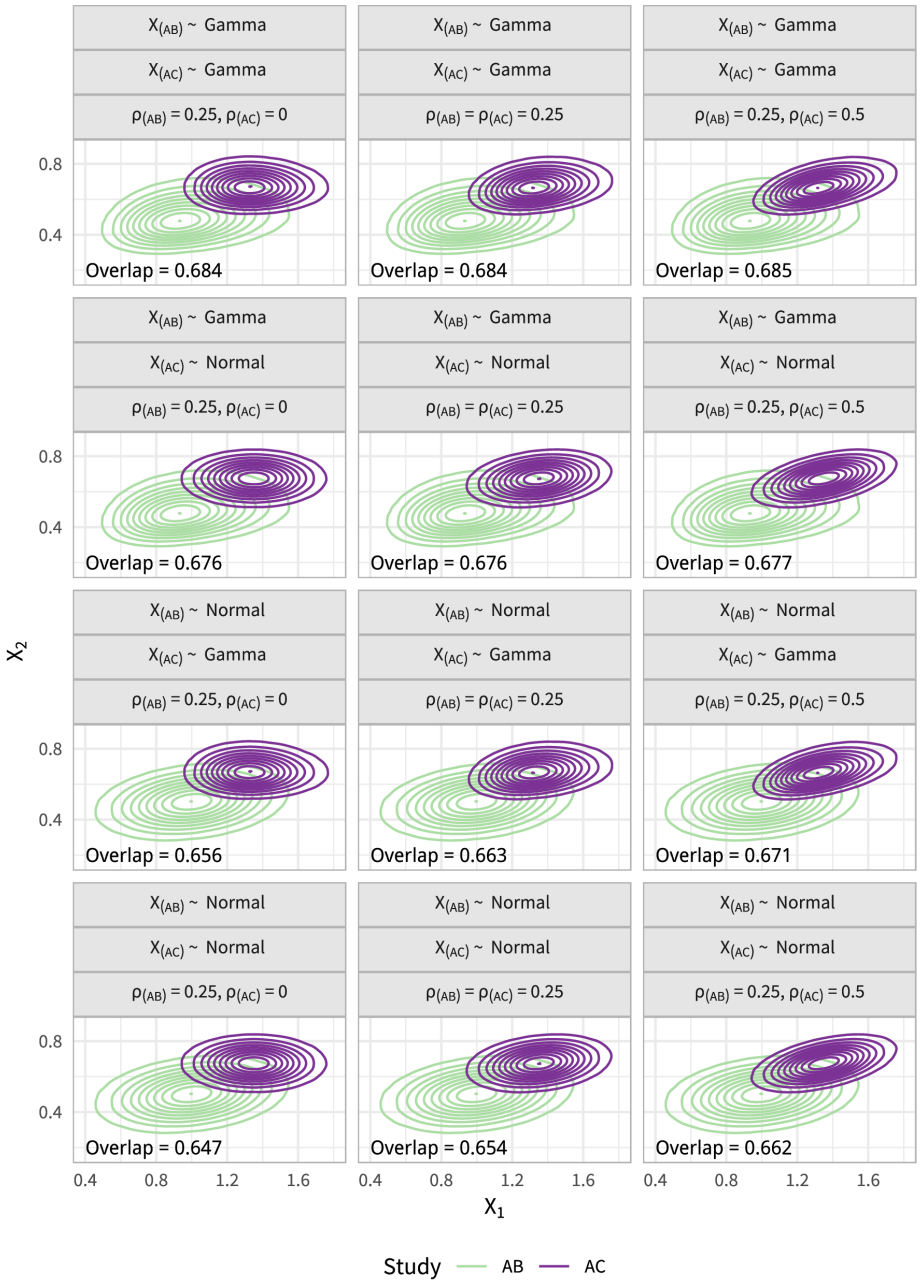


Figure 8.26 Joint covariate distributions in the AB and AC study, as the covariate distributions and correlation between covariates are varied. The true overlap is defined as the proportion of the AC joint density contained within the 95% HDR of the AB joint density, calculated using numerical integration.

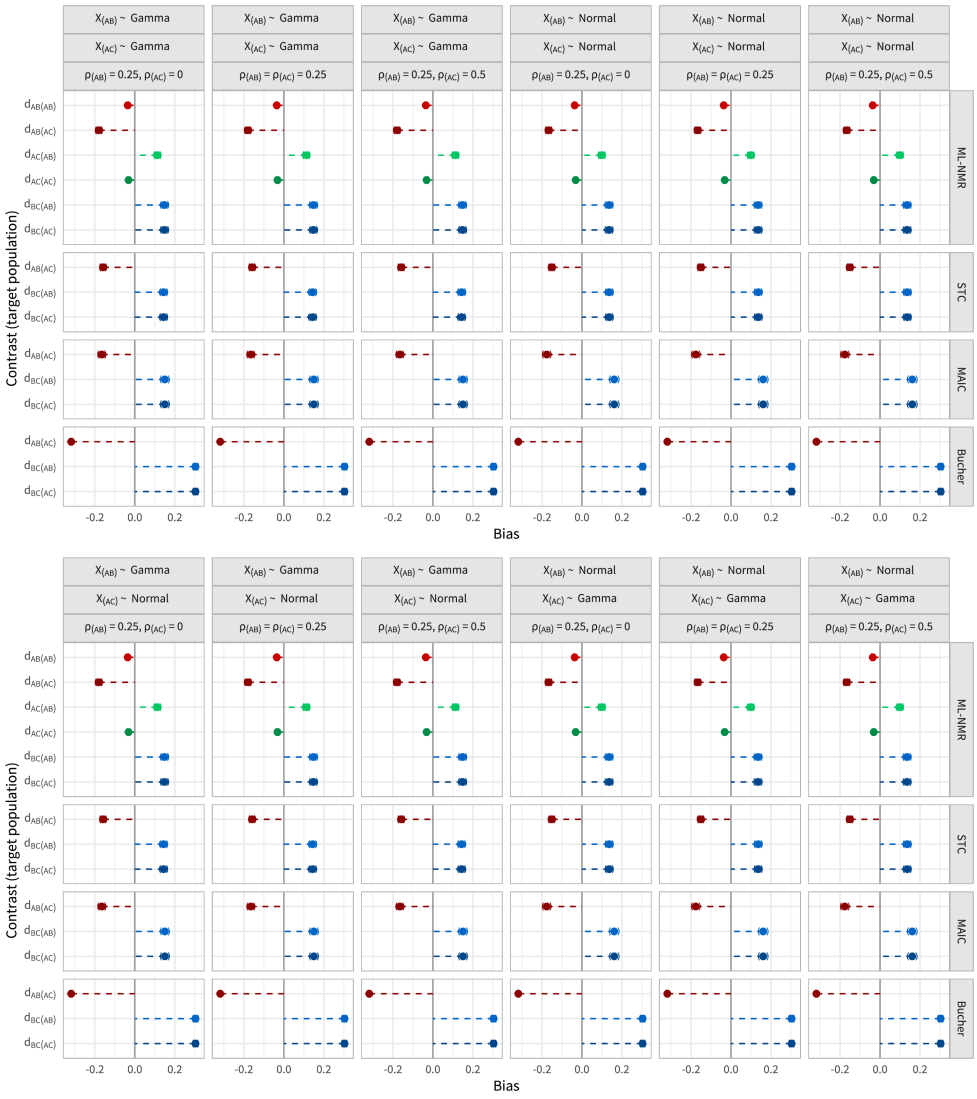


Figure 8.27 Bias in the population-average contrast estimates for scenarios g, h, and i, along with 95% Monte Carlo confidence intervals. One of the two effect modifiers was not adjusted for. The covariate distributions and correlation structures in each study population are varied jointly. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

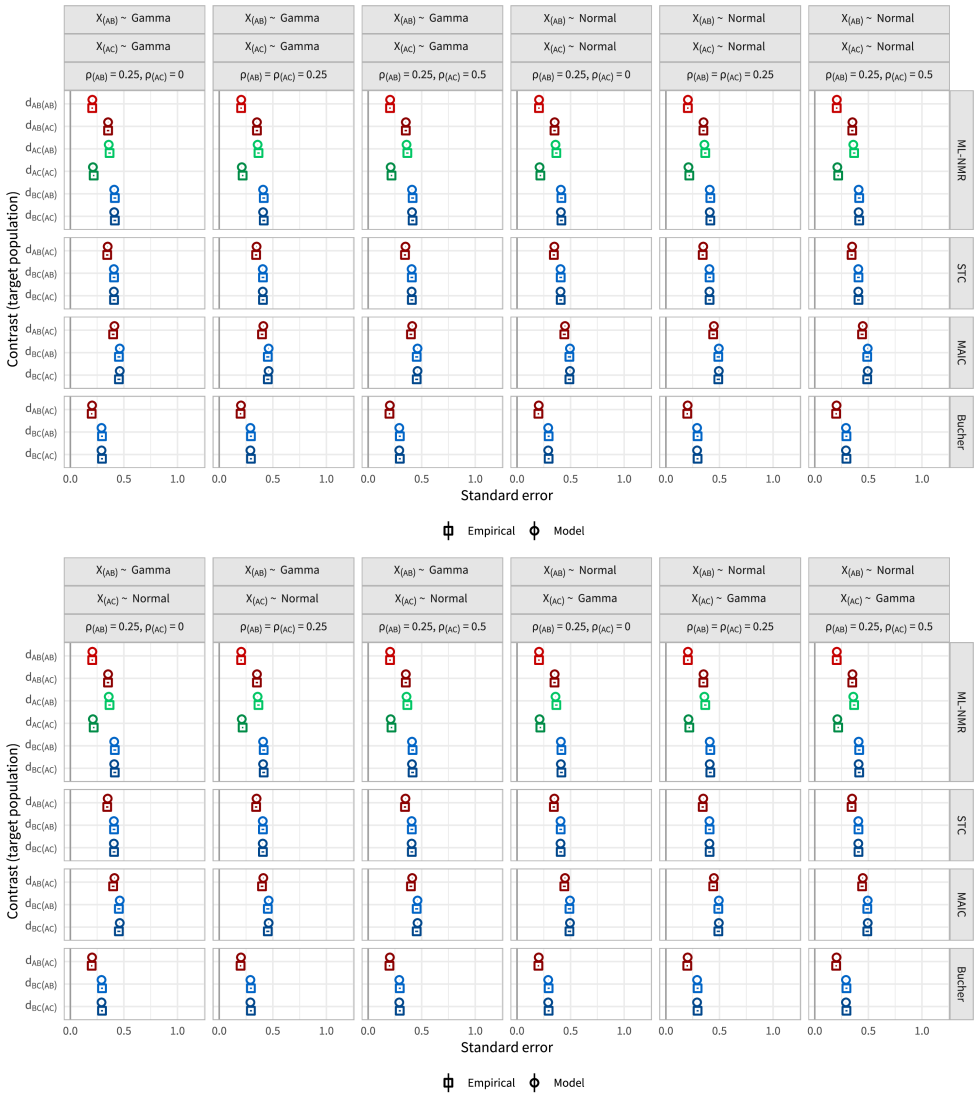


Figure 8.28 Empirical and model standard errors for scenarios g, h, and i, along with 95% Monte Carlo confidence intervals. One of the two effect modifiers was not adjusted for. The covariate distributions and correlation structures in each study population are varied jointly. The points are coloured by contrast, with lighter shades for the AB population and darker for the AC population.

Figure 8.29 Coverage zip plots for the $d_{BC(AC)}$ contrast estimate for scenarios g, h, and i. One of the two effect modifiers was not adjusted for. The covariate distributions and correlation structures in each study population are varied jointly. The 95% confidence/credible intervals are coloured as coverers (green) or non-coverers (purple), and the colour change should occur at the 95th centile (i.e. nominal coverage). The horizontal dashed lines are 95% Monte Carlo confidence intervals for the coverage.

Conclusion 8.3

In this simulation study, we have investigated the performance of ML-NMR in comparison with current population adjustment methods (MAIC and STC), in a wide range of scenarios. ML-NMR and STC perform very similarly throughout the simulation study. This is to be expected, as the two methods are both regression adjustment methods. ML-NMR offers additional advantages over STC, including the ability to synthesise a larger network of treatments, and to produce estimates in any target population. ML-NMR makes further assumptions about the covariate distributions and correlation structure in the aggregate population in order to derive the aggregate likelihood through integration, namely that they are the same as in the IPD population. However, the performance of ML-NMR was not found to be sensitive to this assumption (see scenarios g, h, and i, Section 8.2.6).

Serious questions must be raised about the use of MAIC for population adjustment. MAIC performed poorly in all simulation scenarios, and in some cases even increased the bias compared to a standard indirect comparison. Bootstrap error estimation was also extremely unstable, except for the largest sample sizes. The issues with MAIC stem from the fact that it is a reweighting method, and therefore cannot extrapolate. As such, bias can only be completely removed when the population of the AgD study is entirely contained within the population of the IPD study (see scenario e, Section 8.2.5). However, when the two study populations overlap entirely there is unlikely to be much bias as the imbalance in effect modifiers is necessarily small, and thus population adjustment methods may not be needed. Furthermore, this means that MAIC is only valid from at most one end of the indirect comparison—that with IPD on the study with a broader distribution of covariates. MAIC analysis in the other direction will result in estimates that are biased, possibly by more than a standard indirect comparison (as well as being produced for a different target population). If the two study populations do not overlap at all, then no MAIC analyses are possible.

As regression methods, ML-NMR and STC are able to extrapolate beyond the range of the IPD, producing estimates even when there is no overlap between study populations. However, when extrapolation occurs, estimates will only be unbiased if such extrapolation is valid. For example, in scenarios e and f (Section 8.2.5), ML-NMR and STC produced biased estimates when there was no overlap between populations and the true covariate-outcome relationship was non-linear outside of the range of the IPD, but only a linear relationship was accounted for.

Notably, all population adjustment methods are susceptible to bias (and as a result, under-coverage) when an effect modifier is missing from the adjustment. This highlights the necessity of careful and considered selection of potential effect modifiers prior to analysis (see Section 2.3.1; also Phillippo et al. 2016; Phillippo et al. 2018a). When all effect modifiers have been identified and included in the adjustment model, ML-NMR and STC are both robust techniques for obtaining population-adjusted indirect comparisons. The additional flexibility of ML-NMR makes this an attractive choice for population adjustment in a wide variety of scenarios.

Chapter 9

Discussion

In this chapter, we begin by summarising the main contributions of this thesis (Section 9.1). We then make suggestions for future research (Section 9.2), before a final conclusion (Section 9.3).

Summary and discussion of thesis 9.1

The contributions of this thesis can be split into five broad headings. Firstly, we have reviewed the current literature on methods for population adjustment and related problems, set out their properties and assumptions, and reviewed applications in the published literature and in submissions to NICE. Motivated by these reviews, we then proposed a new method for population-adjusted indirect comparisons and network meta-regression combining IPD and AgD, called Multilevel Network Meta-Regression, which aims to address the issues with current approaches. We implemented ML-NMR using a general method for numerical integration and efficient computational methods in Stan, before applying ML-NMR to a real example of plaque psoriasis treatments with binary outcomes. We then extended the ML-NMR framework to handle general individual-level likelihoods, where the corresponding aggregate-level likelihood may not have a known form, and applied this to an artificial example of survival outcomes and to the plaque psoriasis example with ordered categorical outcomes. Finally, we investigated the performance of ML-NMR alongside current methods in a comprehensive simulation study.

Review of population adjustment methods and surrounding literature 9.1.1

This thesis began with a review of the literature on methods for population adjustment and closely related problems (Chapter 2). Population adjustment

describes a general problem in which we wish to estimate the relative effects of a set of treatments from a collection of studies, adjusting for differences in baseline characteristics between studies, but where IPD are only available in a subset of studies. In the simplest two-study scenario, outcomes from an IPD AB trial are adjusted into the population of an AgD AC trial; a population-adjusted indirect comparison is then formed between treatments B and C in the AC trial population (Section 2.2). If there is a common comparator A arm between the IPD and AgD studies, then an anchored indirect comparison is formed using the relative effects. This is the same as the standard indirect comparison described by Bucher et al. (1997), except that differences in effect modifying covariates are adjusted for. If there is no common comparator arm, then we are forced instead to form an unanchored indirect comparison using the absolute effects on treatments B and C. In this case, all prognostic and effect modifying covariates are required to be adjusted for in order to produce an unbiased comparison—a very strong assumption, which is very hard to verify.

The most widely used of the current methods, MAIC (Ishak et al. 2015; Signorovitch et al. 2010) and STC (Caro and Ishak 2010; Ishak et al. 2015), are based on ideas of reweighting and regression adjustment, respectively, that date back several decades and have been applied to the closely related problems of standardisation, generalisation, and calibration (Section 2.1). They are both designed with the simple two-study scenario in mind, and are not readily generalisable to larger networks of studies and treatments. They are also limited to producing estimates in the AC trial population without further assumptions, which may not match the target population for a decision. Moreover, as a reweighting method MAIC cannot extrapolate, and requires sufficient overlap between the IPD and AgD study populations. This severely limits the range of scenarios in which MAIC can produce a valid comparison, a fact borne out by the simulation study (Chapter 8).

Other approaches based on network meta-regression have also been proposed (see Section 2.2.3; Donegan et al. 2013; Jackson et al. 2006, 2008; Saramago et al. 2012; Sutton et al. 2008; Thom et al. 2015). These specify a model at both the individual level, used to fit individuals in the IPD, and at the aggregate level, used to fit the summary outcomes from the AgD. Several of these approaches define the same model at the individual and aggregate levels, and simply "plug in" mean covariate values from the AgD studies (Donegan et al. 2013; Saramago et al. 2012; Sutton et al. 2008; Thom et al. 2015). However, this incurs aggregation bias when the link function is not

the identity function (Berlin et al. 2002; Rothman et al. 2008). Several authors propose splitting the estimation of EM interaction effects into between- and within-study interactions (Donegan et al. 2013; Hua et al. 2016; Riley et al. 2008; Riley and Steyerberg 2010; Saramago et al. 2012; Thom et al. 2015) in order to account for aggregation bias and investigate potential ecological bias, but these models are not identifiable in small networks due to the additional parameters. An alternative approach derives from the ecological inference literature Jackson et al. (2006, 2008), where the aggregate-level model is an integration of the individual-level model over the aggregate study population (Jansen 2012). This avoids aggregation bias by correctly relating the individual and aggregate levels of the model through integration, and the model scales naturally to networks of all sizes. These models had previously only been derived for the special case of binary covariates (Jansen 2012); however, we later proceeded to generalise this approach and develop a general framework for population-adjusted indirect comparisons and network meta-regression combining IPD and AgD.

A key outcome of this review was to set out the properties and assumptions of these methods, in light of the surrounding literature. All methods for indirect comparison and meta-analysis—population-adjusted or otherwise—rely on some form of constancy assumption (Table 2.1). This assumption means that either relative effects (for anchored indirect comparisons and connected networks) or absolute effects (for unanchored indirect comparisons) are constant across populations, with the latter being a much stronger assumption and very hard to verify. Constancy may be conditional on a set of covariates (for population-adjusted indirect comparisons and meta-regression) or assumed to hold across all populations without adjustment (for standard indirect comparisons and meta-analysis). Another key assumption, required by MAIC and STC to produce estimates in target populations other than that of the AgD trial, is the shared effect modifier assumption (Section 2.5). This means that the relative effects of a given set of active treatments are modified in the same way by the same set of effect modifying covariates. Prior to this review, these assumptions had not been set out in this context, nor had the theoretical properties of each of the respective methods been discussed in the context of population adjustment. Moreover, there was no guidance on the use of population adjustment methods in submissions to NICE, despite the increasing use of such methods in submissions. This review was published as NICE Decision Support Unit Technical Support Document 18 (Phillippo et al. 2016; in abridged form as Phillippo et al. 2018a), alongside additional

guidance on the use of population adjustment methods in submissions to NICE. Since its publication, Technical Support Document 18 has been cited in every technology appraisal in which population adjustment methods were employed (Section 3.2).

In light of these assumptions and properties, we then reviewed the applications of MAIC and STC in the published literature (Section 3.1), and applications of all forms of population adjustment in NICE Technology Appraisal submissions (Section 3.2). These reviews revealed several deficiencies in current practice, which have thus far limited the usefulness of population-adjusted analyses for decision making. These include the prevalence of unanchored analyses that rely on very strong assumptions that are very hard to justify, the inability to produce estimates for the relevant decision target population, and insufficient justification for variable selection or effect modifier status.

These reviews also highlighted limitations of the current methods, motivating the development of a new method, multilevel network meta-regression.

9.1.2 Development of multilevel network meta-regression

In Chapter 4, we proposed a new method for population-adjusted indirect comparisons and network meta-regression combining IPD and AgD, which we call multilevel network meta-regression. There are several key advantages to ML-NMR, particularly in comparison with methods such as MAIC (Ishak et al. 2015; Signorovitch et al. 2010), STC (Caro and Ishak 2010; Ishak et al. 2015), or other network meta-regression based approaches (Donegan et al. 2013; Jackson et al. 2006, 2008; Saramago et al. 2012; Sutton et al. 2008; Thom et al. 2015).

Firstly, ML-NMR is applicable to treatment networks of any size, allowing use of all available information. Unlike other meta-regression approaches, the model remains identifiable in the two-study scenario, whilst avoiding aggregation bias. The approach of Jackson et al. (2006, 2008) (on which ML-NMR is based) also enjoys this property, however the model has only been derived for binary covariates. Being an extension of IPD NMR, ML-NMR scales naturally to larger networks, unlike MAIC or STC. Crucially for decision making, comparisons may be provided in any target population given sufficient information on the covariate distribution. MAIC and STC require the shared EM assumption to produce estimates for populations other than that represented by the AgD trial (Section 2.5). ML-NMR also uses this assumption to identify the model in small networks (Section 4.5.2), but when

larger networks of studies and treatments are available ML-NMR offers the possibility of assessing and relaxing this assumption (Section 4.6.1).

All population adjustment methods rely on some form of constancy assumption; ML-NMR (along with anchored MAIC and STC, and other meta-regression based approaches) relies on conditional constancy of relative effects, which means that all effect modifiers have been suitably adjusted for in the model. The presence of residual heterogeneity or inconsistency indicates a failure of conditional constancy of relative effects, since this means that there are factors that affect the relative effects that have not been suitably accounted for. Possible causes include unobserved effect modifiers or a misspecified model (e.g. using an incorrect functional form of a covariate, or using the shared EM assumption when it is invalid). It is therefore important to use random effects models to assess heterogeneity and node splitting or unrelated mean effects models to assess inconsistency.

Efficient computation of ML-NMR models 9.1.3

In Chapter 5, we described several techniques for the efficient implementation of ML-NMR in Stan, including transformation of covariates (Section 5.2.1) and choosing an efficient parameterisation of hierarchical structures (Sections 5.2.2 and 5.2.3) such as random effects or exchangeable EM interactions. In the extreme case, we must choose a suitable model parameterisation to avoid pathological sampling behaviour and biased posterior estimates. These concepts are equally applicable to implementations of AgD NMA and IPD NMR in Stan, and also to implementations of this family of models in other Bayesian software.

The use of QMC integration to obtain the aggregate-level model is a flexible and widely applicable approach, which can be used regardless of the number or form of covariates or the complexity of the model (Sections 4.3.3.2 and 5.1). Using quasi-random sequences for numerical integration improves upon the convergence rates of standard pseudo-random Monte Carlo integration, and retains this performance in high-dimensions. We have used a Gaussian copula to impart a given correlation structure on the integration points, whilst allowing free choice of the form of the marginal distributions; other methods such as MAIC or STC ignore this correlation structure (Section 2.3.4). Moreover, in Section 4.3.2 we showed (for multivariate Normal covariates) that correlations between covariates are implicated in aggregation bias. Specifically, when AgD are incorporated in a meta-regression with a non-identity link function by "plugging in" mean covariate values, the amount of aggregation

bias depends on the correlation between covariates, the strength of effect modification, and the population variance of the effect modifying covariates. Whilst we have found the use of a Gaussian copula to be very flexible in accounting for observed relationships between covariates, other copulæ could be used to assert relationship structures under different assumptions (Nelsen 2006). In comparison with MAIC and STC, this approach does require greater effort in specifying the form of the marginal distributions and correlation structure. Moreover, the marginal distributions and correlations may have been specified incorrectly; however, ML-NMR was not seen to be sensitive to this misspecification in the simulation study (Section 8.2.6). A limitation of this approach is that uncertainty in specifying the form of the marginal covariate distributions and the correlation structure is not accounted for.

9.1.4 Extension of ML-NMR to general likelihoods

The ML-NMR approach described in Chapter 4 relies on the form of the aggregate-level likelihood being known (Section 4.2), however this may not always be possible. Most notably, the form of the aggregate-level likelihood is unknown when a survival or time-to-event outcome is of interest—as is the case for the large majority of the applications of population adjustment methods to date (Chapter 3). The extension of ML-NMR to handle general likelihoods in Chapter 7 is thus well-motivated. As before, we begin by fully specifying the individual-level model. However, the aggregate-level model is then fitted by considering the likelihood contributions from the AgD, which are obtained directly using numerical integration without the need to specify the form of the aggregate-level likelihood. This approach greatly expands the range of scenarios in which ML-NMR models can be fitted. When the aggregate data consist of individual outcomes and summary covariate information (such as survival data reconstructed from Kaplan-Meier curves), the method is fully general: the individual-level model can take any form, and the aggregate-level model can always be fitted via numerical integration. However, when the aggregate data consist of summary outcomes and summary covariate information, this approach is only applicable when the aggregate marginal likelihood contributions can be expressed in terms of the summary outcomes. This is only straightforward for discrete outcomes. This would appear to limit the generality of the approach for continuous outcomes; however, the aggregate-level likelihood has a known closed form for many continuous individual-level likelihoods common in practice (Section 4.2).

Fitting models where part of the model has no explicit closed form presents

new challenges, in particular for model comparison. Whilst the fit of data points under a given model can always be investigated using the deviance (-2 times the log likelihood), computation of standard model comparison statistics such as the DIC (Spiegelhalter et al. 2002) is complicated when the aggregate-level likelihood has no closed form. In Section 7.2, we proposed instead to use approximate leave-one-out cross validation for model comparison, as described by Vehtari et al. (2016). This approach requires only the posterior samples of the log likelihood contributions, and so can be computed without a closed form aggregate-level likelihood.

Comprehensive simulation study 9.1.5

In Chapter 8, we undertook a comprehensive simulation study to investigate the performance of ML-NMR alongside MAIC, STC, and standard indirect comparison in a wide range of scenarios, including varying sample sizes, strength of effect modification, overlap between studies, and joint covariate distributions. As well as validating the performance of the different methods when assumptions were met, we were particularly interested in how the methods fared when assumptions were broken: namely conditional constancy of relative effects (i.e. no missing effect modifiers), the shared effect modifier assumption, validity of extrapolation beyond the IPD study population, and (for ML-NMR) correctly specifying the form of the marginal distributions and correlations between covariates in the AgD study. The simulation study focused on the two-study scenario for comparison with MAIC, STC, and standard indirect comparison; however, we expect the conclusions regarding ML-NMR to extend to larger networks also, since the underlying assumptions remain the same.

ML-NMR was seen to perform well when the requisite assumptions were met, largely eliminating the bias incurred by a standard indirect comparison and estimating standard errors well. STC performed very similarly, being a regression adjustment method also. ML-NMR additionally requires the joint covariate distribution in the AgD studies to be correctly specified; in practice this is likely to be achieved by inferring the forms of the marginal distributions and the correlations between covariates from the IPD studies (Section 4.5.1). However, the performance of ML-NMR was not found to be sensitive to this assumption, and still performed well even when marginal distributions and correlations were misspecified (Section 8.2.6).

In contrast to ML-NMR and STC, MAIC performed poorly in all simulation scenarios, and in some cases even increased the bias compared to a standard

indirect comparison. The issues with MAIC stem from the fact that it is a reweighting method, and therefore cannot extrapolate. As such, bias can only be completely removed when the population of the AgD study is entirely contained within the population of the IPD study (see Section 8.2.5). However, when the two study populations overlap entirely there is unlikely to be much bias as the imbalance in effect modifiers is likely to be small, and thus population adjustment methods may not be needed. Furthermore, this means that MAIC is only valid from at most one company's perspective of the indirect comparison—that of the company whose study (for which they have IPD) has a broader distribution of covariates. MAIC analysis from the other company's perspective (who has IPD on the other study, with a more restricted covariate distribution) will result in estimates that are biased, possibly by more than a standard indirect comparison (as well as being produced for a different target population). If the two study populations do not overlap at all, then no MAIC analyses are possible. The simulation study therefore raises serious questions regarding whether MAIC is fit for purpose. Previous simulation studies investigating MAIC have not observed this issue regarding overlap (Belger et al. 2015a,b; Hatswell et al. 2018; Leahy 2019; Signorovitch et al. 2013a). This is likely because they we not designed to vary the overlap of continuous covariates between studies, instead basing simulations on scenarios with good overlap where MAIC could work well, or because they focused on binary covariates, where issues only arise when covariate proportions are close to zero or one in the IPD study.

When an effect modifier was missing from the adjustment, and thus the conditional constancy of relative effects assumption was broken, all population adjustment methods were susceptible to bias (and as a result, under-coverage). This highlights the necessity of careful and considered selection of potential effect modifiers prior to analysis (Section 2.3.1). When all effect modifiers have been identified and included in the adjustment model, ML-NMR and STC are both robust techniques for obtaining population-adjusted indirect comparisons. The additional flexibility of ML-NMR makes this an attractive choice for population adjustment in a wide variety of scenarios.

9.2 Suggestions for future research

This thesis has motivated the need for new methods for population adjustment, described ML-NMR as a general and widely-applicable approach, and demonstrated its performance in a wide range of simulated scenarios. However, open

questions remain which warrant future research.

Investigating data requirements 9.2.1

Whilst ML-NMR can in theory be applied to networks with any number of treatments and any number of IPD and AgD studies, in practice there must be sufficient data available to estimate the model. For example, we may need to rely on the shared effect modifier assumption to identify EM interaction coefficients for some treatments where IPD studies or a sufficient number of AgD studies with a range of covariate distributions are not available (Section 4.5.2), rather than fitting exchangeable or independent EM interaction coefficients (Section 4.6.1). The simulation study showed that all population adjustment methods were sensitive to the validity of the shared EM assumption (Section 8.2.3), and thus it is preferable not to rely on the shared EM assumption unless supported by clinical or biological reasoning (Section 2.5). Even then, it is desirable to assess and relax this assumption. Future simulation studies could be undertaken to determine the data requirements for fitting models with exchangeable or independent EM interactions. Such simulation studies could investigate the feasibility and performance of shared and independent EM interaction models with different numbers of IPD and AgD studies, different numbers of treatments in a class with exchangeable EM interactions, and with varying ranges of covariate values between the studies. This would give analysts using ML-NMR better intuition for the data requirements of such models, which may help identify potential issues prior to analysis—perhaps giving the opportunity to pursue greater availability of IPD, or identify more informative prior distributions (here for the EM interaction coefficients and the treatment effects) to supplement weaker data.

Split between- and within-study interactions 9.2.2

The ML-NMR models that we have considered involve an individual-level model where there is a single treatment-covariate interaction term $\boldsymbol{\beta}_{2,k}$, as in equation (4.34b):

$$g(\theta_{ijk}) = \mu_j + \boldsymbol{x}_{ijk}^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k.$$

As we have described in Sections 2.2.3 and 4.7, several authors (Donegan et al. 2013; Hua et al. 2016; Riley et al. 2008; Riley and Steyerberg 2010; Saramago et al. 2012) have suggested to instead split the interaction term into a within-study interaction $\boldsymbol{\beta}_{2,k}^{(w)}$ and a between-study interaction $\boldsymbol{\beta}_{2,k}^{(b)}$, as in equation (4.51):

$$g(\theta_{ijk}) = \mu_j + (\boldsymbol{x}_{ijk} - \bar{\boldsymbol{x}}_j)^{\mathsf{T}}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}^{(w)}) + \bar{\boldsymbol{x}}_j^{\mathsf{T}}\boldsymbol{\beta}_{2,k}^{(b)} + \gamma_k.$$

The motivation is that combining information on interactions from within and between studies can result in ecological bias due to unobserved effect modifiers, and splitting the interaction term in this manner means that $\boldsymbol{\beta}_{2,k}^{(w)}$ should be free from this bias (Hua et al. 2016). Several authors (Donegan et al. 2013; Riley et al. 2008; Riley and Steyerberg 2010; Saramago et al. 2012) use this split interaction model to incorporate AgD studies, where the AgD studies only contribute to the between-study interaction $\boldsymbol{\beta}_{2,k}^{(b)}$ and mean covariate values are "plugged in", as in equation (4.52):

$$g(\theta_{\bullet jk}) = \mu_j + \bar{\boldsymbol{x}}_j^{\mathsf{T}} \boldsymbol{\beta}_{2,k}^{(b)} + \gamma_k.$$

In this case, when $g(\cdot)$ is not the identity function differences between $\boldsymbol{\beta}_{2,k}^{(w)}$ and $\boldsymbol{\beta}_{2,k}^{(b)}$ are additionally due to aggregation bias (Greenland 1992).

ML-NMR avoids aggregation bias since the aggregate-level model is appropriately related to the individual-level model through integration over the covariate distribution in the AgD studies. However, unobserved EMs are still a concern—whether these are individual-level covariates, or study-level differences in treatment regimens or outcome definitions for example. We have not split the interaction terms in the ML-NMR model: we assume that there are no unobserved EMs so that the conditional constancy of relative effects assumption holds, which is required in order to produce unbiased estimates of population-adjusted relative effects.

The ML-NMR model could be modified to include split interaction terms and assess the conditional constancy of relative effects assumption in this manner, however this results in a non-identifiable model in the two-study scenario, and likely requires substantial amounts of data to estimate well. Instead, we have proposed assessing the conditional constancy of relative effects assumption by investigating residual heterogeneity using a random effects model (Section 4.6.2) and inconsistency using unrelated mean effects or node-splitting models (Section 4.6.3), which have lesser data requirements than splitting EM interactions. The presence of either residual heterogeneity or inconsistency indicates a failure of the conditional constancy of relative effects assumption, due to unobserved EMs or other model misspecification (such as an invalid shared EM assumption). Future research could investigate the relative performance of fitting ML-NMR models with split interaction terms versus checking for residual heterogeneity or inconsistency as methods for assessing the conditional constancy of relative effects assumption. For example, it remains to be seen—given sufficient data for either approach—which would be more powerful and better able to detect failings in this crucial assumption.

Including contrast-based aggregate data 9.2.3

The ML-NMR model that we have developed is based on aggregate data being available in an arm-based format, for example event counts or mean outcomes on each treatment arm in an AgD study. However, it is also common for aggregate data to be available in a contrast-based format as a summary relative effect measure between two treatment arms, such as log odds ratios or a difference in means. In this format, studies with two treatment arms report a single relative effect; studies with three treatment arms report two relative effects (which are correlated, by virtue of a shared reference arm), and so on.

In this case, the usual approach (as described in Section 1.2.8 for AgD NMA; Dias et al. 2011c; Salanti et al. 2007; Woods et al. 2010) is to use a multivariate Normal likelihood, with marginal distributions and covariances

$$y_{\bullet jab} \sim \mathrm{N}(\theta_{\bullet jab}, s^2_{jab}) \tag{9.1a}$$

$$\mathrm{cov}(y_{\bullet jab}, y_{\bullet jac}) = v^{(a)}_{jbc}, \tag{9.1b}$$

where $y_{\bullet jab}$ is the summary relative effect of treatment b vs. treatment a in study j, with standard error s_{jab}. For studies with three or more treatment arms, the relative effects $y_{\bullet jab}$ and $y_{\bullet jac}$ for any two treatments b and c both compared to a have covariance $v^{(a)}_{jbc}$. If the covariance $v^{(a)}_{jbc}$ is not reported but the standard error s_{jbc} of the c vs. b relative effect y_{jbc} is available, then the covariance can be derived using the relation

$$v^{(a)}_{jbc} = \frac{s^2_{jab} + s^2_{jac} - s^2_{jbc}}{2}. \tag{9.2}$$

In cases where the relative effects are simple differences between mean outcomes on each treatment, the covariance $v^{(a)}_{jbc}$ is equal to the variance of the outcome in the reference treatment a arm (Dias et al. 2011c). When the reference treatment arm variance is not reported, it may be possible to impute from those reported in other trials (Dakin et al. 2011). When a study reports only a single relative effect, the likelihood is univariate Normal given by (9.1a). In the context of ML-NMR, the expected marginal relative effect $\theta_{\bullet jab}$ is modelled with an identity link function; thus the usual integration of the individual-level model over the covariate distribution reduces to "plugging in" mean covariate values $\bar{\boldsymbol{x}}_j$ like so

$$\begin{aligned} \theta_{\bullet jab} &= \eta_{jb}(\bar{\boldsymbol{x}}_j) - \eta_{ja}(\bar{\boldsymbol{x}}_j) \\ &= \bar{\boldsymbol{x}}^\intercal(\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}) + \gamma_b - \gamma_a, \end{aligned} \tag{9.3}$$

where $\eta_{jk}(\cdot)$ is the linear predictor described previously (Chapter 4):

$$\eta_{jk}(\boldsymbol{x}) = \mu_j + \boldsymbol{x}^\intercal(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}) + \gamma_k. \tag{9.4}$$

This approach, whilst widely used in the general meta-analysis literature, relies upon the assumption of (multivariate) Normality. This is typically justified asymptotically by the Central Limit Theorem, but may not be appropriate in practice—particularly for the results of smaller studies. An alternative approach, requiring further investigation, would be to follow the methods described in Chapter 7 and attempt to derive the aggregate marginal likelihood contributions for these data (Section 7.1). The ML-NMR model for general likelihoods, written in terms of the likelihood contributions from the different levels of the model, was given in equation (7.3). After specifying an individual-level model, defining the individual conditional likelihood contributions given the covariates, the individual marginal likelihood contributions were obtained by integration over the covariate distribution. With summary relative effects data, the aim would be to write the product of these individual marginal likelihood contributions over the individuals in both treatment a and b arms in terms of the summary relative effect $y_{\bullet jab}$. This would avoid the need for a Normality assumption, and instead fit the summary relative effects using their exact likelihood contributions. However, it remains to be seen whether the required derivations are analytically tractable.

9.2.4 Including results from subgroup analyses and reported regression coefficients

As discussed in Section 9.2.1, the most demanding data requirements for ML-NMR often involve the estimation of EM interactions. To estimate these parameters for a given treatment we require either: IPD studies involving the treatment, multiple AgD studies involving the treatment with sufficiently different covariate distributions, or an identifying assumption such as the shared EM assumption. These requirements may be difficult to meet: the acquisition of IPD is often non-trivial, there are only a certain number of relevant trials reported in the literature, and the shared EM assumption may not always be appropriate. Therefore, any additional sources of information on EM interactions are valuable and likely to aid estimation of ML-NMR models. Potential sources of information include the results of subgroup analyses, or reported regression coefficients.

When results of subgroup analyses are available, the one possible solution is to translate any difference in treatment effects between subgroups into an informative prior for the corresponding EM interaction parameter. However, this is not an exact approach, and may not be the most efficient use of such information. Alternatively, the aggregate-level model can be modified to

directly incorporate the subgroup outcomes; in other words, rather than the usual aggregate-level model for a summary outcome $y_{\bullet jk}$, we would fit an aggregate-level model to each summary outcome $y_{\bullet jk;z}$ in a set of $z = 1, \ldots, Z_j$ subgroups. The subgroup variables themselves may be discrete covariates, or continuous covariates that have been split into categories.

If outcomes $y_{\bullet jk;z}$ are reported in a factorial fashion for each distinct combination of subgroup variables, then the aggregate-level model becomes

$$y_{\bullet jk;z} \sim \pi_{\text{Agg}}(\theta_{\bullet jk;z}) \tag{9.5a}$$

$$\theta_{\bullet jk;z} = \int_{\mathfrak{x}_z} g^{-1}\big(\eta_{jk}(\boldsymbol{x})\big) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x}, \tag{9.5b}$$

where $\mathfrak{x}_z$ is now the support of the covariates in subgroup z. The Quasi Monte-Carlo integration approach (Sections 4.3.3.2 and 5.1) can be modified to fit this aggregate-level model. Firstly, $\tilde{N}$ integration points $\tilde{\boldsymbol{x}}_{jk}$ are generated from the full covariate distribution $f_{jk}(\cdot)$ (Section 5.1). Then these integration points are split into each discrete subgroup, labelled as $\tilde{\boldsymbol{x}}_{jk;z}$. Finally, the integration in (9.5b) is carried out by summation over each set of integration points:

$$\int_{\mathfrak{x}_z} g^{-1}\big(\eta_{jk}(\boldsymbol{x})\big) f_{jk}(\boldsymbol{x})\, d\boldsymbol{x} \approx \frac{1}{\tilde{N}_z} \sum_{\tilde{\boldsymbol{x}}_{jk;z}} g^{-1}\big(\eta_{jk}(\tilde{\boldsymbol{x}}_{jk;z})\big). \tag{9.6}$$

The practical properties of this approach require further investigation. For example, it is likely that a greater number of integration points will be required compared to an analysis without subgroups, since the integration points are being split amongst Z_j discrete subgroups. Furthermore, this approach may run in to difficulties if certain subgroups are narrowly-defined or specified at extreme covariate values, since the number of integration points in such subgroups may become too small. Such issues become increasingly likely as the number of subgroup variables increases.

If outcomes are not reported for each discrete combination of subgroup variables, but are instead reported marginally by each subgroup variable, then the above approach is not appropriate since the marginal subgroup outcomes are correlated. It is not immediately apparent how such correlations can be calculated, nor how the correlated subgroup outcomes should be modelled.

When reported estimates of regression coefficients are available, along with their standard errors (and preferably their correlations), these may simply be incorporated into informative prior distributions for the corresponding model parameters. A further potential source of information on model parameters is expert opinion, from which informative prior distributions may be elicited.

An important consideration when including information from either subgroup analyses or reported regression coefficients is the possibility of reporting bias. It is more likely that ad hoc subgroup analyses are reported if they are "significant" or demonstrate a large difference in treatment effects—even if the observed effect modification is entirely chance—and thus any reported ad hoc subgroup analyses are likely to be biased towards larger estimates of effect modification (Brookes et al. 2004; Hahn et al. 2000). There should therefore be a strong preference for using pre-specified subgroup analyses to inform ML-NMR, in an attempt to avoid such bias.

9.2.5 Addressing limitations in included studies

Throughout this thesis, we have assumed that the studies included in the analysis are internally valid; that is, they produce unbiased estimates of relative treatment effects within their respective sample populations. Issues with internal validity are commonly highlighted using a structured checklist, such as the Cochrane Risk of Bias tool (Higgins et al. 2011, 2016).

When IPD are available, it may be possible to account for certain issues of internal validity through an appropriate method of analysis. For example, non-compliance issues are commonly dealt with using instrumental variable methods to obtain complier average causal effects (DiazOrdaz et al. 2018; Imbens and Rubin 1997). Several methods to account for treatment switching in the analysis of survival or time-to-event outcomes are available, including rank-preserving structural failure time models (Robins and Tsiatis 1991), iterative parameter estimation (Branson and Whitehead 2002), and inverse probability of censoring weighting (Robins and Finkelstein 2000). Future research could investigate incorporating such methods into ML-NMR analyses.

Missing covariate and outcome data are also likely to be encountered within the IPD. Multiple imputation is a widely-used method of dealing with missing data (Kenward and Carpenter 2007; Little and Rubin 2002), to which the Bayesian framework is well-suited. Missing data can be imputed at every iteration of the MCMC sampler, incorporating the uncertainty arising from the missing values into the posterior distribution (Jackson et al. 2009; Mason et al. 2012). Such approaches have previously been described for IPD NMA (Quartagno and Carpenter 2016), and apply similarly to ML-NMR by extension.

However, when such issues are present within the AgD studies there may be little that can be done. If an adjustment was performed in the original analysis of the study and subsequently published, then the adjusted summary

relative effect estimates could instead be incorporated into the model following the approach described in Section 9.2.3. Otherwise, ML-NMR (as with AgD NMA) can only note where biases may be present in the included AgD studies. In a decision making context, threshold analysis is a form of sensitivity analysis that aims to determine whether biases in the included evidence could plausibly alter the treatment decision—without requiring the estimation of or adjustment for any such biases—and may also be applied to the results of ML-NMR analyses (Caldwell et al. 2016; Phillippo et al. 2018b, 2019b).

Model fit and model comparison 9.2.6

When fitting NMA models in a Bayesian framework, model fit is often assessed using the residual deviance and different models are often compared using the DIC (Dias et al. 2011c; Spiegelhalter et al. 2002), as described in Section 1.2.6. We can follow this approach for ML-NMR models also, as long as the form of the aggregate likelihood is known (Section 5.3).

However, when using the two-parameter Binomial aggregate-level likelihood to model summary count data, there are some additional complications. These stem from the fact that the likelihood has more than one parameter—the adjusted mean probability $\bar{p}'_{jk}$ and the adjusted number of individuals N'_{jk}—and both are modelled parameters. Firstly, calculating the effective number of parameters p_D requires evaluating the residual deviance at the fitted values. In likelihoods with more than one parameter such as the two-parameter Binomial, the fitted values do not fully determine the values of the parameters. In this case, we need to plug in a suitable value of N'_{jk}; we chose the posterior median, but other choices could be made such as the unadjusted sample size N_{jk}. Secondly, when calculating the saturated deviance, the two modelled likelihood parameters are not uniquely defined at a given fitted value $\hat{y}_{\bullet jk}$; any combination of N'_{jk} and $\bar{p}'_{jk}$ could be chosen to produce a given value of $\hat{y}_{\bullet jk} = N'_{jk}\bar{p}'_{jk}$. We chose to calculate the saturated deviance under a model where $\bar{p}'_{jk}$ perfectly predicts the observed number of events at each posterior sample of N'_{jk}, but other choices include fixing $N'_{jk} = N_{jk}$. This second issue can be avoided by instead using the deviance (i.e. without subtracting the saturated deviance to obtain the residual deviance) to calculate DIC and p_D; however, whilst this does not affect comparisons between candidate models this does inhibit checking absolute model fit, particularly between the IPD and AgD contributions to the model since these will be scaled differently (Section 5.3.1.2). It remains to be seen which approaches to calculating the effective number of parameters and the residual deviance are theoretically or

practically most desirable. However, these choices only relate to the aggregate part of the model, and since the overall model fit and DIC will be dominated by the individual-level model and its fit to the IPD, the choice is unlikely to make much practical difference.

When the form of the aggregate-level likelihood is unknown, as is the case when using the generalised approach in Chapter 7 that calculates the likelihood contributions directly, we need to take an alternative approach to assessing model fit and model comparison. In Section 7.2, we suggested using approximate leave-one-out cross validation instead, following Vehtari et al. (2016) to use Pareto-smoothed importance sampling to approximate the expected log pointwise predictive density. PSIS-LOO requires only the posterior samples of the log likelihood, and can thus be calculated without explicit knowledge of the form of the likelihood. This approach works well when the AgD are in the form of individual outcomes and summary covariate distributions, as is the case for survival data reconstructed from Kaplan-Meier curves, and we successfully used PSIS-LOO to perform model comparison for the artificial survival example in Section 7.3. However, the approximation used by PSIS-LOO breaks down when there are data points that are highly influential; this is the case when the AgD are summary outcomes and there are only a small number of AgD studies per treatment. In this case, each aggregate data point is highly influential, and may even be "saturated" in the model if it is the only data point informing a parameter (e.g. a treatment effect). We suggested some solutions to this issue in Section 7.2, including basing model comparison only on the IPD, or modifying the PSIS-LOO approach to ignore data points that are saturated in the model (since these data points will always be perfectly fit, under any model). The latter solution is preferable, since the fit to AgD studies which are not saturated in the model is taken into account. However, at present this requires manually selecting the data points to ignore from the PSIS-LOO calculation, which is arduous—particularly in larger networks. It should be possible to automatically determine the data points that are saturated in the model and exclude these from the PSIS-LOO calculation, making this approach more practically appealing; this is an area for further research. Indeed, once this practical limitation has been addressed, PSIS-LOO may be readily applied to standard AgD NMA, where the same issue is frequently encountered, as well as IPD NMR or ML-NMR. This is an attractive proposition, since PSIS-LOO has several benefits over DIC: it is a better measure of future predictive performance; it is fully Bayesian, accounting for the full posterior distribution rather than relying on plug-in

posterior estimates; and standard errors facilitate judgement of differences when comparing models (see Section 7.2; Gelman et al. 2013b; Vehtari et al. 2016).

Incorporating data from single-arm or observational studies 9.2.7

The methods developed in this thesis are focused on scenarios where a connected network of randomised controlled trials is present. However, as we have identified in our review of the applied literature and submissions to NICE (Chapter 3), the majority of applications of population adjustment methods such as MAIC and STC to date have involved single-arm or observational studies in disconnected networks. An important area for future research is therefore the extension of ML-NMR to incorporate data from single-arm or observational studies.

One possible approach is to place a random effect distribution on the study-specific baselines μ_j (Dias et al. 2013c; Li and Begg 1994; Thom et al. 2015). This is straightforward under the reference treatment parameterisation (Section 1.2.2), since the baseline parameters all refer to the reference treatment 1. Data from single-arm studies are then incorporated by comparison with the predicted absolute effects on the reference treatment, using the same individual- or aggregate-level model as for the randomised studies in equation (4.34). The assumption here is that absolute treatment effects can be predicted on average given the covariates, and that any remaining differences in absolute outcomes between studies are random. This is a very strong assumption, requiring that there are no systematic differences in prognostic or effect modifying covariates between study populations. Moreover, when this assumption does not hold, biases can be introduced across the treatment network—including for comparisons only informed by randomised studies (placing random effects on the study-specific baselines interferes with randomisation, see Dias et al. 2013c; Senn 2010). Unanchored MAIC or STC make an even stronger assumption (conditional constancy of absolute effects, see Section 2.3.2), requiring that absolute treatment effects can be predicted exactly given the covariates; this would be equivalent to a single fixed baseline μ across studies.

If comparative observational data are available (i.e. including more than one treatment, and thus informing relative rather than absolute effects) then other approaches may be taken (Schmitz et al. 2013). Observational data may be used to create informative prior distributions, potentially down-weighted using a "power prior" to reflect additional uncertainty in the evidence (Prevost et al. 2000). Another approach is to assume that relative effects are exchangeable

between different study types using a hierarchical model (Prevost et al. 2000; Schmitz et al. 2013); relative effects can then be compared between different study types, and it is also possible to incorporate bias adjustment and to allow for additional uncertainty in observational evidence by down-weighting.

Importantly, including single-arm or observational data in unanchored indirect comparisons or in larger networks requires very strong assumptions (Section 2.3.2). Such analyses are susceptible to residual bias due to unmeasured confounding. It is therefore crucial to assess residual bias if such analyses are to be useful for decision-making.

9.2.8 Quantifying residual bias due to unmeasured confounding in unanchored indirect comparisons

Since they respect within-study randomisation, anchored indirect comparisons are unaffected by differences in prognostic variables between studies, and need only to account for differences in effect modifying variables. Conversely, unanchored indirect comparisons do not rely on within-study randomisation, and are therefore susceptible to large amounts of systematic error unless all prognostic variables and effect modifiers are accounted for. If unanchored indirect comparisons are to be used, it is therefore necessary to attempt to quantify the possible extent of any residual systematic error resulting from unobserved prognostic variables and effect modifiers.

The simplest way to quantify residual bias is by comparing observed and predicted outcomes on the treatment of interest in a range of different studies in the target population; however, this might not be a viable option if there are no such studies available. It should be noted however that unobserved covariates are only one source of heterogeneity between studies (and bias in an ensuing indirect comparison); for example, differences in study design and conduct will also introduce heterogeneity, but cannot be accounted for with population adjustment methods. The way in which residual systematic error is quantified is therefore an area that requires further research. We explore some initial suggestions here (also published in Phillippo et al. 2016), focusing on the two-study unanchored indirect comparison scenario, although the ideas generalise to larger networks.

9.2.8.1 Out-of-sample methods

One possible method for quantifying the possible extent of systematic bias present in a unanchored indirect comparison, comparing a treatment B in an IPD study with a treatment C in an AgD study, involves comparing the

heterogeneity observed in a set of studies in the target population to the heterogeneity in predictions from the model.

Firstly, a set of external studies in the target population with aggregate data on the relevant outcome is identified; these need not all be on the same treatment. A random-effects pooling across absolute outcomes on study arms in the target population, controlling for treatment, is then carried out. Predicted outcomes on treatment B in each of the study arms used in the pooling can be obtained using a population adjustment method, and a similar pooling performed. If all prognostic variables and effect modifiers are accounted for, then the between-studies variation of the predicted outcomes, say τ_*^2, will match that of the observed outcomes τ^2 (that is, residual variation will be minimised). Conversely, lower between-studies variation of predicted outcomes would be expected if some prognostic variables and/or effect modifiers remain unaccounted for. The ratio of the between-studies variance in predicted to observed outcomes, τ_*^2/τ^2, could be interpreted as the proportion of systematic error "explained" by the included covariates. It is likely that, in practice, limited study data will be available, and therefore the estimation of between-study variance may be difficult.

Methods based on between-study variance should be underpinned by a protocol-driven systematic review to prevent selection of an overly homogeneous sample of studies for inclusion. Likewise, a company's own trials would be expected to be more homogeneous than a wider selection of trials in the target population.

In-sample methods 9.2.8.2

Other approaches for quantifying the systematic error in unanchored comparisons are possible. For example, if regression-based approaches such as STC or ML-NMR are used, cross-validation methods (e.g. Picard and Cook 1984) enable the estimation of the prediction error in the outcome model, which is largely due to missing prognostic variables and/or effect modifiers. K-fold cross-validation is a frequently used method, in which the IPD are split into K equal-sized sets. Each of the K sets is omitted in turn at the model fitting stage, and used as a validation set to check the model predictions. Prediction error may then be averaged over the K sets. A value of $K = 10$ is often used, although the choice of K should be based on the situation at hand, particularly with reference to the available sample size, as there is a bias-variance trade-off. If regression-based approaches such as STC or ML-NMR are used, R^2 values may be used to assess the predictive performance of the model;

R^2 may be interpreted as the proportion of variance explained by the model, similarly to the between-studies variance ratio described above (e.g. Xu 2003). (A Bayesian approach to calculating R^2 is described by Gelman et al. (2018), which would be appropriate for ML-NMR in a Bayesian framework.) A general disadvantage in our context of cross-validation, R^2, and other "in-sample" methods for checking predictive accuracy (as opposed to the "out-of-sample" methods above), is that the individuals in the IPD trial are likely to be more homogeneous than those of the wider target population, thus leading to overconfidence in the abilities of population adjustment methods to predict outcomes in the target population. In-sample methods in general will most likely underestimate the true amount of residual variation.

9.2.9 An R package for ML-NMR

Stan code for implementing ML-NMR is provided in Appendix A. The code is modular and designed to be easily repurposed: it is straightforward to substitute in different likelihoods and link functions to model other outcomes. However, a working knowledge of Stan and R is still required. We therefore plan to develop an R package that implements the ML-NMR models in a more user-friendly manner. Since ML-NMR is an extension of the standard NMA framework, with AgD NMA and IPD NMR as special cases (Section 4.6), such an R package would also enable a broad range of network meta-analyses with any level of AgD and IPD availability. The package would also simplify the generation of numerical integration points (Section 5.1), support checking of MCMC convergence and numerical integration error, calculate model fit and model comparison statistics (Section 5.3), and facilitate the production of estimates of quantities of interest in different target populations (Section 4.4). Such a package would streamline the process of implementing ML-NMR models, and make the methods accessible to a wider range of users.

9.3 Conclusion

In this thesis, we have reviewed the literature on population adjustment methods and their applications; proposed ML-NMR as a new and general method for population-adjusted indirect comparisons and network meta-regression combining IPD and AgD; provided efficient computational methods for implementing ML-NMR; and investigated the performance of ML-NMR alongside current methods in a comprehensive simulation study. Whilst we have been motivated by the context of population adjustment in network

meta-analysis and indirect comparisons of randomised controlled trials, with a particular focus on health technology assessment and appraisal, there is potential for impact in the wider literature. The problem of synthesising information reported at different levels of aggregation is common to many areas, and the methods developed in this thesis are likely to be applicable to this general problem in a variety of contexts. Of particular note is the ecological inference literature, where integration of an individual-level model over a population has thus far been pursued algebraically (Jackson et al. 2006, 2008; Salway and Wakefield 2005), and which inspired the first applications to network meta-regression (Jansen 2012). As we have discussed in this final chapter, several areas for future research still remain, which would further support the practical implementation of ML-NMR and its extension to applications beyond the scenarios considered here.

Bibliography

Abo-Zaid, G., B. Guo, J. J. Deeks, T. P. A. Debray, E. W. Steyerberg, K. G. M. Moons, R. D. Riley, H. Bang, and J. M. Robins (2013). "Individual participant data meta-analyses should not ignore clustering". In: *Journal of Clinical Epidemiology* 66.8, pp. 865–873. DOI: 10.1016/j.jclinepi.2012.12.017.

Ades, A. E. (2003). "A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence". In: *Statistics in Medicine* 22.19, pp. 2995–3016. DOI: 10.1002/sim.1566.

Bang, H. and J. M. Robins (2005). "Doubly Robust Estimation in Missing Data and Causal Inference Models". In: *Biometrics* 61.4, pp. 962–973. DOI: 10.1111/j.1541-0420.2005.00377.x.

Belger, M., A. Brnabic, Z. Kadziola, H. Petto, and D. Faries (2015a). "Alternative Weighting Approaches for Matching Adjusted Indirect Comparisons (MAIC)". In: *ISPOR 20th Annual International Meeting*. Philadelphia, PA, USA.

— (2015b). "Inclusion of Multiple Studies in Matching Adjusted Indirect Comparisons (MAIC)". In: *ISPOR 20th Annual International Meeting*. Philadelphia, PA, USA.

Bender, R., T. Augustin, and M. Blettner (2005). "Generating survival times to simulate Cox proportional hazards models". In: *Statistics in Medicine* 24.11, pp. 1713–1723. DOI: 10.1002/sim.2059.

Berlin, J. A., J. Santanna, C. H. Schmid, L. A. Szczech, and H. I. Feldman (2002). "Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head". In: *Statistics in Medicine* 21.3, pp. 371–387. DOI: 10.1002/sim.1023.

Betancourt, M. J. and M. Girolami (2013). "Hamiltonian Monte Carlo for Hierarchical Models". In: arXiv: http://arxiv.org/abs/1312.0906 [stat.ME].

Betts, K. A., M. Mittal, J. Song, M. Skup, and A. Joshi (2016). "Relative efficacy of Adalimumab versus Secukinumab in active ankylosing spondylitis:

a matching-adjusted indirect comparison". In: *European League Against Rheumatism*. Vol. 75. Suppl2. London: Ann Rheum Dis, p. 98. DOI: 10.1136/annrheumdis-2016-eular.2754.

Branson, M. and J. Whitehead (2002). "Estimating a treatment effect in survival studies in which patients switch treatment". In: *Statistics in Medicine* 21.17, pp. 2449–2463. DOI: 10.1002/sim.1219.

Brilleman, S. (2018). *simsurv: Simulate Survival Data*. R package version 0.2.2. URL: https://CRAN.R-project.org/package=simsurv.

Brookes, S. T., E. Whitely, M. Egger, G. Davey Smith, P. A. Mulheran, and T. J. Peters (2004). "Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test". In: *Journal of Clinical Epidemiology* 57.3, pp. 229–236. DOI: 10.1016/j.jclinepi.2003.08.009.

Brumback, B. and A. Berg (2008). "On effect-measure modification: relationships among changes in the relative risk, odds ratio, and risk difference". In: *Statistics in Medicine* 27.18, pp. 3453–3465. DOI: 10.1002/sim.3246.

Bucher, H. C., G. H. Guyatt, L. E. Griffith, and S. D. Walter (1997). "The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials". In: *Journal of Clinical Epidemiology* 50.6, pp. 683–691. DOI: 10.1016/s0895-4356(97)00049-8.

Caflisch, R. E. (1998). "Monte Carlo and quasi-Monte Carlo methods". In: *Acta Numerica* 7, pp. 1–49. DOI: 10.1017/S0962492900002804.

Caldwell, D. M., A. E. Ades, S. Dias, S. Watkins, T. Li, N. Taske, B. Naidoo, and N. J. Welton (2016). "A threshold analysis assessed the credibility of conclusions from network meta-analysis". In: *Journal of Clinical Epidemiology* 80, pp. 68–76. DOI: 10.1016/j.jclinepi.2016.07.003.

Caldwell, D. M., A. E. Ades, and J. P. T. Higgins (2005). "Simultaneous comparison of multiple treatments: combining direct and indirect evidence". In: *BMJ* 331.7521, pp. 897–900. DOI: 10.1136/bmj.331.7521.897.

Caldwell, D. M., N. J. Welton, S. Dias, and A. E. Ades (2012). "Selecting the best scale for measuring treatment effect in a network meta-analysis: a case study in childhood nocturnal enuresis". In: *Research Synthesis Methods* 3.2, pp. 126–141. DOI: 10.1002/jrsm.1040.

Caro, J. J. and K. J. Ishak (2010). "No Head-to-Head Trial? Simulate the Missing Arms". In: *Pharmacoeconomics* 28.10, pp. 957–967.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). "Stan: A Probabilistic

Programming Language". In: *Journal of Statistical Software* 76.1. DOI: 10.18637/jss.v076.i01.

Chang, G.-C., M.-J. Ahn, E. Wright, H. T. Kim, J.-H. Kim, J. H. Kang, S.-W. Kim, S. Sherman, and S. Walzer (2011). "Comparative effectiveness of bevacizumab plus cisplatin-based chemotherapy versus pemetrexed plus cisplatin treatment in East Asian non-squamous non-small cell lung cancer patients applying real-life outcomes". In: *Asia-Pacific Journal of Clinical Oncology* 7, pp. 34–40. DOI: 10.1111/j.1743-7563.2011.01400.x.

Christophe, D. and S. Petr (2018). *randtoolbox: Generating and Testing Random Numbers*. R package version 1.17.1.

Claxton, K., M. Sculpher, C. McCabe, A. Briggs, R. Akehurst, M. Buxton, J. Brazier, and T. O'Hagan (2005). "Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra". In: *Health Economics* 14.4, pp. 339–347. DOI: 10.1002/hec.985.

Cochran, W. G. (1968). "The effectiveness of adjustment by subclassification in removing bias in observational studies". In: *Biometrics* 24.2, pp. 295–313. DOI: 10.2307/2528036.

Collett, D. (2003). *Modelling survival data in medical research*. 2nd ed. Texts in statistical science. London: Chapman & Hall. ISBN: 1584883251.

Cox, D. R. (1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

Cox, D. R. (1984). *Analysis of survival data*. Monographs on statistics and applied probability. London: Chapman and Hall. ISBN: 041224490X.

Critchfield, G. C. and K. E. Willard (1986). "Probabilistic Analysis of Decision Trees Using Monte Carlo Simulation". In: *Medical Decision Making* 6.2, pp. 85–92. DOI: 10.1177/0272989x8600600205.

Crowther, M. J., R. D. Riley, J. A. Staessen, J. Wang, F. Gueyffier, and P. C. Lambert (2012). "Individual patient data meta-analysis of survival data using Poisson regression models". In: *BMC Medical Research Methodology* 12.1. DOI: 10.1186/1471-2288-12-34.

Dakin, H. A., N. J. Welton, A. E. Ades, S. Collins, M. Orme, and S. Kelly (2011). "Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension". In: *Statistics in Medicine* 30.20, pp. 2511–2535. DOI: 10.1002/sim.4284.

Deeks, J., J. Dinnes, R. D'Amico, A. Sowden, C. Sakarovitch, F. Song, M. Petticrew, and D. Altman (2003). "Evaluating non-randomised intervention studies". In: *Health Technology Assessment* 7.27. DOI: 10.3310/hta7270.

Deeks, J. J. (2002). "Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes". In: *Statistics in Medicine* 21.11, pp. 1575–1600. DOI: 10.1002/sim.1188.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc. DOI: 10.1002/0471728438.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer-Verlag. ISBN: 0-387-96305-7.

Dias, S. and A. E. Ades (2015). "Absolute or relative effects? Arm-based synthesis of trial data". In: *Research Synthesis Methods* 7.1, pp. 23–28. DOI: 10.1002/jrsm.1184.

Dias, S., A. J. Sutton, A. E. Ades, and N. J. Welton (2013a). "Evidence Synthesis for Decision Making 2: A Generalized Linear Modelling Framework for Pairwise and Network Meta-analysis of Randomized Controlled Trials". In: *Medical Decision Making* 33.5, pp. 607–617. DOI: 10.1177/0272989X12458724.

Dias, S., A. J. Sutton, N. J. Welton, and A. E. Ades (2011a). *NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment*. Tech. rep. National Institute for Health and Care Excellence. URL: http://www.nicedsu.org.uk.

— (2011b). *NICE DSU Technical Support Document 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis: Software choices*. Tech. rep. National Institute for Health and Care Excellence. URL: http://www.nicedsu.org.uk.

— (2013b). "Evidence Synthesis for Decision Making 6: Embedding Evidence Synthesis in Probabilistic Cost-effectiveness Analysis". In: *Medical Decision Making* 33.5, pp. 671–678. DOI: 10.1177/0272989X13487257.

Dias, S., N. J. Welton, D. M. Caldwell, and A. E. Ades (2010). "Checking consistency in mixed treatment comparison meta-analysis". In: *Statistics in Medicine* 29.7-8, pp. 932–944. DOI: 10.1002/sim.3767.

Dias, S., N. J. Welton, A. J. Sutton, and A. E. Ades (2011c). *NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials*. Tech. rep. National Institute for Health and Care Excellence. URL: http://www.nicedsu.org.uk.

— (2013c). "Evidence Synthesis for Decision Making 5: The Baseline Natural History Model". In: *Medical Decision Making* 33.5, pp. 657–670. DOI: 10.1177/0272989X13485155.

Dias, S., N. J. Welton, A. J. Sutton, D. M. Caldwell, G. B. Lu, and A. E. Ades (2013d). "Evidence Synthesis for Decision Making 4: Inconsistency in Networks of Evidence Based on Randomized Controlled Trials". In: *Medical Decision Making* 33.5, pp. 641–656. DOI: 10.1177/0272989X12455847.

Dias, S., N. J. Welton, A. J. Sutton, D. M. Caldwell, G. Lu, and A. E. Ades (2011d). *NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials*. Tech. rep. National Institute for Health and Care Excellence. URL: http://www.nicedsu.org.uk.

Dias, S., A. E. Ades, N. J. Welton, J. P. Jansen, and A. J. Sutton (2018). *Network Meta-Analysis for Decision-Making*. Statistics in Practice. John Wiley & Sons Inc. 488 pp. ISBN: 1118647505.

DiazOrdaz, K., A. J. Franchini, and R. Grieve (2018). "Methods for estimating complier average causal effects for cost-effectiveness analysis". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.1, pp. 277–297. DOI: 10.1111/rssa.12294.

Donegan, S., N. J. Welton, C. Tudur Smith, U. D'Alessandro, and S. Dias (2017). "Network meta-analysis including treatment by covariate interactions: Consistency can vary across covariate values". In: *Research Synthesis Methods* 8.4, pp. 485–495. DOI: 10.1002/jrsm.1257.

Donegan, S., P. Williamson, U. D'Alessandro, P. Garner, and C. Tudur Smith (2013). "Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: Individual patient data may be beneficial if only for a subset of trials". In: *Statistics in Medicine* 32.6, pp. 914–930. DOI: 10.1002/sim.5584.

Doubilet, P., C. B. Begg, M. C. Weinstein, P. Braun, and B. J. McNeil (1985). "Probabilistic Sensitivity Analysis Using Monte Carlo Simulation". In: *Medical Decision Making* 5.2, pp. 157–177. DOI: 10.1177/0272989x8500500205.

Eddy, D. M., V. Hasselblad, and R. Shachter (1990). "An Introduction to a Bayesian Method for Meta-analysis". In: *Medical Decision Making* 10.1, pp. 15–23. DOI: 10.1177/0272989x9001000104.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552.

Ehm, W. (1991). "Binomial approximation to the Poisson binomial distribution". In: *Statistics & Probability Letters* 11.1, pp. 7–16. DOI: 10.1016/0167-7152(91)90170-v.

Ellison, B. E. (1964). "Two Theorems for Inferences about the Normal Distribution with Applications in Acceptance Sampling". In: *Journal of the*

American Statistical Association 59.305, pp. 89–95. DOI: 10.1080/01621459.1964.10480702.

Faria, R., M. Hernandez Alava, A. Manca, and A. J. Wailoo (2015). *NICE DSU Technical Support Document 17: the use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data*. Tech. rep. National Institute for Health and Care Excellence. URL: http://www.nicedsu.org.uk.

Fernandez, M. and S. Williams (2010). "Closed-Form Expression for the Poisson-Binomial Probability Density Function". In: *IEEE Transactions on Aerospace and Electronic Systems* 46.2, pp. 803–817. DOI: 10.1109/taes.2010.5461658.

Freeman, S. C. and J. R. Carpenter (2017). "Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models". In: *Research Synthesis Methods* 8.4, pp. 451–464. DOI: 10.1002/jrsm.1253.

Funk, M. J., D. Westreich, C. Wiesen, T. Sturmer, M. A. Brookhart, and M. Davidian (2011). "Doubly Robust Estimation of Causal Effects". In: *American Journal of Epidemiology* 173.7, pp. 761–767. DOI: 10.1093/aje/kwq439.

Gelfand, A. E., D. Dey, and H. Chang (1992). "Model determination using predictive distributions with implementation via sampling-based methods. Proceedings of the Fourth Valencia International Meeting". In: *Bayesian Statistics*. Ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. 4th ed. Clarendon Press, pp. 147–167. ISBN: 9780198522669.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013a). *Bayesian Data Analysis*. 3rd ed. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. ISBN: 9781439898208.

Gelman, A., B. Goodrich, J. Gabry, and A. Vehtari (2018). "R-squared for Bayesian Regression Models". In: *The American Statistician*, pp. 1–7. DOI: 10.1080/00031305.2018.1549100.

Gelman, A., J. Hwang, and A. Vehtari (2013b). "Understanding predictive information criteria for Bayesian models". In: *Statistics and Computing* 24.6, pp. 997–1016. DOI: 10.1007/s11222-013-9416-2.

Gelman, A. and D. B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4, pp. 457–472. DOI: 10.1214/ss/1177011136.

Glenny, A., D. Altman, F. Song, C. Sakarovitch, and J. Deeks (2005). "Indirect comparisons of competing interventions". In: *Health Technology Assessment* 9.26, p. 148. DOI: 10.3310/hta9260. URL: http://journalslibrary.nihr.ac.uk/hta/hta9260.

Golub, G. H. and C. F. Van Loan (1996). *Matrix computations*. 3rd ed. Johns Hopkins studies in the mathematical sciences. Baltimore: Johns Hopkins University Press. ISBN: 080185413X.

Gordon, K. B., A. Blauvelt, K. A. Papp, R. G. Langley, T. Luger, M. Ohtsuki, K. Reich, D. Amato, S. G. Ball, D. K. Braun, G. S. Cameron, J. Erickson, R. J. Konrad, T. M. Muram, B. J. Nickoloff, O. O. Osuntokun, R. J. Secrest, F. Zhao, L. Mallbris, and C. L. Leonardi (2016). "Phase 3 Trials of Ixekizumab in Moderate-to-Severe Plaque Psoriasis". In: *New England Journal of Medicine* 375.4, pp. 345–356. DOI: 10.1056/nejmoa1512711.

Greenland, S. (1991). "Estimating Standardized Parameters from Generalized Linear-Models". In: *Statistics in Medicine* 10.7, pp. 1069–1074. DOI: 10.1002/sim.4780100707.

Greenland, S. (1992). "Divergent biases in ecologic and individual-level studies". In: *Statistics in Medicine* 11.9, pp. 1209–1223. DOI: 10.1002/sim.4780110907.

Griffiths, C. E. M., K. Reich, M. Lebwohl, P. van de Kerkhof, C. Paul, A. Menter, G. S. Cameron, J. Erickson, L. Zhang, R. J. Secrest, S. Ball, D. K. Braun, O. O. Osuntokun, M. P. Heffernan, B. J. Nickoloff, and K. Papp (2015). "Comparison of ixekizumab with etanercept or placebo in moderate-to-severe psoriasis (UNCOVER-2 and UNCOVER-3): results from two phase 3 randomised trials". In: *The Lancet* 386.9993, pp. 541–551. DOI: 10.1016/s0140-6736(15)60125-8.

Guyot, P., A. E. Ades, M. J. N. M. Ouwens, and N. J. Welton (2012). "Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves". In: *BMC Medical Research Methodology* 12.1. DOI: 10.1186/1471-2288-12-9.

Hahn, S., P. R. Williamson, J. L. Hutton, P. Garner, and E. V. Flynn (2000). "Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies". In: *Statistics in Medicine* 19.24, pp. 3325–3336. DOI: 10.1002/1097-0258(20001230)19:24<3325::aid-sim827>3.0.co;2-d.

Hainmueller, J. (2012). "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies". In: *Political Analysis* 20.1. Polit Anal, pp. 25–46. DOI: 10.1093/pan/mpr025.

Hartman, E., R. Grieve, R. Ramsahai, and J. S. Sekhon (2015). "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate

population treatment effects". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.3, pp. 757–778. DOI: 10.1111/rssa.12094.

Hartman, E. and F. D. Hidalgo (2011). *What's the alternative?: An equivalence approach to balance and placebo tests*. Tech. rep. Department of Political Science, University of California at Berkeley, Berkeley.

Hasselblad, V. (1998). "Meta-analysis of Multitreatment Studies". In: *Medical Decision Making* 18.1, pp. 37–43. DOI: 10.1177/0272989x9801800110.

Hatswell, A. J., G. Baio, J. A. Berlin, A. Irs, and N. Freemantle (2016). "Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014". In: *BMJ Open* 6.6, e011666. DOI: 10.1136/bmjopen-2016-011666.

Hatswell, A. J., N. Freemantle, and G. Baio (2018). "Does matching adjusted indirect comparison (MAIC) work? Results from a simulation study". In: *ISPOR European Meeting*. Barcelona, Spain.

Hawkins, N., D. A. Scott, and B. Woods (2015). "'Arm-based' parameterization for network meta-analysis". In: *Research Synthesis Methods* 7.3, pp. 306–313. DOI: 10.1002/jrsm.1187.

Hedges, L. V. and I. Olkin (1985). *Statistical Methods for Meta-Analysis*. Academic Press. 369 pp. ISBN: 0123363802.

Higgins, J. P. T., D. G. Altman, P. C. Gøtzsche, P. Juni, D. Moher, A. D. Oxman, J. Savović, K. F. Schulz, L. Weeks, and J. A. C. Sterne (2011). "The Cochrane Collaboration's tool for assessing risk of bias in randomised trials". In: *BMJ* 343.oct18 2, pp. d5928–d5928. DOI: 10.1136/bmj.d5928.

Higgins, J. P. T., D. Jackson, J. K. Barrett, G. Lu, A. E. Ades, and I. R. White (2012). "Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies". In: *Research Synthesis Methods* 3.2, pp. 98–110. DOI: 10.1002/jrsm.1044.

Higgins, J. P. T., J. A. C. Sterne, J. Savović, M. J. Page, A. Hróbjartsson, I. Boutron, B. Reeves, and S. Eldridge (2016). *A revised tool for assessing risk of bias in randomized trials*. Cochrane Database of Systematic Reviews. Tech. rep. 10 suppl 1. Cochrane Methods. DOI: 10.1002/14651858.cd201601.

Higgins, J. P. T., S. G. Thompson, and D. J. Spiegelhalter (2009). "A re-evaluation of random-effects meta-analysis". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.1, pp. 137–159. DOI: 10.1111/j.1467-985x.2008.00552.x.

Higgins, J. P. T. and A. Whitehead (1996). "Borrowing strength from external trials in a meta-analysis". In: *Statistics in Medicine* 15.24, pp. 2733–2749. DOI:

10.1002/(sici)1097-0258(19961230)15:24<2733::aid-sim562>3.0.co;2-0.

Hingorani, A. D., D. A. v. d. Windt, R. D. Riley, K. Abrams, K. G. M. Moons, E. W. Steyerberg, S. Schroter, W. Sauerbrei, D. G. Altman, and H. Hemingway (2013). "Prognosis research strategy (PROGRESS) 4: Stratified medicine research". In: *BMJ* 346. DOI: 10.1136/bmj.e5793.

Hoaglin, D. C., N. Hawkins, J. P. Jansen, D. A. Scott, R. Itzler, J. C. Cappelleri, C. Boersma, D. Thompson, K. M. Larholt, M. Diaz, and A. Barrett (2011). "Conducting Indirect-Treatment-Comparison and Network-Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2". In: *Value in Health* 14.4, pp. 429–437. DOI: 10.1016/j.jval.2011.01.011.

Hofert, M. and M. Mächler (2016). "Parallel and Other Simulations in R Made Easy: An End-to-End Study". In: *Journal of Statistical Software* 69.4, pp. 1–44. ISSN: 1548-7660. DOI: 10.18637/jss.v069.i04.

Hoffman, M. D. and A. Gelman (2011). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: arXiv: http://arxiv.org/abs/1111.4246 [stat.CO].

Hong, H., H. Chu, J. Zhang, and B. P. Carlin (2015). "A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons". In: *Research Synthesis Methods* 7.1, pp. 6–22. DOI: 10.1002/jrsm.1153.

Hong, Y. (2013). "On computing the distribution function for the Poisson binomial distribution". In: *Computational Statistics & Data Analysis* 59, pp. 41–51. DOI: 10.1016/j.csda.2012.10.006.

Hua, H., D. L. Burke, M. J. Crowther, J. Ensor, C. Tudur Smith, and R. D. Riley (2016). "One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information". In: *Statistics in Medicine* 36.5, pp. 772–789. DOI: 10.1002/sim.7171.

Hyndman, R. J. (1996). "Computing and Graphing Highest Density Regions". In: *The American Statistician* 50.2, p. 120. DOI: 10.2307/2684423.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2001). *Bayesian Survival Analysis*. Springer Series in Statistics. Springer New York. ISBN: 0387952772.

Imai, K., G. King, and E. A. Stuart (2008). "Misunderstandings between experimentalists and observationalists about causal inference". In: *Journal of the Royal Statistical Society: Series A-Statistics in Society* 171, pp. 481–502. DOI: 10.1111/j.1467-985X.2007.00527.x.

Imbens, G. W. and D. B. Rubin (1997). "Bayesian inference for causal effects in randomized experiments with noncompliance". In: *The Annals of Statistics* 25.1, pp. 305–327. DOI: 10.1214/aos/1034276631.

— (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press. DOI: 10.1017/CBO9781139025751.

Ishak, K. J. (2014). *Indirect Treatment Comparison Without Network Meta-Analysis: Overview of Novel Techniques*. Tech. rep. Evidera. URL: http://www.evidera.com/wp-content/uploads/2015/04/Indirect_Treatment_Comparison_Without_Network_Meta-Analysis-_Overview_of_Novel_Techniques.pdf.

Ishak, K. J., I. Proskorovsky, and A. Benedict (2015). "Simulation and Matching-Based Approaches for Indirect Comparison of Treatments". In: *Pharmacoeconomics* 33.6, pp. 537–549. DOI: 10.1007/s40273-015-0271-1.

Jackson, C. H., N. G. Best, and S. Richardson (2006). "Improving ecological inference using individual-level data". In: *Statistics in Medicine* 25.12, pp. 2136–2159. DOI: 10.1002/sim.2370.

— (2008). "Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171.1, pp. 159–178. DOI: 10.1111/j.1467-985X.2007.00500.x.

— (2009). "Bayesian graphical models for regression on multiple data sets with different variables". In: *Biostatistics* 10.2, pp. 335–351. DOI: 10.1093/biostatistics/kxn041.

Jansen, J. P. (2012). "Network meta-analysis of individual and aggregate level data". In: *Research Synthesis Methods* 3.2, pp. 177–190. DOI: 10.1002/jrsm.1048.

Jansen, J. P. (2011). "Network meta-analysis of survival data with fractional polynomials". In: *BMC Medical Research Methodology* 11.1. DOI: 10.1186/1471-2288-11-61.

Kang, J. D. Y. and J. L. Schafer (2007). "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data". In: *Statistical Science* 22.4, pp. 523–539. DOI: 10.1214/07-Sts227.

Kenward, M. G. and J. Carpenter (2007). "Multiple imputation: current perspectives". In: *Statistical Methods in Medical Research* 16.3, pp. 199–218. DOI: 10.1177/0962280206075304.

Kirson, N. Y., S. Rao, H. G. Birnbaum, E. Kantor, R. S. Wei, and M. Cifaldi (2013). "Matching-adjusted indirect comparison of adalimumab vs etanercept and

infliximab for the treatment of psoriatic arthritis". In: *Journal of Medical Economics* 16.4, pp. 479–489. DOI: 10.3111/13696998.2013.768530.

Lambert, P. C., A. J. Sutton, K. R. Abrams, and D. R. Jones (2002). "A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis". In: *Journal of Clinical Epidemiology* 55.1, pp. 86–94. DOI: 10.1016/S0895-4356(01)00414-0.

Langley, R. G., B. E. Elewski, M. Lebwohl, K. Reich, C. E. M. Griffiths, K. Papp, L. Puig, H. Nakagawa, L. Spelman, B. Sigurgeirsson, E. Rivas, T.-F. Tsai, N. Wasel, S. Tyring, T. Salko, I. Hampele, M. Notter, A. Karpov, S. Helou, and C. Papavassilis (2014). "Secukinumab in Plaque Psoriasis — Results of Two Phase 3 Trials". In: *New England Journal of Medicine* 371.4, pp. 326–338. DOI: 10.1056/nejmoa1314258.

Latimer, N. R. and K. R. Abrams (2014). *NICE DSU Technical Support Document 16: Adjusting survival time estimates in the presence of treatment switching*. Tech. rep. National Institute for Health and Care Excellence. URL: http://www.nicedsu.org.uk.

Le Cam, L. (1960). "An approximation theorem for the Poisson binomial distribution". In: *Pacific Journal of Mathematics* 10.4, pp. 1181–1197. DOI: 10.2140/pjm.1960.10.1181.

Leahy, J. (2019). "The Impact of Performing a Network Meta-Analysis with Imperfect Evidence". PhD thesis. Trinity College Dublin. URL: http://hdl.handle.net/2262/86070.

Li, Z. and C. B. Begg (1994). "Random Effects Models for Combining Results from Controlled and Uncontrolled Studies in a Meta-Analysis". In: *Journal of the American Statistical Association* 89.428, pp. 1523–1527. DOI: 10.1080/01621459.1994.10476892.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. 2nd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc. ISBN: 9780471183860.

Lu, G. B. and A. E. Ades (2004). "Combination of direct and indirect evidence in mixed treatment comparisons". In: *Statistics in Medicine* 23.20, pp. 3105–3124. DOI: 10.1002/sim.1875.

— (2006). "Assessing evidence inconsistency in mixed treatment comparisons". In: *Journal of the American Statistical Association* 101.474, pp. 447–459. DOI: 10.1198/016214505000001302.

— (2009). "Modeling between-trial variance structure in mixed treatment comparisons". In: *Biostatistics* 10.4, pp. 792–805. DOI: 10.1093/biostatistics/kxp032.

Lunn, D., D. Spiegelhalter, N. Best, A. Thomas, and C. Jackson (2010). *The BUGS Book*. Taylor & Francis Inc. 399 pp. ISBN: 1584888490.

Maksymowych, W., V. Strand, D. Baeten, P. Nash, H. Thom, S. Cure, E. Palaka, K. Gandhi, H. Richards, and S. Jugl (2016). "Secukinumab for the treatment of ankylosing spondylitis: comparative effectiveness results versus Adalimumab using a matching-adjusted indirect comparison". In: *European League Against Rheumatism*. Vol. 75. Suppl2. London: Ann Rheum Dis, p. 98. DOI: 10.1136/annrheumdis-2016-eular.2050.

Mason, A., S. Richardson, I. Plewis, and N. Best (2012). "Strategy for Modelling Nonrandom Missing Data Mechanisms in Observational Studies Using Bayesian Methods". In: *Journal of Official Staistics* 28.2, pp. 279–302. URL: http://www.jos.nu/Articles/abstract.asp?article=282279.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. 2nd ed. London: Chapman and Hall. 532 pp. ISBN: 9780412317606.

Moher, D., K. F. Schulz, and D. G. Altman (2001). "The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials". In: *The Lancet* 357.9263, pp. 1191–1194. DOI: 10.1016/s0140-6736(00)04337-3.

Morris, T. P., I. R. White, and M. J. Crowther (2019). "Using simulation studies to evaluate statistical methods". In: *Statistics in Medicine* 38.11, pp. 2074–2102. DOI: 10.1002/sim.8086.

Müller, P., F. A. Quintana, A. Jara, and T. Hanson (2015). *Bayesian Nonparametric Data Analysis*. Springer Series in Statistics. Springer-Verlag GmbH. ISBN: 3319189670.

National Institute for Health and Care Excellence (2013). *Guide to the methods of technology appraisal*. Tech. rep. Version April 2013. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/process/pmg9/.

— (2014a). *Bortezomib for induction therapy in multiple myeloma before high-dose chemotherapy and autologous stem cell transplantation. NICE guideline (TA311)*. URL: www.nice.org.uk/guidance/TA311.

— (2014b). *Ipilimumab for previously untreated advanced (unresectable or metastatic) melanoma. NICE guideline (TA319)*. URL: www.nice.org.uk/guidance/TA319.

— (2015a). *Axitinib for treating advanced renal cell carcinoma after failure of prior systemic treatment. NICE guideline (TA333)*. URL: www.nice.org.uk/guidance/TA333.

— (2015b). *Daclatasvir for treating chronic hepatitis C. NICE guideline (TA364)*. URL: www.nice.org.uk/guidance/TA364.

— (2015c). *Simeprevir in combination with peginterferon alfa and ribavirin for treating genotypes 1 and 4 chronic hepatitis C. NICE guideline (TA331)*. URL: www.nice.org.uk/guidance/TA331.
— (2016a). *Osimertinib for treating locally advanced or metastatic EGFR T790M mutation-positive non-small-cell lung cancer. NICE guideline (TA416)*. URL: www.nice.org.uk/guidance/TA416.
— (2016b). *Panobinostat for treating multiple myeloma after at least 2 previous treatments. NICE guideline (TA380)*. URL: www.nice.org.uk/guidance/TA380.
— (2016c). *Talimogene laherparepvec for treating unresectable metastatic melanoma. NICE guideline (TA410)*. URL: www.nice.org.uk/guidance/TA410.
— (2016d). *TNF-alpha inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis. NICE guideline (TA383)*. URL: www.nice.org.uk/guidance/TA383.
— (2017a). *Atezolizumab for untreated locally advanced or metastatic urothelial cancer when cisplatin is unsuitable. NICE guideline (TA492)*. URL: www.nice.org.uk/guidance/TA492.
— (2017b). *Brentuximab vedotin for treating relapsed or refractory systemic anaplastic large cell lymphoma. NICE guideline (TA478)*. URL: www.nice.org.uk/guidance/TA478.
— (2017c). *Carfilzomib for previously treated multiple myeloma. NICE guideline (TA457)*. URL: www.nice.org.uk/guidance/TA457.
— (2017d). *Everolimus and sunitinib for treating unresectable or metastatic neuroendocrine tumours in people with progressive disease. NICE guideline (TA449)*. URL: www.nice.org.uk/guidance/TA449.
— (2017e). *Everolimus for advanced renal cell carcinoma after previous treatment. NICE guideline (TA432)*. URL: www.nice.org.uk/guidance/TA432.
— (2017f). *Ibrutinib for previously treated chronic lymphocytic leukaemia and untreated chronic lymphocytic leukaemia with 17p deletion or TP53 mutation. NICE guideline (TA429)*. URL: www.nice.org.uk/guidance/TA429.
— (2017g). *Ibrutinib for treating Waldenstrom's macroglobulinaemia. NICE guideline (TA491)*. URL: www.nice.org.uk/guidance/TA491.
— (2017h). *Nivolumab for treating relapsed or refractory classical Hodgkin lymphoma. NICE guideline (TA462)*. URL: www.nice.org.uk/guidance/TA462.
— (2017i). *Pomalidomide for multiple myeloma previously treated with lenalidomide and bortezomib. NICE guideline (TA427)*. URL: www.nice.org.uk/guidance/TA427.
— (2017j). *Ponatinib for treating chronic myeloid leukaemia and acute lymphoblastic leukaemia. NICE guideline (TA451)*. URL: www.nice.org.uk/guidance/TA451.

National Institute for Health and Care Excellence (2018a). *Avelumab for treating metastatic Merkel cell carcinoma. NICE guideline (TA517)*. URL: www.nice.org.uk/guidance/TA517.

— (2018b). *Ceritinib for untreated ALK-positive non-small-cell lung cancer. NICE guideline (TA500)*. URL: www.nice.org.uk/guidance/TA500.

— (2018c). *Daratumumab monotherapy for treating relapsed and refractory multiple myeloma. NICE guideline (TA510)*. URL: www.nice.org.uk/guidance/TA510.

— (2018d). *Technology appraisals guidance*. URL: www.nice.org.uk/guidance/published?type=ta (visited on Apr. 20, 2018).

Neal, R. M. (2003). "Slice sampling". In: *The Annals of Statistics* 31.3, pp. 705–767. DOI: 10.1214/aos/1056562461.

— (2012). "MCMC using Hamiltonian dynamics". In: *Published as Chapter 5 of the Handbook of Markov Chain Monte Carlo, 2011*. arXiv: http://arxiv.org/abs/1206.1901 [stat.CO].

Nelsen, R. B. (2006). *An Introduction to Copulas*. 2nd ed. Springer Series in Statistics. Springer New York. ISBN: 0387286594.

Nie, L. and G. Soon (2010). "A covariate-adjustment regression model approach to noninferiority margin definition". In: *Statistics in Medicine* 29.10, pp. 1107–13. DOI: 10.1002/sim.3871.

Nie, L., Z. Zhang, D. B. Rubin, and J. X. Chu (2013). "Likelihood Reweighting Methods to Reduce Potential Bias in Noninferiority Trials Which Rely on Historical Data to Make Inference". In: *Annals of Applied Statistics* 7.3, pp. 1796–1813. DOI: 10.1214/13-AOAS655.

Niederreiter, H. (1978). "Quasi-Monte Carlo methods and pseudo-random numbers". In: *Bulletin of the American Mathematical Society* 84.6, pp. 957–1041. DOI: 10.1090/S0002-9904-1978-14532-7.

Nixon, R., N. Bergvall, D. Tomic, N. Sfikas, G. Cutter, and G. Giovannoni (2014). "No Evidence of Disease Activity: Indirect Comparisons of Oral Therapies for the Treatment of Relapsing-Remitting Multiple Sclerosis". In: *Advances in Therapy* 31.11, pp. 1134–1154. DOI: 10.1007/s12325-014-0167-z.

Norton, E. C., M. M. Miller, J. J. Wang, K. Coyne, and L. C. Kleinman (2012). "Rank Reversal in Indirect Comparisons". In: *Value in Health* 15.8. Value Health, pp. 1137–1140. DOI: 10.1016/j.jval.2012.06.001.

Owen, D. B. (1980). "A table of normal integrals". In: *Communications in Statistics - Simulation and Computation* 9.4, pp. 389–419. DOI: 10.1080/03610918008812164.

Owen, D. B. (1956). "Tables for Computing Bivariate Normal Probabilities". In: *The Annals of Mathematical Statistics* 27.4, pp. 1075–1090. DOI: 10.1214/aoms/1177728074.

Papaspiliopoulos, O., G. O. Roberts, and M. Sköld (2007). "A General Framework for the Parametrization of Hierarchical Models". In: *Statistical Science* 22.1, pp. 59–73. DOI: 10.1214/088342307000000014.

Parmar, M. K. B., V. Torri, and L. Stewart (1998). "Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints". In: *Statistics in Medicine* 17.24, pp. 2815–2834. DOI: 10.1002/(sici)1097-0258(19981230)17:24<2815::aid-sim110>3.0.co;2-8.

Patefield, M. and D. Tandy (2000). "Fast and Accurate Calculation of Owen's T Function". In: *Journal of Statistical Software* 5.5. DOI: 10.18637/jss.v005.i05.

Peköz, E. A., A. Röllin, V. Čekanavičius, and M. Shwartz (2009). "A three-parameter binomial approximation". In: *Journal of Applied Probability* 46.4, pp. 1073–1085. DOI: 10.1239/jap/1261670689.

Peköz, E. A., M. Shwartz, C. L. Christiansen, and D. Berlowitz (2010). "Approximate models for aggregate data when individual-level data sets are very large or unavailable". In: *Statistics in Medicine* 29.21, pp. 2180–2193. DOI: 10.1002/sim.3979.

Petto, H., Z. Kadziola, A. Brnabic, D. Saure, and M. Belger (2019). "Alternative Weighting Approaches for Anchored Matching-Adjusted Indirect Comparisons via a Common Comparator". In: *Value in Health* 22.1, pp. 85–91. DOI: 10.1016/j.jval.2018.06.018.

Phillippo, D. M., A. E. Ades, S. Dias, S. Palmer, K. R. Abrams, and N. J. Welton (2016). *NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE*. Tech. rep. National Institute for Health and Care Excellence. URL: http://www.nicedsu.org.uk.

Phillippo, D. M., A. E. Ades, S. Dias, S. Palmer, K. R. Abrams, and N. J. Welton (2018a). "Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal". In: *Medical Decision Making* 38.2, pp. 200–211. DOI: 10.1177/0272989x17725740.

Phillippo, D. M., S. Dias, A. E. Ades, V. Didelez, and N. J. Welton (2018b). "Sensitivity of treatment recommendations to bias in network meta-analysis". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.3, pp. 843–867. DOI: 10.1111/rssa.12341.

Phillippo, D. M., S. Dias, A. Elsada, A. E. Ades, and N. J. Welton (2019a). "Population adjustment methods for indirect comparisons: A review of National Institute for Health and Care Excellence technology appraisals".

In: *International Journal of Technology Assessment in Health Care*. Available online 13/06/2019. DOI: 10.1017/S0266462319000333.

Phillippo, D. M., S. Dias, N. J. Welton, D. M. Caldwell, N. Taske, and A. E. Ades (2019b). "Threshold Analysis as an Alternative to GRADE for Assessing Confidence in Guideline Recommendations Based on Network Meta-analyses". In: *Annals of Internal Medicine* 170.8, p. 538. DOI: 10.7326/m18-3542.

Picard, R. R. and R. D. Cook (1984). "Cross-Validation of Regression Models". In: *Journal of the American Statistical Association* 79.387, pp. 575–583. DOI: 10.1080/01621459.1984.10478083.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2007). *Numerical Recipes*. 3rd ed. Cambridge University Press. 1256 pp. ISBN: 0521884071.

Prevost, T. C., K. R. Abrams, and D. R. Jones (2000). "Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening". In: *Statistics in Medicine* 19.24, pp. 3359–3376. DOI: 10.1002/1097-0258(20001230)19:24<3359::aid-sim710>3.0.co;2-n.

Quartagno, M. and J. R. Carpenter (2016). "Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates". In: *Statistics in Medicine* 35.17, pp. 2938–2954. DOI: 10.1002/sim.6837.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Riley, R. D., P. C. Lambert, and G. Abo-Zaid (2010). "Meta-analysis of individual participant data: rationale, conduct, and reporting". In: *British Medical Journal* 340. Brit Med J. DOI: 10.1136/bmj.c221.

Riley, R. D., P. C. Lambert, J. A. Staessen, J. Wang, F. Gueyffier, L. Thijs, and F. Boutitie (2008). "Meta-analysis of continuous outcomes combining individual patient data and aggregate data". In: *Statistics in Medicine* 27.11, pp. 1870–1893. DOI: 10.1002/sim.3165.

Riley, R. D. and E. W. Steyerberg (2010). "Meta-analysis of a binary outcome using individual participant data and aggregate data". In: *Research Synthesis Methods* 1.1, pp. 2–19. DOI: 10.1002/jrsm.4.

Robins, J. M. and D. M. Finkelstein (2000). "Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests". In: *Biometrics* 56.3, pp. 779–788. DOI: 10.1111/j.0006-341x.2000.00779.x.

Robins, J. M., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). "Comment: Performance of Double-Robust Estimators When Inverse Probability Weights Are Highly Variable". In: *Statistical Science*, pp. 544–559. DOI: 10.1214/07-STS227D.

Robins, J. M. and A. A. Tsiatis (1991). "Correcting for non-compliance in randomized trials using rank preserving structural failure time models". In: *Communications in Statistics - Theory and Methods* 20.8, pp. 2609–2631. DOI: 10.1080/03610929108830654.

Rosenbaum, P. R. (1987). "Model-Based Direct Adjustment". In: *Journal of the American Statistical Association* 82.398, pp. 387–394. DOI: 10.2307/2289440.

— (1991). "A Characterization of Optimal Designs for Observational Studies". In: *Journal of the Royal Statistical Society: Series B - Methodological* 53.3, pp. 597–610. URL: http://www.jstor.org/stable/2345589.

Rosenbaum, P. R. and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70.1, pp. 41–55. DOI: 10.1093/biomet/70.1.41.

Rothman, K. J., S. Greenland, and T. L. Lash (2008). *Modern Epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins. ISBN: 9780781755641.

Royston, P. and M. K. B. Parmar (2002). "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects". In: *Statistics in Medicine* 21.15, pp. 2175–2197. DOI: 10.1002/sim.1203.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc. ISBN: 9780471087052. DOI: 10.1002/9780470316696.

— (2001). "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation". In: *Health Services and Outcomes Research Methodology* 2.3, pp. 169–188. DOI: 10.1023/a:1020363010465.

Salanti, G., J. P. T. Higgins, A. E. Ades, and J. P. A. Ioannidis (2007). "Evaluation of networks of randomized trials". In: *Statistical Methods in Medical Research* 17.3, pp. 279–301. DOI: 10.1177/0962280207080643.

Salway, R. and J. Wakefield (2005). "Sources of bias in ecological studies of non-rare events". In: *Environmental and Ecological Statistics* 12.3, pp. 321–347. DOI: 10.1007/s10651-005-1516-5.

Saramago, P., A. J. Sutton, N. J. Cooper, and A. Manca (2012). "Mixed treatment comparisons using aggregate and individual participant level data". In: *Statistics in Medicine* 31.28, pp. 3516–3536. DOI: 10.1002/sim.5442.

Schmitz, S., R. Adams, and C. Walsh (2013). "Incorporating data from various trial designs into a mixed treatment comparison model". In: *Statistics in Medicine* 32.17, pp. 2935–2949. DOI: 10.1002/sim.5764.

Schulz, K. F., D. G. Altman, and D. Moher (2010). "CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials". In: *Journal of Clinical Epidemiology* 63.8, pp. 834–840. DOI: 10.1016/j.jclinepi.2010.02.005.

Senn, S. (2010). "Hans van Houwelingen and the Art of Summing up". In: *Biometrical Journal* 52.1, pp. 85–94. DOI: 10.1002/bimj.200900074.

Severini, T. A. (2005). *Elements of Distribution Theory*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. ISBN: 9780521844727.

Sherman, S., B. Amzal, E. Calvo, X. Wang, J. Park, Z. Liu, C. Lin, and R. Casciano (2015). "An Indirect Comparison of Everolimus Versus Axitinib in US Patients With Advanced Renal Cell Carcinoma in Whom Prior Sunitinib Therapy Failed". In: *Clinical Therapeutics* 37.11, pp. 2552–2559. DOI: 10.1016/j.clinthera.2015.09.013.

Signorovitch, J. E., V. Sikirica, M. H. Erder, J. P. Xie, M. Lu, P. S. Hodgkins, K. A. Betts, and E. Q. Wu (2012). "Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research". In: *Value in Health* 15.6, pp. 940–947. DOI: 10.1016/j.jval.2012.05.004.

Signorovitch, J. E., E. Q. Wu, K. A. Betts, K. Parikh, E. Kantor, A. Guo, V. K. Bollu, D. Williams, L. J. Wei, and D. J. DeAngelo (2011a). "Comparative efficacy of nilotinib and dasatinib in newly diagnosed chronic myeloid leukemia: a matching-adjusted indirect comparison of randomized trials". In: *Current Medical Research and Opinion* 27.6, pp. 1263–1271. DOI: 10.1185/03007995.2011.576238.

Signorovitch, J. E., E. Q. Wu, A. P. Yu, C. M. Gerrits, E. Kantor, Y. J. Bao, S. R. Gupta, and P. M. Mulani (2010). "Comparative Effectiveness Without Head-to-Head Trials A Method for Matching-Adjusted Indirect Comparisons Applied to Psoriasis Treatment with Adalimumab or Etanercept". In: *Pharmacoeconomics* 28.10, pp. 935–945. DOI: 10.2165/11538370-000000000-00000.

Signorovitch, J. E., R. Ayyagari, D. Cheng, and Q. Wu E. (2013a). "Matching-adjusted indirect comparisons: a simulation study of statistical performance". In: *ISPOR 18th Annual International Meeting*. New Orleans, LA, USA.

Signorovitch, J. E., E. Swallow, E. Kantor, X. Wang, J. Klimovsky, T. Haas, B. Devine, and P. Metrakos (2013b). "Everolimus and sunitinib for advanced pancreatic neuroendocrine tumors: a matching-adjusted indirect comparison". In: *Experimental Hematology & Oncology* 2.1, pp. 1–8. DOI: 10.1186/2162-3619-2-32.

Signorovitch, J. E., E. Q. Wu, E. Swallow, E. Kantor, L. Fan, and J.-B. Gruenberger (2011b). "Comparative Efficacy of Vildagliptin and Sitagliptin in Japanese Patients with Type 2 Diabetes Mellitus". In: *Clinical Drug Investigation* 31.9, pp. 665–674. DOI: 10.2165/11592490-000000000-00000.

Sikirica, V., R. L. Findling, J. E. Signorovitch, M. H. Erder, R. Dammerman, P. Hodgkins, M. Lu, J. Xie, and E. Q. Wu (2013). "Comparative Efficacy of Guanfacine Extended Release Versus Atomoxetine for the Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescents: Applying Matching-Adjusted Indirect Comparison Methodology". In: *CNS Drugs* 27.11, pp. 943–953. DOI: 10.1007/s40263-013-0102-x.

Sklar, A. W. (1959). "Fonctions de répartition à n dimension et leurs marges". In: *Publications de l'Institut de Statistique de l'Université de Paris* 8, pp. 229–231.

Smith, T. C., D. J. Spiegelhalter, and A. Thomas (1995). "Bayesian approaches to random-effects meta-analysis: A comparative study". In: *Statistics in Medicine* 14.24, pp. 2685–2699. DOI: 10.1002/sim.4780142408.

Sobol', I. M. (1967). "On the distribution of points in a cube and the approximate evaluation of integrals". In: *USSR Computational Mathematics and Mathematical Physics* 7.4, pp. 86–112. ISSN: 0041-5553. DOI: 10.1016/0041-5553(67)90144-9.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4, pp. 583–639. DOI: 10.1111/1467-9868.00353.

Stan Development Team (2018). *Stan User's Guide and Language Reference Manual*. Version 2.18. URL: https://mc-stan.org/users/documentation/.

Steyerberg, E. W., M. J. C. Eijkemans, and J. D. F. Habbema (1999). "Stepwise Selection in Small Data Sets: A Simulation Study of Bias in Logistic Regression Analysis". In: *Journal of Clinical Epidemiology* 52.10, pp. 935–942. DOI: 10.1016/s0895-4356(99)00103-1.

Strober, B., A. Brnabic, A. Schacht, L. Mallbris, K. See, R. B. Warren, and A. Nast (2016). *Indirect Comparison of Ixekizumab and Secukinumab Using Matched-Adjusted Indirect Comparisons*. Oral presentation presented at 25th

Congress of the European Academy of Dermatology and Venereology. Vienna, Austria.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (2011). "The use of propensity scores to assess the generalizability of results from randomized trials". In: *Journal of the Royal Statistical Society: Series A - Statistics in Society* 174, pp. 369–386. DOI: 10.1111/j.1467-985X.2010.00673.x.

Sutton, A. J., D. Kendrick, and C. A. C. Coupland (2008). "Meta-analysis of individual- and aggregate-level data". In: *Statistics in Medicine* 27.5, pp. 651–669. DOI: 10.1002/sim.2916.

Swallow, E., J. Song, Y. Yuan, A. Kalsekar, C. Kelley, M. Peeples, F. Mu, P. Ackerman, and J. E. Signorovitch (2016). "Daclatasvir and Sofosbuvir Versus Sofosbuvir and Ribavirin in Patients with Chronic Hepatitis C Coinfected with HIV: A Matching-adjusted Indirect Comparison". In: *Clinical Therapeutics* 38.2, pp. 404–412. DOI: 10.1016/j.clinthera.2015.12.017.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. New York: Springer-Verlag. ISBN: 0387987843.

Thom, H. H. Z., G. Capkun, A. Cerulli, R. M. Nixon, and L. S. Howard (2015). "Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension". In: *BMC Medical Research Methodology* 15.1, p. 34. DOI: 10.1186/s12874-015-0007-0.

Tu, Y.-K. (2014). "Use of Generalized Linear Mixed Models for Network Meta-analysis". In: *Medical Decision Making* 34.7, pp. 911–918. DOI: 10.1177/0272989x14545789.

Tudur Smith, C., P. R. Williamson, and A. G. Marson (2005). "Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes". In: *Statistics in Medicine* 24.9, pp. 1307–1319. DOI: 10.1002/sim.2050. URL: http:https://dx.doi.org/10.1002/sim.2050.

United States Food and Drug Administration (2016). *Hematology/Oncology (Cancer) Approvals & Safety Notifications*. URL: http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm279174.htm (visited on Aug. 10, 2016).

Van Valkenhoef, G. and A. E. Ades (2013). "Evidence Synthesis Assumes Additivity on the Scale of Measurement: Response to "Rank Reversal in Indirect Comparisons" by Norton et al". In: *Value in Health* 16.2, pp. 449–451. DOI: http://dx.doi.org/10.1016/j.jval.2012.11.012.

Valkenhoef, G. van, S. Dias, A. E. Ades, and N. J. Welton (2015). "Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis". In: *Research Synthesis Methods* 7.1, pp. 80–93. DOI: 10.1002/jrsm.1167.

Van Sanden, S., M. Pisini, I. Duchesne, A. Mehnert, and J. Belsey (2016). "Indirect comparison of the antiviral efficacy of peginterferon alpha 2a plus ribavirin used with or without simeprevir in genotype 4 hepatitis C virus infection, where common comparator study arms are lacking: a special application of the matching adjusted indirect comparison methodology". In: *Current Medical Research and Opinion* 32.1, pp. 147–154. DOI: 10.1185/03007995.2015.1106934.

Vartivarian, S. and R. J. Little (2004). "Does weighting for nonresponse increase the variance of survey means?" In: *JSM Proceedings, Survey Research Methods Section*. Alexandria, VA: American Statistical Association, pp. 3897–3904. URL: http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000892.pdf.

Vehtari, A., A. Gelman, and J. Gabry (2016). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and Computing* 27.5, pp. 1413–1432. DOI: 10.1007/s11222-016-9696-4.

Veroniki, A. A., S. E. Straus, C. Soobiah, M. J. Elliott, and A. C. Tricco (2016). "A scoping review of indirect comparison methods and applications using individual patient data". In: *BMC Medical Research Methodology* 16.1, pp. 1–14. DOI: 10.1186/s12874-016-0146-y.

Vittinghoff, E. and C. E. McCulloch (2007). "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression". In: *American Journal of Epidemiology* 165.6, pp. 710–718. DOI: 10.1093/aje/kwk052.

Welton, N. J., A. J. Sutton, N. J. Cooper, and K. R. Abrams (2012). *Evidence Synthesis for Decision Making in Healthcare*. Statistics in Practice. Wiley-Blackwell. 282 pp. ISBN: 047006109X.

White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". In: *Econometrica* 48.4, p. 817. DOI: 10.2307/1912934.

White, I. R. (2015). "Network meta-analysis". In: *Stata Journal* 15.4, 951–985(35). URL: http://www.stata-journal.com/article.html?article=st0410.

Williamson, E. J., A. Forbes, and I. R. White (2013). "Variance reduction in randomised trials by inverse probability weighting using the propensity score". In: *Statistics in Medicine* 33.5, pp. 721–737. DOI: 10.1002/sim.5991.

Woods, B. S., N. Hawkins, and D. A. Scott (2010). "Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: A tutorial". In: *BMC Medical Research Methodology* 10.1. DOI: 10.1186/1471-2288-10-54.

Woolacott, N., N. Hawkins, A. Mason, A. Kainth, Z. Khadjesari, Y. Bravo Vergel, K. Misso, K. Light, R. Chalmers, M. Sculpher, and R. Riemsma (2006). "Etanercept and efalizumab for the treatment of psoriasis: a systematic review". In: *Health Technology Assessment* 10.46. DOI: 10.3310/hta10460.

Xu, R. (2003). "Measuring explained variation in linear mixed effects models". In: *Statistics in Medicine* 22.22, pp. 3527–3541. DOI: 10.1002/sim.1572.

Yamaguchi, Y., W. Sakamoto, M. Goto, J. A. Staessen, J. Wang, F. Gueyffier, and R. D. Riley (2014). "Meta-analysis of a continuous outcome combining individual patient data and aggregate data: a method based on simulated individual patient data". In: *Research Synthesis Methods* 5.4, pp. 322–351. DOI: 10.1002/jrsm.1119.

Yan, J. (2007). "Enjoy the Joy of Copulas: With a Package copula". In: *Journal of Statistical Software* 21.4, pp. 1–21. DOI: 10.18637/jss.v021.i04.

Zhang, Z. (2009). "Covariate-Adjusted Putative Placebo Analysis in Active-Controlled Clinical Trials". In: *Statistics in Biopharmaceutical Research* 1.3, pp. 279–290. DOI: 10.1198/sbr.2009.0034.

Zhang, Z., L. Nie, G. Soon, and Z. Hu (2015). "New methods for treatment effect calibration, with applications to non-inferiority trials". In: *Biometrics*. DOI: 10.1111/biom.12388.

Glossary of abbreviations

AgD Aggregate data

BMI Body mass index

CDF Cumulative distribution function

DIC Deviance information criterion

DR Doubly robust

ELPD Expected log predictive density

EM Effect modifier

ESS Effective sample size

ETN Etanercept

FDA Food and Drug Administration

FE Fixed effect

GLM Generalised linear model

HDR Highest density region

HMC Hamiltonian Monte Carlo

HR Hazard ratio

IPD Individual patient data

IPW Inverse propensity weighting

IXE Ixekizumab

LOO Leave-one-out

LOOIC Leave-one-out information criterion

MAIC Matching adjusted indirect comparison

MCMC Markov chain Monte Carlo

MCSE Monte Carlo standard error

MGF Moment generating function

ML-NMR Multilevel network meta-regression

NICE National Institute for Health and Care Excellence

NMA Network meta-analysis

NMR Network meta-regression

OS Overall survival

PASI Psoriasis area and severity index

PATT Population average treatment effect on the treated

PBO Placebo

PFS Progression-free survival

PS Propensity score

PSIS Pareto-smoothed importance sampling

QMC Quasi-Monte Carlo

RE Random effects

SATE Sample average treatment effect

SATT Sample average treatment effect on the treated

SD Standard deviation

SEC Secukinumab

SMD Standardised mean difference

SPATE Superpopulation average treatment effect

STC Simulated treatment comparison

TA Technology appraisal

UME Unrelated mean effects

UST Ustekinumab

WLS Weighted least squares

Glossary of notation

$\mathbf{0}$	A vector of zeros
$\mathbb{I}(\cdot)$	Indicator function
$\boldsymbol{\alpha}$	Parameters in a propensity score model
$\boldsymbol{\beta}_1$	Regression parameter for prognostic effects of covariates
$\boldsymbol{\beta}_{2,ab}$	Regression parameter for interaction of effect modifying covariates with the relative effect of treatment b vs. a; we define $\boldsymbol{\beta}_{2,k} = \boldsymbol{\beta}_{2,1k}$
γ_{ab}	Individual-level (conditional) relative effect of treatment b vs. a; we define $\gamma_k = \gamma_{1k}$
δ_{jab}	Study-specific relative effect of treatment b vs. a in study j; we define $\delta_{jk} = \delta_{j1k}$
$\eta_{jk}(\boldsymbol{x})$	Linear predictor at covariate values $\boldsymbol{x}$, in study j on treatment k
$\theta_{\bullet jk}$	Expected aggregate (marginal) outcome on treatment k in study j; we write simply θ_{jk} when there is no ambiguity
θ_{ijk}	Expected individual (conditional) outcome for individual i in study j on treatment k
$\lambda_{jk}(t)$	Hazard function for treatment k in study j
$\mu_j^{(b)}$	Study-specific baseline (intercept) parameters, with respect to treatment b, for study j; we write μ_j for $\mu_j^{(1)}$ when there is no ambiguity
ν_j	Shape parameter for survival distribution in study j

ξ Parameter vector, consisting of μ_j, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_{2,k}$, and γ_k

$\pi(\cdot)$ A likelihood distribution; for ML-NMR we distinguish between $\pi_{\text{Ind}}(\cdot)$ and π_{Agg} for individual- and aggregate-level likelihoods, respectively

$\rho_{j;l_1l_2}$ Correlation between covariates x_{l_1} and x_{l_2} in study j

σ^2_{jk} Variance for individual-level Normal likelihood, on treatment k in study j

$\sigma^2_{\beta_2;l}$ Variance of exchangeable effect modifier distribution for the l-th covariate

τ^2_{ab} Heterogeneity variance of the random effect δ_{jab}; we define $\tau^2_k = \tau^2_{1k}$

$\Phi(\cdot)$ Standard Normal cumulative distribution function

$\psi^{(k)}_{ab}$ Correlation between random effects δ_{jka} and δ_{jkb}

$\boldsymbol{\Omega}_{jk}$ Correlation matrix between covariates on treatment k in study j

c_m Latent cutpoints for an ordered categorical outcome model

$C(\cdots)$ Copula function

$d_{ab(P)}$ Population-average relative effect of treatment b vs. a in population P, and we define $d_{k(P)} = d_{1k(P)}$; when constancy of relative effects is assumed (e.g. for an NMA), we drop the subscript (P)

$D_{\text{res};\bullet jk}$ Residual deviance contribution for the aggregate data on treatment k in study j

$D_{\text{res};ijk}$ Residual deviance contribution for individual i on treatment k in study j

D_{res} Total residual deviance

$f_{jk}(\cdot)$ Joint covariate distribution on treatment k in study j, with marginal distributions $f_{jk;l}(\cdot)$

$F^{-1}(\cdot)$ Inverse cumulative distribution function

$g(\cdot)$ Link function

$h(\cdot)$ A general function of the covariates and parameters

$H(\cdot)$ An objective function to be minimised; $H_{\text{MAIC}}(\cdot)$ for MAIC, $H_{\text{EB}}(\cdot)$ for entropy balancing

J Number of studies

K Number of treatments

L Number of covariates

$L_{ijk}(\boldsymbol{\xi}; y_{ijk}, \boldsymbol{x}_{ijk})$ Individual conditional likelihood function for an individual i in study j on treatment k, conditional on covariates $\boldsymbol{x}_{ijk}$

$L_{ijk}(\boldsymbol{\xi}; y_{ijk})$ Individual marginal likelihood function for an individual i in study j on treatment k

$L_{\bullet jk}(\boldsymbol{\xi}; y_{\bullet jk})$ Aggregate marginal likelihood function for aggregate data on treatment k in study j

$m_{\beta_2;l}$ Mean of exchangeable effect modifier distribution for the l-th covariate

$M_{x(P)}(\cdot)$ Moment generating function for covariate x in the population P

M Number of categories, for a categorical outcome

N_{jk} Number of individuals in study j on treatment k

$\tilde{N}$ Number of numerical integration points

p_{ijk} Probability of an event for individual i in study j on treatment k

$\bar{p}_{jk}$ Average probability of an event in study j on treatment k

p_D Effective number of parameters, based on residual deviance

$p(\boldsymbol{x})$ Propensity score, as a function of covariates

s_{jk} Standard error of the mean outcome on treatment k in study j

$S_{jk}(t)$ Survival function for treatment k in study j

t_{ijk} Survival time for individual i in study j on treatment k

$\mathcal{T}$ A set of treatments

$\tilde{\boldsymbol{u}}$ Uniform points in the unit hypercube, for numerical integration

$\tilde{\boldsymbol{u}}^*$ Correlated points in the unit hypercube, for numerical integration

w_{ik} For MAIC, the weight assigned to individual i on treatment k in the AB trial

$\boldsymbol{x}$ Vector of covariates, with $\boldsymbol{x}^{\text{EM}}$ denoting the subvector of effect modifiers

$\tilde{\boldsymbol{x}}$ Vector of numerical integration points

$\bar{\boldsymbol{x}}_{jk}$ Vector of mean covariate values on treatment k in study j

$\mathfrak{X}$ Support of $\boldsymbol{x}$

$y_{\bullet jk}$ Summary outcome on treatment k in study j; we write simply y_{jk} when there is no ambiguity

y_{ijk} Outcome for individual i study j on treatment k

Appendix A

Stan code listings

This appendix provides Stan code for the NMA and ML-NMR models used in this thesis. Code for binary outcomes, as used in Chapter 6, is given in Section A.1. Code for ordered categorical outcomes, as used in Section 7.4, is given in Section A.2. Code for survival outcomes, as used in Section 7.3, is given in Section A.3. The Stan code is modular and designed to be easily repurposed: different likelihoods and link functions may be substituted in to model other outcomes. We outline how to run the ML-NMR models in Stan from R in Section A.4.

Binary outcomes A.1

This section provides Stan code for the analysis of binary outcomes, as used in Chapter 6.

Network meta-analysis A.1.1

Firstly, we provide Stan code for fixed effect (Section A.1.1.1) and random effects (Section A.1.1.2) network meta-analysis using aggregate data (i.e. event counts per arm). We also provide code for assessing consistency using an unrelated mean effects model (Section A.1.1.3).

Fixed effect A.1.1.1

The following Stan code can be used to fit a fixed effect AgD NMA. Here we consider a binary outcome reported as event counts per arm, synthesised using a Binomial likelihood and probit link function. Other likelihoods (line 45) and link functions (line 37) may be substituted in. In the generated quantities block,

the variables `resdev_alt` and `totresdev_alt` provide the residual deviance contributions and total residual deviance under the equivalent model where some studies are fitted with IPD—i.e. here using a Bernoulli likelihood for each individual—to facilitate comparison with a ML-NMR model combining IPD and AgD.

```
data {
  // Constants
  int<lower = 2> n_i; // Number of data points

  // Data
  int<lower = 0> y[n_i]; // Number of events
  int<lower = 1> n[n_i]; // Number of individuals
  int<lower = 1> trt[n_i]; // Treatment code
  int<lower = 1> study[n_i]; // Study

  // Priors
  real<lower = 0> prior_sd_mu;
  real<lower = 0> prior_sd_d;

  // For equivalent IPD + AgD deviance calculation
  int<lower = 0, upper = 1> has_ipd[n_i]; // IPD study indicator
}
transformed data {
  int<lower = 2> n_t = max(trt); // Number of treatments
  int<lower = 1> n_s = max(study); // Number of studies
}
parameters {
  vector[n_s] mu; // Study baselines
  vector[n_t - 1] d; // Recoded basic treatment parameters (no d_1)
}
transformed parameters {
  vector[n_i] eta;
  vector[n_i] theta;

  // Linear predictor
  for (i in 1:n_i) {
    if (trt[i] > 1) eta[i] = mu[study[i]] + d[trt[i] - 1];
    else eta[i] = mu[study[i]];
  }

  // Probit model
  theta = Phi(eta);
}
model {
  // Priors
  mu ~ normal(0, prior_sd_mu);
  d ~ normal(0, prior_sd_d);

  // Likelihood
  y ~ binomial(n, theta);
}
generated quantities {
  vector[n_i] log_lik;
```

```
  vector[n_i] resdev;
  vector[n_i] yhat = to_vector(n) .* theta;
  real totresdev;
  // For equivalent IPD + AgD model
  vector[n_i] resdev_alt;
  real totresdev_alt;


  for (i in 1:n_i) {
    // Log likelihood
    log_lik[i] = binomial_lpmf(y[i] | n[i], theta[i]);

    // Residual deviance
    resdev[i] = 2 * (lmultiply(y[i], y[i] / (n[i] * theta[i])) +
                     lmultiply(n[i] - y[i],
                               (n[i] - y[i]) / (n[i] - n[i] * theta[i])));

    // For equivalent IPD + AgD model
    if (has_ipd[i] == 1){
      resdev_alt[i] = -2 * (y[i] * log(theta[i]) +
                            (n[i] - y[i]) * log(1 - theta[i]));
    } else {
      resdev_alt[i] = resdev[i];
    }
  }

  // Total residual deviance
  totresdev = sum(resdev);
  totresdev_alt = sum(resdev_alt);
}
```

Random effects A.1.1.2

The fixed effect model (Section A.1.1.1) can be modified to include random effects. Here, we assume homogeneous heterogeneity standard deviation τ, which is given a half-Normal prior distribution (line 126; note the constraint on the parameter `tau` to have a lower bound of 0 on line 101). The `functions` block defines functions to construct the random effects structure within the Stan program; alternatively, this could be constructed externally (e.g. in R) and passed as data. The non-centered RE parameterisation is used (see Section 5.2.3), via the Cholesky decomposition of the RE correlation matrix (line 90). Transformed random effects are sampled as independent standard Normal (line 129), which are then back-transformed (line 109).

```
functions {
  // Construct RE correlation matrix
  matrix Rho(int[] trt, int[] study, int n_i, int n_s) {
    int ddim[n_s];
```

```
  int s = 1;
  int arms = 0;

  for (i in 1:n_i) {
    if (trt[i] > 1) arms += 1;
    if (i < n_i && study[i] != study[i+1]) {
      ddim[s] = arms;
      arms = 0;
      s += 1;
    }
    if (i == n_i) ddim[s] = arms;
  }

  {
  int totdim = sum(ddim);
  matrix[totdim, totdim] R;
  int cumdim = 0;
  int d = 1;

  for(j in 1:totdim) {
    for(i in 1:totdim) {
      if (i == j) R[i, j] = 1;
      else if (j > cumdim && j <= cumdim + ddim[d] &&
               i > cumdim && i <= cumdim + ddim[d])
        R[i, j] = 0.5;
      else R[i, j] = 0;

      if (i == totdim && j == cumdim + ddim[d]) {
        cumdim += ddim[d];
        d += 1;
      }
    }
  }
  return R;
  }
}

// Index random effects deltas for each data point
int[] whichdelta(int[] trt, int n_i) {
  int des[n_i];
  int s = 1;
  for (i in 1:n_i) {
    if (trt[i] == 1) des[i] = 0;
    else {
      des[i] = s;
      s += 1;
    }
  }
  return des;
}

// Return the total number of random effects deltas
int ndelta(int[] trt, int n_i) {
  int count = 0;
  for (i in 1:n_i) if (trt[i] > 1) count += 1;
  return count;
}
```

```
}
data {
  // Constants
  int<lower = 2> n_i; // Number of data points

  // Data
  int<lower = 0> y[n_i]; // Number of events
  int<lower = 1> n[n_i]; // Number of individuals
  int<lower = 1> trt[n_i]; // Treatment code
  int<lower = 1> study[n_i]; // Study

  // Priors
  real<lower = 0> prior_sd_mu;
  real<lower = 0> prior_sd_d;
  real<lower = 0> prior_sd_tau;

  // For equivalent IPD + AgD deviance calculation
  int<lower = 0, upper = 1> has_ipd[n_i]; // IPD study indicator
}
transformed data {
  int<lower = 2> n_t = max(trt); // Number of treatments
  int<lower = 1> n_s = max(study); // Number of studies
  int<lower = 0> delta_design[n_i];
  int<lower = 1> n_delta = ndelta(trt, n_i);

  // RE MVN mean and correlations
  vector[n_delta] RE_mu = rep_vector(0, n_delta);
  // Cholesky decomposition of RE MVN correlations
  matrix[n_delta, n_delta] RE_L = cholesky_decompose(Rho(trt, study, n_i, n_s));

  // Which arms have RE deltas? Since we are using the reference treatment
  // parameterisation (rather than the baseline shift parameterisation), any arm
  // not on treatment 1 has a random effect
  delta_design = whichdelta(trt, n_i);
}
parameters {
  vector[n_s] mu; // Study baselines
  vector[n_t - 1] d; // Recoded basic treatment parameters (no d_1)
  vector[n_delta] u_delta; // Non-centered random effects
  real<lower = 0> tau; // RE heterogeneity standard deviation
}
transformed parameters {
  vector[n_i] eta;
  vector[n_i] theta;
  vector[n_delta] f_delta;

  // RE deltas
  f_delta = tau * RE_L * u_delta;

  // Linear predictor
  for (i in 1:n_i) {
    if (delta_design[i]) // Note: implies not treatment 1 arm
      eta[i] = mu[study[i]] + d[trt[i] - 1] + f_delta[delta_design[i]];
    else
      eta[i] = mu[study[i]];
  }
```

```
  // Probit model
  theta = Phi(eta);
}
model {
  // Priors
  mu ~ normal(0, prior_sd_mu);
  d ~ normal(0, prior_sd_d);
  tau ~ normal(0, prior_sd_tau);

  // Random effects
  u_delta ~ normal(0, 1);

  // Likelihood
  y ~ binomial(n, theta);
}
generated quantities {
  vector[n_i] log_lik;
  vector[n_i] resdev;
  vector[n_i] delta;
  vector[n_i] rhat = to_vector(n) .* theta;
  real totresdev;
  // For equivalent IPD + AgD model
  vector[n_i] resdev_alt;
  real totresdev_alt;

  for (i in 1:n_i) {
    // Log likelihood
    log_lik[i] = binomial_lpmf(y[i] | n[i], theta[i]);

    // Residual deviance
    resdev[i] = 2 * (lmultiply(y[i], y[i] / (n[i] * theta[i])) +
                     lmultiply(n[i] - y[i],
                               (n[i] - y[i]) / (n[i] - n[i] * theta[i])));

    // Shrunken estimate delta
    delta[i] = delta_design[i] ? (d[trt[i] - 1] + f_delta[delta_design[i]]) : 0;

    // For equivalent IPD + AgD model
    if (has_ipd[i] == 1){
      resdev_alt[i] = -2 * (y[i] * log(theta[i]) +
                            (n[i] - y[i]) * log(1 - theta[i]));
    } else {
      resdev_alt[i] = resdev[i];
    }
  }

  // Total residual deviance
  totresdev = sum(resdev);
  totresdev_alt = sum(resdev_alt);
}
```

Unrelated mean effects A.1.1.3

The unrelated mean effects model treats all contrasts as unrelated parameters with independent prior distributions, without imposing consistency (Section 1.2.7.1). The UME model needs to be written with the study-specific baselines referring to a reference arm in each trial (the baseline shift parameterisation, Section 1.2.1), rather than the reference treatment 1, since the reference treatment parameterisation imposes consistency implicitly. This is reflected in the random effects structure, as specified by the functions defined in the `functions` block.

```
functions {
  // Construct RE correlation matrix
  matrix Rho(int[] trt, int[] study, int n_i, int n_s) {
    int ddim[n_s];
    int arms = 0;
    int s = 1;
    for (i in 2:n_i) {
      if (study[i] == study[i - 1]) arms += 1;
      if (i < n_i && study[i] != study[i+1]) {
        ddim[s] = arms;
        arms = 0;
        s += 1;
      }
      if (i == n_i) ddim[s] = arms;
    }

    {
    int totdim = sum(ddim);
    matrix[totdim, totdim] R;
    int cumdim = 0;
    int d = 1;

    for(j in 1:totdim) {
      for(i in 1:totdim) {
        if (i == j) R[i, j] = 1;
        else if (j > cumdim && j <= cumdim + ddim[d] &&
                 i > cumdim && i <= cumdim + ddim[d])
          R[i, j] = 0.5;
        else R[i, j] = 0;

        if (i == totdim && j == cumdim + ddim[d]) {
          cumdim += ddim[d];
          d += 1;
        }
      }
    }
    return R;
    }
  }

  // Determine reference treatment for each study
```

```
  int[] ref_trt(int[] trt, int[] study, int n_i) {
    int ref[n_i];

    ref[1] = trt[1];
    for (i in 2:n_i) {
      if (study[i] == study[i - 1])
        ref[i] = ref[i - 1];
      else
        ref[i] = trt[i];
    }
    return ref;
  }

  // Index random effects deltas for each data point
  int[] whichdelta(int[] study, int n_i) {
    int des[n_i];
    int s = 1;
    des[1] = 0;
    for (i in 2:n_i) {
      if (study[i] != study[i - 1]) des[i] = 0;
      else {
        des[i] = s;
        s += 1;
      }
    }
    return des;
  }

  // Return the total number of random effects deltas
  int ndelta(int[] study, int n_i) {
    int count = 0;
    for (i in 2:n_i) if (study[i] == study[i - 1]) count += 1;
    return count;
  }
}
data {
  // Constants
  int<lower = 2> n_i; // Number of data points

  // Data
  int<lower = 0> y[n_i]; // Number of events
  int<lower = 1> n[n_i]; // Number of individuals
  int<lower = 1> trt[n_i]; // Treatment code
  int<lower = 1> study[n_i]; // Study

  // For equivalent IPD + AgD deviance calculation
  int<lower = 0, upper = 1> has_ipd[n_i]; // IPD study indicator

  // Priors
  real<lower = 0> prior_sd_mu;
  real<lower = 0> prior_sd_dd;
  real<lower = 0> prior_sd_tau;
}
transformed data {
  int<lower = 2> n_t = max(trt);
  int<lower = 1> n_s = max(study);
```

```
  // Unrelated mean effects
  int<lower = 1> n_dd = n_t * (n_t - 1) / 2;
  // Above line gives integer division warning, but this is fine here
  int<lower = 0> delta_design[n_i] = whichdelta(study, n_i);
  int<lower = 1> n_delta = ndelta(study, n_i);
  int<lower = 1> ref[n_i] = ref_trt(trt, study, n_i);
  int<lower = 0> dd_ind[n_t - 1, n_t] = rep_array(0, n_t - 1, n_t);

  // RE MVN mean and correlations
  vector[n_delta] RE_mu = rep_vector(0, n_delta);
  // Cholesky decomposition of RE MVN correlations
  matrix[n_delta, n_delta] RE_L = cholesky_decompose(Rho(trt, study, n_i, n_s));

  // Construct lookup matrix for dd index from d_ab
  {
  int i = 1;
  for (a in 1:(n_t - 1)) {
    for (b in (a + 1):n_t) {
      dd_ind[a, b] = i;
      i += 1;
    }
  }
  }
}
parameters {
  vector[n_s] mu; // Study baselines
  vector[n_dd] dd; // Unrelated mean treatment effects
  vector[n_delta] u_delta; // Non-centered random effects
  real<lower = 0> tau; // RE heterogeneity standard deviation
}
transformed parameters {
  vector[n_i] eta;
  vector[n_i] theta;
  vector[n_delta] f_delta;

  // RE deltas
  f_delta = tau * RE_L * u_delta;

  // Linear predictor
  for (i in 1:n_i) {
    if (delta_design[i]) {
      if (trt[i] > ref[i])
        eta[i] = mu[study[i]] + dd[dd_ind[ref[i], trt[i]]] +
                                f_delta[delta_design[i]];
      else if (trt[i] < ref[i])
        eta[i] = mu[study[i]] - dd[dd_ind[trt[i], ref[i]]] +
                                f_delta[delta_design[i]];
      else
        eta[i] = mu[study[i]] + f_delta[delta_design[i]];
    }
    else
      eta[i] = mu[study[i]];
  }

  // Probit model
  theta = Phi(eta);
}
```

```
model {
  // Priors
  mu ~ normal(0, prior_sd_mu);
  dd ~ normal(0, prior_sd_dd);
  tau ~ normal(0, prior_sd_tau);

  // Random effects
  u_delta ~ normal(0, 1);

  // Likelihood
  y ~ binomial(n, theta);
}
generated quantities {
  vector[n_i] log_lik;
  vector[n_i] resdev;
  vector[n_i] delta;
  vector[n_i] rhat = to_vector(n) .* theta;
  real totresdev;
  // For equivalent IPD + AgD model
  vector[n_i] resdev_alt;
  real totresdev_alt;

  for (i in 1:n_i) {
    // Log likelihood
    log_lik[i] = binomial_lpmf(y[i] | n[i], theta[i]);

    // Residual deviance
    resdev[i] = 2 * (lmultiply(y[i], y[i] / (n[i] * theta[i])) +
                     lmultiply(n[i] - y[i],
                               (n[i] - y[i]) / (n[i] - n[i] * theta[i])));

    // Shrunken estimate delta
    if (delta_design[i]) {
      if (trt[i] > ref[i])
        delta[i] = dd[dd_ind[ref[i], trt[i]]] + f_delta[delta_design[i]];
      else if (trt[i] < ref[i])
        delta[i] = f_delta[delta_design[i]] - dd[dd_ind[trt[i], ref[i]]];
      else
        delta[i] = f_delta[delta_design[i]];
    }
    else
      delta[i] = 0;

    // For equivalent IPD + AgD model
    if (has_ipd[i] == 1){
      resdev_alt[i] = -2 * (y[i] * log(theta[i]) +
                            (n[i] - y[i]) * log(1 - theta[i]));
    } else {
      resdev_alt[i] = resdev[i];
    }
  }

  // Total residual deviance
  totresdev = sum(resdev);
  totresdev_alt = sum(resdev_alt);
}
```

Multilevel network meta-regression A.1.2

We now provide Stan code for the ML-NMR model for binary outcomes, both fixed effect (Sections A.1.2.1 and A.1.2.2) and random effects (Section A.1.2.3). We also provide code for assessing consistency using an unrelated mean effects model (Section A.1.2.4). The derivation of numerical integration points is performed externally in R (see Section A.4).

Fixed effect, shared or independent effect modifiers A.1.2.1

The following code implements the fixed effect ML-NMR model (Section 4.6). We represent the linear predictor for both individual and aggregate levels together using an augmented design matrix $\boldsymbol{X}^*$, to which we then apply the QR decomposition, as described in Section 5.2.1.2. From the QR decomposition, the matrices $\boldsymbol{Q}^*$ and $(\boldsymbol{R}^*)^{-1}$ are input as data (`Q` and `R_inv`, respectively). EM interaction parameters may either be independent or shared between a set of treatments: the Stan code does not change for these cases, only the (augmented) design matrix and the resulting QR decomposition.

```
data {
  // -- Constants --
  int<lower=1> ns_ipd; // Number of IPD studies
  int<lower=1> ns_agd; // Number of AgD studies
  int<lower=1> ni_ipd; // Total number of IPD individuals
  int<lower=2> ni_agd; // Total number of AgD data points
  int<lower=1> nt; // Number of treatments
  int<lower=1> nint; // Number of samples for numerical integration
  int<lower=0> nPV; // Number of prognostic variables
  int<lower=0> nEM; // Number of effect modifier *interactions*
                    // (NOT number of EM variables)
  int<lower=1> int_thin; // Thinning factor for saved p_ii integration points

  // -- IPD --
  int<lower=0, upper=1> y[ni_ipd]; // Binary outcome

  // -- AgD --
  int<lower=0> ag_n[ni_agd]; // Outcome denominator
  int<lower=0> ag_y[ni_agd]; // Outcome numerator

  // The following are only needed if no PVs are included in the model (improves
  // sampling efficiency by not doing numerical integration on AgD reference
  // treatment 1 arms)

  // int<lower=1> ag_trt[ni_agd]; // Treatment indicator
  // int<lower=2> ag_study[ni_agd]; // Study indicator
```

```
  // -- Thin QR decomposition --
  matrix[ni_ipd + nint * ni_agd, ns_ipd + ns_agd + nPV + nEM + (nt - 1)] Q;
  matrix[ns_ipd + ns_agd + nPV + nEM + (nt - 1),
         ns_ipd + ns_agd + nPV + nEM + (nt - 1)] R_inv;

  // -- Priors --
  real<lower=0> prior_sd_mu;
  real<lower=0> prior_sd_beta1;
  real<lower=0> prior_sd_beta2;
  real<lower=0> prior_sd_gamma;
}
transformed data {
  // Total number of parameters and data points
  int totnpar = ns_ipd + ns_agd + nPV + nEM + (nt - 1);
  int totni = ni_ipd + nint * ni_agd;

  // Split Q matrix into IPD and AgD rows
  matrix[ni_ipd, totnpar] Q_ipd = Q[1:ni_ipd];
  matrix[nint * ni_agd, totnpar] Q_agd = Q[(ni_ipd + 1):totni];

  // nint/int_thin for numerical integration checks
  // This will give a warning about integer division, which cannot be avoided
  int n_int_thin = nint / int_thin;
}
parameters {
  // Parameters on QR scale
  vector[totnpar] beta_tilde;
}
transformed parameters {
  // -- Likelihood parameters needed later for log lik calculation --
  vector[ni_ipd] eta; // IPD linear predictor
  vector[ni_ipd] theta; // IPD predicted probability
  vector[ni_agd] nprime; // AgD adjusted binomial denominator
  vector[ni_agd] pprime; // AgD adjusted binomial probability


  // -- Back-transformed parameters --
  vector[totnpar] allbeta = R_inv * beta_tilde;
  // Study baselines
  vector[ns_ipd + ns_agd] mu = allbeta[1:(ns_ipd + ns_agd)];
  // Treatment effects
  vector[nt - 1] gamma = allbeta[(ns_ipd + ns_agd +1):
                                 (ns_ipd + ns_agd + nt - 1)];
  // Prognostic variables
  vector[nPV] beta1 = allbeta[(ns_ipd + ns_agd + nt):
                              (ns_ipd + ns_agd + nt - 1 + nPV)];
  // EM interactions
  vector[nEM] beta2 = allbeta[(ns_ipd + ns_agd + nt + nPV):totnpar];

  // -- AgD integration --
  vector[nint * ni_agd] p_ii = Phi(Q_agd * beta_tilde);
  vector[ni_agd] p_bar;
  vector[ni_agd] p2_bar;

  // -- IPD model --
  // We define the IPD and AgD models here in the transformed parameters block,
  // as the linear predictors are required to calculate the log likelihood later
```

```
  // on. This is slightly more inefficient than defining the models locally in
  // the model block.
  eta = Q_ipd * beta_tilde;
  theta = Phi(eta);

  // -- AgD model --
  // Using the two-parameter Binomial approximation to the Poisson Binomial
  for (i in 1:ni_agd) {
    // Uncomment if no PVs are included in the model, don't do numerical
    // integration on reference arms

    // if (nPV == 0 && ag_trt[i] == 1) {
    //   p_bar[i] = Phi(mu[ag_study[i]]);
    //   p2_bar[i] = Phi(mu[ag_study[i]])^2;
    //   nprime[i] = ag_n[i];
    //   pprime[i] = p_bar[i];
    // } else {

      p_bar[i] = mean(p_ii[(1 + (i - 1)*nint):(i*nint)]);
      p2_bar[i] = dot_self(p_ii[(1 + (i - 1)*nint):(i*nint)]) / nint;

      // Calculate adjusted n and p
      nprime[i] = ag_n[i] * p_bar[i]^2 / p2_bar[i];
      pprime[i] = p2_bar[i] / p_bar[i];

    // }

    // Reject if nprime less than number of observed events - should only happen
    // when generating initial values
    if (nprime[i] < ag_y[i]) reject("nprime = ", nprime[i],
                                    " less than ag_y = ", ag_y[i]);
  }
}
model {
  // -- Priors --
  // These prior statements will cause Stan to raise warnings regarding
  // transformed parameters possibly needing Jacobian adjustments. These should
  // be ignored, as the transformation is entirely linear
  mu ~ normal(0, prior_sd_mu);
  beta1 ~ normal(0, prior_sd_beta1);
  beta2 ~ normal(0, prior_sd_beta2);
  gamma ~ normal(0, prior_sd_gamma);

  // -- IPD likelihood --
  y ~ bernoulli(theta);

  // -- AgD likelihood --
  // We have to hand code the log likelihood contribution for the adjusted
  // binomial here, as N is not necessarily an integer (which Stan doesn't
  // like). The following is exactly equivalent to:
  //     ag_y ~ binomial(nprime, pprime);
  for (i in 1:ni_agd)
    target += lchoose(nprime[i], ag_y[i]) +
              lmultiply(ag_y[i], pprime[i]) +
              (nprime[i] - ag_y[i]) * log1m(pprime[i]);
}
generated quantities {
```

```
  // -- Log likelihood and residual deviance calculation --
  vector[ni_ipd + ni_agd] log_lik;
  vector[ni_ipd + ni_agd] resdev;

  // -- Estimate integration error --
  vector[ni_agd * n_int_thin] p_bar_cum;
  vector[ni_agd * n_int_thin] p2_bar_cum;

  // -- Predicted probabilities and numbers of events --
  vector[ni_ipd + ni_agd] p_hat;
  vector[ni_ipd + ni_agd] y_hat;

  for (i in 1:ni_ipd) {
    p_hat[i] = theta[i];
    y_hat[i] = theta[i];
    log_lik[i] = bernoulli_lpmf(y[i] | theta[i]);
    resdev[i] = -2 * log_lik[i];
  }

  for (i in 1:ni_agd) {
    log_lik[ni_ipd + i] = lchoose(nprime[i], ag_y[i]) +
                          lmultiply(ag_y[i], pprime[i]) +
                          (nprime[i] - ag_y[i]) * log1m(pprime[i]);

    y_hat[ni_ipd + i] = nprime[i] * pprime[i];
    p_hat[ni_ipd + i] = y_hat[i] / ag_n[i];

    // Approximate residual deviance for AgD, letting nprime be fixed
    resdev[ni_ipd + i] = 2 * (lmultiply(ag_y[i],
                                        ag_y[i] / (nprime[i] * pprime[i])) +
                              lmultiply(ag_n[i] - ag_y[i],
                                        (ag_n[i] - ag_y[i]) /
                                        (ag_n[i] - nprime[i] * pprime[i])));

  for (j in 1:n_int_thin) {
      p_bar_cum[(i - 1)*n_int_thin + j] =
        mean(p_ii[(1 + (i - 1)*nint):((i - 1)*nint + j*int_thin)]);
      p2_bar_cum[(i - 1)*n_int_thin + j] =
        (dot_self(p_ii[(1 + (i - 1)*nint):((i - 1)*nint + j*int_thin)]) /
        (j*int_thin));
    }
  }
}
```

A.1.2.2 Fixed effect, exchangeable effect modifiers

We can attempt to relax the shared effect modifier assumption by fitting a model with exchangeable effect modifier interactions (Section 4.6.1). As we described in Section 5.2.2, we cannot use the QR decomposition with this model; we therefore provide the (centred and scaled) augmented design matrix

X^* as data, instead of its QR decomposition.

```
data {
  // -- Constants --
  int<lower=1> ns_ipd; // Number of IPD studies
  int<lower=1> ns_agd; // Number of AgD studies
  int<lower=1> ni_ipd; // Total number of IPD individuals
  int<lower=2> ni_agd; // Total number of AgD data points
  int<lower=1> nt; // Number of treatments
  int<lower=1> nint; // Number of samples for numerical integration
  int<lower=0> nPV; // Number of prognostic variables
  int<lower=0> nEM; // Number of effect modifier *interactions*
                    // (NOT number of EM variables)
  int<lower=1> int_thin; // Thinning factor for saved p_ii integration points

  // -- RE class EM interaction --
  int<lower=0> reptclass[nEM]; // Class coding vector

  // -- IPD --
  int<lower=0, upper=1> y[ni_ipd]; // Binary outcome

  // -- AgD --
  int<lower=0> ag_n[ni_agd]; // Outcome denominator
  int<lower=0> ag_y[ni_agd]; // Outcome numerator
  // The following are only needed if no PVs are included in the model (improves
  // sampling efficiency by not doing numerical integration on AgD reference
  // treatment 1 arms)
  // int<lower=1> ag_trt[ni_agd]; // Treatment indicator
  // int<lower=2> ag_study[ni_agd]; // Study indicator

  // -- Design matrix --
  matrix[ni_ipd + nint * ni_agd, ns_ipd + ns_agd + nPV + nEM + (nt - 1)] X;

  // -- Priors --
  real prior_sd_mu;
  real prior_sd_beta1;
  real prior_sd_gamma;

  real prior_sd_beta2_nonexch;
  real prior_sd_beta2_exch;
  real prior_sd_mean_beta2_exch;

  // -- Options --
  int<lower=0,upper=1> share_class_sd; // Common class sd (1) or not (0)?
}
transformed data {
  // Total number of parameters and data points
  int totnpar = ns_ipd + ns_agd + nPV + nEM + (nt - 1);
  int totni = ni_ipd + nint * ni_agd;

  // Split X matrix into IPD and AgD rows
  matrix[ni_ipd, totnpar] X_ipd = X[1:ni_ipd];
  matrix[nint * ni_agd, totnpar] X_agd = X[(ni_ipd + 1):totni];

  // nint/int_thin for numerical integration checks
  // This will give a warning about integer division, which cannot be avoided
```

```
  int n_int_thin = nint / int_thin;

  // Number of treatment classess, inferred from class coding vector
  int<lower=0> ntclass = max(reptclass);
}
parameters {
  // Model parameters
  vector[ns_ipd + ns_agd] mu;
  vector[nPV] beta1;
  vector[nt - 1] gamma;

  // Uncentered EM interactions
  vector[nEM] u_beta2;

  // RE class EM interaction means and SDs
  vector[ntclass] classmean_beta2;
  vector<lower=0>[share_class_sd == 1 ? 1 : ntclass] classsd_beta2;
}
transformed parameters {
  // -- Likelihood parameters needed later for log lik calculation --
  vector[ni_ipd] eta; // IPD linear predictor
  vector[ni_ipd] theta; // IPD predicted probability
  vector[ni_agd] nprime; // AgD adjusted binomial denominator
  vector[ni_agd] pprime; // AgD adjusted binomial probability

  // -- Parameters --
  vector[totnpar] allbeta;

  // -- AgD integration --
  vector[nint * ni_agd] p_ii;
  vector[ni_agd] p_bar;
  vector[ni_agd] p2_bar;

  // -- Uncentered EM random interaction effect --
  vector[nEM] beta2;

  // If tclass = 0, don't put class effect on that treatment interaction
  for (k in 1:nEM) {
    if (reptclass[k] == 0) {
      beta2[k] = u_beta2[k] * prior_sd_beta2_nonexch;
      // Equivalent to beta2[k] ~ normal(0, prior_sd_beta2_nonexch)
    }
    else {
      if (share_class_sd == 0)
        beta2[k] = classmean_beta2[reptclass[k]] +
                   u_beta2[k] * classsd_beta2[reptclass[k]];
      // Equivalent to
      //   beta2[k] ~ normal(classmean_beta2[...], classsd_beta2[...])
      else
        beta2[k] = classmean_beta2[reptclass[k]] +
                   u_beta2[k] * classsd_beta2[1];
    }
  }

  // -- All parameters in one vector --
  // Study baselines
  allbeta[1:(ns_ipd + ns_agd)] = mu;
```

```
  // Treatment effects
  allbeta[(ns_ipd + ns_agd +1):(ns_ipd + ns_agd + nt - 1)] = gamma;
  // Prognostic variables
  allbeta[(ns_ipd + ns_agd + nt):(ns_ipd + ns_agd + nt - 1 + nPV)] = beta1;
  // EM interactions
  allbeta[(ns_ipd + ns_agd + nt + nPV):totnpar] = beta2;

  // -- IPD model --
  // We define the IPD and AgD models here in the transformed parameters block,
  // as the linear predictors are required to calculate the log likelihood later
  // on. This is slightly more inefficient than defining the models locally in
  // the model block.
  eta = X_ipd * allbeta;
  theta = Phi(eta);

  // -- AgD model --
  p_ii = Phi(X_agd * allbeta);
  // Using the two-parameter Binomial approximation to the Poisson Binomial.
  for (i in 1:ni_agd) {
    // Uncomment if no PVs are included in the model, don't do numerical
    // integration on reference arms

    // if (nPV == 0 && ag_trt[i] == 1) {
    //   p_bar[i] = inv_logit(mu[ag_study[i]]);
    //   p2_bar[i] = inv_logit(mu[ag_study[i]])^2;
    //   nprime[i] = ag_n[i];
    //   pprime[i] = p_bar[i];
    // } else {

      p_bar[i] = mean(p_ii[(1 + (i - 1)*nint):(i*nint)]);
      p2_bar[i] = dot_self(p_ii[(1 + (i - 1)*nint):(i*nint)]) / nint;

      // Calculate adjusted n and p
      nprime[i] = ag_n[i] * p_bar[i]^2 / p2_bar[i];
      pprime[i] = p2_bar[i] / p_bar[i];

    // }

    // Reject if nprime less than number of observed events - should only happen
    // when generating initial values
    if (nprime[i] < ag_y[i]) reject("nprime = ", nprime[i],
                                    " less than ag_y = ", ag_y[i]);
  }

}
model {
  // -- Priors --
  // These prior statements will cause Stan to raise warnings regarding
  // transformed parameters possibly needing Jacobian adjustments. These should
  // be ignored, as the transformation is entirely linear
  mu ~ normal(0, prior_sd_mu);
  beta1 ~ normal(0, prior_sd_beta1);
  gamma ~ normal(0, prior_sd_gamma);

  // -- Random class effects --
  classmean_beta2 ~ normal(0, prior_sd_mean_beta2_exch);
  classsd_beta2 ~ normal(0, prior_sd_beta2_exch);
```

```
  u_beta2 ~ normal(0, 1);

  // -- IPD likelihood --
  y ~ bernoulli(theta);

  // -- AgD likelihood --
  // We have to hand code the log likelihood contribution for the adjusted
  // binomial here, as N is not necessarily an integer (which Stan doesn't
  // like). The following is exactly equivalent to:
  // ag_y ~ binomial(nprime, pprime);
  for (i in 1:ni_agd)
    target += lchoose(nprime[i], ag_y[i]) +
              lmultiply(ag_y[i], pprime[i]) +
              (nprime[i] - ag_y[i]) * log1m(pprime[i]);
}
generated quantities {
  // -- Log likelihood and residual deviance calculation --
  vector[ni_ipd + ni_agd] log_lik;
  vector[ni_ipd + ni_agd] resdev;

  // -- Estimate integration error --
  // vector[ni_agd] p_bar_diff;
  // vector[ni_agd] p2_bar_diff;
  vector[ni_agd * n_int_thin] p_bar_cum;
  vector[ni_agd * n_int_thin] p2_bar_cum;

  // -- Predicted probabilities and numbers of events --
  vector[ni_ipd + ni_agd] p_hat;
  vector[ni_ipd + ni_agd] y_hat;

  for (i in 1:ni_ipd) {
    p_hat[i] = theta[i];
    y_hat[i] = theta[i];
    log_lik[i] = bernoulli_lpmf(y[i] | theta[i]);
    resdev[i] = -2 * log_lik[i];
  }

  for (i in 1:ni_agd) {
    log_lik[ni_ipd + i] = lchoose(nprime[i], ag_y[i]) +
                          lmultiply(ag_y[i], pprime[i]) +
                          (nprime[i] - ag_y[i]) * log1m(pprime[i]);

    y_hat[ni_ipd + i] = nprime[i] * pprime[i];
    p_hat[ni_ipd + i] = y_hat[i] / ag_n[i];

    // Approximate residual deviance for AgD, letting nprime be fixed
    resdev[ni_ipd + i] = 2 * (lmultiply(ag_y[i],
                                        ag_y[i] / (nprime[i] * pprime[i])) +
                              lmultiply(ag_n[i] - ag_y[i],
                                        (ag_n[i] - ag_y[i]) /
                                        (ag_n[i] - nprime[i] * pprime[i])));

  for (j in 1:n_int_thin) {
      p_bar_cum[(i - 1)*n_int_thin + j] =
        mean(p_ii[(1 + (i - 1)*nint):((i - 1)*nint + j*int_thin)]);
      p2_bar_cum[(i - 1)*n_int_thin + j] =
```

```
        (dot_self(p_ii[(1 + (i - 1)*nint):((i - 1)*nint + j*int_thin)]) /
        (j*int_thin));
    }
  }
}
```

Random effects A.1.2.3

We can assess residual heterogeneity by fitting a model with random effects (Section 4.6.2). As before for NMA (Section A.1.1), the fixed effect ML-NMR model (Section A.1.2.1) can be modified to include random effects. Here, we assume homogeneous heterogeneity standard deviation τ, which is given a half-Normal prior distribution (line 231; note the constraint on the parameter `tau` to have a lower bound of 0 on line 133). The `functions` block defines functions to construct the random effects structure within the Stan program (these functions are identical to those used for the NMA code in Section A.1.1.2); alternatively, this could be constructed externally (e.g. in R) and passed as data. The non-centered RE parameterisation is used (see Section 5.2.3), via the Cholesky decomposition of the RE correlation matrix (line 125). Transformed random effects are sampled as independent standard Normal (line 234), which are then back-transformed (line 143).

```
functions {
  // Construct RE correlation matrix
  matrix Rho(int[] trt, int[] study, int n_i, int n_s) {
    int ddim[n_s];
    int s = 1;
    int arms = 0;

    for (i in 1:n_i) {
      if (trt[i] > 1) arms = arms + 1;
      if (i < n_i && study[i] != study[i+1]) {
        ddim[s] = arms;
        arms = 0;
        s = s + 1;
      }
      if (i == n_i) ddim[s] = arms;
    }

    {
    int totdim = sum(ddim);
    matrix[totdim, totdim] R;
    int cumdim = 0;
    int d = 1;

    for(j in 1:totdim) {
```

```
      for(i in 1:totdim) {
        if (i == j) R[i, j] = 1;
        else if (j > cumdim && j <= cumdim + ddim[d] &&
                 i > cumdim && i <= cumdim + ddim[d])
          R[i, j] = 0.5;
        else R[i, j] = 0;

        if (i == totdim && j == cumdim + ddim[d]) {
          cumdim = cumdim + ddim[d];
          d = d + 1;
        }
      }
    }
    return R;
    }
  }

  // Index random effects deltas for each data point
  int[] whichdelta(int[] trt, int n_i) {
    int des[n_i];
    int s = 1;
    for (i in 1:n_i) {
      if (trt[i] == 1) des[i] = 0;
      else {
        des[i] = s;
        s = s + 1;
      }
    }
    return des;
  }

  // Return the total number of random effects deltas
  int ndelta(int[] trt, int n_i) {
    int count = 0;
    for (i in 1:n_i) if (trt[i] > 1) count = count + 1;
    return count;
  }
}
data {
  // -- Constants --
  int<lower=1> ns_ipd; // Number of IPD studies
  int<lower=1> ns_agd; // Number of AgD studies
  int<lower=1> ni_ipd; // Total number of IPD individuals
  int<lower=2> ni_agd; // Total number of AgD data points
  int<lower=1> nt; // Number of treatments
  int<lower=1> nint; // Number of samples for numerical integration
  int<lower=0> nPV; // Number of prognostic variables
  int<lower=0> nEM; // Number of effect modifier *interactions*
                    // (NOT number of EM variables)
  int<lower=1> int_thin; // Thinning factor for saved p_ii integration points

  // -- IPD --
  int<lower=0, upper=1> y[ni_ipd]; // Binary outcome

  // -- AgD --
  int<lower=0> ag_n[ni_agd]; // Outcome denominator
  int<lower=0> ag_y[ni_agd]; // Outcome numerator
```

```
  // -- Treatment and study indicators to construct RE terms --
  int<lower=1> narm_ipd; // Number of IPD arms
  int<lower=1> ipd_arm[ni_ipd]; // IPD arm indicator
  int<lower=1> trt[narm_ipd + ni_agd]; // Treatment indicator
  int<lower=1> study[narm_ipd + ni_agd]; // Study indicator

  // -- Thin QR decomposition --
  matrix[ni_ipd + nint * ni_agd, ns_ipd + ns_agd + nPV + nEM + (nt - 1)] Q;
  matrix[ns_ipd + ns_agd + nPV + nEM + (nt - 1),
         ns_ipd + ns_agd + nPV + nEM + (nt - 1)] R_inv;

  // -- Priors --
  real prior_sd_mu;
  real prior_sd_beta1;
  real prior_sd_beta2;
  real prior_sd_gamma;
  real prior_sd_tau;
}
transformed data {
  // Total number of parameters and data points
  int totnpar = ns_ipd + ns_agd + nPV + nEM + (nt - 1);
  int totni = ni_ipd + nint * ni_agd;

  // Split Q matrix into IPD and AgD rows
  matrix[ni_ipd, totnpar] Q_ipd = Q[1:ni_ipd];
  matrix[nint * ni_agd, totnpar] Q_agd = Q[(ni_ipd + 1):totni];

  // nint/int_thin for numerical integration checks
  // This will give a warning about integer division, which cannot be avoided
  int n_int_thin = nint / int_thin;

  // Which arms have RE deltas? Since we are using the reference treatment
  // parameterisation (rather than the baseline shift parameterisation), any arm
  // not on treatment 1 has a random effect
  int<lower = 0> delta_design[narm_ipd + ni_agd] =
    whichdelta(trt, narm_ipd + ni_agd);
  int<lower = 1> n_delta = ndelta(trt, narm_ipd + ni_agd);

  // RE MVN mean and correlations
  vector[n_delta] RE_mu = rep_vector(0, n_delta);
  // Cholesky decomposition of RE MVN correlations
  matrix[n_delta, n_delta] RE_L =
    cholesky_decompose(Rho(trt, study, narm_ipd + ni_agd, ns_ipd + ns_agd));
}
parameters {
  // Parameters on QR scale
  vector[totnpar] beta_tilde;

  // Non-centered random effects
  vector[n_delta] u_delta;
  real<lower = 0> tau;
}
transformed parameters {
  // -- Likelihood parameters needed later for log lik calculation --
  vector[ni_ipd] eta; // IPD linear predictor
  vector[ni_ipd] theta; // IPD predicted probability
```

```
  vector[ni_agd] nprime; // AgD adjusted binomial denominator
  vector[ni_agd] pprime; // AgD adjusted binomial probability

  // -- RE deltas --
  vector[n_delta] f_delta = tau * RE_L * u_delta;


  // -- Back-transformed parameters --
  vector[totnpar] allbeta = R_inv * beta_tilde;
  // Study baselines
  vector[ns_ipd + ns_agd] mu = allbeta[1:(ns_ipd + ns_agd)];
  // Treatment effects
  vector[nt - 1] gamma = allbeta[(ns_ipd + ns_agd +1):
                                 (ns_ipd + ns_agd + nt - 1)];
  // Prognostic variables
  vector[nPV] beta1 = allbeta[(ns_ipd + ns_agd + nt):
                              (ns_ipd + ns_agd + nt - 1 + nPV)];
  // EM interactions
  vector[nEM] beta2 = allbeta[(ns_ipd + ns_agd + nt + nPV):totnpar];

  // -- AgD integration --
  vector[nint * ni_agd] p_ii;
  vector[ni_agd] p_bar;
  vector[ni_agd] p2_bar;

  // -- IPD model --
  // We define the IPD and AgD models here in the transformed parameters block,
  // as the linear predictors are required to calculate the log likelihood later
  // on. This is slightly more inefficient than defining the models locally in
  // the model block.

  {
    vector[ni_ipd] eta_ipd_noRE = Q_ipd * beta_tilde;
    for (i in 1:ni_ipd) {
      if (delta_design[ipd_arm[i]])
        eta[i] = eta_ipd_noRE[i] + f_delta[delta_design[ipd_arm[i]]];
      else
        eta[i] = eta_ipd_noRE[i];
    }
  }
  theta = Phi(eta);

  // -- AgD model --
  // Using the two-parameter Binomial approximation to the Poisson Binomial.
  {
    vector[nint * ni_agd] eta_agd_noRE = Q_agd * beta_tilde;

    for (i in 1:ni_agd) {
      if (delta_design[narm_ipd + i])
        p_ii[(1 + (i-1)*nint):(i*nint)] =
          Phi(eta_agd_noRE[(1 + (i-1)*nint):(i*nint)] +
          f_delta[delta_design[narm_ipd + i]]);
      else
        p_ii[(1 + (i-1)*nint):(i*nint)] =
          Phi(eta_agd_noRE[(1 + (i-1)*nint):(i*nint)]);

      // Uncomment if no PVs are included in the model, don't do numerical
```

```
      // integration on reference arms

      // if (nPV == 0 && trt[narm_ipd + i] == 1) {
      //   p_bar[i] = inv_logit(mu[study[narm_ipd + i]]);
      //   p2_bar[i] = inv_logit(mu[study[narm_ipd + i]])^2;
      //   nprime[i] = ag_n[i];
      //   pprime[i] = p_bar[i];
      // } else {

        p_bar[i] = mean(p_ii[(1 + (i-1)*nint):(i*nint)]);
        p2_bar[i] = dot_self(p_ii[(1 + (i-1)*nint):(i*nint)]) / nint;

        // Calculate adjusted n and p
        nprime[i] = ag_n[i] * p_bar[i]^2 / p2_bar[i];
        pprime[i] = p2_bar[i] / p_bar[i];

      // }

      // Reject if nprime less than number of observed events - should only
      // happen when generating initial values
      if (nprime[i] < ag_y[i]) reject("nprime = ", nprime[i],
                                      " less than ag_y = ", ag_y[i]);
    }
  }
}
model {
  // -- Priors --
  // These prior statements will cause Stan to raise warnings regarding
  // transformed parameters possibly needing Jacobian adjustments. These should
  // be ignored, as the transformation is entirely linear
  mu ~ normal(0, prior_sd_mu);
  beta1 ~ normal(0, prior_sd_beta1);
  beta2 ~ normal(0, prior_sd_beta2);
  gamma ~ normal(0, prior_sd_gamma);

  tau ~ normal(0, prior_sd_tau);

  // -- Random effects --
  u_delta ~ normal(0, 1);

  // -- IPD likelihood --
  y ~ bernoulli(theta);

  // -- AgD likelihood --
  // We have to hand code the log likelihood contribution for the adjusted
  // binomial here, as N is not necessarily an integer (which Stan doesn't
  // like). The following is exactly equivalent to:
  //     ag_y ~ binomial(nprime, pprime);
  for (i in 1:ni_agd)
    target += lchoose(nprime[i], ag_y[i]) +
              lmultiply(ag_y[i], pprime[i]) +
              (nprime[i] - ag_y[i]) * log1m(pprime[i]);
}
generated quantities {
  // -- Log likelihood and residual deviance calculation --
  vector[ni_ipd + ni_agd] log_lik;
  vector[ni_ipd + ni_agd] resdev;
```

```
  // -- Estimate integration error --
  vector[ni_agd * n_int_thin] p_bar_cum;
  vector[ni_agd * n_int_thin] p2_bar_cum;

  // -- RE shrunken estimate delta --
  // Note: These are the individual-level trial-specific treatment effects
  vector[narm_ipd + ni_agd] delta;

  // -- Predicted probabilities and numbers of events --
  vector[ni_ipd + ni_agd] p_hat;
  vector[ni_ipd + ni_agd] y_hat;

  // For the shrunken estimates, since treatment 1 is the reference and REs are
  // treatment based rather than arm based, any treatment 1 arm has delta = 0
  for (i in 1:(narm_ipd + ni_agd)) {
    delta[i] = trt[i] == 1 ? 0 : gamma[trt[i] - 1] + f_delta[delta_design[i]];
  }

  for (i in 1:ni_ipd) {
    p_hat[i] = theta[i];
    y_hat[i] = theta[i];
    log_lik[i] = bernoulli_lpmf(y[i] | theta[i]);
    resdev[i] = -2 * log_lik[i];
  }

  for (i in 1:ni_agd) {
    log_lik[ni_ipd + i] = lchoose(nprime[i], ag_y[i]) +
                          lmultiply(ag_y[i], pprime[i]) +
                          (nprime[i] - ag_y[i]) * log1m(pprime[i]);

    y_hat[ni_ipd + i] = nprime[i] * pprime[i];
    p_hat[ni_ipd + i] = y_hat[i] / ag_n[i];

    // Approximate residual deviance for AgD, letting nprime be fixed
    resdev[ni_ipd + i] = 2 * (lmultiply(ag_y[i],
                                        ag_y[i] / (nprime[i] * pprime[i])) +
                              lmultiply(ag_n[i] - ag_y[i],
                                        (ag_n[i] - ag_y[i]) /
                                        (ag_n[i] - nprime[i] * pprime[i])));

  for (j in 1:n_int_thin) {
      p_bar_cum[(i-1)*n_int_thin + j] =
        mean(p_ii[(1 + (i-1)*nint):((i-1)*nint + j*int_thin)]);
      p2_bar_cum[(i-1)*n_int_thin + j] =
        (dot_self(p_ii[(1 + (i-1)*nint):((i-1)*nint + j*int_thin)]) /
        (j*int_thin));
    }
  }
}
```

Unrelated mean effects A.1.2.4

We can assess consistency by fitting an unrelated mean effects model (Section 4.6.3.1). Since the ML-NMR Stan code takes the augmented design matrix (or its QR decomposition) as its input, it is possible to implement an UME model by re-specifying the design matrix in terms of the independent contrast parameters, one for each contrast for which data are available. However, the UME model needs to be written with a reference arm in each trial (i.e. the baseline shift parameterisation), rather than an overall reference treatment. The following code therefore implements a random effects ML-NMR using the baseline-shift parameterisation, which allows either a standard (consistency) model or an UME (inconsistency) model to be fitted using the same code, simply by changing the design matrix.

```
functions {
  // Construct RE correlation matrix, using baseline shift parameterisation
  matrix Rho(int[] study, int n_i, int n_s) {
    int ddim[n_s];
    int arms = 0;
    int s = 1;
    for (i in 2:n_i) {
      if (study[i] == study[i - 1]) arms += 1;
      if (i < n_i && study[i] != study[i+1]) {
        ddim[s] = arms;
        arms = 0;
        s += 1;
      }
      if (i == n_i) ddim[s] = arms;
    }

    {
    int totdim = sum(ddim);
    matrix[totdim, totdim] R;
    int cumdim = 0;
    int d = 1;

    for(j in 1:totdim) {
      for(i in 1:totdim) {
        if (i == j) R[i, j] = 1;
        else if (j > cumdim && j <= cumdim + ddim[d] &&
                 i > cumdim && i <= cumdim + ddim[d])
          R[i, j] = 0.5;
        else R[i, j] = 0;

        if (i == totdim && j == cumdim + ddim[d]) {
          cumdim += ddim[d];
          d += 1;
        }
      }
    }
    return R;
```

```
    }
  }

  // Index random effects deltas for each data point
  int[] whichdelta(int[] study, int n_i) {
    int des[n_i];
    int s = 1;
    des[1] = 0;
    for (i in 2:n_i) {
      if (study[i] != study[i - 1]) des[i] = 0;
      else {
        des[i] = s;
        s += 1;
      }
    }
    return des;
  }

  // Return the total number of random effects deltas
  int ndelta(int[] study, int n_i) {
    int count = 0;
    for (i in 2:n_i) if (study[i] == study[i - 1]) count += 1;
    return count;
  }
}
data {
  // -- Constants --
  int<lower=1> ns_ipd; // Number of IPD studies
  int<lower=1> ns_agd; // Number of AgD studies
  int<lower=1> ni_ipd; // Total number of IPD individuals
  int<lower=2> ni_agd; // Total number of AgD data points
  int<lower=1> nt; // Number of treatments
  int<lower=1> nint; // Number of samples for numerical integration
  int<lower=0> nPV; // Number of prognostic variables
  int<lower=0> nEM; // Number of effect modifier *interactions*
                    // (NOT number of EM variables)
  int<lower=1> int_thin; // Thinning factor for saved p_ii integration points

  // -- IPD --
  int<lower=0, upper=1> y[ni_ipd]; // Binary outcome

  // -- AgD --
  int<lower=0> ag_n[ni_agd]; // Outcome denominator
  int<lower=0> ag_y[ni_agd]; // Outcome numerator

  // -- Treatment and study indicators to construct RE terms --
  int<lower=1> narm_ipd; // Number of IPD arms
  int<lower=1> ipd_arm[ni_ipd]; // IPD arm indicator
  int<lower=1> study[narm_ipd + ni_agd]; // Study indicator

  // -- Thin QR decomposition --
  matrix[ni_ipd + nint * ni_agd, ns_ipd + ns_agd + nPV + nEM + (nt - 1)] Q;
  matrix[ns_ipd + ns_agd + nPV + nEM + (nt - 1),
         ns_ipd + ns_agd + nPV + nEM + (nt - 1)] R_inv;

  // -- Priors --
  real prior_sd_mu;
```

```
  real prior_sd_beta1;
  real prior_sd_beta2;
  real prior_sd_gamma;
  real prior_sd_tau;
}
transformed data {
  // Total number of parameters and data points
  int totnpar = ns_ipd + ns_agd + nPV + nEM + (nt - 1);
  int totni = ni_ipd + nint * ni_agd;

  // Split Q matrix into IPD and AgD rows
  matrix[ni_ipd, totnpar] Q_ipd = Q[1:ni_ipd];
  matrix[nint * ni_agd, totnpar] Q_agd = Q[(ni_ipd + 1):totni];

  // nint/int_thin for numerical integration checks
  // This will give a warning about integer division, which cannot be avoided
  int n_int_thin = nint / int_thin;

  // Which arms have RE deltas? Using the baseline shift parameterisation
  int<lower = 0> delta_design[narm_ipd + ni_agd] =
    whichdelta(study, narm_ipd + ni_agd);
  int<lower = 1> n_delta = ndelta(study, narm_ipd + ni_agd);

  // RE MVN mean and correlations
  vector[n_delta] RE_mu = rep_vector(0, n_delta);
  // Cholesky decomposition of RE MVN correlations
  matrix[n_delta, n_delta] RE_L =
    cholesky_decompose(Rho(study, narm_ipd + ni_agd, ns_ipd + ns_agd));
}
parameters {
  // Parameters on QR scale
  vector[totnpar] beta_tilde;

  // Non-centered random effects
  vector[n_delta] u_delta;
  real<lower = 0> tau;
}
transformed parameters {
  // -- Likelihood parameters needed later for log lik calculation --
  vector[ni_ipd] eta; // IPD linear predictor
  vector[ni_ipd] theta; // IPD predicted probability
  vector[ni_agd] nprime; // AgD adjusted binomial denominator
  vector[ni_agd] pprime; // AgD adjusted binomial probability

  // -- RE deltas --
  vector[n_delta] f_delta = tau * RE_L * u_delta;

  // -- Back-transformed parameters --
  vector[totnpar] allbeta = R_inv * beta_tilde;
  // Study baselines
  vector[ns_ipd + ns_agd] mu = allbeta[1:(ns_ipd + ns_agd)];
  // Treatment effects
  vector[nt - 1] gamma = allbeta[(ns_ipd + ns_agd +1):
                                 (ns_ipd + ns_agd + nt - 1)];
  // Prognostic variables
  vector[nPV] beta1 = allbeta[(ns_ipd + ns_agd + nt):
                              (ns_ipd + ns_agd + nt - 1 + nPV)];
```

```
  // EM interactions
  vector[nEM] beta2 = allbeta[(ns_ipd + ns_agd + nt + nPV):totnpar];

  // -- AgD integration --
  // Can save these directly to check how the integration error decreases, or
  // just p_bar_diff and p2_bar_diff in the generated quantities block below
  vector[nint * ni_agd] p_ii;
  vector[ni_agd] p_bar;
  vector[ni_agd] p2_bar;

  // -- IPD model --
  // We define the IPD and AgD models here in the transformed parameters block,
  // as the linear predictors are required to calculate the log likelihood later
  // on. This is slightly more inefficient than defining the models locally in
  // the model block.

  {
    vector[ni_ipd] eta_ipd_noRE = Q_ipd * beta_tilde;
    for (i in 1:ni_ipd) {
      if (delta_design[ipd_arm[i]])
        eta[i] = eta_ipd_noRE[i] + f_delta[delta_design[ipd_arm[i]]];
      else
        eta[i] = eta_ipd_noRE[i];
    }
  }
  theta = Phi(eta);

  // -- AgD model --
  // Using the two-parameter Binomial approximation to the Poisson Binomial.
  {
    vector[nint * ni_agd] eta_agd_noRE = Q_agd * beta_tilde;

    for (i in 1:ni_agd) {
      if (delta_design[narm_ipd + i])
        p_ii[(1 + (i-1)*nint):(i*nint)] =
          Phi(eta_agd_noRE[(1 + (i-1)*nint):(i*nint)] +
          f_delta[delta_design[narm_ipd + i]]);
      else
        p_ii[(1 + (i-1)*nint):(i*nint)] =
          Phi(eta_agd_noRE[(1 + (i-1)*nint):(i*nint)]);

      // Uncomment if no PVs are included in the model, don't do numerical
      // integration on reference arms

      // if (nPV == 0 && delta_design[narm_ipd + i] == 0) {
      //   p_bar[i] = inv_logit(mu[study[narm_ipd + i]]);
      //   p2_bar[i] = inv_logit(mu[study[narm_ipd + i]])^2;
      //   nprime[i] = ag_n[i];
      //   pprime[i] = p_bar[i];
      // } else {

        p_bar[i] = mean(p_ii[(1 + (i-1)*nint):(i*nint)]);
        p2_bar[i] = dot_self(p_ii[(1 + (i-1)*nint):(i*nint)]) / nint;

        // Calculate adjusted n and p
        nprime[i] = ag_n[i] * p_bar[i]^2 / p2_bar[i];
        pprime[i] = p2_bar[i] / p_bar[i];
```

```
      // }

      // Reject if nprime less than number of observed events - should only
      // happen when generating initial values
      if (nprime[i] < ag_y[i]) reject("nprime = ", nprime[i],
                                      " less than ag_y = ", ag_y[i]);
    }
  }
}
model {
  // -- Priors --
  // These prior statements will cause Stan to raise warnings regarding
  // transformed parameters possibly needing Jacobian adjustments. These should
  // be ignored, as the transformation is entirely linear
  mu ~ normal(0, prior_sd_mu);
  beta1 ~ normal(0, prior_sd_beta1);
  beta2 ~ normal(0, prior_sd_beta2);
  gamma ~ normal(0, prior_sd_gamma);

  tau ~ normal(0, prior_sd_tau);

  // -- Random effects --
  u_delta ~ normal(0, 1);

  // -- IPD likelihood --
  y ~ bernoulli(theta);

  // -- AgD likelihood --
  // We have to hand code the log likelihood contribution for the adjusted
  // binomial here, as N is not necessarily an integer (which Stan doesn't
  // like). The following is exactly equivalent to:
  // ag_y ~ binomial(nprime, pprime);
  for (i in 1:ni_agd)
    target += lchoose(nprime[i], ag_y[i]) +
              lmultiply(ag_y[i], pprime[i]) +
              (nprime[i] - ag_y[i]) * log1m(pprime[i]);
}
generated quantities {
  // -- Log likelihood and residual deviance calculation --
  vector[ni_ipd + ni_agd] log_lik;
  vector[ni_ipd + ni_agd] resdev;

  // -- Estimate integration error --
  vector[ni_agd * n_int_thin] p_bar_cum;
  vector[ni_agd * n_int_thin] p2_bar_cum;

  // -- Predicted probabilities and numbers of events --
  vector[ni_ipd + ni_agd] p_hat;
  vector[ni_ipd + ni_agd] y_hat;

  for (i in 1:ni_ipd) {
    p_hat[i] = theta[i];
    y_hat[i] = theta[i];
    log_lik[i] = bernoulli_lpmf(y[i] | theta[i]);
    resdev[i] = -2 * log_lik[i];
  }
```

```
  for (i in 1:ni_agd) {
    log_lik[ni_ipd + i] = lchoose(nprime[i], ag_y[i]) +
                          lmultiply(ag_y[i], pprime[i]) +
                          (nprime[i] - ag_y[i]) * log1m(pprime[i]);

    y_hat[ni_ipd + i] = nprime[i] * pprime[i];
    p_hat[ni_ipd + i] = y_hat[i] / ag_n[i];

    // Approximate residual deviance for AgD, letting nprime be fixed
    resdev[ni_ipd + i] = 2 * (lmultiply(ag_y[i],
                                        ag_y[i] / (nprime[i] * pprime[i])) +
                              lmultiply(ag_n[i] - ag_y[i],
                                        (ag_n[i] - ag_y[i]) /
                                        (ag_n[i] - nprime[i] * pprime[i])));

  for (j in 1:n_int_thin) {
      p_bar_cum[(i-1)*n_int_thin + j] =
        mean(p_ii[(1 + (i-1)*nint):((i-1)*nint + j*int_thin)]);
      p2_bar_cum[(i-1)*n_int_thin + j] =
        (dot_self(p_ii[(1 + (i-1)*nint):((i-1)*nint + j*int_thin)]) /
        (j*int_thin));
    }
  }
}
```

A.2 Ordered categorical outcomes

This section provides Stan code for the analysis of ordered categorical outcomes, as used in Section 7.4. The code is largely identical to that for binary outcomes in Section A.1, except for the changes required for the ordered categorical (or for AgD, multinomial) likelihood.

A.2.1 Network meta-analysis

Firstly, we provide Stan code for fixed effect (Section A.2.1.1) and random effects (Section A.2.1.2) network meta-analysis using aggregate data (i.e. event counts in each category per arm).

A.2.1.1 Fixed effect

The following Stan code can be used to fit a fixed effect AgD NMA, using a multinomial likelihood and probit link function. Prior distributions on the differences between latent cutoffs $c_m - c_{m-1}$ are specified as half-Normal with standard deviation `prior_cc_sd` (line 67), but if a negative value of `prior_cc_sd` is passed to Stan an improper uniform prior distribution $\mathrm{U}(-\infty, \infty)$ is used in-

stead. This is possible since Stan automatically truncates the prior distribution to satisfy the ordering constraint on the latent cutoffs (line 29).

```
data {
  // Constants
  int<lower = 2> n_i; // Number of data points
  int<lower=2> ncat; // Number of ordered outcome categories

  // Data
  int<lower=0> y[n_i, ncat]; // Outcome category counts
  int<lower=1> trt[n_i]; // Treatment indicator
  int<lower=1> study[n_i]; // Study indicator

  // For equivalent IPD + AgD deviance calculation
  int<lower = 0, upper = 1> has_ipd[n_i]; // IPD study indicator

  // Priors
  real<lower=0> prior_sd_mu;
  real<lower=0> prior_sd_d;
  real prior_cc_sd;
}
transformed data {
  int<lower = 2> n_t = max(trt); // Number of treatments
  int<lower = 1> n_s = max(study); // Number of studies
}
parameters {
  vector[n_s] mu; // Study baselines
  vector[n_t -1] d; // Recoded basic treatment parameters (no d_1)

  // Ordered cutoffs on underlying probit-PASI scale
  // "Fixed effect" cutoffs, the same across trials
  positive_ordered[ncat - 2] cc;
}
transformed parameters {
  vector[n_i] eta;
  vector[ncat] theta[n_i];

  // Linear predictor
  for (i in 1:n_i) {
    if (trt[i] > 1) eta[i] = mu[study[i]] + d[trt[i] - 1];
    else eta[i] = mu[study[i]];
  }

  // Probit model, cutoffs for each category
  for (i in 1:n_i) {
    vector[ncat - 1] phi_temp;
    phi_temp[1] = Phi(eta[i]);
    theta[i, 1] = 1 - phi_temp[1];
    for (k in 2:(ncat - 1)) {
      phi_temp[k] = Phi(eta[i] - cc[k - 1]);
      theta[i, k] = phi_temp[k - 1] - phi_temp[k];
    }
    theta[i, ncat] = phi_temp[ncat - 1];
  }
}
```

```
model {
  // -- Priors --
  mu ~ normal(0, prior_sd_mu);
  d ~ normal(0, prior_sd_d);

  // Implied improper uniform prior on cutpoints if given negative prior_cc_sd:
  //   cc ~ uniform(-inf, inf)
  // Even if we place priors on cc ourselves, Stan will impose the ordering
  // constraint for us, so these are effectively half-normal priors on the
  // differences
  if (prior_cc_sd >= 0) {
    vector[ncat - 2] diff_cc;
    for (k in 1:(ncat - 2))
      diff_cc[k] = cc[k + 1] - cc[k];
    diff_cc  ~ normal(0, prior_cc_sd);
  }

  // -- Likelihood --
  for (i in 1:n_i)
    y[i] ~ multinomial(theta[i]);
}
generated quantities {
  // -- Log likelihood and residual deviance calculation --
  vector[n_i] log_lik;
  vector[ncat] y_hat[n_i];
  vector[n_i] resdev;
  real totresdev;
  // For equivalent IPD + AgD model
  vector[n_i] resdev_alt;
  real totresdev_alt;


  for (i in 1:n_i) {
    log_lik[i] = multinomial_lpmf(y[i] | theta[i]);
    y_hat[i] = sum(y[i]) * theta[i];

    // Multinomial residual deviance
    {
      vector[ncat] dv;
      for (k in 1:ncat) {
        dv[k] = y[i, k] == 0 ? 0 : y[i, k] * (log(y[i, k]) - log(y_hat[i, k]));
      }
      resdev[i] = 2 * sum(dv);
    }

    // For equivalent IPD + AgD model
    if (has_ipd[i] == 1){
      resdev_alt[i] = -2 * to_row_vector(y[i]) * log(theta[i]);
    } else {
      resdev_alt[i] = resdev[i];
    }
  }

  // Total residual deviance
  totresdev = sum(resdev);
  totresdev_alt = sum(resdev_alt);
}
```

Random effects A.2.1.2

The fixed effect model (Section A.2.1.1) can be modified to include random effects, in exactly the same manner as for binary outcomes (Section A.1.2.3). Here, we assume homogeneous heterogeneity standard deviation τ, which is given a half-Normal prior distribution.

```
functions {
  // Construct RE correlation matrix
  matrix Rho(int[] trt, int[] study, int n_i, int n_s) {
    int ddim[n_s];
    int s = 1;
    int arms = 0;

    for (i in 1:n_i) {
      if (trt[i] > 1) arms += 1;
      if (i < n_i && study[i] != study[i+1]) {
        ddim[s] = arms;
        arms = 0;
        s += 1;
      }
      if (i == n_i) ddim[s] = arms;
    }

    {
    int totdim = sum(ddim);
    matrix[totdim, totdim] R;
    int cumdim = 0;
    int d = 1;

    for(j in 1:totdim) {
      for(i in 1:totdim) {
        if (i == j) R[i, j] = 1;
        else if (j > cumdim && j <= cumdim + ddim[d] &&
                 i > cumdim && i <= cumdim + ddim[d])
          R[i, j] = 0.5;
        else R[i, j] = 0;

        if (i == totdim && j == cumdim + ddim[d]) {
          cumdim += ddim[d];
          d += 1;
        }
      }
    }
    return R;
    }
  }

  // Index random effects deltas for each data point
  int[] whichdelta(int[] trt, int n_i) {
    int des[n_i];
```

```
    int s = 1;
    for (i in 1:n_i) {
      if (trt[i] == 1) des[i] = 0;
      else {
        des[i] = s;
        s += 1;
      }
    }
    return des;
  }

  // Return the total number of random effects deltas
  int ndelta(int[] trt, int n_i) {
    int count = 0;
    for (i in 1:n_i) if (trt[i] > 1) count += 1;
    return count;
  }
}
data {
  // Constants
  int<lower = 2> n_i; // Number of data points
  int<lower=2> ncat; // Number of ordered outcome categories

  // Data
  int<lower=0> y[n_i, ncat]; // Outcome category counts
  int<lower=1> trt[n_i]; // Treatment indicator
  int<lower=1> study[n_i]; // Study indicator

  // Priors
  real<lower=0> prior_sd_mu;
  real<lower=0> prior_sd_d;
  real<lower=0> prior_sd_tau;
  real prior_cc_sd;

  // For equivalent IPD + AgD deviance calculation
  int<lower = 0, upper = 1> has_ipd[n_i]; // IPD study indicator
}
transformed data {
  int<lower = 2> n_t = max(trt); // Number of treatments
  int<lower = 1> n_s = max(study); // Number of studies

  int<lower = 0> delta_design[n_i];
  int<lower = 1> n_delta = ndelta(trt, n_i);

  // RE MVN mean and correlations
  vector[n_delta] RE_mu = rep_vector(0, n_delta);
  // Cholesky decomposition of RE MVN correlations
  matrix[n_delta, n_delta] RE_L = cholesky_decompose(Rho(trt, study, n_i, n_s));

  // Which arms have RE deltas? Since we are using the reference treatment
  // parameterisation (rather than the baseline shift parameterisation), any arm
  // not on treatment 1 has a random effect
  delta_design = whichdelta(trt, n_i);
}
parameters {
  vector[n_s] mu; // Study baselines
  vector[n_t -1] d; // Recoded basic treatment parameters (no d_1)
```

```
  vector[n_delta] u_delta; // Non-centered random effects
  real<lower = 0> tau; // RE heterogeneity standard deviation

  // Ordered cutoffs on underlying probit-PASI scale
  // "Fixed effect" cutoffs, the same across trials
  positive_ordered[ncat - 2] cc;
}
transformed parameters {
  vector[n_i] eta;
  vector[ncat] theta[n_i];
  vector[n_delta] f_delta;

  // RE deltas
  f_delta = tau * RE_L * u_delta;

  // Linear predictor
  for (i in 1:n_i) {
    if (delta_design[i]) // Note: implies not treatment 1 arm
      eta[i] = mu[study[i]] + d[trt[i] - 1] + f_delta[delta_design[i]];
    else
      eta[i] = mu[study[i]];
  }

  // Probit model, cutoffs for each category
  for (i in 1:n_i) {
    vector[ncat - 1] phi_temp;
    phi_temp[1] = Phi(eta[i]);
    theta[i, 1] = 1 - phi_temp[1];
    for (k in 2:(ncat - 1)) {
      phi_temp[k] = Phi(eta[i] - cc[k - 1]);
      theta[i, k] = phi_temp[k - 1] - phi_temp[k];
    }
    theta[i, ncat] = phi_temp[ncat - 1];
  }
}
model {
  // Priors
  mu ~ normal(0, prior_sd_mu);
  d ~ normal(0, prior_sd_d);
  tau ~ normal(0, prior_sd_tau);

  // Implied improper uniform prior on cutpoints if given negative prior_cc_sd:
  //   cc ~ uniform(-inf, inf)
  // Even if we place priors on cc ourselves, Stan will impose the ordering
  // constraint for us, so these are effectively half-normal priors on the
  // differences
  if (prior_cc_sd >= 0) {
    vector[ncat - 2] diff_cc;
    for (k in 1:(ncat - 2))
      diff_cc[k] = cc[k + 1] - cc[k];
    diff_cc  ~ normal(0, prior_cc_sd);
  }

  // Random effects
  u_delta ~ normal(0, 1);

  // Likelihood
```

```
  for (i in 1:n_i)
    y[i] ~ multinomial(theta[i]);
}
generated quantities {
  // Log likelihood and residual deviance calculation
  vector[n_i] log_lik;
  vector[ncat] y_hat[n_i];
  vector[n_i] resdev;
  real totresdev;
  // For equivalent IPD + AgD model
  vector[n_i] resdev_alt;
  real totresdev_alt;
  vector[n_i] delta;


  for (i in 1:n_i) {
    log_lik[i] = multinomial_lpmf(y[i] | theta[i]);
    y_hat[i] = sum(y[i]) * theta[i];

    // Multinomial residual deviance
    {
      vector[ncat] dv;
      for (k in 1:ncat) {
        dv[k] = y[i, k] == 0 ? 0 : y[i, k] * (log(y[i, k]) - log(y_hat[i, k]));
      }
      resdev[i] = 2 * sum(dv);
    }

    // For equivalent IPD + AgD model
    if (has_ipd[i] == 1){
      resdev_alt[i] = -2 * to_row_vector(y[i]) * log(theta[i]);
    } else {
      resdev_alt[i] = resdev[i];
    }

    // Shrunken estimates delta
    delta[i] = delta_design[i] ? d[trt[i] - 1] + f_delta[delta_design[i]] : 0;
  }

  // Total residual deviance
  totresdev = sum(resdev);
  totresdev_alt = sum(resdev_alt);
}
```

A.2.2 Multilevel network meta-regression

We now provide Stan code for the fixed effect ML-NMR model for ordered categorical outcomes. Again, the derivation of numerical integration points is performed externally in R (see Section A.4). The modifications of this code to incorporate random effects or to fit an unrelated mean effects model are identical to those for binary outcomes, given in Sections A.1.2.3 and A.1.2.4

respectively.

```
data {
  // -- Constants --
  int<lower=1> ns_ipd; // Number of IPD studies
  int<lower=1> ns_agd; // Number of AgD studies
  int<lower=1> ni_ipd; // Total number of IPD individuals
  int<lower=2> ni_agd; // Total number of AgD data points
  int<lower=1> nt; // Number of treatments
  int<lower=1> nint; // Number of samples for numerical integration
  int<lower=0> nPV; // Number of prognostic variables
  int<lower=0> nEM; // Number of effect modifier *interactions*
                    // (NOT number of EM variables)
  int<lower=2> ncat; // Number of ordered outcome categories
  int<lower=1> int_thin; // Thinning factor for saved p_ii integration points

  // -- IPD --
  int<lower=1, upper=ncat> y[ni_ipd]; // Multinomial outcome

  // -- AgD --
  int<lower=0> ag_y[ni_agd, ncat]; // Outcome category counts

  // The following are only needed if no PVs are included in the model (improves
  // sampling efficiency by not doing numerical integration on AgD reference
  // treatment 1 arms)

  // int<lower=1> ag_trt[ni_agd]; // Treatment indicator
  // int<lower=2> ag_study[ni_agd]; // Study indicator

  // -- Thin QR decomposition --
  matrix[ni_ipd + nint * ni_agd, ns_ipd + ns_agd + nPV + nEM + (nt - 1)] Q;
  matrix[ns_ipd + ns_agd + nPV + nEM + (nt - 1),
         ns_ipd + ns_agd + nPV + nEM + (nt - 1)] R_inv;

  // -- Priors --
  real<lower=0> prior_sd_mu;
  real<lower=0> prior_sd_beta1;
  real<lower=0> prior_sd_beta2;
  real<lower=0> prior_sd_gamma;
  real prior_cc_sd;
}
transformed data {
  // Total number of parameters and data points
  int totnpar = ns_ipd + ns_agd + nPV + nEM + (nt - 1);
  int totni = ni_ipd + nint * ni_agd;

  // Split Q matrix into IPD and AgD rows
  matrix[ni_ipd, totnpar] Q_ipd = Q[1:ni_ipd];
  matrix[nint * ni_agd, totnpar] Q_agd = Q[(ni_ipd + 1):totni];

  // nint/int_thin for numerical integration checks
  // This will give a warning about integer division, which cannot be avoided
  int n_int_thin = nint / int_thin;
}
parameters {
  // Parameters on QR scale
```

```
  vector[totnpar] beta_tilde;

  // Ordered cutoffs on underlying probit-PASI scale
  // "Fixed effect" cutoffs, the same across trials
  positive_ordered[ncat - 2] f_cc;
}
transformed parameters {
  // -- Cut offs --
  vector[ncat - 1] cc;

  // -- Likelihood parameters needed later for log lik calculation --
  vector[ni_ipd] eta; // IPD linear predictor
  vector[ncat] theta[ni_ipd]; // IPD predicted probabilities

  // -- Back-transformed parameters --
  vector[totnpar] allbeta = R_inv * beta_tilde;
  // Study baselines
  vector[ns_ipd + ns_agd] mu = allbeta[1:(ns_ipd + ns_agd)];
  // Treatment effects
  vector[nt - 1] gamma = allbeta[(ns_ipd + ns_agd +1):
                                 (ns_ipd + ns_agd + nt - 1)];
  // Prognostic variables
  vector[nPV] beta1 = allbeta[(ns_ipd + ns_agd + nt):
                              (ns_ipd + ns_agd + nt - 1 + nPV)];
  // EM interactions
  vector[nEM] beta2 = allbeta[(ns_ipd + ns_agd + nt + nPV):totnpar];

  // -- AgD integration --
  // Can save these directly to check how the integration error decreases
  matrix[nint * ni_agd, ncat - 1] q_ii;
  vector[ncat] p_bar[ni_agd];

  cc[1] = 0;
  cc[2:] = f_cc;

  // -- IPD model --
  // We define the IPD and AgD models here in the transformed parameters block,
  // as the linear predictors are required to calculate the log likelihood later
  // on. This is slightly more inefficient than defining the models locally in
  // the model block.
  eta = Q_ipd * beta_tilde;

  for (i in 1:ni_ipd) {
    vector[ncat - 1] phi_temp;
    phi_temp[1] = Phi(eta[i] - cc[1]);
    theta[i, 1] = 1 - phi_temp[1];
    for (k in 2:(ncat - 1)) {
      phi_temp[k] = Phi(eta[i] - cc[k]);
      theta[i, k] = phi_temp[k - 1] - phi_temp[k];
    }
    theta[i, ncat] = phi_temp[ncat - 1];
  }

  // -- AgD model --
  {
    vector[nint * ni_agd] eta_ii = Q_agd * beta_tilde;
    vector[ncat - 1] q_bar[ni_agd];
```

```
    // k == 1
    for (i in 1:ni_agd) {
      q_ii[(1 + (i - 1)*nint):(i*nint), 1] =
        Phi(eta_ii[(1 + (i - 1)*nint):(i*nint)] - cc[1]);
      q_bar[i, 1] = mean(q_ii[(1 + (i - 1)*nint):(i*nint), 1]);
      p_bar[i, 1] = 1 - q_bar[i, 1];
    }
    for (k in 2:(ncat - 1)) {
      for (i in 1:ni_agd) {
        q_ii[(1 + (i - 1)*nint):(i*nint), k] =
          Phi(eta_ii[(1 + (i - 1)*nint):(i*nint)] - cc[k]);
        q_bar[i, k] = mean(q_ii[(1 + (i - 1)*nint):(i*nint), k]);
        p_bar[i, k] = q_bar[i, k - 1] - q_bar[i, k];
      }
    }
    // k == ncat
    p_bar[, ncat] = q_bar[, ncat - 1];
  }
}
model {
  // -- Priors --
  // These prior statements will cause Stan to raise warnings regarding
  // transformed parameters possibly needing Jacobian adjustments. These should
  // be ignored, as the transformation is entirely linear
  mu ~ normal(0, prior_sd_mu);
  beta1 ~ normal(0, prior_sd_beta1);
  beta2 ~ normal(0, prior_sd_beta2);
  gamma ~ normal(0, prior_sd_gamma);

  // Implied improper uniform prior on cutpoints if given negative prior_cc_sd:
  //   cc ~ uniform(-inf, inf)
  // Even if we place priors on cc ourselves, Stan will impose the ordering
  // constraint for us, so these are effectively half-normal priors on the
  // differences
  if (prior_cc_sd >= 0) {
    vector[ncat - 2] diff_cc;
    for (k in 1:(ncat - 2))
      diff_cc[k] = cc[k + 1] - cc[k];
    diff_cc  ~ normal(0, prior_cc_sd);
  }


  // -- IPD likelihood --
  for (i in 1:ni_ipd)
    y[i] ~ categorical(theta[i]);

  // -- AgD likelihood --
  for (i in 1:ni_agd)
    ag_y[i] ~ multinomial(p_bar[i]);
}
generated quantities {
  // -- Log likelihood and residual deviance calculation --
  vector[ni_ipd + ni_agd] log_lik;
  vector[ni_ipd + ni_agd] resdev;

  // -- Estimate integration error --
```

```
    matrix[ni_agd * n_int_thin, ncat - 1] q_bar_cum;

  // -- Predicted probabilities and numbers of events --
  row_vector[ncat] p_hat[ni_ipd + ni_agd];
  row_vector[ncat] y_hat[ni_ipd + ni_agd];

  for (i in 1:ni_ipd) {
    log_lik[i] = categorical_lpmf(y[i] | theta[i]);
    p_hat[i] = theta[i]';
    y_hat[i] = theta[i]';
    resdev[i] = -2 * log_lik[i];
  }

  for (i in 1:ni_agd) {
    log_lik[ni_ipd + i] = multinomial_lpmf(ag_y[i] | p_bar[i]);
    p_hat[ni_ipd + i] = p_bar[i]';
    y_hat[ni_ipd + i] = sum(ag_y[i]) * p_hat[ni_ipd + i];

    // Multinomial residual deviance
    {
      vector[ncat] dv;
      for (k in 1:ncat) {
        dv[k] = ag_y[i, k] == 0 ?
          0 :
          ag_y[i, k] * (log(ag_y[i, k]) - log(y_hat[ni_ipd + i, k]));
      }
      resdev[ni_ipd + i] = 2 * sum(dv);
    }

    // Cumulative integration - note this is of the q_ii intermediates, NOT p_ii
    for (k in 1:(ncat - 1)) {
      for (j in 1:n_int_thin) {
        q_bar_cum[(i - 1)*n_int_thin + j, k] =
          mean(q_ii[(1 + (i - 1)*nint):((i - 1)*nint + j*int_thin), k]);
      }
    }
  }
}
```

A.3 Survival outcomes

This section provides Stan code for the analysis of survival or time-to-event outcomes, as used in Section 7.3.

A.3.1 IPD network meta-regression

Firstly, we consider an IPD network meta-regression. The following code allows the user to select the survival distribution (Exponential, Weibull, and Gompertz are implemented) by setting the value of `dist`. There is an implicit

improper uniform prior distribution $\mathrm{U}(0, \infty)$ on the shape parameters ν_j, but a proper prior distribution could be specified if desired.

```
data {
  // -- Constants --
  int<lower=1> ns; // Number of studies
  int<lower=1> ni; // Total number of individuals
  int<lower=1> nt; // Number of treatments
  int<lower=0> nPV; // Number of prognostic variables
  int<lower=0> nEM; // Number of effect modifier *interactions*
                    // (NOT number of EM variables)

  // -- Survival data --
  vector<lower=0>[ni] y; // Observation time
  int<lower=0, upper=1> status[ni]; // Event status (0 = censored, 1 = event)
  int<lower=1> study[ni]; // Study id

  // -- Distribution flag --
  int<lower=1, upper=3> dist; // 1 = Exponential, 2 = Weibull, 3 = Gompertz

  // -- Thin QR decomposition --
  matrix[ni, ns + nPV + nEM + (nt - 1)] Q;
  matrix[ns + nPV + nEM + (nt - 1), ns + nPV + nEM + (nt - 1)] R_inv;

  // -- Priors --
  real<lower=0> prior_sd_mu;
  real<lower=0> prior_sd_beta1;
  real<lower=0> prior_sd_beta2;
  real<lower=0> prior_sd_gamma;
  // NOTE: There is an implicit improper uniform prior on the shape parameter,
  // but a proper prior distribution could be specified if desired
}
transformed data {
  // Total number of parameters and data points
  int totnpar = ns + nPV + nEM + (nt - 1);
  int totni = ni;

  // Exponential model indicator, ZERO when exponential
  int<lower=0, upper=1> nonexp = dist == 1 ? 0 : 1;
}
parameters {
  // Parameters on QR scale
  vector[totnpar] beta_tilde;

  // Shape for parametric model
  // Exponential model has shape = 1 so parameter is removed (zero dimension)
  vector<lower=0>[ns*nonexp] shape;
}
transformed parameters {
  vector[ni] eta; // log rates
  vector[ni] theta; // Rates
  vector[ni] itheta; // Scales

  // -- Back-transformed parameters --
  vector[totnpar] allbeta = R_inv * beta_tilde;
```

```
  // Study baselines (equivalent to log scales)
  vector[ns] mu = allbeta[1:ns];
  // Treatment effects
  vector[nt - 1] gamma = allbeta[(ns + 1):(ns + nt - 1)];
  // Prognostic variables
  vector[nPV] beta1 = allbeta[(ns + nt):(ns + nt - 1 + nPV)];
  // EM interactions
  vector[nEM] beta2 = allbeta[(ns + nt + nPV):totnpar];

  // -- Model on log rate --
  if (dist == 1) { // Exponential
    eta = Q * beta_tilde;
  } else if (dist == 2) { // Weibull
    eta = (Q * beta_tilde) ./ shape[study];
  } else if (dist == 3) { // Gompertz
    eta = Q * beta_tilde;
  }
  theta = log(eta);
  itheta = log(-eta);
}
model {
  // -- Priors --
  // These prior statements will cause Stan to raise warnings regarding
  // transformed parameters possibly needing Jacobian adjustments. These should
  // be ignored, as the transformation is entirely linear.
  mu ~ normal(0, prior_sd_mu);
  beta1 ~ normal(0, prior_sd_beta1);
  beta2 ~ normal(0, prior_sd_beta2);
  gamma ~ normal(0, prior_sd_gamma);

  // NOTE: implied improper uniform prior U(0, inf) on shape

  // -- Likelihood --
  if (dist == 1) { // Exponential model
    for (i in 1:ni) {
      if (status[i]==1)
        y[i] ~ exponential(theta[i]);
      else
        target += exponential_lccdf(y[i] | theta[i]);
    }
  } else if (dist == 2) { // Weibull model
    for (i in 1:ni) {
      if (status[i]==1)
        y[i] ~ weibull(shape[study[i]], itheta[i]);
      else
        target += weibull_lccdf(y[i] | shape[study[i]], itheta[i]);
    }
  } else if (dist == 3) { // Gompertz model
    // Stan does not have Gompertz distribution, instead code the log likelihood
    // contributions directly
    for (i in 1:ni) {
      if (status[i]==1)
        target += -theta[i]/shape[study[i]] * expm1(shape[study[i]] * y[i]) +
                  eta[i] + (shape[study[i]] * y[i]);
      else
        target += -theta[i]/shape[study[i]] * expm1(shape[study[i]] * y[i]);
    }
```

```
  }
}
generated quantities {
  // Transform intercepts back to scales
  vector[ns] scale = exp(mu);
  // Log likelihood contributions
  vector[ni] log_lik;

  if (dist == 1) { // Exponential model
    for (i in 1:ni) {
      if (status[i]==1)
        log_lik[i] = exponential_lpdf(y[i] | theta[i]);
      else
        log_lik[i] = exponential_lccdf(y[i] | theta[i]);
    }
  } else if (dist == 2) { // Weibull model
    for (i in 1:ni) {
      if (status[i]==1)
        log_lik[i] = weibull_lpdf(y[i] | shape[study[i]], itheta[i]);
      else
        log_lik[i] = weibull_lccdf(y[i] | shape[study[i]], itheta[i]);
    }
  } else if (dist == 3) { // Gompertz model
    for (i in 1:ni) {
      if (status[i]==1)
        log_lik[i] = -theta[i]/shape[study[i]] * expm1(shape[study[i]] * y[i]) +
                     eta[i] + (shape[study[i]] * y[i]);
      else
        log_lik[i] = -theta[i]/shape[study[i]] * expm1(shape[study[i]] * y[i]);
    }
  }
}
```

Multilevel network meta-regression A.3.2

We now provide Stan code for the fixed effect ML-NMR model for survival outcomes. Again, the derivation of numerical integration points is performed externally in R (see Section A.4). As for the IPD NMR model (Section A.3.1), the code allows the user to select the survival distribution (Exponential, Weibull, and Gompertz are implemented) by setting the value of `dist`. Alternative survival and hazard functions could be specified by the user in the `functions` block. There is an implicit improper uniform prior distribution $\mathrm{U}(0, \infty)$ on the shape parameters ν_j, but a proper prior distribution could be specified if desired.

```
functions {
  // -- Exponential --
  // Survival
```

```
  vector S_exp(real y, vector rate) {
    return exp(-y * rate);
  }

  // Hazard
  vector h_exp(vector rate) {
    return rate;
  }

  // -- Weibull --
  vector S_weib(real y, vector rate, real shape) {
    vector[num_elements(rate)] S;
    for (i in 1:num_elements(rate)) {
      S[i] = exp(-pow(y * rate[i], shape));
    }
    return S;
  }

  // Hazard
  vector h_weib(real y, vector rate, real shape) {
    vector[num_elements(rate)] h;
    for (i in 1:num_elements(rate)) {
      h[i] = shape * rate[i] * pow(y * rate[i], shape - 1);
    }
    return h;
  }

  // -- Gompertz --
  // Survival
  vector S_gomp(real y, vector rate, real shape) {
    vector[num_elements(rate)] S;
    for (i in 1:num_elements(rate)) {
      S[i] = exp(-rate[i]/shape * expm1(shape * y));
    }
    return S;
  }

  // Hazard
  vector h_gomp(real y, vector rate, real shape) {
    return rate * exp(shape * y);
  }
}
data {
  // -- Constants --
  int<lower=1> ns_ipd; // Number of IPD studies
  int<lower=1> ns_agd; // Number of AgD studies
  int<lower=1> ni_ipd; // Total number of IPD individuals
  int<lower=2> ni_agd; // Total number of AgD data points
  int<lower=1> nt; // Number of treatments
  int<lower=1> nint; // Number of samples for numerical integration
  int<lower=0> nPV; // Number of prognostic variables
  int<lower=0> nEM; // Number of effect modifier *interactions*
                    // (NOT number of EM variables)
  int<lower=1> int_thin; // Thinning factor for saved integration points

  // -- Survival data --
  vector<lower=0>[ni_ipd + ni_agd] y; // Observation time
```

```
  int<lower=0, upper=1> status[ni_ipd + ni_agd]; // Event status
                                                 // (0 = censored, 1 = event)
  int<lower=1> study[ni_ipd + ni_agd]; // Study id

  // -- Integration point details --
  int<lower=1> int_id[ni_agd]; // Integration id (i.e. for an arm or study) for
                               // each AgD individual event time
  int<lower=1> int_study[max(int_id)]; // Study id for integration points

  // -- Distribution flag --
  int<lower=1, upper=3> dist; // 1 = Exponential, 2 = Weibull, 3 = Gompertz

  // -- Thin QR decomposition --
  matrix[ni_ipd + nint * max(int_id), ns_ipd + ns_agd + nPV + nEM + (nt - 1)] Q;
  matrix[ns_ipd + ns_agd + nPV + nEM + (nt - 1),
         ns_ipd + ns_agd + nPV + nEM + (nt - 1)] R_inv;

  // -- Priors --
  real<lower=0> prior_sd_mu;
  real<lower=0> prior_sd_beta1;
  real<lower=0> prior_sd_beta2;
  real<lower=0> prior_sd_gamma;
  // NOTE: There is an implicit improper uniform prior on the shape parameter,
  // but a proper prior distribution could be specified if desired
}
transformed data {
  // Total number of studies
  int ns = ns_ipd + ns_agd;

  // Total number of parameters
  int totnpar = ns + nPV + nEM + (nt - 1);

  // For numerical integration checks
  // This will give a warning about integer division, which cannot be avoided
  int n_int_thin = nint / int_thin;

  // Exponential model indicator, ZERO when exponential
  int<lower=0, upper=1> nonexp = dist == 1 ? 0 : 1;

  // Status to vector (from array)
  vector<lower=0, upper=1>[ni_ipd + ni_agd] status_v = to_vector(status);

  // Study design vector, with expanded AgD integration points
  int study_long[ni_ipd + nint * max(int_id)];
  study_long[1:ni_ipd] = study[1:ni_ipd];
  for (i in 1:max(int_id)) {
    study_long[(ni_ipd + (i-1)*nint + 1):(ni_ipd + i*nint)] =
      rep_array(int_study[i], nint);
  }
}
parameters {
  // Parameters on QR scale
  vector[totnpar] beta_tilde;

  // Shape for parametric model
  // Exponential model has shape = 1 so parameter is removed (zero dimension)
  vector<lower=0>[ns*nonexp] shape;
```

```
}
transformed parameters {
  vector[ni_ipd + nint * max(int_id)] eta; // log rates
  vector[ni_ipd + nint * max(int_id)] theta; // Rates
  vector[ni_ipd + nint * max(int_id)] itheta; // Scales

  // -- Back-transformed parameters --
  vector[totnpar] allbeta = R_inv * beta_tilde;
  // Study baselines (equivalent to log scales)
  vector[ns] mu = allbeta[1:ns];
  // Treatment effects
  vector[nt - 1] gamma = allbeta[(ns + 1):(ns + nt - 1)];
  // Prognostic variables
  vector[nPV] beta1 = allbeta[(ns + nt):(ns + nt - 1 + nPV)];
  // EM interactions
  vector[nEM] beta2 = allbeta[(ns + nt + nPV):totnpar];

  // -- AgD integration --
  vector[nint * ni_agd] S_ii;
  vector[nint * ni_agd] h_ii;
  vector[ni_agd] P_bar;

  // -- Model on log rate --
  if (dist == 1) { // Exponential
    eta = Q * beta_tilde;
  } else if (dist == 2) { // Weibull
    eta = (Q * beta_tilde) ./ shape[study_long];
  } else if (dist == 3) { // Gompertz
    eta = Q * beta_tilde;
  }
  theta = exp(eta);   // Rates
  itheta = exp(-eta); // Scales

  // -- Perform AgD integration --
  // NOTE: Integration of hazard only needed for observed (not censored) events
  for (i in 1:ni_agd) {
    if (dist == 1) { // Exponential
      S_ii[((i-1)*nint + 1):i*nint]
        = S_exp(y[ni_ipd + i],
                theta[(ni_ipd + (int_id[i]-1)*nint + 1):
                      (ni_ipd + int_id[i]*nint)]);
      // Exponential hazard is just the rate
      // NOTE: equivalent to using function h_exp but probably more efficient
      h_ii[((i-1)*nint + 1):i*nint]
        = status[ni_ipd + i] == 1
          ? theta[(ni_ipd + (int_id[i]-1)*nint + 1):(ni_ipd + int_id[i]*nint)]
          : rep_vector(0, nint);
    } else if (dist == 2) { // Weibull
      S_ii[((i-1)*nint + 1):i*nint]
        = S_weib(y[ni_ipd + i],
                 theta[(ni_ipd + (int_id[i]-1)*nint + 1):
                       (ni_ipd + int_id[i]*nint)],
                 shape[study[ni_ipd + i]]);
      h_ii[((i-1)*nint + 1):i*nint]
        = status[ni_ipd + i] == 1
          ? h_weib(y[ni_ipd + i],
                   theta[(ni_ipd + (int_id[i]-1)*nint + 1):
```

```
                        (ni_ipd + int_id[i]*nint)],
                  shape[study[ni_ipd + i]])
          : rep_vector(0, nint);
    } else if (dist == 3) { // Gompertz
      S_ii[((i-1)*nint + 1):i*nint]
        = S_gomp(y[ni_ipd + i],
                 theta[(ni_ipd + (int_id[i]-1)*nint + 1):
                       (ni_ipd + int_id[i]*nint)],
                 shape[study[ni_ipd + i]]);
      h_ii[((i-1)*nint + 1):i*nint]
        = status[ni_ipd + i] == 1
          ? h_gomp(y[ni_ipd + i],
                   theta[(ni_ipd + (int_id[i]-1)*nint + 1):
                         (ni_ipd + int_id[i]*nint)],
                   shape[study[ni_ipd + i]])
          : rep_vector(0, nint);
    }

    // Take average to calculate marginal survival and hazard
    // Again, only take mean of h_ii if uncensored
    // NOTE: Set h_bar to 1 if censored, so that log(h_bar) is 0 (i.e. no log
    // likelihood contribution)
    if (status[ni_ipd + i] == 1) {
      P_bar[i] = mean(S_ii[((i-1)*nint + 1):i*nint] .*
                      h_ii[((i-1)*nint + 1):i*nint]);
    } else {
      P_bar[i] = mean(S_ii[((i-1)*nint + 1):i*nint]);
    }
  }
}
model {
  // -- Priors --
  // These prior statements will cause Stan to raise warnings regarding
  // transformed parameters possibly needing Jacobian adjustments. These should
  // be ignored, as the transformation is entirely linear.
  mu ~ normal(0, prior_sd_mu);
  beta1 ~ normal(0, prior_sd_beta1);
  beta2 ~ normal(0, prior_sd_beta2);
  gamma ~ normal(0, prior_sd_gamma);

  // NOTE: implied improper uniform prior U(0, inf) on shape

  // -- IPD likelihood --
  if (dist == 1) { // Exponential model
    for (i in 1:ni_ipd) {
      if (status[i]==1)
        y[i] ~ exponential(theta[i]);
      else
        target += exponential_lccdf(y[i] | theta[i]);
    }
  } else if (dist == 2) { // Weibull model
    for (i in 1:ni_ipd) {
      if (status[i]==1)
        y[i] ~ weibull(shape[study[i]], itheta[i]);
      else
        target += weibull_lccdf(y[i] | shape[study[i]], itheta[i]);
    }
```

```
  } else if (dist == 3) { // Gompertz model
    // Stan does not have Gompertz distribution, instead code the log likelihood
    // contributions directly
    for (i in 1:ni_ipd) {
      if (status[i]==1)
        target += -theta[i]/shape[study[i]] * expm1(shape[study[i]] * y[i]) +
                  eta[i] + (shape[study[i]] * y[i]);
      else
        target += -theta[i]/shape[study[i]] * expm1(shape[study[i]] * y[i]);
    }
  }

  // -- AgD likelihood --
  // Simply add up the log marginal likelihood contributions
  target += log(P_bar);
}
generated quantities {
  // Transform intercepts back to scales
  vector[ns] scale = exp(mu);
  // Log likelihood contributions
  vector[ni_ipd + ni_agd] log_lik;

  // -- Estimate integration error --
  matrix[ni_agd, n_int_thin] P_bar_cum;

  // -- IPD log likelihood --
  if (dist == 1) { // Exponential model
    for (i in 1:ni_ipd) {
      if (status[i]==1)
        log_lik[i] = exponential_lpdf(y[i] | theta[i]);
      else
        log_lik[i] = exponential_lccdf(y[i] | theta[i]);
    }
  } else if (dist == 2) { // Weibull model
    for (i in 1:ni_ipd) {
      if (status[i]==1)
        log_lik[i] = weibull_lpdf(y[i] | shape[study[i]], itheta[i]);
      else
        log_lik[i] = weibull_lccdf(y[i] | shape[study[i]], itheta[i]);
    }
  } else if (dist == 3) { // Gompertz model
    for (i in 1:ni_ipd) {
      if (status[i]==1)
        log_lik[i] = -theta[i]/shape[study[i]] * expm1(shape[study[i]] * y[i]) +
                     eta[i] + (shape[study[i]] * y[i]);
      else
        log_lik[i] = -theta[i]/shape[study[i]] * expm1(shape[study[i]] * y[i]);
    }
  }

  // -- AgD log likelihood --
  log_lik[(ni_ipd + 1):] = log(P_bar);

  // Estimate integration error
  for (j in 1:n_int_thin) {
    for (i in 1:ni_agd) {
      if (status[ni_ipd + i] == 1) {
```

```
        P_bar_cum[i,j] =
          mean(S_ii[((i-1)*nint + 1):((i-1)*nint + j*int_thin)] .*
               h_ii[((i-1)*nint + 1):((i-1)*nint + j*int_thin)]);
      } else {
        P_bar_cum[i,j] = mean(S_ii[((i-1)*nint + 1):((i-1)*nint + j*int_thin)]);
      }
    }
  }
}
```

Running ML-NMR in R A.4

Stan interfaces are available in a variety of computing environments, including R, Stata, and Python.[1] We use R for the analyses throughout this thesis, and run Stan via the `rstan` package. In this section, we outline how to prepare and run ML-NMR models from R.

Generating QMC integration points A.4.1

Firstly, we need to generate QMC integration points (following Section 5.1), which will be passed to the ML-NMR model in Stan as data. We use the `sobol` function from the `randtoolbox` package to generate points from a Sobol' sequence. A Gaussian copula is implemented using the `copula` package (which also includes functions for other copulæ), with which we apply the correlations observed in the IPD to the Sobol' points before transforming to the required marginal distributions using the inverse CDF method. We load the `tidyverse` suite of packages to simplify the process of data manipulation. The following code assumes that all data from the IPD studies are contained in a data frame called `ip_dat`, and the AgD study summaries are contained in a data frame called `ag_dat`. We illustrate with covariates from the plaque psoriasis example in Chapter 6: duration of psoriasis (`durnpso`), previous systemic treatment (`prevsys`), body surface area (`bsa`), weight (`weight`), and psoriatic arthritis (`psa`).

```
# Load packages
library(tidyverse)
library(randtoolbox)
library(copula)

# Assume that IPD are contained in ip_dat, AgD in ag_dat
```

[1] For a current list of interfaces, see `https://mc-stan.org/users/interfaces/`.

```r
# Column studyn is a numeric study indicator
# Column trtn is a numeric treatment indicator

# Scenario setup
ns_ipd <- length(unique(ip_dat$studyn))  # Number of IPD studies
ns_agd <- length(unique(ag_dat$studyn))  # Number of AgD studies
ntrt <- length(unique(c(ip_dat$trtn, ag_dat$trtn)))  # Number of treatments

# Names of covariates to include in analysis (column names in ip_dat)
X_names <- c("durnpso", "prevsys", "bsa", "weight", "psa")

# Set parameters for numerical integration
n_X <- length(X_names)  # Number of covariates
n_int <- 5000  # Number of sample points for numerical integration

# Draw n_int Sobol points in n_X dimensions
sobol_points <- sobol(n_int, n_X)

# Compute the correlation matrix of the covariates in the IPD using a
# Z-transformed weighted average
ipd_cors <-
  ip_dat %>%
  group_by(study) %>%
  do(w = nrow(.) - 3, r = cor(select(., !! X_names), method = "spearman")) %>%
  mutate(wcor = list(w * log((1 + r) / (1 - r)) / 2)) %>%
  unnest(w) %>%
  ungroup()

ipd_cor <- {Reduce(+, ipd_cors$wcor) / sum(ipd_cors$w)} %>%
           {(exp(2 * .) - 1) / (exp(2 * .) + 1)}

diag(ipd_cor) <- 1

# Create copula object
ipd_copula <- normalCopula(P2p(ipd_cor), dim = n_X, dispstr = "un")

# Apply copula to the Sobol points
sobol_points_cor <- as.tibble(cCopula(sobol_points, ipd_copula,
                                      inverse = TRUE)) %>%
                    setNames(X_names)

# Redefine logitnorm::qlogitnorm, the inverse CDF of logit Normal distribution,
# to be parameterised using mean and standard deviation
.lndiff <- function(est, smean, ssd){
  x <- logitnorm::momentsLogitnorm(est[1], est[2])
  (x[1] - smean)^2 + (sqrt(x[2]) - ssd)^2
}

.lnopt <- function(sample_mean, sample_sd) {
  opt <- optim(c(sample_mean, sample_sd), .lndiff,
               smean = sample_mean, ssd = sample_sd)

  if (opt$convergence != 0) {
    warning("Optimisation did not converge, NAs produced.")
    c("mu" = NA, "sigma" = NA)
  } else {
    c("mu" = opt$par[1], "sigma" = opt$par[2])
```

```
  }
}

pars_logitnorm <- function(sample_mean, sample_sd) {
  if (length(sample_mean) != length(sample_sd))
    stop("Parameters not same length.")
  if (any(sample_mean > 1 | sample_mean < 0))
    stop("Sample mean not in [0,1]. Have you rescaled?")

  as.data.frame(do.call(rbind, mapply(.lnopt, sample_mean, sample_sd,
                                      SIMPLIFY = FALSE)))
}

qlogitnorm <- function(p, mean, sd){
  pars <- pars_logitnorm(mean, sd)
  logitnorm::qlogitnorm(p, pars$mu, pars$sigma)
}

# Use inverse CDF method to transform points to required marginal distributions
ag_xpoints <- ag_dat %>%
  rowwise() %>%
  do(durnpso = qgamma(sobol_points_cor$durnpso,
                      (.$durnpso_mean / .$durnpso_sd)^2 ,
                      .$durnpso_mean / .$durnpso_sd^2),
     prevsys = qbinom(sobol_points_cor$prevsys, 1, .$prevsys / 100),
     bsa = qlogitnorm(sobol_points_cor$bsa,
                      .$bsa_mean / 100, .$bsa_sd / 100) * 100,
     weight = qgamma(sobol_points_cor$weight,
                     (.$weight_mean / .$weight_sd)^2 ,
                     .$weight_mean / .$weight_sd^2),
     psa = qbinom(sobol_points_cor$psa, 1, .$psa / 100)
     )  %>%
  unnest(.id = "ag_id") %>%
  select(ag_id, everything(), -matches("^ag_id[0-9]+$"))
```

Calling Stan A.4.2

Now that we have the QMC integration points, we can run the ML-NMR model in Stan. The following code illustrates the process of fitting a FE ML-NMR model for the plaque psoriasis example in Chapter 6, where binary PASI 75 outcomes are modelled using a probit link function, and shared EM interactions are assumed for the class of interleukin-17A blockers (ixekizumab and secukinumab). The code implements centring and QR decomposition, for efficient computation (Section 5.2.1). The corresponding ML-NMR Stan code is given in Section A.1.2.1.

```
# Load rstan package and set up parallel processing
library(rstan)
rstan_options(auto_write = TRUE)
```

```r
options(mc.cores = parallel::detectCores())

# Function to get overall weighted mean of a covariate from IPD and AgD
global_mean <- function(v,  # Name of covariate
                        ss,  # Sample size variable in AgD
                        ipd,  # IPD data frame
                        agd,  # AgD data frame
                        na.rm = FALSE) { # Remove missing values?

  v <- enquo(v)
  ss <- enquo(ss)

  v_ipd <- pull(ipd, !! v)

  if (na.rm & any(is.na(v_ipd))) {
    num_na <- sum(is.na(v_ipd))
    v_ipd <- v_ipd[!is.na(v_ipd)]
    message("Removed ", num_na, " missing values in IPD.")
  } else if (any(is.na(v_ipd))) {
    warning("IPD has missing values, NA returned.", call. = FALSE)
    return(NA)
  }

  v_agd <- pull(agd, !! paste0(quo_name(enquo(v)), "_mean"))
  ss_agd <- pull(agd, !! ss)

  drop((sum(v_ipd) + ss_agd %*% v_agd) / (length(v_ipd) + sum(ss_agd)))
}

# Calculate an overall mean value for each covariate, to use for centring
gmean_durnpso <- global_mean(durnpso, sample_size_w0,
                             ip_dat, ag_dat)
gmean_bsa <- global_mean(bsa, sample_size_w0,
                         ip_dat, ag_dat)
gmean_weight <- global_mean(weight, sample_size_w0,
                            ip_dat, ag_dat)

# Create a big data frame with the AgD integration points added on to the bottom
# of the IPD
stan_xdat <- ip_dat %>%
  select(studyn, trtn,
         prevsys, durnpso, bsa, weight, psa) %>%
  bind_rows(
    ag_dat %>% transmute(ag_id = 1:n(), studyn, trtn) %>%
      full_join(ag_xpoints, by = "ag_id") %>%
      select(studyn, trtn, prevsys, durnpso, bsa, weight, psa)
  ) %>%
  mutate(study = as.factor(studyn), trt = as.factor(trtn),
         trtclass = recode_factor(trtn,
                      "1" = 1,  # Placebo
                      "2" = 2, "3" = 2, "5" = 2, "6" = 2,  # IL-17 blockers
                      "4" = 3,  # TNFa blocker
                      "7" = 4),  # IL-12,23 blocker
         prevsys = prevsys,
         durnpso = durnpso - gmean_durnpso,
         bsa = bsa - gmean_bsa,
         weight = weight - gmean_weight,
```

```r
         psa = psa)

# Now we can simply use the model.matrix function to create the model matrix.
# To be compatible with the Stan code, the resulting model matrix columns should
# be in the order: study baselines, treatment parameters, PVs, EM interactions.
X_all <- model.matrix(~ -1 + study + trt +
                        prevsys + durnpso + bsa + weight + psa +
                        (prevsys + durnpso + bsa + weight + psa):trtclass,
                      data = stan_xdat)

# Then get the thin QR decomposition
X_all_qr <- qr(X_all)
X_all_Q <- qr.Q(X_all_qr) * sqrt(nrow(X_all) - 1)
X_all_R <- qr.R(X_all_qr)[, sort.list(X_all_qr$pivot)] / sqrt(nrow(X_all) - 1)
X_all_R_inv <- solve(X_all_R)

# Construct the data list for Stan
pasi75_shared_standat <- list(
  # Constants
  ns_ipd = ns_ipd,
  ns_agd = ns_agd,
  ni_ipd = nrow(ip_dat),
  ni_agd = nrow(ag_dat),
  nt = ntrt,
  nint = n_int,
  nPV = n_X,
  nEM = 3 * n_X,  # Shared EM model, so number of active trt classes * n_X
  int_thin = 100,
  # IPD
  y = ip_dat$pasi75,
  # AgD
  ag_n = ag_dat$pasi75_n,
  ag_r = ag_dat$pasi75_r,
  ag_trt = ag_dat$trtn,
  ag_study = ag_dat$studyn,
  # QR decomposition
  Q = X_all_Q,
  R_inv = X_all_R_inv,
  # Priors
  prior_sd_beta0 = 100,
  prior_sd_beta1 = 100,
  prior_sd_beta2 = 100,
  prior_sd_gamma = 100)

# Run Stan
# The ML-NMR model code is saved in file ML-NMR_probit_indep_twoparbin_qr.stan
pasi75_shared_stan <- stan("./ML-NMR_probit_indep_twoparbin_qr.stan",
  data = pasi75_shared_standat,
  pars = c("beta0", "beta1", "beta2", "gamma", "nprime", "pprime",
           "p_bar_cum", "p2_bar_cum",
           "log_lik", "resdev", "r_hat", "lp__"),
  iter = 2000,
  chains = 4,
  init_r = 0.2)

# Diagnostic checks
check_hmc_diagnostics(pasi75_shared_stan)
```

```
get_sampler_params(pasi75_shared_stan, inc_warmup = FALSE) %>%
  map_dfr(~ as_tibble(.) %>% summarise_at(1:4, mean))

# The shinystan package can also be used to assess convergence interactively
# library(shinystan)
# pasi75_shared_sso <- drop_parameters(as.shinystan(pasi75_shared_stan),
#   pars = c("log_lik", "p_bar_cum", "p2_bar_cum", "resdev", "r_hat"))
# launch_shinystan(pasi75_shared_sso)

# Print parameter estimates
print(pasi75_shared_stan, pars = c("beta0", "beta1", "beta2", "gamma"))
```

Appendix B

Tables of simulation study results

B.1 Scenario a

Table B.1 Simulation results for scenario a, adjusting for all effect modifiers. Sample size N is varied between 100, 500, and 1000. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario	Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$N = 100$	-0.356 (0.013)	0.603 (0.010)	0.528 (0.002)	87.5 (0.7)
		$N = 500$	-0.056 (0.005)	0.215 (0.003)	0.207 (<0.001)	93.8 (0.5)
		$N = 1000$	-0.032 (0.003)	0.148 (0.002)	0.145 (<0.001)	94.4 (0.5)
	$d_{AC(AB)}$	$N = 100$	-0.286 (0.032)	1.420 (0.022)	1.219 (0.004)	91.0 (0.6)
		$N = 500$	-0.038 (0.011)	0.487 (0.008)	0.471 (<0.001)	94.6 (0.5)
		$N = 1000$	-0.005 (0.008)	0.336 (0.005)	0.328 (<0.001)	94.1 (0.5)
	$d_{BC(AB)}$	$N = 100$	0.070 (0.034)	1.513 (0.024)	1.301 (0.003)	90.5 (0.7)
		$N = 500$	0.019 (0.012)	0.522 (0.008)	0.507 (<0.001)	94.1 (0.5)
		$N = 1000$	0.026 (0.008)	0.361 (0.006)	0.354 (<0.001)	95.0 (0.5)
	$d_{AB(AC)}$	$N = 100$	-0.258 (0.032)	1.416 (0.022)	1.204 (0.004)	90.2 (0.7)
		$N = 500$	-0.041 (0.011)	0.474 (0.007)	0.462 (<0.001)	94.7 (0.5)
		$N = 1000$	-0.040 (0.007)	0.323 (0.005)	0.322 (<0.001)	95.0 (0.5)
	$d_{AC(AC)}$	$N = 100$	-0.188 (0.012)	0.542 (0.009)	0.517 (0.001)	93.1 (0.6)
		$N = 500$	-0.023 (0.005)	0.216 (0.003)	0.210 (<0.001)	94.3 (0.5)
		$N = 1000$	-0.014 (0.003)	0.153 (0.002)	0.148 (<0.001)	94.1 (0.5)
	$d_{BC(AC)}$	$N = 100$	0.070 (0.034)	1.513 (0.024)	1.301 (0.003)	90.5 (0.7)
		$N = 500$	0.019 (0.012)	0.522 (0.008)	0.507 (<0.001)	94.1 (0.5)
		$N = 1000$	0.026 (0.008)	0.361 (0.006)	0.354 (<0.001)	95.0 (0.5)
STC	$d_{AB(AC)}$	$N = 100$	-0.124 (0.028)	1.232 (0.019)	1.123 (0.003)	94.0 (0.5)
		$N = 500$	-0.020 (0.010)	0.463 (0.007)	0.457 (<0.001)	95.3 (0.5)
		$N = 1000$	-0.029 (0.007)	0.319 (0.005)	0.320 (<0.001)	95.2 (0.5)
	$d_{BC(AC)}$	$N = 100$	0.092 (0.030)	1.326 (0.021)	1.220 (0.003)	93.7 (0.5)
		$N = 500$	0.021 (0.011)	0.510 (0.008)	0.502 (<0.001)	94.2 (0.5)
		$N = 1000$	0.027 (0.008)	0.357 (0.006)	0.352 (<0.001)	95.2 (0.5)
MAIC	$d_{AB(AC)}$	$N = 100$	-3.215 (0.228)	10.203 (0.161)	- (-)	79.1 (0.9)
		$N = 500$	-0.199 (0.028)	1.249 (0.020)	- (-)	88.3 (0.7)
		$N = 1000$	-0.101 (0.015)	0.683 (0.011)	0.666 (0.005)	89.5 (0.7)
	$d_{BC(AC)}$	$N = 100$	3.176 (0.228)	10.214 (0.162)	- (-)	- (-)
		$N = 500$	0.200 (0.028)	1.260 (0.020)	- (-)	94.6 (0.5)
		$N = 1000$	0.099 (0.016)	0.699 (0.011)	0.682 (0.005)	90.7 (0.7)
Bucher	$d_{AB(AC)}$	$N = 100$	-0.359 (0.011)	0.475 (0.008)	0.461 (<0.001)	87.5 (0.7)
		$N = 500$	-0.316 (0.005)	0.207 (0.003)	0.202 (<0.001)	65.0 (1.1)
		$N = 1000$	-0.317 (0.003)	0.145 (0.002)	0.143 (<0.001)	39.7 (1.1)
	$d_{BC(AC)}$	$N = 100$	0.327 (0.015)	0.678 (0.011)	0.664 (<0.001)	92.0 (0.6)
		$N = 500$	0.318 (0.007)	0.295 (0.005)	0.290 (<0.001)	80.2 (0.9)
		$N = 1000$	0.315 (0.005)	0.210 (0.003)	0.205 (<0.001)	66.5 (1.1)

Table B.2 Simulation results for scenario a, only adjusting for one of two effect modifiers. Sample size N is varied between 100, 500, and 1000. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario	Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$N = 100$	-0.223 (0.012)	0.544 (0.009)	0.498 (0.001)	92.2 (0.6)
		$N = 500$	-0.035 (0.005)	0.212 (0.003)	0.205 (<0.001)	94.3 (0.5)
		$N = 1000$	-0.020 (0.003)	0.146 (0.002)	0.144 (<0.001)	94.4 (0.5)
	$d_{AC(AB)}$	$N = 100$	-0.025 (0.022)	0.962 (0.015)	0.892 (0.002)	93.2 (0.6)
		$N = 500$	0.115 (0.008)	0.370 (0.006)	0.358 (<0.001)	92.5 (0.6)
		$N = 1000$	0.140 (0.006)	0.258 (0.004)	0.251 (<0.001)	90.8 (0.6)
	$d_{BC(AB)}$	$N = 100$	0.198 (0.025)	1.098 (0.017)	1.004 (0.002)	92.7 (0.6)
		$N = 500$	0.150 (0.009)	0.416 (0.007)	0.408 (<0.001)	93.3 (0.6)
		$N = 1000$	0.160 (0.007)	0.293 (0.005)	0.286 (<0.001)	91.3 (0.6)
	$d_{AB(AC)}$	$N = 100$	-0.329 (0.022)	0.982 (0.016)	0.877 (0.003)	91.7 (0.6)
		$N = 500$	-0.165 (0.008)	0.353 (0.006)	0.351 (<0.001)	93.0 (0.6)
		$N = 1000$	-0.170 (0.006)	0.247 (0.004)	0.245 (<0.001)	90.1 (0.7)
	$d_{AC(AC)}$	$N = 100$	-0.131 (0.012)	0.519 (0.008)	0.502 (0.001)	93.8 (0.5)
		$N = 500$	-0.015 (0.005)	0.215 (0.003)	0.209 (<0.001)	94.2 (0.5)
		$N = 1000$	-0.010 (0.003)	0.152 (0.002)	0.147 (<0.001)	93.9 (0.5)
	$d_{BC(AC)}$	$N = 100$	0.198 (0.025)	1.098 (0.017)	1.004 (0.002)	92.7 (0.6)
		$N = 500$	0.150 (0.009)	0.416 (0.007)	0.408 (<0.001)	93.3 (0.6)
		$N = 1000$	0.160 (0.007)	0.293 (0.005)	0.286 (<0.001)	91.3 (0.6)
STC	$d_{AB(AC)}$	$N = 100$	-0.235 (0.020)	0.885 (0.014)	0.833 (0.003)	94.9 (0.5)
		$N = 500$	-0.149 (0.008)	0.347 (0.005)	0.347 (<0.001)	94.3 (0.5)
		$N = 1000$	-0.162 (0.005)	0.245 (0.004)	0.244 (<0.001)	91.0 (0.6)
	$d_{BC(AC)}$	$N = 100$	0.203 (0.022)	1.002 (0.016)	0.960 (0.002)	94.8 (0.5)
		$N = 500$	0.151 (0.009)	0.409 (0.006)	0.405 (<0.001)	93.7 (0.5)
		$N = 1000$	0.160 (0.007)	0.291 (0.005)	0.285 (<0.001)	91.6 (0.6)
MAIC	$d_{AB(AC)}$	$N = 100$	-0.454 (0.032)	1.417 (0.022)	- (-)	83.3 (0.8)
		$N = 500$	-0.169 (0.010)	0.439 (0.007)	0.444 (0.001)	91.0 (0.6)
		$N = 1000$	-0.178 (0.007)	0.311 (0.005)	0.303 (<0.001)	89.3 (0.7)
	$d_{BC(AC)}$	$N = 100$	0.422 (0.034)	1.498 (0.024)	- (-)	94.9 (0.5)
		$N = 500$	0.171 (0.011)	0.490 (0.008)	0.490 (0.001)	92.7 (0.6)
		$N = 1000$	0.176 (0.008)	0.346 (0.005)	0.337 (<0.001)	91.5 (0.6)
Bucher	$d_{AB(AC)}$	$N = 100$	-0.359 (0.011)	0.475 (0.008)	0.461 (<0.001)	87.5 (0.7)
		$N = 500$	-0.316 (0.005)	0.207 (0.003)	0.202 (<0.001)	65.0 (1.1)
		$N = 1000$	-0.317 (0.003)	0.145 (0.002)	0.143 (<0.001)	39.7 (1.1)
	$d_{BC(AC)}$	$N = 100$	0.327 (0.015)	0.678 (0.011)	0.664 (<0.001)	92.0 (0.6)
		$N = 500$	0.318 (0.007)	0.295 (0.005)	0.290 (<0.001)	80.2 (0.9)
		$N = 1000$	0.315 (0.005)	0.210 (0.003)	0.205 (<0.001)	66.5 (1.1)

B.2 Scenario b

Table B.3 Simulation results for scenario b, adjusting for all effect modifiers. Strength of effect modification is varied from weak to strong. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario	Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.056 (0.005)	0.215 (0.003)	0.207 (<0.001)	93.8 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.066 (0.005)	0.225 (0.004)	0.216 (<0.001)	93.2 (0.6)
	$d_{AC(AB)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.038 (0.011)	0.487 (0.008)	0.471 (<0.001)	94.6 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.137 (0.012)	0.545 (0.009)	0.503 (<0.001)	92.2 (0.6)
	$d_{BC(AB)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.019 (0.012)	0.522 (0.008)	0.507 (<0.001)	94.1 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.071 (0.012)	0.545 (0.009)	0.516 (<0.001)	93.0 (0.6)
	$d_{AB(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.041 (0.011)	0.474 (0.007)	0.462 (<0.001)	94.7 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.037 (0.011)	0.501 (0.008)	0.470 (<0.001)	93.5 (0.5)
	$d_{AC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.023 (0.005)	0.216 (0.003)	0.210 (<0.001)	94.3 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.034 (0.005)	0.213 (0.003)	0.214 (<0.001)	94.8 (0.5)
	$d_{BC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.019 (0.012)	0.522 (0.008)	0.507 (<0.001)	94.1 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.071 (0.012)	0.545 (0.009)	0.516 (<0.001)	93.0 (0.6)
STC	$d_{AB(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.020 (0.010)	0.463 (0.007)	0.457 (<0.001)	95.3 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.023 (0.011)	0.488 (0.008)	0.464 (<0.001)	94.2 (0.5)
	$d_{BC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.021 (0.011)	0.510 (0.008)	0.502 (<0.001)	94.2 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.033 (0.012)	0.529 (0.008)	0.509 (<0.001)	93.9 (0.5)
MAIC	$d_{AB(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.199 (0.028)	1.249 (0.020)	- (-)	88.3 (0.7)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.128 (0.022)	0.968 (0.015)	- (-)	92.4 (0.6)
	$d_{BC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.200 (0.028)	1.260 (0.020)	- (-)	94.6 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.118 (0.022)	0.985 (0.016)	- (-)	95.0 (0.5)
Bucher	$d_{AB(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.316 (0.005)	0.207 (0.003)	0.202 (<0.001)	65.0 (1.1)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-1.468 (0.004)	0.200 (0.003)	0.200 (<0.001)	0.0 (0.0)
	$d_{BC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.318 (0.007)	0.295 (0.005)	0.290 (<0.001)	80.2 (0.9)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	1.458 (0.006)	0.283 (0.004)	0.289 (<0.001)	<0.1 (<0.1)

Table B.4 Simulation results for scenario b, only adjusting for one of two effect modifiers. Strength of effect modification is varied from weak to strong. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario	Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.035 (0.005)	0.212 (0.003)	0.205 (<0.001)	94.3 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.006 (0.005)	0.213 (0.003)	0.209 (<0.001)	94.7 (0.5)
	$d_{AC(AB)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.115 (0.008)	0.370 (0.006)	0.358 (<0.001)	92.5 (0.6)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.670 (0.008)	0.377 (0.006)	0.370 (<0.001)	56.4 (1.1)
	$d_{BC(AB)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.150 (0.009)	0.416 (0.007)	0.408 (<0.001)	93.3 (0.6)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.664 (0.009)	0.402 (0.006)	0.405 (<0.001)	61.9 (1.1)
	$d_{AB(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.165 (0.008)	0.353 (0.006)	0.351 (<0.001)	93.0 (0.6)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.690 (0.008)	0.346 (0.005)	0.346 (<0.001)	46.9 (1.1)
	$d_{AC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.015 (0.005)	0.215 (0.003)	0.209 (<0.001)	94.2 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.026 (0.005)	0.210 (0.003)	0.212 (<0.001)	95.2 (0.5)
	$d_{BC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.150 (0.009)	0.416 (0.007)	0.408 (<0.001)	93.3 (0.6)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.664 (0.009)	0.402 (0.006)	0.405 (<0.001)	61.9 (1.1)
STC	$d_{AB(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.149 (0.008)	0.347 (0.005)	0.347 (<0.001)	94.3 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.690 (0.008)	0.339 (0.005)	0.343 (<0.001)	46.7 (1.1)
	$d_{BC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.151 (0.009)	0.409 (0.006)	0.405 (<0.001)	93.7 (0.5)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.679 (0.009)	0.393 (0.006)	0.401 (<0.001)	59.8 (1.1)
MAIC	$d_{AB(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.169 (0.010)	0.439 (0.007)	0.444 (0.001)	91.0 (0.6)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-0.708 (0.009)	0.417 (0.007)	0.426 (0.001)	56.7 (1.1)
	$d_{BC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.171 (0.011)	0.490 (0.008)	0.490 (0.001)	92.7 (0.6)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	0.698 (0.010)	0.456 (0.007)	0.474 (0.001)	69.4 (1.0)
Bucher	$d_{AB(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	-0.316 (0.005)	0.207 (0.003)	0.202 (<0.001)	65.0 (1.1)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	-1.468 (0.004)	0.200 (0.003)	0.200 (<0.001)	0.0 (0.0)
	$d_{BC(AC)}$	$\beta_B = \beta_C = 0.1\sigma_{X(AB)}$	0.318 (0.007)	0.295 (0.005)	0.290 (<0.001)	80.2 (0.9)
		$\beta_B = \beta_C = 0.5\sigma_{X(AB)}$	1.458 (0.006)	0.283 (0.004)	0.289 (<0.001)	<0.1 (<0.1)

B.3 Scenario c

Table B.5 Simulation results for scenario c, adjusting for all effect modifiers. The shared effect modifier assumption is broken. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario	Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	-0.061 (0.005)	0.213 (0.003)	0.207 (<0.001)	93.2 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	-0.066 (0.005)	0.225 (0.004)	0.216 (<0.001)	93.1 (0.6)
	$d_{AC(AB)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	1.192 (0.011)	0.502 (0.008)	0.472 (<0.001)	30.2 (1.0)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	-1.397 (0.012)	0.549 (0.009)	0.503 (<0.001)	19.8 (0.9)
	$d_{BC(AB)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	1.253 (0.012)	0.537 (0.008)	0.508 (<0.001)	30.9 (1.0)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	-1.331 (0.012)	0.549 (0.009)	0.516 (<0.001)	27.2 (1.0)
	$d_{AB(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	-0.028 (0.011)	0.494 (0.008)	0.463 (<0.001)	93.3 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	0.037 (0.011)	0.501 (0.008)	0.470 (<0.001)	93.5 (0.5)
	$d_{AC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	-0.035 (0.005)	0.207 (0.003)	0.211 (<0.001)	94.8 (0.5)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	-0.034 (0.005)	0.216 (0.003)	0.214 (<0.001)	95.0 (0.5)
	$d_{BC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	-0.007 (0.012)	0.537 (0.008)	0.508 (<0.001)	93.0 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	-0.071 (0.012)	0.549 (0.009)	0.516 (<0.001)	92.6 (0.6)
STC	$d_{AB(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	-0.007 (0.011)	0.483 (0.008)	0.457 (<0.001)	93.5 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	0.023 (0.011)	0.488 (0.008)	0.464 (<0.001)	94.2 (0.5)
	$d_{BC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	-0.003 (0.012)	0.525 (0.008)	0.502 (<0.001)	93.5 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	-0.033 (0.012)	0.531 (0.008)	0.509 (<0.001)	93.5 (0.6)
MAIC	$d_{AB(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	-0.212 (0.023)	1.029 (0.016)	- (-)	88.0 (0.7)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	-0.128 (0.022)	0.968 (0.015)	- (-)	92.4 (0.6)
	$d_{BC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	0.202 (0.023)	1.047 (0.017)	- (-)	95.2 (0.5)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	0.118 (0.022)	0.987 (0.016)	- (-)	95.0 (0.5)
Bucher	$d_{AB(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	-0.320 (0.005)	0.205 (0.003)	0.202 (<0.001)	66.1 (1.1)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	-1.468 (0.004)	0.200 (0.003)	0.200 (<0.001)	0.0 (0.0)
	$d_{BC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}$, $\beta_C = 0.5\sigma_{X(AB)}$	0.309 (0.006)	0.286 (0.005)	0.290 (<0.001)	81.3 (0.9)
		$\beta_B = 0.5\sigma_{X(AB)}$, $\beta_C = 0.1\sigma_{X(AB)}$	1.459 (0.006)	0.283 (0.004)	0.289 (<0.001)	0.1 (<0.1)

Table B.6 Simulation results for scenario c, only adjusting for one of two effect modifiers. The shared effect modifier assumption is broken. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario	Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	-0.038 (0.005)	0.209 (0.003)	0.205 (<0.001)	93.9 (0.5)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	0.006 (0.005)	0.213 (0.003)	0.209 (<0.001)	95.0 (0.5)
	$d_{AC(AB)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	1.355 (0.008)	0.368 (0.006)	0.359 (<0.001)	4.3 (0.5)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	-0.590 (0.009)	0.381 (0.006)	0.370 (<0.001)	64.4 (1.1)
	$d_{BC(AB)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	1.394 (0.009)	0.414 (0.007)	0.409 (<0.001)	7.8 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	-0.596 (0.009)	0.405 (0.006)	0.405 (<0.001)	68.3 (1.0)
	$d_{AB(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	-0.160 (0.008)	0.362 (0.006)	0.351 (<0.001)	92.0 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	-0.690 (0.008)	0.346 (0.005)	0.346 (<0.001)	46.9 (1.1)
	$d_{AC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	-0.026 (0.005)	0.206 (0.003)	0.210 (<0.001)	95.0 (0.5)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	-0.026 (0.005)	0.211 (0.003)	0.212 (<0.001)	95.0 (0.5)
	$d_{BC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	0.134 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.9 (0.5)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	0.664 (0.009)	0.405 (0.006)	0.405 (<0.001)	62.0 (1.1)
STC	$d_{AB(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	-0.145 (0.008)	0.355 (0.006)	0.348 (<0.001)	93.0 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	-0.690 (0.008)	0.339 (0.005)	0.343 (<0.001)	46.7 (1.1)
	$d_{BC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	0.134 (0.009)	0.407 (0.006)	0.405 (<0.001)	94.5 (0.5)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	0.680 (0.009)	0.396 (0.006)	0.401 (<0.001)	60.1 (1.1)
MAIC	$d_{AB(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	-0.163 (0.010)	0.441 (0.007)	0.445 (0.001)	91.1 (0.6)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	-0.708 (0.009)	0.417 (0.007)	0.426 (0.001)	56.7 (1.1)
	$d_{BC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	0.153 (0.011)	0.479 (0.008)	0.491 (0.001)	93.9 (0.5)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	0.698 (0.010)	0.459 (0.007)	0.474 (0.001)	69.0 (1.0)
Bucher	$d_{AB(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	-0.320 (0.005)	0.205 (0.003)	0.202 (<0.001)	66.1 (1.1)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	-1.468 (0.004)	0.200 (0.003)	0.200 (<0.001)	0.0 (0.0)
	$d_{BC(AC)}$	$\beta_B = 0.1\sigma_{X(AB)}, \beta_C = 0.5\sigma_{X(AB)}$	0.309 (0.006)	0.286 (0.005)	0.290 (<0.001)	81.3 (0.9)
		$\beta_B = 0.5\sigma_{X(AB)}, \beta_C = 0.1\sigma_{X(AB)}$	1.459 (0.006)	0.283 (0.004)	0.289 (<0.001)	0.1 (<0.1)

B.4 Scenario d

Table B.7 Simulation results for scenario d, adjusting for all effect modifiers. The correlation between covariates is varied between 0, 0.25, and 0.5. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario	Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.059 (0.005)	0.211 (0.003)	0.207 (<0.001)	93.7 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.056 (0.005)	0.215 (0.003)	0.207 (<0.001)	93.8 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.059 (0.005)	0.212 (0.003)	0.207 (<0.001)	93.6 (0.5)
	$d_{AC(AB)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.051 (0.012)	0.530 (0.008)	0.514 (<0.001)	94.5 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.038 (0.011)	0.487 (0.008)	0.471 (<0.001)	94.6 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.051 (0.010)	0.455 (0.007)	0.442 (<0.001)	94.0 (0.5)
	$d_{BC(AB)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.008 (0.013)	0.571 (0.009)	0.548 (<0.001)	94.3 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.019 (0.012)	0.522 (0.008)	0.507 (<0.001)	94.1 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.008 (0.011)	0.503 (0.008)	0.481 (<0.001)	94.5 (0.5)
	$d_{AB(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.045 (0.012)	0.524 (0.008)	0.506 (<0.001)	94.4 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.041 (0.011)	0.474 (0.007)	0.462 (<0.001)	94.7 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.046 (0.010)	0.449 (0.007)	0.432 (<0.001)	94.4 (0.5)
	$d_{AC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.037 (0.005)	0.211 (0.003)	0.211 (<0.001)	94.4 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.023 (0.005)	0.216 (0.003)	0.210 (<0.001)	94.3 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.038 (0.005)	0.211 (0.003)	0.211 (<0.001)	94.5 (0.5)
	$d_{BC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.008 (0.013)	0.571 (0.009)	0.548 (<0.001)	94.3 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.019 (0.012)	0.522 (0.008)	0.507 (<0.001)	94.1 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.008 (0.011)	0.503 (0.008)	0.481 (<0.001)	94.5 (0.5)
STC	$d_{AB(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.024 (0.011)	0.511 (0.008)	0.500 (<0.001)	95.2 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.020 (0.010)	0.463 (0.007)	0.457 (<0.001)	95.3 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.025 (0.010)	0.439 (0.007)	0.427 (<0.001)	95.2 (0.5)
	$d_{BC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.011 (0.012)	0.557 (0.009)	0.542 (<0.001)	94.8 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.021 (0.011)	0.510 (0.008)	0.502 (<0.001)	94.2 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.011 (0.011)	0.492 (0.008)	0.475 (<0.001)	94.8 (0.5)
MAIC	$d_{AB(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.621 (0.067)	2.975 (0.047)	- (-)	87.6 (0.7)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.199 (0.028)	1.249 (0.020)	- (-)	88.3 (0.7)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.139 (0.018)	0.792 (0.013)	- (-)	90.8 (0.6)
	$d_{BC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.609 (0.067)	2.991 (0.047)	- (-)	100.0 (0.0)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.200 (0.028)	1.260 (0.020)	- (-)	94.6 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.126 (0.019)	0.828 (0.013)	- (-)	93.4 (0.6)
Bucher	$d_{AB(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.320 (0.005)	0.203 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.316 (0.005)	0.207 (0.003)	0.202 (<0.001)	65.0 (1.1)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.317 (0.005)	0.204 (0.003)	0.202 (<0.001)	65.8 (1.1)
	$d_{BC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.307 (0.007)	0.296 (0.005)	0.290 (<0.001)	81.0 (0.9)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.318 (0.007)	0.295 (0.005)	0.290 (<0.001)	80.2 (0.9)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.304 (0.007)	0.298 (0.005)	0.290 (<0.001)	81.2 (0.9)

Table B.8 Simulation results for scenario d, only adjusting for one of two effect modifiers. The correlation between covariates is varied between 0, 0.25, and 0.5. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario	Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.037 (0.005)	0.208 (0.003)	0.205 (<0.001)	94.1 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.035 (0.005)	0.212 (0.003)	0.205 (<0.001)	94.3 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.037 (0.005)	0.208 (0.003)	0.205 (<0.001)	94.2 (0.5)
	$d_{AC(AB)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.145 (0.008)	0.366 (0.006)	0.359 (<0.001)	92.2 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.115 (0.008)	0.370 (0.006)	0.358 (<0.001)	92.5 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.071 (0.008)	0.365 (0.006)	0.359 (<0.001)	94.1 (0.5)
	$d_{BC(AB)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.182 (0.009)	0.422 (0.007)	0.410 (<0.001)	92.2 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.150 (0.009)	0.416 (0.007)	0.408 (<0.001)	93.3 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.107 (0.009)	0.421 (0.007)	0.408 (<0.001)	93.2 (0.6)
	$d_{AB(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.211 (0.008)	0.359 (0.006)	0.352 (<0.001)	90.3 (0.7)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.165 (0.008)	0.353 (0.006)	0.351 (<0.001)	93.0 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.137 (0.008)	0.357 (0.006)	0.351 (<0.001)	92.8 (0.6)
	$d_{AC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.029 (0.005)	0.209 (0.003)	0.210 (<0.001)	94.7 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.015 (0.005)	0.215 (0.003)	0.209 (<0.001)	94.2 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.030 (0.005)	0.210 (0.003)	0.210 (<0.001)	94.5 (0.5)
	$d_{BC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.182 (0.009)	0.422 (0.007)	0.410 (<0.001)	92.2 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.150 (0.009)	0.416 (0.007)	0.408 (<0.001)	93.3 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.107 (0.009)	0.421 (0.007)	0.408 (<0.001)	93.2 (0.6)
STC	$d_{AB(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.195 (0.008)	0.352 (0.006)	0.349 (<0.001)	91.4 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.149 (0.008)	0.347 (0.005)	0.347 (<0.001)	94.3 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.122 (0.008)	0.351 (0.006)	0.347 (<0.001)	93.7 (0.5)
	$d_{BC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.182 (0.009)	0.415 (0.007)	0.406 (<0.001)	92.4 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.151 (0.009)	0.409 (0.006)	0.405 (<0.001)	93.7 (0.5)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.109 (0.009)	0.414 (0.007)	0.405 (<0.001)	93.4 (0.6)
MAIC	$d_{AB(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.222 (0.010)	0.449 (0.007)	0.446 (0.001)	88.3 (0.7)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.169 (0.010)	0.439 (0.007)	0.444 (0.001)	91.0 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.147 (0.010)	0.444 (0.007)	0.444 (0.001)	91.1 (0.6)
	$d_{BC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.209 (0.011)	0.499 (0.008)	0.492 (0.001)	91.8 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.171 (0.011)	0.490 (0.008)	0.490 (0.001)	92.7 (0.6)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.134 (0.011)	0.494 (0.008)	0.491 (0.001)	93.1 (0.6)
Bucher	$d_{AB(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	-0.320 (0.005)	0.203 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.316 (0.005)	0.207 (0.003)	0.202 (<0.001)	65.0 (1.1)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	-0.317 (0.005)	0.204 (0.003)	0.202 (<0.001)	65.8 (1.1)
	$d_{BC(AC)}$	$\rho_{(AB)} = \rho_{(AC)} = 0$	0.307 (0.007)	0.296 (0.005)	0.290 (<0.001)	81.0 (0.9)
		$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.318 (0.007)	0.295 (0.005)	0.290 (<0.001)	80.2 (0.9)
		$\rho_{(AB)} = \rho_{(AC)} = 0.5$	0.304 (0.007)	0.298 (0.005)	0.290 (<0.001)	81.2 (0.9)

B.5 Scenarios e and f

Table B.9 Simulation results for scenarios e and f, adjusting for all effect modifiers. The between-study overlap and covariate-outcome relationship are varied jointly. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario		Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$\kappa = 0$	linear $q(\cdot)$	-0.050 (0.005)	0.212 (0.003)	0.207 (<0.001)	93.3 (0.6)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.018 (0.004)	0.197 (0.003)	0.197 (<0.001)	94.7 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.050 (0.005)	0.212 (0.003)	0.207 (<0.001)	93.4 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.018 (0.004)	0.197 (0.003)	0.197 (<0.001)	94.5 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	-0.050 (0.005)	0.212 (0.003)	0.207 (<0.001)	93.8 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.018 (0.004)	0.198 (0.003)	0.196 (<0.001)	94.5 (0.5)
	$d_{AC(AB)}$	$\kappa = 0$	linear $q(\cdot)$	-0.083 (0.031)	1.394 (0.022)	1.338 (0.001)	94.0 (0.5)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.317 (0.029)	1.319 (0.021)	1.280 (0.001)	93.0 (0.6)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.052 (0.011)	0.488 (0.008)	0.470 (<0.001)	93.7 (0.5)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.029 (0.011)	0.476 (0.008)	0.464 (<0.001)	94.2 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	-0.042 (0.006)	0.247 (0.004)	0.243 (<0.001)	94.5 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.011 (0.006)	0.264 (0.004)	0.260 (<0.001)	94.8 (0.5)
	$d_{BC(AB)}$	$\kappa = 0$	linear $q(\cdot)$	-0.033 (0.032)	1.409 (0.022)	1.346 (0.001)	94.1 (0.5)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.299 (0.030)	1.338 (0.021)	1.297 (0.001)	93.0 (0.6)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.002 (0.012)	0.532 (0.008)	0.507 (<0.001)	93.3 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.011 (0.012)	0.520 (0.008)	0.506 (<0.001)	94.5 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	0.008 (0.007)	0.326 (0.005)	0.316 (<0.001)	94.3 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.007 (0.007)	0.331 (0.005)	0.326 (<0.001)	94.4 (0.5)
	$d_{AB(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.006 (0.031)	1.390 (0.022)	1.330 (0.001)	94.0 (0.5)

Table B.9 (continued)

Method	Contrast	Scenario		Bias	Empirical SE	Model SE	Coverage
		$\kappa = 0$	non-linear $q(\cdot)$	0.315 (0.029)	1.313 (0.021)	1.276 (0.001)	92.8 (0.6)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.036 (0.011)	0.485 (0.008)	0.462 (<0.001)	93.5 (0.5)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.009 (0.010)	0.460 (0.007)	0.449 (<0.001)	94.0 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	-0.046 (0.006)	0.246 (0.004)	0.236 (<0.001)	93.2 (0.6)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.014 (0.005)	0.233 (0.004)	0.229 (<0.001)	94.6 (0.5)
	$d_{AC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.038 (0.005)	0.210 (0.003)	0.211 (<0.001)	95.0 (0.5)
		$\kappa = 0$	non-linear $q(\cdot)$	0.016 (0.006)	0.266 (0.004)	0.265 (<0.001)	94.7 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.038 (0.005)	0.210 (0.003)	0.211 (<0.001)	94.7 (0.5)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.002 (0.005)	0.245 (0.004)	0.244 (<0.001)	95.0 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	-0.038 (0.005)	0.210 (0.003)	0.211 (<0.001)	95.0 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.007 (0.005)	0.237 (0.004)	0.236 (<0.001)	94.9 (0.5)
	$d_{BC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.033 (0.032)	1.409 (0.022)	1.346 (0.001)	94.1 (0.5)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.299 (0.030)	1.338 (0.021)	1.297 (0.001)	93.0 (0.6)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.002 (0.012)	0.532 (0.008)	0.507 (<0.001)	93.3 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.011 (0.012)	0.520 (0.008)	0.506 (<0.001)	94.5 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	0.008 (0.007)	0.326 (0.005)	0.316 (<0.001)	94.3 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.007 (0.007)	0.331 (0.005)	0.326 (<0.001)	94.4 (0.5)
STC	$d_{AB(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.004 (0.030)	1.356 (0.021)	1.313 (0.001)	94.3 (0.5)
		$\kappa = 0$	non-linear $q(\cdot)$	0.303 (0.029)	1.283 (0.020)	1.262 (0.001)	93.8 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.015 (0.011)	0.474 (0.007)	0.457 (<0.001)	94.1 (0.5)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.012 (0.010)	0.451 (0.007)	0.444 (<0.001)	94.6 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	-0.019 (0.005)	0.241 (0.004)	0.234 (<0.001)	94.1 (0.5)

Table B.9 (continued)

Method	Contrast	Scenario		Bias	Empirical SE	Model SE	Coverage
		$\kappa = 1$	non-linear $q(\cdot)$	-0.006 (0.005)	0.229 (0.004)	0.227 (<0.001)	95.0 (0.5)
	$d_{BC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.010 (0.031)	1.375 (0.022)	1.330 (0.001)	94.3 (0.5)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.300 (0.029)	1.315 (0.021)	1.289 (0.001)	93.5 (0.6)
		$\kappa = 0.5$	linear $q(\cdot)$	0.002 (0.012)	0.521 (0.008)	0.502 (<0.001)	93.8 (0.5)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.019 (0.012)	0.515 (0.008)	0.505 (<0.001)	94.8 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	0.005 (0.007)	0.321 (0.005)	0.313 (<0.001)	94.8 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.002 (0.007)	0.329 (0.005)	0.325 (<0.001)	94.5 (0.5)
MAIC	$d_{AB(AC)}$	$\kappa = 0.5$	linear $q(\cdot)$	-0.184 (0.023)	1.046 (0.017)	- (-)	88.2 (0.7)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.028 (0.022)	0.998 (0.016)	- (-)	92.6 (0.6)
		$\kappa = 1$	linear $q(\cdot)$	-0.005 (0.006)	0.254 (0.004)	0.253 (<0.001)	94.7 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.006 (0.005)	0.240 (0.004)	0.243 (<0.001)	95.5 (0.5)
	$d_{BC(AC)}$	$\kappa = 0.5$	linear $q(\cdot)$	0.170 (0.024)	1.064 (0.017)	- (-)	94.8 (0.5)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.021 (0.023)	1.028 (0.016)	- (-)	93.5 (0.6)
		$\kappa = 1$	linear $q(\cdot)$	-0.009 (0.007)	0.330 (0.005)	0.328 (<0.001)	95.4 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.002 (0.008)	0.337 (0.005)	0.337 (<0.001)	94.8 (0.5)
Bucher	$d_{AB(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.984 (0.005)	0.204 (0.003)	0.202 (<0.001)	0.1 (<0.1)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.670 (0.004)	0.191 (0.003)	0.193 (<0.001)	6.3 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.309 (0.005)	0.204 (0.003)	0.202 (<0.001)	67.9 (1.0)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.295 (0.004)	0.191 (0.003)	0.193 (<0.001)	66.8 (1.1)
		$\kappa = 1$	linear $q(\cdot)$	-0.084 (0.005)	0.204 (0.003)	0.202 (<0.001)	93.0 (0.6)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.091 (0.004)	0.191 (0.003)	0.193 (<0.001)	92.8 (0.6)

Table B.9 (continued)

Method	Contrast	Scenario		Bias	Empirical SE	Model SE	Coverage
	$d_{BC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	0.971 (0.007)	0.292 (0.005)	0.290 (<0.001)	8.7 (0.6)
		$\kappa = 0$	non-linear $q(\cdot)$	0.673 (0.007)	0.326 (0.005)	0.325 (<0.001)	45.6 (1.1)
		$\kappa = 0.5$	linear $q(\cdot)$	0.296 (0.007)	0.292 (0.005)	0.290 (<0.001)	81.9 (0.9)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.288 (0.007)	0.310 (0.005)	0.309 (<0.001)	85.0 (0.8)
		$\kappa = 1$	linear $q(\cdot)$	0.071 (0.007)	0.292 (0.005)	0.290 (<0.001)	94.5 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.083 (0.007)	0.303 (0.005)	0.302 (<0.001)	94.2 (0.5)

Table B.10 Simulation results for scenarios e and f, only adjusting for one of two effect modifiers. The between-study overlap and covariate-outcome relationship are varied jointly. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario		Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$\kappa = 0$	linear $q(\cdot)$	-0.027 (0.005)	0.209 (0.003)	0.205 (<0.001)	93.6 (0.5)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.012 (0.004)	0.195 (0.003)	0.195 (<0.001)	95.1 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.027 (0.005)	0.209 (0.003)	0.205 (<0.001)	93.8 (0.5)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.012 (0.004)	0.195 (0.003)	0.195 (<0.001)	94.7 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	-0.028 (0.005)	0.209 (0.003)	0.205 (<0.001)	93.8 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.013 (0.004)	0.195 (0.003)	0.195 (<0.001)	94.7 (0.5)
	$d_{AC(AB)}$	$\kappa = 0$	linear $q(\cdot)$	0.414 (0.022)	0.973 (0.015)	0.936 (<0.001)	91.7 (0.6)
		$\kappa = 0$	non-linear $q(\cdot)$	0.157 (0.021)	0.919 (0.015)	0.904 (<0.001)	94.0 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	0.111 (0.008)	0.368 (0.006)	0.358 (<0.001)	92.9 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.121 (0.008)	0.369 (0.006)	0.365 (<0.001)	93.2 (0.6)
		$\kappa = 1$	linear $q(\cdot)$	0.010 (0.005)	0.227 (0.004)	0.226 (<0.001)	94.8 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.032 (0.006)	0.248 (0.004)	0.247 (<0.001)	94.5 (0.5)
	$d_{BC(AB)}$	$\kappa = 0$	linear $q(\cdot)$	0.441 (0.022)	0.997 (0.016)	0.953 (<0.001)	91.8 (0.6)
		$\kappa = 0$	non-linear $q(\cdot)$	0.169 (0.021)	0.943 (0.015)	0.927 (<0.001)	94.6 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	0.139 (0.010)	0.425 (0.007)	0.408 (<0.001)	92.8 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.134 (0.009)	0.421 (0.007)	0.415 (<0.001)	93.4 (0.6)
		$\kappa = 1$	linear $q(\cdot)$	0.038 (0.007)	0.310 (0.005)	0.303 (<0.001)	94.3 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.045 (0.007)	0.318 (0.005)	0.315 (<0.001)	94.2 (0.5)
	$d_{AB(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.472 (0.022)	0.972 (0.015)	0.929 (<0.001)	90.8 (0.6)

Table B.10 (continued)

Method	Contrast	Scenario		Bias	Empirical SE	Model SE	Coverage
		$\kappa = 0$	non-linear $q(\cdot)$	-0.155 (0.020)	0.909 (0.014)	0.893 (<0.001)	94.5 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.169 (0.008)	0.367 (0.006)	0.351 (<0.001)	91.3 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.137 (0.008)	0.345 (0.005)	0.340 (<0.001)	93.0 (0.6)
		$\kappa = 1$	linear $q(\cdot)$	-0.068 (0.005)	0.227 (0.004)	0.219 (<0.001)	93.2 (0.6)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.052 (0.005)	0.212 (0.003)	0.212 (<0.001)	94.4 (0.5)
	$d_{AC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.030 (0.005)	0.209 (0.003)	0.210 (<0.001)	95.1 (0.5)
		$\kappa = 0$	non-linear $q(\cdot)$	0.014 (0.006)	0.265 (0.004)	0.264 (<0.001)	94.8 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.030 (0.005)	0.209 (0.003)	0.210 (<0.001)	95.0 (0.5)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.003 (0.005)	0.244 (0.004)	0.243 (<0.001)	95.0 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	-0.030 (0.005)	0.209 (0.003)	0.210 (<0.001)	95.0 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.007 (0.005)	0.237 (0.004)	0.235 (<0.001)	95.2 (0.5)
	$d_{BC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	0.441 (0.022)	0.997 (0.016)	0.953 (<0.001)	91.8 (0.6)
		$\kappa = 0$	non-linear $q(\cdot)$	0.169 (0.021)	0.943 (0.015)	0.927 (<0.001)	94.6 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	0.139 (0.010)	0.425 (0.007)	0.408 (<0.001)	92.8 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.134 (0.009)	0.421 (0.007)	0.415 (<0.001)	93.4 (0.6)
		$\kappa = 1$	linear $q(\cdot)$	0.038 (0.007)	0.310 (0.005)	0.303 (<0.001)	94.3 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.045 (0.007)	0.318 (0.005)	0.315 (<0.001)	94.2 (0.5)
STC	$d_{AB(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.464 (0.021)	0.952 (0.015)	0.920 (<0.001)	91.7 (0.6)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.158 (0.020)	0.891 (0.014)	0.885 (<0.001)	94.8 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.153 (0.008)	0.361 (0.006)	0.348 (<0.001)	92.4 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.134 (0.008)	0.339 (0.005)	0.338 (<0.001)	93.7 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	-0.050 (0.005)	0.223 (0.004)	0.218 (<0.001)	93.7 (0.5)

Table B.10 (continued)

Method	Contrast	Scenario		Bias	Empirical SE	Model SE	Coverage
		$\kappa = 1$	non-linear $q(\cdot)$	-0.046 (0.005)	0.210 (0.003)	0.210 (<0.001)	94.8 (0.5)
	$d_{BC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	0.451 (0.022)	0.976 (0.015)	0.943 (<0.001)	92.2 (0.6)
		$\kappa = 0$	non-linear $q(\cdot)$	0.161 (0.021)	0.930 (0.015)	0.923 (<0.001)	94.9 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	0.140 (0.009)	0.418 (0.007)	0.405 (<0.001)	93.3 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.127 (0.009)	0.418 (0.007)	0.415 (<0.001)	93.8 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	0.036 (0.007)	0.307 (0.005)	0.301 (<0.001)	94.3 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.038 (0.007)	0.316 (0.005)	0.314 (<0.001)	94.4 (0.5)
MAIC	$d_{AB(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-2.902 (0.087)	3.913 (0.062)	- (-)	44.4 (1.1)
		$\kappa = 0$	non-linear $q(\cdot)$	-2.239 (0.084)	3.770 (0.060)	- (-)	22.2 (0.9)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.175 (0.010)	0.447 (0.007)	0.447 (0.001)	91.0 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.135 (0.009)	0.423 (0.007)	0.432 (0.001)	91.7 (0.6)
		$\kappa = 1$	linear $q(\cdot)$	-0.041 (0.005)	0.229 (0.004)	0.228 (<0.001)	93.6 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.046 (0.005)	0.213 (0.003)	0.218 (<0.001)	94.7 (0.5)
	$d_{BC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	2.709 (0.087)	3.872 (0.061)	- (-)	- (-)
		$\kappa = 0$	non-linear $q(\cdot)$	2.044 (0.081)	3.643 (0.058)	- (-)	- (-)
		$\kappa = 0.5$	linear $q(\cdot)$	0.162 (0.011)	0.490 (0.008)	0.493 (0.001)	93.5 (0.6)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.128 (0.011)	0.488 (0.008)	0.495 (0.001)	93.9 (0.5)
		$\kappa = 1$	linear $q(\cdot)$	0.027 (0.007)	0.313 (0.005)	0.309 (<0.001)	95.5 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.038 (0.007)	0.319 (0.005)	0.319 (<0.001)	94.5 (0.5)
Bucher	$d_{AB(AC)}$	$\kappa = 0$	linear $q(\cdot)$	-0.984 (0.005)	0.204 (0.003)	0.202 (<0.001)	0.1 (<0.1)
		$\kappa = 0$	non-linear $q(\cdot)$	-0.670 (0.004)	0.191 (0.003)	0.193 (<0.001)	6.3 (0.5)
		$\kappa = 0.5$	linear $q(\cdot)$	-0.309 (0.005)	0.204 (0.003)	0.202 (<0.001)	67.9 (1.0)

Table B.10 (continued)

Method	Contrast	Scenario		Bias	Empirical SE	Model SE	Coverage
		$\kappa = 0.5$	non-linear $q(\cdot)$	-0.295 (0.004)	0.191 (0.003)	0.193 (<0.001)	66.8 (1.1)
		$\kappa = 1$	linear $q(\cdot)$	-0.084 (0.005)	0.204 (0.003)	0.202 (<0.001)	93.0 (0.6)
		$\kappa = 1$	non-linear $q(\cdot)$	-0.091 (0.004)	0.191 (0.003)	0.193 (<0.001)	92.8 (0.6)
	$d_{BC(AC)}$	$\kappa = 0$	linear $q(\cdot)$	0.971 (0.007)	0.292 (0.005)	0.290 (<0.001)	8.7 (0.6)
		$\kappa = 0$	non-linear $q(\cdot)$	0.673 (0.007)	0.326 (0.005)	0.325 (<0.001)	45.6 (1.1)
		$\kappa = 0.5$	linear $q(\cdot)$	0.296 (0.007)	0.292 (0.005)	0.290 (<0.001)	81.9 (0.9)
		$\kappa = 0.5$	non-linear $q(\cdot)$	0.288 (0.007)	0.310 (0.005)	0.309 (<0.001)	85.0 (0.8)
		$\kappa = 1$	linear $q(\cdot)$	0.071 (0.007)	0.292 (0.005)	0.290 (<0.001)	94.5 (0.5)
		$\kappa = 1$	non-linear $q(\cdot)$	0.083 (0.007)	0.303 (0.005)	0.302 (<0.001)	94.2 (0.5)

Scenarios g, h, and i B.6

Table B.11 Simulation results for scenarios g, h, and i, adjusting for all effect modifiers. The covariate distributions and correlation structures in each study are varied jointly. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.058 (0.005)	0.207 (0.003)	0.207 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.058 (0.005)	0.207 (0.003)	0.207 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.058 (0.005)	0.207 (0.003)	0.207 (<0.001)	94.4 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.058 (0.005)	0.207 (0.003)	0.207 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.058 (0.005)	0.207 (0.003)	0.207 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.058 (0.005)	0.207 (0.003)	0.207 (<0.001)	94.1 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.059 (0.005)	0.208 (0.003)	0.207 (<0.001)	94.5 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.059 (0.005)	0.208 (0.003)	0.207 (<0.001)	94.3 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.059 (0.005)	0.208 (0.003)	0.207 (<0.001)	94.5 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.059 (0.005)	0.208 (0.003)	0.207 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.059 (0.005)	0.208 (0.003)	0.207 (<0.001)	94.3 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.059 (0.005)	0.208 (0.003)	0.207 (<0.001)	94.4 (0.5)
	$d_{AC(AB)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.028 (0.011)	0.476 (0.008)	0.470 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.028 (0.011)	0.477 (0.008)	0.470 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.028 (0.011)	0.476 (0.008)	0.470 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.028 (0.011)	0.477 (0.008)	0.470 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.028 (0.011)	0.477 (0.008)	0.470 (<0.001)	94.9 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.028 (0.011)	0.477 (0.008)	0.470 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.051 (0.011)	0.478 (0.008)	0.471 (<0.001)	94.9 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.051 (0.011)	0.478 (0.008)	0.472 (<0.001)	94.8 (0.5)

Table B.11 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.051 (0.011)	0.478 (0.008)	0.471 (<0.001)	94.6 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.051 (0.011)	0.478 (0.008)	0.472 (<0.001)	94.9 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.051 (0.011)	0.478 (0.008)	0.472 (<0.001)	94.7 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.051 (0.011)	0.478 (0.008)	0.471 (<0.001)	94.8 (0.5)
	$d_{BC(AB)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.5 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.6 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.7 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.7 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.029 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.008 (0.012)	0.514 (0.008)	0.508 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.008 (0.012)	0.515 (0.008)	0.508 (<0.001)	94.7 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.008 (0.012)	0.514 (0.008)	0.508 (<0.001)	94.5 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.008 (0.012)	0.514 (0.008)	0.509 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.007 (0.012)	0.515 (0.008)	0.509 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.008 (0.012)	0.514 (0.008)	0.508 (<0.001)	94.5 (0.5)
	$d_{AB(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.069 (0.011)	0.470 (0.007)	0.464 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.069 (0.011)	0.471 (0.007)	0.464 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.069 (0.011)	0.470 (0.007)	0.464 (<0.001)	94.4 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.069 (0.011)	0.471 (0.007)	0.465 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.069 (0.011)	0.470 (0.007)	0.464 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.069 (0.011)	0.470 (0.007)	0.464 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.047 (0.010)	0.468 (0.007)	0.463 (<0.001)	94.4 (0.5)

Table B.11 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.047 (0.010)	0.468 (0.007)	0.463 (<0.001)	94.4 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.047 (0.010)	0.468 (0.007)	0.463 (<0.001)	94.5 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.047 (0.010)	0.468 (0.007)	0.463 (<0.001)	94.3 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.047 (0.010)	0.468 (0.007)	0.463 (<0.001)	94.4 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.047 (0.010)	0.468 (0.007)	0.463 (<0.001)	94.3 (0.5)
	$d_{AC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.039 (0.005)	0.219 (0.003)	0.211 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.040 (0.005)	0.220 (0.003)	0.211 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.039 (0.005)	0.220 (0.003)	0.211 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.039 (0.005)	0.220 (0.003)	0.211 (<0.001)	94.3 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.040 (0.005)	0.220 (0.003)	0.211 (<0.001)	93.9 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.039 (0.005)	0.220 (0.003)	0.211 (<0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.039 (0.005)	0.219 (0.003)	0.211 (<0.001)	94.1 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.039 (0.005)	0.220 (0.003)	0.211 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.039 (0.005)	0.220 (0.003)	0.211 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.039 (0.005)	0.220 (0.003)	0.211 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.040 (0.005)	0.220 (0.003)	0.211 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.039 (0.005)	0.220 (0.003)	0.211 (<0.001)	93.8 (0.5)
	$d_{BC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.5 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.6 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.7 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.7 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.029 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.030 (0.012)	0.516 (0.008)	0.509 (<0.001)	94.8 (0.5)

Table B.11 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.008 (0.012)	0.514 (0.008)	0.508 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.008 (0.012)	0.515 (0.008)	0.508 (<0.001)	94.7 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.008 (0.012)	0.514 (0.008)	0.508 (<0.001)	94.5 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.008 (0.012)	0.514 (0.008)	0.509 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.007 (0.012)	0.515 (0.008)	0.509 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.008 (0.012)	0.514 (0.008)	0.508 (<0.001)	94.5 (0.5)
STC	$d_{AB(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.038 (0.010)	0.458 (0.007)	0.458 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.037 (0.010)	0.458 (0.007)	0.458 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.037 (0.010)	0.457 (0.007)	0.458 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.038 (0.010)	0.458 (0.007)	0.458 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.037 (0.010)	0.458 (0.007)	0.458 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.037 (0.010)	0.457 (0.007)	0.458 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.026 (0.010)	0.457 (0.007)	0.458 (<0.001)	95.1 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.026 (0.010)	0.457 (0.007)	0.458 (<0.001)	95.1 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.026 (0.010)	0.457 (0.007)	0.458 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.026 (0.010)	0.457 (0.007)	0.458 (<0.001)	95.1 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.026 (0.010)	0.457 (0.007)	0.458 (<0.001)	95.1 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.026 (0.010)	0.457 (0.007)	0.458 (<0.001)	95.2 (0.5)
	$d_{BC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.023 (0.011)	0.504 (0.008)	0.503 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.022 (0.011)	0.504 (0.008)	0.503 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.022 (0.011)	0.504 (0.008)	0.503 (<0.001)	95.3 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.023 (0.011)	0.504 (0.008)	0.503 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.022 (0.011)	0.504 (0.008)	0.503 (<0.001)	95.2 (0.5)

Table B.11 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.023 (0.011)	0.504 (0.008)	0.503 (<0.001)	95.3 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.011 (0.011)	0.503 (0.008)	0.503 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.011 (0.011)	0.503 (0.008)	0.503 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.011 (0.011)	0.503 (0.008)	0.503 (<0.001)	94.9 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.012 (0.011)	0.503 (0.008)	0.503 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.011 (0.011)	0.503 (0.008)	0.503 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.011 (0.011)	0.503 (0.008)	0.503 (<0.001)	94.9 (0.5)
MAIC	$d_{AB(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.138 (0.019)	0.850 (0.013)	- (-)	90.1 (0.7)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.138 (0.019)	0.849 (0.013)	- (-)	90.2 (0.7)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.139 (0.019)	0.849 (0.013)	- (-)	90.2 (0.7)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.138 (0.019)	0.850 (0.013)	10.701 (4.733)	90.2 (0.7)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.139 (0.019)	0.850 (0.013)	49.602 (22.145)	90.2 (0.7)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.139 (0.019)	0.849 (0.013)	- (-)	90.2 (0.7)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.219 (0.024)	1.091 (0.017)	- (-)	89.4 (0.7)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.221 (0.025)	1.112 (0.018)	- (-)	89.5 (0.7)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.220 (0.024)	1.089 (0.017)	- (-)	89.5 (0.7)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.219 (0.024)	1.083 (0.017)	- (-)	89.4 (0.7)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.219 (0.024)	1.087 (0.017)	- (-)	89.5 (0.7)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.220 (0.025)	1.097 (0.017)	- (-)	89.5 (0.7)
	$d_{BC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.123 (0.020)	0.875 (0.014)	- (-)	95.4 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.123 (0.020)	0.875 (0.014)	- (-)	95.1 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.124 (0.020)	0.875 (0.014)	- (-)	95.1 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.123 (0.020)	0.875 (0.014)	10.703 (4.732)	95.1 (0.5)

Table B.11 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.123 (0.020)	0.875 (0.014)	49.603 (22.145)	95.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.124 (0.020)	0.875 (0.014)	- (-)	95.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.205 (0.025)	1.106 (0.017)	- (-)	96.0 (0.4)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.206 (0.025)	1.126 (0.018)	- (-)	95.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.204 (0.025)	1.104 (0.017)	- (-)	95.6 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.204 (0.025)	1.098 (0.017)	- (-)	95.8 (0.4)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.204 (0.025)	1.103 (0.017)	- (-)	95.9 (0.4)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.206 (0.025)	1.112 (0.018)	- (-)	95.4 (0.5)
Bucher	$d_{AB(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
	$d_{BC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.7 (0.9)

Table B.11 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.304 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.7 (0.9)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.7 (0.9)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.6 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.304 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)

Table B.12 Simulation results for scenarios g, h, and i, only adjusting for one of two effect modifiers. The covariate distributions and correlation structures in each study are varied jointly. Monte Carlo standard errors for each statistic are shown in brackets.

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
ML-NMR	$d_{AB(AB)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.036 (0.005)	0.204 (0.003)	0.205 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.036 (0.005)	0.204 (0.003)	0.205 (<0.001)	95.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.036 (0.005)	0.204 (0.003)	0.205 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.036 (0.005)	0.204 (0.003)	0.205 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.036 (0.005)	0.204 (0.003)	0.205 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.036 (0.005)	0.204 (0.003)	0.205 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.036 (0.005)	0.205 (0.003)	0.205 (<0.001)	94.9 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.036 (0.005)	0.205 (0.003)	0.205 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.036 (0.005)	0.205 (0.003)	0.205 (<0.001)	94.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.036 (0.005)	0.205 (0.003)	0.205 (<0.001)	95.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.036 (0.005)	0.205 (0.003)	0.205 (<0.001)	94.9 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.036 (0.005)	0.205 (0.003)	0.205 (<0.001)	94.8 (0.5)
	$d_{AC(AB)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.113 (0.008)	0.366 (0.006)	0.358 (<0.001)	93.5 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.112 (0.008)	0.366 (0.006)	0.358 (<0.001)	93.5 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.112 (0.008)	0.366 (0.006)	0.358 (<0.001)	93.5 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.113 (0.008)	0.366 (0.006)	0.358 (<0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.112 (0.008)	0.366 (0.006)	0.358 (<0.001)	93.5 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.112 (0.008)	0.366 (0.006)	0.358 (<0.001)	93.7 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.099 (0.008)	0.367 (0.006)	0.359 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.099 (0.008)	0.367 (0.006)	0.359 (<0.001)	94.3 (0.5)

Table B.12 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.099 (0.008)	0.367 (0.006)	0.359 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.099 (0.008)	0.367 (0.006)	0.359 (<0.001)	94.1 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.099 (0.008)	0.367 (0.006)	0.359 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.099 (0.008)	0.367 (0.006)	0.359 (<0.001)	94.2 (0.5)
	$d_{BC(AB)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.148 (0.009)	0.416 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.5 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.4 (0.6)
	$d_{AB(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.180 (0.008)	0.352 (0.006)	0.352 (<0.001)	92.5 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.180 (0.008)	0.352 (0.006)	0.352 (<0.001)	92.5 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.179 (0.008)	0.352 (0.006)	0.352 (<0.001)	92.7 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.180 (0.008)	0.351 (0.006)	0.352 (<0.001)	92.3 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.180 (0.008)	0.352 (0.006)	0.352 (<0.001)	92.3 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.180 (0.008)	0.352 (0.006)	0.352 (<0.001)	92.3 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.167 (0.008)	0.350 (0.006)	0.351 (<0.001)	92.8 (0.6)

Table B.12 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.167 (0.008)	0.350 (0.006)	0.351 (<0.001)	92.8 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.167 (0.008)	0.350 (0.006)	0.351 (<0.001)	92.8 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.167 (0.008)	0.350 (0.006)	0.351 (<0.001)	92.8 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.167 (0.008)	0.350 (0.006)	0.351 (<0.001)	92.7 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.167 (0.008)	0.350 (0.006)	0.351 (<0.001)	92.7 (0.6)
	$d_{AC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.031 (0.005)	0.218 (0.003)	0.210 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.032 (0.005)	0.219 (0.003)	0.210 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.032 (0.005)	0.218 (0.003)	0.210 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.031 (0.005)	0.218 (0.003)	0.210 (<0.001)	94.3 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.032 (0.005)	0.218 (0.003)	0.210 (<0.001)	93.9 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.032 (0.005)	0.218 (0.003)	0.210 (<0.001)	93.9 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.031 (0.005)	0.218 (0.003)	0.210 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.032 (0.005)	0.219 (0.003)	0.210 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.032 (0.005)	0.218 (0.003)	0.210 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.031 (0.005)	0.218 (0.003)	0.210 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.032 (0.005)	0.218 (0.003)	0.210 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.031 (0.005)	0.218 (0.003)	0.210 (<0.001)	93.8 (0.5)
	$d_{BC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.148 (0.009)	0.416 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.5 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.148 (0.009)	0.415 (0.007)	0.409 (<0.001)	93.2 (0.6)

Table B.12 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.3 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.135 (0.009)	0.414 (0.007)	0.409 (<0.001)	93.4 (0.6)
STC	$d_{AB(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.159 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.159 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.159 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.159 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.159 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.159 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.151 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.151 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.151 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.151 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.151 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.151 (0.008)	0.344 (0.005)	0.348 (<0.001)	94.2 (0.5)
	$d_{BC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.144 (0.009)	0.408 (0.006)	0.406 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.143 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.143 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.144 (0.009)	0.408 (0.006)	0.406 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.143 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.9 (0.5)

Table B.12 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.144 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.136 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.136 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.9 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.136 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.136 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.136 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.7 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.136 (0.009)	0.408 (0.006)	0.406 (<0.001)	93.8 (0.5)
MAIC	$d_{AB(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.165 (0.009)	0.399 (0.006)	0.411 (<0.001)	91.7 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.165 (0.009)	0.399 (0.006)	0.411 (<0.001)	91.7 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.165 (0.009)	0.399 (0.006)	0.411 (<0.001)	91.7 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.165 (0.009)	0.399 (0.006)	0.411 (<0.001)	91.7 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.165 (0.009)	0.399 (0.006)	0.411 (<0.001)	91.7 (0.6)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.165 (0.009)	0.399 (0.006)	0.411 (<0.001)	91.7 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.176 (0.010)	0.442 (0.007)	0.448 (0.001)	91.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.176 (0.010)	0.442 (0.007)	0.448 (0.001)	91.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.176 (0.010)	0.442 (0.007)	0.448 (0.001)	91.2 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	-0.176 (0.010)	0.441 (0.007)	0.447 (0.001)	91.3 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.176 (0.010)	0.441 (0.007)	0.447 (0.001)	91.3 (0.6)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	-0.176 (0.010)	0.441 (0.007)	0.447 (0.001)	91.3 (0.6)
	$d_{BC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.150 (0.010)	0.453 (0.007)	0.461 (<0.001)	94.1 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.150 (0.010)	0.453 (0.007)	0.461 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.150 (0.010)	0.453 (0.007)	0.461 (<0.001)	94.0 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.150 (0.010)	0.453 (0.007)	0.461 (<0.001)	94.2 (0.5)

Table B.12 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.150 (0.010)	0.453 (0.007)	0.461 (<0.001)	94.2 (0.5)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.150 (0.010)	0.453 (0.007)	0.461 (<0.001)	94.1 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.161 (0.011)	0.489 (0.008)	0.494 (0.001)	93.7 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.161 (0.011)	0.489 (0.008)	0.494 (0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.161 (0.011)	0.490 (0.008)	0.494 (0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.161 (0.011)	0.489 (0.008)	0.494 (0.001)	93.5 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.161 (0.011)	0.489 (0.008)	0.494 (0.001)	93.8 (0.5)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.161 (0.011)	0.489 (0.008)	0.494 (0.001)	93.5 (0.5)
Bucher	$d_{AB(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.5 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	-0.318 (0.004)	0.199 (0.003)	0.202 (<0.001)	65.3 (1.1)
	$d_{BC(AC)}$	$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25, \rho_{(AC)} = 0.5$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.7 (0.9)

Table B.12 (continued)

Method	Contrast	Scenario			Bias	Empirical SE	Model SE	Coverage
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.304 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.7 (0.9)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.7 (0.9)
		$X_{(AB)} \sim$ Gam	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.6 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ Gam	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.303 (0.007)	0.296 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0$	0.304 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = \rho_{(AC)} = 0.25$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)
		$X_{(AB)} \sim$ N	$X_{(AC)} \sim$ N	$\rho_{(AB)} = 0.25$, $\rho_{(AC)} = 0.5$	0.303 (0.007)	0.295 (0.005)	0.291 (<0.001)	81.8 (0.9)

Appendix C

Computing environment

Hardware C.1

Laptop

Intel i7-6600U CPU, dual core 2.6 GHz. 16 GB RAM.

Desktop

Intel i7-8700 CPU, hex core 3.2 GHz. 16 GB RAM.

BlueCrystal Phase 3 supercomputer

Each node has 16 Sandy Bridge cores at 2.6 GHz, with 64 GB RAM.

Software C.2

R versions 3.4.1 to 3.5.3 (R Core Team 2018) and Stan versions 2.16.0 to 2.18.1 (Carpenter et al. 2017) were used for analysis.