Special Issue Paper

Statistics in Medicine

Received 31 October 2014, Accepted 5 October 2015 Published online 3 November 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6776

Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process

Sylwia Bujkiewicz,[a*] John R. Thompson,[b] Richard D. Riley[c] and Keith R. Abrams[a]

A number of meta-analytical methods have been proposed that aim to evaluate surrogate endpoints. Bivariate meta-analytical methods can be used to predict the treatment effect for the final outcome from the treatment effect estimate measured on the surrogate endpoint while taking into account the uncertainty around the effect estimate for the surrogate endpoint. In this paper, extensions to multivariate models are developed aiming to include multiple surrogate endpoints with the potential benefit of reducing the uncertainty when making predictions. In this Bayesian multivariate meta-analytic framework, the between-study variability is modelled in a formulation of a product of normal univariate distributions. This formulation is particularly convenient for including multiple surrogate endpoints and flexible for modelling the outcomes which can be surrogate endpoints to the final outcome and potentially to one another. Two models are proposed, first, using an unstructured between-study covariance matrix by assuming the treatment effects on all outcomes are correlated and second, using a structured between-study covariance matrix by assuming treatment effects on some of the outcomes are conditionally independent. While the two models are developed for the summary data on a study level, the individual-level association is taken into account by the use of the Prentice's criteria (obtained from individual patient data) to inform the within study correlations in the models. The modelling techniques are investigated using an example in relapsing remitting multiple sclerosis where the disability worsening is the final outcome, while relapse rate and MRI lesions are potential surrogates to the disability progression. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: Bayesian analysis; multivariate meta-analysis; multiple outcomes; surrogate endpoints; multiple sclerosis

1. Introduction

Surrogate endpoints are increasingly being investigated as candidate endpoints in randomised controlled trials where measuring a primary outcome of interest may be too costly, too difficult to measure or require long follow-up time. Prior to the use of surrogate endpoint for trial design or decision-making for regulatory or reimbursement purposes, such endpoints need to be validated. The validation takes place on three levels: by establishing a biological plausibility of the association between outcomes, assessing association between outcomes at the individual level and validating the surrogate endpoints at the study level to assess them as predictors of clinical benefit measured by the final outcome [1]. It has been established that methods based on a single clinical trial are not sufficient and surrogate endpoints have to be validated based on a number of clinical trials in a meta-analytic framework [1, 2].

A number of meta-analytical methods have been proposed that aim to evaluate treatment effects on surrogate endpoints as predictors of the effect on a target outcome. Methods by Buyse *et al*. [3] were

[a]*Biostatistics Research Group, Department of Health Sciences, University of Leicester, University Road, Leicester, LE1 7RH, U.K.*
[b]*Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, University Road, Leicester, LE1 7RH, U.K.*
[c]*Research Institute of Primary Care and Health Sciences, Keele University, Staffordshire, ST5 5BG U.K.*
[*]*Correspondence to: Sylwia Bujkiewicz, Department of Health Sciences, University of Leicester, University Road, Leicester, LE1 7RH, U.K.*
[†]*E-mail: sb309@le.ac.uk*

developed to model surrogate endpoints at the arm level by extending the ideas developed by Prentice [4] to a meta-analytic framework, while those by Daniels and Hughes [2] are focussed on modelling the relationship between relative treatment effects on outcomes. The former are developed in a frequentist approach while the latter in a Bayesian framework. Various extensions to meta-analytic approaches to evaluating surrogate endpoints have been developed, for example, for the time-to-event data by Burzykowki *et al*. [5] extended to a Bayesian framework by Renfro *et al*. [6].

Most methods developed to date are designed to evaluate single surrogate endpoints. In the summary of a National Institutes of Health Workshop on the use of surrogate endpoints, Gruttola *et al*. [7] made a number of recommendations for future research that included, for example, development of models that can accommodate measurement error, missing data and multiple surrogate endpoints and/or multiple clinical outcomes. Methods for evaluating multiple surrogate endpoints were proposed by Xu and Zeger [8] for time-to-event data modelled jointly with multiple biomarkers measured longitudinally but were mainly limited to individual-level data. 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 [9, 10] and relapse rate and number of active lesions in the brain as predictors of disability progression in relapsing remitting multiple sclerosis (RRMS) [11], which used methods that did not allow for inclusion of the measurement error for the surrogate endpoint.

Bivariate meta-analytical methods can be used to predict the treatment effect on the target outcome from the effect on surrogate endpoint (while taking into account the uncertainty around the treatment effect on surrogate endpoint ignoring of which can impact on predictions [12]) as well as to combine evidence on treatment effect on both outcomes by 'borrowing of strength' across outcomes when evaluating new health technologies [13, 14]. Extending such methods to multivariate models can be used to evaluate multiple surrogate endpoints as joint mediators of clinical benefit with a potential advantage of increasing the precision of predictions. Multivariate meta-analysis models require within-study correlations between multiple effect estimates which are usually not available but need to be accounted for [13, 15]. If individual patient data (IPD) are not available, but results of the Prentice's criteria for surrogacy are, then we propose to use those criteria to obtain the within-study correlations.

Models considered in this paper build on a multivariate meta-analysis model of mixed outcomes developed by Bujkiewicz *et al*. [16], where the between-study covariance is parameterised in a formulation of a product of univariate normal distributions. In this model, an assumption of conditional independence between outcomes was used to simplify the model by putting a structure on the between-study covariance matrix. Building on this model, we extend it to two alternative models. In the first case, the assumption is relaxed to allow for full unstructured covariance to be modelled in a product normal formulation, where multiple surrogate endpoints and the final outcome are correlated and hence one surrogate endpoint can also act as a surrogate to the other. In the second case, we propose an alternative parameterisation to the one by Bujkiewicz *et al*. [16], which also assumes conditional independence between some of the outcomes (by putting a structure on the between-study covariance) but is more suitable for defining criteria for surrogacy. In both models, the product normal parameterisation offers the possibility of describing the criteria for surrogacy in a greater detail compared with providing a between-study correlation only, which would be the case when modelling the covariance structure directly. The modelling techniques are investigated using example in RRMS, where the disability worsening is the final outcome, while relapse rate and number of active MRI lesions have been considered potentially good surrogates to the disability progression [11, 17].

In the remainder of this paper, the two models (with unstructured and structured covariance matrix) are introduced in Section 2, where in addition to this a model for obtaining the within-study correlation from published Prentice's criteria is described. In Section 2.7, the use of software is briefly reported, which is then followed by the application of the methods to data specific to the example in RRMS which is presented together with the results in Section 3. Simulation study investigating the robustness of the methods to the normality assumption as well as the performance of the methods is described in Section 4. Extensions of the two approaches to the multivariate case are developed in Section 5. The paper concludes with a discussion of methods in Section 6.

2. Trivariate random effects meta-analysis with application to surrogate endpoints

Suppose in each study i, we have three estimates of treatment effect observed on each of the three outcomes Y_{1i}, Y_{2i} and Y_{3i}, where Y_3 is the treatment effect estimate for the final clinical outcome, while

Y_1 and Y_2 are intermediate surrogate endpoints. Assuming the treatment effect estimates for the three outcomes in each study have a trivariate normal sampling distribution, the model can be written as

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \\ Y_{3i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \\ \mu_{3i} \end{pmatrix}, \mathbf{\Sigma_i}\right), \ \mathbf{\Sigma_i} = \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi}^{12} & \sigma_{1i}\sigma_{3i}\rho_{wi}^{13} \\ \sigma_{2i}\sigma_{1i}\rho_{wi}^{12} & \sigma_{2i}^2 & \sigma_{2i}\sigma_{3i}\rho_{wi}^{23} \\ \sigma_{3i}\sigma_{1i}\rho_{wi}^{13} & \sigma_{3i}\sigma_{2i}\rho_{wi}^{23} & \sigma_{3i}^2 \end{pmatrix} \tag{1}$$

$$\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \\ \mu_{3i} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \mathbf{T}\right), \ \mathbf{T} = \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho_b^{12} & \tau_1\tau_3\rho_b^{13} \\ \tau_2\tau_1\rho_b^{12} & \tau_2^2 & \tau_2\tau_3\rho_b^{23} \\ \tau_3\tau_1\rho_b^{13} & \tau_3\tau_2\rho_b^{23} & \tau_3^2 \end{pmatrix}. \tag{2}$$

In the aforementioned model, Y_{1i}, Y_{2i} and Y_{3i} are assumed to be estimates of the correlated true treatment effects μ_{1i}, μ_{2i} and μ_{3i} with corresponding within-study covariance matrices $\mathbf{\Sigma_i}$ of the estimates (comprising of the within-study correlations ρ_{wi}^{jk} between the estimates Y_{ji} and Y_{ki} and within-study standard deviations σ_{ji} corresponding to each estimate Y_j, for each outcome $j = 1, 2, 3$ and study i). These true study-level effects are assumed to follow a trivariate normal distribution with means $(\beta_1, \beta_2, \beta_3)$ and covariance $\mathbf{T}$ (comprising of the between-study correlations ρ_b^{jk} between the true treatment effects μ_j and μ_k and between-study standard deviations τ_j corresponding to each true effect μ_{ji}, $j = 1, 2, 3$ in each study i). In this hierarchical framework, Equations (1) and (2) describe the within-study and the between-study models, respectively.

In the remainder of this Section, the procedure for the use of the aforementioned model for the purpose of the validation of the surrogate endpoints is described in Section 2.1. The estimation of the variances for missing outcomes (omitted for the purpose of the validation) and the within-study correlations (which are not reported for any of the studies but can be obtained from reported individual-level surrogacy criteria) is described in Section 2.2. This is followed by setting out scenarios for modelling surrogate endpoints, by the use of an unstructured or structured between-study covariance matrix, in Section 2.3 followed by the details of the modelling in the product normal formulation with the unstructured covariance matrix T in Section 2.4 and structured covariance in Section 2.5. The section is concluded by defining surrogacy criteria for the models in Section 2.6 and remark on the software use in Section 2.7.

2.1. Application of multivariate meta-analysis to evaluating surrogate endpoints

2.1.1. Using multivariate meta-analysis to make predictions from surrogate outcomes. Traditionally, multivariate meta-analysis is used to estimate pooled effects for multiple outcomes from multiple studies by taking into account the correlation between the outcomes. Studies included in a multivariate meta-analysis may all report all of the outcomes, or alternatively, it is permitted that some of the studies report only a subset of outcomes under a missing at random assumption. In the latter case, the unreported outcomes can be predicted for each of the studies by taking into account the correlation between the outcomes. In a Bayesian framework, this can be achieved by coding the unreported outcomes as missing, which are then predicted by the Markov chain Monte Carlo (MCMC) simulation of the model (1)–(2) [16]. This framework can be adopted to validating surrogate endpoints (at a study level).

2.1.2. Using cross-validation to examine the accuracy of predictions from multivariate meta-analysis. The study-level validation is conducted in the form of cross-validation similar to the 'leave-one-out' approach as for example in Daniels and Hughes [2], except instead of taking one study out, only the estimate for the final outcome in one study is taken out. For all of the studies in the data set, one study at a time, the following procedure is conducted. For a given study, the estimate of the treatment effect on the final outcome is omitted (and coded as missing as if this particular study was a new study with the final outcome yet unknown). Then the multivariate meta-analytic framework (1)–(2) is applied to the data including the study with the estimates of the effects on candidate surrogate endpoints known but on the final outcome missing. This missing effect is then predicted by the MCMC simulation from observed effect(s) on surrogate endpoint(s) by taking into account the data on all outcomes from the remaining studies and the relationship between the effects on all outcomes defined by the multivariate meta-analytic model. The observed estimate is then compared with the predicted value by checking

whether the value of the observed estimate falls within the predicted interval. To see if this framework will give reliable predictions for the omitted final outcome in one study, this procedure is repeated for each study i ($i = 1, \ldots, I$), giving I comparisons of how the predictions from the surrogate endpoints using the multivariate modelling approach perform. Ideally, we want to compare the performance of the model in predicting the true effect μ_{3n} in a new study n. But we do not know what the true effect is, so instead, we predict $\hat{Y}_{3n}$ and then compare the estimate Y_{3n} with the predicted interval with the variance $\sigma^2_{3n} + var(\hat{\mu}_{3n}|Y_{1n}, Y_{2n}, \sigma_{1n}, \sigma_{2n}, Y_{1(-n)}, Y_{2(-n)}, Y_{3(-n)})$, where $Y_{1(2,3)(-n)}$ denote the data from the remaining studies without the validation study n.

Post the validation process, when endpoints are established as surrogate endpoints to the final outcome, this multivariate meta-analytic framework can be used to predict treatment effect on the final outcome from those on the surrogate endpoints in a new study, for example, for the purpose of early decision making process, in particular when the treatment effect(s) on surrogate endpoint(s) is(are) measured early compared with the final outcome.

2.2. *Obtaining within-study variances and covariances*

The within-study variability in each study i is represented by the within-study covariance matrices Σ_i in model (1) which are assumed to be known. In practice, only the variances can be obtained from the reported data by calculating standard errors squared, while the within-study correlations between the treatment effect estimates are usually not reported. Also the within-study variance of the treatment effect for the final outcome in the study in which this effect is omitted in the cross-validation (Section 2.1) will be unknown at the validation stage. In this case, the variance will be coded as missing and a distribution has to be placed over a missing node (required if using WinBUGS for MCMC simulation). Uniform prior distribution was placed on the missing variance, $\sigma^2_{1,(2,3)} \sim unif(0.001, 1000)$, which was recommended as a non-informative prior distribution for variances by Lambert *et al* [18] (who also list other suitable prior distributions).

2.2.1. Within-study covariances. Calculation of the within-study covariances between estimates Y_{ji} and Y_{ki}, $\Sigma_i[j,k]$, for each pair of outcomes j, k in each study i requires knowledge of the estimate of the within-study correlations ρ^{jk}_{wi}. These correlations between estimates of the treatment effects can be obtained by bootstrapping [2] (or double bootstrapping for correlation with uncertainty [16]) of the IPD from all of the studies or a subset of studies in the meta-analysis. Here, we consider an alternative approach where IPD is not available for any of the studies, but surrogacy on individual level has been investigated and reported by the use of Prentice's criteria.

To obtain the within-study correlations ρ^{jk}_{wi} between each pair of the treatment effect estimates Y_{1i}, Y_{2i} and Y_{3i}, calculated as a difference between two measurements in the experimental and control arms, we adopt an approach similar to the one developed by Wei and Higgins [19] who used a bivariate delta method to express the within-study covariance between treatment effects, such as, for example, log odds ratios, in terms of the covariances between outcomes, such as the probabilities (or risks) and using a correlation between the outcomes from the literature. Here, this approach is adopted in a simpler form by representing the within-study covariance between treatment effects in terms of the variances and correlations between the effects on specific treatment arms. In the case considered here, this is sufficient (and does not require the use of the bivariate delta method) as the correlations between effects on arm level can be obtained from Prentice's criteria of association of outcomes modelled on the absolute normal scale. However, to estimate the variances of effects in each arm, sufficient data need to be available, such as two by two table (or log odds of event with corresponding standard error) for Binomial outcomes or event rates and number of individuals (or log rates with corresponding standard errors) for Poisson outcomes.

The within-study covariances between treatment effects on each pair of normally distributed outcomes j, k ($j \neq k; j, k = 1, 2, 3$) can be expressed in terms of the covariances between arm-specific effects:

$$\begin{aligned}\Sigma_i[j,k] = \Sigma_i[k,j] &= Cov\left(Y_{ji}, Y_{ki}\right) \\ &= Cov\left(X_{eji} - X_{cji}, X_{eki} - X_{cki}\right) \\ &= Cov\left(X_{eji}, X_{eki}\right) - Cov\left(X_{eji}, X_{cki}\right) - Cov\left(X_{cji}, X_{eki}\right) + Cov\left(X_{cji}, X_{cki}\right) \\ &= Cov\left(X_{eji}, X_{eki}\right) + Cov\left(X_{cji}, X_{cki}\right)\end{aligned} \tag{3}$$

where Y_{ji} (Y_{ki}) is the treatment difference on outcome j (k) (such as a mean difference or log odds ratio), X_{aji} (X_{aki}) are the absolute treatment effects on outcome j (k) (such as mean or log odds) in arm a (where e and c stand for an experimental and control arm, respectively) in study i and

$$Cov(X_{aji}, X_{aki}) = \sqrt{Var(X_{aji})Var(X_{aki})}\rho^*_{jki} \tag{4}$$

where ρ^*_{jki} is the correlation between the absolute effects measured on outcomes j and k in each study i, which can be obtained from reported surrogacy criteria as discussed in the succeeding section.

2.2.2. Within-study correlations. Taking into account the above derivation of the within-study covariance, the within-study correlation can be obtain from

$$\rho^{jk}_{wi} = \frac{\Sigma_i[j,k]}{\sqrt{\Sigma_i[j,j]\Sigma_i[k,k]}} = \frac{\left(\sqrt{Var(X_{eji})Var(X_{eki})} + \sqrt{Var(X_{cji})Var(X_{cki})}\right)\rho^*_{jki}}{\sqrt{\sigma^2_{ji}\sigma^2_{ki}}}, \tag{5}$$

where, as noted in the previous section, it is assumed that the variances of $X_{aj(k)i}$ can be obtained from the data (for example from two by two tables for binary outcomes). Therefore, only the correlations ρ^*_{jki} between the effects on arm level need to be estimated in order to calculate the within-study correlations, ρ^{jk}_{wi}, between treatment effects on each pair of outcomes j and k. Consider that X_{ji} is an average measurement on surrogate endpoint S_i, that is, $X_{ji} = E(S_i)$ and X_{ki} is an average measurement on final endpoint F_i, that is, $X_{ki} = E(F_i)$, and ρ^*_{jki} is the correlation between those average effects. Noting that for normally distributed outcomes, the correlation between the normally distributed individual observations equals the correlation between the means ensures that ρ^*_{jki} should equal the correlation between the individual-level responses for S and F that can be obtained by modelling individual patient data.

When investigating the association on individual level between surrogate endpoint S_{im} and final outcome F_{im} for patient m under treatment Z_{im} in a study i, Prentice's criteria can be used to evaluate surrogacy. The four Prentice's criteria require that (i) there is significant treatment effect on the surrogate endpoint, (ii) there is significant treatment effect on the final outcome, (iii) surrogate endpoint has a significant impact on the final outcome and (iv) the effect of treatment on the final outcome is fully mediated by the surrogate endpoint [1, 4]. Similarly as by Burzykowski, Molenberghs and Buyse [1], the first two criteria for the normally distributed outcomes can be written for each study i in the following model:

$$\begin{aligned} S_{im} &= \mu_{Si} + \alpha_i Z_{im} + \epsilon_{Sim}, \\ F_{im} &= \mu_{Fi} + \beta_i Z_{im} + \epsilon_{Fim} \end{aligned} \tag{6}$$

with correlated error structure

$$\Omega = \begin{pmatrix} \omega_{iSS} & \omega_{iSF} \\ & \omega_{iFF} \end{pmatrix},$$

and hence, the correlation between S_i and F_i, referred to by Buyse and Mollenberghs [20] as adjusted association (between S_i and F_i after adjustment for the treatment Z_i), equals $\rho_{ZiSF} = \frac{\omega_{iSF}}{\sqrt{\omega_{iSS}\omega_{iFF}}}$. This is the correlation in the individual-level response for S and F which as noted above equals the correlation between the means, in this case between $E(S_{im})$ and $E(F_{im})$ and hence between X_{ji} and X_{ki}. Thus the correlation ρ_{ZiSF} equals the correlation ρ^*_{jki} between the absolute effects in (4) if, for example, $j = S$ and $k = F$. Also the average treatment effects α_i and β_i on the two outcomes are the treatment effect estimates Y_{ji} and Y_{ki} in study i. The correlation ρ_{ZiSF} is required to estimate the within-study correlation ρ^{jk}_{wi}, which can be obtained from the fourth Prentice's criteria [4] which Burzykowski, Molenberghs and Buyse [1] propose to verify through the conditional distribution of the final outcome conditional on the treatment and the surrogate endpoint,

$$F_{im} = \tilde{\mu}_{iF} + \beta_{iS} Z_{im} + \gamma_{iZ} S_{im} + \tilde{\epsilon}_{Fim}, \tag{7}$$

where $\beta_{iS} = \beta_i - \alpha_i\omega_{iFS}/\omega_{iSS}$ and $\gamma_{iZ} = \omega_{iFS}/\omega_{iSS}$. This criterion requires that if all treatment effect is mediated by the surrogate endpoint S then β_{iS} should be zero. Rearranging the terms for β_{iS} gives $\omega_{iFS} = \omega_{iSS}(\beta_i - \beta_{iS})/\alpha_i$, which substituted into the formula for the adjusted association ρ_{ZiSF} gives

$$\rho_{ZiSF} = \frac{\beta_i - \beta_{iS}}{\alpha_i}\sqrt{\omega_{iSS}/\omega_{iFF}}. \tag{8}$$

The parameters α_i, β_i, ω_{iSS} and ω_{iFF} can be obtained by fitting model (6) to the patient data on responses S and F.

To complete the individual-level validation, the third Prentice's criteria can be verified by

$$F_{im} = \mu_i + \gamma_i S_{im} + \epsilon_{im}, \tag{9}$$

with further details discussed by Burzykowski, Molenberghs and Buyse [1].

In the trivariate meta-analysis, the within-study correlations between treatment effects need to be estimated for each pair of outcomes. To do this using the above criteria, one of the outcomes in each pair is considered a surrogate endpoint to other outcome. Ideally, the within-study correlations (or the Prentice's criteria from which the correlations derive) should be obtained from the IPD from each study in the meta-analysis. When IPD is available from each of the studies, the within-study correlation can be calculated directly. However, in the absence of the data, it may be possible to obtain the Prentice's criteria corresponding to the data from each (or some) of the studies (or some of the studies in the meta-analysis). The Prentice's criteria provide both the assessment of surrogacy on the individual level as well as sufficient information to obtain the within-study correlations. This approach is illustrated in Section 3, where the Prentice's criteria published in literature for one of the studies included in the meta-analysis for the motivating example in RRMS are used to estimate the within-study correlation. Note that establishing the within-study correlation does not guarantee causal relationship between the treatment effects on surrogate and final outcomes. Research developments on causality of surrogate endpoints are highlighted in the Discussion section.

2.3. Scenarios for modelling of surrogate endpoints

Two scenarios for modelling the surrogate endpoints are considered here. In the first scenario, true treatment effects, μ_{ji}, on all the outcomes are assumed correlated as shown in Figure 1a. In this case, the between-study covariance matrix T in the model (2) is unstructured and can be modelled directly, by either placing inverse Wishart prior distribution on T, by Cholesky or spherical decomposition [21] (and placing separate prior distributions on the between-study correlations and standard deviations) or by re-parameterising the between-study model in the product normal formulation of the series of univariate conditional distributions. The latter approach has a number of advantages. In contrast to modelling it directly by the use of Wishart prior distribution, the product normal formulation allows direct control over the prior distributions on all elements of the between-study covariance matrix (between-study

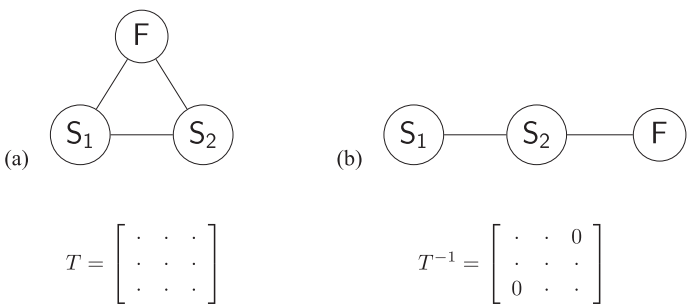

Figure 1. Scenarios for modelling surrogates endpoints: (a) all outcomes correlated giving unstructured covariance matrix T, (b) final outcome conditionally independent from the first surrogate endpoint conditional on the second giving structured covariance matrix equivalent to the precision matrix T^{-1} with element [1, 3] equal to zero.

standard deviations and correlations). It also allows to describe some association criteria which helps to assess the surrogacy in more detail compared with obtaining only between-study correlation.

In the second scenario, shown in Figure 1b, in which the treatment effects on multiple outcomes may (but do not have to) be assumed to be measured sequentially in time, assumption is made that the treatment effect on the final outcome is conditionally independent from the effect on the first surrogate endpoint conditional on the effect on the second surrogate. This assumption puts a structure on the between-study covariance matrix T resulting in element $[1,3]$ of the precision matrix T^{-1} being equal to zero. This approach leads to a reduced number of parameters to estimate and is easier to implement, in particular when dealing with multiple outcomes beyond the trivariate case (for details see Sections 5.1 and 5.2).

2.4. *Product normal formulation with unstructured covariance matrix*

For the first scenario of the true treatment effects μ_{ji} on all outcomes being correlated, represented graphically in Figure 1a, the between-study covariance has an unstructured form. The between study model (2) for this scenario is re-parameterised in the product normal formulation of the series of univariate conditional distributions:

$$\begin{cases} \mu_{1i} \sim N\left(\eta_1, \psi_1^2\right) \\ \mu_{2i} \mid \mu_{1i} \sim N\left(\eta_{2i}, \psi_2^2\right) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\mu_{1i} \\ \mu_{3i} \mid \mu_{1i}, \mu_{2i} \sim N\left(\eta_{3i}, \psi_3^2\right) \\ \eta_{3i} = \lambda_{30} + \lambda_{31}\mu_{1i} + \lambda_{32}\mu_{2i}. \end{cases} \tag{10}$$

Instead of placing independent non-informative prior distributions on all the parameters and hyper-parameters of the model, relationships between these parameters and the elements of the between-study covariance matrix are derived to allow to take into account the inter-relationship between the parameters. These relationships have the following forms:

$$\begin{aligned} &\psi_1^2 = \tau_1^2, \quad \psi_2^2 = \tau_2^2 - \lambda_{21}^2\tau_1^2, \quad \psi_3^2 = \tau_3^2 - \lambda_{31}^2\tau_1^2 - \lambda_{32}^2\tau_2^2, \\ &\lambda_{21} = \frac{\tau_2}{\tau_1}\rho_b^{12}, \quad \lambda_{31} = \rho_b^{13}\frac{\tau_3}{\tau_1} - \lambda_{32}\lambda_{21}, \quad \lambda_{32} = \left(\rho_b^{23}\tau_2\tau_3 - \rho_b^{31}\tau_3\tau_1\lambda_{21}\right) / \left(\tau_2^2 - \lambda_{21}^2\tau_1^2\right) \end{aligned} \tag{11}$$

which are obtained by following the procedure described in detail in Section 5.1 for N-dimensional case. Having established them allows to place prior distributions on the between-study standard deviations and correlations, which is an easier task 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 [22] or Bujkiewicz *et al*. [16]). By placing prior distributions on these parameters, for example, $\rho_b^{12(13,23)} \sim dunif(-1,1)$, $\tau_{1(2,3)} \sim N(0,10)I(0,)$ (normal distribution truncated at value zero), the above derived relationships give the implied prior distribution on the parameters λ_{21}, λ_{31} and λ_{32} and hyper-parameters ψ_1, ψ_2 and ψ_3. The remaining parameters are given 'vague' prior distributions, $\eta_1 \sim N(0, 1000)$, $\lambda_{20(30)} \sim N(0, 1000)$.

Note that the pooled effects $\beta_{1(2,3)}$ on the three outcomes in model (2) are also directly linked to the model (10); $\beta_1 = \eta_1$, $\beta_2 = \lambda_{20} + \lambda_{21}\beta_1$ and $\beta_3 = \lambda_{30} + \lambda_{31}\beta_1 + \lambda_{32}\beta_2$. It is possible to center the true effects on surrogate endpoints (by replacing μ_{ji} with $(\mu_{ji} - \overline{\mu_{ji}})$ in the third and fifth line of formula 10, which can be useful if there are problems with autocorrelation) in which case the intercepts would equal the pooled effects; $\beta_2 = \lambda_{20}$ and $\beta_3 = \lambda_{30}$.

2.5. *Product normal formulation with structured covariance matrix*

A simplified model can be used by assuming conditional independence between treatment effects on some of the outcomes. For example, 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 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, described in Figure 1b, leads to a

simplified between-study model with the true effect on the final outcome now conditional on the effect on the second surrogate endpoint only:

$$\begin{cases} \mu_{1i} \sim N\left(\eta_1, \psi_1^2\right) \\ \mu_{2i} \mid \mu_{1i} \sim N\left(\eta_{2i}, \psi_2^2\right) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\ \mu_{1i} \\ \mu_{3i} \mid \mu_{2i} \sim N\left(\eta_{3i}, \psi_3^2\right) \\ \eta_{3i} = \lambda_{30} + \lambda_{32}\ \mu_{2i}, \end{cases} \tag{12}$$

The assumption of conditional independence of the true treatment effects on outcomes three and one, μ_3 and μ_1, conditional on the effect on outcome two, μ_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 $\rho_b^{13|2} = 0$. Because the partial correlation between μ_1 and μ_3 (adjusted for μ_2) equals $\rho_b^{13|2} = \left(\rho_b^{13} - \rho_b^{12} * \rho_b^{23}\right) / \sqrt{1-(\rho_b^{12})^2}\sqrt{1-(\rho_b^{23})^2} = 0$, it implies that $\rho_b^{13} = \rho_b^{12} * \rho_b^{23}$. This reduces the number of parameters in the model that need to be estimated and also simplifies the relationships between the parameters and hyperparameters of the model with the elements of the between-study covariance matrix:

$$\begin{aligned} &\psi_1^2 = \tau_1^2, \quad \psi_2^2 = \tau_2^2 - \lambda_{21}^2\tau_1^2, \quad \psi_3^2 = \tau_3^2 - \lambda_{32}^2\tau_2^2, \\ &\lambda_{21} = \rho_b^{12}\frac{\tau_2}{\tau_1}, \quad \lambda_{32} = \rho_b^{23}\frac{\tau_3}{\tau_2}, \end{aligned} \tag{13}$$

which both can have an advantage in scenarios with multiple outcomes beyond trivariate case as discussed in Section 5. Similarly as in the case of the full unstructured covariance matrix, placing prior distributions directly on the between-study standard deviations $\tau_{1(2,3)} \sim N(0,10)I(0,)$ and correlations $\rho_b^{12(23)} \sim dunif(-1,1)$ gives implied prior distributions placed on the parameters of the model (12), $\psi_{1(2,3)}$ 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 $\beta_{1(2,3)}$ on the three outcomes in model (2) are also directly linked to the model (12); $\beta_1 = \eta_1$, $\beta_2 = \lambda_{20} + \lambda_{21}\beta_1$ and $\beta_3 = \lambda_{30} + \lambda_{32}\beta_2$. As in the model in Section 2.4, centering of the true effects on surrogate endpoints can be applied to this models in which case the intercepts would equal the pooled effects.

Assuming such sequential structure of the treatment effects, that could be measured on the same outcome at multiple time points, the model can lead to removing some of the measurement error, for example at time two using time one. We cannot measure the true effect μ_{2i} at time two (only have an estimate Y_{2i}), so having measurement at time one can improve estimate of measurement at time two. This can be useful in particular when the treatment effect at time one is measured precisely and at time two inaccurately, then the prediction at time two may be more precise (due to accounting for the correlation with outcome one) leading to better prediction at time three.

2.6. Criteria for surrogate markers

Consider first a bivariate case (with one surrogate endpoint, where the first endpoint is surrogate to the second and the third outcome is removed) as described in the first three lines of Equation (10) or (12). We can then follow the criteria set out by Daniels and Hughes [2], by which λ_{21} indicates the association between the treatment effect measured by the surrogate endpoint and the treatment effect measured by the second clinical outcome (final outcome in the bivariate case with a single surrogate endpoint), therefore, we require $\lambda_{21} \neq 0$. For the association to be perfect, the conditional variance should be zero; $\psi_2^2 = 0$. Also, we would expect $\lambda_{20} = 0$ (no treatment effect on the surrogate endpoint gives no treatment effect on the target outcome), otherwise not all of the treatment effect on the target outcome is mediated by the effect on the surrogate endpoint. In the trivariate case with two surrogate markers, the effects of treatment on all biomarkers may jointly mediate the treatment effect on the final outcome. For the combined effect on the biomarkers to fully mediate the effect on the target outcome, we expect the intercept $\lambda_{30} = 0$ and the conditional variance $\psi_3^2 = 0$. The association between the effect on target outcome and each of the surrogate endpoints is expected not to be zero; $\lambda_{31,32} \neq 0$. In the sequential scenario of conditionally independent effects (described in Section 2.5), the same criteria as in the bivariate case apply, where the

effect on the first outcome is a surrogate to the effect on the second and the effect on the second outcome is a surrogate to the effect on the third (final) outcome, but with additional 'borrowing of strength' across outcomes by taking into account the correlation structure between all of the outcomes (i.e. only conditional independence assumed).

2.7. Implementation in WinBUGS and R

All models were implemented in WinBUGS [23] where the estimates were obtained using MCMC simulation using 50 000 iterations (including 20 000 burn-in). Convergence was checked by visually assessing the history, chains and autocorrelation using graphical tools in WinBUGS. All posterior estimates are presented as means with the 95% credible intervals (CrI). R was used for data manipulation and to execute WinBUGS code multiple times (for validation of surrogates for each study) using the R2WinBUGS package [24]. OpenBUGS and R2OpenBUGS version of the software was used for the simulation study which was conducted using Linux (Red Hat, Inc., Raleigh, North Carolina)-based high performance computer.

WinBUGS programs corresponding to the two TRMA models (applied to data in Tables I and II) are included in Web Supplements A1 and A2.

3. Application: relapsing remitting multiple sclerosis

To illustrate the application of the methods described in Section 2, the models developed there are applied to the motivating example in RRMS described in Section 3.1. Section 3.2 (along with the Appendix A) contains details of how data in Table I are used to populate the models for outcomes specific to this motivating example. Results are presented in Section 3.3.

3.1. Introduction to the motivating example

To illustrate the use of the modelling techniques, we applied them to an example in RRMS. Multiple sclerosis (MS) is an inflammatory disease of the brain and spinal cord. RRMS is a most common type of MS. During the course of the disease, patients experience a series of periods of exacerbations (relapses) and remission. A large proportion of patients (25%) eventually progresses to secondary progressive disease [25]. The disability progression is considered the final outcome, whereas the number of active (new or enlarging) T2 lesions in the brain obtained from the magnetic resonance imaging (MRI) and the annualised relapse rate are two potential surrogate endpoints. This example is based on work by Sormani *et al*. [17] who used meta-analytic approach to evaluate whether the effects on relapse rate and number of active MRI lesions are good predictors of the treatment effects on disability progression (one surrogate endpoint at a time). In another paper, Sormani *et al*. [26] investigated estimates of the treatment effect on number of active MRI lesions as predictors of the effects on relapse rate. Subsequently, Sormani and colleagues used IPD to investigate individual-level association between outcomes, where they validated the number of MRI lesions as a surrogate to the number of relapses [27]. In another paper, Sormani *et al*. also used IPD to validate both the number of active MRI lesions and the number of relapses as surrogate endpoints to the disability progression, as individual surrogates as well as joint mediators of the treatment effect on progression [11]. In the papers mentioned previously, the meta-analytic work on study level was conducted using weighted linear regression [17, 26], whereas the association on the individual level was conducted using Prentice's criteria [11, 27].

This example with two surrogate endpoints, where one of the candidate surrogate endpoints (to the final outcome) is also a potential surrogate endpoint to the second surrogate, serves as a desirable illustration of modelling techniques investigated in this paper. We collect data on the treatment effects on those outcomes from studies included in the analysis by Sormani *et al*. [17]. To investigate the surrogate endpoints jointly, we chose to include only the studies reporting treatment effect estimates on all of the three outcomes and those that reported sufficient information to obtain uncertainty estimates around the observed treatment effect. As a result, we obtained data from 13 studies, 11 placebo-controlled trials and two active-controlled trials, which are listed in Table I and displayed graphically in Figure 2. To obtain the within-study correlations between the treatment effect estimates for the outcomes, we use the estimates of Prentice's criteria, reported by Sormani *et al*. [27] for the association between the number of MRI lesions and the number of relapses who adopted the model (6) to those outcomes on log scale and for the association of both the number of MRI lesions and the relapses with the disability progression

Table I. Data on disability progression, relapse rate and number of active MRI lesions included in the meta-analysis.

Study	Follow-up (months)	Disability progression				Relapse rate				MRI			
		Nd_E	Rd_E	Nd_C	Rd_C	Nr_E	ARr_E	Nr_C	ARr_C	Nm_E	Rm_E (SE)	Nm_C	Rm_C (SE)
Paty (A)	24	124	35	124	35	124	145	124	157	124	1.80 (0.40)	124	4.9 (1.30)
Paty (B)	24	124	25	124	35	124	104	124	157	124	2.00 (0.70)	124	4.9 (1.30)
Jacobs	24	85	18	87	29	85	52	87	78	85	3.20 (0.41)	87	4.8 (0.49)
Millefiorini	24	27	2	24	9	27	12	24	31	23	3.50 (0.71)	19	7.3 (1.84)
Li (C)	24	189	57	187	69	189	172	187	240	189	9.00 (4.00)	187	15.5 (2.90)
Li (D)	24	184	50	187	69	184	159	187	240	184	5.50 (0.50)	187	15.5 (2.90)
Polman	24	627	107	315	91	627	144	315	230	627	1.90 (0.37)	315	11.0 (0.88)
Comi (E)	24	433	62	437	90	433	61	437	144	433	0.38 (0.07)	437	1.43 (0.06)
Comi (F)	24	456	69	437	90	456	68	437	144	456	0.33 (0.06)	437	1.43 (0.06)
Rudick	24	589	135	582	169	589	200	582	437	589	0.90 (0.09)	582	5.4 (0.36)
Sorensen	24	66	11	64	16	66	15	64	38	66	2.70 (0.46)	64	3.5 (0.51)
Clanet	36	400	148	402	149	400	324	402	310	400	8.00 (0.88)	402	9.0 (0.74)
Mikol	24	386	45	378	33	386	116	378	110	172	0.58 (0.11)	178	0.77 (0.18)

Nd_E (Nd_C), total number of patients in experimental (control) arm with disability status recorded.
Rd_E (Rd_C), number of patients in experimental (control) arm who progressed.
Nr_E (Nr_C), total number of patients in experimental (control) arm with number of relapses recorded.
ARr_E (ARr_C), number of relapses per year in experimental (control) arm.
Nm_E (Nm_C), total number of patients in experimental (control) arm with number of active MRI lesions in experimental (control) arm.
Rm_E (Rm_C), mean number (standard error) of active MRI lesions in experimental (control) arm.

Table II. Indicators of individual-level surrogacy for number of active MRI lesions and relapse rate as surrogate endpoints to disablity progression (reproduced from Sormani *et al* [11]) and number of active MRI lesions as a surrogate endpoint to relapse rate (reproduced from Sormani *et al* [27]).

Surrogate endpoint	Final outcome	Prentice's criteria		
		1st criterion*	2nd criterion†	4th criterion‡
active T2 lesions	disability progression	$\alpha_1 = -0.93(0.12)$	$\beta_1 = -0.37(0.19)$	$\beta_{S1} = -0.14(0.19)$
relapses	disability progression	$\alpha_2 = -0.44(0.08)$	$\beta_2 = -0.37(0.19)$	$\beta_{S2} = -0.15(0.20)$
active T2 lesions	relapses	$\alpha_3 = -0.90(0.13)$	$\beta_3 = -0.36(0.08)$	$\beta_{S3} = -0.17(0.09)$

Coefficients are reported with standard errors.
*1st Prentice's criterion, treatment is effective on surrogate endpoint.
†2nd Prentice's criterion, treatment is effective on final clinical outcome.
‡4th Prentice's criterion, treatment effect on final clinical outcome fully mediated by surrogate.
α_1, treatment effect on log number of MRI lesions over 1 year; α_2, treatment effect on log relapse rate over 1 year; α_3, treatment effect on log number of MRI lesions over 2 years; β_1, β_2, treatment effect on log odds disability progression over 2 years; β_3, treatment effect on log relapse rate over 2 years.
β_{S1} and β_{S2}, treatment effect on log odds disability progression over 2 years adjusted for treatment effect on log number of MRI lesions and log relapse rate, respectively; β_{S3}, treatment effect on log relapse rate over 2 years adjusted for treatment effect on log number of MRI lesions.

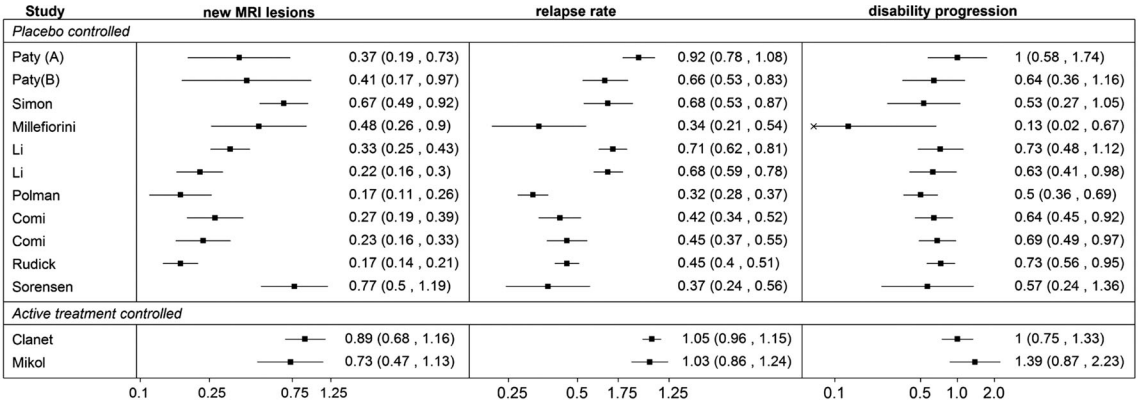

Figure 2. Graphical representation of data for treatment effects on MRI (surrogate endpoint 1), relapse rate (surrogate endpoint 2) and disability progression (final clinical outcome).

reported by Sormani *et al.* [11], also on log scale. Relevant estimates are listed in Table II. Because of limitation of the data, in this paper, the correlation is assumed constant across studies.

3.2. Scale of the outcomes and the within-study model

The relative treatment effects on each outcome (MRI, relapse rate and disability progression) and the within-study variances for the treatment effects in each study are calculated using data listed (and notation described) in Table I.

The MRI effect is modelled on the log rate ratio (RR) scale, $Y_{1i} = \log MRIRR = \log(Rm_E/Rm_C)$. The relative treatment effect on relapses is modelled by the log annualised relapse rate ratio (ARRR) scale, $Y_{2i} = \log ARRR = \log\left(\frac{ARR_E}{ARR_C}\right)$. The relative treatment effect on disability progression is modelled on the log odds ratio (OR) scale, $Y_{3i} = \log OR = \log\left(\frac{Rd_E(Nd_C - Rd_C)}{Rd_C(Nd_E - Rd_E)}\right)$.

The corresponding variances are obtained from the data by the use of delta method for the effects on MRI and relapses and using the standard formulae for the standard error of log *OR*, as shown in Appendix A.1. The within-study correlations are obtained following methods described in Section 2.2, with details of the algebra for the data in RRMS in Appendix A.2. The correlations were obtained for one study and assumed the same in all the studies.

Note that some authors used relative risk to quantify the treatment effect on the disability progression. However, log OR scale is applied here in order to combine the summary data on effectiveness with

reported Prentice's criteria, which were modelled for the progression on the log odds scale. Typically, progression is reported cross-sectionally as a number (or proportion) of patients who progressed (i.e. whose expanded disability status scale score had increased by at least one point at the follow up). As such, the reported outcome is not suitable to model as time to event outcome on the hazard ratio scale.

3.3. Results

Applying the aformentioned formulae to the RRMS data in Table I and combining it with data of individual-level surrogacy criteria in Table II gives the within-study correlations: $\rho_{wi}^{12} = 0.25$, $\rho_{wi}^{13} = 0.09$, $\rho_{wi}^{23} = 0.09$. The data are applied using the above models to validate surrogate endpoints as predictors of the clinical benefit on the final outcome using a cross-validation procedure as described in Section 2.1. First, the models (1) and (10), assuming that all true effects are correlated, is applied. Results of this application are presented in Section 3.3.1. Then the models (1) and (12), assuming the conditional independence between the effects on the final outcome and the first surrogate endpoint is applied and results presented in Section 3.3.2.

3.3.1. Results from the model with unstructured between-study covariance matrix. To quantify the surrogacy criteria overall, the model was first applied to the full data (no outcome missing for any of the studies). The results (parameters with the 95% CrIs) are shown in Table III (top). The coefficients λ_{20} and λ_{21} and variance ψ_2^2 represent the association between the treatment effect on the number of active MRI lesions and the effect on the relapse rate, whereas the coefficients λ_{30}, λ_{31}, λ_{32} and ψ_3^2 describe how the effects on number of MRI lesions and relapse jointly mediate the effect on the disability progression (as described in Section 2.6). The parameters on the left-hand-side indicate that there was a poor association between the effects on MRI and effects on relapses, as the interval of the slope contains zero and the variance ψ_3^2 appears greater than zero. The parameters on the right-hand-side suggest that the association between the effects on both surrogate endpoints (MRI and relapse rates) are not strongly associated with the effect on disability progression (slopes contain zero), however, the variance appears small suggesting that the effect on MRI and relapse rate largely mediate the effect of the treatment on disability progression. In both cases, the CrIs for the intercepts contain zero, which is encouraging in the sense that zero treatment effect on MRI would be expected to predict zero effect on relapse rate in the first case and the zero effects measured by the MRI and relapse rate would lead to predicted zero effect on the disability progression in the second case. However, in both cases the CrIs are wide, hence, the predictions may not be accurate.

After estimating the surrogacy criteria using the full data, the cross-validation process was carried out. The results of the validation are presented in Table IV (column four, along with the results from the bivariate model and trivariate with structured covariance matrix), which shows the values of observed effects on the final outcome in each study (column two) and the effects on the final outcome predicted by the model from the effects on the surrogates (given full data on all other studies) with the 95% CrIs. In each row, the values correspond to the validation of prediction for a study whose corresponding observed effect on the final outcome is in the second column. In the take-one-out approach of the cross-validation procedure, the regression parameters (the intercepts, slopes and conditional variances, as in the between-study model (10)) did not change substantially. The full set of those values is included in the Web Supplement B. The predicted intervals contain the observed value of the treatment effect on the final outcome for all of the studies confirming good fit of the model.

Table III. Surrogacy criteria obtained from the trivariate models applied to the full data (no missing outcomes).

Parameter	Mean (95% CrI)	Parameter	Mean (95% CrI)
Unstructured between-study covariance			
λ_{20}	−0.28 (−0.73, 0.16)	λ_{30}	−0.09 (−0.38, 0.21)
λ_{21}	0.26 (−0.12, 0.65)	λ_{31}	0.00 (−0.38, 0.31)
		λ_{32}	0.43 (−0.02, 1.06)
ψ_2^2	0.15 (0.05, 0.37)	ψ_3^2	0.02 (0.00, 0.10)
Structured between-study covariance			
λ_{20}	−0.24 (−0.70, 0.20)	λ_{30}	−0.06 (−0.31, 0.19)
λ_{21}	0.30 (−0.09, 0.70)	λ_{32}	0.48 (0.11, 0.88)
ψ_2^2	0.15 (0.05, 0.36)	ψ_3^2	0.02 (0.00, 0.10)

Table IV. Comparison of results of validation obtained from the two trivariate models and a bivariate model.

		Predicted				
Study	Observed	Bivariate	Trivariate Unstructured	% red.	Trivariate Structured	% red.
Paty (A)	1.00 (0.57, 1.74)	0.86 (0.42 ,1.74)	0.86 (0.43, 1.73)	0.67	0.88 (0.44, 1.73)	3.16
Paty (B)	0.64 (0.36, 1.16)	0.76 (0.37, 1.56)	0.77 (0.38, 1.54)	2.24	0.77 (0.38, 1.54)	3.29
Simon	0.53 (0.27, 1.06)	0.78 (0.36, 1.73)	0.80 (0.36, 1.77)	0.19	0.79 (0.37, 1.72)	2.10
Millefiorini	0.13 (0.02, 0.67)	0.62 (0.11, 3.46)	0.65 (0.12, 3.62)	−0.29	0.62 (0.11, 3.44)	0.32
Li (C)	0.73 (0.47, 1.12)	0.79 (0.43, 1.45)	0.79 (0.45, 1.41)	4.46	0.80 (0.45, 1.42)	5.19
Li (D)	0.63 (0.41, 0.98)	0.79 (0.43, 1.44)	0.81 (0.44, 1.50)	−1.07	0.79 (0.45, 1.39)	7.23
Polman	0.50 (0.36, 0.69)	0.59 (0.32, 1.11)	0.61 (0.35, 1.07)	9.67	0.58 (0.33, 1.03)	8.62
Comi (E)	0.64 (0.45, 0.92)	0.62 (0.35, 1.11)	0.62 (0.36, 1.08)	3.71	0.62 (0.36, 1.07)	5.13
Comi (F)	0.69 (0.49, 0.97)	0.64 (0.36, 1.13)	0.64 (0.38, 1.09)	5.77	0.63 (0.37, 1.08)	5.93
Rudick	0.73 (0.56, 0.95)	0.62 (0.37, 1.05)	0.61 (0.37, 1.01)	4.18	0.61 (0.38, 0.98)	10.01
Sorensen	0.57 (0.24, 1.36)	0.62 (0.24, 1.64)	0.64 (0.23, 1.82)	−7.52	0.63 (0.24, 1.66)	0.63
Clanet	1.00 (0.75, 1.33)	0.89 (0.49, 1.63)	0.88 (0.49, 1.58)	3.21	0.92 (0.53, 1.60)	8.30
Mikol	1.39 (0.87, 2.23)	0.82 (0.44, 1.55)	0.83 (0.45, 1.53)	2.65	0.85 (0.46, 1.57)	3.45
average % reduction in CrI				2.14%		4.87%
DIC			350.1		347.9	

The % red refers to the percentage reduction in the width of the credible interval corresponding to the prediction from the trivariate model, with the unstructured (columns 4 and 5) or structured (columns 6 and 7) between-study covariance matrix, compared with the width of the interval corresponding to the prediction from the bivariate model (column 3). DIC, deviance information criteria.

3.3.2. Results from the model with structured between-study covariance matrix. Models (1) and (12) with structured between-study covariance matrix, where the effect on the disability progression (the final outcome) is conditionally independent from the effect on the MRI (the first surrogate endpoint) conditional on the effect on relapse rate (the second surrogate endpoint) is now used for validation. Similarly as in the case of the model with unstructured covariance matrix, the surrogacy criteria are estimated using the full data, which is then followed by the cross-validation. In Table III (bottom) the parameters are shown together with the 95% CrIs for the model applied to the full data. The parameters λ_{20}, λ_{21} and ψ_2^2 describe the association between the treatment effects on the MRI and the relapse rate, while λ_{30}, λ_{31} and ψ_3^2 the association between the treatment effects on the relapse rate and the disability progression (conditional on the effect on MRI). The association between the effects on the MRI and the relapse rate is not strong which is indicated by the interval of λ_{21} containing zero and the variance ψ_2^2 significantly larger than zero. However, the association between the effect on the relapse rate, as a surrogate endpoint, with the effect on the disability progression, as the final outcome, (conditional on the effect on MRI) appears strong as indicated by the non-zero slope λ_{31} and the small variance ψ_3^2. Also the interval of the intercept λ_{30} containing zero indicates that zero effect on the relapse rate is likely to imply zero effect on the disability progression. In the take-one-out approach of the cross-valdation procedure, the regression parameters (the intercepts, slopes and conditional variances, as in the between-study model (12)) did not change substantially. The full set of those values is included in the Web Supplement B. Table IV shows the results of cross-validation (with column six corresponding to predictions from the model with structured between-study covariance). The results of predictions are similar to those obtained from the model with unstructured between-study covariance matrix.

3.3.3. Comparison of the results from the two models and those from a bivariate model. When carrying out the cross-validation process, we want to ensure that not only predicted CrIs contain the actual observed values but also that the intervals are narrow. Inclusion of multiple surrogate endpoints can potentially lead to reduced intervals and hence better predictions. To compare the aforementioned results of the validation from the two trivariate models, they are shown side by side in Table IV as well as graphically in Figure 3 alongside those from a bivariate model with the effect on the relapse rate as a surrogate for the effect on the disability progression. In Table IV, apart from the predicted values, the percentage reduction in the width of the credible interval relative to the width of the interval obtained from the bivariate model is shown for both trivariate models. On average, the model with the unstructured between-study

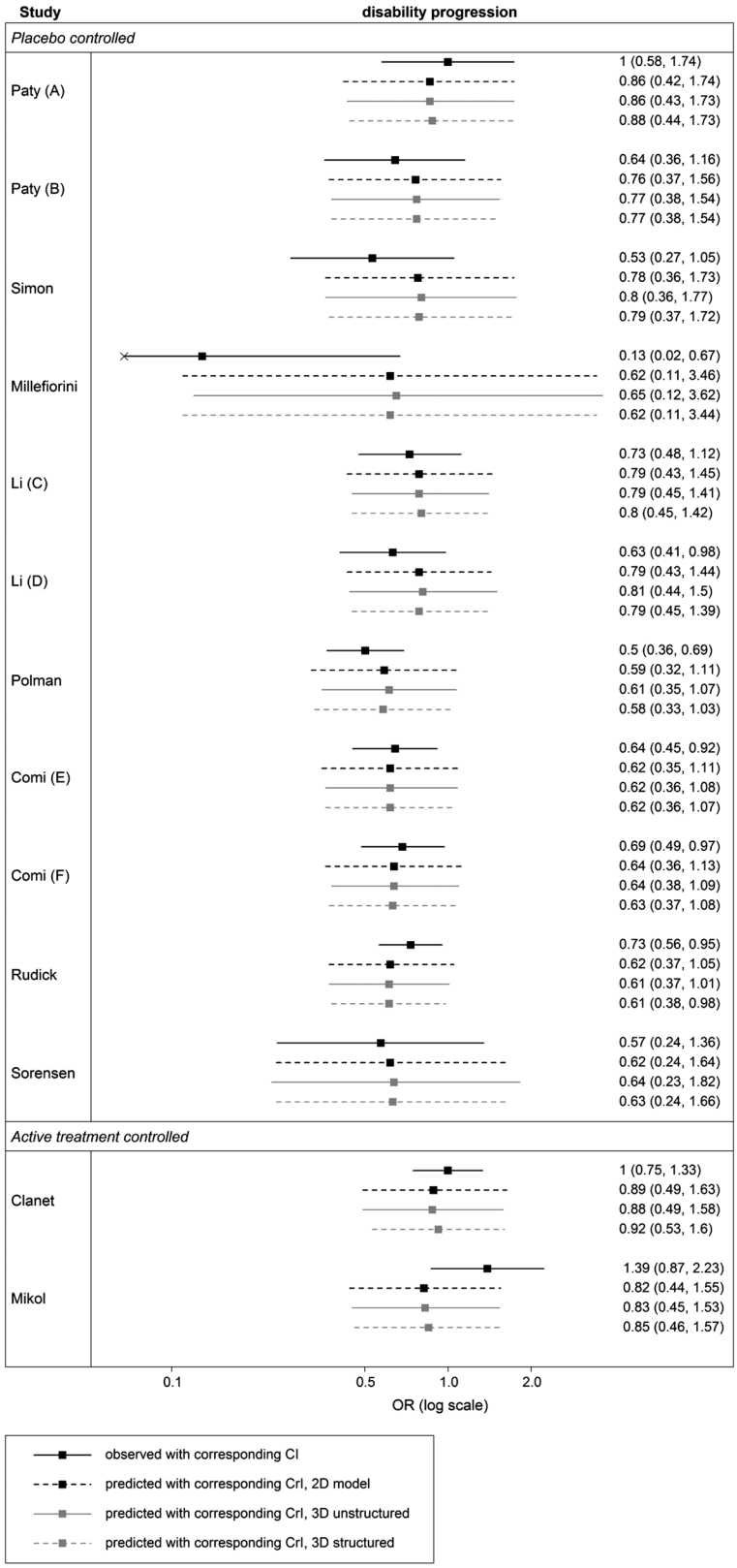

Figure 3. Forest plot, showing for each study the observed value of the OR of disability progression with corresponding confidence interval (CI) and the predicted values with corresponding credible intervals (CrIs) from a bivariate model, trivariate model with unstructured covariance matrix and from the trivariate model with structured covariance matrix.

covariance matrix gave intervals 2% narrower than those from the model using a single surrogate endpoint, whereas the model with the structured covariance matrix led to 5% reduction in uncertainty. Although the average gain in precision is modest, the inclusion of both surrogates does improve the predictions and this reduction in uncertainty may be sufficient to improve the decision making process based on such predictions. This may be the case in particular for larger studies, such as by Polman, Rudick and Clanet in our illustrative example, where the gain in precision was also larger, up to 10%, likely due to the treatment effect on the surrogate endpoints being measured with larger precision compared with the smaller studies. Inclusion of multiple surrogate endpoints may lead to a more substantial gain in precision in other disease areas and when data on some outcomes in some of the studies are missing at random [28]. Deviance information criteria (DIC) obtained for the two trivariate models using the complete data, showed at the bottom of Table IV, suggests that the fit by both models is comparable. The model assuming conditional independence is simpler and easier to implement as shown in Section 5. It also requires fewer parameters to estimate as discussed further in Section 6. If such assumption can be justified, the model with the structured covariance may be more practical.

4. Simulation and sensitivity analysis

4.1. Methods

A simulation study was carried out to compare the performance of the three models: a bivariate model with a single surrogate endpoint and the two trivariate models with two surrogate endpoints, the model with unstructured between-study covariance matrix and with the structured covariance matrix. Data were simulated under two scenarios considering alternative covariance structures: unstructured, by simulating data from the model (1)+(10) and structured by simulating data from the model (1)+(12). The parameters of the between-study model were set to be comparable with those corresponding to the RRMS data, namely, $\tau_{1,2,3} = 0.5$ and $\rho_b^{12,13,23} = 0.8$ and $\eta_1 = -0.3$.

The regression coefficients and the conditional variances were obtained from the between-study standard deviations and correlations following the formulae (11) and (13) for the models with unstructured and structured covariance matrices, respectively. The within-study correlations ρ_{wi}^{jk} were set to the same values as in the example in RRMS (obtained from the Prentice's criteria). The within-study variances were simulated by drawing the precisions (inverse variances) from the gamma distributions; $\sigma_{1(2,3)i} = 1/P_{1(2,3)i}$, $P_{1(2,3)i} \sim \Gamma(\alpha_{1(2,3)}, \theta_{1(2,3)})$, where $\alpha_{1(2,3)}$ are the shape parameters and $\theta_{1(2,3)}$ the scale parameters, which were obtained using the method of moments: $E(P_{1,2,3}) = \alpha_{1,2,3}/\xi_{1,2,3}$, $V(P_{1,2,3}) = \alpha_{1,2,3}/\xi_{1,2,3}^2$, where $\xi_{1,2,3} = 1/\theta_{1,2,3}$ is a rate parameter. By summarising the inverse variances from the RRMS data, the following parameters were obtained: $E(P_1) = 30$, $E(P_2) = 150$, $E(P_3) = 25$, $V(P_1) = 420$, $V(P_2) = 15\,000$ and $V(P_2) = 275$, giving the following shape and rate parameters: $\alpha_1 = 2.14$, $\alpha_2 = 1.5$, $\alpha_3 = 2.3$, $\xi_1 = 0.07$, $\xi_2 = 0.01$ and $\xi_3 = 0.09$. Because of the structure of the gamma distribution, some of the simulated precisions were very close to zero, resulting in very large variances. This led to some problems with the estimation. To overcome this issue, a constraint was placed on the simulated value of the precision by discarding the precisions resulting in variances larger than 2 (this number was taken as an arbitrary cut off, large enough to be much larger than the variances in the RRMS data and hence including all plausible variances in the population but small enough not to produce problems with the estimation). R code for the data simulation is included in the Web Supplement C.

The cross-validation procedure is applied to each simulated data set. This time, however, the true effect μ_{3n} in a validation study n is known as it has been simulated, so the cross-validation can be performed on the true effects (which in real circumstances we would like to predict) by comparing the simulated μ_{3n} with predicted interval of $\hat{\mu}_{3n}$ with the corresponding variance $var(\hat{\mu}_{3n}|Y_{1n}, Y_{2n}, \sigma_{1n}, \sigma_{2n}, Y_{1(-n)}, Y_{2(-n)}, Y_{3(-n)})$. The two trivariate models and also the bivariate model are applied to the simulated data sets to compare their performance, by estimating bias of the mean $\hat{\mu}_{3n}$, root-mean-square error (RMSE), coverage of 95% credible intervals (CrI) and the potential reduction of the width of the predicted interval by calculating the ratios of the width of the interval from a trivariate model w_{3d} (with two surrogate endpoints) with the width of the predicted interval from the bivariate model w_{2d}.

4.2. Sensitivity analysis

To investigate the sensitivity of the methods to the assumption of normality of the data, another simulation study was carried out where the data were simulated from multivariate *t*-distribution as well as mixture normal distribution.

4.2.1. Simulation study with t-distribution. To simulate data from the t-distribution with unstructured covariance matrix, the true treatment effects on the three outcomes were generated from the following model:

$$\begin{cases} \mu_{1i} \sim t(\eta_1, \nu_1, df) \\ \mu_{2i} \mid \mu_{1i} \sim t(\eta_{2i}, \nu_2, df) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\mu_{1i} \\ \mu_{3i} \mid \mu_{1i}, \mu_{2i} \sim t(\eta_{3i}, \nu_3, df) \\ \eta_{3i} = \lambda_{30} + \lambda_{31}\mu_{1i} + \lambda_{32}\mu_{2i}, \end{cases} \tag{14}$$

and from the t-distribution with structured covariance matrix, from the following model

$$\begin{cases} \mu_{1i} \sim t(\eta_1, \nu_1, df) \\ \mu_{2i} \mid \mu_{1i} \sim t(\eta_{2i}, \nu_2, df) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\mu_{1i} \\ \mu_{3i} \mid \mu_{1i}, \mu_{2i} \sim t(\eta_{3i}, \nu_3, df) \\ \eta_{3i} = \lambda_{30} + \lambda_{32}\mu_{2i}, \end{cases} \tag{15}$$

where $\nu_{1,2,3} = \left(\psi^2_{1,2,3}(df-2)\right)/df$ and $df = 4$. The individual study estimates were simulated from the trivariate t-distribution

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \\ Y_{3i} \end{pmatrix} \sim \mathrm{MV}t\left(\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \\ \mu_{3i} \end{pmatrix}, \boldsymbol{\Sigma}_\mathbf{i}, \mathrm{df}\right), \ \boldsymbol{\Sigma}_\mathbf{i} = \begin{pmatrix} \sigma^2_{1i} & \sigma_{1i}\sigma_{2i}\rho^{12}_{wi} & \sigma_{1i}\sigma_{3i}\rho^{13}_{wi} \\ \sigma_{2i}\sigma_{1i}\rho^{12}_{wi} & \sigma^2_{2i} & \sigma_{2i}\sigma_{3i}\rho^{23}_{wi} \\ \sigma_{3i}\sigma_{1i}\rho^{13}_{wi} & \sigma_{3i}\sigma_{2i}\rho^{23}_{wi} & \sigma^2_{3i} \end{pmatrix}. \tag{16}$$

The simulation was conducted in R software using the `rt` command (and scaling the simulated data by $\psi_{1,2,3}\sqrt{((df-2)/df)} + \eta_1$) for univariate t-distributions in the between-study models (14) and (15) and the `rmvt` in the within-study model (16).

4.2.2. Simulation study with mixture normal distribution. Data with more severe departure from normality were generated by the use of mixture normal distributions. This was achieved by replacing the univariate normal distribution for the true treatment effect on the first outcome in the models (10) and (12) with the mixture normal distribution

$$\mu_{1i} \sim p_1 * N\left(\eta_1, \psi^2_1\right) + p_2 * N\left(\eta_1 - 4 * \psi_1, \psi^2_1\right) + p_3 * N\left(\eta_1 + 4 * \psi_1, \psi^2_1\right), \tag{17}$$

with $p_1 = 0.5$, $p_2 = 0.3$ and $p_3 = 0.2$. This deviation from normality feeds through to the true effects μ_{2i} and μ_{3i} by the linear association of those effects with μ_{1i}. These now non-normal true effects $\mu_{1(2,3)i}$ are then used as mean values when generating the within study data from the multivariate normal distribution giving data with 'distorted' normality.

4.3. Results

Data sets including the treatments effects (and corresponding sampling variances) on three outcomes in 15 studies were generated in 1000 simulations for each scenario. 0.1% of simulation runs were discarded because of precisions resulting in too high variances, as explained in the methods Section 4.1. Simulation results are presented in Table V. The bias of mean predicted effect $\hat{\mu}_{3n}$ for a validation study n was comparable across all models and data scenarios. The RMSE was larger when data were simulated from a model with unstructured covariance matrix, regardless of the distribution or model fitted to the data. All models (the bivariate and the two trivariate) seemed to perform equally well, giving coverage of 95% credible interval close to 95% for most scenarios, except for the data generated from the mixture normal distribution where the coverage was slightly inflated to 98% (because of the three normal distributions being approximated by models with a single normal distribution, leading to the inflated variance of predictions). For data generated from either multivariate normal or t-distribution with unstructured covariance matrix, both trivariate models gave on average 4% reduction of the width of the predicted interval

Table V. Results of simulation studies.

Meta-analysis model	Bias of mean $\hat{\mu}_{3n}$	RMSE of $\hat{\mu}_{3n}$	Coverage of 95% CrI for $\hat{\mu}_{3n}$	Median w_{3d}/w_{2d}
Scenario 1: Data simulated from normal TRMA with UCM				
BRMA	−0.002	0.46	0.96	
TRMA UCM	0.003	0.45	0.96	0.96
TRMA SCM	−0.001	0.46	0.95	0.96
Scenario 2: Data simulated from normal TRMA with SCM				
BRMA	0.006	0.36	0.95	
TRMA UCM	0.003	0.36	0.95	0.98
TRMA SCM	0.006	0.35	0.94	0.95
Scenario 3: Data simulated from TRMA with UCM and *t*-distribution				
BRMA	−0.002	0.48	0.95	
TRMA UCM	−0.004	0.47	0.95	0.96
TRMA SCM	−0.002	0.48	0.94	0.96
Scenario 4: Data simulated from TRMA with SCM and *t*-distribution				
BRMA	−0.0001	0.36	0.95	
TRMA UCM	0.002	0.37	0.95	0.98
TRMA SCM	0.0002	0.36	0.94	0.95
Scenario 5: Data simulated from TRMA with UCM and mixture normal				
BRMA	−0.007	0.49	0.98	
TRMA UCM	0.003	0.47	0.98	1.00
TRMA SCM	−0.001	0.47	0.97	0.91
Scenario 6: Data simulated from TRMA with SCM and mixture normal				
BRMA	−0.002	0.36	0.98	
TRMA UCM	0.0001	0.37	0.98	1.01
TRMA SCM	−0.0001	0.36	0.97	0.91

RMSE, root-mean-squared-error; CrI, credible interval
UCM, unstructured covariance matrix; SCM, structured covariance matrix; TRMA, trivariate random effects meta-analysis;
w_{3d} (w_{2d}), width of the predicted interval from TRMA (BRMA)
MC errors of the predicted mean effects were less than 0.012 in scenarios 1–4 and less than 0.015 and 0.025 in scenarios 5 and 6, respectively

compared with the intervals obtained from the bivariate model. When data was simulated using normal or *t*-distribution with structured covariance matrix, the trivariate model with unstructured covariance gave on average 2% reduction in uncertainty around the predicted effect compared with the uncertainty around predictions obtained from the bivariate model, while the model with structured covariance matrix gave predictions with intervals 5% narrower compared with those obtained from the model with a single surrogate endpoint. When data were simulated from the mixture of normal distributions, the trivariate model with unstructured covariance did not produce any gain in precision of predictions, while the model with structured covariance matrix gave predicted intervals that were on average 9% narrower compared with those obtained from the bivariate model.

4.4. Discussion of the results

The outcomes of the simulation study were broadly in agreement with those obtained from the case study. The predicted intervals obtained from TRMA models were narrower compared with those obtained from BRMA, but this reduction was less pronounced when using the TRMA UCM model on the data simulated from a model with structured covariance matrix (one of the surrogates correlated to the other surrogate but less so to the final outcome; scenarios 2, 4 and 6 in Table V and also in the RRMS case study). Using TRMA UCM on data simulated from the same model gives the same reduction in uncertainty as when using TRMA SCM but this effect of the addition of the second surrogate on the uncertainty around the predicted effects diminished with the departure from normality (scenario 5). There was no effect of adding the second surrogate endpoint on the RMSE which was almost the same

across the three methods within each scenario. When data were simulated from non-normal distribution (t-distribution or mixture normal) with structured covariance matrix, the RMSE was slightly larger when using TRMA UCM (compared with BRMA or TRMA SCM). This is likely due to the TRMA UCM model forcing too rigid correlation structure on the data leading to bias when making predictions for outlying observations in data with not as strong correlation pattern. When data represents all outcomes correlated (from a distribution with unstructured covariance matrix), both TRMA models seem to perform equally well, with slightly smaller RMSE when using TRMA UCM if the data are normally distributed. However, for non-normally distributed data, gain in precision is only present when using TRMA SCM . When data corresponds to the scenario with the structured covariance matrix the TRMA SCM model seems to perform better than TRMA UCM in terms of both RMSE and the uncertainty of the predictions.

5. Multivariate random effects meta-analysis

Methodology introduced in Section 2 can be extended to a scenario with multiple surrogate endpoints. Suppose, we have estimates of treatment effects observed on N outcomes, $Y_{1i}, Y_{2i}, \dots, Y_{Ni}$ in each study i, and Y_N is the final clinical outcome of interest, while $Y_1, \dots, Y_{N-1}$ are intermediate surrogate endpoints. If the estimates of the treatment effects on all of the outcomes are assumed normally distributed and correlated, then they follow a multivariate normal distribution:

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{Ni} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \\ \vdots \\ \mu_{Ni} \end{pmatrix}, \mathbf{\Sigma_i}\right), \ \mathbf{\Sigma_i} = \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi}^{12} & \cdots & \sigma_{1i}\sigma_{Ni}\rho_{wi}^{1N} \\ \sigma_{2i}\sigma_{1i}\rho_{wi}^{12} & \sigma_{2i}^2 & \cdots & \sigma_{2i}\sigma_{Ni}\rho_{wi}^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{Ni}\sigma_{1i}\rho_{wi}^{1N} & \sigma_{Ni}\sigma_{2i}\rho_{wi}^{2N} & \cdots & \sigma_{Ni}^2 \end{pmatrix} \tag{18}$$

In the aforementioned model, outcomes Y_{1i}, Y_{2i}, … and Y_{Ni} are assumed to be estimates of correlated true effects μ_{1i}, μ_{2i}, … and μ_{Ni} with corresponding within-study covariance matrices $\mathbf{\Sigma_i}$ of the estimates. These study-level effects follow a multivariate normal distribution with means $(\beta_1, \beta_2, \dots, \beta_N)$ and covariance $\mathbf{T}$,

$$\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \\ \vdots \\ \mu_{Ni} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix}, \mathbf{T}\right), \ \mathbf{T} = \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho_b^{12} & \cdots & \tau_1\tau_N\rho_b^{1N} \\ \tau_2\tau_1\rho_b^{12} & \tau_2^2 & \cdots & \tau_2\tau_N\rho_b^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_N\tau_1\rho_b^{1N} & \tau_N\tau_2\rho_b^{2N} & \cdots & \tau_N^2 \end{pmatrix}. \tag{19}$$

In this hierarchical framework, Equations (18) and (19) describe the within-study and the between-study models, respectively. The between-study model can be reparameterised to extend the scenarios for modelling of surrogate endpoints described in Sections 2.3, 2.4 and 2.5 to the multivariate case, using the unstructured and structured covariance matrices.

5.1. Product normal formulation with unstructured covariance matrix

Assuming that true treatment effects on all the outcomes are correlated, we can parameterise the between-study model (19) in the form of product normal formulation by extending model (10) to

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

In this model, prior distributions need to be placed 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-study correlations and between-study standard deviations (elements of matrix T in (19) for which we are more likely to anticipate a range of values or, in some applications can obtain an external information to construct informative prior distributions for them). The relationships between the model hyper-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-study parameters (correlations and standard deviations) give implied prior distributions for those hyper-parameters and also ensure that the between-study covariance is positively defined,

$$\begin{cases} \psi_1^2 = \tau_1^2 \\ \psi_2^2 = \tau_2^2 - \lambda_{21}^2 \tau_1^2 \\ \psi_3^2 = \tau_3^2 - \lambda_{31}^2 \tau_1^2 - \lambda_{32}^2 \tau_2^2 \\ \vdots \\ \psi_N^2 = \tau_N^2 - \lambda_{N1}^2 \tau_1^2 - \lambda_{N2}^2 \tau_2^2 - \cdots - \lambda_{N(N-1)}^2 \tau_{N-1}^2. \end{cases} \tag{21}$$

$$\begin{cases} \lambda_{21} = \rho_b^{12} \tau_2 / \tau_1 \\ \lambda_{31} = \rho_b^{13} \tau_3 / \tau_1 - \lambda_{21} \\ \lambda_{32} = \left(\rho_b^{23} \tau_2 \tau_3 - \lambda_{21} \lambda_{31} \tau_1^2 \right) / \tau_2^2 \\ \vdots \\ \lambda_{NQ} = \left(\rho_b^{QN} \tau_Q \tau_N - \sum_{P=1, P \neq Q}^{N-1} \lambda_{NQ} Cov(\mu_P, \mu_Q) \right) / \tau_Q^2 \end{cases} \tag{22}$$

These relationships are obtained by calculating the variances and correlations in terms of the hyperparameters and rearranging the equations for the variances and solving the set of simultaneous equations for correlations. Details are given in Appendix B.1. This becomes a complex task in higher dimensions. Alternative and simpler model can be used by assuming a structure of the between-study covariance matrix as described in Section 5.2.

5.1.1. Criteria for surrogate markers. In the case of multiple endpoints with $N - 1$ surrogate markers, the effects of treatment on all biomarkers may jointly mediate the treatment effect on the target outcome. For the combined effect on the biomarkers to fully mediate the effect on the target outcome, we expect the intercept $\lambda_{N0} = 0$ and the conditional variance $\psi_N^2 = 0$. The association between the effect on target outcome and each of the surrogate endpoints is expected not to be zero; $\lambda_{NX} \neq 0$, $(X = 1, \ldots, N - 1)$. Similarly, the relationship between effects on other outcomes can be investigated. For example, if we consider outcomes Y_1 to Y_{M-1} to be potential surrogates to outcome Y_M $(M < N)$, then coefficients for η_{Mi}, namely, $\lambda_{M0}, \lambda_{M1}, \ldots, \lambda_{M(M-1)}$ and ψ_M^2, are investigated in the same manner, giving an option for considering multiple 'final' endpoints at the same time.

5.2. Product normal formulation with structured covariance matrix

Scenario with two surrogate endpoints and final outcome measured 'in sequence', described in Section 2.5 can be extended to the multivariate case. If we imagine that the 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-study covariance matrix in such a way to fully take into account of the correlations between the treatment effect on pairs of outcomes (for example those that are measured one after another in a time sequence, but assume conditional independence of other effects). The elements of the precision matrix T^{-1} corresponding to the effects that are conditionally independent become zero and only those on diagonal and immediate off-diagonals are non-zero. The between-study model (19) is then parameterised in the product normal,

$$\begin{cases} \mu_{1i} \sim N\left(\eta_1, \psi_1^2\right) \\ \mu_{2i} \mid \mu_{1i} \sim N\left(\eta_{2i}, \psi_2^2\right) \\ \eta_{2i} = \lambda_{20} + \lambda_{21}\mu_{1i} \\ \mu_{3i} \mid \mu_{2i} \sim N\left(\eta_{3i}, \psi_3^2\right) \\ \eta_{3i} = \lambda_{30} + \lambda_{32}\mu_{2i} \\ \vdots \\ \mu_{Ni} \mid \mu_{(N-1)i} \sim N\left(\eta_{Ni}, \psi_N^2\right) \\ \eta_{Ni} = \lambda_{N0} + \lambda_{N(N-1)}\mu_{(N-1)i}, \end{cases} \tag{23}$$

Similarly as in previous models (the trivariate and the multivariate with unstructured between-study covariance), the parameters of the above model can be expressed in terms of the elements of the between-study covariance matrix $\mathbf{T}$ (19):

$$\psi_1^2 = \tau_1^2, \quad \psi_2^2 = \tau_2^2 - \lambda_{21}^2\tau_1^2, \quad \psi_3^2 = \tau_3^2 - \lambda_{32}^2\tau_2^2, \quad \ldots, \quad \psi_N^2 = \tau_N^2 - \lambda_{N(N-1)}^2\tau_{N-1}^2 \tag{24}$$

and

$$\begin{cases} \lambda_{21} = \rho_b^{12}\frac{\tau_2}{\tau_1} \\ \lambda_{32} = \rho_b^{23}\frac{\tau_3}{\tau_2} \\ \vdots \\ \lambda_{N(N-1)} = \rho_b^{(N-1)N}\frac{\tau_N}{\tau_{(N-1)}} \end{cases} \tag{25}$$

obtained from the formulae for the correlations in Appendix B.1. Non-informative prior distributions are then placed on the between-study correlations and standard deviations, $\rho_b^{12}, \rho_b^{23}, \rho_b^{34}, \ldots, \rho_b^{(N-1)N} \sim dunif(-1, 1)$, $\tau_1, \ldots, \tau_N \sim N(0, 10)I(0,)$, as well as the remaining, independent parameters, $\eta_1 \sim N(0, 1000)$, $\lambda_{20}, \ldots, \lambda_{N0} \sim N(0, 1000)$. In this form, the model is much easier to implement compared with the model with the unstructured covariance matrix in Section 5.1.

6. Discussion

We have developed a multivariate meta-analytic framework to include multiple surrogate endpoints when predicting the treatment effect on the final clinical outcome in an early drug evaluation process. The validation process discussed here aims to evaluate how well the effect of a treatment on multiple surrogate endpoints can jointly predict treatment effect on the final clinical outcome. Two approaches were developed, one assuming all effects being correlated giving an unstructured between-study covariance matrix and a model assuming conditional independence between the effects on some of the endpoints giving a structured covariance matrix. The first model makes fewer assumptions about the data but requires estimation of a large number of parameters (which may be difficult if the number of studies is relatively low) and also may be difficult to implement in higher dimensions. The second model makes an assumption of conditional independence of some effects but leads to a reduced number of parameters to estimate and is easier to implement as the relationships between the parameters of the model and the elements of the between-study covariance matrix have a simple form. For example, in a scenario with five endpoints (four surrogate endpoints and one final outcome), the model with unstructured covariance matrix is set up to estimate five between study variances and ten between-study correlations, while the model assuming conditional independence with structured between-study covariance matrix estimates only four between-study correlations. However the assumption of conditional independence of some of the outcomes may lead to underestimating some correlations which may impact on predictions. This modelling framework, however, allows for flexible modelling of the correlation structure where a choice can be made about which correlations need to be fully taken into account. A model 'in between' those two discussed can be implemented, such as the one showed schematically in Figure 4 which extends the scenario with five endpoints in sequence (structured covariance matrix model) by taking into account of the correlation between the final and the second to last outcome (now removing the zero from the element [3,5] of the precision matrix T^{-1} which was present in the sequential model with conditional independence of effects on outcomes three and five). The desirable correlation is taken into account, while an

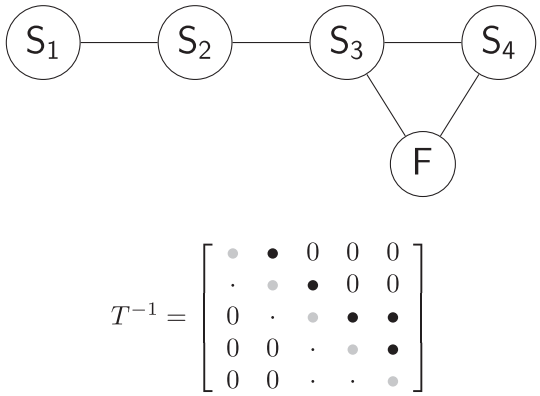

Figure 4. Example of a scenario of modelling multiple surrogate endpoints with a choice of a correlation structure.

assumption of conditional independence of the remaining outcomes is still made thereby reducing the number of correlations from ten in the model with unstructured covariance matrix to only five. See all three scenarios represented graphically in the Web Supplement D. The results of the case study and the simulation study suggest that the choice of the model (in terms of the covariance structure) when made in line with the correlation structure of the data will lead to better predictions.

Further research is required to extend the aforementioned methodology to more exact methods for Binomial or count data to relax the assumption of normality, for example, by adopting methods similar to those introduced by Stijnen *et al*. [29] using exact likelihood, by means of a generalised linear mixed models. However, our simulation study showed robustness of the methods to the departures from normality of the data. Another limitation is the prediction of the treatment effect 'estimate' rather than the true effect. Ideally, we want to know how our predicted treatment effect compares with the true treatment effect. However, although the models have the ability to predict the true effects, such effect cannot be used for validation based on the real data as the true treatment effect is unknown which limits the cross-validation to comparing the predicted treatment effect 'estimates' with the observed (but pretended missing) treatment effects for the final outcome.

The approach to validation of surrogate endpoints presented here focusses mainly on validating the endpoints as good (joint) predictors of the clinical benefit on the final outcome. It is now well established that the surrogacy has to be validated on both individual and study level. Prentice's criteria can be used to validate surrogate endpoints on the individual level when IPD are available. The use of the Prentice's criteria to estimate the within-study correlations between the treatment effects on different outcomes (by assuming one endpoint in each pair of outcomes can be considered a surrogate to the other) provides a bridge between validation on the two levels. This approach may be useful when the sharing of patient data from multiple clinical trials is problematic, but owners of the data may be willing to share Prentice's criteria.

There are some limitations of Prentice's criteria for surrogacy. As discussed by a number of authors [30–32], satisfying those criteria does not guarantee a causal relationship between the treatment effects on the surrogate and the final outcome. An example of data scenario where Prentice's fourth criterion is satisfied but it does not support a claim of causality can be found in Buyse *et al*, who discuss a number of approached to evaluation of surrogate endpoints [32]. As pointed out by Joffe and Greene [30], the statistical model describing the Prentice's criteria does not account for the common causes of the surrogate and the final outcomes. The fourth Prentice's criteria, sometimes referred to as the surrogacy criteria, is based on the assumption that the treatment effect can be partitioned into direct and indirect effects with the indirect effect being the part mediated by the surrogate and the direct being the part of the effect not mediated by the surrogate. Ignoring the common causes by modelling the direct effect on the final outcome by conditioning on the effect on the surrogate is, however, not equivalent with experimentally controlling the surrogate, as pointed out by Pearl [33]. As discussed by VanderWeele [31], this can lead to the 'surrogate paradox' where despite the positive association between treatment and the surrogate and the surrogate and the final outcome, the association between the treatment and the final outcome may be negative. Chen *et al* [34] referred to this phenomena as effect reversal.

This issue of the causal effect mostly affects surrogate evaluation when such analysis is based on a single study. The meta-analytic approach, such as described in our paper, 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) [30]. 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 as the causal association does not require experimental manipulation of the surrogate. 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, such as proposed in our study, 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. Such a single trial validation cannot guarantee that an association between effects confirmed based on individual data under one treatment will hold in other interventions. Moreover, Alonso *et al*, who investigated the relationship between the causal inference and meta-analytic approaches to the surrogate endpoint evaluation, have shown that causal effect of the surrogate on the final outcome validated based on a single study may not be confirmed in the meta-analytic setting, in particular, when the between-study heterogeneity is large and the causal effect is weak [35]. The authors have also concluded that a surrogate endpoint that successfully validated in a meta-analytic setting, on both individual and trial level, is likely to be confirmed when evaluated based on causal inference framework.

In this paper, we used the Prentice's criteria to estimate the within-study correlation between the treatment effects, as ignoring the within-study correlation has known consequences [15]. Moreover, ideally, the correlation should be obtained for each study independently to account for potential differences in the association depending on the treatment contrast. The main goal in the development of these models was building a framework under which all available evidence can be used to predict future treatment effect on the final outcome while taking into account of uncertainty around relevant parameters. In particular, these approaches have advantage over simple regression models as they take into account measurement error of the treatment effect on surrogate endpoints ignoring which can impact on predictions [12]. Further research on putting this work in the framework of causal mediation could provide solution to some of the aforementioned limitations.

Appendix A. Within-study model for relapsing remitting multiple sclerosis

In this Section, the details of how the data in Tables I and II are used to obtain the within-study standard variances and correlations.

A.1. Within-study variances

The MRI effect is modelled on the log rate ratio (RR) scale,

$$Y_{1i} = \log MRIRR = \log(Rm_E/Rm_C) \tag{A.1}$$

with

$$\sigma^2_{1i} = Var(\log MRIRR) = Var(\log Rm_E - \log Rm_C) = Var(\log Rm_E) + Var(\log Rm_C), \tag{A.2}$$

where subscript C and E stand for the control and experimental arms, respectively. From the delta method

$$Var(\log Rm_X) = \frac{1}{(Rm_X)^2} Var(Rm_X) = \left(\frac{SE(Rm_X)}{Rm_X}\right)^2 \tag{A.3}$$

in each arm X.

The relative treatment effect on relapses is modelled by the log annualised relapse rate ratio (ARRR) scale,

$$Y_{2i} = \log ARRR = \log\left(\frac{ARR_E}{ARR_C}\right) \tag{A.4}$$

with

$$\sigma^2_{2i} = Var(\log ARRR) = Var(\log ARR_E - \log ARR_C) = Var(\log ARR_E) + Var(\log ARR_C), \tag{A.5}$$

where $ARR_X = ARr_X/Nr_X$ is the annualised relapse rate in arm X. From delta method

$$\begin{aligned} Var(\log ARR_X) &= Var(\log(RelRate_X/Ft)) = Var(\log(RelRate_X) - \log(Ft)) \\ &= Var(\log(RelRate_X)) \\ &= \frac{1}{(mean(RelRate_X))^2} Var(RelRate_X) \\ &= \frac{1}{mean(RelRate_X)} = \frac{12}{ARr_x \times Ft}, \end{aligned} \tag{A.6}$$

where Ft is the follow-up time in months and $RelRate_X$ is relapse rate in arm X.

The relative treatment effect on disability progression is modelled on the log odds ratio (OR) scale,

$$Y_{3i} = \log OR = \log\left(\frac{Rd_E(Nd_C - Rd_C)}{Rd_C(Nd_E - Rd_E)}\right) \tag{A.7}$$

with the corresponding variance

$$\sigma^2_{3i} = Var(\log OR) = \frac{1}{Rd_E} + \frac{1}{Nd_E - Rd_E} + \frac{1}{Rd_C} + \frac{1}{Nd_C - Rd_C}. \tag{A.8}$$

The within-study correlations ρ^{jk}_{wi} are required to complete the within-study model. Because the correlations are not available for any of the studies and the IPD are not available for any of the studies, but the Prentice's criteria are available for one of the studies (Li for combined cohorts C and D), the approach derived in Section 2.2.2 is adopted here to calculate the within-study correlations for this particular study which is then assumed to be the same in all of the studies in meta-analysis.

A.2. Within-study correlations

Let ξ correspond to the index of the study by Li *et al*.

A.2.1. The within-study correlation between the estimates of the treatment effects on MRI and on the relapse rate. In this study,

$$\rho^{12}_{w\xi} = \Sigma_\xi[1,2]/\sqrt{\sigma^2_{1\xi}\sigma^2_{2\xi}}$$

and the covariance between log MRI RR and log ARRR, $\Sigma_\xi[1,2]$ can be expressed as

$$\begin{aligned} \Sigma_\xi[1,2] = \Sigma_\xi[2,1] &= Cov(\log ARRR, \log MRIRR) \\ &= Cov(\log ARR_E - \log ARR_C, \log Rm_E - \log Rm_C) \\ &= Cov(\log ARR_E, \log Rm_E) + Cov(\log ARR_C, \log Rm_C) \end{aligned} \tag{A.9}$$

and

$$\begin{aligned} Cov(\log ARR_X, \log Rm_X) &= Cov(\log ARr_X - \log Nr_X, \log Rm_X) \\ &= Cov(\log ARr_X, \log MRm_X) - Cov(\log Nr_X, \log Rm_X) \\ &= \sqrt{Var(\log ARr_X)Var(\log Rm_X)}\rho^*_{MR} \\ &\quad -\sqrt{Var(\log Nr_X)Var(\log Rm_X)}\hat{\rho}_{MR} \\ &= \sqrt{Var(\log ARr_X)Var(\log Rm_X)}\rho^*_{MR}. \end{aligned} \tag{A.10}$$

where the correlation ρ^*_{MR}, as discussed in Section 2, equals to the adjusted association obtained using the formula (8)

$$\rho^*_{MR} = \frac{\beta_3 - \beta_{S3}}{\alpha_3}\sqrt{\omega_{\xi MM}/\omega_{\xi RR}}. \tag{A.11}$$

Prentice's criteria for the association between the log MRI rate and log relapse rate (α_3, β_3 and β_{S3} reported in Table II) were obtained by fitting the model (6) to the full cohort of the study by Li *et al.* (rather than separately to the treatment arms). We assume that the correlation is the same in the treatment arms and the whole cohort and hence, the variances $\omega_{\xi MM} = Var(\log Rm_E) + Var(\log Rm_C)$ and $\omega_{\xi RR} = Var(\log ARr_E) + Var(\log ARr_C)$.

A.2.2. The within-study correlation between the estimates of the treatment effects on relapse rate and on the disability progression. In this study the correlation can be obtained in a similar manner:

$$\rho^{23}_{w\xi} = \Sigma_\xi[2,3]/\sqrt{\sigma^2_{2\xi}\sigma^2_{3\xi}}$$

and the covariance between log ARRR and log OR, $\Sigma_\xi[2,3]$, can be expressed as

$$\begin{aligned}\Sigma_\xi[2,3] = \Sigma_\xi[3,2] &= Cov(\log ARRR, \log OR)\\ &= Cov(\log ARR_E - \log ARR_C, \log OP_E - \log OP_C)\\ &= Cov(\log ARR_E, \log OP_E) + Cov(\log ARR_C, \log OP_C)\end{aligned} \tag{A.12}$$

and

$$\begin{aligned}Cov(logARR_X, logOP_X) &= Cov(\log ARr_X - \log Nr_X, \log OP_X)\\ &= Cov(\log ARr_X, \log OP_X) - Cov(\log Nr_X, \log OP_X)\\ &= \sqrt{Var(\log ARr_X)Var(\log OP_X)}\rho^*_{RP}\\ &- \sqrt{Var(\log Nr_X)Var(\log OP_X)}\hat{\rho}_{RP}\\ &= \sqrt{Var(\log ARr_X)Var(\log OP_X)}\rho^*_{RP}\end{aligned} \tag{A.13}$$

where OP_X refers to the odds of progression in arm X and $Var(logOP_X) = \frac{1}{Rd_X} + \frac{1}{Nd_X - Rd_X}$. The correlation ρ^*_{RP} is the adjusted association obtained using the formula (8)

$$\rho^*_{RP} = \frac{\beta_2 - \beta_{S2}}{\alpha_2}\sqrt{\omega_{\xi RR}/\omega_{\xi PP}}. \tag{A.14}$$

Prentice's criteria for the association between the log relapse rate and log odds of disability (α_2, β_2 and β_{S2} reported in Table II) were obtained by fitting the model (6) to the full cohort of the study by Li *et al.* hence, the variances $\omega_{\xi RR} = Var(\log ARR_E) + Var(\log ARR_C)$ and $\omega_{\xi PP} = Var(\log OP_E) + Var(\log OP_C)$.

A.2.3. The within-study correlation between the estimates of the treatment effects on MRI and on the disability progression. In this study the correlation is obtained using similar approach:

$$\rho^{13}_{w\xi} = \Sigma_\xi[1,3]/\sqrt{\sigma^2_{1\xi}\sigma^2_{3\xi}}$$

and the covariance between log MRI RR and log OR, $\Sigma_\xi[1,3]$ can be expressed as

$$\begin{aligned}\Sigma_i[1,3] = \Sigma_i[3,1] &= Cov(\log MRIRR, \log OR)\\ &= Cov(\log Rm_E - \log Rm_C, \log OP_E - \log OP_C)\\ &= Cov(\log Rm_E, \log OP_E) + Cov(\log Rm_C, \log OP_C)\end{aligned} \tag{A.15}$$

and

$$Cov(\log Rm_X, \log OP_X) = \sqrt{Var(\log Rm_X)Var(\log OP_X)}\rho^*_{MP}, \tag{A.16}$$

where

$$\rho^*_{MP} = \frac{\beta_1 - \beta_{S1}}{\alpha_1}\sqrt{\omega_{\xi MM}/\omega_{\xi PP}}. \tag{A.17}$$

Prentice's criteria for the association between the log MRI rate and log odds of disability (α_1, β_1 and β_{S1} reported in Table II) were obtained by fitting the model (6) to the full cohort of the study by Li *et al.* hence, the variances $\omega_{\xi MM} = Var(\log Rm_E) + Var(\log Rm_C)$ and $\omega_{\xi PP} = Var(\log OP_E) + Var(\log OP_C)$.

Appendix B. Relationships between parameters of the N-dimensional model

In this section, we derive the relationships between the model hyper-parameters (conditional variances $\psi^2_1, \psi^2_2, \ldots, \psi^2_N$ and slopes $\lambda_{21}, \lambda_{31}, \lambda_{32}, \ldots, \lambda_{N1}, \lambda_{N2}, \ldots, \lambda_{N(N-1)}$) and the between-study parameters (correlations and standard deviations).

B.1. Model with unstructured covariance

In the model with the unstructured covariance the between-study variances have the following forms:

$$\begin{cases} Var(\mu_1) = \tau^2_1 = \psi^2_1 \\ Var(\mu_2) = \tau^2_2 = \psi^2_2 + \lambda^2_{21}\psi^2_1 = \psi^2_2 + \lambda^2_{21}\tau^2_1 \\ Var(\mu_3) = \tau^2_3 = \psi^2_3 + \lambda^2_{31}\psi^2_1 + \lambda^2_{32}(\psi^2_2 + \lambda^2_{21}\psi^2_1) = \psi^2_3 + \lambda^2_{31}\tau^2_1 + \lambda^2_{32}\tau^2_2 \\ \vdots \\ Var(\mu_N) = \tau^2_N = \psi^2_N + \lambda^2_{N1}\tau^2_1 + \cdots + \lambda^2_{N(N-1)}\tau^2_{N-1}, \end{cases} \tag{B.1}$$

the covariances are

$$\begin{cases} Cov(\mu_1, \mu_2) = \lambda_{21} Var(\mu_1) \\ Cov(\mu_1, \mu_3) = \lambda_{31} Var(\mu_1) + \lambda_{32} Cov(\mu_2, \mu_1) \\ Cov(\mu_2, \mu_3) = \lambda_{32} Var(\mu_2) + \lambda_{31} Cov(\mu_2, \mu_1) \\ \vdots \\ Cov(\mu_Q, \mu_N) = \lambda_{NQ} Var(\mu_Q) + \sum\limits_{P=1, P\neq Q}^{N-1} \lambda_{NP} Cov(\mu_P, \mu_Q), \quad Q = 1, \ldots, N-1, \end{cases} \tag{B.2}$$

and corresponding correlations

$$\begin{cases} \rho^{12}_b = \frac{\lambda_{21}\tau^2_1}{\sqrt{\tau^2_1\tau^2_2}} \\ \rho^{13}_b = \frac{\lambda_{31}\tau^2_1 + \lambda_{32}\lambda_{21}\tau^2_1}{\sqrt{\tau^2_1\tau^2_3}} \\ \rho^{23}_b = \frac{\lambda_{32}\tau^2_2 + \lambda_{31}\lambda_{21}\tau^2_1}{\sqrt{\tau^2_2\tau^2_3}} \\ \vdots \\ \rho^{QN}_b = \frac{\lambda_{NQ}\tau^2_Q + \sum\limits_{P=1, P\neq Q}^{N-1} \lambda_{NQ} Cov(\mu_P, \mu_Q)}{\sqrt{\tau^2_Q\tau^2_N}}, \quad Q = 1, \ldots, N-1. \end{cases} \tag{B.3}$$

To obtain implied prior distribution for the hyper-parameters, we need to express them in terms of the between-study correlations and standard deviation. The conditional variances are obtained by rearranging terms in Equation (B.1) which gives formulae in (21). To obtain formulae for the slopes, the set of simultaneous Equations (B.3) needs to be solved giving the formulae in (22).

B.1. Model with structured covariance

In the model with structured covariance matrix, the correlations have a simple form,

$$\begin{cases} \rho_b^{12} = \lambda_{21}\frac{\tau_1}{\tau_2} \\ \rho_b^{23} = \lambda_{32}\frac{\tau_2}{\tau_3} \\ \vdots \\ \rho_b^{(N-1)N} = \lambda_{N(N-1)}\frac{\tau_{(N-1)}}{\tau_N}, \end{cases} \tag{B.4}$$

which lead to also simpler (and easier to calculate and implement) relationships (25), between the slopes and the between-study parameters.

Acknowledgements

The authors thank the two anonymous reviewers and the associate editor for their comments which helped to improve the quality of the manuscript. This research used the ALICE High Performance Computing Facility at the University of Leicester. This work was supported by the Medical Research Council (MRC) Methodology Research Programme [New Investigator Research Grant MR/L009854/1 to Sylwia Bujkiewicz]. Richard Riley is supported by funding from a multivariate meta-analysis grant from the MRC Methodology Research Programme [grant reference number: MR/J013595/1]. Keith Abrams is partially supported by the National Institute for Health Research (NIHR) as a Senior Investigator [NF-SI-0512-10159].

References

1. Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints*. Springer: New York, 2006.
2. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**:1965–1982.
3. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; **1**:49–67.
4. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 1989; **8**: 431–440.
5. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. 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)* 2001; **50**:405–422.
6. Renfro LA, Shi Q, Sargent DJ, Carlin BP. Bayesian adjusted R^2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Statistics in Medicine* 2012; **31**:743–761.
7. De Gruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL. Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a National Institutes of Health workshop. *Controlled Clinical Trials* 2001; **22**:485–502.
8. Xu J, Zeger SL. The evaluation of multiple surrogate endpoints. *Biometrics* 2001; **57**:81–87.
9. O'Brien WA, Hartigan PM, Martin D, Esinhart J, Hill A, Benoit S, Rubin M, Simberkoff MS, Hamilton JD. Changes in plasma HIV-1 amd CD4+ lymphocyte counts and the risk of progression to AIDS. *The New England Journal of Medicine* 1996; **334**:426–431.
10. Mellors JW, Muñoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P, Kingsley LA, Todd JA, Saah AJ, Detels R, Phair JP, Rinaldo Jr. CR. Plasma viral load and CD4+ lumphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine* 1997; **126**:946–954.
11. Sormani MP, Li DK, Bruzzi P, Stubinski B, Cornelisse P, Rocak S, De Stefano N. Combined MRI lesions and relapses as a surrogate for disability in multiple sclerosis. *Neurology* 2011; **77**:1684–1690.
12. Bujkiewicz S, Thompson JR, Spata E, Abrams KR. Uncertainty in the Bayesian meta-analysis of normally distributed surrogate endpoints. *Statistical Methods in Medical Research* 2015. E-pub ahead of print. DOI: 10.1177/0962280215597260.
13. Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine* 2011; **30**: 2481–2498.
14. Riley RD, Abrams KR, Lambert PC, Sutton AJ. Thompson JR An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**:78–97.
15. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A* 2009; **172**:789–811.
16. Bujkiewicz S, Thompson JR, Sutton AJ, Cooper NJ, Harrison MJ, Symmons DPM, Abrams KR. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Statistics in Medicine* 2013; **32**:3926–3943.
17. Sormani MP, Bonzano L, Roccatagliata L, Mancardi GL, Uccelli A, Bruzzi P. Surrogate endpoints for EDSS worsening in multiple sclerosis: a meta-analytic approach. *Neurology* 2010; **75**:302–309.
18. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* 2005; **24**:2401–2428.

19. Wei Y, Higgins JPT. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* 2013; **32**:1191–1205.
20. Buyse M, Molenberghs G. The validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**: 2797–2812.
21. Wei Y, Higgins JPT. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* 2013; **32**: 2911–34.
22. Higgins JPT, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* 1996; **15**:2733–2749.
23. Spiegelhalter D, Thomas A, Best N, Lunn D. 2003. WinBUGS User Manual (Version 1.4). http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf (accessed 16 Aug 1012).
24. Sturtz S, Ligges U, Gelman A. R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software* 2005; **12**:1–16.
25. Compston A, Coles A. Multiple sclerosis. *The Lancet* 2008; **372**:1502–1517.
26. Sormani MP, Bonzano L, Roccatagliata L, Cutter GR, Mancardi GL, Bruzzi P. Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: a meta-analytic approach. *Annals of Neurology* 2009; **65**:268–275.
27. Sormani MP, Stubinski B, Cornelisse P, Rocak S, Li D, De Stefano N. Magnetic resonance active lesions as individual-level surrogate for relapses in multiple sclerosis. *Multiple Sclerosis* 2011; **17**:541–549.
28. Riley RD, Abrams KR, Sutton AJ, Lambert PC. Thompson JR Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3.
29. Stijnen T, Hamzab TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine* 2010; **29**:3046–3067.
30. Joffe MM, Greene T. Related Causal Frameworks for Surrogate Outcomes. *Biometrics* 2009; **65**:530–538.
31. VanderWeele TJ. Surrogate Measures and Consistent Surrogates. *Biometrics* 2013; **69**:561–581.
32. Buyse M, Molenberghs G, Paoletti X, Oba K, Alonso A, Van der Elst W, Burzykowski T. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal* 2015. E-pub ahead of print. DOI: 10.1002/bimj.201400049.
33. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K: Cambridge University Press, 2000.
34. Chen H, Geng Z, Jia J. Criteria for surrogate end points. *Journal of the Royal Statistical Society, Series B* 2007; **69**:919–932.
35. Alonso A, Van der Elst W, Molenberghs G, Buyse M, Burzykowski T. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics* 2014; **71**:15–24.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

