NICE DSU TECHNICAL SUPPORT DOCUMENT 20:

MULTIVARIATE META-ANALYSIS OF SUMMARY DATA FOR COMBINING TREATMENT EFFECTS ON CORRELATED OUTCOMES AND EVALUATING SURROGATE ENDPOINTS

Version 2

REPORT BY THE DECISION SUPPORT UNIT

22 October 2019

Sylwia Bujkiewicz[1], Felix Achana[2], Tasos Papanikos[1], Richard D Riley[3], and Keith R Abrams[1]

[1] Biostatistics Research Group, Department of Health Sciences, University of Leicester

[2] Warwick Medical School, Population Evidence and Technologies, University of Warwick

[3] Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street Sheffield, S1 4DA

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
Twitter @NICE_DSU

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Centre is based at the University of Sheffield with members at York, Bristol, Leicester and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information www.nicedsu.org.uk.

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

ABOUT THE TECHNICAL SUPPORT DOCUMENT SERIES

The NICE Guide to the Methods of Technology Appraisal[i] is a regularly updated document that provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The Methods Guide does not provide detailed advice on how to implement and apply the methods it describes. This DSU series of Technical Support Documents (TSDs) is intended to complement the Methods Guide by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in each topic area, and make clear recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE Technology Appraisals, whether companies, assessment groups or any other stakeholder type.

We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Dr Allan Wailoo, Director of DSU and TSD series editor.

---

i National Institute for Health and Care Excellence. Guide to the methods of technology appraisal, 2013 (updated April 2013), London

Acknowledgements

We thank the independent reviewers, Ian White, Malcolm Price, Beth Woods, Sandro Gsteiger and Jamie Elvidge, whose comments helped to improve the quality of this document. The editor for the TSD series is Allan Wailoo.

Research that led to developing this TSD was funded by the Medical Research Council (MRC) Methodology Research Programme (New Investigator Research Grant MR/L009854/1 awarded to Sylwia Bujkiewicz). Keith Abrams was partially supported as a UK National Institute for Health Research (NIHR) Senior Investigator Emeritus (NI-SI-0512-10159).

We thank Prof G.J. Melendez-Torres (Peninsula Technology Assessment Group, College of Medicine and Health, University of Exeter) for sharing data from systematic review in relapsing remitting multiple sclerosis which we used as an illustrative example in multivariate network meta-analysis. We also thank Marta Soares for useful discussion on use of bivariate meta-analysis in the NICE technology appraisal of tumour necrosis factor–α inhibitors for ankylosing spondylitis and non–radiographic axial spondyloarthritis. The authors are also grateful to Dr Georgios Nikolaidis and Anastasios Tasoulas at IQVIA for bringing to our attention an error in one of the trivariate meta-analysis models, which helped us to improve the quality of the document.

This report should be referenced as follows:

Bujkiewicz, S., Achana, F., Papanikos, T., Riley, R.D., Abrams, K.R. NICE DSU Technical Support Document 20: Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. 2019; available from http://www.nicedsu.org.uk

Executive summary

Health technology assessment (HTA) agencies such as the National Institute for Health and Care Excellence (NICE) require evidence synthesis of existing studies to inform their decisions, for example about the best available treatments with respect to multiple efficacy and safety outcomes. However, relevant studies may not provide direct evidence about all the treatments or outcomes of interest. Studies that do not provide direct evidence about a particular outcome or treatment of interest are often excluded from a meta-analysis evaluating that outcome or treatment. This is unwelcome, especially if their participants are otherwise representative of the population, clinical settings and condition of interest. Research studies require considerable costs and time, and involve precious patient participation, and simply discarding them could be viewed as research waste. Statistical models for multivariate and network meta-analysis address this by simultaneously analysing multiple outcomes and multiple treatments, respectively. This allows more studies to contribute toward each outcome and treatment comparison.

This Technical Support Document (TSD) describes the key methods of multivariate meta-analysis and their extensions to network meta-analysis of multiple outcomes. We focus on the use of multivariate meta-analytic methods for combining data from multiple correlated outcomes with the aim of including all relevant evidence and borrowing of information across outcomes, in particular when not all of the relevant studies report an outcome of interest. We also devote considerable attention to the use of multivariate meta-analysis for the purpose of surrogate endpoint evaluation.

Surrogate endpoints (such as, for example, progression free survival as an early marker of overall survival in cancer) play an increasingly important role in the drug development process as new health technologies are increasingly being licensed by the regulatory agencies, such as European Medicines Agency (EMA) in Europe or Food and Drug Administration (FDA) in the US, based on evidence obtained by measuring effectiveness on a surrogate marker. When data on the final clinical outcome are not available or limited at the licensing stage, and therefore also for the HTA decision-making process, a modelling framework is required to establish the strength of the surrogate relationship between the treatment effects on the surrogate and the final outcome and to predict the likely treatment effect on the final outcome for the new health technology. Multivariate meta-analytic methods provide such a framework as they, by definition, take into account the correlation between the treatment effects on the surrogate and final outcomes as well as the uncertainty related to all parameters describing the surrogate relationship.

We describe and apply the methods in the Bayesian framework of estimation and provide WinBUGS code for a Bayesian analysis using Markov chain Monte Carlo (MCMC) simulation.

Contents

1 Introduction **11**
- 1.1 Bayesian approaches to multivariate meta-analysis . . . 12
- 1.2 Examples of multivariate meta-analysis used in the context of NICE . . . 13
- 1.3 Use of multivariate meta-analysis for surrogate endpoint evaluation . . . 14
  - 1.3.1 Surrogate endpoints in regulatory decision making and HTA . . . 14
- 1.4 Implementation . . . 15
- 1.5 Structure of the TSD and the level of required expertise . . . 15

2 Bivariate random-effects meta-analysis (BRMA) **17**
- 2.1 BRMA in a Bayesian framework with considerations of appropriate prior distributions 17
  - 2.1.1 Within-study covariance . . . 17
  - 2.1.2 Between-studies model parameters . . . 18
  - 2.1.3 Accounting for missing data . . . 19
- 2.2 BRMA in the product normal formulation (PNF) . . . 20
- 2.3 Example: rheumatoid arthritis . . . 21
  - 2.3.1 Data . . . 21
  - 2.3.2 Results . . . 22
- 2.4 Illustration of the use of informative prior distributions for combining diverse sources if evidence . . . 25
  - 2.4.1 Using external IPD to inform within-study correlations . . . 25
  - 2.4.2 Using external summary data to inform between-studies correlation . . . 25
  - 2.4.3 Logic of the meta-analysis model and notation . . . 25
  - 2.4.4 Statistical methods . . . 26
  - 2.4.5 Results . . . 27

3 Surrogate endpoint evaluation with bivariate meta-analysis **28**
- 3.1 Surrogate endpoints, their importance and validity . . . 28
  - 3.1.1 Importance . . . 28
  - 3.1.2 Validity . . . 28
  - 3.1.3 Meta-analytic approach to surrogate endpoint evaluation . . . 29
  - 3.1.4 Data requirements for surrogate endpoint validation . . . 29
- 3.2 Standard surrogacy model by Daniels and Hughes . . . 30
- 3.3 BRMA in product normal formulation . . . 30
- 3.4 BRMA in the standard form . . . 31
- 3.5 Relationships between the models and with other models in the literature . . . 31
- 3.6 Validation and making predictions . . . 32
  - 3.6.1 Cross-validation procedure . . . 32
  - 3.6.2 Predicting treatment effect on the final outcome from the effect measured on a surrogate endpoint in a new study . . . 33
- 3.7 Example: surrogacy validation in relapsing remitting multiple sclerosis . . . 34
  - 3.7.1 Data . . . 34
  - 3.7.2 Implementation . . . 34
  - 3.7.3 Results of surrogate endpoint validation . . . 36
  - 3.7.4 Making prediction for a new study . . . 39
  - 3.7.5 Discussion of the results for RRMS . . . 39
- 3.8 Discussion of surrogacy criteria, other surrogacy models and further work . . . 40
  - 3.8.1 Further work . . . 40

4 Multivariate random effects meta-analysis (MRMA) 42
4.1 MRMA in the standard form . . . 42
4.1.1 Prior distribution on the between-studies covariance matrix . . . 42
4.2 MRMA in the product normal formulation . . . 43
4.2.1 TRMA PNF unstructured model . . . 43
4.2.2 TRMA PNF structured model . . . 44
4.3 Example: rheumatoid arthritis . . . 45
4.4 Benefits of multivariate meta-analysis . . . 47
4.5 Application to multiple surrogate endpoints . . . 49

5 Network meta-analysis of multiple correlated outcomes 50
5.1 Bivariate network meta-analysis (bvNMA) for contrast level data entry . . . 51
5.2 Bivariate network meta-analysis (bvNMA) for arm level data entry . . . 52
5.3 Multivariate network meta-analysis (mvNMA) for arm level data entry . . . 53
5.4 Example: relapsing remitting multiple sclerosis (RRMS) . . . 54
5.4.1 Data . . . 54
5.4.2 Models fitted . . . 55
5.4.3 Results . . . 56
5.5 Discussion of mvNMA methodology and ongoing research . . . 60

6 Discussion and extensions 61
6.1 Normality assumption for random effects . . . 61
6.2 Binary, nested and mutually exclusive outcomes . . . 61
6.3 Bayesian versus frequentist methods and related software . . . 61
6.4 Other application areas . . . 62

7 Summary and recommendations for use of multivariate meta-analysis to inform decision modelling 63
7.1 Recommendations . . . 63
7.2 Conclusions . . . 64

8 Disclosures 64

9 References 64

A Supplementary materials for application of bivariate meta-analysis to the example in rheumatoid arthritis 73
A.1 WinBUGS code for BRMA in the standard form . . . 73
A.1.1 Data requirements . . . 73
A.1.2 Code . . . 73
A.2 WinBUGS code for BRMA in product normal formulation . . . 76
A.3 WinBUGS code for the analysis with informative prior distributions for the correlations (as discussed in Section 2.4, with methods described in Section 2.4.4) . . . 78

B Supplementary materials for evaluation of surrogate endpoints in relapsing remitting multiple sclerosis (RRMS) 81
B.1 WinBUGS code for model by Daniels and Hughes . . . 81
B.2 WinBUGS code for BRMA PNF model for surrogate endpoints . . . 83
B.3 WinBUGS code for BRMA standard model for surrogate endpoints . . . 84
B.4 R code (with models in bugs) for cross-validation (Sec. 3.6.1 and 3.7) . . . 85
B.5 Example of the analysis (using method by Daniels and Hughes) for predicting the treatment effect on the final outcome in a new study reporting only the treatment effect on the surrogate endpoint . . . 93

C Supplementary materials for multivariate meta-analysis **96**
C.1 Constructing prior distribution for the between-studies covariance matrix 96
C.2 WinBUGS code for multivariate meta-analysis with Cholesky decomposition of the between-studies covariance matrix with application to the example in rheumatoid arthritis 98
C.3 WinBUGS code for multivariate meta-analysis with spherical decomposition of the between-studies covariance matrix with application to the example in rheumatoid arthritis 102
C.4 WinBUGS code for trivariate meta-analysis with in PNF (unstructured model) with application to the example in rheumatoid arthritis 105
C.5 WinBUGS code for trivariate meta-analysis with in PNF (structured model) with application to the example in rheumatoid arthritis 108
C.6 Additional results for the example in RA 111
C.7 Extension of multivariate-meta-analytic models in product normal formulation (PNF) to multiple outcomes (beyond the trivariate case) 113
C.7.1 MRMA PNF: unstructured model 113
C.7.2 MRMA PNF: structured model 114

D Supplementary materials for network meta-analysis of multiple outcomes **115**
D.1 Additional technical details for mvNMA of multi-arm trials 115
D.2 WinBUGS code for multivariate NMA applied to the example in multiple sclerosis . 116

List of Tables

1 Studies in RA data reporting outcomes: ACR20, DAS-28 and HAQ. 21
2 Results of the univariate meta-analyses of treatment effects on HAQ and DAS-28 separately and bivariate meta-analysis of the treatment effects on HAQ and DAS-28. Negative values (reduction from baseline for HAQ or DAS-28) represent positive effect of the treatment. 22
3 Results of the univariate meta-analyses of treatment effects on HAQ and DAS-28 separately and bivariate meta-analysis of the treatment effects on HAQ and DAS-28. Negative values (reduction from baseline for HAQ or DAS-28) represent positive effect of the treatment. 27
4 Parameters in the surrogacy models and the relationships between them 32
5 Studies in RRMS data reporting the annualised relapse rate ratio and the disability progression risk ratio. 35
6 Summary results for placebo-controlled studies for the treatment effects on the risk of disability progression and annualized relapse rate ratio 37
7 Predictions obtained from all models for all studies in the "Sormani data" 38
8 Results of the comparison of the models for predicting the treatment effect on disability progression from the treatment effect on relapse rate 39
9 Results of the univariate meta-analyses of HAQ and DAS-28 separately, bivariate meta-analysis of HAQ and DAS-28 and trivariate meta-analysis of HAQ, ACR20 and DAS-28. 46
10 RRMS network data from systematic review by Melendez-Torres et al. 55
11 Estimates of the between-studies correlation and the between-studies standard deviation parameters (medians and 95% CrIs) obtained from fitting univariate and multivariate network meta-analysis models to the multiple sclerosis data. 56
12 Bivariate and univariate NMA estimates of all pairwise odd-ratios comparing the effectiveness of 8 interventions relative to one another on proportion relapse free. The upper triangle displays the bvNMA results whilst the lower triangle are the estimates from the uvNMA 59

13 Bivariate and univariate NMA estimates of all pairwise odd-ratios comparing the effectiveness of 6 interventions relative to one another on discontinuation due to adverse events. The upper triangle displays the bvNMA results whilst the lower triangle are the estimates from the uvNMA. 59
14 Results of the univariate meta-analyses of HAQ and DAS-28 separately, bivariate meta-analyses of HAQ and DAS-28 and trivariate meta-analyses of HAQ, ACR20 and DAS-28. Results include mean (SD) and [95% credible interval]. 112

List of Figures

1 Forest plots for HAQ: from URMA (left) and from BRMA of HAQ and DAS-28 (middle) and for DAS-28 also from BRMA (right). Graph shows estimates from the systematic review with 95% CIs (grey solid lines), predicted missing estimates from BRMA with 95% CrIs (grey dashed lines), "shrunken" estimates with 95% CrIs (black solid lines) and the pooled estimates with 95% CrIs (black solid lines for pooled effect from each of the meta-analyses and black dashed lines representing results from URMA for comparison). 24
2 Sources of evidence and the role of the data sets in the BRMA model. 26
3 Summary of the RRMS data. The point estimates and corresponding 95% confidence intervals, presented graphically and numerically, represent the annualised relapse rate ratio (left) and the disability progression risk ratio (right). 36
4 Scenarios for modelling multiple outcomes; (a) true treatment effects on all three outcomes are correlated and modelled using unstructured covariance matrix T, (b) true effects on outcomes one and three, δ_1 and δ_3 are conditionally independent, conditional on δ_2, which is modelled with structured covariance matrix equivalent to the precision matrix T^{-1} with element $\{1, 3\}$ equal to zero. 43
5 Network diagrams for proportion remaining relapse free (left graph) and discontinuation due to adverse events at 24 months (right graph). Treatment names: Placebo (plac), Glatiramer 20 mg daily (ga20), Glatiramer 40 mg thrice weekly (ga40), IFN-B-1a 22ug SC 3x daily (ifn1a22), IFN-B-1a 30ug IM weekly (ifn1a30), IFN-B-1a 44ug SC 3x weekly (ifn1a44), IFN-B-1a 250ug SC every other day (ifn1a250) and IFN-B-1a pegylated 125ug every 2 weeks. 54
6 Forest plot intervention effects relative to placebo on proportion of patients remaining relapse free estimated from bivariate (Models 2 to 4) and univariate (model 1) random-effects NMA models. Model 2 assumes zero within study-correlation whilst Models 3 and 4 assume $Unif(0, 1)$ and $Unif(-1, 1)$ prior distributions for the within-study correlation common to all-studies. 57
7 Forest plot intervention effects relative to placebo on discontinuation to due to adverse effect of treatment from bivariate random-effect network meta-analyses (Models 2 to 4) and univariate network meta-analysis (model 1). Model 2 assumes zero within-study correlation whilst Models 3 and 4 assume $Unif(0, 1)$ and $Unif(-1, 1)$ prior distributions for the within-study correlation common to all-studies. 58

Abbreviations and Definitions

ACR American College of Rheumatology. 21

BRMA bivariate random-effect meta-analysis. 17

bvNMA bivariate network meta-analysis. 50

DAS-28 Disease Activity Score. 21

EMA European Medicines Agency. 14

EUnetHTA European Network for Health Technology Assessment. 15

FDA Food and Drug Administration. 14

HAQ Health Assessment Questionaire. 21

HRQoL health related quality of life. 12

HTA health technology assessment. 11

ICER inceremental cost-effectiveness ratio. 12

MCMC Markov chain Monte Carlo. 15

MRMA multivariate random-effect meta-analysis. 42

MVMA multivariate meta-analysis. 11

mvNMA multivariate network meta-analysis. 50

NICE National Institute for Health and Care Excellence. 13

NMA network meta-analysis. 11

RA rheumatoid arthritis. 21

RRMS relapsing remitting multiple sclerosis. 28

SMC Scottish Medicines Consortium. 14

TRMA trivariate random-effect meta-analysis. 43

URMA univariate random-effect meta-analysis. 22

uvNMA univariate network meta-analysis. 56

1 Introduction

Evidence-based decision making requires careful synthesis of available evidence. When assessing new health technologies to, for example, make reimbursement decisions, the health technology assessment (HTA) process relies heavily on meta-analysis of effectiveness of new interventions. The evidence base is typically obtained from a systematic literature review of randomised controlled trials (RCTs) and sometimes real world evidence (RWE) such as from observational studies. There is often a lot of heterogeneity in reporting of clinical outcomes due to, for example, variety of scales on which effectiveness can be measured, different time points at which different studies report their outcomes, different outcome definitions, selective outcome reporting or disagreement about core outcomes. In addition to this, for some endpoints long follow up time is required before measurement of the treatment effect can be made and surrogate endpoints are reported by RCTs instead. These issues lead to obvious challenges in HTA decision making by limiting the evidence base when estimating parameters required in a decision modelling framework.

Bayesian statistics provides a flexible framework for modelling complex data structures by allowing multiple parameters to be modelled simultaneously. Network meta-analysis (NMA) facilitates simultaneous comparison of multiple treatment options whereas multivariate meta-analysis (MVMA) allows to jointly model treatment effects on multiple correlated outcomes. Whilst NMA has already become a standard tool for synthesising evidence in HTA, MVMA of multiple outcomes is gradually being introduced due to substantial methodological developments in this area in recent years [1, 2, 3, 4, 5, 6]. The uptake of this methodology has also increased due to many advantages and applications of this approach to evidence synthesis which includes evaluation of surrogate endpoints [7, 8, 9, 10, 11, 12, 13].

As discussed by Riley et al. [14], many clinical outcomes are correlated with each other, such as a hypertensive patient's systolic and diastolic blood pressure, level of pain and nausea in patients experiencing migraine, and a cancer patient's disease-free and overall survival times. Such correlation at the individual level will lead to correlation between effects at the population (study) level. For example, in a randomised trial of anti-hypertensive treatment, the estimated treatment effects for systolic and diastolic blood pressure are likely to be highly correlated. Similarly, in a cancer cohort study the estimated prognostic effects of a biomarker are likely to be highly correlated for disease-free survival and overall survival. Correlated effects also arise in many other situations, for example when there are multiple time-points (longitudinal data) [15]; multiple biomarkers and genetic factors that are interrelated [16]; multiple effect obtained from analyses using overlapping sets of covariates [17]; multiple measures of accuracy or performance (e.g. in regard to a diagnostic test or prediction model) [18], and multiple measures of the same construct (e.g. scores from different pain scoring scales, or biomarker values from different laboratory measurement techniques [19]). We broadly refer to these as multiple correlated outcomes.

Multivariate meta-analytic methods have a number of advantages [1]. They provide a natural framework for extending the evidence base to a larger number of studies (compared to the scenario where only a single outcome is analysed at a time). This can potentially lead to increased precision of the estimate of pooled effect on the outcome of interest [1, 5], in particular when the number of studies in the meta-analysis is small or there are missing data on an outcome of interest in a proportion of studies identified through a systematic review of literature. We discuss these issues in more detail in Section 4.4. Whether or not use of multivariate meta-analysis will result in more precise estimates, inclusion of evidence on other outcomes has other benefits.

In practice, when multiple outcomes are reported, they often are synthesised separately; that is, researchers conduct a standard univariate meta-analysis for each outcome separately. A consequence is that studies that do not provide direct evidence about a particular outcome are excluded from the meta-analysis of that outcome. This may be unwelcome, especially if their participants are otherwise representative of the population, clinical settings and condition of interest. Research studies require considerable costs and time, and entail precious participant involvement, and simply discarding them could be viewed as research waste if they still contain other outcomes that are

correlated with (and thus contain indirect information about) the outcome of interest. Statistical models for multivariate meta-analysis address this by simultaneously analysing multiple outcomes and accounting for their within-study and between-studies correlations. This allows more studies to contribute toward the meta-analysis for each outcome, which may improve efficiency and even decrease bias (e.g. due to selective outcome reporting [20]) in the summary estimates compared to separate univariate meta-analyses. Specifically, in addition to using direct evidence, the summary result for each outcome now depends on correlated results from related outcomes. The rationale is that by observing the related outcomes we learn something about the missing outcomes of interest, and thus gain some information that is otherwise lost.

One of the applications of multivariate meta-analysis, beyond synthesising evidence for purpose of obtaining pooled effects, is to model relationships between treatment effects on surrogate endpoints. Where lack of data on an outcome of interest presents an issue with populating a decision model, this problem may be ameliorated by predicting the treatment effect on this endpoint from another outcome (treated as a surrogate endpoint to the final outcome of interest). In a similar spirit, disease-specific measures of treatment effect on health related quality of life (HRQoL) may be synthesized jointly with treatment effects reported on a common scale of HRQoL, to make predictions if such outcome is unreported for an intervention under consideration but is desirable in a decision-making context. In the context of HRQoL, multivariate meta-analysis has been used, for example, to synthesize jointly multiple instrument-specific preference-based values in coronary heart disease and the underlying disease-subgroups, stable angina and post-acute coronary syndrome [21].

When multiple clinical endpoints are included in an economic model, the multivariate meta-analysis accounting for the correlation between these endpoints will be important for two reasons: (1) it will change the point estimate of the inceremental cost-effectiveness ratio (ICER) if the model includes a non-linear function of the two endpoints (these ICERs will in fact be biased unless the correlation is reflected); and (2) it will change the estimates of decision uncertainty generated by the the models i.e. probabilities that interventions are cost-effective or value of information estimates. Taking account of the correlation between multiple endpoints could, therefore, have important implications for decision making, as many models utilise two or more decision endpoints (e.g. PFS and OS in cancer models).

In the remainder of this introduction, we discuss in more detail a number of scenarios where the methodology proofs to be useful. Before we do so, we also highlight some advantages of the Bayesian framework of implementing these methods, which this TSD is focussed on.

1.1 Bayesian approaches to multivariate meta-analysis

As mentioned above, one of the advantages of integrating data on multiple outcomes through a multivariate approach is that of borrowing of strength across outcomes, which can potentially lead to reduced uncertainty around the resulting effectiveness estimates. In a Bayesian framework, a wide range of sources of evidence can be integrated, which is an important factor in evidence-based medicine [22, 23, 24]. For example, additional data (external data from observational studies, clinical trials or systematic reviews) or experts' opinions can be incorporated in the form of prior distributions, which can further inform a multivariate meta-analysis model. The use of external data can potentially lead to a further reduction of uncertainty around the estimates. A Bayesian approach to the analysis also allows for making direct probabilistic statements about the parameters of interest, such as, for example, the probability of a hazard ratio being less than 1 (probability that a new treatment reduces risk of death or progression compared to a control treatment arm). Such direct probability statements are conditional on the model and on the chosen prior distributions. Multivariate meta-analytic methods in the Bayesian framework are particularly useful in the HTA context due to their flexibility in modelling uncertainty around all relevant parameters. Such methods have been well developed [3, 5, 6] in a number of application areas, including surrogate endpoints [8, 11, 12, 13].

1.2 Examples of multivariate meta-analysis used in the context of NICE

MRMA is often used to obtain pooled treatment effects on two or more correlated outcomes when the number of studies reporting an outcome of interest is limited but additional data are available from studies reporting alternative, correlated endpoints. For example, bivariate meta-analysis has been used in a technology appraisal of continuous positive airway pressure devices for the treatment of obstructive sleep apnoea—hypopnoea syndrome [25, 26] conducted by the National Institute for Health and Care Excellence (NICE). In this appraisal, a range of endpoints were considered when evaluating the clinical effectiveness. The endpoints included Epworth Sleepiness Scale (ESS) and ambulatory blood pressure monitoring (ABPM). However, only a subset of studies, identified by a systematic literature review, reported ABPM. Modelling the two endpoints jointly in a bivariate meta-analysis allowed the analysts to include all of the studies in the meta-analysis. In this case results were not dramatically different from those obtained using conventional univariate meta-analysis, however such analysis is still valuable as it provides a sensitivity analysis for potential impact of outcome reporting bias. That is, by borrowing information from correlated outcomes that are reported, the meta-analyst can reduce the impact of unreported outcomes, and see if conclusions are robust.

Bivariate approach to meta-analysis was also used in another NICE technology appraisal of tumour necrosis factor-α inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis [27, 28]. A decision model was developed with a generalised framework for evidence synthesis that pooled change in disease activity, measured by Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), and simultaneously synthesised information on function, measured by Bath Ankylosing Spondylitis Functional Index (BASFI) to determine the long-term quality-adjusted life-year and cost burden of the disease in the economic model. In this work, BASFI and BASDAI were synthesised jointly (using the bivariate meta-analytic approach) in order to generate appropriate effect-size estimates and their associated uncertainty to inform the main input parameters of the economic model. In the decision model, prognosis, costs and QALYs were determined by absolute BASDAI and BASFI scores. Correlation between treatment effects on these two endpoints was important for the cost-effectiveness model that relied heavily on these two measures. Extending the synthesis modelling to consider BASFI scores not only allowed all relevant evidence to contribute to the synthesis but also ensured that all measures were synthesised together to reflect the expected correlations between the effects on the two outcomes. Uncertainty was also more appropriately quantified compared to synthesising each outcome separately.

When not all relevant estimates are available from the relevant clinical studies, not only may data from some studies have to be discarded but also the health economic model may have to be simplified. Multi-parameter evidence synthesis can be used to combine data from diverse sources of evidence, which results in obtaining estimates, required in HTA decision making, that otherwise may not be available. For example, Tan et al. [29] demonstrated how bivariate meta-analysis can be used to predict an unreported estimate of a treatment effect enabling implementation of a multi-state Markov model, which otherwise needed to be simplified. To illustrate this, the authors used an example of cost-effectiveness analysis for docetaxel in combination with prednisolone in metastatic hormone-refractory prostate cancer. Bivariate meta-analysis was used to model jointly available data on treatment effects on overall survival and PFS to predict the unreported effect on PFS in a study evaluating docetaxel with prednisolone. The predicted treatment effect on PFS enabled implementation of a three-state Markov model comprising of stable disease, progressive disease and dead states, whilst lack of the estimate restricted the model to a two-state model (with alive and dead states). The two-state and three-state models were compared by calculating the incremental cost-effectiveness ratio (which was much lower in the three-state model: £22,148 per QALY gained compared to £30,026 obtained from the two-state model) and the expected value of perfect information (which increased with the three-state model). The three-state model has the advantage of distinguishing surviving patients who progressed from those who did not progress. Hence, the use of advanced meta-analytic techniques allowed obtaining relevant parameter estimates

to populate a model describing disease pathway more appropriately, whilst helping to prevent valuable clinical data from being discarded.

1.3 Use of multivariate meta-analysis for surrogate endpoint evaluation

One of the most useful applications of bivariate and multivariate meta-analysis include the evaluation of surrogate endpoints [7, 8, 9, 10, 11, 12, 13]. Surrogate endpoints have become an important part of HTA and regulatory decision making [30, 31]. Bivariate meta-analysis can be used to model treatment effects on surrogate and final outcomes jointly, with the advantage of obtaining the estimates of clinical effectiveness early. When the treatment effect of a new intervention under consideration is not yet reported on the final clinical outcome, licensing of the treatment can be agreed conditional on a surrogate endpoint. When making such conditional licensing decisions, a bivariate meta-analysis can be used to predict a likely treatment effect on the final outcome from the treatment effect measured on the surrogate endpoint. Such yet unmeasured (or for other reasons unreported) effect predicted from the effect measured on the surrogate endpoint (or from other highly correlated effect) can inform a health economic model, which otherwise may have to be simplified in a way that compromises its usefulness and external validity [29].

1.3.1 Surrogate endpoints in regulatory decision making and HTA

Regulatory agencies such as the European Medicines Agency (EMA) and the Food and Drug Administration (FDA) in the US have introduced flexible licensing pathways, for example by allowing conditional licensing based on treatment effect measured on a surrogate endpoint [32, 33]. The effectiveness measured on the final outcome is obtained when data become more mature and the drug is then re-evaluated. For example, between Jan 2008 and Dec 2012, FDA made 36 of 54 cancer drug approvals (67%) on the basis of a surrogate endpoint: 19 based on response rate and 17 based on PFS or disease free survival [34]. Conditional licensing often takes place, for example, for orphan medicines (used against rare, life-threatening or chronically debilitating conditions) where the number of patients in trials is likely to be small leading to uncertainty around early measurement of treatment effect. For example olaratumab (Lartruvo), used in patients with advanced soft tissue sarcoma, was granted a conditional marketing authorisation by the EMA, based on early evidence of benefits of the treatment measured on progression-free survival (PSF) and some evidence of increased overall survival (OS) time. Due to small number of patients in the study, this licensing decision was conditional on data collected from an ongoing study confirming the efficacy and safety of the medicine [35]. There are many other examples of accelerated approval made by regulatory bodies based on a surrogate or early measurement. FDA, for example, granted accelerated approval of bevacizumab (Avastin) for use in combination with paclitaxel as the first line treatment of patients with metastatic HER2-negative breast cancer. The decision was based on early measurement of the treatment effect on PFS, conditional on verification of the treatment effect on PFS and additional information on the effects on OS. This decision was subsequently withdrawn as the clinical trials failed to show sufficient benefit and there were substantial adverse reactions observed in patients [36].

Although the use of surrogate endpoints in the regulatory processes speeds up the licensing of new health technologies, the new therapies then go through another scrutiny at the HTA stage. The reimbursement decisions conducted by the HTA agencies, such as NICE and the Scottish Medicines Consortium (SMC) in the UK, are typically made based on estimates of cost-effectiveness of new health technologies. To obtain such estimates, the long term estimates of effectiveness are required, amongst many parameters including costs and utilities, to populate a health economic decision model. When new therapies are licensed based on surrogate endpoints, the long term estimates are not available until more mature data are available for reassessment. To expedite access of new therapies to patients, technology appraisals need to be based on an estimate of the effect of the therapy measured on the surrogate endpoint. For example, a NICE technology appraisal committee

has approved venetoclax in combination with rituximab for treating relapsed or refractory chronic lymphocytic leukaemia using an estimate of the treatment effects on PFS [37]. This was made on the basis that treatment effect on PFS was deemed a surrogate for the effects on OS.

HTA agencies, however, are cautious about the use of surrogate outcome data and highlight the importance of an appropriate use of such endpoints in their guidelines. This is particularly reflected in the guidelines published by the European Network for Health Technology Assessment (EUnetHTA) [38] as well as NICE's current guidance on methods for manufacturers [39]. The guidelines recommend that HTA analysts and decision-makers should be cautious about using surrogate endpoints and use them only if they have been appropriately validated. The additional uncertainty associated with using surrogate endpoints to predict cost-effectiveness should also be fully explored.

NICE guidelines (section 4.4.3) highlight the importance of identifying appropriate outcome and surrogate outcome measures. Modelling methods (as discussed in section 5.7 of the NICE guidelines) provide *"an important framework for synthesising available evidence and generating estimates of clinical and cost effectiveness in a format relevant to the Appraisal Committee's decision-making process. Models are required for most technology appraisals"*. The guidelines list a number of situations when modelling is likely to be required. With regards to endpoints, the guidelines state that *"When the use of "final" clinical end points is not possible and "surrogate" data on other outcomes are used to infer the effect of treatment on mortality and health-related quality of life, evidence in support of the surrogate-to-final end point outcome relationship must be provided together with an explanation of how the relationship is quantified for use in modelling. The usefulness of the surrogate end point for estimating QALYs will be greatest when there is strong evidence that it predicts health-related quality of life and/or survival. In all cases, the uncertainty associated with the relationship between the end point and health-related quality of life or survival should be explored and quantified"*.

Bivariate meta-analytic methods are needed to evaluate the surrogate-to-final relationships because these methods by nature take into account not only the correlation between these outcomes (or between the treatment effects measured on these outcomes) but also all related uncertainty required in decision modelling as stated by the guidelines. Without taking the correlation into account (either directly or through some functional relationship between the correlated effects), it is not possible to quantify the surrogate relationship. Other methods, such as linear or meta regression do not take into account all relevant uncertainty (for example ignore the uncertainty associated with the treatment effect on the surrogate endpoint) whilst the bivariate meta-analysis appropriately accounts for all relevant uncertainty, both within and between studies [12].

1.4 Implementation

In this TSD, we present a number of examples to demonstrate how to fit the models and interpret the results. All models were implemented in WinBUGS [40], where the estimates were obtained using Markov chain Monte Carlo (MCMC) simulation. Convergence was checked by visually assessing the history, chains and autocorrelation using graphical tools in WinBUGS. All posterior estimates are presented as means with standard deviations and 95% credible intervals (CrIs). Issues related to MCMC and use of WinBUGS are discussed in a TSD by Dias et al. [41]. Some methods presented in this TSD also make use of R, in particular where multiple runs of a model is recommended (when, for example, validating surrogate endpoints). We then also use R2WinBUGS.

1.5 Structure of the TSD and the level of required expertise

In the remainder of this TSD, we discuss the methods of multivariate meta-analysis in the order of increased complexity. We begin with the use of bivariate meta-analysis in Section 2 where we discuss methodological aspects of the method including those specific to the Bayesian approach and present application of the method to an example in rheumatoid arthritis. We then expand

on the use of bivariate meta-analysis in Section 3, where we discuss its application to modelling surrogate endpoints. We provide guidance on how to evaluate surrogate endpoints as predictors of clinical benefit. The methods we include are for modelling surrogate relationships between the treatment effects on two outcomes (a surrogate endpoint and a final clinical outcome) at the study level using aggregate data. We give an example of use of the methods in multiple sclerosis. In Section 4, we extend the methods to the multivariate meta-analysis of treatment effects on multiple outcomes beyond bivariate case and demonstrate the use of the methods on the extended example in rheumatoid arthritis. We also include a discussion on use of this methodology for evaluating multiple surrogate endpoints as joint predictors of clinical benefit. Finally, in Section 5 we show how the multivariate meta-analytic methods extend to multivariate network meta-analysis and demonstrate the use of the method in an example in multiple sclerosis.

All methods are demonstrated by applying them to the illustrative examples. The purpose of the examples is three-fold (i) to show how to implement the methods, (ii) to show how to interpret the results, and (iii) to show how the results of using multivariate approach differ from simpler univariate methods. We provide the code for the methods in the appendices which also contain data to allow the reader to reproduce the results included in the TSD. The only exception is for a method of analysis of individual level data described in Section 2.4.4 where, due to lack of sharing permission (as well as space in the manuscript), we only show the structure of the data (but this should be sufficient for the reader to be able to apply the method to their own data).

The complexity of the methods is relatively high, in particular in the later chapters. We are aware that many analysts, both in academia and industry have high level technical competence to understand the technical details and they may appreciate this level of detail. Having said that, analysts of less expertise in terms of mathematical description of methods will still be able to use the TSD. In particular analyst who are comfortable using the previous TSDs in evidence synthesis (in particular TSD 2 [41]) and running WinBUGS programmes included in them should have no difficulty reproducing analyses presented in this TSD and adapting the methods to their own data. For those readers with less background knowledge in statistics, we recommend skimming through the formulae (but still reading through the sections to understand the differences between the alternative models and corresponding assumptions) and focusing their attention on the examples.

2 Bivariate random-effects meta-analysis (BRMA)

In this Section, we introduce the bivariate random-effect meta-analysis (BRMA) model beginning with the technical description of the model, a discussion of the components of the model and the availability of data required to populate the model. We discuss in detail two alternative parameterisations of the method and apply them to an example in rheumatoid arthritis. In the final part of this section we demonstrate extended analysis based on this example, this time including additional external data in the form of a prior distribution, illustrating advantages of the Bayesian framework.

2.1 BRMA in a Bayesian framework with considerations of appropriate prior distributions

The BRMA model for correlated and normally distributed treatment effects on two outcomes Y_{1i} and Y_{2i} is usually presented in the form described by van Houwelingen *et al.* [2] and Riley *et al.* [4]:

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix}, \mathbf{\Sigma_i} \right), \ \mathbf{\Sigma_i} = \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi} \\ \sigma_{1i}\sigma_{2i}\rho_{wi} & \sigma_{2i}^2 \end{pmatrix} \tag{1}$$

$$\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \mathbf{T} \right), \ \mathbf{T} = \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho \\ \tau_1\tau_2\rho & \tau_2^2 \end{pmatrix}. \tag{2}$$

In this model, the treatment effects on the two endpoints Y_{1i} and Y_{2i} are assumed to estimate the correlated true effects δ_{1i} and δ_{2i} with corresponding within-study variances σ_{1i}^2 and σ_{2i}^2 of the estimates and the within-study correlation ρ_{wi} between them in each study i. In this hierarchical framework, these true study-level effects are assumed exchangeable. This means that all δ_{1i} and δ_{2i} are "similar" in a way which assumes that the trial labels, i, attached to these treatment effects are irrelevant [41]. Since they are exchangeable and correlated, they follow a common bivariate normal distribution with means (d_1, d_2) corresponding to the treatment effects on the two outcomes, between-studies variances τ_1^2 and τ_2^2 and a between-studies correlation ρ. Equation (1) represents the within-study model and (2) is the between-studies model. The true effects δ_{1i} and δ_{2i} are Bayesian "shrunken" estimates, in which inference for the treatment effect in each particular study is performed by borrowing strength from the other studies [42]. In the multivariate meta-analysis this inference is performed by also borrowing strength from the other outcomes.

2.1.1 Within-study covariance

Within-study correlation between treatment effects on two (or multiple) outcomes occurs because these effects are measured on the same patients. When modelling correlated effects in the multivariate-meta-analysis, the correlation needs to be properly accounted for. Ignoring the within-study correlation may lead to increased uncertainty around the estimated pooled effects and potentially biased results [43].

In the above model, the elements of the within-study covariance matrix Σ_i are assumed to be known. Whilst the estimates of the within-study variances are usually obtainable from published studies as the squares of the standard errors of the treatment effects for each outcome, the estimates of the within-study correlations between the treatment effects on the two outcomes are more difficult to obtain as they would not be reported in the original articles. When individual participant data (IPD) are available, the correlation can be obtained by bootstrapping [8] or alternatively by fitting a regression model for the two outcomes with correlated errors [44]. For continuous outcomes, the latter approach is relatively straightforward (we include an example of modelling IPD for two outcomes with correlated errors in Section 2.4). However, when dealing with other types of data, such as binary or time to event data, such analysis becomes more complex and bootstrapping technique may be preferable. Bootstrapping involves generating multiple data sets by sampling with replacement from the data the same number of observations as in the data set. Note that

we sample pairs of data for the bivariate case or triplets for the meta-analysis of treatment effects on three outcomes and so on, to preserve the correlation structure. Having simulated a number of data sets (perhaps a 1000 data sets), in each data set we calculate treatment effect on each outcome on the same scale as is used in the meta-analysis (the scale of Y_{ij}), for example log odds ratios for binary outcomes. This way we will obtain 1000 pairs (for the bivariate case, or triplets for trivariate analysis) of log odds ratios from each bootstrap sample. Having obtained 1000 log odds ratios on one outcome and 1000 log odds ratios on the second outcome, we can now calculate the correlation between them. This is an estimate of the within-study correlation between the treatment effects on the two outcomes. This is not possible to achieve simply by analysing the data as only a single pair of log odds ratios can be obtained from a single data set, making it impossible to calculate the correlation. Similarly, bootstrapping can be used to obtain the within-study correlation between treatment effects on other types of outcomes. For time to event data, a survival model, such as Cox or Weibull regression can be used to estimate log hazard ratio on each outcome within each bootstrap sample. These multiple pairs of log hazard ratios are then used to obtain the correlation between them. It is also possible to carry out this procedure for mixed outcomes, for example to obtain a correlation between log odds ratio on a binary outcome and log hazard ratio on a time to event outcome.

Often IPD are available only from a small subset of studies or not available at all. When, for example, IPD are available for a single study, an assumption can be made that the correlation is the same across the studies. In the Bayesian framework, prior distributions can be placed on the within-study correlations to account for uncertainty around the estimate. This can be useful when the within-study correlations either are missing from all studies or are available only from a subset of studies and prior distributions are placed on the correlations for studies that do not provide them. Informative prior distributions for the correlations can be constructed, for example, by double bootstrap of IPD (either from studies included in the meta-analysis, or from an external study in a similar population). Double bootstrapping involves bootstrapping first from the data (sampling with replacement) and then another level of bootstrapping from each bootstrap sample. This results in a number of correlations rater than a single correlation as it is the case in a single bootstrap. This set of correlations forms a distribution of correlation values (correlations with uncertainty) as implemented by Bujkiewicz et al [5].

In the absence of IPD, a range of approaches can be explored to obtain missing within-study correlation. A straightforward method is to investigate a range of values (between -1 and 1) in a sensitivity analysis. Another approach to estimating the within-study correlation was proposed by Wei and Higgins who derived formulae expressing the within-study correlations between the treatment effects (such as log odds ratios) on the two outcomes in terms of, perhaps more likely reported, correlations between other measures (such as probabilities of events on the two outcomes for binomial data) [45]. Moreover, an alternative formulation of bivariate meta-analysis for studies with unknown within-study correlation, proposed by Riley et al [46], can be used. It combines covariances from both the within-study and between-studies models in a single term, which avoids the need to specify the within-study correlations. A limitation of this later method is that this single term cannot be interpreted as either within-study or between-studies correlation and as such has a limited use, for example in surrogate endpoint evaluation.

2.1.2 Between-studies model parameters

To implement the model in the Bayesian framework, prior distributions are placed on the mean effects, for example vague prior distributions with large variances $d_1 \sim N(0, 10^4)$, $d_2 \sim N(0, 10^4)$ and on the between-studies variances and correlation. In the general case, for any number of outcomes, a prior distribution has to be placed on the whole variance-covariance matrix to ensure that it is properly defined. We discuss this in more detail in Section 4. In the bivariate case, such as considered in this section, a proper definition can be achieved by placing prior distributions on the variances ensuring that the variances are restricted to plausible positive values, for example

by placing uniform prior distributions on the corresponding standard deviations $\tau_j \sim Unif(0, 2)$, and by choosing a prior distribution for the correlation which restricts it to values between -1 and $+1$, for example $\rho \sim Unif(-1, 1)$. Alternatively, a prior distribution with less weight on the edges of the plausible range (values ± 1) can be constructed, for example by using a beta distribution: $\frac{\rho+1}{2} \sim Beta(1.5, 1.5)$ [47, 48], which may help with estimation. When additional information (possibly based on expert knowledge) about the between-studies correlation exists, a weakly informative prior distribution, for example limited to positive or negative values can be used as recommended by Burke et al [47]. Fully informative prior distribution based on, for example, external data, as implemented by Bujkiewicz et al [5], can be constructed. This can potentially result in additional borrowing of information both across studies (when the correlation is known to be high) and from the external data.

2.1.3 Accounting for missing data

When summarising evidence across multiple endpoints, it is common to encounter instances where some studies do not report information for all outcomes of interest leading to incomplete vectors with missing study-specific effects for the outcomes not reported [5,10]. Such studies can be included in a meta-analytic model under the assumption that the effects for outcomes not reported are missing at random (i.e. there might be systematic differences between the missing and observed values, but these can be entirely explained by other observed variables [49]). When implemented using MCMC, for example in the WinBUGS software, the missing study effects and standard errors are coded as NA in the data, a strategy previously outlined in Bujkiewicz et al. [5] and Dakin et al. [50]. Missing values require distributions to be placed on them, which enables MCMC to automatically "impute" values for the missing information under missing at random assumption with predicted distributions.

Following the strategy by Bujkiewicz et all [5], defining the bivariate model, as in (1), already provides distributions for missing values of treatment effect Y_{ji} on outcome j in study i. However, corresponding missing standard errors σ_{ji} still need to be estimated. Independent prior distributions can be placed on the missing standard errors (or corresponding population variances). A WinBUGS code for this approach can be found in Appendix B.4 for cross-validation procedure when using bivariate meta-analytic methods for evaluating surrogate endpoints, described in the next chapter.

Additional borrowing of information can be achieved by assuming exchangeability of the population variances $var_{ji} = \sigma_{ji}^2 N_i$ (note that the assumption of exchangeability of the variances of the mean, σ_{ji}^2, won't hold as these variances depend on the study size N_i). The population variances are assumed to come from the same distribution, for example:

$$\begin{aligned} var_{ji} &\sim N(0, h_j)I(0,), \\ h_j &\sim \Gamma(1.0, 0.01). \end{aligned} \tag{3}$$

By borrowing of information from the studies reporting the Y_{ji} for outcome j and the corresponding standard errors, the variances (and hence standard errors) for studies with missing Y_{ji} are predicted by the MCMC simulation, conditional upon both the data and the posterior estimates of the model parameters. WinBUGS code corresponding to this model (with data for the example introduced in section 2.3) is included in Appendix A.1.

Note that we used the total sample size N_i when calculating the population variances var_{ji} in (3). This approach is suitable for single arm studies or comparative studies with equal allocation in the arms. For studies with unbalanced treatment allocation, we can instead calculate population variances in the arms, which (assuming, as in a t-test, that they are the same in both arms A and B) are equal $var_{ji} = \sigma_{ji}^2 / \left(\frac{1}{N_{Ai}} + \frac{1}{N_{Bi}}\right)$, where N_{Ai} and N_{Bi} are the numbers of individuals in arms A and B respectively and σ_{ji} is, as in the above model, the SE of the treatment difference. We then assume that these population variances in the arms are exchangeable using the formulae (3). This approach is illustrated in Section 3.7.4 discussing prediction of the treatment effect on the

final outcome from the treatment effect on a surrogate endpoint in a new study, with an example code in Appendix B.5.

An alternative is to enter an arbitrary treatment effect estimate for the missing outcome, and give a very large variance (e.g. 1000000) and associated covariances as zero.

2.2 BRMA in the product normal formulation (PNF)

As proposed by Bujkiewicz et al. [5] and extended to the context of surrogate endpoints by Bujkiewicz et al. [13], the BRMA model (1)–(2) can be parameterised in an alternative form. Specifically, the between-studies model (2) is represented in the product normal formulation (PNF) [51, 5] (a product of univariate conditional normal distributions), whereas the within-study model remains the same:

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix}, \mathbf{\Sigma_i} \right), \ \mathbf{\Sigma_i} = \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi} \\ \sigma_{1i}\sigma_{2i}\rho_{wi} & \sigma_{2i}^2 \end{pmatrix} \tag{4}$$

$$\begin{cases} \delta_{1i} \sim N(\eta_1, \psi_1^2) \\ \delta_{2i} \mid \delta_{1i} \sim N(\eta_{2i}, \psi_2^2) \\ \eta_{2i} = \lambda_0 + \lambda_1 \delta_{1i}. \end{cases} \tag{5}$$

As for the BRMA model, Y_{1i} and Y_{2i} are the estimates of the correlated treatment effects measured by two outcomes, and the δ_{1i} and δ_{2i} are the true effects in the population which are correlated and are assumed exchangeable, and therefore are modelled here as random effects with a linear relationship.

Instead of placing independent non-informative prior distributions on all the parameters of model (5), relationships between these parameters and the elements of the between-studies covariance matrix $\mathbf{T}$ in the between-studies model (2) are derived to allow to take into account the inter-relationship between the parameters. These relationships have the following forms

$$\psi_1^2 = \tau_1^2, \ \ \psi_2^2 = \tau_2^2 - \lambda_1^2 \tau_1^2, \ \ \lambda_1 = \frac{\tau_2}{\tau_1}\rho, \tag{6}$$

Having established these relationships allows us to place prior distributions on the between-studies standard deviations and correlations. This is an easier task (compared with placing prior distributions on all the parameters of model (5)) as the plausible range of values for these parameters are known or can be obtained from external sources of information (see for example Higgins and Whitehead [52] or Turner et al [53, 54] for the heterogeneity parameter or Bujkiewicz et al. [5] for the correlations). By placing prior distributions on these parameters, for example for the between-studies correlation $\rho \sim Unif(-1, 1)$, and the between-studies standard deviations for each outcome $j = 1, 2$; $\tau_j \sim Unif(0, 2)$, the above derived relationships give the implied prior distribution on the parameters λ_1, ψ_1 and ψ_2 [5, 13]. The remaining parameters are given "vague" prior distributions, $\eta_1 \sim N(0, 1000)$, $\lambda_0 \sim N(0, 1000)$.

The pooled effects d_j on the two outcomes in model (2) are directly linked to the parameters of model (5); $d_1 = \eta_1$, $d_2 = \lambda_0 + \lambda_1 d_1$. It is possible to center the true effects on the first outcome (by replacing δ_{1i} with $(\delta_{1i} - \overline{\delta_{1i}})$ in the third line of formulae (5)) in which case the pooled effect on the second outcome will be equal to the intercept; $d_2 = \lambda_0$. This can be useful if there are problems with autocorrelation.

This model is equivalent to the BRMA model described in the previous section when all prior distributions are the same and the data are available on both outcomes. In the presence of missing data, the results from these two methods will differ. Moreover, in the presence of missing data results from BRMA PNF may depend on the order of outcomes.

The BRMA PNF model is particularly useful when applied to surrogate endpoint evaluation as discussed in Sections 3.3 and 4.5. This parameterisation also becomes very useful in higher dimensions where many parameters need to be estimated, potentially requiring data from many studies.

The parameterisation in a simplified form may be used, by assuming conditional independence between treatment effects on some pairs of outcomes, reducing the number of parameters. Models for multiple outcomes are discussed in Section 4.

WinBUGS code corresponding to this model (with data for the example introduced in section 2.3) is included in Appendix A.2.

2.3 Example: rheumatoid arthritis

2.3.1 Data

A systematic review and meta-analysis was carried out to investigate the effectiveness of anti-TNF-α inhibitors such as etanercept, infliximab and adalimumab used as second line treatments in patients with rheumatoid arthritis (RA) [55]. Standard instruments for measuring response to treatment in RA were considered: the Health Assessment Questionaire (HAQ) and Disease Activity Score (DAS-28) measures, and the American College of Rheumatology (ACR) response criteria. The results of the meta-analyses of all three outcomes individually showed that the biologic interventions are effective when used sequentially. Data collected in this systematic review are used to investigate how multivariate meta-analysis can be applied to incorporate multiple outcomes, such as ACR20 (20% response according to the ACR criteria; a binary outcome modelled on log odds scale) and changes from baseline of DAS-28 and HAQ scores (continuous outcomes), in evidence synthesis which aims to estimate the change from baseline of the HAQ score. The estimate of the HAQ is of particular interest to clinicians and decision-makers in health care as it is often used to estimate quality of life of patients following treatments of RA.

Table 1: Studies in RA data reporting outcomes: ACR20, DAS-28 and HAQ.

Study		ACR20	DAS-28	HAQ
	n	r	mean* (se)	mean* (se)
Bennet 2005	26	–	-1.7 (0.25)‡	-0.31 (0.13)
Bingham 2009	188	85	-1.6 (0.1)	-0.35 (0.05)
Bombardieri 2007	810	486	-1.9 (0.05)	-0.48 (0.02)
Buch 2005	25	18	–	–
Buch 2007	72	55	-1.47 (0.18)	–
Cohen 2005	30	–	-1.87 (0.24)	–
Di Poi 2007	18	–	-2.1 (0.29)	–
Finckh 2007	66	–	-0.98 (0.18)	–
Haroui 2004	22	14	–	-0.45 (0.14)
Hjardem 2007	123	–	-1 (0.11)	–
Hyrich 2008	331	–	–	-0.12 (0.03)
Iannone 2009	37	–	–	0.15 (0.13)
Karlsson 2008	337	172	–	–
Laas (InTol) 2008	6	–	-1.17 (0.66)	–
Laas (InEff) 2008	20	–	-1.26 (0.35)	–
Navarro-Sarabia 2009	83	–	-1.1 (0.18)	-0.21 (0.07)
Nikas 2006	24	18	-2.4 (0.16)	–
Van der Bijl 2008	41	19	-1.5 (0.25)	-0.21 (0.08)
Van Vollenhoven 2003	18	12	–	–
Wick (EA) 2005	9	7	-1.9 (0.22)	–
Wick (IA) 2005	27	19	-1.3 (0.28)	–

$*$ mean change from baseline; se – standard error, r – number of responders
n – total number of participants in the study.

Table 1 gives details of the three outcomes which were reported in each of the studies within

the systematic review. Some of the listed outcomes or treatment effects (and in particular the standard error) were obtained from other measures or imputed. The details of how these values were obtained can be found in Appendix 2 of the original meta-analysis of this data [55]. We will refer to these data as the "RA data" throughout this document. In this section, for illustration of BRMA, we use data on DAS-28 and HAQ. Data on all three outcomes will be used in Section 4.

Note that all studies in this example were single arm (with the treatment effect measured as change from baseline). This example is used for convenience and the same method can be applied for data on the relative treatment effects (for example data containing treatment effects on the log odds ratio scale with the corresponding standard errors). For clarity, we emphasise that the TSD is not advocating use of single trial arm change from baseline data as an ideal approach for inferring treatment effects.

2.3.2 Results

The results of applying the univariate and bivariate meta-analyses of HAQ and DAS-28 are included in Table 2. The BRMA model allowed us to include 10 cohorts (from 8 studies reporting DAS-28 but not HAQ in Table 1) in addition to those 8 that could be used in the univariate random-effect meta-analysis (URMA) of HAQ (studies reporting HAQ in Table 1). The within-study correlation between the mean change from baseline in DAS-28 and HAQ, obtained by bootstrapping of IPD from a study external to the meta-analysis, was relatively weak; $\rho_{wi}^{das,haq} = 0.24$. It was assumed the same across all studies in the meta-analysis.

Table 2: Results of the univariate meta-analyses of treatment effects on HAQ and DAS-28 separately and bivariate meta-analysis of the treatment effects on HAQ and DAS-28. Negative values (reduction from baseline for HAQ or DAS-28) represent positive effect of the treatment.

	posterior mean treatment effect (SD) [95% CrI]			
	univariate analyses		bivariate analyses	
	HAQ	DAS-28	HAQ & DAS-28	
			BRMA	BRMA (PNF)
HAQ	-0.25 (0.09) [-0.43,-0.07]		-0.27 (0.08) [-0.42,-0.11]	-0.27 (0.07) [-0.39,-0.13]
DAS		-1.57 (0.13) [-1.84,-1.31]	-1.53 (0.13) [-1.78,-1.26]	-1.53 (0.13) [-1.78,-1.26]
τ_H	0.21 (0.09) [0.10,0.44]		0.21 (0.08) [0.11,0.41]	0.22 (0.08) [0.11,0.41]
τ_D		0.44 (0.11) [0.25,0.67]	0.44 (0.11) [0.27,0.71]	0.44 (0.11) [0.27,0.71]
ρ^{DH}			0.58 (0.42) [-0.51,0.99]	0.59 (0.43) [-0.51,0.99]

When using BRMA, the estimate of the mean effect on HAQ shifted towards a more extreme value (change in HAQ from baseline shifts from -0.25 in URMA to -0.27 in BRMA) with uncertainty reduced by 14% of the width of the 95% CrI when using BRMA and 28% when using BRMA PNF compared to the interval from URMA. The estimate of DAS-28, however, moved to a less extreme value when using BRMA (from -1.57 to -1.53), with only marginally reduced uncertainty. Adding more data to the meta-analysis by using BRMA reduced uncertainty but not heterogeneity; the between-studies standard deviations τ_H and τ_D, corresponding to the change from baseline in HAQ

and DAS-28 respectively, remain almost the same, with only reduced uncertainty for τ_H. The between-studies correlation was estimated with a large credible interval, 0.59 (-0.51, 0.99) form BRMA PNF and a similar results was obtained from BRMA in the standard form. The large interval is likely to result from a small number of studies in the data, thus being dominated by the non-informative prior distribution. Example of using external data to construct an informative prior distribution is discussed in Section 2.4, where much narrower interval was obtained when using such informative prior compared to non-informative prior distribution. The results are based on 50,000 iterations after a burn-in of 50,000.

Figure 1 shows three forest plots representing pooled effects estimates on HAQ from URMA (left) and BRMA (middle), and DAS-28 from BRMA (right). Black solid lines correspond to the "shrunken" estimates δ_{ji} and pooled estimates d_j, the grey solid lines show the estimates obtained from the systematic review (data used in this meta-analysis) and the grey dashed lines (estimates also marked with a $*$ on their right) correspond to the predicted estimates for the studies that did not report the HAQ but reported the DAS-28 (or reverse in plots for the DAS-28). Dashed black lines below the estimates obtained from BRMA represent the pooled estimates obtained from URMA for comparison.

In the forest plot of the estimates from URMA, the estimates are shrunken towards the mean, which is especially noticeable for those studies with higher uncertainty around the known estimates (i.e. Bennet, Haroui and Iannone). For example, in the case of the Haroui study, the HAQ estimate of -0.45 (95% confidence interval (CI): -0.73, -0.17) is now shifted towards the overall mean (which for URMA is -0.25) with shrunken estimate equal -0.38 (95% CrI: -0.6, -0.15). Borrowing of strength across studies led to both the shift towards the mean and decrease in uncertainty reducing the width of the credible interval by 16% for the Haroui estimate.

In the middle forest plot in Figure 1 (representing estimates of HAQ from BRMA) the predicted estimates (including those for studies not reporting HAQ) contribute to the pooled estimate even though there is considerable uncertainty associated with them. This borrowing of information across outcomes leads to the aforementioned reduction in uncertainty around the pooled estimate of the HAQ.

As can be seen in Figure 1 (middle and right-hand-side forest plots of estimates from BRMA), the predicted effects for the HAQ (DAS-28) follow the heterogeneity pattern of the corresponding known estimates from the DAS-28 (HAQ) due to the accounting for the correlation between the effects on the two outcomes. Studies that reported the DAS-28, but not the HAQ, had on average relatively high positive effects (negative values of the estimates represent positive effects) which led to the high positive predicted estimates for the HAQ. Especially extreme values predicted for the HAQ were those for studies by Cohen, Di-Poi, Nikas and one cohort of the Wick study which had extreme values for the DAS-28. Also, two studies out of those three reporting the HAQ, but not the DAS-28, had more extreme estimates, showing little or no effect (only small improvement in HAQ in the Hyrich study and no improvement in Iannone) which led to the more extreme predicted estimates for the DAS-28 and in consequence reduced the pooled effect measured by the change from baseline of the DAS-28.

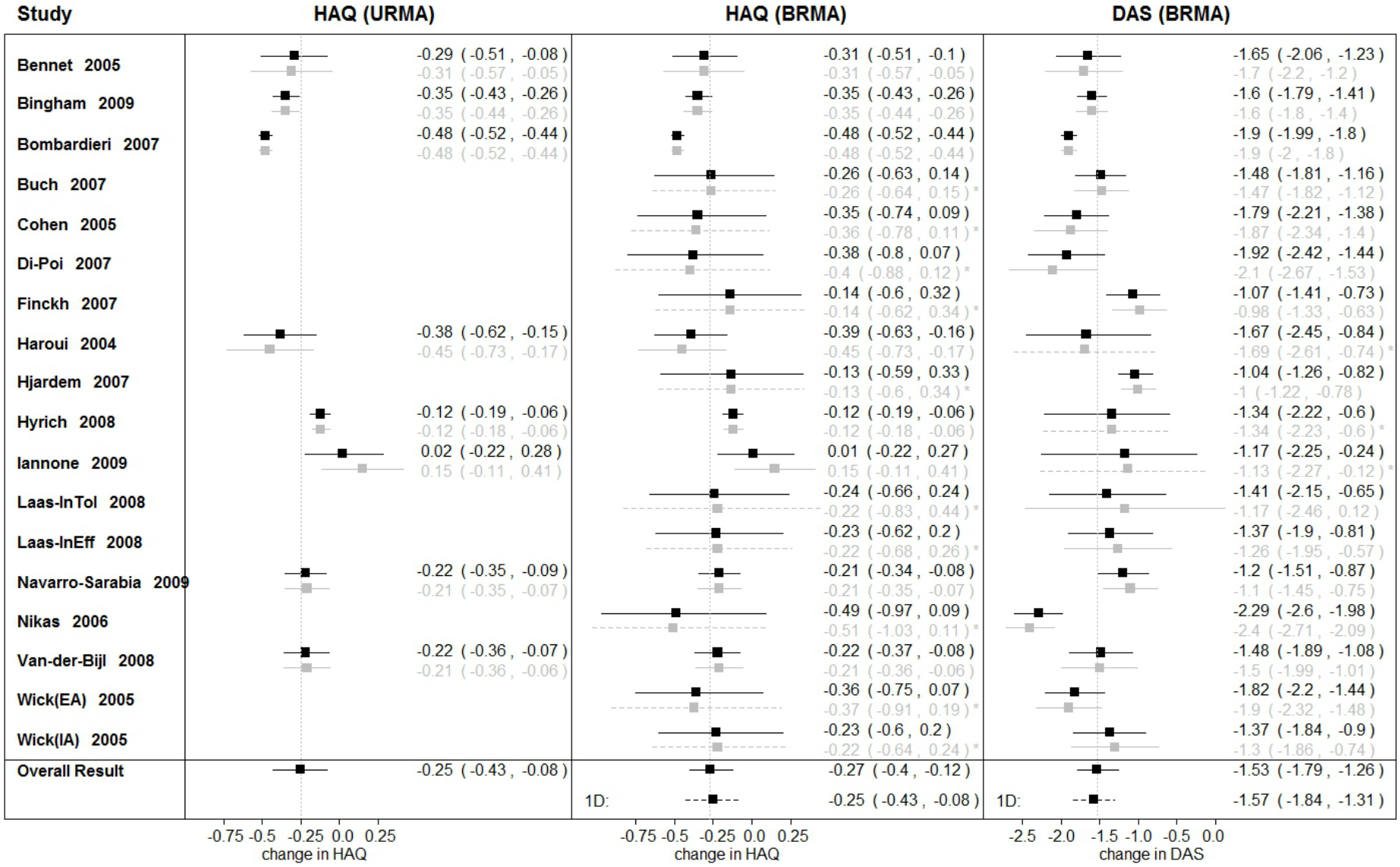


Figure 1: Forest plots for HAQ: from URMA (left) and from BRMA of HAQ and DAS-28 (middle) and for DAS-28 also from BRMA (right). Graph shows estimates from the systematic review with 95% CIs (grey solid lines), predicted missing estimates from BRMA with 95% CrIs (grey dashed lines), "shrunken" estimates with 95% CrIs (black solid lines) and the pooled estimates with 95% CrIs (black solid lines for pooled effect from each of the meta-analyses and black dashed lines representing results from URMA for comparison).

2.4 Illustration of the use of informative prior distributions for combining diverse sources if evidence

We will now illustrate how external evidence can be incorporated into the meta-analytic framework in the form of informative prior distributions. To do so, we will use external IPD and external summary data (ESD) in RA to inform the within-study and between-studies correlations.

2.4.1 Using external IPD to inform within-study correlations

None of the studies included in the "RA data" reported correlations between the treatment effects on the two outcomes. These correlations cannot be obtained directly from summary data and none of the studies listed the IPD, thus external data were required to estimate the correlations.

External individual patient data (EIPD) were obtained from the British Rheumatoid Outcome Study Group (BROSG) trial which was designed to assess the benefit of aggressive disease-modifying anti-rheumatic drug treatments in patients with established RA [56]. It was a randomised trial, which recruited 466 patients with stable RA. The trial assessed clinical outcomes (i.e. HAQ, DAS-28 and ACR20) in two cohorts of patients managed using either a regime focused on (1) symptomatic control of pain and stiffness in the shared care setting, or (2) a more aggressive regime focusing on control of symptoms and joint inflammation in the hospital setting. Bujkiewicz et al. used data from 293 patients in the BROSG trial (subset of the individuals reporting all three outcomes) to obtain the within-study correlations between the treatment effects on the multiple outcomes to populate the multivariate meta-analytic models [5, 57]. They assumed that in this external study the correlation between the treatment effects on the three outcomes is expected to be of similar magnitude to the correlations within the studies included in the meta-analysis. In this data example with single-arm studies, the within-study correlations are between the estimates of change from baseline on the continuous outcomes (HAQ and DAS-28) or log odds of response for ACR20. In general case these are the correlations between the treatment effects as described in Section 2.1.1. We will use data on DAS-28 and HAQ to obtain the correlation for the bivariate case. In the previous section, we used the mean correlation obtained for these data as a fixed value and assumed that it was the same across all studies in the meta-analysis. Here we will use the parameters of the posterior correlation to construct a prior distribution for the analysis of RA data.

2.4.2 Using external summary data to inform between-studies correlation

In contrast to the within-study correlations, the between-studies correlations can be estimated from the study level summary data. In a Bayesian framework this could entail placing non-informative prior distributions on these correlations. However, one of the advantages of Bayesian approach we want to exploit here is the possibility of incorporating external information in the analysis. This can be achieved by using external information to construct an informative prior distribution. To illustrate this, we conducted a bivariate meta-analysis of ESD to obtain an estimate of the between-studies correlation which can be used as a prior distribution in a BRMA model. The ESD included studies of the same type of treatment as in the "RA data", but used as the first line treatment. Therefore it was acceptable to assume that the between-studies correlation estimated from this external data should be reasonably similar to inform the prior distribution for the between-studies correlation in the main analysis. Further details with the full list of studies included in the ESD are reported by Bujkiewicz et al. [5, 57].

2.4.3 Logic of the meta-analysis model and notation

In our motivating example, we aim to model the summary data of the correlated outcomes from the "RA data" using a multivariate meta-analysis in a Bayesian framework. To do so, we need to place prior distributions on the within-study correlations (which are not known in the "RA data") and on the between-studies correlations. We use the EIPD, described in Section 2.4.1, to construct

prior distributions for the within-study correlations and the ESD, described in Section 2.4.2, to construct the prior distributions for the between-studies correlations. Figure 2 illustrates this data structure and the role of each element within it. The external data are used to construct prior distributions for the within-study and between-studies correlations. The remaining parameters of the model, such as the pooled effects and the between-studies standard deviations, are given non-informative prior distributions [22]. Note that the external data set used in this example was not very large. However, in more general circumstances the relevance and rigour of the external evidence can be taken into account. For example, the variance of the prior distribution can be adjusted to construct a less informative distribution. In addition, when there are multiple external data sources, a random effects meta-analysis can be carried out. A number of authors have advocated using posterior predictive distribution from such external meta-analysis as a source of external evidence incorporated in the analysis in the form of a prior distribution [22, 58].

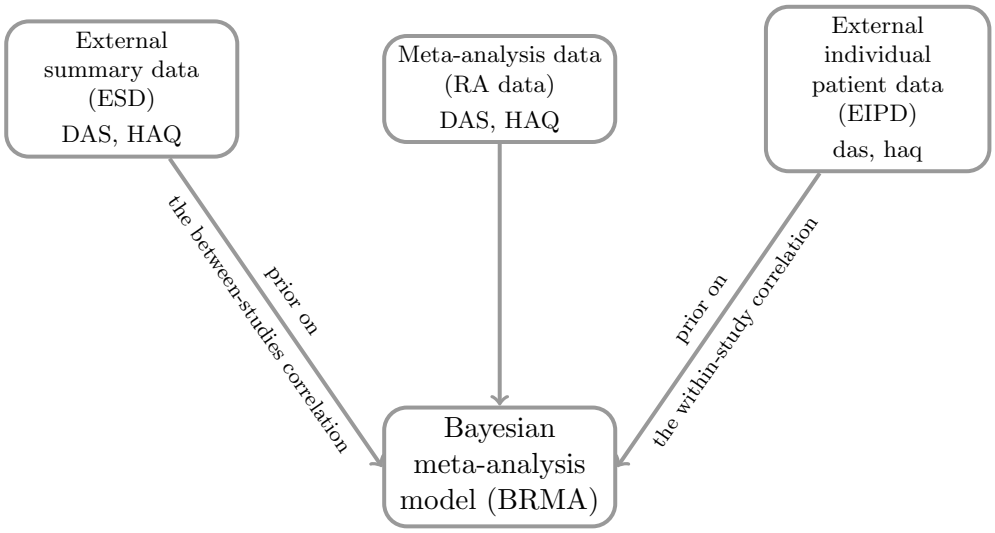


Figure 2: Sources of evidence and the role of the data sets in the BRMA model.

2.4.4 Statistical methods

EIPD on treatment effects HAQ and DAS-28 (measured as change from baseline) are analysed jointly to obtain the within-study correlation. If we have data on change from baseline in HAQ_j and DAS_j from each patient j, then we can model them jointly by assuming that they are correlated and normally distributed either using a bivariate normal distribution or in the product normal formulation:

$$\begin{aligned} HAQ_j &\sim N(\theta_H, \xi_H^2) \\ DAS_j &\sim N(\theta_{Dj}, \xi_D^2) \\ \theta_{Dj} &= \alpha_0 + \alpha_1 \left(HAQ_j - \overline{HAQ_j}\right) \end{aligned}$$

with prior distributions on the parameters, for example $\theta_H \sim N(0, 10^3)$, $\alpha_0 \sim N(0, 10^3)$, $\alpha_1 \sim N(0, 10^3)$, $\xi_D \sim Unif(0, 10)$ and $\xi_H \sim Unif(0, 10)$. Then the correlation is $\rho_{IPD} = \alpha_1 var_H / var_D$, where $var_H = \xi_H^2$ and $var_D = \alpha_1^2 \xi_H^2 + \xi_D^2$. The correlation can be transformed using Fisher transformation to construct a prior distribution on the normal scale [59, 5]. This is achieved by calculating the Fisher transformation $FT = \frac{1}{2} \log\left(\frac{1+\rho_{IPD}}{1-\rho_{IPD}}\right)$. The Fisher transformed correlation is then assumed to be normally distributed; $\rho^{FT} \sim N(FT, FT_{SD}^2)$ and then the prior distribution for the between-studies correlation is obtained by back-transforming ρ^{FT} into $\rho_w = [\exp(2\rho^{FT}) - 1]/[\exp(2\rho^{FT}) + 1]$. WinBUGS code corresponding to this analysis is included in the first section of Appendix A.3.

External summary data are analysed using BRMA (for this example we used BRMA PNF) with non-informative prior distributions. The resulting between studies correlation is transformed using Fisher transformation. The posterior mean and standard deviation of the Fisher transformed correlation are then used in the main BRMA model of RA data to construct prior distribution on the Fisher transformed correlation, which is then back-transformed to give implied prior distribution for the between-studies correlation.

WinBUGS coding for these analyses is included in Appendix A.3.

2.4.5 Results

The posterior correlation between the treatment effects on DAS-28 and HAQ obtained from the EIPD (used as a prior distribution for the within-study correlation in each study, as discussed in Section 2.4.1) was relatively weak, with mean $\rho_w^{das,haq} = 0.24$ (95% CrI: 0.13, 0.35) and the Fisher transformed posterior mean (SD), used to construct the prior distribution, was 0.245 (0.06). The posterior between-studies correlation between treatment effects on DAS-28 and HAQ obtained from BRMA of the ESD (used here as a prior distribution for the between-studies correlation) had a higher mean; $\rho^{DH} = 0.91$ (95% CrI: 0.50, 1.00).

Results of applying BRMA PNF with informative prior distributions constructed based on these posterior estimates for the within-study and between-studies correlations are shown in Table 3. They are presented along with the results obtained in the previous section, using non-informative prior distribution for the between-studies correlation for comparison.

Table 3: Results of the univariate meta-analyses of treatment effects on HAQ and DAS-28 separately and bivariate meta-analysis of the treatment effects on HAQ and DAS-28. Negative values (reduction from baseline for HAQ or DAS-28) represent positive effect of the treatment.

	posterior mean treatment effect (SD) [95% CrI]				
	univariate analyses		bivariate analyses		
	HAQ	DAS-28	HAQ & DAS-28		
			BRMA	BRMA (PNF)	BRMA (PNF/IP)
HAQ	-0.25 (0.09) [-0.43,-0.07]		-0.27 (0.08) [-0.42,-0.11]	-0.27 (0.07) [-0.39,-0.13]	-0.29 (0.04) [-0.36,-0.20]
DAS		-1.57 (0.13) [-1.84,-1.31]	-1.53 (0.13) [-1.78,-1.26]	-1.53 (0.13) [-1.78,-1.26]	-1.50 (0.13) [-1.75,-1.23]
τ_H	0.21 (0.09) [0.10,0.44]		0.21 (0.08) [0.11,0.41]	0.22 (0.08) [0.11,0.41]	0.22 (0.07) [0.12,0.39]
τ_D		0.44 (0.11) [0.25,0.67]	0.44 (0.11) [0.27,0.71]	0.44 (0.11) [0.27,0.71]	0.47 (0.12) [0.28,0.74]
ρ^{DH}			0.58 (0.42) [-0.51,0.99]	0.59 (0.43) [-0.51,0.99]	0.97 (0.05) [0.84,1.00]

PNF/IP – product normal formulation with informative prior distributions

Use of an informative prior distribution on the between-studies correlation resulted in estimation of this correlation with much higher precision; 0.97 (95% CrI: 0.84, 1.0). This also led to the estimate for HAQ obtained with uncertainty reduced by 56%, in terms of the width of the credible interval, compared to the estimate obtained from the univariate analysis of HAQ alone. This gave an additional increase in precision of 38% reduction in the width of the CrI compared to the estimate from BRMA PNF with non-informative prior distribution for the between-studies correlation.

3 Surrogate endpoint evaluation with bivariate meta-analysis

3.1 Surrogate endpoints, their importance and validity

3.1.1 Importance

Surrogate endpoints are very important in the drug development process, at both the trial design and the evaluation stage. They are particularly useful when they can provide early or more accurate measurement of the treatment effect, in settings where a long follow up time is required before an accurate measurement of the final clinical outcome can be made [7]. This is often the case in cancer where overall survival (OS) is of primary interest whilst other outcomes such as progression-free survival (PFS) potentially can be used to measure the effect of a treatment earlier. In particular for highly successful new generation of targeted therapies, the number of events on the final outcome (deaths in the case of OS) may be small resulting in large uncertainty around the treatment effect on this outcome, whilst the number of progressions may be sufficiently high to estimate the effect of the treatment on PFS with reasonable precision. Alternatively, PFS may be of primary interest and tumour response (TR) is then investigated as a short term surrogate endpoint to PFS. In other disease areas, other outcomes have been investigated as potential surrogate endpoints to a final clinical outcome, for example relapse rate as a surrogate endpoint to disease progression in relapsing remitting multiple sclerosis (RRMS) [60]. As discussed in the introduction Section 1.3.1, surrogate endpoints play an increasingly important role in regulatory approval of new health technologies at the licensing stage and therefore are of interest in HTA decision-making.

3.1.2 Validity

Before they can be used in evaluation of new health technologies, candidate surrogate endpoints have to be assessed for their predictive value of the treatment effect on the final clinical outcome. The International Conference on Harmonisation guidelines for the conduct of clinical trials for the registration of drugs (ICH-9) [61] stated that "in practice, the strength of the evidence for surrogacy depends upon (i) the biological plausibility of the relationship, (ii) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (iii) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome". These and other similar criteria for surrogate endpoint validation, by Bucher et al in JAMA Users' Guides to the Medical Literature [62] and Lassere et al for the Outcomes Measures in Rheumatology Clinical Trials (OMERACT) Working Group [63], are summarised by Taylor and Elston [30] and more recently Ciani et al [31].

Based on these guidelines, Taylor and Elston proposed a three-step framework for evaluation of a surrogate endpoint, designed for health care policy makers, each depending on availability of evidence supporting the validity of a surrogate endpoint [30]:

- **level 3:** evidence based on biological plausibility alone
- **level 2:** evidence of a strong association between the surrogate and the final endpoint at the individual patient level
- **level 1:** evidence showing that health technologies improving the surrogate also improve the final clinical outcome across many randomized controlled trials.

There has been a substantial degree of criticism of relying solely on level 2 of evidence when evaluating surrogate endpoints, in particular when individual level association is evaluated based on data from a single trial. Fleming and DeMets famously said that "A correlate does not a surrogate make" [64]. The authors discuss surrogacy criteria developed by Prentice [65], which require that the surrogate must be correlated with the true clinical outcome and (in the fourth criterion sometimes referred to as the surrogacy criteria) fully mediate the treatment effect on the final clinical outcome. Fleming and DeMets describe a number of scenarios where, depending on

the mechanism of action of a treatment, the latter criteria may not be satisfied. Limitations of Prentice's criteria for surrogacy have been discussed by a number of other authors [66, 67, 68], who emphasised that satisfying those criteria does not guarantee a causal relationship between the treatment effects on the surrogate and the final outcome.

The issue of the causal effect mostly affects surrogate evaluation when such analysis is based on a single study. A meta-analytic approach, based on data from a number of trials, follows the causal association paradigm which, as discussed by Joffe and Greene, is based on establishing the association between the treatment effects on the candidate surrogate endpoint and on the final outcome (rather than modelling the effect of surrogate on the final outcome) [66]. The authors point out that this approach is more useful for evaluation of surrogate endpoints as it is free from the restrictions of the causal effect paradigm. The meta-analytic approach to the causal association study involves associations between quantities that derive directly from randomisation and as such are average causal effects. Meta-analytic approach is based on data from a number of studies or subgroups and is likely to include heterogeneous treatment contrasts which is an obvious advantage over evaluation based on a single study. A single trial validation cannot guarantee that an association between effects confirmed based on individual data under one treatment will hold in other interventions.

3.1.3 Meta-analytic approach to surrogate endpoint evaluation

A range of meta-analytic techniques have been proposed to evaluate surrogate outcomes. Putative surrogate endpoints are validated, using these techniques, by estimating the pattern of association between the treatment effects on surrogate and final endpoints across trials, in different populations and/or investigating different treatments [8, 9, 10, 11, 12, 13]. Multivariate meta-analysis methods are used to obtain average treatment effects on multiple endpoints while taking account of the correlations between them [2, 4, 6, 5] and, as such, are suitable tools for modelling surrogate endpoints [12, 13]. Such methods are superior to, for example, meta-regression models as they take into account the uncertainty around the treatment effects – not only on the final outcome but also on the surrogate endpoint (which would be treated as a fixed covariate in a meta-regression). Bivariate meta-analysis of treatment effects on a surrogate and a final outcome allows for both the validation of a surrogate endpoint and for making predictions of an unobserved treatment effect on the final clinical outcome from observed treatment effects on a surrogate endpoint. In this TSD we focus on the application of bivariate meta-analytic methods to surrogate evaluation at the study level. IPD models evaluating surrogacy at both individual level and study level are discussed briefly in Section 3.8.

3.1.4 Data requirements for surrogate endpoint validation

To use bivariate meta-analytic techniques to evaluate surrogate endpoints, data from all relevant studies on the treatment effect on both outcomes (the surrogate endpoint and the final clinical outcome) and corresponding standard errors will be required for the analysis. When more than one surrogate endpoint is investigated one at a time, multiple data sets (as many as the number of candidate surrogate endpoints) will be requited that include the same information: the treatment effects on the candidate surrogate endpoint and treatment effects on the final clinical outcome along with the corresponding standard errors. Example of such data set is included in the Appendix B.1. Studies reporting only treatment effect on one of the outcomes typically would not be included in this analysis, although they could be included as the unreported effects in these studies are predicted by the models and these predicted effects can contribute to the estimation of the between-studies correlation. More details on this issues are included in Section 2.1.3 discussing missing data and also in the discussion of the results of the example in rheumatoid arthritis in Section 2.3.2. Multiple surrogate endpoints may also be investigated jointly by the use of multivariate meta-analysis, as discussed in Section 4.5. In this case, data containing multiple treatment effects (on all surrogate

endpoints and on the final outcome) and corresponding standard errors will be needed.

Data in the same format is required when carrying out cross-validation procedure described in detail in Section 3.6.1.

When predicting treatment effect on the final clinical outcome from the treatment effect on a surrogate endpoint in a new study, as in Section 3.6.2, unlike in the cross-validation procedure the treatment effect on the final outcome will not be known and should be included as missing effect. When making prediction, we want to account for the uncertainty around the predicted effect and, therefore, the interval around this effect will also need to be estimated. To do so, we will need to make additional assumptions about the variances of the treatment effect at the population level, therefore it will be more convenient to have data on the treatment effects with the corresponding population variances along with the numbers of patients in each study. This is discussed in detail in Section 3.6.2 and an example of such data set is included in the Appendix B.5, which corresponds to the example described in Section 3.7.4.

Relevant studies are typically identified through a systematic review and the scope for this literature review may be much larger than, for example, for the technology under evaluation by a HTA body. For a strong surrogate endpoint (a good predictor of clinical benefit) the surrogate relationship will not depend on a treatment or a subpopulation and data from all trials in all subgroups of patients in a given disease area would be used. Often, however, subsets of interventions or population may only be included. This may be the case when the differences in mechanism of action between treatment types or patients subgroups affect the estimates of the treatment effect on the surrogate and final outcomes in different ways, thus affecting the estimates of the surrogate relationship. Data from all studies including all treatments may be used together in more complex models which are discussed in Section 3.8.

3.2 Standard surrogacy model by Daniels and Hughes

In a model proposed by Daniels and Hughes [8], the estimates of the treatment effects measured by the surrogate endpoint Y_{1i} and the final outcome Y_{2i} are assumed to come from a bivariate normal distribution and they estimate underlying true effects on the surrogate and final outcomes δ_{1i} and δ_{2i}, respectively, from each study i with corresponding within-study standard deviations σ_{1i} and σ_{2i} and the within-study correlation ρ_{wi}^{12}:

$$\begin{gathered}\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi}^{12} \\ \sigma_{1i}\sigma_{2i}\rho_{wi}^{12} & \sigma_{2i}^2 \end{pmatrix}\right) \\ \delta_{2i} \mid \delta_{1i} \sim N(\lambda_0 + \lambda_1\delta_{1i}, \psi_2^2),\end{gathered} \tag{7}$$

where the true effects δ_{1i} measured by the surrogate endpoint are assumed to be study-specific fixed effects (which here means that they are independent effects in each study) and to have a linear relationship with the true final outcomes δ_{2i}. Prior distributions are given to all parameters, for example on the fixed effects $\delta_{1i} \sim N(0, 1000)$, and the regression parameters: $\lambda_0 \sim N(0, 1000)$, $\lambda_1 \sim N(0, 1000)$, $\psi_2 \sim Unif(0, 2)$.

Daniels and Hughes described a strong surrogate relationship as one with $\lambda_0 = 0$ (to ensure that no treatment effect on the surrogate endpoint implies no treatment effect on the final clinical outcome) and $\lambda_1 \neq 0$ (establishing a relationship between treatment effects on the surrogate and final clinical outcomes). The conditional variance ψ_2^2 measures the strength of the association and for a perfect surrogate relationship it will be equal to 0. We will refer to these three criteria describing the strength of the surrogate relationship as "surrogacy criteria." The use of these criteria is further discussed in Section 3.8.

3.3 BRMA in product normal formulation

The BRMA model in product normal formulation, described in Section 2.2, can also be used for surrogate endpoint evaluation, in a similar way as the above model adopted from the paper by

Daniels and Hughes. The main difference is in the assumption of random effects (exchangeability) for the true treatment effects on the surrogate endpoints $\delta_{1i} \sim N(\eta_1, \psi_1^2)$ in the BRMA model, whilst Daniels and Hughes assumed fixed affects on the surrogate endpoint. When this assumption of exchangeability is reasonable, this model can lead to more precise predictions due to increased borrowing of information across studies (for both outcomes). However, as discussed by Daniels and Hughes, the choice of the distribution for the random effects may be difficult and depend on the ordering of the active and control treatments. It also can be complex when, for example, the distribution of the effects on the surrogate endpoint is bimodal (for example when two classes of treatments of very different effectiveness, both against the same control, are investigated). As shown in a simulation study by Bujkiewicz et al. [12], when the exchangeability assumption does not hold, the predictions obtained from BRMA PNF model may be biased.

The surrogacy criteria for this model are the same as in the model by Daniels and Hughes in the previous section.

3.4 BRMA in the standard form

BRMA in the standard form, described in Section 2.1, can also be used to evaluate surrogate endpoints. In this case, the measure of the strength of the surrogacy pattern is the between-studies correlation ρ. When the surrogate relationship is perfect the correlation is $\rho = \pm 1$. Similarly as for BRMA PNF discussed above, this model assumes random effect on the surrogate endpoint. Therefore analysts should exercise caution when using this model.

3.5 Relationships between the models and with other models in the literature

The balance between the advantages and potential limitations of each method needs to be considered when evaluating surrogate endpoints and a sensitivity analysis is advised to investigate models' fit. This can be achieved by comparing models using deviance information criteria (DIC) [69] and by cross validation procedure, described in more detail in Section 3.6.1.

The parameters of models discussed in Sections 3.2–3.4 and the relationships between them are summarised in Table 4. The parameters of the between-studies model of BRMA, the means d_j, heterogeneity parameters τ_j on both outcomes and the between-studies correlation ρ between the true effects on the two outcomes relate directly to the parameters of BRMA PNF model and the corresponding surrogacy criteria. The slope and the conditional variance are

$$\lambda_1 = \rho \frac{\tau_2}{\tau_1} \tag{8}$$

and

$$\psi_2^2 = \tau_2^2 - \lambda_1^2 \tau_1^2 = \tau_2^2 (1 - \rho^2). \tag{9}$$

Since the slope is also $\lambda_1 = (d_2 - \lambda_0)/d_1$ (which can be easily seen by drawing a picture with a regression line, two mean values d_1 and d_2 on each axis and a non-zero intercept), we can express the intercept in terms of the parameters of BRMA:

$$\lambda_0 = d_2 - d_1 \rho \tau_2 / \tau_1. \tag{10}$$

In the BRMA PNF model (similarly as in the model by Daniels and Hughes)), the surrogate relationship was perfect when the conditional variance was zero: $\psi_2^2 = 0$ (Daniels and Hughes [8], Bujkiewicz et al [12]). From (9), $\rho^2 = 1 - \frac{\psi_2^2}{\tau_2^2}$, which implies that if $\psi_2^2 = 0$ then the correlation $\rho = \pm 1$. Therefore the criteria for perfect surrogacy for the two BRMA models are equivalent. The criteria have the same meaning in terms of surrogacy for the model by Daniels and Hughes but the corresponding parameters do not relate directly to those of the BRMA model. This is because of the fixed effects assumption in the model by Daniels and Hughes, and hence no heterogeneity parameters or means.

Table 4: Parameters in the surrogacy models and the relationships between them

D&H	BRMA PNF	BRMA
λ_0	λ_0	ρ
λ_1	λ_1	τ_1
ψ_2^2	ψ_1^2	τ_2
	ψ_2^2	d_1
	η_1	d_2
	$\eta_1 = d_1$	$d_1 = \eta_1$
	$\lambda_1 = \rho\tau_2/\tau_1$	$^*d_2 = \lambda_0 + \lambda_1\eta_1$
	$\lambda_0 = d_2 - d_1\rho\tau_2/\tau_1$	$\rho = \lambda_1\tau_1/\tau_2$
	$\psi_1 = \tau_1$	$\tau_1 = \psi_1$
	$\psi_2^2 = \tau_2^2\left(1-\rho^2\right)$	$\tau_2^2 = \psi_2^2 + \lambda_1^2\psi_1^2$
		$\rho^2 = \lambda_1^2\psi_1^2/\left(\psi_2^2 + \lambda_1^2\psi_1^2\right)$

D&H – model by Daniels and Hughes, BRMA – bivariate random effects meta-analysis,
BRMA PNF – BRMA in product normal formulation
λ_0 – intercept, λ_1 – slope, ψ_2^2 – conditional variance, ρ – between-studies correlation,
η_1, d_1, d_2 – mean values, τ_1, τ_2 – between-studies heterogeneity parameters
* if the effects are centred $d_2 = \lambda_0$ (see Section 2.2).

The measure of surrogacy can be also expressed as ρ^2 ($\rho^2 = 1$ for perfect surrogacy), which some authors refer to as the study-level adjusted R-squared (Burzykowski et al [10] and Renfro et al [11]).

3.6 Validation and making predictions

3.6.1 Cross-validation procedure

To evaluate surrogate endpoints, in the first instance the surrogacy criteria are obtained from the meta-analysis of all of the data: the regression coefficient for models by Daniels and Hughes and BRMA PNF or the between-studies correlation (or adjusted R-squared) when using BRMA in the standard form. Following this, predicted values (and corresponding predicted intervals) of the treatment effects on the final outcome are compared to the observed estimates in take-one-out cross-validation procedure. In one study at a time, the estimate of the treatment effect on the final outcome Y_{2i} is removed (and treated as missing at random) and then this treatment effect is predicted from the treatment effect on the surrogate endpoint, conditional of the data on both outcomes from all the remaining studies in the meta-analysis. The mean predicted effect is equal to the mean predicted true effect $\hat{\delta}_{2i}$, predicted by MCMC simulation, and the variance of the predicted effect is equal to $\sigma_{2i}^2 + var(\hat{\delta}_{2i}|Y_{1i}, \sigma_{1i}, Y_{1(-i)}, Y_{2(-i)})$, where $Y_{1(-i)}$ and $Y_{2(-i)}$ denote the data from the remaining studies without the validation study i [8]. For a valid surrogate endpoint, the predicted interval, constructed using this variance, should contain the observed estimate, at least in 95% of the studies if using 95% predicted intervals.

As discussed in Section 2.1.3, when missing data are present in the bivariate meta-analysis, in the Bayesian framework these missing treatment effects are directly predicted from the meta-analytic model using MCMC simulation. To implement this, prior distributions need to be placed on the missing data. The model used to synthesise the treatment effects already takes care of this for missing values of treatment effect Y_{ji} on the missing outcome j in study i. However, the corresponding missing standard errors σ_{ji} still need to be given prior distributions. For the cross validation, only one treatment effect is missing; on the final outcome in the validation study, so there is also only one missing standard error corresponding this this treatment effect. A prior distribution can be placed on this standard error, such as, for example, $\sigma_{ji} \sim Unif(0.0001, 15)$.

In Section 2.1.3, discussing use of bivariate meta-analysis for estimation of pooled effects, we

also suggested a model allowing for exchangeability of the population variances as a method for estimating the missing standard errors. In the cross-validation procedure, however, this would not be a suitable approach as this may lead to either overestimating or underestimating the predicted intervals. An independent prior distribution in the "missing" standard error is sufficient, as we do not need to estimate this parameter. When constructing the predicted interval (as in the formula in the first paragraph of this section), we use the actual standard error (which we do know, only treat as unreported for the purpose of the cross-validation). However, in a situation when we want to predict the treatment effect on the final outcome in a new study where we only have data on the treatment effect on the surrogate endpoint, but not on the final outcome, we do need to estimate this parameter to account for the uncertainty around the predicted estimate appropriately. We discuss this issue in the next section.

3.6.2 Predicting treatment effect on the final outcome from the effect measured on a surrogate endpoint in a new study

When our validation procedure was successful and we concluded that an endpoint can be used as a surrogate endpoint to the final clinical outcome, then we can predict the treatment effect on the final outcome in a new study reporting only the treatment effect on the surrogate endpoint. To do so, we use a bivariate meta-analytic model. It can be the model by Daniels and Hughes, described in Section 3.2, or a BRMA model as discussed in Sections 3.3 and 3.4.

The main difference between the model by Daniels and Hughes and BRMA (in either formulation) is that the former assumes fixed effects (independent prior distributions) and the latter random effects for the treatment effects on the surrogate endpoint. When the normality assumption of the random effects, i.e. the true effects δ_{1j} on the surrogate endpoint is not satisfied (for example plotting a histogram of the estimated effects of the surrogate endpoints shows clear deviation from normality such as a bimodal distribution for these effects) then the model by Daniels and Hughes will be more appropriate. However, if the normality assumption of these effects is reasonable, using BRMA may give more precise predictions, as the assumption of the random effects gives additional borrowing of information across studies.

When predicting the treatment effect on the final outcome in a new study reporting only the treatment effect on the surrogate endpoint, we use all data from the previous relevant studies (see Section 3.1.4) reporting the treatment effects on both outcomes with corresponding standard errors, together with the data on the treatment effects on the surrogate endpoint and the corresponding standard error for the new study. As in BRMA with missing data, we code the unreported treatment effect on the final outcome in the new study and the corresponding standard error as missing. However, as in BRMA with missing data, we will want to assume exchangeability of the population variances to be able to estimate the interval for the predicted effect on the final outcome. To do this, we need data on the population variances (rather than the standard errors) corresponding to the treatment effects on both outcomes and the numbers of participants. See Section 2.1.3 for details. By assuming that the two treatment effects, the effect on the surrogate endpoint and the unreported treatment effect on the final outcome, and their corresponding population level variances follow the same model as the data from all other studies reporting the treatment effects on both outcomes, the unreported effect and variance corresponding to the final outcome in the new study are predicted from the model (conditional on all of the data) by the MCMC simulation. The predicted effect on the final outcome in the new study will be obtained as the mean and the credible interval of Y_{2k} for the new study k. This is illustrated in Section 3.7.4.

There may be more than one new study reporting the treatment effects on the surrogate endpint but not on the final outcome. We can then predict the treatment effects on the final outcome for all of those new studies and either treat them as separate predicted estimates that can be used in a decision-making framework individually or jointly by obtaining an average predicted effect by using, for example, a standard (univariate) meta-analysis of the predicted effects. When the new studies are investigating different treatments, individual predicted effects for each study are likely

to be of interest.

3.7 Example: surrogacy validation in relapsing remitting multiple sclerosis

To illustrate the use of the modeling techniques discussed in sections 3.2-3.6, we applied them to an example in RRMS. Multiple sclerosis is an inflammatory disease of the brain and spinal cord, and RRMS is the most common type of the disease. During the course of the disease, patients experience a series of periods of exacerbations (relapses) and remission. A large proportion of patients eventually progresses to secondary progressive disease. Treatments in RRMS aim to reduce the relapses and delay disease progression, therefore the disability progression can be considered the final clinical outcome. The annualised relapse rate is another endpoint typically used in clinical trials in MS and potential surrogate endpoint to the disability progression. This example is based on work by Sormani et al. [60] who showed that in studies investigating treatment effect in patients with RRMS, the treatment effect on relapse rate can potentially be used as a surrogate endpoint to the treatment effect on the disability progression. We use data from this study as an illustrative example of surrogate endpoint evaluation process and refer to these data as the "RRMS data" in the remainder of this section.

3.7.1 Data

The annualized relapse rate ratio, the ratio between the relapse rate in the experimental and the control arms, was used as the summary estimate of the treatment effect on relapses (the surrogate endpoint measuring the treatment effect). The disability progression risk ratio, the ratio between the proportion of patients with a disability progression in the experimental and the control arms at year 2 (or at year 3 for trials of longer follow up time which do not report the outcome at year 2), was used as the summary estimate of the treatment effect on disability progression, which was the final clinical outcome. We model these treatment effects on the log scale. Details of the specific treatment regimens are included in Table 5. Figure 3 shows data on both outcomes presented graphically in the form of a forest plot, revealing similar heterogeneity patterns between the effects for both outcomes across studies, and hence implying a strong correlation between the effects on these outcomes. The studies are grouped as placebo-controlled and active-treatment-controlled.

Every study in this example included both endpoints as only those studies contribute directly to estimation of the surrogate association (through the correlation in BRMA or equivalently the slope and conditional variance in the model by Daniels and Hughes). However, additional studies that include only one endpoint could be included in such analyses. In particular, when using a BRMA model assuming random effects for the treatment effects on the surrogate endpoint, the data from studies reporting only the treatment effect on surrogate endpoint may contribute to the individual "shrunken" estimates in each study and ultimately to the estimation of the surrogate relationship parameters.

Within-study correlations were not available for any of the studies in this example. We placed prior distributions on all of these correlations assuming that they are all positive between 0 and 1.

3.7.2 Implementation

All three models, described in Sections 3.2–3.4, were applied to the RRMS data. WinBUGS codes including data for all three models are included in appendices B.1–B.3. Appendix B.4 includes R code for cross-validation procedure (along with the WinBUGS code for each model and data in a suitable format). Appendix B.5 includes WinBUGS code for making prediction of the treatment effect on the final outcome in a new study.

Table 5: Studies in RRMS data reporting the annualised relapse rate ratio and the disability progression risk ratio.

Study	contrast	N	follow-up (months)	annualised relapse rate ratio	disability progression risk ratio
Paty (1) 1993	IFNbeta-1b 1.6 MIU vs PBO	248	24	0.92 (0.82, 1.03)	1.00 (0.67, 1.49)
Paty (2) 1993	IFNbeta-1b 8 MIU vs PBO	247	24	0.66 (0.58, 0.75)	0.71 (0.46, 1.12)
Miligan 1994	MMPS vs PBO	26	24	0.81 (0.50, 1.30)	1.14 (0.26, 5.03)
Johnson 1995	GA vs PBO	251	24	0.71 (0.61, 0.82)	0.88 (0.57, 1.35)
Jacobs 1996	IFNbeta-1a 6 MIU vs PBO	172	24	0.68 (0.57, 0.81)	0.63 (0.38, 1.04)
Fazekas 1997	IVIg vs PBO	150	24	0.41 (0.34, 0.49)	0.70 (0.36, 1.35)
Millefiorini 1997	Mitoxantrone vs PBO	51	24	0.34 (0.24, 0.47)	0.19 (0.05, 0.78)
Achiron 1998	IVIg vs PBO	40	24	0.37 (0.27, 0.52)	0.82 (0.19, 3.50)
Li (1) 1998	IFNbeta1a 22 μg vs PBO	376	24	0.71 (0.64, 0.78)	0.81 (0.61, 1.08)
Li (2) 1998	IFNbeta1a 44 μg vs PBO	371	24	0.68 (0.62, 0.75)	0.73 (0.54, 0.99)
Baumhackl 2005	Hydrolytic enzymes vs PBO	306	24	0.85 (0.74, 0.97)	1.08 (0.74, 1.57)
Polman 2006	NAT vs PBO	942	24	0.32 (0.29, 0.36)	0.59 (0.46, 0.75)
Comi (1) 2009	Cladribine 3.5 mg/kg vs PBO	870	24	0.42 (0.36, 0.49)	0.69 (0.52, 0.93)
Comi (2) 2009	Cladribine 5.25 mg/kg vs PBO	893	24	0.45 (0.39 , 0.52)	0.73 (0.55 , 0.97)
Sorensen 2009	IFNbeta-1a and oral MPS vs IFNbeta-1a and PBO	130	24	0.37 (0.27, 0.50)	0.64 (0.32, 1.28)
Clanet 2002	IFNbeta-1a 60 μg vs 30μg	802	36	1.05 (0.99, 1.12)	1.00 (0.84, 1.20)
Durelli 2002	IFNbeta1b vs IFNbeta1a	188	24	0.71 (0.59, 0.86)	0.43 (0.24, 0.78)
Rudick 2006	NAT + IFNbeta-1a vs IFNbeta-1a	1171	24	0.45 (0.41, 0.49)	0.79 (0.65, 0.96)
Coles (1) 2008	ALE 12 mg vs IFNbeta-1a	223	36	0.31 (0.24, 0.40)	0.35 (0.16, 0.73)
Coles (2) 2008	ALE 24 mg vs IFNbeta-1a	221	36	0.22 (0.16, 0.30)	0.38 (0.19, 0.76)
Mikol 2008	IFNbeta vs GA	764	24	1.03 (0.90, 1.17)	1.34 (0.88, 2.06)
Havrdova (1) 2009	IFNbeta-1a 30 μg + AZA 50 mg vs IFNbeta-1a 30 μg	118	24	0.87 (0.73, 1.04)	1.23 (0.58, 2.62)
Havrdova (2) 2009	IFNbeta-1a 30 μg IM + AZA 50 mg + prednisone 10 mg vs IFNbeta-1a 30 μg	123	24	0.70 (0.58, 0.85)	1.04 (0.48, 2.27)
O'Connor (1) 2009	IFNbeta-1b 250 μg vs GA	1345	24	1.06 (0.97, 1.16)	1.05 (0.84, 1.31)
O'Connor (2) 2009	IFNbeta-1b 500 μg vs GA	1347	24	0.97 (0.88, 1.06)	1.10 (0.88, 1.37)

AZA = azathioprine; GA = glatiramer acetate; IFNβ = interferon-β; IVIg = IV immunoglobulin;
MPS = Methylprednisolone; PBO = placebo; N – number of patients.

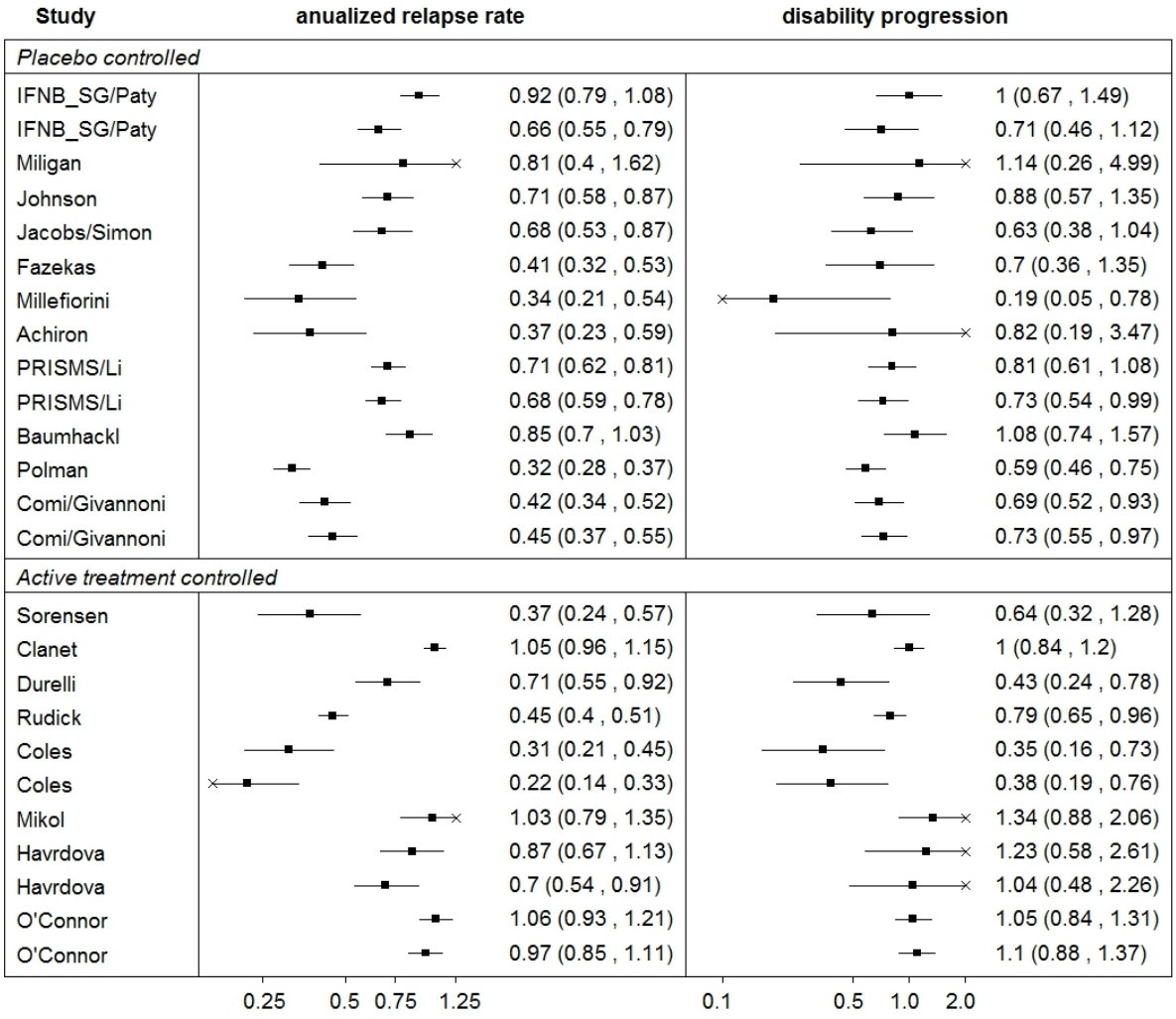


Figure 3: Summary of the RRMS data. The point estimates and corresponding 95% confidence intervals, presented graphically and numerically, represent the annualised relapse rate ratio (left) and the disability progression risk ratio (right).

3.7.3 Results of surrogate endpoint validation

To discuss the results of applying the models to RRMS data, in the first instance we present the summary estimates of the parameters on both outcomes, which include the pooled effects on both outcomes; the log annualized relapse rate ratio d_1 and the log disability progression risk ratio d_2, the between-studies correlations between them ρ, the heterogeneity parameters τ_1 and τ_2 corresponding to the two outcomes, the intercept λ_0, slope λ_1 and the conditional variance ψ_2^2 as the parameters describing the linear relationship between the effects on the two outcomes (for BRMA these parameters were derived from other parameters as discussed in section 3.4), and the between-studies correlation squared ρ^2 (or adjusted R-squared). The estimates of these parameters, obtained from the three methods, are listed in Table 6. Due to the heterogeneity of the control arm between the studies (and the fact that an intervention which is a control arm in one study may be an experimental arm in the other) there is limited clinical interpretation of the pooled effects. We present them for completeness, but the interest here lies in the heterogeneity patterns and the parameters describing the shape and strength of the surrogate relationship. The results are based on 50,000 iterations after a burn-in of 50,000.

Both forms of BRMA allowed for the estimation of the pooled effect of both outcomes, in

contrast to the model by Daniels and Hughes (however the estimation of the pooled effect on the disability progression could be made by centering the effect of the surrogate endpoint, resulting in the mean effect on disability progression being equal to the intercept [12]).

For the model by Daniels and Hughes, the intercept was close to zero with the 95% CrI including zero. The slope was positive and the conditional variance relatively small. These three parameters indicate a good surrogacy pattern between the treatment effects on the two outcomes, the annualized relapse rate and the disability progression. Similar estimates were obtained from BRMA PNF model, which in addition gave high between-studies correlation and R-squared, although the CrI around these parameter estimates was a little large. Results obtained from BRMA in the standard form were similar to those obtained from BRMA PNF. This was expected as there were no missing data and the same prior distributions were used on all parameters in the two models. The slopes obtained from BRMA were reduced compared to the slope obtained from D&H model. This was likely due to the over-shrinkage caused by the assumption of random treatment effects made in the BRMA models, possibly also resulting in the underestimated between-studies correlations and R-squared statistics.

Table 6: Summary results for placebo-controlled studies for the treatment effects on the risk of disability progression and annualized relapse rate ratio

	D&H	BRMA (PNF)	BRMA
d_1	NA	-0.50 (-0.68, -0.33)	-0.50 (-0.68, -0.33)
d_2	NA	-0.22 (-0.35, -0.11)	-0.22 (-0.34, -0.11)
ρ	NA	0.90 (0.63, 1.00)	0.90 (0.62, 1.00)
τ_1	NA	0.42 (0.30, 0.58)	0.41 (0.30, 0.57)
τ_2	NA	0.21 (0.11, 0.34)	0.20 (0.11, 0.33)
λ_0	0.03 (-0.07, 0.14)	0.001 (-0.12, 0.12)	-0.003 (-0.12, 0.11)
λ_1	0.52 (0.35, 0.70)	0.44 (0.24, 0.65)	0.43 (0.24, 0.64)
ψ_2^2	0.01 (0.00, 0.03)	0.01 (0.00, 0.03)	0.01 (0.00, 0.03)
ρ^2 (R^2)	NA	0.82 (0.40, 0.99)	0.81 (0.39, 0.99)

NA – not applicable, D&H – model by Daniels & Hughes, BRMA – bivariate random-effects meta-analysis, PNF – product normal formulation

All three models were then applied to make predictions in a cross-validation procedure. The treatment effect on the final outcome (log of the disease progression risk ratio) in the 25 studies was assumed unknown (in one study at a time which in that case became a validation study) and then predicted from the treatment effect on the surrogate endpoint (log annualized relapse rate ratio), conditional on the data for treatment effects on both outcomes from the remaining 24 studies. The cross-validation was carried out by each model to make comparison.

Table 7 lists all the predictions obtained from the cross-validation procedure, using all of the models for all of the studies. For most studies, all models gave predicted effects on disability progression $\hat{Y}_{2i}$ with the 95% predicted intervals containing the corresponding observed estimates, except for one study by Durelli et al.

The discrepancies between the observed and predicted values were obtained for all studies (by taking the absolute difference between the observed estimate of the treatment effect and the predicted effect on the log risk ratio scale) and summarised in Table 8, which also summarises the degree of uncertainty around the predicted estimate compared to the uncertainty around the observed value (by calculating the ratio $w_{\hat{Y}_{2j}}/w_{Y_{2j}}$ of the length of the 95% predicted interval to the length of the 95% confidence interval of the observed estimate, shown in the last column of the table). Note that the intervals of the predicted $\hat{Y}_{2j}$ were inflated compared to those corresponding to the observed effects Y_{2j} due to the additional between-studies variability. The accuracy of predictions for the point estimate was similar across models.

Table 7: Predictions obtained from all models for all studies in the "Sormani data"

	Disability progression risk ratio, mean (95% CrI)				
	Paty (1)	Paty (2)	Miligan	Johnson	Jacobs
Observed	1.00 (0.67, 1.49)	0.71 (0.45, 1.12)	1.14 (0.26, 5.03)	0.88 (0.57, 1.35)	0.63 (0.38, 1.05)
Daniels & Hughes	0.99 (0.63, 1.55)	0.84 (0.51, 1.37)	0.93 (0.20, 4.35)	0.86 (0.53, 1.40)	0.85 (0.49, 1.47)
BRMA (PNF)	0.95 (0.60, 1.50)	0.83 (0.51, 1.37)	0.86 (0.19, 3.93)	0.85 (0.52, 1.39)	0.85 (0.49, 1.47)
BRMA	0.95 (0.60, 1.51)	0.84 (0.51, 1.37)	0.86 (0.19, 3.93)	0.85 (0.52, 1.40)	0.85 (0.49, 1.47)
	Fazekas	Millefiorini	Achiron	Li (1)	Li (2)
Observed	0.70 (0.36, 1.35)	0.19 (0.05, 0.79)	0.82 (0.19, 3.50)	0.81 (0.61, 1.08)	0.73 (0.54, 0.99)
Daniels & Hughes	0.65 (0.32, 1.32)	0.60 (0.14, 2.59)	0.62 (0.14, 2.73)	0.87 (0.61, 1.22)	0.87 (0.59, 1.22)
BRMA (PNF)	0.68 (0.34, 1.39)	0.66 (0.15, 2.85)	0.68 (0.16, 3.01)	0.86 (0.60, 1.23)	0.85 (0.59, 1.22)
BRMA	0.68 (0.34, 1.40)	0.66 (0.15, 2.85)	0.68 (0.16, 3.01)	0.86 (0.60, 1.23)	0.85 (0.60, 1.22)
	Clanet	Durelli	Baumhackl	Polman	Rudick
Observed	1.00 (0.83, 1.20)	0.43 (0.24, 0.78)	1.08 (0.74, 1.57)	0.59 (0.46, 0.75)	0.79 (0.65, 0.96)
Daniels & Hughes	1.08 (0.81, 1.44)	0.88 (0.47, 1.65)∗	0.94 (0.61, 1.46)	0.56 (0.39, 0.81)	0.66 (0.50, 0.87)
BRMA (PNF)	1.03 (0.76, 1.39)	0.86 (0.46, 1.61)∗	0.91 (0.59, 1.42)	0.61 (0.42, 0.89)	0.68 (0.51, 0.91)
BRMA	1.02 (0.76, 1.40)	0.86 (0.46, 1.62)∗	0.91 (0.59, 1.42)	0.61 (0.42, 0.89)	0.68 (0.51, 0.91)
	Coles (1)	Coles (2)	Mikol	Comi (1)	Comi (2)
Observed	0.35 (0.16, 0.74)	0.38 (0.19, 0.77)	1.34 (0.88, 2.06)	0.69 (0.52, 0.93)	0.73 (0.55, 0.97)
Daniels & Hughes	0.58 (0.26, 1.30)	0.48 (0.22, 1.04)	1.03 (0.63, 1.67)	0.65(0.44, 0.95)	0.67 (0.47, 0.97)
BRMA (PNF)	0.64 (0.29, 1.43)	0.57 (0.27, 1.23)	0.98 (0.59, 1.58)	0.68 (0.47, 1.01)	0.70 (0.48, 1.02)
BRMA	0.64 (0.29, 1.43)	0.57 (0.27, 1.23)	0.97 (0.59, 1.58)	0.68 (0.47, 1.01)	0.70 (0.48, 1.02)
	Havrdova (1)	Havrdova (2)	Sorensen	O'Connor (1)	O'Connor (2)
Observed	1.23 (0.58, 2.62)	1.04 (0.48, 2.27)	0.64 (0.32, 1.28)	1.05 (0.84, 1.31)	1.10 (0.88, 1.37)
Daniels & Hughes	0.96 (0.44, 2.11)	0.86 (0.38, 1.92)	0.62 (0.29, 1.32)	1.06 (0.78, 1.47)	1.00 (0.74, 1.36)
BRMA (PNF)	0.92 (0.42, 2.03)	0.85 (0.38, 1.91)	0.68 (0.32, 1.43)	1.02 (0.73, 1.42)	0.96 (0.70, 1.31)
BRMA	0.92 (0.42, 2.03)	0.85 (0.38, 1.91)	0.68 (0.32, 1.43)	1.02 (0.73, 1.42)	0.95 (0.70, 1.31)

Table 8: Results of the comparison of the models for predicting the treatment effect on disability progression from the treatment effect on relapse rate

	absolute discrepancy	$w_{\hat{Y}_{2j}}/w_{Y_{2j}}$
Model	median (range)	median (range)
Daniels & Hughes	0.16 (0.01, 1.15)	1.12 (1.02, 1.60)
BRMA PNF	0.15 (0.01, 1.26)	1.12 (1.02, 1.67)
BRMA	0.15 (0.00, 1.26)	1.12 (1.02, 1.67)

3.7.4 Making prediction for a new study

To demonstrate the use of the model for predicting a treatment effect in a new study, let's imagine that the second study by Paty was a new study where we only have the treatment effect on the relapse rate recorded, but not on the disability progression.

To make a prediction, we follow the approach described in Section 3.6.2. To predict the variance we make an assumption of exchangeability of population variances. In this case we will assume exchangeability of population variances in the arms, as some of the studies in the MS data were unbalanced (see end of Section 2.1.3), and we only assume exchangeability of the variances for the final outcome, because only the effect on this final outcome is missing. These variances are obtained by calculating $var_i = \sigma_{2i}^2 / \left(\frac{1}{N_{Ai}} + \frac{1}{N_{Bi}}\right)$, where N_{Ai} and N_{Bi} are the numbers of individuals in arms A and B and σ_{2i} are the SEs corresponding to the treatment effects on the final outcome in each study i. The population variances in the arms, var_i, along with the numbers of individuals in each arm, N_{Ai} and N_{Bi}, are added to the data set. The WinBUGS code (using the model by Daniels and Hughes) with the full data set is included in Appendix B.5. By assuming exchangeability of the population variances in the arms (see formulae (3)), the missing population variance var_2 for the disability progression in the "new study" $k = 2$ is predicted from the MCMC simulation and then used to calculate the predicted SEs for this outcome ($j = 2$) in the new study: $\sigma_{22} = \sqrt{var_2 \left(\frac{1}{N_{A2}} + \frac{1}{N_{B2}}\right)}$ (which is needed to populate the within-study covariance matrix for the new study). This in turn leads to predicting the unmeasured treatment effect on the final outcome Y_{22} in the new study.

Applying the model by Daniels and Hughes (all details of the implementation are included in Appendix B.5) gave the predicted effect on disability progression 0.86 (0.52, 1.36). This predicted effect is similar to the one obtained from the cross-validation procedure (Table 7), where the interval was constructed using the actual standard error of the observed estimate. In a similar way, other models (BRMA and BRMA PNF) can be used to make such prediction.

3.7.5 Discussion of the results for RRMS

Based on our results we can conclude that relapse rate is a reasonably good surrogate endpoint for disability progression as the cross-validation procedure gave good prediction for almost all of the studies (it only failed for the study by Durelli using all methods). Considering that we use 95% prediction intervals, we would expect one study (out of 25 included in the data set) be outside this range.

Note that the effect on the final outcome in the data set investigated here is measured at the same time point as the effect on the surrogate endpoint. Since the disability progression is considered a long term endpoint, when measured early it is measured with a relatively large uncertainty due to low number of events. Further research may establish whether the relapse rate is a good surrogate endpoint and in particular an early marker of longer term disability progression. Such further research could include disability progression reported later compared to relapse rate, but potentially also consider both outcomes on alternative scales such as the hazard ratio for the time to disability progression. However, as indicated earlier in this section, despite the surrogate

relationship not being 100% perfect in this setting, a successful cross-validation in a majority of studies can be considered to be sufficient to accept the validity of the surrogate endpoint.

Alternative candidate surrogate endpoints may be investigated when validation of a surrogate endpoint fails or an endpoint is deemed a weak surrogate. However, less perfect surrogate endpoints may be considered reasonable in a situation, for example, where the need for timely evaluation of a new intervention outweighs the need for further research. This may be the case when in a regulatory setting a licensing decision about a new intervention is based conditional on a surrogate marker, but the treatment is re-evaluated when data on the final clinical outcome become available. We discuss these issues further in the next section.

3.8 Discussion of surrogacy criteria, other surrogacy models and further work

As described above, the between-studies correlation equal to ± 1 indicates perfect surrogacy. In practice, it is difficult to quantify how large the correlation should be in order to consider the surrogate endpoint suitable to make the prediction. Some authors claimed that a high level of association is required to demonstrate surrogacy. For example, Lassere et al in their Biomarker-Surrogacy Evaluation Schema defined such high association by the square of the between-studies correlation (or so-called adjusted R-squared) above 0.6 [70], and the German Institute of Quality and Efficiency in Health Care (IQWiG) requires high correlation with the lower limit of the 95% confidence interval above 0.85 [71]. Other authors emphasised that the decision of whether the surrogate endpoint may be used to make the prediction of the clinical benefit should be based on the balance between the strength of the surrogate relationship and the need for the decision to be made about the effectiveness of the new treatment, for example for regulatory purposes [72]. Moreover, the strength (or weakness) of the surrogate relationship will manifest itself in the width of the predicted interval of the treatment effect on the final outcome. A smaller value of the correlation will result in a larger interval and hence increased uncertainty about the regulatory or clinical decision made based on such prediction. The implication of this is that perhaps we don't need criteria about the correlation and instead we need only look at the predictions [48]. Such predicted estimate (along with the uncertainty) of the treatment effect may be used in HTA decision making. The evaluation of the quality of predictions can be achieved through a cross-validation procedure, discussed in Section 3.6.1.

In this TSD, we focused on meta-analytic methods for summary data which, when applied to surrogate endpoint evaluation, describe level 1 of evidence in the hierarchy described in Section 3.1.2 (Taylor and Elston [30]), which is most relevant to regulatory decision making. When IPD are available, surrogate endpoints can be evaluated at both individual- and study-level (levels 2 and 1 of evidence). Methods by Buyse et al. [9] were developed to model surrogate endpoints at the arm level by extending the ideas developed by Prentice [65] to a meta-analytic framework. The authors developed a mixed model framework where two measures, the individual-level R-squared and the trial-level R-squared, were developed to validate candidate surrogate endpoint at both levels simultaneously. These methods are developed in a frequentist approach. Various extensions to this approach have been developed [7], for example for the time-to-event data by Burzykowki et al. [10] extended further to a Bayesian framework by Renfro et al. [11].

3.8.1 Further work

As discussed in Section 3.1.4, surrogate relationship may depend on the mechanism of action of treatments or treatment classes. When this is the case, surrogate relationship may be investigated in subgroups. Data included in such analysis will be limited to a certain class of treatments. This may dramatically reduce evidence base for surrogate endpoint evaluation. To overcome this limitation, new methods have recently been developed. Bujkiewicz et al, developed a bivariate network meta-analytic method for surrogate endpoint evaluation [48]. The method allows for modelling surrogate relationships in each treatment contrast individually whilst borrowing information

from other treatment contrasts by taking into account the network structure of the data. The authors proposed an extension of the method that in addition to modelling the study-level surrogate relationship (within each treatment contrast), a treatment-level surrogacy is also modelled by assuming additional similarity between the treatments. This extended method allows for predicting treatment effect on the final outcome for a new study and a new treatment. Another method has recently been developed by Papanikos at al which allows for borrowing information about surrogate relationships between treatment classes [73]. Two versions of the method are discussed by the authors, one assuming exchangeability (similarity) of the surrogate relationships across the treatment classes and a model which relaxes this assumption by allowing for partial exchangeability, i.e. the level of exchangeability is defined by a probability of similarity which is learned from the data.

4 Multivariate random effects meta-analysis (MRMA)

In this Section, we extend the bivariate meta-analytic methods to a more general case for multiple outcomes. We begin with the technical description of the general model and discuss additional complexities regarding the prior distributions for the between-studies covariances. We then discuss two alternative parameterisations of the method, in the product normal formulation. We then apply the methods to the example in rheumatoid arthritis introduced in Section 2.3. We follow this by a discussion of the benefits of the multivariate meta-analysis and its application to modelling multiple surrogate endpoints.

4.1 MRMA in the standard form

Suppose our data consists of treatment effects measured on N outcomes (Y_1, Y_2, ..., Y_N). If we assume that these effect are normally distributed (i.e. they are continuous outcomes or transformations of other outcomes) and correlated, we can model them simultaneously in the multivariate random-effect meta-analysis (MRMA) model, by extending BRMA model described in Section 2.1, as follows:

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{Ni} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \\ \vdots \\ \delta_{Ni} \end{pmatrix}, \mathbf{\Sigma_i}\right), \ \mathbf{\Sigma_i} = \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi}^{12} & \cdots & \sigma_{1i}\sigma_{Ni}\rho_{wi}^{1N} \\ \sigma_{2i}\sigma_{1i}\rho_{wi}^{12} & \sigma_{2i}^2 & \cdots & \sigma_{2i}\sigma_{Ni}\rho_{wi}^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{Ni}\sigma_{1i}\rho_{wi}^{1N} & \sigma_{Ni}\sigma_{2i}\rho_{wi}^{2N} & \cdots & \sigma_{Ni}^2 \end{pmatrix} \quad (11)$$

$$\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \\ \vdots \\ \delta_{Ni} \end{pmatrix} \sim N\left(\begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix}, \mathbf{T}\right), \ \mathbf{T} = \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho^{12} & \cdots & \tau_1\tau_N\rho^{1N} \\ \tau_2\tau_1\rho^{12} & \tau_2^2 & \cdots & \tau_2\tau_N\rho^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_N\tau_1\rho^{1N} & \tau_N\tau_2\rho^{2N} & \cdots & \tau_N^2 \end{pmatrix}. \quad (12)$$

In the above model, the treatment effects Y_{1i}, Y_{2i}, ..., Y_{Ni} on N outcomes in each study i are the estimates of correlated true effects δ_{1i}, δ_{2i}, ..., δ_{Ni} with corresponding within-study covariance matrices $\mathbf{\Sigma_i}$ of the estimates. The elements of $\mathbf{\Sigma_i}$ are assumed known as discussed in Section 2.1. The study-level true effects δ_{ji} ($j = 1, \ldots N$) are assumed exchangeable and hence follow a common multivariate normal distribution with means $(d_1, d_2, \ldots, d_N)$ and covariance $\mathbf{T}$ in this hierarchical framework. Equations (11) and (12) describe the within-study and the between-studies models respectively. In the Bayesian framework, prior distributions need to be placed on the mean effects, for example $d_j \sim N(0.0, 1000)$ and the between-studies covariance matrix $\mathbf{T}$.

4.1.1 Prior distribution on the between-studies covariance matrix

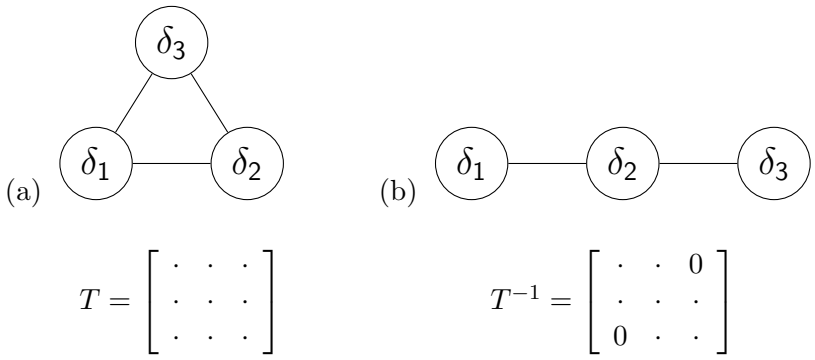
There is an additional complexity in multivariate meta-analysis of more than two outcomes, in particular when it comes to placing prior distributions on the between-studies covariances. A matrix of covariances has to satisfy certain conditions to be a proper covariance matrix. In linear algebra, matrix satisfying such conditions is called positive semi-definite. A prior distribution has to be placed on the covariance matrix in such a way to ensure that the matrix satisfies this condition. In contrast to the bivariate case, where it is sufficient to ensure that each element of the covariance matrix is given a prior distribution restricting the parameters to possible values (positive for the variances and between -1 and 1 for the correlation), in the general case a prior distribution has to be placed on the whole variance-covariance matrix or the correlation matrix. Placing independent prior distributions on the elements of the covariance matrix in general case, may lead to set of sampled values (in a MCMC simulation) amounting to the covariance matrix that is not positive semi-definite. There are a number of approaches to constructing a prior distribution on the between-studies covariance matrix, for example by using so called separation strategy with either Cholesky or spherical decomposition, which we describe in Appendix C.1.

4.2 MRMA in the product normal formulation

Multivariate meta-analysis can be parameterised in the product normal formulation by extending the model described in Section 2.2 to multiple outcomes. This approach has a number of advantages. In contrast to placing a prior distribution on the covariance matrix as a whole, the product normal formulation allows direct control over the prior distributions on all elements of the between-studies covariance matrix (between-studies standard deviations and correlations). It also describes the association patterns between the treatment effects on all outcomes which is useful when modelling multiple surrogate endpoints which we discuss in Section 4.5.

For simplicity, we focus here on the trivariate random-effect meta-analysis (TRMA), i.e. meta analysis of treatment effects on three correlated outcomes and describe MRMA for correlated treatment effects on any number of outcomes in the product normal formulation (MRMA PNF) in Appendix C.7. One of the advantages of the PNF for the multivariate meta-analysis is that it allows for a flexible modelling of the covariance structure. To illustrate this we describe two scenarios, one with fully unstructured covariance matrix and one with some structure imposed on the covariance. The two scenarios are illustrated graphically in Figure 4.

Figure 4: Scenarios for modelling multiple outcomes; (a) true treatment effects on all three outcomes are correlated and modelled using unstructured covariance matrix T, (b) true effects on outcomes one and three, δ_1 and δ_3 are conditionally independent, conditional on δ_2, which is modelled with structured covariance matrix equivalent to the precision matrix T^{-1} with element $\{1, 3\}$ equal to zero.



4.2.1 TRMA PNF unstructured model

For the first scenario where the true treatment effects δ_{ji} on all three outcomes are correlated, represented graphically in Figure 4a, the between-studies covariance has an unstructured form (this means that all three treatment effects are correlated and there are no special assumptions made about the covariance structure). The between study model (12) for three outcomes for this scenario is re-parameterised in the product normal formulation as a series of univariate conditional distributions

$$\begin{cases} \delta_{1i} \sim N(\eta_1, \psi_1^2) \\ \delta_{2i} \mid \delta_{1i} \sim N(\eta_{2i}, \psi_2^2) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\delta_{1i} \\ \delta_{3i} \mid \delta_{1i}, \delta_{2i} \sim N(\eta_{3i}, \psi_3^2) \\ \eta_{3i} = \lambda_{30} + \lambda_{31}\delta_{1i} + \lambda_{32}\delta_{2i}, \end{cases} \tag{13}$$

where λ_{20} and λ_{21} represent the intercept and slope for the association between the treatment effect on outcomes 2 and 1, and λ_{30}, λ_{31} and λ_{32} denote the intercept and the slopes for the association between treatment effects on outcome 3 and effects on outcomes 1 and 2, specifically λ_{31} is the

slope between the treatment effects on outcomes 3 and 1 and λ_{32} is the slope between the effects on outcomes 3 and 2 respectively. Instead of placing independent non-informative prior distributions on all the parameters of the model, relationships between these parameters and the elements of the between-studies covariance matrix are derived to allow to take into account the inter-relationship between the parameters. These relationships have the following forms

$$\begin{gathered}\psi_1^2 = \tau_1^2,\ \ \psi_2^2 = \tau_2^2 - \lambda_{21}^2\tau_1^2,\ \ \psi_3^2 = \tau_3^2 - \lambda_{31}^2\tau_1^2 - \lambda_{32}^2\tau_2^2 - 2\lambda_{31}\lambda_{32}\lambda_{21}\tau_1^2, \\ \lambda_{21} = \frac{\tau_2}{\tau_1}\rho^{12},\ \ \lambda_{31} = \frac{\tau_3\left(\rho^{13} - \rho^{12}\rho^{23}\right)}{\tau_1\left(1 - (\rho^{12})^2\right)},\ \ \lambda_{32} = \frac{\tau_3\left(\rho^{23} - \rho^{12}\rho^{13}\right)}{\tau_2\left(1 - (\rho^{12})^2\right)}\end{gathered} \tag{14}$$

which are obtained by following the procedure described in detail in Section C.7.1 for N-dimensional case. Having established them, allows us to place prior distributions on the between-studies standard deviations and correlations. By placing prior distributions on these parameters, for example $\rho^{12}, \rho^{13}, \rho^{23} \sim Unif(-1, 1)$, $\tau_{1(2,3)} \sim Unif(0, 2)$, the above derived relationships give the implied prior distribution on the parameters λ_{21}, λ_{31}, λ_{32}, ψ_1, ψ_2 and ψ_3. Note that the between-studies correlations and standard deviations are inter-related and their relationships with each other and with other parameters are defined by the formulaes (4.2.1). The remaining parameters are given "vague" prior distributions, $\eta_1 \sim N(0, 1000)$, $\lambda_{20(30)} \sim N(0, 1000)$.
Note also that the pooled effects $d_{1(2,3)}$ on the three outcomes in model (12) are also directly linked to the model (13); $d_1 = \eta_1$, $d_2 = \lambda_{20} + \lambda_{21}d_1$ and $d_3 = \lambda_{30} + \lambda_{31}d_1 + \lambda_{32}d_2$. It is possible to centre the true effects on outcomes 1 and 2 (by replacing δ_{ji} with $(\delta_{ji} - \overline{\delta_{ji}})$ in the third and fifth line of equation (13), which can be useful if there are problems with autocorrelation) in which case the pooled effects will be equal to the intercepts; $d_2 = \lambda_{20}$ and $d_3 = \lambda_{30}$.

4.2.2 TRMA PNF structured model

The aforementioned TRMA PNF model may be computationally demanding and even more so for MRMA of many outcomes, as it requires many parameters to be estimated and relatively large data sets may be needed to fit the model. Another option, that helps to reduce the number of parameters to be estimated, is a simplified model assuming conditional independence between treatment effects on some of the outcomes. For TRMA, such assumption of conditional independence can be made, for example, between the treatment effects on outcomes 1 and 3. We can consider a situation where treatment effect is measured on different outcomes sequentially in time (or on the same outcome repeatedly in time). For example, if treatment effect on the first outcome is measured at 12 months, second at 24 months and third at 36 months, then it may be conceivable to assume that the effect on the final outcome may be conditionally independent from the effect on the first outcome conditional on the second. This scenario, illustrated in Figure 4b, can be described by a simplified between-studies model with the true effect on the third outcome now conditional on the effect on the second outcome only:

$$\begin{cases} \delta_{1i} \sim N(\eta_1, \psi_1^2) \\ \delta_{2i} \mid \delta_{1i} \sim N(\eta_{2i}, \psi_2^2) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\ \mu_{1i} \\ \delta_{3i} \mid \delta_{2i} \sim N(\eta_{3i}, \psi_3^2) \\ \eta_{3i} = \lambda_{30} + \lambda_{32}\ \delta_{2i}, \end{cases} \tag{15}$$

The assumption of conditional independence of the true treatment effects on outcomes three and one, δ_3 and δ_1, conditional on the effect on outcome two, δ_2 puts a structure on the covariance matrix giving the corresponding element $\{1, 3\}$ of the inverse covariance (precision) matrix equal to zero. This means that partial correlation between the true treatment effects on outcomes one and three conditional on the true treatment effect on outcome two $\rho^{13|2} = 0$. Since the partial correlation between δ_1 and δ_3 (adjusted for δ_2) equals $\rho^{13|2} = \left(\rho^{13} - \rho^{12}\rho^{23}\right) / \sqrt{1 - (\rho^{12})^2}\sqrt{1 - (\rho^{23})^2} = 0$, it implies that $\rho^{13} = \rho^{12}\rho^{23}$. This reduces the number of parameters in the model that need to

be estimated and also simplifies the relationships between the parameters of the model with the elements of the between-studies covariance matrix:

$$\psi_1^2 = \tau_1^2, \;\; \psi_2^2 = \tau_2^2 - \lambda_{21}^2\tau_1^2, \;\; \psi_3^2 = \tau_3^2 - \lambda_{32}^2\tau_2^2,$$
$$\lambda_{21} = \rho^{12}\frac{\tau_2}{\tau_1}, \;\; \lambda_{32} = \rho^{23}\frac{\tau_3}{\tau_2}, \tag{16}$$

which both can have an advantage in scenarios with multiple outcomes beyond trivariate case as discussed in Appendix C.7. As in the case of the unstructured covariance matrix, placing prior distributions directly on the between-studies standard deviations $\tau_{1(2,3)} \sim N(0, 10)I(0,)$ and correlations $\rho^{12(23)} \sim Unif(-1, 1)$ gives implied prior distributions placed on the parameters of the model (15), $\psi_{1(2,3)}$ and $\lambda_{21(32)}$ obtained from the the derived relationship between the two sets of parameters. The remaining parameters are given non-informative prior distributions $\eta_1 \sim N(0, 1000)$, $\lambda_{20(30)} \sim N(0, 1000)$. The pooled effects $d_{1(2,3)}$ on the three outcomes in model (12) are also directly linked to the model (15); $d_1 = \eta_1$, $d_2 = \lambda_{20} + \lambda_{21}d_1$ and $d_3 = \lambda_{30} + \lambda_{32}d_2$. As in the model in Section 4.2.1, centering of the true effects on outcomes 1 and 2 can be applied to this model in which case the pooled effects would equal the intercepts.

4.3 Example: rheumatoid arthritis

To illustrate the use of MRMA methods, we return to the example in rheumatoid arthritis described in Section 2.3. We apply the methods discussed in this section to meta-analyse treatment effects on three outcomes, HAQ, DAS-28 and ACR20 from RA data included in Table 1. The results of applying MRMA with Cholesky decomposition and PNF with unstructured covariance, to the three outcomes in RA, are presented in Table 9 along with the results from univariate and bivariate meta-analysis, which were listed in Table 2, for comparison. Results of applying MRMA models with spherical decomposition and PNF with structured covariance are also available and can be found in Appendix C. The code for all four MRMA models and data corresponding to the trivariate case (of DAS-28, ACR20 and HAQ) are included in Appendix C.

The results are based on 50,000 iterations after a burn-in of 50,000 and applying thinning every 10 iteration (500,000 iterations were run after burn-in).

When applying MRMA to jointly synthesise the treatment effects on the three outcomes (extending application of BRMA to RA data in Section 2.3 to TRMA), we were able to include three additional studies (which reported the ACR20, but not the HAQ or DAS-28 scores). By incorporating an additional outcome, ACR20, we not only extended the data by those three studies, but also incorporated the data on ACR20 from eight studies already in BRMA, which reported the HAQ and/or DAS-28 as well as ACR20 (the details can be found in Table 1). Adding ACR20 in TRMA did not, however, lead to any change in the point estimate of HAQ or its uncertainty; they remained almost the same as in BRMA, perhaps with slightly increased uncertainty. The between-studies heterogeneity increased slightly (τ_H increased from 0.21 to 0.22 when using BRMA and TRMA in the standard form and from 0.22 to 0.23 when using the PNF) after inclusion of the 3 studies reporting the ACR20. This is likely due to the relatively high between-studies heterogeneity of the treatment effects on ACR20 across the studies and lack of correlation between the true treatment effects on ACR20 and HAQ or DAS28.

The between-studies correlation was estimated higher for TRMA with the Cholesky decomposition compared to TRMA in the product normal formulation. This is likely due to the fact that these models do not use the same prior distributions for the correlations. However, the uncertainty around the correlation was comparable, with more precise estimate obtained from TRMA with Cholesky decomposition.

Table 9: Results of the univariate meta-analyses of HAQ and DAS-28 separately, bivariate meta-analysis of HAQ and DAS-28 and trivariate meta-analysis of HAQ, ACR20 and DAS-28.

	posterior mean (SD) [95% CrI]						
	univariate analyses			bivariate analyses		trivariate analyses	
	HAQ	DAS-28	ACR20	HAQ & DAS-28		HAQ & DAS-28 & ACR20	
				BRMA	BRMA (PNF)	TRMA*	TRMA (PNF)†
HAQ	-0.25 (0.09) [-0.43,-0.07]			-0.27 (0.08) [-0.42,-0.11]	-0.27 (0.07) [-0.39,-0.13]	-0.27 (0.08) [-0.43, -0.10]	-0.26 (0.08) [-0.42, -0.09]
DAS-28		-1.57 (0.13) [-1.84,-1.31]		-1.53 (0.13) [-1.78,-1.26]	-1.53 (0.13) [-1.78,-1.26]	-1.53 (0.13) [-1.78, -1.26]	-1.54 (0.13) [-1.80, -1.28]
ACR20‡			0.62 (0.05) [0.53, 0.72]			0.61 (0.05) [0.51, 0.71]	0.61 (0.05) [0.52, 0.71]
τ_H	0.21 (0.09) [0.10,0.44]			0.21 (0.08) [0.11,0.41]	0.22 (0.08) [0.11,0.41]	0.22 (0.08) [0.11, 0.43]	0.23 (0.09) [0.11, 0.45]
τ_D		0.44 (0.11) [0.25,0.67]		0.44 (0.11) [0.27,0.71]	0.44 (0.11) [0.27,0.71]	0.45 (0.11) [0.27,0.72]	0.44 (0.11) [0.27, 0.71]
τ_A			0.52 (0.19) [0.25, 0.99]			0.56 (0.21) [0.26, 1.06]	0.56 (0.21) [0.26, 1.06]
ρ^{DH}				0.58 (0.42) [-0.51,0.99]	0.59 (0.43) [-0.51,0.99]	0.63 (0.40) [-0.45, 0.99]	0.46 (0.42) [-0.51, 0.97]
ρ^{AH}						-0.06 (0.43) [-0.83, 0.73]	-0.01 (0.43) [-0.82, 0.73]
ρ^{DA}						-0.16 (0.40) [-0.82, 0.64]	-0.14 (0.37) [-0.78, 0.61]

* with prior distribution for the between-studies covariance matrix using Cholesky decomposition,
† using PNF model corresponding to the unstructured between-studies covariance matrix,
‡ transformed to proportion of responders (modelled using log odds scale).

We focussed here on the estimate of the treatment of HAQ as an outcome of particular interest in HTA as it is a disease-specific measure of quality of life in RA. However, in general terms, all outcomes may be of interest in multivariate meta-analysis and can be investigated in a similar manned. In our example in RA, the average treatment effect on other outcomes did not benefit from the multivariate approach in terms of reduced uncertainty. The point estimate for the treatment effect on DAS-28 has shifted slightly, but the uncertainty remained the same when using both BRMA and TRMA. Treatment effect on ACR20 remained unchanged.

4.4 Benefits of multivariate meta-analysis

The benefits of multivariate meta-analysis are diverse but mostly are related to the potential gain in precision when estimating effects of treatments, as discussed by Riley et al. [14]. Such gain has been quantified by measures of efficiency [74] and borrowing of strength (BoS) [75]. Copas et al. [74] propose that, in comparison to a multivariate (or network) meta-analysis with the same magnitude of between-trial heterogeneity, a standard (univariate) meta-analysis (of just the direct evidence in NMA) is similar to throwing away $100 \times (1-E)\%$ of the available studies. The efficiency (E) is defined by,

$$E = \frac{\text{variance of summary result based on direct and related evidence}}{\text{variance of summary result based on only direct evidence}}$$

where "related evidence" refers to either indirect or correlated evidence (or both), and the variance relates to the original scale of the meta-analysis (so typically the log relative risk, log odds ratio, log hazard ratio, or mean difference). For example, if $E = 0.9$ then a standard meta-analysis is similar to throwing away 10% of available studies and patients (and events). Let us also define n as the number of available studies with direct evidence (i.e. those that would contribute toward a standard meta-analysis). Then, the extra information gained toward a particular summary meta-analysis result by using indirect or correlated evidence can also be considered similar to having found direct evidence from a further $n \times ((1 - E))/E$ studies of a similar size to the n trials. For example, if there are nine studies providing direct evidence about an outcome for a standard univariate meta-analysis and E = 0.9, then the advantage of using a multivariate meta-analysis is like finding direct evidence for that outcome from a further $9 \times ((1 - 0.9))/0.9 = 1$ study. We thus gain the considerable time, effort and money invested in about one research study.

Jackson et al. also propose the borrowing of strength (BoS) statistic, [75] which can be calculated for each summary result within a multivariate or network meta-analysis by $BoS = 100 \times (1 - E)\%$. BoS provides the percentage reduction in the variance of a summary result that is due to (borrowed from) correlated or indirect evidence. An equivalent way of interpreting BoS is the percentage weight in the meta-analysis that is given to the correlated or indirect evidence [76]. For example, in a network meta-analysis, a BoS of 0% indicates that the summary result is based only on direct evidence, whereas a BoS of 100% indicates that it is based entirely on indirect evidence. Riley et al. show how to derive percentage study weights for multi-parameter meta-analysis models, including network and multivariate meta-analysis [77].

Following Riley et al. [14], the potential importance of a multivariate meta-analysis of multiple outcomes is greatest when borrowing of information and gain in precision are large, which is more likely when:

- the proportion of studies without direct evidence for an outcome of interest is large;
- results for other outcomes are available in studies where an outcome of interest is not reported;
- and the magnitude of correlation amongst outcomes is large (e.g. > 0.5 or < -0.5), either within-studies or between-studies.

Riley et al. [14] suggest that BoS and (1-E) are usually greatest in a network meta-analysis of multiple treatments; that is, more information is usually gained about multiple treatments via

the consistency assumption than is gained about multiple outcomes via correlation. A multivariate meta-analysis of multiple outcomes is best reserved for a set of highly correlated outcomes, as otherwise BoS and E are usually small. Such outcomes should be identified and specified in advance of analysis, for example using clinical judgement and statistical knowledge, so as to avoid data dredging across different sets of outcomes. A multivariate meta-analysis of multiple outcomes is also best reserved for a situation with missing outcomes (at the study-level), as anecdotal evidence suggests that BoS for an outcome is approximately bounded by the percentage of missing data for that outcome. For example, in multivariate meta-analysis to examine the prognostic effect of a 1-unit increase of fibrinogen on CVD rate, the percentage of trials with a missing fully adjusted outcome is $55\%(= 100\% \times 17/31)$, and thus the multivariate approach is flagged as worthwhile as BoS could be as high as 55% for the fully adjusted pooled result. As discussed, the actual BoS was 53% and thus very close to 55%, due to the near perfect correlation between partially and fully adjusted effects. In contrast, in situations with complete data or a low percentage of missing outcomes, BoS (and thus a multivariate meta-analysis) is unlikely to be important. Multivariate meta-analysis is likely to be most advantageous in terms of large BoS when there are missing outcome data. In situations with complete data (i.e. all outcomes are available in all studies) BoS will usually be small (e.g. $< 10\%$). Also, multivariate meta-analysis cannot handle trials that do not report any of the outcomes of interest. Therefore, although it can reduce the impact of selective outcome reporting in published trials, it cannot reduce the impact of non-publication of entire trials (publication bias).

In addition to increasing precision, the extra information used in multivariate meta-analysis may lead to different summary effect estimates compared to univariate meta-analysis. This is most likely to occur in situations with selectively missing data for some outcomes. For example Kirkham et al. [20] conclude that multivariate meta-analysis can reduce the impact of outcome reporting bias, where some outcomes are selectively missing (for example, based on their p-value) but other correlated outcomes (from which information can be borrowed) are more consistently available.

If a formal comparison of correlated treatment effects is of interest (e.g. to estimate the difference between the treatment effects on systolic and diastolic blood pressure), then this should always be done in a multivariate framework regardless of the amount of missing data, in order to account for correlations between treatment effects and thus avoid erroneous confidence intervals and p-values [4]. Similarly, a network meta-analysis of multiple treatments is preferable even if all trials examine all treatments, as we require a single analysis framework for estimating and comparing the effects of each treatment.

The benefits of multivariate (and network) meta-analysis depend on missing study results being missing at random [78]. We are assuming that the relationships that we do observe in some trials are transferable to other trials where they are unobserved. For example, in a multivariate meta-analysis of multiple outcomes the observed linear association (correlation) of effects for pairs of outcomes (both within-studies and between-studies) is assumed to be transferable to other studies where only one of the outcomes is available. This relationship is also used to justify surrogate outcomes [79, 12], but often receives criticism and debate therein [80]. Missing not at random may be more appropriate when results are missing due to selective outcome reporting [81], or selective choice of analyses [82]. Missing at random assumes that the distribution of missing values is known conditional on other variables that are available. However, with selective outcome reporting, it is likely that the full set of factors needed (for the missing at random assumption to hold) is unavailable. However, even by conditioning on one of the factors related to the missingness (here a correlated outcome) we may do better (in terms of statistical properties of the summary meta-analysis results) in a multivariate meta-analysis, than in a univariate meta-analysis (which does not condition on any of the factors). A multivariate approach may therefore still reduce selective reporting biases in this situation [20], but not completely.

4.5 Application to multiple surrogate endpoints

Most meta-analytic methods for surrogate endpoint evaluation are designed to evaluate a single surrogate endpoint. However, methods for multiple surrogate endpoints evaluated as joint predictors of clinical benefit have also been proposed. The idea of evaluating multiple surrogate endpoints jointly is not new. In the summary of a National Institutes of Health Workshop on the use of surrogate endpoints, Gruttola et al. [83] made a number of recommendations for future research that included development of models that can accommodate multiple surrogate endpoints and/or multiple clinical outcomes. Methods for evaluating multiple surrogate endpoints were proposed by Xu and Zeger [84] for time-to-event data modelled jointly with multiple biomarkers measured longitudinally. Other examples of validating multiple surrogate endpoints include the plasma HIV-1 RNA and $CD4^+$ lymphocytes as predictors of progression to AIDS in HIV-positive patients [85, 86] and relapse rate and number of active lesions in the brain as predictors of disability progression in relapsing remitting multiple sclerosis (RRMS) [87]. These studies investigated study-level surrogacy using individual-level data.

Bujkiewicz et al developed methods for multivariate meta-analysis in product normal formulation, described in Sections 4.2.1 and 4.2.2, to evaluate multiple surrogate endpoints in a meta-analytic framework [13]. This can be achieved in the same way as using the bivariate meta-analysis, by extending BRMA to multiple outcomes: the number of surrogate endpoints plus the final outcome. Both scenarios depicted graphically in Figure 4, and many more for any number of outcomes or covariance structures, could be investigated. Similarly as for BRMA, the regression coefficients, the intercept, slope and the conditional variance, can be used to describe and evaluate surrogate relationships. When using MRMA in the standard form, the between-studies correlations (or correlations squared) can be used to quantify the strength of the surrogate relationship. The choice between the models (and appropriate structure for the covariance matrix) can be made based on the biological plausibility of the correlations supported by the exploratory analysis of the data and verified by fitting alternative models and making comparison using the deviance information criteria (DIC) or cross-validation procedure. As in bivariate case discussed in Section 3.4, the assumption of random effect for the surrogate endpoints may not be satisfied and it may be preferable to assume fixed effects for the true treatment effects on the surrogate endpoints, as in the model by Daniels and Hughes (7) for a single surrogate endpoint. A similar approach was investigated by Pozzi et al. [88].

5 Network meta-analysis of multiple correlated outcomes

The methods and examples considered in the previous sections were focused on multiple outcomes. The heterogeneity of evidence is also often related to multiple treatments. Data from RCTs investigating multiple treatments can be synthesised simultaneously using NMA, which is well explained in the NICE DSU TSD by Dias et al. [41]. However, there is a growing interest in accommodating both multiple outcomes and multiple treatments together, in order to help identify the best treatment across multiple clinically relevant outcomes [89, 90, 91, 92, 93, 94]. This is achievable, but challenging due to the extra complexity of the statistical models required.

There is a growing body of literature on methods for multivariate network meta-analysis (mvNMA). For example, Efthimiou *et al.* [89] proposed a model for the joint modelling of odds ratios on multiple endpoints. Efthimiou et al. then developed another model [91] that is a network extension of an alternative multivariate meta-analytic model that was originally proposed by Riley et al.[46]. The authors perform a network meta-analysis of 68 studies comparing 13 active antimanic drugs and placebo for acute mania. Two primary outcomes of interest were efficacy (defined as the proportion of patients with at least a 50% reduction in manic symptoms from baseline to week 3) and non-acceptability (defined as the proportion of patients with treatment discontinuation before 3 weeks). These are likely to be negatively correlated (as patients often discontinue treatment due to lack of efficacy), so the authors extend a network meta-analysis framework to jointly analyse these outcomes and account for their correlation (estimated to be about -0.5; by applying an extension of Riley's overall correlation modelling approach [46, 91], which avoids having to know the within-study correlations which were unavailable here). This is especially important as 19 of the 68 studies provided data on only one of the two outcomes. Compared to considering each outcome separately, this approach produces narrower confidence intervals for summary treatment effects and has an impact on the relative ranking of some of the treatments. In particular, carbamazepine ranks as the most effective treatment in terms of response when considering outcomes separately, but falls to fourth place when accounting for their correlation.

Achana *et al.* [90] developed a model for multiple correlated outcomes in multi-arm studies in public health. The authors extend the standard NMA model to multiple outcome settings in two stages. In the first stage, information is borrowed across outcomes as well as across studies through modelling the within-study and between-studies correlation structure. In the second stage, an additional assumption is made, that intervention effects are exchangeable between outcomes, to predict effect estimates for all outcomes. This enables prediction of treatment effects on outcomes for which evidence is either sparse or the treatment effects had not been considered by any one of the studies included in the analysis. Achana et al. applied the methods to binary outcome data from a systematic review evaluating the effectiveness of nine home safety interventions on uptake of three poisoning prevention practices (safe storage of medicines, safe storage of other household products, and possession of poison centre control telephone number) in households with children. The first stage multivariate models produced broadly similar point estimates of intervention effects as the univariate approach to NMA, but the uncertainty around the multivariate estimates varied depending on the prior distribution specified for the between-studies covariance structure. The second stage multivariate analyses produced more precise effect estimates while enabling intervention effects to be predicted for all outcomes, including intervention effects on outcomes not directly considered by the studies included in the analysis.

Bujkiewicz et al adapted and developed further these ideas for use in surrogate endpoint evaluation, where surrogate relationships may vary across treatment contrasts [48].

In this TSD, we start by introducing a simple bivariate network meta-analysis (bvNMA) for data from two-arm studies by showing how BRMA, introduced in Section 2.1, extends to bvNMA by taking into account the network structure of the data (entered at the contrast level, i.e. as relative effects). We then describe an alternative formulation for data entered at the arm level (absolute effects in each treatment arm). To clarify, both models are contrast based, only the data entry is at either contrast or arm level leading to different parameterisations of the within-study

model. We then move onto describing a general model of mvNMA for multi-arm studies of multiple outcomes. We illustrate the use of this method in an example in multiple sclerosis.

5.1 Bivariate network meta-analysis (bvNMA) for contrast level data entry

In the pairwise meta-analytic methods we model treatment effects corresponding to the same treatment contrast (or sometimes class of treatments) where the treatment effects follow a common distribution at the between-studies level. In contrast to this, the NMA methods allow us to model simultaneously effects corresponding to multiple treatment contrasts and only the effects corresponding to the same treatment contrast follow the same distribution. In other words, only the treatment effects corresponding to the same treatment contrast are assumed similar (exchangeable). The consistency assumption then brings all the evidence together by combining the direct effects (for example of treatments B vs A and C vs A, obtained from studies of these specific treatment contrasts) and indirect effects (C vs B, obtained from treatment effects B and C vs a common reference treatment A). This is a common approach to NMA, described in detail for the univariate case by Dias et al. [41]. The extension from pairwise model to NMA model is made by distinguishing between studies/effects of different contrasts and allowing for consistency assumption. In a similar way, pairwise BRMA model, discussed in Section 2.1, can be generalised to bvNMA.

To model correlated treatment effects on two outcomes for multiple treatment contrasts, the assumption made by BRMA, that the true effects follow a common distribution, can be replaced by an assumption that the true effects corresponding to different treatment contrasts follow separate distributions (but common to studies from the same treatment contrasts). This naturally leads to allowing the elements of the between-studies covariance matrix (the between-studies correlations ρ_{kl}^{12} between the treatment effects l vs. k on two outcomes and the heterogeneity parameters for the treatment effects on the two outcomes $\tau^2_{(kl)1}$ and $\tau^2_{(kl)2}$) to vary across the treatment contrasts kl.

To take into account the network structure of the data, we model the treatment effect differences $Y_{(kl)ij}$ between treatments k and l in study i for outcome $j = 1, 2$, following an approach by Bujkiewicz et al [48]:

$$\begin{pmatrix} Y_{(kl)i1} \\ Y_{(kl)i2} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \delta_{(kl)i1} \\ \delta_{(kl)i2} \end{pmatrix}, \mathbf{\Sigma_i} \right\}, \ \mathbf{\Sigma_i} = \begin{pmatrix} \sigma^2_{1kli} & \sigma_{1kli}\sigma_{2kli}\rho_{wkli} \\ \sigma_{1kli}\sigma_{2kli}\rho_{wkli} & \sigma^2_{2kli} \end{pmatrix} \tag{17}$$

$$\begin{pmatrix} \delta_{(kl)i1} \\ \delta_{(kl)i2} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} d_{(kl)1} \\ d_{(kl)2} \end{pmatrix}, \begin{pmatrix} \tau^2_{(kl)1} & \tau_{(kl)1}\tau_{(kl)2}\rho^{12}_{kl} \\ \tau_{(kl)1}\tau_{(kl)2}\rho^{12}_{kl} & \tau^2_{(kl)2} \end{pmatrix} \right\} \tag{18}$$

where k and l denote baseline (control) and experimental treatments respectively in study i, $\delta_{(kl)ij}$ denote the random true treatment effects (differences between the effects of treatments k and l) on outcome j in study i, and the $d_{(kl)j}$ are mean treatment effect differences between treatments k and l for each outcome j. The first-order consistency assumption, described by Lu and Ades [95], is extended here to the bivariate case. For any three treatments (b, k, l), the treatment differences $(\delta_{(kl)ij})$ satisfy the following transitivity relations

$$\begin{pmatrix} \delta_{(kl)i1} \\ \delta_{(kl)i2} \end{pmatrix} = \begin{pmatrix} \delta_{(bl)i1} - \delta_{(bk)i1} \\ \delta_{(bl)i2} - \delta_{(bk)i2} \end{pmatrix}. \tag{19}$$

These relationships imply (by taking their mean) the first order consistency equations

$$\begin{pmatrix} d_{(kl)1} \\ d_{(kl)2} \end{pmatrix} = \begin{pmatrix} d_{(bl)1} - d_{(bk)1} \\ d_{(bl)2} - d_{(bk)2} \end{pmatrix} \tag{20}$$

which represent the relationships between the treatment contrasts in the population. When $b = 1$ is a common reference treatment in the network, the treatment effects of each treatment k relative to

this common reference treatment 1; the $d_{(1k)j}$ are referred to as basic parameters for each outcome j, with $d_{(11)j} = 0$ and the others are given prior distributions:

$$d_{(1k)j} \sim N(0, 10^3). \tag{21}$$

Prior distributions are also placed on the elements of the between-studies variance-covariance matrices. As in BRMA, prior distributions for the heterogeneity parameters are selected to ensure that they are restricted to plausible positive values, such as $\tau_{(kl)j} \sim Unif(0, 2)$ and for the correlations to ensure restriction to the values between -1 and 1, e.g. $\frac{\rho^{12}_{kl}+1}{2} \sim Beta(1.5, 1.5)$, thus guaranteeing that the variance-covariance matrix for each treatment contrast is positive semi-definite.

5.2 Bivariate network meta-analysis (bvNMA) for arm level data entry

The above model was applicable to data representing the estimates of treatment effects $Y_{(kl)ij}$ at the contrast level, comparing the effect of treatment l vs k for each outcome j and study i. When data are available at the arm level, we can reformulate the bvNMA as in Achana et al [90]:

$$\begin{pmatrix} Y_{ik1} \\ Y_{ik2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_{ik1} \\ \theta_{ik2} \end{pmatrix}, \Sigma_{\mathrm{ik}} = \begin{pmatrix} \sigma^2_{ik1} & \sigma_{ik1}\sigma_{ik2}\rho^{12}_{wi} \\ \sigma_{ik1}\sigma_{ik2}\rho^{12}_{wi} & \sigma^2_{ik2} \end{pmatrix} \right)$$

$$\begin{pmatrix} \theta_{ik1} \\ \theta_{ik2} \end{pmatrix} = \begin{cases} \begin{pmatrix} \mu_{ib1} \\ \mu_{ib2} \end{pmatrix}, & b = k \\ \begin{pmatrix} \mu_{ib1} + \delta_{(bk)i1} \\ \mu_{ib2} + \delta_{(bk)i2} \end{pmatrix}, & k > b \end{cases}$$

$$\begin{pmatrix} \delta_{(bk)i1} \\ \delta_{(bk)i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} d_{(bk)1} \\ d_{(bk)2} \end{pmatrix}, \begin{pmatrix} \tau^2_{(bk)1} & \tau_{(bk)1}\tau_{(bk)2}\rho^{12}_{bk} \\ \tau_{(bk)1}\tau_{(bk)2}\rho^{12}_{bk} & \tau^2_{(bk)2} \end{pmatrix} \right)$$

where the within-study correlations ρ^{12}_{wi} between the estimates of treatment effects Y_{ikj} in each arm k on the two outcomes $j = 1, 2$ in each study i are assumed known. The between-studies correlations between effect on outcomes 1 and 2 for each contrast bk are given prior distributions, for example $\rho^{12}_{bk} \sim Unif(-1, 1)$. The average effects can be represented in terms of the basic parameters $d_{(1k)j}$ for the effect of each treatment k compared to the common reference treatment in the network $b = 1$ for each outcome j:

$$d_{(bk)j} = d_{(1k)j} - d_{(1b)j}, \tag{22}$$

and the basic parameters are given prior distributions: $d_{(1k)j} \sim N(0, 10^3)$, $d_{11} = 0$. We place prior distributions on other parameters: the baseline effects in each study i, $\mu_{ibj} \sim N(0, 10^3)$ and the heterogeneity parameters $\tau_{(bk)j} \sim Unif(0, 2)$, $j = 1, 2$.

Often homogeneity of the between-studies parameters (correlations and standard deviations) is assumed: $\rho^{12}_{bk} = \rho^{12}$ and $\tau_{(bk)j} = \tau_j$, $j = 1, 2$ resulting in a simplified form of the between-studies model:

$$\begin{pmatrix} \delta_{(bk)i1} \\ \delta_{(bk)i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} d_{(bk)1} \\ d_{(bk)2} \end{pmatrix}, \begin{pmatrix} \tau^2_1 & \tau_1\tau_2\rho^{12} \\ \tau_1\tau_2\rho^{12} & \tau^2_2 \end{pmatrix} \right)$$

with prior distributions $\rho^{12} \sim Unif(-1, 1)$, $\tau_m \sim Unif(0, 2)$, $d_{(1k)m} \sim N(0.0, 10^3)$, $\mu_{ibm} \sim N(0.0, 10^3)$, $m = 1, 2$.

5.3 Multivariate network meta-analysis (mvNMA) for arm level data entry

We present here a mvNMA model described by Achana et al [90] assuming that the treatment effects are normally distributed. We assume that in each study i and for each k-th arm, the estimates of the treatment effects Y_{ikm} (such as log odds of an event) on all outcomes m ($m = 1, 2, \ldots, M$) jointly follow a multivariate normal distribution:

$$\begin{pmatrix} Y_{ik1} \\ \vdots \\ Y_{ikM} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_{ik1} \\ \vdots \\ \theta_{ikM} \end{pmatrix}, \Sigma_{\mathrm{ik}} = \begin{pmatrix} \sigma^2_{ik1} & \cdots & \sigma_{ik1}\sigma_{ikM}\rho^{1M}_{ik} \\ \vdots & \ddots & \vdots \\ \sigma_{ik1}\sigma_{ikM}\rho^{1M}_{ik} & \cdots & \sigma^2_{ikM} \end{pmatrix}\right) \tag{23}$$

$$\begin{pmatrix} \theta_{ik1} \\ \vdots \\ \theta_{ikM} \end{pmatrix} = \begin{cases} \begin{pmatrix} \mu_{ib1} \\ \vdots \\ \mu_{ibM} \end{pmatrix}, & b = k \\ \begin{pmatrix} \mu_{ib1} + \delta_{(bk)i1} \\ \vdots \\ \mu_{ibM} + \delta_{(bk)iM} \end{pmatrix}, & k > b \end{cases} \tag{24}$$

$$\begin{pmatrix} \delta_{(bk)i1} \\ \vdots \\ \delta_{(bk)iM} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} d_{(bk)1} \\ \vdots \\ d_{(bk)M} \end{pmatrix}, \mathrm{T_{M\times M}} = \begin{pmatrix} \tau^2_1 & \cdots & \tau_1\tau_M\rho^{1M} \\ \vdots & \ddots & \vdots \\ \tau_1\tau_M\rho^{1M} & \cdots & \tau^2_M \end{pmatrix}\right) \tag{25}$$

where $(Y_{ik1}, \ldots, Y_{ikM})$ and $(\theta_{ib1}, \theta_{ib2}, \ldots, \theta_{ibM})$ represent vectors of observed and true treatment effects respectively (for example log-odds of response) in arm k of study i and Σ_{ik} is the associated within-study covariance matrix usually assumed known but estimated in practice from the data [2]. The within-study correlations ρ^{mn}_{ik} between treatment effects on outcomes m and n ($m \neq n$) in arm k of study i can be obtained from IPD as discussed in Section 2.1.1. Vectors $(\mu_{ib1}, \ldots, \mu_{ibM})$ represent study-specific true baseline treatment effects and $(\delta_{(bk)i1}, \ldots, \delta_{(bk)iM})$ are relative effects of treatment in arm k versus baseline treatment b in study i for outcomes $(1, \ldots, M)$. Equation (25) describes the between-studies model for the network of two-arm trials. The true effects $\delta_{i(bk)m}$ ($m = 1, 2, \ldots, M$) jointly follow a Normal distribution with mean effects $d_{(bk)m}$, which are the pooled effects of treatment k relative to treatment b and τ^2_m is the between-studies variance or heterogeneity parameter corresponding to outcome m (assuming homogeneity of the between-studies variances and correlations across treatments as discussed in the previous section). The pooled effects $d_{(bk)m}$ can be expressed as functions of basic parameters as in the bivariate case (i.e. $d_{(bk)m} = d_{(1k)m} - d_{(1b)m}$). This multivariate NMA extends to allow for the multi-arm trials by assuming that all effects in trial i relative to a common baseline treatment in this trial are correlated and normally distributed. We discuss this in more detail in Appendix D.1.

The baseline effects and the basic parameters are given minimally informative prior distributions: $\mu_{ibm} \sim N(0, 10^3)$ and $d_{(1k)m} \sim N(0, 10^3)$. A prior distribution also needs to be specified for the between-studies covariance matrix $T_{M\times M}$ ensuring it is positive semi-definite. This can be achieved, for example, by using a separation strategy, with spherical [95, 6] or Cholesky decomposition [6] of the correlation matrix, as outlined by Lu and Ades [95] and more recently by Wei and Higgins [6] (see Section 4.1.1 for details). The separation strategy following Barnard et al. [96] expresses the covariance matrix: $T_{M\times M} = V^{1/2}RV^{1/2}$, where $V^{1/2}$ is a diagonal matrix of the standard deviations and R is a positive semi-definite matrix of correlations. In Cholesky decomposition, matrix R is represented as $R = L^TL$ with L being a $M \times M$ upper triangular matrix. The spherical parameterization technique [6, 95] can be used to express R in terms of sine and cosine functions of the elements in L. Using this later technique, we specified uniform prior distributions for the spherical coordinate in our model, $\phi_{mn} \sim Unif(0, \pi)$ to ensure that elements of the correlation matrix R lie in the interval $(-1, 1)$. Finally, the elements of $V^{1/2}$ correspond to the between-studies standard deviation terms in $T_{M\times M}$ and are given independent uniform prior distributions; $Unif(0, 2)$.

5.4 Example: relapsing remitting multiple sclerosis (RRMS)

5.4.1 Data

Melendez-Torres et al. conducted a systematic review and economic evaluation to investigate the clinical and cost-effectiveness of beta-interferon and glatiramer acetate as treatment options for adults with RRMS [97]. Thirty-two studies identified in the review evaluated the impact of various interventions across a range of RRMS outcomes: annualised relapse rate, disability progression, proportion remaining relapsed free at end of follow-up, treatment related adverse events, and discontinuation due to adverse events 24 months after randomisation. For purpose of illustrating multivariate NMA modelling, we consider models for arm-level data and the bivariate case (bvNMA) with proportion remaining relapsed free and discontinuation due to adverse events as outcomes. Review by Melendez-Torres et al. provided summary data for these two outcomes in the form of the number of events and the total number randomised to treatment for 22 of the 32 studies comparing a total of 8 interventions. Table 10 summarises the available data and Figure 5 shows network diagrams representing the data structure. Data were available for both outcomes from 11 studies with the remaining 11 studies only reporting data for the proportion relapse free. Two of the 22 studies are 3-arm trials, one reporting data for the proportion relapse free [98] and the other reporting both outcomes [99]. We added 0.5 continuity correction when arms had zero events. The network for proportion relapse free shows comparisons between 8 interventions from 22 studies whilst discontinuation due to adverse events shows comparisons between 6 interventions from 11 studies that reported both outcomes. We refer to these data as the "RRMS network data" in the remainder of this TSD.

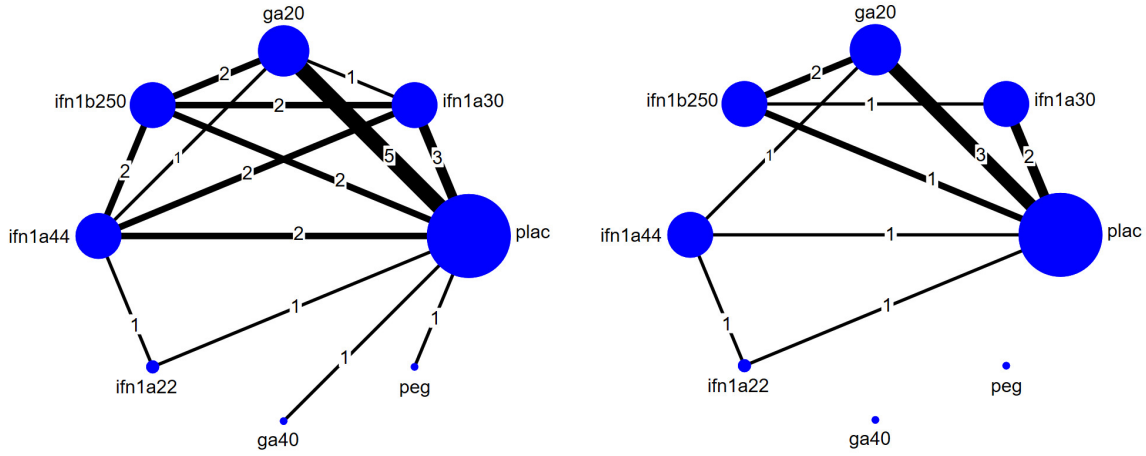


Figure 5: Network diagrams for proportion remaining relapse free (left graph) and discontinuation due to adverse events at 24 months (right graph). Treatment names: Placebo (plac), Glatiramer 20 mg daily (ga20), Glatiramer 40 mg thrice weekly (ga40), IFN-B-1a 22ug SC 3x daily (ifn1a22), IFN-B-1a 30ug IM weekly (ifn1a30), IFN-B-1a 44ug SC 3x weekly (ifn1a44), IFN-B-1a 250ug SC every other day (ifn1a250) and IFN-B-1a pegylated 125ug every 2 weeks.

Table 10: RRMS network data from systematic review by Melendez-Torres et al.

Study Name	Treatment	Number	r1	n1	r2	n2	ID	Arm
Euro-Canadian 2001	Placebo	1	59	120	NA	NA	1	1
Euro-Canadian 2001	Glatiramer 20 mg daily	2	66	119	NA	NA	1	2
GATE 2015	Placebo	1	62	84	NA	NA	2	1
GATE 2015	Glatiramer 20 mg daily	2	264	357	NA	NA	2	2
GALA 2013	Placebo	1	302	461	NA	NA	3	1
GALA 2013	Glatiramer 40 mg thrice weekly	7	726	943	NA	NA	3	2
Kappos 2011	Placebo	1	41	54	NA	NA	4	1
Kappos 2011	IFN B-1a 30 mcg IM weekly	3	42	54	NA	NA	4	2
REMAIN 2012	Placebo	1	7	15	NA	NA	5	1
REMAIN 2012	IFN B-1a 44 mcg SC thrice weekly	4	11	15	NA	NA	5	2
ADVANCE 2014	Placebo	1	370	500	NA	NA	6	1
ADVANCE 2014	IFN B-1a pegylated 125 mcg every 2 weeks	8	422	512	NA	NA	6	2
Knobler 1993	Placebo	1	1	6	NA	NA	7	1
Knobler 1993	IFN B-1b 250 mcg SC every other day	5	10	24	NA	NA	7	2
EVIDENCE 2007	IFN B-1a 30 mcg IM weekly)	3	162	338	NA	NA	8	1
EVIDENCE 2007	IFN B-1a 44 mcg SC thrice weekly	4	190	339	NA	NA	8	2
CombiRx 2013	Glatiramer 20 mg daily	2	206	259	NA	NA	9	1
CombiRx 2013	IFN B-1a 30 mcg IM weekly	3	185	250	NA	NA	9	2
REFORMS 2012	IFN B-1a 44 mcg SC thrice weekly	4	6	65	NA	NA	10	1
REFORMS 2012	IFN B-1b 250 mcg SC every other day	5	0.5	64	NA	NA	10	2
Etemadifar 2006	IFN B-1a 30 mcg IM weekly	3	17	30	NA	NA	11	1
Etemadifar 2006	IFN B-1a7 44 mcg SC thrice weekly	4	6	30	NA	NA	11	2
Etemadifar 2006	IFN B-1b 250 mcg SC every other day	5	13	30	NA	NA	11	3
CONFIRM 2012	Placebo	1	149	363	38	363	12	1
CONFIRM 2012	Glatiramer 20 mg daily	2	112	350	35	351	12	2
Bornstein 1987	Placebo	1	6	23	0.5	23	13	1
Bornstein 1987	Glatiramer 20 mg daily	2	14	25	2	25	13	2
Copolymer 1 1995	Placebo	1	34	126	1	126	14	1
Copolymer 1 1995	Glatiramer 20 mg daily	2	42	125	5	125	14	2
BRAVO 2014	Placebo	1	275	450	19	450	15	1
BRAVO 2014	IFN B-1a 30 mcg IM weekly	3	308	447	26	447	15	2
MSCRG 1996	Placebo	1	23	87	2	143	16	1
MSCRG 1996	IFN B-1a 30 mcg IM weekly	3	32	85	7	158	16	2
IFNB group 1993	Placebo	1	17	123	1	123	17	1
IFNB group 1993	IFN B-1b 250 mcg SC every other day	5	27	124	10	124	17	2
REGARD 2008	Glatiramer 20 mg daily	2	132	347	19	378	18	1
REGARD 2008	I FN B-1a 44 mcg SC thrice weekly	4	126	332	23	386	18	2
BEYOND	Glatiramer 20 mg daily	2	264	448	8	448	19	1
BEYOND	IFN B-1b 250 mcg SC every other day	5	520	897	13	897	19	2
BECOME 2009	Glatiramer 20 mg daily	2	28	39	0.5	39	20	1
BECOME 2009	IFN B-1b 250 mcg SC every other day	5	19	36	1	36	20	2
INCOMIN 2002	IFN B-1a 30 mcg IM weekly	3	33	92	1	92	21	1
INCOMIN 2002	IFN B-1b 250 mcg SC every other day	5	49	96	5	96	21	2
PRISMS 1998	Placebo	1	30	187	1	187	22	1
PRISMS 1998	IFN B-1a 44 mcg SC thrice weekly	4	59	184	7	184	22	2
PRISMS 1998	IFN B-1a 22 mcg SC thrice weekly	6	51	189	3	189	22	3

5.4.2 Models fitted

Fitting the mvNMA models described in this TSD requires a vector of continuous and approximately normally distributed responses. Non-normal data should therefore be transformed so that the data are approximately normal on the transformed scale of measurement. Regardless of data type, however, the statistics required for analysis are exactly the same and come in the form of the arm-specific vector of mean-responses and matrices carrying covariance information between outcomes for each study. For our illustrative example, binary data were available as number of events and total number randomised in each treatment-arm for proportion relapse free and discontinuation due to adverse events. We model on the log odds scale by taking logistic transformation of the underlying event probabilities of each outcome. Corresponding estimated standard errors on the log odds scale were calculated using standard 2×2 table formulae. Within-study correlations

were not available, hence we fitted models reflecting this lack of information by formulating prior distributions for the within-study correlation based on assumptions about the correlation between the treatment effect:

- In Model 1, we assumed that outcomes are uncorrelated by setting the within-study and between-studies correlation parameters to zero. This model is therefore equivalent to fitting separate univariate network meta-analysis (uvNMA) models.
- Model 2 assumes outcomes are uncorrelated within-studies by setting the within-study correlation to zero but take count of the between-studies correlation of treatment effects on different outcomes.
- In Model 3 we place a uniform prior distribution on the positive scale on the within-study correlations, $Unif(0, 1)$ encoding the assumption that more effective treatments (as measured by the number of patients remaining relapse free) are also more likely to be associated with increased adverse effect profile leading to discontinuation of treatment.
- Model 4 expresses a complete lack of information about the nature of the within-study correlation between the treatment effects on the two outcomes by specifying a uniform prior distribution over all possible range of correlations, $Unif(-1, 1)$.

We chose the above scenarios as an illustration for modelling the correlation. Other scenarios may also be considered. For example, there may be a rationale for modelling the correlation on the negative scale, $Unif(-1, 0)$, reflecting both worse effectiveness and increased adverse events.

5.4.3 Results

Estimates of the between-studies standard deviations along with the between-studies correlation parameters from fitting all four models, assuming homogeneity of between-studies variances across treatment contrasts, are presented in Table 11. The estimates of the between-studies correlation were similar across the models (apart form model 1 which assumed no correlation); they were stable to prior specification of the within-study correlation, with a median estimate close to 0.9 across alternative prior specifications. Overall, the results suggest borrowing of information from modelling the between-studies correlation. The resultant impact of being able to borrow information across outcomes is a decrease in the between-studies standard deviation from 0.84 (95% CrI 0.17 to 2.08) on the log odds ratio scale in Model 1 (univariate NMA) to 0.78 (95% CrI 0.15 to 1.87) in Model 4 (bvNMA model assuming $\rho_w \sim U(-1, 1)$) for discontinuation due to adverse events where we have relatively few studies reporting this outcome. This represents a small reduction in heterogeneity on the log odds ratio scale which might indicate a meaningful reduction in between-studies heterogeneity on odds ratio scale. For proportion relapse free where borrowing of strength did not take place, the between-studies standard deviation remained relatively unchanged in both univariate and bivariate models. All results are based on 70,000 iterations after a burn-in of 20,000 and applying thinning every 10 iteration.

Table 11: Estimates of the between-studies correlation and the between-studies standard deviation parameters (medians and 95% CrIs) obtained from fitting univariate and multivariate network meta-analysis models to the multiple sclerosis data.

model	τ_1	τ_2	ρ_b
Model 1: $\rho_w = 0$, ρ_b=0	0.40 (0.20, 0.78)	0.84 (0.17, 2.08)	0.00 (0.00, 0.00)
Model 2: $\rho_w = 0$	0.40 (0.18, 0.79)	0.77 (0.17, 1.82)	0.90 (-0.14, 1.00)
Model 3: $\rho_w \sim Unif(0, 1)$	0.43 (0.20, 0.83)	0.74 (0.08, 1.77)	0.83 (-0.50, 1.00)
Model 4: $\rho_w \sim Unif(-1, 1)$	0.42 (0.20, 0.82)	0.78 (0.15, 1.87)	0.87 (-0.31, 1.00)

Figure 6 and Figure 7 are summary forest plots of intervention effects relative to placebo on proportion relapse free and discontinuation due to adverse events respectively. All models produced estimates of treatment effect with a similar degree of precision for proportion relapse free in Figure 6 but the multivariate models produced more precise estimates of treatment effect on discontinuation than the univariate model.

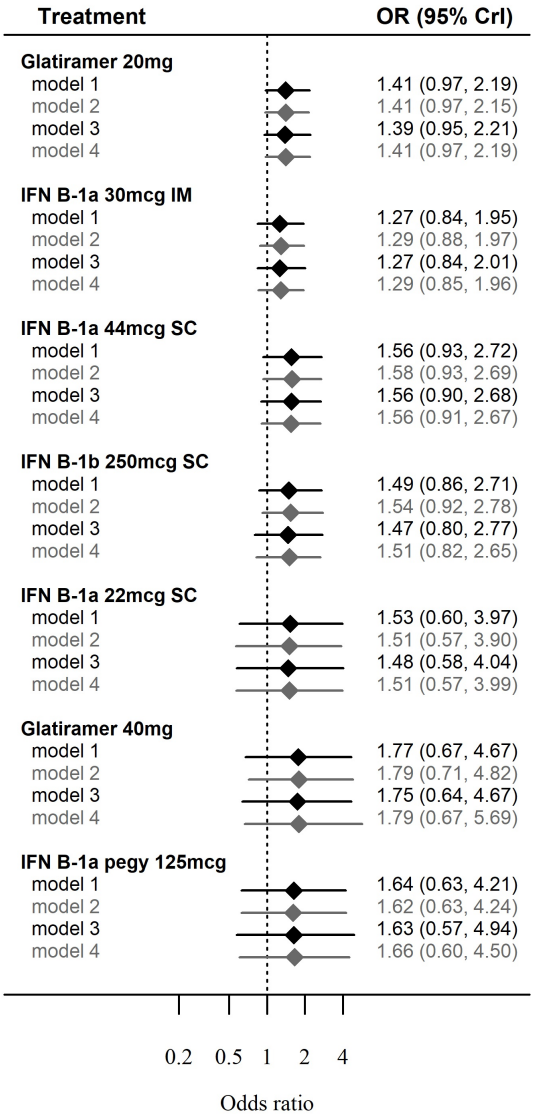


Figure 6: Forest plot intervention effects relative to placebo on proportion of patients remaining relapse free estimated from bivariate (Models 2 to 4) and univariate (model 1) random-effects NMA models. Model 2 assumes zero within study-correlation whilst Models 3 and 4 assume $Unif(0,1)$ and $Unif(-1,1)$ prior distributions for the within-study correlation common to all-studies.

Table 12 presents pairwise odds ratios comparing all 8 interventions relative to one another, on the proportion relapse free from Model 1 (uvNMA) and Model 4 (bvNMA assuming $\rho_w \sim Unif(-1,1)$). The corresponding pairwise odds ratios for discontinuation due to adverse events are presented in Table 13. The upper triangle of Table 12 and Table 13 display the bvNMA results whilst the lower triangle are the estimates from the uvNMA. Interventions are consecutively arranged top-down and left-right such that rows represent comparator interventions for estimates in the upper half of the table whilst columns the comparators for estimates presented in the lower half. For example, the estimated odds ratios for ifnb1a44 versus ifnb1a30 on proportion relapse free

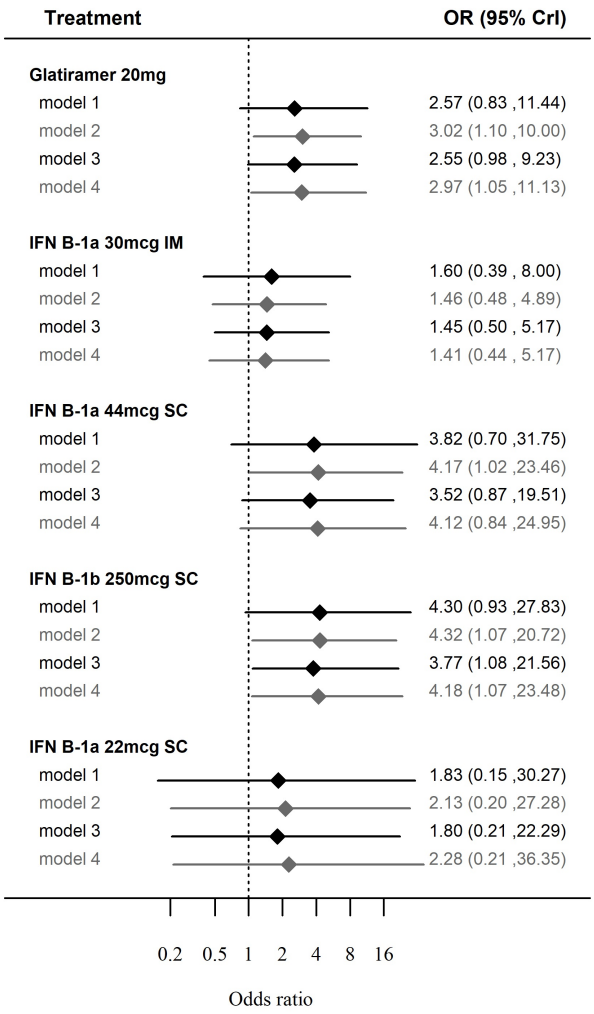


Figure 7: Forest plot intervention effects relative to placebo on discontinuation to due to adverse effect of treatment from bivariate random-effect network meta-analyses (Models 2 to 4) and univariate network meta-analysis (model 1). Model 2 assumes zero within-study correlation whilst Models 3 and 4 assume $Unif(0, 1)$ and $Unif(-1, 1)$ prior distributions for the within-study correlation common to all-studies.

are 1.24 (95%CrI: 0.69, 2.15) and 1.22 (95% CrI: 0.67, 2.21) from univariate and bivariate models respectively (Table 12). The corresponding estimates for ifnb1a44 vs ifnb1a30 on discontinuation due to adverse events are 2.32 (95% CrI: 0.24, 28.05) and 2.84 (95% CrI: 0.45, 24.53) in the univariate and bvNMA models respectively (Table 13). In the network for proportion relapse free where we have a relatively large body of evidence, including head-to-head comparisons between 44mcg SC versus ifnb1a30 (Figure 5) from two studies [98, 100], univariate and bivariate models produced broadly identical estimate of the treatment effect. In contrast to the proportion relapse free, few studies reported discontinuation due to adverse events including the two studies that directly compared ifnb1a44 vs ifnb1a30. The consequent borrowing of information between the two outcomes in the bvNMA model suggest a worse but more precise estimate of the side-effect profile for ifnb1a44 vs ifnb1a30 in comparison with the estimate from the uvNMA model. The comparison between ifnb1a44 vs ifnb1a30 illustrates the situation where there is direct and indirect evidence on the pairwise contrasts on one outcome but only indirect evidence for the same contrasts on another. In such situation, it may be advantageous to fit a multivariate NMA and borrow information from the network with direct and indirect evidence on pairwise to inform estimates of the corresponding contrasts in the "indirect evidence only" network.

Table 12: Bivariate and univariate NMA estimates of all pairwise odd-ratios comparing the effectiveness of 8 interventions relative to one another on proportion relapse free. The upper triangle displays the bvNMA results whilst the lower triangle are the estimates from the uvNMA

Treatment	Placebo	Glatiramer 1	IFN 1	IFN 2	IFN 3	IFN 4	Glatiramer 2	IFN 5
Placebo	NA	1.41 (0.97, 2.19)	1.29 (0.85, 1.96)	1.56 (0.91, 2.67)	1.51 (0.82, 2.65)	1.51 (0.57, 3.99)	1.79 (0.67, 5.69)	1.66 (0.60, 4.50)
Glatiramer 1	1.41 (0.97, 2.19)	NA	0.91 (0.53, 1.42)	1.11 (0.61, 1.86)	1.07 (0.59, 1.81)	1.08 (0.37, 3.00)	1.29 (0.41, 3.82)	1.18 (0.38, 3.34)
IFN 1	1.27 (0.84, 1.95)	0.90 (0.54, 1.44)	NA	1.22 (0.67, 2.21)	1.18 (0.66, 2.05)	1.19 (0.42, 3.26)	1.41 (0.49, 4.39)	1.31 (0.42, 3.76)
IFN 2	1.56 (0.93, 2.72)	1.12 (0.62, 1.87)	1.24 (0.69, 2.15)	NA	0.97 (0.49, 1.87)	0.99 (0.37, 2.57)	1.16 (0.39, 3.64)	1.07 (0.32, 3.19)
IFN 3	1.49 (0.86, 2.71)	1.07 (0.60, 1.80)	1.18 (0.70, 2.03)	0.96 (0.51, 1.82)	NA	1.02 (0.35, 3.02)	1.19 (0.36, 4.01)	1.09 (0.34, 3.42)
IFN 4	1.53 (0.60, 3.97)	1.09 (0.40, 2.89)	1.20 (0.45, 3.27)	0.97 (0.38, 2.51)	1.01 (0.36, 2.85)	NA	1.19 (0.30, 4.96)	1.11 (0.26, 4.29)
Glatiramer 2	1.77 (0.67, 4.67)	1.26 (0.42, 3.39)	1.41 (0.48, 3.92)	1.13 (0.37, 3.44)	1.18 (0.37, 3.59)	1.16 (0.28, 4.46)	NA	0.91 (0.22, 3.71)
IFN 5	1.64 (0.63, 4.21)	1.17 (0.39, 3.07)	1.29 (0.44, 3.56)	1.05 (0.34, 3.12)	1.10 (0.37, 3.11)	1.07 (0.28, 3.88)	0.92 (0.24, 3.69)	NA

Glatiramer 1 = Glatiramer 20mg, Glatiramer 2 = Glatiramer 40mg, IFN 1 = IFN B-1a 30mcg IM,
IFN 2 = IFN B-1a 44mcg SC, IFN 3 = IFN B-1b 250mcg SC, IFN 4 = IFN B-1a 22mcg SC, IFN 5 = IFN B-1a pegy 125mcg

Table 13: Bivariate and univariate NMA estimates of all pairwise odd-ratios comparing the effectiveness of 6 interventions relative to one another on discontinuation due to adverse events. The upper triangle displays the bvNMA results whilst the lower triangle are the estimates from the uvNMA.

Treatment	Placebo	Glatiramer 1	IFN 1	IFN 2	IFN 3	IFN 4
Placebo	NA	2.97 (1.05, 11.13)	1.41 (0.44, 5.17)	4.12 (0.84, 24.95)	4.18 (1.07, 23.48)	2.28 (0.21, 36.35)
Glatiramer 1	2.57 (0.83, 11.44)	NA	0.49 (0.08, 2.12)	1.43 (0.30, 5.77)	1.43 (0.36, 5.69)	0.79 (0.07, 9.29)
IFN 1	1.60 (0.39, 8.00)	0.63 (0.08, 3.79)	NA	2.84 (0.45, 24.53)	2.92 (0.62, 17.52)	1.62 (0.11, 27.86)
IFN 2	3.82 (0.70, 31.75)	1.46 (0.24, 8.62)	2.32 (0.24, 28.05)	NA	1.03 (0.18, 7.40)	0.56 (0.06, 5.17)
IFN 3	4.30 (0.93, 27.83)	1.63 (0.36, 8.39)	2.57 (0.46, 19.83)	1.08 (0.12, 10.84)	NA	0.59 (0.03, 8.06)
IFN 4	1.83 (0.15, 30.27)	0.71 (0.05, 10.27)	1.13 (0.06, 25.67)	0.48 (0.04, 5.11)	0.44 (0.02, 7.68)	NA

Glatiramer 1 = Glatiramer 20mg, Glatiramer 2 = Glatiramer 40mg, IFN 1 = IFN B-1a 30mcg IM,
IFN 2 = IFN B-1a 44mcg SC, IFN 3 = IFN B-1b 250mcg SC, IFN 4 = IFN B-1a 22mcg SC.

5.5 Discussion of mvNMA methodology and ongoing research

Network meta-analysis in the univariate form already leads to borrowing of information from other treatment contrasts (compared to analysing the data in each contrast separately using pairwise meta-analysis). Therefore, we may not observe additional borrowing of strength when using multivariate meta-analysis methods. However, when a network on one outcome is sparse, such as the network on discontinuation due to adverse events in the RRMS example, the borrowing of information from other outcomes (in this case proportion relapse free) can reduce uncertainty around pooled estimates. Moreover, a joint posterior distribution for the correlated average treatment effects, used in health economic decision models may also result in more appropriate estimates of the cost-effectiveness models [101]. Another novel application of bivariate NMA is in the area of surrogate endpoint evaluation developed by Bujkiewicz et al [48], as discussed in Section 3.8. Such methodology will be useful when an effect on a decision endpoint is not available from any study for one or more interventions of interest and evidence is available on a surrogate endpoint. Predictions then can be made about the effect on the outcome of interest for specific intervention.

Other methods for multivariate NMA have also been reported. A network meta-analysis of multiple outcomes with individual patient data has also been proposed by Hong *et al.* [92] under both contrast-based and arm-based parameterizations (where by contrast-based approach we mean an approach aiming to estimate the average relative treatment effect of an experimental treatment compared control, whilst the arm-based approach models absolute treatment effects on individual treatment arms), and Hong *et al.* [93] developed a Bayesian framework for multivariate network meta-analysis. These multivariate network meta-analysis models are based on the assumption of consistency of treatment effects in the network (as described in Section 5.1), extending the approach introduced by Lu and Ades [102]. Jackson et al. [94] developed a mvNMA model that allows for the inconsistency, i.e the assumption of consistency in the network of treatment effects is not satisfied and additional inconsistency term is modelled to account for this. In other words, different forms of direct and indirect evidence may not agree, even after taking between-studies heterogeneity into account. The inconsistency is taken into account using a design-by-treatment interaction, where design is the composition of a trial in the network such as A vs B or three-arm trial of treatments C, D and E compared against one another. The aim of all these methods is to obtain pooled effects for correlated multiple outcomes and treatment contrasts in a network of studies.

6 Discussion and extensions

6.1 Normality assumption for random effects

The methods considered here are models with random effects to reflect the assumption that the modelled true treatment effects are different (but still similar) between the studies. The differences in the true effects may be due to the varying populations, different treatments under investigation in those studies or perhaps heterogeneity in the definitions of the outcomes and follow-up length [42]. Typically, a normal distribution of the between-studies random effects is assumed to reflect the similarity of the effects. The assumption that the true treatment effects on both outcomes (such as log odds ratios on two outcomes) are normally distributed may, however, not always be reasonable [103]. When dealing with departures from normality of the modelled data, this assumption can lead to limitations of modelling and restricted inferences [104], especially when making predictive inferences about the potential magnitude of effects in new studies or populations. For example, as discussed by Marshall and Spiegelhalter, inadequate use of normality assumption about the random effects may lead to "overshrinkage" of the true effects and hence to misleading inferences [105].

One way of relaxing this assumption is to use a t-distribution as recommended, for example, by Smith, Spiegelhalter and Thomas [106] or by Lee and Thompson [104]. In contrast to the normal distribution, the t-distribution gives more weight in the tails which is more likely to be better at modelling extreme effects such as outlying observations [105]. If the distribution of the data is, for example, bimodal or skewed, other approaches can be investigated such as a convolution of normal distributions [107] or skewed t-distribution as proposed by Lee and Thompson [104].

6.2 Binary, nested and mutually exclusive outcomes

Often, data are collected from multiple binary outcomes. A standard approach is to model such data on the log odds ratio scale (or log odds scale for single arm data) using normal approximation, as we did in this TSD, for example for the ACR20 response in the example in rheumatoid arthritis or for the multivariate NMA example in multiple sclerosis. Exact methods, using binomial likelihood, may be preferable, in particular when the number of studies in the meta-analysis is small or probability of an event is close to 0 or 1. In the bivariate case, however, the model is limited to one that ignores the within study correlation, which may lead to biased results [43]. This issue is further discussed by Chen et al. [108] who also propose alternative techniques for modelling correlated binomial data. However, binomial models can be used when outcomes are mutually exclusive or have a "is-subset-of" relationship, as described by Trikalinos et al. [109], who also discuss multivariate meta-analytic models for categorical data.

6.3 Bayesian versus frequentist methods and related software

We used Bayesian approach to modelling multiple outcomes. Bayesian methods are most suited to flexibly model the uncertainty for the multiple parameters in such models. Multivariate meta-analytic models can also be implemented using frequentist approach. In Stata, for example, the model can be implemented using the command `mvmeta` [110]. Bujkiewicz et al. [12] also provide additional Stata code for cross-validation procedure for surrogate endpoint evaluation which can be used alongside of the `mvmeta` command. However, only BRMA in the standard form is available in Stata (not BRMA PNF or model by Daniels and Hughes for surrogate endpoints). Multivariate meta-analysis can also be performed in R using `mvmeta` package [111]. Software for multivariate network meta-analysis is available in R, which was developed by Jackson et al. [94] for model accounting for inconsistency, but it also provides estimates from a model assuming consistency. Polanin et al provide review of other packages for meta-analysis in R which include multivariate approaches [112].

6.4 Other application areas

As discussed in the introduction, multivariate meta-analysis has a wide range of application areas. In addition to synthesising data on clinical effectiveness of treatments, in can also include data of effects of treatment on adverse events (as we shown in our example of applying multivariate network meta-analysis in multiple sclerosis). Equally, multivariate meta-analysis can be used for obtaining estimates of the treatment effect on a common scale of HRQoL, by modelling such estimates jointly or through mapping techniques. Bujkiewicz et al [57] showed that when use of multivariate meta-analysis leads to more precise estimates of disease-specific measures of HRQoL, then mapping such estimates onto a common scale, such as EQ-5D, will also result in these estimates on the common scale obtained with increased precision, potentially leading to reduced uncertainty when making decisions based on cost-effectiveness models (or even change in the decision about reimbursement of a new health technology). Multivariate meta-analysis can also be used to model jointly estimates of clinical effectiveness and HRQoL outcomes (both disease-specific and generic such as EQ-5D) which can lead to obtaining not only pooled estimates on all outcomes but also predicting estimates for desirable scale in a similar way as is done for surrogate endpoints, where treatment effect on the final clinical outcome is predicted from the treatment effect on the surrogate endpoint. Alternative mapping techniques that involve a form of multivariate meta-analysis were also proposed by Lu et al. [113].

Dias et al. discuss a number of other applications of multivariate meta-analysis in the decision making context [114]. They describe a number of scenarios where evidence structure can be modelled by capturing functional relationship between outcomes. These scenarios include, for example, a "chain of evidence" structures that lead to relationships between outcomes when modelling the natural history of a disease or use of multiple time points of follow-up which can generate different network structures of evidence at different time points.

7 Summary and recommendations for use of multivariate meta-analysis to inform decision modelling

In health technology assessment, decisions are often based on complex cost-effectiveness models which require numerous input parameters. Multivariate meta-analysis can facilitate inclusion of broader range of data in the analysis, potentially avoiding excluding valuable evidence. In the introduction to this TSD we discussed two examples of use of bivariate meta-analysis in technology appraisals by NICE, where this methodology allowed for inclusion of broader evidence base or accounting for relevant correlations between treatment effects on two outcomes. In our illustrative example in rheumatoid arthritis, we demonstrated how multivariate meta-analysis can help include a larger number of studies when estimating treatment effect. We showed that this approach can also potentially lead to the effectiveness estimates obtained with higher precision, which was the case in particular when external data were incorporated into the analysis in the form of informative prior distributions, thus highlighting the advantage of the Bayesian approach to the analysis. This may be particularly beneficial when interest lies in the treatment effect on outcomes for which trials have not been powered, such as adverse events. Our illustration of applying multivariate network meta-analysis to the data in multiple sclerosis showed that modeling the effects on these outcomes together with the effects on primary outcome can result in borrowing of information, increasing the precision of the underpowered outcome. Multivariate meta-analytic techniques are particularly useful in modelling surrogate endpoints or similar correlated outcomes. When treatment effects on outcomes required for populating a decision model are not reported for a treatment under investigation, these effects can be predicted from other outcomes using multivariate meta-analytic framework, as we have shown in our example in multiple sclerosis.

7.1 Recommendations

When conducting an analysis for HTA decision making or other areas of evidence based medicine the following aspects of multiple outcomes should be considered

- At the scoping stage, a careful consideration should be made when deciding on the outcomes of interest to ensure that sources of relevant evidence are not missed.
- Correlated treatment effects on multiple endpoints should be analysed jointly when the synthesis aims to obtain robust pooled estimates of the treatment effect on one or more of these outcomes. This approach has benefits in particular when
  - there are relatively few studies and borrowing of strength from other outcomes is more likely to have an impact on the final estimate (relevant methods and examples are described in Sections 2.1 and 4)
  - a proportion of studies identified through the systematic review do not report an outcome of interest (or missing data suggests outcome reporting bias), in which case borrowing of strength is also more likely to have impact on the uncertainty of the estimates, but also a more appropriate estimates can be obtained in the presence of outcome reporting bias (see methods and examples described in Sections 2.1 and 4)
  - when multiple treatment comparisons are needed and only a few studies reporting treatment effect on an outcome of interest are available for some of the treatments; then taking into account the network structure of the data will help borrow information from data on other treatments and outcomes (see methods of multivariate NMA and an example described in Section 5)
  - a health-economic model takes inputs from multiple outcomes which are propagated through the model using multivariate posterior distribution and accounting for the correlation between the outcomes may reduce bias of the resulting cost-effectiveness estimates (see methods and examples described in Section 2.1 for pairwise meta-analysis

of treatment effects on two outcomes, Section 4 for pairwise meta-analysis of treatment effects on multiple (at least three) outcomes and Section 5 for multiple comparisons (network) meta-analysis of treatment effects on two or more outcomes).

- When study (or studies) investigating a new treatment does not report the treatment effect on the final clinical outcome needed to make a decision by, for example, informing a health-economic decision model, but treatment effect on another outcome (or multiple outcomes), such as a surrogate endpoint, is reported, then joint synthesis of the treatment effects on the two (or multiple) outcomes can be preformed. Such joint synthesis can be used
  - to assess the strength of the surrogate relationship between the treatment effects on the surrogate endpoint (or multiple surrogate endpoints) and the final outcome using data from other studies reporting the treatment effects on both outcomes (see relevant methods in Section 3.4 or for multiple surrogate endpoints in Section 4)
  - and to predict the treatment effect on the final clinical outcome from an observed treatment effect on the surrogate endpoint using data from the study (or multiple studies) reporting the treatment effect on surrogate endpoint along with data from other studies reporting the treatment effects on both outcomes (which provide evidence base for the surrogate relationship between the treatment effects on the two outcomes (see Sections 3.6.2 and 3.1.4).

7.2 Conclusions

Multivariate meta-analysis provides a range of opportunities for effective synthesis of diverse sources of evidence. This TSD offers recommendations on the use of different parameterisations of multivariate meta-analysis in different settings of application. When used appropriately, the methods can help model available evidence more effectively in future HTAs and other settings of evidence based decision making. In the context of HTA, such more effective modelling should increase the likelihood of making appropriate reimbursement decisions.

8 Disclosures

Keith Abrams has served as a paid consultant, providing methodological and strategic HTA advice to the pharmaceutical industry and international HTA agencies, as well as being a partner and director of Visible Analytics Limited, a HTA consultancy company. He has received research funding from Association of the British Pharmaceutical Industry (ABPI), European Federation of Pharmaceutical Industries & Associations (EFPIA), Pfizer and Sanofi.
Sylwia Bujkiewicz has served as a paid consultant, providing methodological advice to Roche and received research funding from EFPIA and Johnson & Johnson.

9 References

[1] Dan Jackson, Richard Riley, and Ian R White. Multivariate meta-analysis: potential and promise. *Statistics in medicine*, 30(20):2481–2498, 2011.

[2] H. C. van Houwelingen, L. R. Arends, and T. Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21(4):589–624, 2002.

[3] I.-S. Nam, K. Mengersen, and P. Garthwaite. Multivariate meta-analysis. *Statistics in Medicine*, 22(14):2309–2333, 2003.

[4] R. D. Riley, K. R. Abrams, P. C. Lambert, A. J. Sutton, and J. R. Thompson. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, 26(1):78–97, 2007.

[5] S. Bujkiewicz, J. R. Thompson, A. J. Sutton, N. J. Cooper, M. J. Harrison, D. P. M. Symmons, and K. R. Abrams. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Statistics in Medicine*, 32(22):3926–3943, 2013.

[6] Y. Wei and J. Higgins. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(17):2911–2934, 2013.

[7] T. Burzykowski, G. Molenberghs, and M. Buyse. *The evaluation of surrogate endpoints*. Springer, 2006.

[8] M. J. Daniels and M. D. Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, 16(17):1965–1982, 1997.

[9] M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1):49–67, 2000.

[10] T. Burzykowski, G. Molenberghs, M. Buyse, H. Geys, and D. Renard. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(4):405–422, 2001.

[11] L. A. Renfro, Q. Shi, D. J. Sargent, and B. P. Carlin. Bayesian adjusted r2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Statistics in Medicine*, 31(8):743–761, 2012.

[12] S. Bujkiewicz, J. R. Thompson, E. Spata, and K. R. Abrams. Uncertainty in the Bayesian meta-analysis of normally distributed surrogate endpoints. *Statistical methods in medical research*, 26(5):2287–2318, 2017.

[13] S. Bujkiewicz, J. R. Thompson, R. D. Riley, and K. R. Abrams. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Statistics in Medicine*, 35(7):1063–1089, 2016.

[14] Richard D Riley, Dan Jackson, Georgia Salanti, Danielle L Burke, Malcolm Price, Jamie Kirkham, and Ian R White. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *bmj*, 358:j3932, 2017.

[15] Ashley P Jones, Richard D Riley, Paula R Williamson, and Anne Whitehead. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical Trials*, 6(1):16–27, 2009.

[16] John R Thompson, Cosetta Minelli, Keith R Abrams, Martin D Tobin, and Richard D Riley. Meta-analysis of genetic studies using mendelian randomization—a multivariate approach. *Statistics in medicine*, 24(14):2241–2254, 2005.

[17] Fibrinogen Studies Collaboration. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Statistics in Medicine*, 28(8):1218–1237, 2009.

[18] Kym IE Snell, Harry Hua, Thomas PA Debray, Joie Ensor, Maxime P Look, Karel GM Moons, and Richard D Riley. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *Journal of clinical epidemiology*, 69:40–50, 2016.

[19] Richard D Riley, Eleni G Elia, Gemma Malin, Karla Hemming, and Malcolm P Price. Multivariate meta-analysis of prognostic factor studies with multiple cut-points and/or methods of measurement. *Statistics in medicine*, 34(17):2481–2496, 2015.

[20] Jamie J Kirkham, Richard D Riley, and Paula R Williamson. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in medicine*, 31(20):2179–2195, 2012.

[21] Jelena Stevanović, Petros Pechlivanoglou, Marthe A Kampinga, Paul FM Krabbe, and Maarten J Postma. Multivariate meta-analysis of preference-based quality of life values in coronary heart disease. *PloS one*, 11(3):e0152030, 2016.

[22] David J Spiegelhalter, Keith R Abrams, and Jonathan P Myles. *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons, 2004.

[23] David J Spiegelhalter and Nicola G Best. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in medicine*, 22(23):3687–3709, 2003.

[24] Deborah Ashby and Adrian FM Smith. Evidence-based medicine as Bayesian decision-making. *Statistics in medicine*, 19(23):3291–3305, 2000.

[25] Catriona McDaid, S Griffin, H Weatherly, Kate Durée, M Van der Burgt, S Van Hout, J Akers, R Davies, Mark Sculpher, and Marie Westwood. Continuous positive airway pressure devices for the treatment of obstructive sleep apnoea–hypopnoea syndrome: a systematic review and economic analysis. *Health Technology Assessment*, 13(4), 2009.

[26] National Institute for Health and Care Excellence. Continuous positive airway pressure for the treatment of obstructive sleep apnoea/hypopnoea syndrome. *Available at: https://www.nice.org.uk/guidance/ta139, Accessed May 2017*, 2008.

[27] Jhuti G Rice S Spackman E Sideris E et al. Corbett M, Soares M. Tumour necrosis factor-α inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis: a systematic review and economic evaluation. *Health Technology Assessment*, 20(9), 2016.

[28] National Institute for Health and Care Excellence. Tnf-alpha inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis. 2016.

[29] Sze Huey Tan, Keith R Abrams, and Sylwia Bujkiewicz. Bayesian multiparameter evidence synthesis to inform decision making: A case study in metastatic hormone-refractory prostate cancer. *Medical Decision Making*, 38(7):834–848, 2018.

[30] R. S. Taylor and J. Elston. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of uk health technology assessment reports. *Health Technology Assessment*, 13(6), 2009.

[31] Oriana Ciani, Marc Buyse, Michael Drummond, Guido Rasi, Everardo D Saad, and Rod S Taylor. Time to review the role of surrogate end points in health policy: state of the art and the way forward. *Value in Health*, 20(3):487–495, 2017.

[32] European Medicines Agency. Final report from the emea/chmp-think-tank group on innovative drug development. *available at http://bit.ly/2rCmLAH*, 2007.

[33] Stephen Joel Coons. The fda's critical path initiative: a brief introduction. *Clinical therapeutics*, 31(11):2572–2573, 2009.

[34] Chul Kim and Vinay Prasad. Cancer drugs approved on the basis of a surrogate end point and subsequent overall survival: an analysis of 5 years of us food and drug administration approvals. *JAMA internal medicine*, 175(12):1992–1994, 2015.

[35] European Medicines Agency. European public assessment report of lartruvo (olaratumab). *available at https://bit.ly/2VnLv9Z*, 2016.

[36] FDA Center for Drug Evaluation and Research. Regulatory decision to withdraw avastin (bevacizumab) first-line metastatic breast cancer indication. *available at https://bit.ly/2HpeLrR*, 2010.

[37] National Institute for Health and Care Excellence. Venetoclax in combination with rituximab for treating relapsed or refractory chronic lymphocytic leukaemia [id1097]. *available at https://www.nice.org.uk/guidance/ta561*, 2019.

[38] EUnetHTA. Endpoints used in relative effectiveness assessment of pharmaceuticals - surrogate endpoints. *available at http://bit.ly/2rBRXzX*, 2013.

[39] National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. *available at https://www.nice.org.uk/process/pmg9/chapter/foreword*, 2013.

[40] David Spiegelhalter, Andrew Thomas, Nicky Best, and Dave Lunn. Winbugs user manual, 2003.

[41] Sofia Dias, Nicky J Welton, Alex J Sutton, and AE Ades. Nice dsu technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. 2011, available from http://www.nicedsu.org.uk.

[42] Julian PT Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

[43] Richard D Riley. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):789–811, 2009.

[44] RD Riley, MJ Price, D Jackson, M Wardle, F Gueyffier, J Wang, Jan A Staessen, and IR White. Multivariate meta-analysis using individual participant data. *Research synthesis methods*, 6(2):157–174, 2015.

[45] Yinghui Wei and Julian PT Higgins. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(7):1191–1205, 2013.

[46] Richard D Riley, John R Thompson, and Keith R Abrams. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, 9(1):172–186, 2007.

[47] D. L. Burke, S. Bujkiewicz, and R. D. Riley. Bayesian bivariate meta-analysis of correlated effects: Impact of the prior distributions on the between-study correlation, borrowing of strength, and joint inferences. *Statistical methods in medical research*, 27(2):428–450, 2018.

[48] Sylwia Bujkiewicz, Dan Jackson, John R Thompson, Rebecca M Turner, Nicolas Städler, Keith R Abrams, and Ian R White. Bivariate network meta-analysis for surrogate endpoint evaluation. *Statistics in medicine*, 38(18):3322, 2019.

[49] Krishnan Bhaskaran and Liam Smeeth. What is the difference between missing completely at random and missing at random? *International journal of epidemiology*, 43(4):1336–1339, 2014.

[50] Helen A Dakin, Nicky J Welton, AE Ades, Sarah Collins, Michelle Orme, and Steven Kelly. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Statistics in medicine*, 30(20):2511–2535, 2011.

[51] David J Spiegelhalter. Bayesian graphical modelling: a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1):115–133, 1998.

[52] Julian PT Higgins and Anne Whitehead. Borrowing strength from external trials in a meta-analysis. *Statistics in medicine*, 15(24):2733–2749, 1996.

[53] Rebecca M Turner, Jonathan Davey, Mike J Clarke, Simon G Thompson, and Julian PT Higgins. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the cochrane database of systematic reviews. *International journal of epidemiology*, 41(3):818–827, 2012.

[54] Rebecca M Turner, Dan Jackson, Yinghui Wei, Simon G Thompson, and Julian PT Higgins. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in medicine*, 34(6):984–998, 2015.

[55] Suzanne Lloyd, Sylwia Bujkiewicz, Allan J Wailoo, Alex J Sutton, and David Scott. The effectiveness of anti-tnf-α therapies when used sequentially in rheumatoid arthritis patients: a systematic review and meta-analysis. *Rheumatology*, 49(12):2313–2321, 2010.

[56] D Symmons, K Tricker, M Harrison, C Roberts, M Davis, P Dawes, A Hassell, S Knight, D Mulherin, and DL Scott. Patients with stable long-standing rheumatoid arthritis continue to deteriorate despite intensified treatment with traditional disease modifying anti-rheumatic drugs—results of the british rheumatoid outcome study group randomized controlled clinical trial. *Rheumatology*, 45(5):558–565, 2005.

[57] Sylwia Bujkiewicz, John R Thompson, Alex J Sutton, Nicola J Cooper, Mark J Harrison, Deborah PM Symmons, and Keith R Abrams. Use of Bayesian multivariate meta-analysis to estimate the HAQ for mapping onto the EQ-5D questionnaire in rheumatoid arthritis. *Value in Health*, 17(1):109–115, 2014.

[58] AE Ades, G Lu, and JPT Higgins. The interpretation of random-effects meta-analysis in decision models. *Medical Decision Making*, 25(6):646–654, 2005.

[59] Keith R Abrams, Clare L Gillies, and Paul C Lambert. Meta-analysis of heterogeneously reported trials assessing change from baseline. *Statistics in medicine*, 24(24):3823–3844, 2005.

[60] MP Sormani, L Bonzano, L Roccatagliata, GL Mancardi, A Uccelli, and P Bruzzi. Surrogate endpoints for edss worsening in multiple sclerosis a meta-analytic approach. *Neurology*, 75(4):302–309, 2010.

[61] International Conference on Harmonisation guidelines for the conduct of clinical trials for the registration of drugs. E9: statistical principles for clinical trials, https://bit.ly/2ty3rdf. Published 1998. Accessed February 22, 2019.

[62] Heiner C Bucher, Gordon H Guyatt, Deborah J Cook, Anne Holbrook, Finlay A McAlister, Evidence-Based Medicine Working Group, et al. Users' guides to the medical literature: Xix. applying clinical trial results a. how to use an article measuring the effect of an intervention on surrogate end points. *Jama*, 282(8):771–778, 1999.

[63] Marissa N Lassere, Kent R Johnson, Maarten Boers, Peter Tugwell, Peter Brooks, Lee Simon, Vibeke Strand, Philip G Conaghan, Mikkel Ostergaard, Walter P Maksymowych, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *The Journal of rheumatology*, 34(3):607–615, 2007.

[64] Thomas R Fleming and David L DeMets. Surrogate end points in clinical trials: are we being misled? *Annals of internal medicine*, 125(7):605–613, 1996.

[65] Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

[66] Marshall M Joffe and Tom Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538, 2009.

[67] Tyler J VanderWeele. Surrogate measures and consistent surrogates. *Biometrics*, 69(3):561–565, 2013.

[68] Marc Buyse, Geert Molenberghs, Xavier Paoletti, Koji Oba, Ariel Alonso, Wim Van der Elst, and Tomasz Burzykowski. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal*, 58(1):104–132, 2016.

[69] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[70] Marissa N Lassere, Kent R Johnson, Michal Schiff, and David Rees. Is blood pressure reduction a valid surrogate endpoint for stroke prevention? an analysis incorporating a systematic review of randomised controlled trials, a by-trial weighted errors-in-variables regression, the surrogate threshold effect (ste) and the biomarker-surrogacy (biosurrogate) evaluation schema (bses). *BMC medical research methodology*, 12(1):27, 2012.

[71] Institute for Quality and Efficiency in Health Care. Validity of surrogate endpoints in oncology. executive summary of rapid report a10-05, version 1.1., https://www.ncbi.nlm.nih.gov/books/nbk198799/, Published 2011. Accessed December 7, 2018.

[72] A. Alonso, T. Bigirumurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N.J. Perualila, Z. Shkedy, and W. Van der Elst. *Applied Surrogate Endpoint Evaluation Methods with SAS and R.* Chapman & Hall/CRC Biostatistics Series. CRC Press, 2016.

[73] Tasos Papanikos, John R Thompson, Keith R Abrams, Nicolas Städler, Oriana Ciani, Rod Taylor, and Sylwia Bujkiewicz. A Bayesian hierarchical meta-analytic method for modelling surrogate relationships that vary across treatment classes. *arXiv:1905.07194*, 2019, available from https://arxiv.org/pdf/1905.07194.pdf.

[74] John B Copas, Dan Jackson, Ian R White, and Richard D Riley. The role of secondary outcomes in multivariate meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1177–1205, 2018.

[75] Dan Jackson, Ian R White, Malcolm Price, John Copas, and Richard D Riley. Borrowing of strength and study weights in multivariate and network meta-analysis. *Statistical methods in medical research*, 26(6):2853–2868, 2017.

[76] Orestis Efthimiou, Thomas PA Debray, Gert van Valkenhoef, Sven Trelle, Klea Panayidou, Karel GM Moons, Johannes B Reitsma, Aijing Shang, Georgia Salanti, and GetReal Methods Review Group. Getreal in network meta-analysis: a review of the methodology. *Research synthesis methods*, 7(3):236–263, 2016.

[77] Richard D Riley, Joie Ensor, Dan Jackson, and Danielle L Burke. Deriving percentage study weights in multi-parameter meta-analysis models: with application to meta-regression, network meta-analysis and one-stage individual participant data models. *Statistical methods in medical research*, 27(10):2885–2905, 2018.

[78] Shaun Seaman, John Galati, Dan Jackson, John Carlin, et al. What is meant by "missing at random"? *Statistical Science*, 28(2):257–268, 2013.

[79] Richard M Nixon, Stephen W Duffy, and Guy RK Fender. Imputation of a true endpoint from a surrogate: application to a cluster randomized controlled trial with partial information on the true endpoint. *BMC medical research methodology*, 3(1):17, 2003.

[80] Ralph B D'Agostino. Debate: The slippery slope of surrogate outcomes. *Trials*, 1(2):76, 2000.

[81] Jamie J Kirkham, Kerry M Dwan, Douglas G Altman, Carrol Gamble, Susanna Dodd, Rebecca Smyth, and Paula R Williamson. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *Bmj*, 340:c365, 2010.

[82] Kerry Dwan, Douglas G Altman, Mike Clarke, Carrol Gamble, Julian PT Higgins, Jonathan AC Sterne, Paula R Williamson, and Jamie J Kirkham. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. *PLoS medicine*, 11(6):e1001666, 2014.

[83] Victor G De Gruttola, Pamela Clax, David L DeMets, Gregory J Downing, Susan S Ellenberg, Lawrence Friedman, Mitchell H Gail, Ross Prentice, Janet Wittes, and Scott L Zeger. Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a national institutes of health workshop. *Controlled clinical trials*, 22(5):485–502, 2001.

[84] Jane Xu and Scott L Zeger. The evaluation of multiple surrogate endpoints. *Biometrics*, 57(1):81–87, 2001.

[85] William A O'brien, Pamela M Hartigan, David Martin, James Esinhart, Andrew Hill, Sharon Benoit, Marc Rubin, Michael S Simberkoff, John D Hamilton, and Veterans Affairs Cooperative Study Group on AIDS. Changes in plasma hiv-1 rna and cd4+ lymphocyte counts and the risk of progression to aids. *New England Journal of Medicine*, 334(7):426–431, 1996.

[86] John W Mellors, Alvaro Munoz, Janis V Giorgi, Joseph B Margolick, Charles J Tassoni, Phalguni Gupta, Lawrence A Kingsley, John A Todd, Alfred J Saah, Roger Detels, et al. Plasma viral load and cd4+ lymphocytes as prognostic markers of hiv-1 infection. *Annals of internal medicine*, 126(12):946–954, 1997.

[87] MP Sormani, DK Li, P Bruzzi, B Stubinski, P Cornelisse, S Rocak, and N De Stefano. Combined mri lesions and relapses as a surrogate for disability in multiple sclerosis. *Neurology*, 77(18):1684–1690, 2011.

[88] Luca Pozzi, Heinz Schmidli, and David I Ohlssen. A Bayesian hierarchical surrogate outcome model for multiple sclerosis. *Pharmaceutical statistics*, 15(4):341–348, 2016.

[89] O. Efthimiou, D. Mavridis, A. Cipriani, S. Leucht, P. Bagos, and G. Salanti. An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. *Statistics in Medicine*, 33(13):2275–2287, 2014.

[90] F. A. Achana, N. J. Cooper, S. Bujkiewicz, S. J. Hubbard, D. Kendrick, D. R. Jones, and A. J. Sutton. Network meta-analysis of multiple outcome measures accounting for borrowing of information across outcomes. *BMC medical research methodology*, 14(1):92, 2014.

[91] Orestis Efthimiou, Dimitris Mavridis, Richard D Riley, Andrea Cipriani, and Georgia Salanti. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics*, 16(1):84–97, 2014.

[92] H. Hong, H. Fu, K. L. Price, and B. P. Carlin. Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment. *Statistics in Medicine*, 34(20):2794–2819, 2015.

[93] Hwanhee Hong, Haitao Chu, Jing Zhang, and Bradley P Carlin. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research synthesis methods*, 7(1):6–22, 2016.

[94] D. Jackson, S. Bujkiewicz, M. Law, R. D. Riley, and I. R. White. A matrix-based method of moments for fitting multivariate network meta-analysis models with multiple outcomes and random inconsistency effects. *Biometrics*, pages e–pub ahead of print, 2017.

[95] G. B. Lu and A. Ades. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805, 2009.

[96] John Barnard, Robert McCulloch, and Xiao-Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311, 2000.

[97] Armoiry X Maheswaran H Court R Madan J Kan A Lin S Counsell C Patterson J Rodrigues J Ciccarelli O Fraser H Melendez-Torres GJ, Auguste P and Clarke A. Clinical effectiveness and cost-effectiveness of beta-interferon and glatiramer acetate for treating multiple sclerosis: systematic review and economic evaluation. *NIHR Journals Library*, Health Technology Assessment, 21(52), 2017.

[98] M Etemadifar, M Janghorbani, and V Shaygannejad. Comparison of betaferon, avonex, and rebif in treatment of relapsing–remitting multiple sclerosis. *Acta Neurologica Scandinavica*, 113(5):283–287, 2006.

[99] Clarence Liu and Lance D Blumhardt. Randomised, double blind, placebo controlled study of interferon β-1a in relapsing-remitting multiple sclerosis analysed by area under disability/time curves. *Journal of Neurology, Neurosurgery & Psychiatry*, 67(4):451–456, 1999.

[100] Steven R Schwid and Hillel S Panitch. Full results of the evidence of interferon dose-response-european north american comparative efficacy (evidence) study: a multicenter, randomized, assessor-blinded comparison of low-dose weekly versus high-dose, high-frequency interferon β-1a for relapsing multiple sclerosis. *Clinical therapeutics*, 29(9):2031–2048, 2007.

[101] Sofia Dias, Alex J Sutton, Nicky J Welton, and AE Ades. Nice dsu technical support document 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis: software choices. 2012, available from http://www.nicedsu.org.uk.

[102] G. Lu and A. E. Ades. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, 23(20):3105–3124, 2004.

[103] Dan Jackson and Ian R White. When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, 60(6):1040–1058, 2018.

[104] Katherine J Lee and Simon G Thompson. Flexible parametric models for random-effects distributions. *Statistics in medicine*, 27(3):418–434, 2008.

[105] Emma Clare Marshall and David J Spiegelhalter. Comparing institutional performance using Markov chain Monte Carlo methods. *Statistical analysis of medical data: new developments*, pages 229–249, 1998.

[106] Teresa C Smith, David J Spiegelhalter, and Andrew Thomas. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in medicine*, 14(24):2685–2699, 1995.

[107] Raymond J Carroll, Kathryn Roeder, and Larry Wasserman. Flexible parametric measurement error models. *Biometrics*, 55(1):44–54, 1999.

[108] Yong Chen, Chuan Hong, Yang Ning, and Xiao Su. Meta-analysis of studies with bivariate binary outcomes: a marginal beta-binomial model approach. *Statistics in medicine*, 35(1):21–40, 2016.

[109] Thomas A Trikalinos, David C Hoaglin, and Christopher H Schmid. An empirical comparison of univariate and multivariate meta-analyses for categorical outcomes. *Statistics in medicine*, 33(9):1441–1459, 2014.

[110] Ian R White. Multivariate random-effects meta-analysis. *The Stata Journal*, 9(1):40–56, 2009.

[111] Antonio Gasparrini. Package 'mvmeta', https://cran.r-project.org/web/packages/mvmeta/mvmeta.pdf. 2018.

[112] Joshua R Polanin, Emily A Hennessy, and Emily E Tanner-Smith. A review of meta-analysis packages in r. *Journal of Educational and Behavioral Statistics*, 42(2):206–242, 2017.

[113] Guobing Lu, Daphne Kounali, and AE Ades. Simultaneous multioutcome synthesis and mapping of treatment effects to a common scale. *Value in Health*, 17(2):280–287, 2014.

[114] Nicky J. Welton Jeroen P. Jansen Alexander J. Sutton Sofia Dias, A. E. Ades. *Network meta-analysis for decision making*. John Wiley & Sons, 2018.

[115] Mark J Schervish. *Theory of statistics*. Springer Science & Business Media, 2012.

[116] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[117] Michael J Daniels and Mohsen Pourahmadi. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566, 2002.

[118] José C Pinheiro and Douglas M Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3):289–296, 1996.

A Supplementary materials for application of bivariate meta-analysis to the example in rheumatoid arthritis

In this supplement we provide code for the example in rheumatoid arthritis introduced in Section 2.3. In the next two sections (A.1 and A.2), we provide the details of the implementation of two methods of BRMA: in the standard form and in the product normal formulation, applied to the example in RA. In the last part of the appendix (Section A.3), we provide details of the analysis with informative prior distributions described in Section 2.4.

A.1 WinBUGS code for BRMA in the standard form

A.1.1 Data requirements

In this section we provide WinBUGS code for BRMA for synthesis of data on two outcomes with some data missing. To account for the missing data, we use model (3) described in Section 2.1.3. The data for this model is required from the treatment effect estimates along with the population variances (standard errors squared multiplied by the number of patients) and the number of patients. We include the data in this format below. In the below code, we also give alternative version - with comments in grey that follow the #. When there are no missing data, data would be given as treatment effects and corresponding standard errors, and the first part of the model (assuming the exchangeability of population variances) will be omitted and the within-study covariance matrix Σ constructed using SEs. The data in this format (using SEs) can also be used when missing data are present. However, the exchangeability model for the variances (first 6 lines) will not apply and other approaches to missing data can then be used instead (see Section 2.1.3). **Please ensure that appropriate data are used.**

A.1.2 Code

```
model{

#exchangeability of variances to predict missing SEs

#next 6 lines should be omitted if there are no missing data
#(or using different approach for missing data)

for (k in 1:num){
var[k,1]~dnorm(0,h1)I(0,)
var[k,2]~dnorm(0,h2)I(0,)
}
h1~dgamma(1.0,0.01)
h2~dgamma(1.0,0.01)


#hirerarchical structure of the model:

for(i in 1:num) {
rho_w[i]<-corr_w # we use the same within-study correlation for all studies
prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
#covariance matrix for the j-th study
sigma[i,1,1]<-var[i,1]/n[i,1]
sigma[i,2,2]<-var[i,2]/n[i,2]
sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w[i]
sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w[i]
```

```
#when data on SEs provided (instead of var and n),
#(i.e. above exchangeability of variances model not used) then use this version:
#sigma[i,1,1]<-pow(se[i,1],2)
#sigma[i,2,2]<-pow(se[i,2],2)
#sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w[i]
#sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w[i]

Y[i,1:2] ~ dmnorm(delta[i,1:2],prec_w[i,1:2,1:2]) # within-study model
delta[i,1:2] ~ dmnorm(d[1:2],prec_b[1:2,1:2]) # between-studies model

}

#prior distributions:

for (j in 1:2) {
d[j] ~ dnorm(0.0,0.001)
}

prec_b[1:2,1:2]<-inverse(Cov_b[,])
for (j in 1:2) {
tau[j]~dunif(0,2)
tau.sq[j]<-pow(tau[j],2)
Cov_b[j,j]<-tau.sq[j]
}
Cov_b[1,2]<-rho*tau[1]*tau[2]
Cov_b[2,1]<-Cov_b[1,2]
rho~dunif(-0.999,0.999)
} #model end

#data
#uncertainty represented in the data as population variances and sample sizes
#used to calculate standard errors
#and to allow for exchangeability of the population variances as in Section 2.1.3

If not using this approach, data will have the following structure:
Y[,1]        se[,1]        Y[,2]        se[,2]
...
END

list(num=18, corr_w=0.24) #number of studies and within-study correlation

#n_das Y_das var_das n_haq Y_haq var_haq

n[,1] Y[,1] var[,1] n[,2] Y[,2] var[,2]
26 -1.7 1.683457 26 -0.31 0.4410603
188 -1.6 1.88 188 -0.352 0.388497
810 -1.9 1.96 810 -0.48 0.36
72 -1.47 2.337441 72 NA NA
30 -1.87 1.698847 30 NA NA
18 -2.1 1.532179 18 NA NA
66 -0.98 2.165142 66 NA NA
22 NA NA 22 -0.45 0.443637
110 -1 1.425636 110 NA NA
```

```
331 NA NA 331 -0.12 0.3628
37 NA NA 37 0.15 0.6399
6 -1.17 2.602115 6 NA NA
20 -1.26 2.483544 20 NA NA
83 -1.1 2.638806 83 -0.21 0.4161627
24 -2.4 0.6029454 24 NA NA
41 -1.5 2.56 41 -0.21 0.25
9 -1.9 0.4222091 9 NA NA
27 -1.3 2.169941 27 NA NA
END

#initial values
#in the presence of missing data, initial values need to be specified for the missing items
#(the missing effects Y and the missing variances var)

list(
Y = structure(.Data = c(
 NA,  NA,  NA,  NA,  NA, NA,  NA,  0.0, NA, 0.0, NA,  0.0, NA,0.0, 0.0,
 NA,  NA,  0.0,0.0, NA, 0.0, NA,  NA,0.0, NA, 0.0, NA,  NA,  NA, 0.0,
 NA,  NA,  NA,0.0, NA, 0.0),.Dim = c(18,2)),
rho = 0.5, tau = c(0.25, 0.25), d=c(0.0,0.0),
h1 = 0.1, h2 = 0.1,
delta = structure(.Data = c(
 -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5,
 -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5,
-0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5,
-0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25,-1.5),
.Dim = c(18,2)),
var = structure(.Data = c(
 NA,  NA,  NA,  NA,  NA, NA,  NA,  0.5, NA, 0.5, NA,  0.5, NA,0.5, 0.5,
 NA,  NA,  0.5,0.5, NA, 0.5, NA,  NA,0.5, NA, 0.5, NA,  NA,  NA, 0.5,
 NA,  NA,  NA,0.5, NA, 0.5),.Dim = c(18,2))
)
```

A.2 WinBUGS code for BRMA in product normal formulation

```
Model{
#exchangeability of variances to predict missing SEs (if using - see previous appendix)
for (k in 1:num){
var[k,1]~dnorm(0,h1)I(0,)
var[k,2]~dnorm(0,h2)I(0,)
}
h1~dgamma(1.0,0.01)
h2~dgamma(1.0,0.01)

#within study precision matrix
for (i in 1:num) {
rho_w[i]<-corr_w
Prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
#covariance matrix for the j-th study
sigma[i,1,1]<-var[i,1]/n[i,1]
sigma[i,2,2]<-var[i,2]/n[i,2]
sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w[i]
sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w[i]
}

# Hierarchical structure of the random effects model:
for (i in 1:num) {
Y[i,1:2]~dmnorm(delta[i,1:2], Prec_w[i,1:2,1:2]) #within-study model

# between-studies model in the product normal formulation:
delta[i,1]~dnorm(eta1,prec1)
delta[i,2]~dnorm(eta2[i],prec2)
eta2[i]<-lambda20+lambda21*(delta[i,1] - mean(delta[,1]))

}
eta1~dnorm(0.0, 0.001)
lambda20~dnorm(0.0, 1.0E-3)
tau.1~dunif(0,2)
tau.2~dunif(0,2)
corr.1.2~dunif(-0.999,0.999)
tau1.sq<-pow(tau.1,2)
prec1<-1/tau1.sq
tau2.sq<-pow(tau.2,2)
psi2.sq<-tau2.sq-pow(lambda21,2)*tau1.sq
prec2<-1/psi2.sq
lambda21<-corr.1.2*tau.2/tau.1
mean.das<-eta1
mean.haq<-lambda20
sd.das<-tau.1
sd.haq<-tau.2
}
```

```
#data

list(num=18, corr_w=0.24)

n[,1] Y[,1] var[,1] n[,2] Y[,2] var[,2]
26 -1.7 1.683457 26 -0.31 0.4410603
188 -1.6 1.88 188 -0.352 0.388497
810 -1.9 1.96 810 -0.48 0.36
72 -1.47 2.337441 72 NA NA
30 -1.87 1.698847 30 NA NA
18 -2.1 1.532179 18 NA NA
66 -0.98 2.165142 66 NA NA
22 NA NA 22 -0.45 0.443637
110 -1 1.425636 110 NA NA
331 NA NA 331 -0.12 0.3628
37 NA NA 37 0.15 0.6399
6 -1.17 2.602115 6 NA NA
20 -1.26 2.483544 20 NA NA
83 -1.1 2.638806 83 -0.21 0.4161627
24 -2.4 0.6029454 24 NA NA
41 -1.5 2.56 41 -0.21 0.25
9 -1.9 0.4222091 9 NA NA
27 -1.3 2.169941 27 NA NA
END

#initial values
list(
Y = structure(.Data = c(
 NA,  NA,  NA,  NA,  NA, NA,  NA,  0.0, NA, 0.0, NA,  0.0, NA,0.0, 0.0,
 NA,  NA,  0.0,0.0, NA, 0.0, NA,  NA,0.0, NA, 0.0, NA,  NA,  NA, 0.0,
 NA,  NA,  NA,0.0, NA, 0.0),.Dim = c(18,2)),
corr.1.2 = 0.5,
eta1 = 0.0,
h1 = 0.1,
h2 = 0.1,
lambda20 = 0.0,
delta = structure(.Data = c(
 -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5,
 -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5,
-0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5,
-0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25,-1.5),
.Dim = c(18,2)),
tau.1 = 0.25,
tau.2 = 0.25,
var = structure(.Data = c(
 NA,  NA,  NA,  NA,  NA, NA,  NA,  0.5, NA, 0.5, NA,  0.5, NA,0.5, 0.5,
 NA,  NA,  0.5,0.5, NA, 0.5, NA,  NA,0.5, NA, 0.5, NA,  NA,  NA, 0.5,
 NA,  NA,  NA,0.5, NA, 0.5),.Dim = c(18,2))
)
```

A.3 WinBUGS code for the analysis with informative prior distributions for the correlations (as discussed in Section 2.4, with methods described in Section 2.4.4)

Analysis of EIPD to obtain the within-study correlation

```
Model {
for (j in 1:n.ipd){
HAQ[j]~dnorm(thetaH,prec1)
DAS[j]~dnorm(thetaD[j],prec2)
thetaD[j]<-alpha0+alpha1*(HAQ[j]-mean(HAQ[]))
}
thetaH~dnorm(0.0, 0.001)
alpha0~ dnorm(0.0,1.0E-3)
alpha1~ dnorm(0.0,1.0E-3)

xiD~dunif(0,10)
xiH~dunif(0,10)
xiD.sq<-xiD*xiD
xiH.sq<-xiH*xiH
prec1<-1/xiH.sq
prec2<-1/xiD.sq

meanH <- thetaH
#meanH <-alpha0+alpha1*mean.das
meanD <-alpha0

varH<-xiH.sq
varD<-pow(alpha1,2)*varH+xiD.sq
covDH<-alpha1*varH
corrDH<-covDH/sqrt(varD*varH)
ftcorr.ipd<- 0.5*log((1.0+corrDH)/(1.0-corrDH))
}

#data format:
list(n.ipd=293}

HAQ[] DAS[]
-0.0773258    0.000
-0.8360777    0.125
-0.8287334    0.000
0.8641753     0.250
...
END
```

Analysis of ESD to obtain estimates for constructing prior distribution for the between-studies correlation

The same code as for BRMA (PNF) included in Section A.2 is applied to the ESD (listed below) with removed first part dealing with missing variances (full data included in the ESD set). We only highlight the differences (or additions) when applying the code to the external data.

```
#remove the following lines:
#var[k,1]~dnorm(0,h1)I(0,)
#var[k,2]~dnorm(0,h2)I(0,)
#}
#h1~dgamma(1.0,0.01)
#h2~dgamma(1.0,0.01)

#replace the within-study correlation with a prior (if assume unknown):
# rho_w[i]<-corr_w
rho_w[i]~dunif(-1,1)


#include additional line to calculate Fisher transform
#of the between-studies correlation:
ftcorr<- 0.5*log((1.0+corr.1.2)/(1.0-corr.1.2))

#data
list(num=9)
#variances in the data are standard errors squared
 Y[,2]  var[,2] Y[,1] var[,1]
-0.55 0.0000698 -2.2 0.0003171
-0.6 0.0222836 -2.5 0.0295069
-0.27 0.0222836 -0.82 0.0122722
-0.1 0.0222836 -0.65 0.0338
-0.29 0.0037443 -1.3 0.0241509
-0.39 0.0034321 -1.6 0.0258036
-0.38 0.0031858 -1.7 0.0226549
-0.46 0.0007451 -2 0.0059507
-0.49 0.0028311 -2 0.0248544
END

#initial values:
list(
rho_w=c(0,0,0,0,0,0,0,0,0), corr.1.2 = 0.5, eta1 = 0.0, lambda20 = 0.0,
delta = structure(.Data = c(
-1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25,
-1.5, -0.25, -1.5, -0.25, -1.5, -0.25, -1.5, -0.25),.Dim = c(9,2)),
tau.1 = 0.25, tau.2 = 0.25)
```

Analysis of RA data (the main part of the analysis) with informative prior distributions for the correlations obtained from the above two sets of analyses

The code for BRMA (PNF) included in Section A.2 is run on the RA data with the following amendments:

Replace the within-study correlation with a prior:

```
#rho_w[i]<-corr_w
#prior distribution for the within-study correlations
#estimated from Fisher transformed correlation from IPD
ftcorr.ipd<-0.2476
ftcorr.ipd_se<-0.058
ftcorr.ipd_prec<-1.0/(ftcorr.ipd_se*ftcorr.ipd_se)
fcorr.w~ dnorm(ftcorr.ipd,ftcorr.ipd_prec)
rho_w<-(exp(2.0*fcorr.w)-1.0)/(exp(2.0*fcorr.w)+1.0)
```

Replace non-informative prior distribution for the between-studies correlation with informative prior distribution:

```
#corr.1.2~dunif(-0.999,0.999)
ftcorrDH<-2.0
ftcorrDH_se<-0.76
FcorrDH_prec<-1.0/(ftcorrDH_se*ftcorrDH_se)
FcorrDH~ dnorm(ftcorrDH,FcorrDH_prec)
corr.1.2<-(exp(2.0*FcorrDH)-1.0)/(exp(2.0*FcorrDH)+1.0)
```

Remove the within-study correlation from the data (using prior distribution here).

Remove the entry for the between-studies correlation from the initial values and include the entries for the Fisher transformed correlations:

```
#corr.1.2 = 0.5,
fcorr.w=0.0,
FcorrDH=0.5,
```

B Supplementary materials for evaluation of surrogate endpoints in relapsing remitting multiple sclerosis (RRMS)

In this appendix, we provide examples of code for methods of surrogate endpoint evaluation, described in Section 3 and applied to the illustrative example in RRMS described in Section 3.7. The next three sections (B.1–B.3) give details of implementing methods by Daniels and Hughes, BRMA PNF and BRMA in the standard form respectively. In Section B.4 we present R code (with embedded bugs code for the meta-analytic models) for the cross-validation procedure introduced in Section 3.6.1 and applied to the RRMS example in Section 3.7. Section B.5 gives an example of the analysis for predicting the treatment effect on the final outcome in a new study reporting only the treatment effect on the surrogate endpoint. We use model by Daniels and Hughes for this final analysis, but the analysis can be easily adapted to use one of the BRMA methods using the code from the earlier sections in this appendix.

B.1 WinBUGS code for model by Daniels and Hughes

```
Model{
#within study precision matrix
#rho_w<-corr_w  #if known, a fixed correlation can be used
rho_w~dunif(0,0.999)  #we assume positive within-study correlation
for (i in 1:num) {
  Prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
  #covariance matrix for the j-th study
  sigma[i,1,1]<-pow(se[i,1],2)
  sigma[i,2,2]<-pow(se[i,2],2)
  sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
  sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
}

# Bivariate model for surrogacy:
for (i in 1:num) {

Y[i,1:2]~dmnorm(delta[i,1:2], Prec_w[i,1:2,1:2]) # within-study model

# Daniels and Hughes formulation for the between-studies model:
delta[i,1]~dnorm(0.0, 0.001)
delta[i,2]~dnorm(eta2[i],prec2)
eta2[i]<-lambda20+lambda21*delta[i,1]

}

# prior distributions:
lambda20~dnorm(0.0, 1.0E-3)
lambda21~dnorm(0.0, 1.0E-3)
psi.2~dunif(0,2)
psi2.sq<-pow(psi.2,2)
prec2<-1/psi2.sq
}
```

```
#data

list(num=25)
#logrelef logrelef_se logdisef logdisef_se

Y[,1] se[,1] Y[,2] se[,2]
-0.0833816 0.0813719 0 0.2036533
-0.4155154 0.0893046 -0.3364722 0.2302073
-0.210721 0.3563859 0.1278334 0.7592028
-0.3424903 0.1067276 -0.1300531 0.2210844
-0.3856625 0.1266053 -0.461035 0.258417
-0.8915981 0.1351165 -0.3629056 0.3393006
-1.07881 0.2388631 -1.665008 0.7296625
-0.9942523 0.2406259 -0.2000212 0.74131
-0.3424903 0.0706827 -0.2097205 0.1464617
-0.3856625 0.0723215 -0.315081 0.1542706
0.0487901 0.0458876 0 0.0920902
-0.3424903 0.1348356 -0.8362481 0.3022435
-0.1625189 0.1005056 0.074108 0.1935841
-1.139434 0.0751103 -0.5340825 0.124761
-0.7985078 0.0603509 -0.2318016 0.0995418
-1.171183 0.1881434 -1.060872 0.3856541
-1.514128 0.2149895 -0.9555115 0.3515206
0.0295588 0.139198 0.2962659 0.2174605
-0.8675006 0.1082371 -0.3650315 0.1504737
-0.7985078 0.1038128 -0.3105963 0.145336
-0.1392621 0.1319462 0.2087549 0.3854737
-0.356675 0.1371439 0.040822 0.3966984
-0.9942523 0.2183502 -0.4462871 0.3555565
0.0582689 0.0695049 0.0487901 0.1145532
-0.0304592 0.0704836 0.0953102 0.1134564
END


#initial values
list(
lambda20=0.0,lambda21=0.0,
delta = structure(.Data = c(
-0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5,
-0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,
-0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5,  -0.5,
-0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5,
-0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5, -0.5),.Dim = c(25,2)),
psi.2=0.25, rho_w=0.25)
```

B.2 WinBUGS code for BRMA PNF model for surrogate endpoints

```
Model{
#within study precision matrix
#rho_w<-corr_w  #if known, a fixed correlation can be used
rho_w~dunif(0,0.999)  #we assume positive within-study correlation
for (i in 1:num) {
  Prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
  #covariance matrix for the j-th study
  sigma[i,1,1]<-pow(se[i,1],2)
  sigma[i,2,2]<-pow(se[i,2],2)
  sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
  sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
}
# Random effects model
for (i in 1:num) {

Y[i,1:2]~dmnorm(delta[i,1:2], Prec_w[i,1:2,1:2]) #within-study model

# product normal formulation for the between-studies model:
delta[i,1]~dnorm(eta1,prec1)
delta[i,2]~dnorm(eta2[i],prec2)
eta2[i]<-lambda20+lambda21*delta[i,1]

}
eta1~dnorm(0.0, 0.001)
lambda20~dnorm(0.0, 1.0E-3)
tau1~dunif(0,2)
tau2~dunif(0,2)
rho~dunif(-0.999,0.999)
tau1.sq<-pow(tau1,2)
prec1<-1/tau1.sq
tau2.sq<-pow(tau2,2)
psi2.sq<-tau2.sq-pow(lambda21,2)*tau1.sq
prec2<-1/psi2.sq
lambda21<-rho*tau2/tau1
d1<-eta1
d2<-lambda20+lambda21*eta1
R2<-pow(rho,2)
}
```

Data in the same format as in appendix B.1.

```
#initial values:
list(
rho = 0.5, eta1=0.0, lambda20=0.0,
delta = structure(.Data = c(
-0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5,
-0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,
-0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5,  -0.5,
-0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5,
-0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5, -0.5),.Dim = c(25,2)),
tau1 = 0.25, tau2=0.25, rho_w=0.25)
```

B.3 WinBUGS code for BRMA standard model for surrogate endpoints

```
Model{
#within study precision matrix
#rho_w<-corr_w  #if known, a fixed correlation can be used
rho_w~dunif(0,0.999)  #we assume positive within-study correlation
for (i in 1:num) {
  prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
  #covariance matrix for the j-th study
  sigma[i,1,1]<-pow(se[i,1],2)
  sigma[i,2,2]<-pow(se[i,2],2)
  sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
  sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w

Y[i,1:2] ~ dmnorm(delta[i,1:2],prec_w[i,1:2,1:2]) # within-study model

delta[i,1:2] ~ dmnorm(d[1:2],prec_b[1:2,1:2]) # between-studies model

}

for (j in 1:2) {
d[j] ~ dnorm(0.0,0.001)
}

prec_b[1:2,1:2]<-inverse(Cov_b[,])
for (j in 1:2) {
tau[j]~dunif(0,2)
tau.sq[j]<-pow(tau[j],2)
Cov_b[j,j]<-tau.sq[j]
}
Cov_b[1,2]<-rho*tau[1]*tau[2]
Cov_b[2,1]<-Cov_b[1,2]

rho~dunif(-0.999,0.999)
R2<-pow(rho,2)
psi.sq<-pow(tau[2],2)*(1.0-pow(rho,2))
lambda1<-rho*tau[2]/tau[1]
lambda0<-d[2]-d[1]*rho*tau[2]/tau[1]
}
```

Data in the same format as in appendix B.1.

```
#initial values:
list(
rho_w=0.25, rho = 0.5, tau = c(0.25, 0.25), d=c(0.0,0.0),
delta = structure(.Data = c(
-0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5,
-0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,
-0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5,  -0.5,
-0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5,
-0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5, -0.5),.Dim = c(25,2))
)
```

B.4 R code (with models in bugs) for cross-validation (Sec. 3.6.1 and 3.7)

```
#Load Package & directory where WinBUGS program is
library(R2WinBUGS)
bugs.dir<-"Z:/My Documents/WinBUGS14"
#Define working directory where source files are located
wd<-"Z:/My documents/Research/TSD/MSexample"
setwd(wd)

#set number of iterations for MCMC simulations
n.iter<- 100000  # Number of iterations
n.burnin<-50000  # Burn-in

# RRMS data (data for surrogate endpoints, reformatted in R for cross-validation):

num<-25
# data consist of lists of: authors, effects on relapse, corresponding standard errors,
# effects on disability progression ad corresponding SEs form all (num=25) studies:
author<-c("Paty1", "Paty2","Miligan","Johnson",
"Jacobs","Fazekas","Millewfiorini","Achiron",
"Li1","Li2","Clanet","Durelli",
"Baumhackl","Polman","Rudick","Coles",
"Coles","Mikol","Comi1","Comi2",
"Havrdova1","Havrdova2","Sorensen","O'Connor1", "O'Connor2")

logrelef<-c(-0.0833816, -0.4155154, -0.210721, -0.3424903,
-0.3856625, -0.8915981, -1.07881, -0.9942523,
-0.3424903, -0.3856625, 0.0487901, -0.3424903,
-0.1625189, -1.139434, -0.7985078, -1.171183,
-1.514128, 0.0295588, -0.8675006, -0.7985078,
-0.1392621, -0.356675, -0.9942523, 0.0582689, -0.0304592)

logrelef_se<-c(0.0813719, 0.0893046, 0.3563859, 0.1067276,
0.1266053, 0.1351165, 0.2388631, 0.2406259,
0.0706827, 0.0723215, 0.0458876, 0.1348356,
0.1005056, 0.0751103, 0.0603509, 0.1881434,
0.2149895, 0.139198, 0.1082371, 0.1038128,
0.1319462, 0.1371439, 0.2183502, 0.0695049, 0.0704836)

logdisef<-c(0, -0.3364722, 0.1278334, -0.1300531,
-0.461035, -0.3629056, -1.665008, -0.2000212,
-0.2097205, -0.315081, 0, -0.8362481,
0.074108, -0.5340825, -0.2318016, -1.060872,
-0.9555115, 0.2962659, -0.3650315, -0.3105963,
0.2087549, 0.040822, -0.4462871, 0.0487901, 0.0953102)

logdisef_se<-c(0.2036533, 0.2302073, 0.7592028, 0.2210844,
0.258417, 0.3393006, 0.7296625, 0.74131,
0.1464617, 0.1542706, 0.0920902, 0.3022435,
0.1935841, 0.124761, 0.0995418, 0.3856541,
0.3515206, 0.2174605, 0.1504737, 0.145336,
0.3854737, 0.3966984, 0.3555565, 0.1145532, 0.1134564)
```

```
#effects on the final outcome (log relative risk and corresponding CIs)
#for comparison with predicted effects
fin<-logdisef
fin.lci<-logdisef-1.96*logdisef_se
fin.uci<-logdisef+1.96*logdisef_se
#effects on the final outcome as relative risk (no log) and corresponding CIs:
e.fin<-exp(logdisef)
e.fin.lci<-exp(logdisef-1.96*logdisef_se)
e.fin.uci<-exp(logdisef+1.96*logdisef_se)


##############################################################################################

# Daniels & Hughes model

#create text file for BUGS code of the Daniels & Hughes model:

dh_model_pred<-"model{


se[index,2]~dunif(0.0001,15) #prior on the missing SE in the validation study i=index
pred.delta2<-delta[index,2] #predicted effect on the final outcome in the validation study

#within study precision matrix
#rho_w<-corr_w  #if known, a fixed correlation can be used
rho_w~dunif(0,0.999)  #we assume positive within-study correlation
for (i in 1:num) {
  Prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
  #covariance matrix for the j-th study
  sigma[i,1,1]<-pow(se[i,1],2)
  sigma[i,2,2]<-pow(se[i,2],2)
  sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
  sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
}

# Random effects model
for (i in 1:num) {
Y[i,1:2]~dmnorm(delta[i,1:2], Prec_w[i,1:2,1:2])
# product normal formulation for the between study part:
delta[i,1]~dnorm(0.0, 0.001)
delta[i,2]~dnorm(eta2[i],prec2)
eta2[i]<-lambda0+lambda1*delta[i,1]
}

lambda0~dnorm(0.0, 1.0E-3)
lambda1~dnorm(0.0, 1.0E-3)
psi.2~dunif(0,2)
psi2.sq<-pow(psi.2,2)
prec2<-1/psi2.sq
}"
writeLines(dh_model_pred,"model_dh.txt")
```

```
#cross-validation code for D&H model:

#setting new variables for results of cross-validation:
pred.delta2.dh<-pred.cri.dh<-bias.dh<-cil.dh<-ciu.dh<-sd.dh<-array(0,num)

for(i in 1:num){
index<-i

Y<-se<-array(0,dim=c(num,2))
#data
Y[,1]<-logrelef
Y[,2]<-logdisef
se[,1]<-logrelef_se
se[,2]<-logdisef_se

#setting final outcome in the validation study i to misisng value
#to be predicted
Y[i,2]<-NA
se[i,2]<-NA
data=list(Y=Y, se=se, num=num, index=i)

#inits
#setting initial values for missing final outcome in the validation study
Y0<-se0<-structure(rep(NA,num*2),dim=c(num,2))
Y0[index,2]<- -0.5
se0[index,2]<- 0.2
inits=list(list(psi.2 = 0.25, lambda1 = 0.0, lambda0 = 0.0, rho_w=0.25,
delta = structure(.Data = c(rep(-0.5,num),rep(-0.5,num)), .Dim=c(num,2)),
Y=Y0,se=se0))

#monitoring
para=c("psi2.sq", "pred.delta2")

#run D&H
fit.dh<-bugs(data=data, inits=inits, para=para,
model.file="model_dh.txt",
n.chains=1, n.burnin=n.burnin, n.iter=n.iter, n.thin=1,
DIC=F, debug=F, save.history =F,
bugs.directory = bugs.dir, working.directory = wd)

x <- fit.dh$sims.matrix[,grep("pred.delta2",colnames(fit.dh$sims.matrix))]
pred.delta2.dh[i]<-mean(x)
bias.dh[i]<-logdisef[i]-pred.delta2.dh[i]
sd.dh[i] <- fit.dh$summary[grep("pred.delta2",rownames(fit.dh$summary)),c(2)]
pred.cri.dh[i]<-sqrt(logdisef_se[i]^2+sd.dh[i]^2)*2.00*1.96
cil.dh[i] <- pred.delta2.dh[i]-pred.cri.dh[i]/2.0
ciu.dh[i] <- pred.delta2.dh[i]+pred.cri.dh[i]/2.0
}

e.pred.dh<-exp(pred.delta2.dh)
e.pred.dh.lci<-exp(cil.dh)
e.pred.dh.uci<-exp(ciu.dh)
```

```
e.predictions.brma.dh<-data.frame(author,e.fin,e.fin.lci,e.fin.uci,
e.pred.dh,e.pred.dh.lci,e.pred.dh.uci,
fin,fin.lci,fin.uci,pred.delta2.dh,cil.dh,ciu.dh)
write.csv(e.predictions.brma.dh, file = "epredictions.dh.ms.csv", row.names = FALSE)

##############################################################################################

### BRMA PNF

#create text file for BUGS code of the BRMA PNF model
pnf_model_pred<-"model{

se[index,2]~dunif(0.0001,15) #prior on the missing SE in the validation study i=index
pred.delta2<-delta[index,2]  #predicted effect on the final outcome in the validation study


#within study precision matrix
#rho_w<-corr_w  #if known, a fixed correlation can be used
rho_w~dunif(0,0.999)  #we assume positive within-study correlation
for (i in 1:num) {
  Prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
  #covariance matrix for the j-th study
  sigma[i,1,1]<-pow(se[i,1],2)
  sigma[i,2,2]<-pow(se[i,2],2)
  sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
  sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
}

# Random effects model
for (i in 1:num) {
Y[i,1:2]~dmnorm(delta[i,1:2], Prec_w[i,1:2,1:2])
# product normal formulation for the between study part:
delta[i,1]~dnorm(eta1,prec1)
delta[i,2]~dnorm(eta2[i],prec2)
eta2[i]<-lambda0+lambda1*delta[i,1]
}
eta1~dnorm(0.0, 0.001)
lambda0~dnorm(0.0, 1.0E-3)
tau1~dunif(0,2)
tau2~dunif(0,2)
rho~dunif(-0.999,0.999)
tau1.sq<-pow(tau1,2)
prec1<-1/tau1.sq
tau2.sq<-pow(tau2,2)
psi2.sq<-tau2.sq-pow(lambda1,2)*tau1.sq
prec2<-1/psi2.sq
lambda1<-rho*tau2/tau1
d1<-eta1
d2<-lambda0+lambda1*eta1
R2<-pow(rho,2)
}"
writeLines(pnf_model_pred,"model_pnf.txt")
```

```
#cross-validation code for BRMA PNF model:

pred.delta2.pnf<-pred.cri.pnf<-bias.pnf<-cil.pnf<-ciu.pnf<-sd.pnf<-array(0,num)
for(i in 1:num){
index<-i

Y<-se<-array(0,dim=c(num,2))
#data
Y[,1]<-logrelef
Y[,2]<-logdisef
se[,1]<-logrelef_se
se[,2]<-logdisef_se

#setting final outcome in the validation study i to misisng value
#to be predicted
Y[i,2]<-NA
se[i,2]<-NA
data=list(Y=Y, se=se, num=num, index=i)

#inits
#setting initial values for missing final outcome in the validation study
Y0<-se0<-structure(rep(NA,num*2),dim=c(num,2))
Y0[index,2]<- -0.5
se0[index,2]<- 0.2
inits=list(list(rho = 0.5, eta1=0.0, lambda0 = 0.0,
tau1 = 0.25, tau2=0.25, rho_w=0.25,
delta = structure(.Data = c(rep(-0.5,num),rep(-0.5,num)), .Dim=c(num,2)),
Y=Y0,se=se0))

#monitoring
para=c("psi2.sq", "pred.delta2")

#run PNF
fit.pnf<-bugs(data=data, inits=inits, para=para,
model.file="model_pnf.txt",
n.chains=1, n.burnin=n.burnin, n.iter=n.iter, n.thin=1,
DIC=F, debug=F, save.history =F,
bugs.directory = bugs.dir, working.directory = wd)

x <- fit.pnf$sims.matrix[,grep("pred.delta2",colnames(fit.pnf$sims.matrix))]
pred.delta2.pnf[i]<-mean(x)
bias.pnf[i]<-logdisef[i]-pred.delta2.pnf[i]
sd.pnf[i] <- fit.pnf$summary[grep("pred.delta2",rownames(fit.pnf$summary)),c(2)]
pred.cri.pnf[i]<-sqrt(logdisef_se[i]^2+sd.pnf[i]^2)*2.00*1.96
cil.pnf[i] <- pred.delta2.pnf[i]-pred.cri.pnf[i]/2.0
ciu.pnf[i] <- pred.delta2.pnf[i]+pred.cri.pnf[i]/2.0
}

e.pred.pnf<-exp(pred.delta2.pnf)
e.pred.pnf.lci<-exp(cil.pnf)
e.pred.pnf.uci<-exp(ciu.pnf)
```

```
e.predictions.brma.pnf<-data.frame(author,e.fin,e.fin.lci,e.fin.uci,
e.pred.pnf,e.pred.pnf.lci,e.pred.pnf.uci,
fin,fin.lci,fin.uci,pred.delta2.pnf,cil.pnf,ciu.pnf)
write.csv(e.predictions.brma.pnf, file = "epredictions.pnf.ms.csv", row.names = FALSE)

#############################################################################################
### BRMA

#create text file for BUGS code of the BRMA model
brma_model_pred<-"model{

se[index,2]~dunif(0.0001,15) #prior on the missing SE in the validation study i=index
pred.delta2<-delta[index,2] #predicted effect on the final outcome in the validation study


#within study precision matrix
#rho_w<-corr_w  #if known, a fixed correlation can be used
rho_w~dunif(0,0.999)  #we assume positive within-study correlation
for (i in 1:num) {
  Prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
  #covariance matrix for the j-th study
  sigma[i,1,1]<-pow(se[i,1],2)
  sigma[i,2,2]<-pow(se[i,2],2)
  sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
  sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
}

# Random effects model
for (i in 1:num) {
Y[i,1:2]~dmnorm(delta[i,1:2], Prec_w[i,1:2,1:2])
# product normal formulation for the between study part:
delta[i,1]~dnorm(eta1,prec1)
delta[i,2]~dnorm(eta2[i],prec2)
eta2[i]<-lambda0+lambda1*delta[i,1]
}
eta1~dnorm(0.0, 0.001)
lambda0~dnorm(0.0, 1.0E-3)
tau1~dunif(0,2)
tau2~dunif(0,2)
rho~dunif(-0.999,0.999)
tau1.sq<-pow(tau1,2)
prec1<-1/tau1.sq
tau2.sq<-pow(tau2,2)
psi2.sq<-tau2.sq-pow(lambda1,2)*tau1.sq
prec2<-1/psi2.sq
lambda1<-rho*tau2/tau1
d1<-eta1
d2<-lambda0+lambda1*eta1
R2<-pow(rho,2)
}"
writeLines(brma_model_pred,"model_brma.txt")
```

```
#cross-validation code for BRMA model:

pred.delta2.brma<-pred.cri.brma<-bias.brma<-cil.brma<-ciu.brma<-sd.brma<-array(0,num)
for(i in 1:num){
index<-i

Y<-se<-array(0,dim=c(num,2))
#data
Y[,1]<-logrelef
Y[,2]<-logdisef
se[,1]<-logrelef_se
se[,2]<-logdisef_se

#setting final outcome in the validation study i to misisng value
#to be predicted
Y[i,2]<-NA
se[i,2]<-NA
data=list(Y=Y, se=se, num=num, index=i)


#inits
#setting initial values for missing final outcome in the validation study
Y0<-se0<-structure(rep(NA,num*2),dim=c(num,2))
Y0[index,2]<- -0.5
se0[index,2]<- 0.2
inits=list(list(rho = 0.5, eta1=0.0, lambda0 = 0.0,
tau1 = 0.25, tau2=0.25, rho_w=0.25,
delta = structure(.Data = c(rep(-0.5,num),rep(-0.5,num)), .Dim=c(num,2)),
Y=Y0,se=se0))

#monitoring
para=c("psi2.sq", "pred.delta2")

#run BRMA
fit.brma<-bugs(data=data, inits=inits, para=para,
model.file="model_brma.txt",
n.chains=1, n.burnin=n.burnin, n.iter=n.iter, n.thin=1,
DIC=F, debug=F, save.history =F,
bugs.directory = bugs.dir, working.directory = wd)

x <- fit.brma$sims.matrix[,grep("pred.delta2",colnames(fit.brma$sims.matrix))]
pred.delta2.brma[i]<-mean(x)
bias.brma[i]<-logdisef[i]-pred.delta2.brma[i]
sd.brma[i] <- fit.brma$summary[grep("pred.delta2",rownames(fit.brma$summary)),c(2)]
pred.cri.brma[i]<-sqrt(logdisef_se[i]^2+sd.brma[i]^2)*2.00*1.96
cil.brma[i] <- pred.delta2.brma[i]-pred.cri.brma[i]/2.0
ciu.brma[i] <- pred.delta2.brma[i]+pred.cri.brma[i]/2.0
}

e.pred.brma<-exp(pred.delta2.brma)
e.pred.brma.lci<-exp(cil.brma)
e.pred.brma.uci<-exp(ciu.brma)
```

```
e.predictions.brma.brma<-data.frame(author,e.fin,e.fin.lci,e.fin.uci,
e.pred.brma,e.pred.brma.lci,e.pred.brma.uci,
fin,fin.lci,fin.uci,pred.delta2.brma,cil.brma,ciu.brma)

write.csv(e.predictions.brma.brma, file = "epredictions.brma.ms.csv", row.names = FALSE)
```

B.5 Example of the analysis (using method by Daniels and Hughes) for predicting the treatment effect on the final outcome in a new study reporting only the treatment effect on the surrogate endpoint

In this Appendix, the WinBUGS implementation for predicting the treatment effect on the final outcome in a new study reporting only the treatment effect on the surrogate endpoint is discussed. It refers to the prediction in the illustrative example in MS, discussed in Section 3.7.4.

```
Model{
#defining the within-study precision matrix

# model assuming exchangeable variances in the arms for the final outcome
# to predict missing standard error for the effect on the final outcome in the new study
for (k in 1:num){
var[k]~dnorm(0,h1)I(0,)
}
h1~dgamma(1.0,0.01)
se[index,2]<-sqrt(var[index]*(1/nA[index] + 1/nB[index]))

#rho_w<-corr_w  #if known, a fixed correlation can be used
rho_w~dunif(0,0.999)  #we assume positive within-study correlation
for (i in 1:num) {
  Prec_w[i,1:2,1:2] <- inverse(sigma[i,1:2,1:2])
  #covariance matrix for the j-th study
  sigma[i,1,1]<-pow(se[i,1],2)
  sigma[i,2,2]<-pow(se[i,2],2)
  sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
  sigma[i,2,1]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w
}

# Model for surrogate relationship (Daniels and Hughes):
for (i in 1:num) {
Y[i,1:2]~dmnorm(delta[i,1:2], Prec_w[i,1:2,1:2])
# product normal formulation for the between study part:
delta[i,1]~dnorm(0.0, 0.001)
delta[i,2]~dnorm(eta2[i],prec2)
eta2[i]<-lambda20+lambda21*delta[i,1]
}
lambda20~dnorm(0.0, 1.0E-3)
lambda21~dnorm(0.0, 1.0E-3)
psi.2~dunif(0,2)
psi2.sq<-pow(psi.2,2)
prec2<-1/psi2.sq

#predicted effect on the final outcome 2 in the new study "index":
predicted.final<-exp(Y[index,2])

}
```

Variable "index" in the data denotes the number/index of the new study in the data, for which we are making the prediction. Data set includes treatment effects on both outcomes Y[,1] and Y[,2] and corresponding standard errors se[,1] and se[,2]. Data on the final outcome are missing for the second study which is assumed here a new study, for which the prediction is made. To make the

prediction, we need to predict the standard error for the missing effect on the final outcome in the second study. We assume exchangeability of the population variances var[] in the arms (see Section 2.1.3, paragraph second to last). These variances are included in the data along with the numbers of individuals in each arm nA[] and nB[], which are needed to calculate the missing standard error of the unreported effect on the final outcome in the new study.

In this case (where only one effect on one outcomes is missing) only numbers for the second study are needed and only the variances for the second outcome are included. However, in general case, a similar approach can be used for variances on both outcomes when data are missing on both outcomes in some studies and treatment allocation is unbalanced.

```
#data
list(num=25, index=2)

Y[,1] se[,1] Y[,2] se[,2] var[] nA[] nB[]
-0.0833816 0.0814432 0 0.2036533 2.571428 124 124
-0.4155154 0.1165316 NA NA NA 124 124
-0.210721 0.348466 0.1278334 0.7592028 3.32 9 16
-0.3424903 0.1045693 -0.1300531 0.2210844 3.067065 126 125
-0.3856625 0.1265924 -0.461035 0.258417 2.871124 87 85
-0.8915981 0.1351122 -0.3629056 0.3393006 4.258843 73 75
-1.07881 0.2404074 -1.665008 0.7296625 6.764706 24 27
-0.9942523 0.2393568 -0.2000212 0.74131 5.495405 20 20
-0.3424903 0.0706421 -0.2097205 0.1464617 2.016341 187 189
-0.3856625 0.0723048 -0.315081 0.1542706 2.207251 187 184
0.0487901 0.0458702 0 0.0920902 1.700351 402 400
-0.3424903 0.1350154 -0.8362481 0.3022435 4.29156 92 96
-0.1625189 0.1006188 0.074108 0.1935841 2.707521 144 145
-1.139434 0.0751408 -0.5340825 0.124761 3.263509 315 627
-0.7985078 0.0603669 -0.2318016 0.0995418 2.900627 582 589
-1.171183 0.1900292 -1.060872 0.3856541 8.291479 111 112
-1.514128 0.2130032 -0.9555115 0.3515206 6.826923 111 110
0.0295588 0.138675 0.2962659 0.2174605 9.031223 378 386
-0.8675006 0.1080229 -0.3650315 0.1504737 4.924605 437 433
-0.7985078 0.104044 -0.3105963 0.145336 4.713475 437 456
-0.1392621 0.1317971 0.2087549 0.3854737 4.382145 60 58
-0.356675 0.1371352 0.040822 0.3966984 4.836237 60 63
-0.9942523 0.2156183 -0.4462871 0.3555565 4.107692 64 66
0.0582689 0.0695519 0.0487901 0.1145532 3.920694 448 897
-0.0304592 0.0705193 0.0953102 0.1134564 3.848822 448 899
END

#initial values
#the new study has missing Y and se (var) on the second outcome
#and these missing values (and all parameters) are given initial values for MCMC
list(
Y = structure(.Data = c(
            NA, NA,            NA,            0.0,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
```

```
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA),
.Dim = c(25,2)),
h1 = 0.5,
lambda20=0.0,lambda21=0.0,
delta = structure(.Data = c(
 -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5,
 -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,
 -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5,  -0.5,
 -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5,
 -0.5,  -0.5, -0.5, -0.5, -0.5,  -0.5, -0.5, -0.5, -0.5, -0.5),.Dim = c(25,2)),
psi.2=0.25, rho_w=0.25,
var = c(
            NA,   0.01,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,            NA)))
```

C Supplementary materials for multivariate meta-analysis

C.1 Constructing prior distribution for the between-studies covariance matrix

There are a number of approaches to constructing a prior distribution on the between-studies covariance matrix in MRMA. Wei and Higgins, for example, investigated placing inverse Wishart prior distribution on the covariance matrix [6] or by using a separation strategy, with spherical [95, 6] or Cholesky decomposition [6] of the correlation matrix.

Inverse Wishart distribution $IW(\Upsilon, df)$ is the conjugate prior distribution for the between-studies variance-covariance matrix, where Υ is a $N \times N$ scale matrix and $df \geq N$ is the number of degrees of freedom [115, 6]. Using the inverse Wishart distribution with the following parameters: the scale matrix equal to the identity matrix and the degrees of freedom equal to $N + 1$ (the dimension of the matrix plus 1) will induce a uniform prior distribution (with values ranging between -1 and 1) for the between-studies correlation ρ^{jk} between treatment effects on outcomes j and k [116]. However, the implied prior distributions on the variances may not be suitably non-informative. A number of authors found that Wishart priors had an influential impact on the posterior distribution [117, 6, 12] therefore we do not recommend this approach.

A preferred way of representing the between-studies variance-covariance matrix is by decomposing it as $\mathbf{T} = \mathbf{V}^{1/2}\mathbf{R}\mathbf{V}^{1/2}$, where $\mathbf{V}^{1/2}$ is a $N \times N$ diagonal matrix of the standard deviations and $\mathbf{R}$ is a positive semi-definite $N \times N$ matrix of correlations. In this separation strategy, described by Barnard et al. [96], the standard deviations and correlations are separated out allowing for independent prior distributions to be placed on these parameters. Prior distributions are placed on the standard deviations, which need to be restricted to positive values, for example $\tau_j \sim Unif(0, 2)$. The correlation matrix needs to be parameterized in such a way to ensure the positive semi-definiteness.

Separation by Cholesky decomposition

Wei and Higgins further reparameterize the correlation matrix to enforce the positive semi-definite constraints [6]. Matrix $\mathbf{R}$ is represented as $\mathbf{R} = \mathbf{L}^T\mathbf{L}$ with $\mathbf{L}$ being a $N \times N$ upper triangular matrix. Because R is a correlation matrix, ensuring that its elements lie in the range (-1,1) will induce constraints on possible values for the $N(N + 1)/2$ Cholesky factors L_{jk} $(j, k = 1, \ldots, N, j \leq k)$. The diagonal elements of R are each 1, so $\rho^{jj} = \mathbf{L}_{j\cdot}^T\mathbf{L}_{j\cdot} = 1$, where $\mathbf{L}_{j\cdot} = (L_{j1}, L_{j2}, \ldots, L_{jj})$ denotes column j in matrix $\mathbf{L}$. The Cholesky factors must lie in the intersection of $(-1, 1)$ and (L_{jk}^l, L_{jk}^r), where (L_{jk}^l, L_{jk}^r) satisfies the conditions that $\rho^{jj} = \mathbf{L}_{j\cdot}^T\mathbf{L}_{j\cdot} = 1$. The prior distributions are then placed on the Cholesky factors L_{ij}.

To obtain the elements of the matrix $\mathbf{L}$ and the corresponding prior distributions, we follow the method by Wei and Higgins (2013). The elements of the top row of the correlation matrix are

$$R_{1j} = L_{11}L_{1j},$$

the diagonal elements are

$$R_{jj} = \sum_{k=1}^{j} L_{kj}^2$$

and the remaining elements are

$$R_{ij} = \sum_{k=1}^{i} L_{ki}L_{kj}$$

where $j > i,\ i = 1, \ldots, N - 1,\ j = 1, \ldots, N$. Prior distributions are placed on the elements of matrix L in such a way to ensure the correlations are constrained to the range of values between -1 and 1. This is achieved, following Wei and Higgins (2013), by selecting plausible intervals for these elements. For the top row of matrix L we set uniform prior distributions on the following intervals:

$$L_{1j} \in [-1, 1]$$

and the intervals for the remaining off-diagonal elements are

$$L_{ij} \in \left[-\sqrt{1 - \sum_{k=1}^{i-1} L_{kl}^2}, \sqrt{1 - \sum_{k=1}^{i-1} L_{kl}^2} \right]$$

which gives implied prior distributions for the diagonal elements:

$$L_{jj} = \sqrt{1 - \sum_{k=1}^{j-1} L_{kj}^2}.$$

A limitation related to specifying the prior distribution for the variance-covariance matrix using the Cholesky decomposition it that the resulting prior distributions for the between-studies correlations are dependent on the ordering of the outcomes. A sensitivity analysis of the effect of the ordering of outcomes is therefore recommended.

Separation by spherical decomposition

Lu and Ades [95] introduced spherical decomposition of the variance-covariance matrix to network meta-analysis of multi-arm trials, where treatment effects relative to the same baseline treatment in the same study are correlated. Wei and Higgins adopted this decomposition technique in multivariate meta-analysis [6]. Spherical decomposition is a reparameterization of the Cholesky decomposition in terms of sine and cosine functions for the Cholesky factors L_{jk}. The products of these sine and cosine functions are convenient for parameterization of a correlation matrix as they lie within the interval $(-1, 1)$. In this parameterization, the products of diagonal elements satisfy the condition $\rho^{jj} = \mathbf{L}_{j\cdot}^T \mathbf{L}_{j\cdot} = 1$, where $\mathbf{L}_{j\cdot}$ denotes column j in matrix $\mathbf{L}$. The first element of the diagonal is set $L_{11} = 1$ and for $j = 1, \ldots, N$,

$$\begin{aligned}
L_{l1} &= \cos(\phi_{l2}) && (26)\\
L_{l2} &= \sin(\phi_{l2}) \cos(\phi_{l3}) && (27)\\
&\vdots && (28)\\
L_{l,l-1} &= \sin(\phi_{l2}) \sin(\phi_{l3}) \cdots \cos(\phi_{ll}) && (29)\\
L_{l,l} &= \sin(\phi_{l2}) \sin(\phi_{l3}) \cdots \sin(\phi_{ll}) && (30)
\end{aligned}$$

To ensure the uniqueness of the spherical parameterization, the parameters are restricted $\phi_{lk} \in (0, \pi)$, for $k = 1, \ldots, l$ [118, 6]. This ensures that, the (j, k) element of $\mathbf{R}$ can be represented as the inner product $\rho^{jk} = \mathbf{L}_{j\cdot}^T \mathbf{L}_{k\cdot}$.

C.2 WinBUGS code for multivariate meta-analysis with Cholesky decomposition of the between-studies covariance matrix with application to the example in rheumatoid arthritis

```
model{

corr_w[1,2]<-corrDA
corr_w[1,3]<-corrDH
corr_w[2,3]<-corrAH

#exchangeability of variances model:
#it can be replaced with priro distributions on individual missing SEs
#and data with SEs (instead of variances and number of participant) - see appendix A.1
for (i in 1:num){
for (j in 1:no){
#se[i,j]<--sqrt(var[i,j]/n[i,j])
var[i,j]~dnorm(0,h[j])I(0,)
}
}
for (j in 1:no){
h[j]~dgamma(1.0,0.01)
}

#Multivariate hierarchical model:
for(i in 1:num) {
prec_w[i,1:no,1:no] <- inverse(Sigma[i,1:no,1:no])
#covariance matrix for the j-th study
for (j in 1:no){
Sigma[i,j,j]<-var[i,j]/n[i,j]
}
for (j in 1:no){
for (k in j+1:no){
rho_w[i,j,k]<-corr_w[j,k]    #assuming the same correlations across studies
rho_w[i,k,j]<-corr_w[j,k]
Sigma[i,j,k]<-sqrt(Sigma[i,j,j])*sqrt(Sigma[i,k,k])*rho_w[i,j,k]
Sigma[i,k,j]<-Sigma[i,j,k]
}
}
y[i,1:no] ~ dmnorm(delta[i,1:no],prec_w[i,1:no,1:no]) # within-study model
delta[i,1:no] ~ dmnorm(d[1:no],prec_b[1:no,1:no]) # between-studies model
}
prec_b[1:no,1:no]<-inverse(Cov_b[,])

for (j in 1:no){
d[j] ~ dnorm(0,0.001)
Cov_b[j,j]<-tau.sq[j]
tau.sq[j]<-pow(tau[j],2)
tau[j]~dunif(0,2)
#rho[j,j]<-1
}
for (j in 1:no){
for (k in j+1:no){
```

```
Cov_b[j,k]<- tau[j]*tau[k]*rho[j,k]
Cov_b[k,j]<-Cov_b[j,k]
}
}
```

constructing a prior distribution on the between-studies correlation matrix:

```
# assigning priors to the upper triangular matrix in Cholesky decomposition

L[1,1]<-1.0

for (k in 2:no){
L.u[1,k]~dunif(-0.999,0.999)
L[1,k] <- L.u[1,k]
}

for (x in 1:no-1){
for (k in x+1:no){
  p[x,k]<-pow(L[x,k],2)
}
}

for (x in 3:no){
 for (k in x:no){
s[x-1,k]<-sum(p[1:x-2,k])
lim[x-1,k]<- sqrt(1-s[x-1,k])
L.u[x-1,k]~dunif(-0.999,0.999)
L[x-1,k]<- lim[x-1,k] * L.u[x-1,k]
 }
}

L.u[2,2]<-sqrt(1-pow(L[1,2],2))
L[2,2]<-L.u[2,2]

for (k in 3:no){
s2[k]<-sum(p[1:k-1,k])
 L.u[k,k]<-sqrt(1-s2[k])
 L[k,k]<-L.u[k,k]
}

#assigning values for the correlations:

#rho[1,1]<-L[1,1]

for (k in 2:no){
rho[1,k]<-L[1,k]
rho[k,1]<-L[1,k]
}
for (x in 1:no){
  for (k in x:no){
    for (j in 1:x){
LL[j,x,k]<-L[j,x]*L[j,k]
```

```
}
  }
}
for (x in 2:no-1){
  for (k in x+1:no){
rho[x,k]<-sum(LL[1:x,x,k])
rho[k,x]<-rho[x,k]
}
}

for (x in 2:no){
rho[x,x]<-sum(LL[1:x,x,x])
}

#end of constructing the prior distribution

Calculating precentage responders ACR20
perc.acr<-exp(d[2])/(1+exp(d[2]))

}

#data
list(num=22, no=3, corrDA=-0.2, corrDH=0.24, corrAH=-0.13)

#n_acr n_das n_haq log_acr_odds var_acr_lodds das28meanchange das28changevar haqmeanchange
#haqchangevar

n[,2] n[,1] n[,3] y[,2] var[,2] y[,1] var[,1] y[,3] var[,3]
26 26 26 NA NA -1.7 1.683457 -0.31 0.4410603
188 188 188 -0.1920777 4.037007 -1.6 1.88 -0.352 0.388497
810 810 810 0.4054652 4.166667 -1.9 1.96 -0.48 0.36
25 25 25 0.9444618 4.960318 NA NA NA NA
72 72 72 1.17412 5.544385 -1.47 2.337441 NA NA
30 30 30 NA NA -1.87 1.698847 NA NA
18 18 18 NA NA -2.1 1.532179 NA NA
66 66 66 NA NA -0.98 2.165142 NA NA
22 22 22 0.5596157 4.321429 NA NA -0.45 0.443637
110 110 110 NA NA -1 1.425636 NA NA
331 331 331 NA NA NA NA -0.12 0.3628
37 37 37 NA NA NA NA 0.15 0.6399
337 337 337 0.041549 4.001727 NA NA NA NA
411 411 411 NA NA NA NA NA NA
6 6 6 NA NA -1.17 2.602115 NA NA
20 20 20 NA NA -1.26 2.483544 NA NA
83 83 83 NA NA -1.1 2.638806 -0.21 0.4161627
24 24 24 1.098612 5.333333 -2.4 0.6029454 NA NA
18 18 18 0.6931473 4.5 NA NA NA NA
41 41 41 -0.1466035 4.021531 -1.5 2.56 -0.21 0.25
9 9 9 1.252763 5.785714 -1.9 0.4222091 NA NA
27 27 27 0.8649974 4.796052 -1.3 2.169941 NA NA
END
```

```
#initial values

list(
L.u = structure(.Data = c(
            NA,0.5,0.5,
NA,      NA,0.5 ,
NA,            NA,            NA),
.Dim = c(3,3)),
d = c(0.5, -1.5, -0.25),

delta = structure(.Data = c(
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25),
.Dim = c(22,3)),
h = c(0.1,0.1,0.1),
tau = c(0.5,0.5,0.5),
var = structure(.Data = c(
            NA,0.5,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,0.5,
            NA,0.5,            NA,            NA,0.5,
            NA,0.5,  0.5,            NA,0.5,
0.5,            NA,0.5,0.5,0.5,
            NA,            NA,            NA,0.5,0.5,
0.5,0.5,            NA,0.5,0.5,
            NA,0.5,            NA,0.5,0.5,
0.5,0.5,            NA,0.5,0.5,
            NA,0.5,0.5,            NA,   0.5,
            NA,            NA,            NA,0.5,0.5,
            NA,0.5,            NA,            NA,            NA,
            NA,            NA,0.5,            NA,            NA,
0.5),.Dim = c(22,3)),
y = structure(.Data = c(
            NA,-0.5,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,-0.5,
            NA,-0.5,            NA,            NA,-0.5,
            NA,-0.5,  -0.5,            NA,-0.5,
-0.5,            NA,-0.5,-0.5,-0.5,
            NA,            NA,            NA,-0.5,-0.5,
-0.5,-0.5,            NA,-0.5,-0.5,
            NA,-0.5,            NA,-0.5,-0.5,
-0.5,-0.5,            NA,-0.5,-0.5,
            NA,-0.5,-0.5,            NA,   -0.5,
            NA,            NA,            NA,-0.5,-0.5,
            NA,-0.5,            NA,            NA,            NA,
            NA,            NA,-0.5,            NA,            NA,
0.5),.Dim = c(22,3)))
```

C.3 WinBUGS code for multivariate meta-analysis with spherical decomposition of the between-studies covariance matrix with application to the example in rheumatoid arthritis

```
model{

corr_w[1,2]<-corrDA
corr_w[1,3]<-corrDH
corr_w[2,3]<-corrAH

#exchangeability of variances model:
for (i in 1:num){
for (j in 1:no){
#se[i,j]<--sqrt(var[i,j]/n[i,j])
var[i,j]~dnorm(0,h[j])I(0,)
}
}
for (j in 1:no){
h[j]~dgamma(1.0,0.01)
}

# Multivariate hierarchical model:
for(i in 1:num) {
prec_w[i,1:no,1:no] <- inverse(Sigma[i,1:no,1:no])
#covariance matrix for the j-th study
for (j in 1:no){
Sigma[i,j,j]<-var[i,j]/n[i,j]
}
for (j in 1:no-1){
for (k in j+1:no){
rho_w[i,j,k]<-corr_w[j,k]    #assuming the same correlations across studies
rho_w[i,k,j]<-corr_w[j,k]
Sigma[i,j,k]<-sqrt(Sigma[i,j,j])*sqrt(Sigma[i,k,k])*rho_w[i,j,k]
Sigma[i,k,j]<-Sigma[i,j,k]
}
}
y[i,1:no] ~ dmnorm(delta[i,1:no],prec_w[i,1:no,1:no]) # within-study model
delta[i,1:no] ~ dmnorm(d[1:no],prec_b[1:no,1:no]) # between-studies model
}
prec_b[1:no,1:no]<-inverse(Cov_b[,])


for (j in 1:no){
d[j] ~ dnorm(0,0.001)
Cov_b[j,j]<-tau.sq[j]
tau.sq[j]<-pow(tau[j],2)
tau[j]~dunif(0,2)
#rho[j,j]<-1
}
```

```
# assigning spherical priors to the upper triangular matrix in covariance decomposition

for (l in 1:no-1){
for (j in (l+1):no){
Cov_b[l,j]<-rho[l,j]*tau[l]*tau[j]
Cov_b[j,l]<-Cov_b[l,j]
g[j,l]<-0.00000E+00
a[l,j]~dunif(0.00000E+00,pi)
rho[l,j]<-inprod(g[,l],g[,j])
}
}

g[1,1]<-1
g[1,2]<-cos(a[1,2])
g[2,2]<-sin(a[1,2])
g[1,3]<-cos(a[1,3])
g[2,3]<-sin(a[1,3])*cos(a[2,3])
g[3,3]<-sin(a[1,3])*sin(a[2,3])

#for four outcome version:
#g[1,4]<-cos(a[1,4])
#g[2,4]<-sin(a[1,4])*cos(a[2,4])
#g[3,4]<-sin(a[1,4])*sin(a[2,4])*cos(a[3,4])
#g[4,4]<-sin(a[1,4])*sin(a[2,4])*sin(a[3,4])

pi<-3.1415

perc.acr<-exp(d[2])/(1+exp(d[2]))
}



# data is the same format as for the model with Cholesky decomposition



#initial values
list(
a = structure(.Data = c(
            NA, 1.0, 1.0,
NA,  NA, 1.0),.Dim = c(2,3)),
d = c(-1.5, 0.5, -0.25),
delta = structure(.Data = c(
-1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25,
-1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25,
-1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25,
-1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25, -1.5, 0.5, -0.25,
-1.5, 0.5, -0.25, -1.5, 0.5, -0.25),
.Dim = c(22,3)),
h = c(0.1,0.1,0.1),
tau = c(0.5,0.5,0.5),
var = structure(.Data = c(
```

```
            NA,0.5,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,0.5,
            NA,0.5,            NA,            NA,0.5,
            NA,0.5,  0.5,           NA,0.5,
0.5,            NA,0.5,0.5,0.5,
            NA,            NA,            NA,0.5,0.5,
0.5,0.5,           NA,0.5,0.5,
            NA,0.5,            NA,0.5,0.5,
0.5,0.5,           NA,0.5,0.5,
            NA,0.5,0.5,           NA,   0.5,
            NA,            NA,            NA,0.5,0.5,
            NA,0.5,            NA,            NA,            NA,
            NA,            NA,0.5,            NA,            NA,
0.5),.Dim = c(22,3)),
y = structure(.Data = c(
            NA,-0.5,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,-0.5,
            NA,-0.5,            NA,            NA,-0.5,
            NA,-0.5,  -0.5,           NA,-0.5,
-0.5,           NA,-0.5,-0.5,-0.5,
            NA,            NA,            NA,-0.5,-0.5,
-0.5,-0.5,           NA,-0.5,-0.5,
            NA,-0.5,            NA,-0.5,-0.5,
-0.5,-0.5,           NA,-0.5,-0.5,
            NA,-0.5,-0.5,           NA,   -0.5,
            NA,            NA,            NA,-0.5,-0.5,
            NA,-0.5,            NA,            NA,            NA,
            NA,            NA,-0.5,            NA,            NA,
0.5),.Dim = c(22,3)))
```

C.4 WinBUGS code for trivariate meta-analysis with in PNF (unstructured model) with application to the example in rheumatoid arthritis

```
Model {
#exchangeability of variances model:
for (k in 1:num){
for (j in 1:no){
var[k,j]~dnorm(0,h[j])I(0,)
}
}
for (j in 1:no){
h[j]~dgamma(1.0,0.01)
}

for (i in 1:num) {
Prec_w[i,1:3,1:3] <- inverse(sigma[i,1:3,1:3])
#covariance matrix for the j-th study
rho_w_12[i]<-rho12
rho_w_13[i]<-rho13
rho_w_23[i]<-rho23
sigma[i,1,1]<-var[i,1]/n[i,1]
sigma[i,2,2]<-var[i,2]/n[i,2]
sigma[i,3,3]<-var[i,3]/n[i,3]
sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w_12[i]
sigma[i,2,1]<-sigma[i,1,2]
sigma[i,1,3]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,3,3])*rho_w_13[i]
sigma[i,3,1]<-sigma[i,1,3]
sigma[i,2,3]<-sqrt(sigma[i,2,2])*sqrt(sigma[i,3,3])*rho_w_23[i]
sigma[i,3,2]<-sigma[i,2,3]
}

# Random effects model
for (i in 1:num) {
y[i,1:3]~dmnorm(delta[i,1:3], Prec_w[i,1:3,1:3]) #within-study model

#  product normal formulation for the between-studied model (unstructured covariance):
delta[i,1]~dnorm(eta1,prec1)
delta[i,2]~dnorm(eta2[i],prec2)
#eta2[i]<-lambda20+lambda21*delta[i,1]                        # no centering
eta2[i]<-lambda20+lambda21*(delta[i,1]-mean(delta[,1])) # centering on the mean
delta[i,3]~dnorm(eta3[i],prec3)
#eta3[i]<-lambda30+lambda31*delta[i,1] +lambda32*delta[i,2]  # no centering
#centering on the mean:
eta3[i]<-lambda30+lambda31*(delta[i,1]-mean(delta[,1]))+lambda32*(delta[i,2]-mean(delta[,2]))

}

#prior distributions
eta1~dnorm(0.0, 0.001)
lambda20~dnorm(0.0, 1.0E-3)
lambda30~dnorm(0.0, 1.0E-3)
corr.1.2~dunif(-0.99,0.99)
corr.2.3~dunif(-0.99,0.99)
```

```
corr.1.3~dunif(-0.99,0.99)
tau.1~dunif(0,2)
tau.2~dunif(0,2)
tau.3~dunif(0,2)
psi1.sq<-pow(tau.1,2)
prec1<-1/psi1.sq
psi2.sq<-pow(tau.2,2) - pow(lambda21,2)*pow(tau.1,2)
prec2<-1/psi1.sq
psi3.sq<-pow(tau.3,2) *(1-(pow(corr.1.3,2)+pow(corr.2.3,2)
                           -2*corr.1.2*corr.1.3*corr.2.3)/(1-pow(corr.1.2,2)))
prec3<-1/psi3.sq

mean.1<-eta1
#mean.2<-lambda20+lambda21*mean.1                      # no centering in product normal
#mean.3<-lambda30+lambda31*mean.1+lambda32*mean.2      # no centering
mean.2<-lambda20                #when centering
mean.3<-lambda30                #when centering

lambda21<-corr.1.2*tau.2/tau.1
lambda31<-(corr.1.3-corr.1.2*corr.2.3)*tau.3/tau.1/(1-pow(corr.1.2,2))
lambda32<-(corr.2.3-corr.1.2*corr.1.3)*tau.3/tau.2/(1-pow(corr.1.2,2))

perc.acr<-exp(mean.2)/(1+exp(mean.2))
}

#data
list(num=22, no=3, rho12=-0.2, rho23=0.24, rho13=-0.13)

n[,2] n[,1] n[,3] y[,2] var[,2] y[,1] var[,1] y[,3] var[,3]
26 26 26 NA NA -1.7 1.683457 -0.31 0.4410603
188 188 188 -0.1920777 4.037007 -1.6 1.88 -0.352 0.388497
810 810 810 0.4054652 4.166667 -1.9 1.96 -0.48 0.36
25 25 25 0.9444618 4.960318 NA NA NA NA
72 72 72 1.17412 5.544385 -1.47 2.337441 NA NA
30 30 30 NA NA -1.87 1.698847 NA NA
18 18 18 NA NA -2.1 1.532179 NA NA
66 66 66 NA NA -0.98 2.165142 NA NA
22 22 22 0.5596157 4.321429 NA NA -0.45 0.443637
110 110 110 NA NA -1 1.425636 NA NA
331 331 331 NA NA NA NA -0.12 0.3628
37 37 37 NA NA NA NA 0.15 0.6399
337 337 337 0.041549 4.001727 NA NA NA NA
411 411 411 NA NA NA NA NA NA
6 6 6 NA NA -1.17 2.602115 NA NA
20 20 20 NA NA -1.26 2.483544 NA NA
83 83 83 NA NA -1.1 2.638806 -0.21 0.4161627
24 24 24 1.098612 5.333333 -2.4 0.6029454 NA NA
18 18 18 0.6931473 4.5 NA NA NA NA
41 41 41 -0.1466035 4.021531 -1.5 2.56 -0.21 0.25
9 9 9 1.252763 5.785714 -1.9 0.4222091 NA NA
27 27 27 0.8649974 4.796052 -1.3 2.169941 NA NA
END
```

```
#initial values
# the initial values for the elements of the between-study covariance matrix
# need to ensure the resulting initial matrix is positive semi-definite

list(
delta = structure(.Data = c(
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25),
.Dim = c(22,3)),
h = c(0.1,0.1,0.1),
tau.1 = 0.5, tau.2 = 0.5, tau.3 = 0.5,
corr.1.2 = 0.5, corr.1.3 = 0.5, corr.2.3 = 0.0,
eta1 = 0.0,
lambda20=0.0, lambda30=0.0,
var = structure(.Data = c(
            NA,0.5,           NA,             NA,             NA,
            NA,             NA,             NA,             NA,0.5,
            NA,0.5,           NA,             NA,0.5,
            NA,0.5,  0.5,           NA,0.5,
0.5,            NA,0.5,0.5,0.5,
            NA,             NA,             NA,0.5,0.5,
0.5,0.5,            NA,0.5,0.5,
            NA,0.5,           NA,0.5,0.5,
0.5,0.5,            NA,0.5,0.5,
            NA,0.5,0.5,           NA,   0.5,
            NA,             NA,             NA,0.5,0.5,
            NA,0.5,           NA,             NA,             NA,
            NA,             NA,0.5,           NA,             NA,
0.5),.Dim = c(22,3)),
y = structure(.Data = c(
            NA,-0.5,           NA,             NA,             NA,
            NA,             NA,             NA,             NA,-0.5,
            NA,-0.5,           NA,             NA,-0.5,
            NA,-0.5,  -0.5,           NA,-0.5,
-0.5,            NA,-0.5,-0.5,-0.5,
            NA,             NA,             NA,-0.5,-0.5,
-0.5,-0.5,            NA,-0.5,-0.5,
            NA,-0.5,           NA,-0.5,-0.5,
-0.5,-0.5,            NA,-0.5,-0.5,
            NA,-0.5,-0.5,           NA,   -0.5,
            NA,             NA,             NA,-0.5,-0.5,
            NA,-0.5,           NA,             NA,             NA,
            NA,             NA,-0.5,           NA,             NA,
0.5),.Dim = c(22,3)))
```

C.5 WinBUGS code for trivariate meta-analysis with in PNF (structured model) with application to the example in rheumatoid arthritis

```
Model {
#exchangeability of variances model:
for (k in 1:num){
for (j in 1:no){
var[k,j]~dnorm(0,h[j])I(0,)
}
}
for (j in 1:no){
h[j]~dgamma(1.0,0.01)
}

for (i in 1:num) {
Prec_w[i,1:3,1:3] <- inverse(sigma[i,1:3,1:3])
#covariance matrix for the j-th study
rho_w_12[i]<-rho12
rho_w_13[i]<-rho13
rho_w_23[i]<-rho23
sigma[i,1,1]<-var[i,1]/n[i,1]
sigma[i,2,2]<-var[i,2]/n[i,2]
sigma[i,3,3]<-var[i,3]/n[i,3]
sigma[i,1,2]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,2,2])*rho_w_12[i]
sigma[i,2,1]<-sigma[i,1,2]
sigma[i,1,3]<-sqrt(sigma[i,1,1])*sqrt(sigma[i,3,3])*rho_w_13[i]
sigma[i,3,1]<-sigma[i,1,3]
sigma[i,2,3]<-sqrt(sigma[i,2,2])*sqrt(sigma[i,3,3])*rho_w_23[i]
sigma[i,3,2]<-sigma[i,2,3]
}

# Random effects model
for (i in 1:num) {
y[i,1:3]~dmnorm(delta[i,1:3], Prec_w[i,1:3,1:3]) # within-study model

# product normal formulation for the between-studies model (structured covariance):
delta[i,1]~dnorm(eta1,prec1)
delta[i,2]~dnorm(eta2[i],prec2)
#eta2[i]<-lambda20+lambda21*delta[i,1]
eta2[i]<-lambda20+lambda21*(delta[i,1]-mean(delta[,1])) # centering on the mean
delta[i,3]~dnorm(eta3[i],prec3)
#eta3[i]<-lambda30+lambda32*delta[i,2]
eta3[i]<-lambda30 +lambda32*(delta[i,2]-mean(delta[,2])) #centering

}

#prior distributions
eta1~dnorm(0.0, 0.001)
lambda20~dnorm(0.0, 1.0E-3)
lambda30~dnorm(0.0, 1.0E-3)
corr.1.2~dunif(-0.99,0.99)
corr.2.3~dunif(-0.99,0.99)
#corr.1.3~dunif(-0.99,0.99)
```

```
tau.1~dunif(0,2)
tau.2~dunif(0,2)
tau.3~dunif(0,2)
psi1.sq<-pow(tau.1,2)
prec1<-1/psi1.sq
psi2.sq<-pow(tau.2,2) - pow(lambda21,2)*pow(tau.1,2)
prec2<-1/psi2.sq
psi3.sq<-pow(tau.3,2) - pow(lambda32,2)*pow(tau.2,2)
prec3<-1/psi3.sq


mean.1<-eta1
#mean.2<-lambda20+lambda21*mean.1        # no centering in product normal
#mean.3<-lambda30+lambda32*mean.2        # no centering
mean.2<-lambda20                         #when centering
mean.3<-lambda30                         #when centering
lambda21<-corr.1.2*tau.2/tau.1
lambda32<-corr.2.3*tau.3*tau.2

perc.acr<-exp(mean.2)/(1+exp(mean.2))
}

#data
list(num=22, no=3, rho12=-0.2, rho23=0.24, rho13=-0.13)

#order of variables in the data:
#n_acr n_das n_haq log_acr_odds var_acr_lodds das28meanchange das28changevar
#haqmeanchange haqchangevar

n[,2] n[,1] n[,3] y[,2] var[,2] y[,1] var[,1] y[,3] var[,3]
26 26 26 NA NA -1.7 1.683457 -0.31 0.4410603
188 188 188 -0.1920777 4.037007 -1.6 1.88 -0.352 0.388497
810 810 810 0.4054652 4.166667 -1.9 1.96 -0.48 0.36
25 25 25 0.9444618 4.960318 NA NA NA NA
72 72 72 1.17412 5.544385 -1.47 2.337441 NA NA
30 30 30 NA NA -1.87 1.698847 NA NA
18 18 18 NA NA -2.1 1.532179 NA NA
66 66 66 NA NA -0.98 2.165142 NA NA
22 22 22 0.5596157 4.321429 NA NA -0.45 0.443637
110 110 110 NA NA -1 1.425636 NA NA
331 331 331 NA NA NA NA -0.12 0.3628
37 37 37 NA NA NA NA 0.15 0.6399
337 337 337 0.041549 4.001727 NA NA NA NA
411 411 411 NA NA NA NA NA NA
6 6 6 NA NA -1.17 2.602115 NA NA
20 20 20 NA NA -1.26 2.483544 NA NA
83 83 83 NA NA -1.1 2.638806 -0.21 0.4161627
24 24 24 1.098612 5.333333 -2.4 0.6029454 NA NA
18 18 18 0.6931473 4.5 NA NA NA NA
41 41 41 -0.1466035 4.021531 -1.5 2.56 -0.21 0.25
9 9 9 1.252763 5.785714 -1.9 0.4222091 NA NA
27 27 27 0.8649974 4.796052 -1.3 2.169941 NA NA
```

```
END

#initial values

list(
delta = structure(.Data = c(
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25, 0.5, -1.5, -0.25,
0.5, -1.5, -0.25, 0.5, -1.5, -0.25),
.Dim = c(22,3)),
h = c(0.1,0.1,0.1),
tau.1 = 0.5, tau.2 = 0.5, tau.3 = 0.5,
corr.1.2 = 0.5, corr.2.3 = 0.5,
eta1 = 0.0,
lambda20=0.0, lambda30=0.0,
var = structure(.Data = c(
            NA,0.5,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,0.5,
            NA,0.5,            NA,            NA,0.5,
            NA,0.5,  0.5,            NA,0.5,
0.5,            NA,0.5,0.5,0.5,
            NA,            NA,            NA,0.5,0.5,
0.5,0.5,            NA,0.5,0.5,
            NA,0.5,            NA,0.5,0.5,
0.5,0.5,            NA,0.5,0.5,
            NA,0.5,0.5,            NA,   0.5,
            NA,            NA,            NA,0.5,0.5,
            NA,0.5,            NA,            NA,            NA,
            NA,            NA,0.5,            NA,            NA,
0.5),.Dim = c(22,3)),
y = structure(.Data = c(
            NA,-0.5,            NA,            NA,            NA,
            NA,            NA,            NA,            NA,-0.5,
            NA,-0.5,            NA,            NA,-0.5,
            NA,-0.5,  -0.5,            NA,-0.5,
-0.5,            NA,-0.5,-0.5,-0.5,
            NA,            NA,            NA,-0.5,-0.5,
-0.5,-0.5,            NA,-0.5,-0.5,
            NA,-0.5,            NA,-0.5,-0.5,
-0.5,-0.5,            NA,-0.5,-0.5,
            NA,-0.5,-0.5,            NA,   -0.5,
            NA,            NA,            NA,-0.5,-0.5,
            NA,-0.5,            NA,            NA,            NA,
            NA,            NA,-0.5,            NA,            NA,
0.5),.Dim = c(22,3)))
```

C.6 Additional results for the example in RA

Table 14 shows results of applying TRMA of treatment effects on HAQ, ACR20 and DAS-28 using spherical decomposition and TRMA in PNF with structured covariance, along the results previously presented in Table 9 of univariate, bivariate and trivariate (with Cholesky decomposition and PNF with unstructured covariance) analyses. The results of all TRMA analyses are based on 50,000 iterations after a burn-in of 50,000 and applying thinning every 10 iteration (500,000 iterations were run after burn-in).

The results of applying TRMA with spherical decomposition of the correlation matrix resulted in similar estimates as from TRMA with Cholesky decomposition and PNF with unstructured covariance, apart from the between-studies correlation between the effects on HAQ and DAS-28 estimated with higher mean of 0.84 compared to 0.63 and 0.44 from the other two models. Results from TRMA PNF with structured covariance were almost the same as those obtained form the univariate analyses. This is due to the assumption of the effects on DAS-28 and HAQ being conditionally independent, whilst in fact in these data these were the two outcomes with the highest correlation. Changing the order of the outcomes from (DAS-28, ACR20 and HAQ and first, second and third outcomes to, for example, ACR20, DAS-28 and HAQ, would account for the correlation between the effects on DAS-28 and HAQ more effectively.

Table 14: Results of the univariate meta-analyses of HAQ and DAS-28 separately, bivariate meta-analyses of HAQ and DAS-28 and trivariate meta-analyses of HAQ, ACR20 and DAS-28. Results include mean (SD) and [95% credible interval].

	HAQ	DAS-28	ACR20	HAQ & DAS-28		TRMA of HAQ, DAS-28 & ACR20			
				BRMA	BRMA (PNF)	ChD	PNF-USC	SphD	PNF-SC
HAQ	-0.25 (0.09) [-0.43,-0.07]			-0.27 (0.08) [-0.42,-0.11]	-0.27 (0.07) [-0.39,-0.13]	-0.27 (0.08) [-0.43, -0.10]	-0.26 (0.08) [-0.42, -0.09]	-0.28 (0.08) [-0.43, -0.12]	-0.25 (0.09) [-0.43, -0.08]
DAS-28		-1.57 (0.13) [-1.84,-1.31]		-1.53 (0.13) [-1.78,-1.26]	-1.53 (0.13) [-1.78,-1.26]	-1.53 (0.13) [-1.78, -1.26]	-1.54 (0.13) [-1.80, -1.28]	-1.51 (0.13) [-1.76, -1.25]	-1.57 (0.13) [-1.82, -1.31]
ACR20‡			0.62 (0.05) [0.53, 0.72]			0.61 (0.05) [0.51, 0.71]	0.61 (0.05) [0.52, 0.71]	0.61 (0.05) [0.50, 0.71]	0.61 (0.05) [0.52, 0.71]
τ_H	0.21 (0.09) [0.10,0.44]			0.21 (0.08) [0.11,0.41]	0.22 (0.08) [0.11,0.41]	0.22 (0.08) [0.11, 0.43]	0.23 (0.09) [0.11, 0.45]	0.23 (0.08) [0.12, 0.41]	0.22 (0.09) [0.11, 0.44]
τ_D		0.44 (0.11) [0.25,0.67]		0.44 (0.11) [0.27,0.71]	0.44 (0.11) [0.27,0.71]	0.45 (0.11) [0.27,0.72]	0.44 (0.11) [0.27, 0.71]	0.46 (0.12) [0.28,0.74]	0.44 (0.11) [0.27, 0.71]
τ_A			0.52 (0.19) [0.25, 0.99]			0.56 (0.21) [0.26, 1.06]	0.56 (0.21) [0.26, 1.06]	0.58 (0.22) [0.26, 1.10]	0.54 (0.19) [0.25, 1.00]
ρ^{DH}				0.58 (0.42) [-0.51,0.99]	0.59 (0.43) [-0.51,0.99]	0.63 (0.40) [-0.45, 0.99]	0.46 (0.42) [-0.51, 0.97]	0.84 (0.28) [-0.08, 1.00]	
ρ^{AH}						-0.06 (0.43) [-0.83, 0.73]	-0.01 (0.43) [-0.82, 0.73]	-0.15 (0.46) [-0.90, 0.73]	0.03 (0.57) [-0.94, 0.95]
ρ^{DA}						-0.16 (0.40) [-0.82, 0.64]	-0.14 (0.37) [-0.78, 0.61]	-0.20 (0.43) [-0.87, 0.68]	-0.12 (0.40) [-0.80, 0.68]

ChD – Cholesky decomposition, SphD – spherical decomposition, USC (SC) – unstructured (structured) between-studies covariance matrix,
‡ transformed to proportion of responders (modelled using log odds scale).

C.7 Extension of multivariate-meta-analytic models in product normal formulation (PNF) to multiple outcomes (beyond the trivariate case)

C.7.1 MRMA PNF: unstructured model

Assuming that true treatment effects on all the outcomes are correlated, we can parameterise the between-studies model (12) in the form of product normal formulation by extending model (13) to

$$\begin{cases} \delta_{1i} \sim N(\eta_1, \psi_1^2) \\ \delta_{2i} \mid \delta_{1i} \sim N(\eta_{2i}, \psi_2^2) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\delta_{1i} \\ \delta_{3i} \mid \delta_{1i}, \delta_{2i} \sim N(\eta_{3i}, \psi_3^2) \\ \eta_{3i} = \lambda_{30} + \lambda_{31}\delta_{1i} + \lambda_{32}\delta_{2i} \\ \vdots \\ \delta_{Ni} \mid \delta_{1i}, \ldots, \delta_{(N-1)i} \sim N(\eta_{Ni}, \psi_N^2) \\ \eta_{Ni} = \lambda_{N0} + \lambda_{N1}\delta_{1i} + \cdots + \lambda_{N(N-1)}\delta_{(N-1)i}. \end{cases} \tag{31}$$

In this model designed in a Bayesian framework, we need to place prior distributions on all the parameters. Non-informative normal distributions are placed on the mean effect $\eta_1 \sim N(0, 1000)$ and the intercepts $\lambda_{20}, \ldots, \lambda_{N0} \sim N(0, 1000)$. Similarly as in the trivariate case, we place prior distributions on the between-studies correlations and between-studies standard deviations (elements of matrix T in (12) for which we are more likely to have an intuition about a reasonable range of values (compared to the slopes and conditional variances of (34)) or, in some applications, we can obtain some external information to construct informative prior distributions for them). The relationships between the model parameters (conditional variances $\psi_1^2, \psi_2^2, \ldots, \psi_N^2$ and slopes $\lambda_{21}, \lambda_{31}, \lambda_{32}, \ldots, \lambda_{N1}, \lambda_{N2}, \ldots, \lambda_{N(N-1)}$) and the between-studies parameters (correlations and standard deviations) give implied prior distributions for those parameters and also ensure that the between-studies covariance is positively defined,

$$\begin{cases} \psi_1^2 = \tau_1^2 \\ \psi_2^2 = \tau_2^2 - \lambda_{21}^2\tau_1^2 \\ \psi_3^2 = \tau_3^2 - \lambda_{31}^2\tau_1^2 - \lambda_{32}^2\tau_2^2 - 2\lambda_{31}\lambda_{32}\lambda_{21}\tau_1^2 \\ \vdots \\ \psi_N^2 = \tau_N^2 - \sum_{i=1}^{N-1} \lambda_{Ni}^2\tau_i^2 - 2\sum_{1\le i<j\le N-1} \lambda_{Ni}\lambda_{Nj}Cov(\delta_i, \delta_j). \end{cases} \tag{32}$$

$$\begin{cases} \lambda_{21} = \rho^{12}\tau_2/\tau_1 \\ \lambda_{31} = \frac{\tau_3\left(\rho^{13} - \rho^{12}\rho^{23}\right)}{\tau_1\left(1-(\rho^{12})^2\right)} \\ \lambda_{32} = \frac{\tau_3\left(\rho^{23} - \rho^{12}\rho^{13}\right)}{\tau_2\left(1-(\rho^{12})^2\right)} \\ \vdots \\ \lambda_{NQ} = \left(\rho^{QN}\tau_Q\tau_N - \sum_{P=1,P\neq Q}^{N-1} \lambda_{NP}Cov(\delta_P, \delta_Q)\right)/\tau_Q^2. \end{cases} \tag{33}$$

These relationships are obtained by calculating the variances and correlations in terms of the parameters and the rearranging the equations for the variances and solving the set of simultaneous equations for correlations. Some more details are given in Bujkiewicz et al [13]. This becomes a complex task in higher dimensions. Alternative and simpler model can be used by assuming a structure of the between-studies covariance matrix, which we describe in Section C.7.2.

C.7.2 MRMA PNF: structured model

If we imagine that N outcomes are ordered in a sequence (for example according to measurement time or other reasons that would impose such correlation structure), a conditional independence between any pair of outcomes that are not "neighbours" can be assumed, conditional on the outcomes placed in the sequence in between that particular pair.

This leads to a structure being placed on the between-studies covariance matrix in such a way to fully take into account of the correlations between the treatment effect on some pairs of outcomes (for example those that are measured one after another in a time sequence), but assume conditional independence between the effects on other pairs of outcomes. The elements of the precision matrix T^{-1} corresponding to the pairs of the effects that are conditionally independent become zero and (if assuming that only effects on consecutive outcomes are correlated) only those on diagonal and immediate off-diagonals are non-zero. The between-studies model (12) is then parameterised in the product normal,

$$\begin{cases} \delta_{1i} \sim N(\eta_1, \psi_1^2) \\ \delta_{2i} \mid \delta_{1i} \sim N(\eta_{2i}, \psi_2^2) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\delta_{1i} \\ \delta_{3i} \mid \delta_{2i} \sim N(\eta_{3i}, \psi_3^2) \\ \eta_{3i} = \lambda_{30} + \lambda_{32}\delta_{2i} \\ \vdots \\ \delta_{Ni} \mid \delta_{(N-1)i} \sim N(\eta_{Ni}, \psi_N^2) \\ \eta_{Ni} = \lambda_{N0} + \lambda_{N(N-1)}\delta_{(N-1)i}. \end{cases} \tag{34}$$

As in previous models (the trivariate and the multivariate with unstructured between-studies covariance), the parameters of the above model can be expressed in terms of the elements of the between-studies covariance matrix $\mathbf{T}$ (12):

$$\psi_1^2 = \tau_1^2, \quad \psi_2^2 = \tau_2^2 - \lambda_{21}^2\tau_1^2, \quad \psi_3^2 = \tau_3^2 - \lambda_{32}^2\tau_2^2, \quad \ldots, \quad \psi_N^2 = \tau_N^2 - \lambda_{N(N-1)}^2\tau_{N-1}^2 \tag{35}$$

and

$$\begin{cases} \lambda_{21} = \rho^{12}\frac{\tau_2}{\tau_1} \\ \lambda_{32} = \rho^{23}\frac{\tau_3}{\tau_2} \\ \vdots \\ \lambda_{N(N-1)} = \rho^{(N-1)N}\frac{\tau_N}{\tau_{(N-1)}} \end{cases} \tag{36}$$

Non-informative prior distributions are then placed on the between-studies correlations and standard deviations, $\rho^{12}, \rho^{23}, \rho^{34}, \ldots, \rho^{(N-1)N} \sim dunif(-1, 1)$, $\tau_1, \ldots, \tau_N \sim N(0, 10)I(0,)$, as well as the remaining, independent parameters, $\eta_1 \sim N(0, 1000)$, $\lambda_{20}, \ldots, \lambda_{N0} \sim N(0, 1000)$. In this from, the model is much easier to implement compared to the model with the unstructured covariance matrix in Section C.7.1.

D Supplementary materials for network meta-analysis of multiple outcomes

D.1 Additional technical details for mvNMA of multi-arm trials

The multivariate NMA extends to account for the correlation in multi-arm trials by assuming that all effects in trial i relative to a common baseline treatment in this trial are correlated and normally distributed, resulting in the full between-studies model in the following form

$$\begin{pmatrix} \begin{pmatrix} \delta_{i(bk_1)1} \\ \vdots \\ \delta_{i(bk_1)M} \end{pmatrix} \\ \begin{pmatrix} \delta_{i(bk_2)1} \\ \vdots \\ \delta_{i(bk_2)M} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \delta_{i(bk_p)1} \\ \vdots \\ \delta_{i(bk_p)M} \end{pmatrix} \end{pmatrix} \sim N \left(\begin{pmatrix} \begin{pmatrix} d_{(bk_1)1} \\ \vdots \\ d_{(bk_1)M} \end{pmatrix} \\ \begin{pmatrix} d_{(bk_2)1} \\ \vdots \\ d_{(bk_2)M} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} d_{(bk_p)1} \\ \vdots \\ d_{(bk_p)M} \end{pmatrix} \end{pmatrix}, T_{Mp\times Mp} \right) \tag{37}$$

for $(p+1)$-arm trial (p comparisons) reported on M outcomes. The between-studies covariance matrix has the following form

$$T_{Mp\times Mp} = \tag{38}$$

$$\begin{pmatrix} \begin{pmatrix} \tau_1^2 & \dots & \rho^{1M}\tau_1\tau_M \\ & \ddots & \vdots \\ & & \tau_M^2 \end{pmatrix} & \frac{1}{2}\begin{pmatrix} \tau_1^2 & \dots & \rho^{1M}\tau_1\tau_M \\ & \ddots & \vdots \\ & & \tau_M^2 \end{pmatrix} & \dots & \frac{1}{2}\begin{pmatrix} \tau_1^2 & \dots & \rho^{1M}\tau_1\tau_M \\ & \ddots & \vdots \\ & & \tau_M^2 \end{pmatrix} \\ & \begin{pmatrix} \tau_1^2 & \dots & \rho^{1M}\tau_1\tau_M \\ & \ddots & \vdots \\ & & \tau_M^2 \end{pmatrix} & \vdots & \frac{1}{2}\begin{pmatrix} \tau_1^2 & \dots & \rho^{1M}\tau_1\tau_M \\ & \ddots & \vdots \\ & & \tau_M^2 \end{pmatrix} \\ & & \ddots & \vdots \\ & & & \begin{pmatrix} \tau_1^2 & \dots & \rho^{1M}\tau_1\tau_M \\ & \ddots & \vdots \\ & & \tau_M^2 \end{pmatrix} \end{pmatrix}$$

The model assumes homogeneity of the between-studies variances and correlations across treatment contrasts. On the block diagonal of the matrix $T_{Mp\times Mp}$, the correlations are between the treatment effects on different outcomes within the same arm contrast and on the block off-diagonal the correlations are between the treatment effects on different arms and also different outcomes (except for the diagonal elements of the blocks).

This model can be simplified by representing the blocks of the multivariate distributions, corresponding to different treatment arms as a series of conditional distributions which are univariate at the arm level but multivariate at the outcome level, introducing a multivariate version of the "correction" for multi-arm trials (Dias et al [41])

$$\begin{pmatrix} \delta_{i(bk_1)1} \\ \vdots \\ \delta_{i(bk_1)M} \end{pmatrix} \sim N \left(\begin{pmatrix} d_{(bk_1)1} \\ \vdots \\ d_{(bk_1)M} \end{pmatrix}, T_{M\times M} = \begin{pmatrix} \tau_1^2 & \dots & \rho^{1M}\tau_1\tau_M \\ & \ddots & \vdots \\ & & \tau_M^2 \end{pmatrix} \right) \tag{39}$$

$$\begin{pmatrix} \delta_{i(bk_j)1} \\ \vdots \\ \delta_{i(bk_j)M} \end{pmatrix} \mid \begin{pmatrix} \begin{pmatrix} \delta_{i(bk_1)1} \\ \vdots \\ \delta_{i(bk_1)M} \end{pmatrix} \\ \begin{pmatrix} \delta_{i(bk_2)1} \\ \vdots \\ \delta_{i(bk_2)M} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \delta_{i(bk_{(j-1)})1} \\ \vdots \\ \delta_{i(bk_{(j-1)}M} \end{pmatrix} \end{pmatrix} \sim N \left(\begin{pmatrix} d_{(bk_j)1} + \frac{1}{j} \sum_{t=1}^{j-1} (\delta_{i(bk_t)1} - d_{(bk_t)1}) \\ \vdots \\ d_{(bk_j)M} + \frac{1}{j} \sum_{t=1}^{j-1} (\delta_{i(bk_t)M} - d_{(bk_t)M}) \end{pmatrix}, \frac{j+1}{2j} T_{M \times M} \right) \tag{40}$$

D.2 WinBUGS code for multivariate NMA applied to the example in multiple sclerosis

We describe below the data preparations steps required to fit bvNMA to multiple sclerosis data. This involves restructuring the data in Table 10 into three data files for the WinBUGS code:

- Datafile1 is a list specifying: the number of studies ns, number of outcomes no, total number of treatments in each network nt.total, number of arms in each study na1 and variables nobs1 and nobs2 that indicate the total number of study-arms across studies reporting one and two outcomes respectively. In our example, 11 studies have data for one outcome only, one of which is a three-arm study giving nobs1 = 23 as the total number of observations that are single. The remaining 11 studies of which one is a 3-arm study report paired outcomes given nobs2 = 23. Finally na1 is a vector of length ns that specifies the number of study arms, arranged such that, for example, the number of arms for study 1 and study 11 in datafile2 and datafile3 are 2 and 3 respectively.

- Datafile2 contains treatment indicators numbered: 1 to 8 in the network for outcome 1 (proportion remaining relapse free) and 1 to 6 in the network for outcome 2 (discontinuation due to adverse events) where 8 and 6 are the total number of interventions in network 1 and 2 respectively (Figure 5). The following variables are included: s[] is a study identifier (same as study[] in datafile2), t[,1], t[,2] and t[,3] carry treatment indicators in arm 1, 2 and 3 of study i, na2[] is the number of arms, o[,1] identifies outcome/network and o[,2] equals o[i,1] if the i^{th} row defines treatments for outcome o[i,1] and zero otherwise. This enables treatments to be numbered consecutively within each network and independent of treatment numbering in other networks.

- Datafile3 contains arm-level summary data arranged such that studies reporting only one outcome are placed on top followed by those reporting two outcomes and so forth. The variable study[] indicates the study number consecutively numbered from 1 to ns, arm[] indicates study-arm, y[,1] and y[,2] indicate mean-response for outcome 1 and 2 respectively with corresponding standard-errors se[,1] and se[,2]. Finally v[] indicates the correlation between y[,1] and y[,2] specific to treatment-arms within-studies. The variable v[] has been set to NA because the within-study correlations were not available for our illustrative example. Instead, we specified prior-distributions for the within-study correlations in the model likelihood and used this to explore the impact of various assumptions about the correlations between outcomes on parameter estimates.

```
Datafile1
list(
  ns=22,                # number of studies
  no=2,                  # number of outcomes
  nobs1 = 23,        # number of observations from studies reporting one outcome
```

```
  obs2 = 23,          # number of observations from studies report both outcomes
  nt.total = c(8,6), # total number of treatments in network 1 and network 2 respectively
  na1 = c(2,2,2,2,2,  2,2,2,2,2,  3,2,2,2,2,  2,2,2,2,2,  2,3)
                        # number of study arms in each study
)

Datafile2
s[] t[,1] t[,2] t[,3] na2 o[,1] o[,2]
1 1 2 NA 2 1 1
1 1 2 NA 2 2 0
2 1 2 NA 2 1 1
2 1 2 NA 2 2 0
3 1 7 NA 2 1 1
3 1 7 NA 2 2 0
4 1 3 NA 2 1 1
4 1 3 NA 2 2 0
5 1 4 NA 2 1 1
5 1 4 NA 2 2 0
6 1 8 NA 2 1 1
6 1 8 NA 2 2 0
7 1 5 NA 2 1 1
7 1 5 NA 2 2 0
8 3 4 NA 2 1 1
8 3 4 NA 2 2 0
9 2 3 NA 2 1 1
9 2 3 NA 2 2 0
10 4 5 NA 2 1 1
10 4 5 NA 2 2 0
11 3 4 5 3 1 1
11 3 4 5 3 2 0
12 1 2 NA 2 1 1
12 1 2 NA 2 2 2
13 1 2 NA 2 1 1
13 1 2 NA 2 2 2
14 1 2 NA 2 1 1
14 1 2 NA 2 2 2
15 1 3 NA 2 1 1
15 1 3 NA 2 2 2
16 1 3 NA 2 1 1
16 1 3 NA 2 2 2
17 1 5 NA 2 1 1
17 1 5 NA 2 2 2
18 2 4 NA 2 1 1
18 2 4 NA 2 2 2
19 2 5 NA 2 1 1
19 2 5 NA 2 2 2
20 2 5 NA 2 1 1
20 2 5 NA 2 2 2
21 3 5 NA 2 1 1
21 3 5 NA 2 2 2
22 1 4 6 3 1 1
22 1 4 6 3 2 2
```

```
END

Data file 3

study[] arm[] y[,1] y[,2] se[,1] se[,2] v[]
1 1 -0.033 NA 0.183 NA NA
1 2 0.219 NA 0.184 NA NA
2 1 1.036 NA 0.248 NA NA
2 2 1.043 NA 0.121 NA NA
3 1 0.642 NA 0.098 NA NA
3 2 1.208 NA 0.077 NA NA
4 1 1.149 NA 0.318 NA NA
4 2 1.253 NA 0.327 NA NA
5 1 -0.134 NA 0.518 NA NA
5 2 1.012 NA 0.584 NA NA
6 1 1.046 NA 0.102 NA NA
6 2 1.545 NA 0.116 NA NA
7 1 -1.609 NA 1.095 NA NA
7 2 -0.336 NA 0.414 NA NA
8 1 -0.083 NA 0.109 NA NA
8 2 0.243 NA 0.109 NA NA
9 1 1.358 NA 0.154 NA NA
9 2 1.046 NA 0.144 NA NA
10 1 -2.286 NA 0.429 NA NA
10 2 -4.844 NA 1.42 NA NA
11 1 0.268 NA 0.368 NA NA
11 2 -1.386 NA 0.456 NA NA
11 3 -0.268 NA 0.368 NA NA
12 1 -0.362 -2.146 0.107 0.171 NA
12 2 -0.754 -2.2 0.115 0.178 NA
13 1 -1.041 -3.807 0.475 1.43 NA
13 2 0.241 -2.442 0.403 0.737 NA
14 1 -0.995 -4.828 0.201 1.004 NA
14 2 -0.681 -3.178 0.189 0.456 NA
15 1 0.452 -3.122 0.097 0.234 NA
15 2 0.796 -2.785 0.102 0.202 NA
16 1 -1.023 -4.256 0.243 0.712 NA
16 2 -0.505 -3.071 0.224 0.387 NA
17 1 -1.83 -4.804 0.261 1.004 NA
17 2 -1.279 -2.434 0.218 0.33 NA
18 1 -0.488 -2.939 0.111 0.235 NA
18 2 -0.492 -2.759 0.113 0.215 NA
19 1 0.361 -4.007 0.096 0.357 NA
19 2 0.322 -4.22 0.068 0.279 NA
20 1 0.934 -4.344 0.356 1.423 NA
20 2 0.111 -3.555 0.334 1.014 NA
21 1 -0.581 -4.511 0.217 1.005 NA
21 2 0.042 -2.901 0.204 0.459 NA
22 1 -1.655 -5.226 0.199 1.003 NA
22 2 -0.751 -3.23 0.158 0.385 NA
22 3 -0.995 -4.127 0.164 0.582 NA
END
```

```
model{

#Likilihood for studies reporting a single outcome (arm level data)
for(i in 1:nobs1){
    tmp[i] <- v[i] #dummy variable within-study corr in datafile3
    omega1[i,1] <- pow(se[i,1],-2) #precision
    y[i,1] ~ dnorm(mean.y[study[i],arm[i],1],omega1[i,1]) # Normal dist.
}

#Likilihood for studies reporting two outcomea (arm level data )
rhoW ~ dunif(-1,1)  # uniform(a,b) prior distr. for within-study corr.
for(i in (nobs1+1):(nobs1+nobs2)){
    omega2[i,1:no,1:no] <- inverse(cov.mat[i,,])
    y[i,1:no] ~ dmnorm(mean.y[study[i],arm[i],1:no],omega2[i,,])#mvNorm distr.

#Define within-study covariance matrix
    cov.mat[i,1,1] <- pow(se[i,1],2)
    cov.mat[i,2,2] <- pow(se[i,2],2)
    cov.mat[i,1,2] <- se[i,1]*se[i,2] *rhoW
    cov.mat[i,2,1] <- cov.mat[i,1,2]
}

#Transform unobserved effects to one-row per study and take contrasts
for(j in 1:ns) {
    for(k in 1:na1[j]) {
        for(m in 1:no)    {
            mean.y[j,k,m] <- mu[j,m] + delta[j,k,m]
        }
    }
}

#Multivariate random-effects (homogenous variance model)
for(j in 1:ns){
    for(m in 1:no) {
        delta[j,1,m] <-0 #set delta in treatment 1/baseline to zero
        w[j,1,m] <-0 #set multi-arm adjustment in trt 1 to zero
    }
    for(k in 2:na1[j]){
        delta[j,k,1:no] ~ dmnorm(md[j,k,1:no],PREC[j,k,1:no,1:no])

#Set precision matrix T  for multiple outcomes
        for(m in 1:no) {
            for(mm in 1:no) {
                PREC[j,k,m,mm] <- T[m,mm]*2*(k-1)/k
            }
        }
    }
}

#Consistency relations between basic parameters
for(i in 1:(ns*no)) {
```

```
  for(k in 2:na2[i])  {
     md[s[i],k,o[i,1]] <- (d[o[i,1],t[i,k]] - d[o[i,1],t[i,1]]) *equals(o[i,1],o[i,2])
                          + sw[s[i],k,o[i,1]]
     w[s[i],k,o[i,1]] <- (delta[s[i],k,o[i,1]] - (d[o[i,1],t[i,k]] - d[o[i,1],t[i,1]]))
                          *equals(o[i,1],o[i,2])
     sw[s[i],k,o[i,1]]  <- sum(w[s[i],1:k-1,o[i,1]])/(k-1)
  }
}

############################################################
#Constraints.
#set effect in trt 1 on both outcome to zero.
#outcome 2 has 8 treatments buts outcome 2 has only 6 treatments#
#so need to set d[2,7] and d[2,1]
d[1,1] <-0
d[2,1] <-0
d[2,7] <-0
d[2,8] <-0
d[1,1] <-0

#trt effects exponentiated and prior distributions
for(m in 1:no) {
    sd[m] ~ dunif(0, 5)
    sigma[m,m] <- pow(sd[m],2)

    for(k in 2: nt.total[m]){
        or[m,k] <- exp(d[m,k])
        d[m,k]  ~ dnorm(0,0.0001)
    }
}

#Prior distributions and estimated between study correlation matrix
for(j in 1:ns){
    for(m in 1:no) {
        mu[j, m] ~ dnorm(0,0.0001)
    }
}

#Parameterization of the between-studies covariance based
# on the separation strategy by spherical decomposition (Wei et al 2013)
T[1:no,1:no] <- inverse(sigma[,])
pi <- 3.1415
for(i in 1:2) {
    for(j in (i+1):no) {
        sigma[i,j] <- rhoB[i,j]*sd[i]*sd[j]
        sigma[j,i] <- sigma[i,j]
        g[j,i] <- 0
        a[i,j] ~ dunif(0, pi)
        rhoB[i,j] <- inprod(g[,i], g[,j])
    }
}
```

```
g[1,1] <- 1
g[1,2] <- cos(a[1,2])
g[2,2] <- sin(a[1,2])

#Additional parameterization in 3-outcome problem
#g[1,3] <- cos(a[1,3])
#g[2,3] <- sin(a[1,3])*cos(a[2,3])
#g[3,3] <- sin(a[1,3])*sin(a[2,3])

#pairwise ORs and LORs for all possible pair-wise comparisons, if nt>2
#outcome 1
for (c in 1:(nt.total[1]-1)) {
    for (k in (c+1):nt.total[1]) {
        or2[c,k]     <- exp(d[1,k] - d[1,c])
        lor1[c,k] <- (d[1,k]-d[1,c])
    }
}

# outcome 2
for (c in 1:(nt.total[2]-1)) {
    for (k in (c+1):nt.total[2]) {
        or2[c,k]     <- exp(d[2,k] - d[2,c])
        lor2[c,k] <- (d[2,k]-d[2,c])
    }
    }
}
```