STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:78–97
Published online 8 March 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2524

An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes

R. D. Riley*,†, K. R. Abrams‡, P. C. Lambert§, A. J. Sutton¶ and J. R. Thompson‖

Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, U.K.

SUMMARY

Often multiple outcomes are of interest in each study identified by a systematic review, and in this situation a separate univariate meta-analysis is usually applied to synthesize the evidence for each outcome independently; an alternative approach is a single multivariate meta-analysis model that utilizes any correlation between outcomes and obtains all the pooled estimates jointly. Surprisingly, multivariate meta-analysis is rarely considered in practice, so in this paper we illustrate the benefits and limitations of the approach to provide helpful insight for practitioners.

We compare a bivariate random-effects meta-analysis (BRMA) to two independent univariate random-effects meta-analyses (URMA), and show how and why a BRMA is able to 'borrow strength' across outcomes. Then, on application to two examples in healthcare, we show: (i) given complete data for both outcomes in each study, BRMA is likely to produce *individual* pooled estimates with very similar standard errors to those from URMA; (ii) given some studies where one of the outcomes is missing at random, the 'borrowing of strength' is likely to allow BRMA to produce *individual* pooled estimates with noticeably smaller standard errors than those from URMA; (iii) for either complete data or missing data, BRMA will produce a more appropriate standard error of the pooled *difference* between outcomes as it incorporates their correlation, which is not possible using URMA; and (iv) despite its advantages, BRMA may often not be possible due to the difficulty in obtaining the within-study correlations required to fit the model. Bivariate meta-regression and further research priorities are also discussed. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: multivariate meta-analysis; meta-regression; multiple outcomes; missing data; prognostic marker

*Correspondence to: R. D. Riley, Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, 22-28 Princess Road West, Leicester LE1 6TP, U.K.
†E-mail: rdr3@leicester.ac.uk
‡E-mail: kra1@le.ac.uk
§E-mail: pl4@le.ac.uk
¶E-mail: ajs22@le.ac.uk
‖E-mail: trj@le.ac.uk

Contract/grant sponsor: NHS HTA
Contract/grant sponsor: Department of Health

Received 1 June 2005
Accepted 9 January 2006

1. INTRODUCTION

Meta-analysis methods combine the quantitative evidence across studies to produce 'pooled' results that can facilitate evidence-based clinical practice and public heath policies [1]. Often multiple outcomes may be of interest for synthesis, and in this situation it is common for a number of separate meta-analyses to be performed, one for each outcome. For example, a recent systematic review of prognostic marker MYCN in neuroblastoma sought to extract log-hazard ratio estimates for both overall survival and disease-free survival, and then a separate meta-analysis was used to synthesize the evidence for each outcome independently [2]. The need to produce such multiple pooled results often requires the synthesis of multiple summary statistics that are correlated [3]. For instance, within each study the log-hazard ratio for disease-free survival is likely to be correlated with the log-hazard ratio for overall survival because a patient's time to recurrence of disease will often be associated with their time of death. Performing a separate meta-analysis for each outcome ignores such correlation. In contrast a multivariate meta-analysis model [4] utilizes the correlation and jointly synthesizes the outcomes to estimate the multiple pooled effects simultaneously. Glas *et al.* [5] apply a bivariate meta-analysis model to a systematic review of tumour markers used for the diagnosis of primary bladder cancer, where sensitivity and specificity were the two outcomes of interest. However, this is a rare application of multivariate meta-analysis in practice and it is more common for meta-analysts to apply a separate univariate meta-analysis to each outcome independently. The reasons for this may include tradition, the increased complexity of the multivariate approach, and perhaps a lack of understanding as to why and when multivariate meta-analysis is beneficial over and above separate univariate analyses.

In this paper our aim is to clearly illustrate the benefits and limitations of multivariate meta-analysis to provide helpful insight for practitioners considering the approach. For simplicity, we will focus on whether bivariate meta-analysis is a useful tool when two correlated outcomes are to be synthesized. Both complete data (where both outcomes are available for each study) and missing data (where one of the two outcomes is unavailable for some studies) will be assessed in relation to the model assumptions required, the data needed to fit the model, and the standard error of the pooled estimates. We also discuss areas where further research of multivariate meta-analysis is needed.

2. THE BERKEY DATA

We begin with a motivating example. Berkey *et al.* [4] were one of the first to consider meta-analysis of multiple outcomes and so for consistency let us revisit their data set involving 5 studies each assessing the difference in a surgical and non-surgical procedure for treating periodontal disease (Table I), with improvement in *probing depth* ($j=1$) and improvement in *attachment level* ($j=2$) the two outcomes of interest (measured in mm one year after treatment). In each of the $i=1$ to 5 studies, each patient received both the surgical and non-surgical procedure, using two different areas of his or her mouth. The evidence-based results of interest from a meta-analysis of these studies include: (i) an estimate ($\widehat{\beta}_1$) of the pooled difference in mean improvement in probing depth between groups, and (ii) an estimate

Table I. Details of two meta-analysis data sets, each containing 5 studies ($i = 1$ to 5): (i) the data from Berkey *et al.* [4] ('Berkey data'), and (ii) a hypothetical, modified version of the Berkey data set ('data set B').

		Berkey data			Data set B		
Study	Outcome	Y_{ij}	s_{ij}^2	λ_i (ρ_{Wi})	Y_{ij}	s_{ij}^2	λ_i (ρ_{Wi})
1	PD	0.47	0.0075	0.0030 (0.39)	0.47	0.010	0 (0)
1	AL	−0.32	0.0077		−0.32	0.0077	
2	PD	0.20	0.0057	0.0009 (0.42)	0.20	0.010	0 (0)
2	AL	−0.60	0.0008		−0.60	0.0008	
3	PD	0.40	0.0021	0.0007 (0.41)	0.40	0.010	0 (0)
3	AL	−0.12	0.0014		−0.12	0.0014	
4	PD	0.26	0.0029	0.0009 (0.43)	0.26	0.010	0 (0)
4	AL	−0.31	0.0015		−0.31	0.0015	
5	PD	0.56	0.0148	0.0072 (0.34)	0.56	0.010	0 (0)
5	AL	−0.39	0.0304		−0.39	0.0304	

PD = probing depth ($j = 1$), AL = attachment level ($j = 2$). Y_{ij} represents the difference in mean outcome improvement (surgical treatment minus non-surgical treatment), one year after treatment; i.e. Y_{i1} represents the difference in mean reduction in PD (in mm) and Y_{i2} represents the difference in mean increase in AL (in mm) between groups, one year after treatment. As Y_{ij} measures improvement, a positive Y_{ij} for either PD or AL indicates that the surgical group produces a better patient outcome. Also s_{ij} is the standard error of Y_{ij}, λ_i is the within-study covariance between Y_{i1} and Y_{i2}, and ρ_{Wi} is the within-study correlation.

($\widehat{\beta}_2$) of the pooled difference in mean improvement in attachment level between groups. As the differences relate to 'improvement' for the surgical group minus 'improvement' from the non-surgical group, a positive value of $\widehat{\beta}_j$ would indicate that the surgical group is providing the better patient outcome.

Unless individual patient data (IPD) are available, meta-analysis usually involves the extraction and then synthesis of summary statistics presented in the individual study publications. Let Y_{ij} represent the difference in mean outcome improvement (surgical treatment minus non-surgical treatment), one year after treatment; i.e. Y_{i1} represents the difference in mean reduction in probing depth (in mm) and Y_{i2} represents the difference in mean increase in attachment level (in mm) between groups, one year after treatment. To fit two independent univariate meta-analyses (one to each outcome) in the Berkey example one requires from each study Y_{i1} with associated standard error, s_{i1}, and Y_{i2} with associated standard error, s_{i2} (Table I). Alternatively, to fit a single bivariate meta-analysis model one additionally requires the within-study covariance (λ_i) to be available from each study. The statistics Y_{i1} and Y_{i2} are correlated within a study (Table I) because they both relate to outcome differences measured on the same set of patients. In addition, the bivariate approach can estimate any between-study correlation, which may exist as the outcome effects may change in a related way across studies according to each study's characteristics (e.g. age of patients, year of publication). The decision thus facing the meta-analyst here is whether to use two independent univariate meta-analyses, and ignore the correlation between outcomes, or alternatively utilize the correlation by performing a joint synthesis using a bivariate

meta-analysis. To facilitate this decision, we now formally introduce and then compare these two options.

3. UNIVARIATE RANDOM-EFFECTS META-ANALYSIS (URMA)

3.1. Random versus fixed-effects meta-analysis

In the Berkey example, the authors adopt a *random-effects* rather than a *fixed-effects* meta-analysis because the Q-statistic (a test for heterogeneity [6]) indicated that between-study heterogeneity exists for both outcomes [4]. In a random-effects meta-analysis, each study's summary statistic (Y_{ij}) ($i=1$ to n studies, $j=1$ to 2 outcomes) is assumed an estimate of a different underlying true value (θ_{ij}) in each study. In addition, each θ_{ij} is assumed to be drawn from a distribution with mean value β_j and between-study variance τ_j^2. Some authors argue that these assumptions are unjustified [7], while others insist the random-effects model is much more realistic than the fixed-effects model because between-study heterogeneity is likely to exist in practice [6], particularly across observational studies [8, 9]. In this paper we concentrate on models that take a random-effects approach.

3.2. Specification and estimation of two independent URMAs

Assuming normality of the Y_{ij}s and the θ_{ij}s, a URMA for outcome $j=1$ and a URMA for outcome $j=2$ can be written as

$$\begin{aligned} &\text{Outcome 1:} && Y_{i1} \sim \mathrm{N}(\theta_{i1}, s_{i1}^2) \\ & && \theta_{i1} \sim \mathrm{N}(\beta_1, \tau_1^2) \\ &\text{Outcome 2:} && Y_{i2} \sim \mathrm{N}(\theta_{i2}, s_{i2}^2) \\ & && \theta_{i2} \sim \mathrm{N}(\beta_2, \tau_2^2) \end{aligned} \tag{1}$$

This is the usual approach to meta-analysis of two outcomes and, as there are no correlation terms linking the separate URMAs, it is equivalent to assuming the correlations between outcomes are all zero (see Section 4). It is common practice in the meta-analysis literature to assume the s_{ij}^2s are known (even though they are estimates themselves), as this assumption makes little difference in practice [10]. Assume there is complete data for both outcomes, i.e. Y_{i1}, s_{i1}^2, Y_{i2} and s_{i2}^2 are available from each study. Using Generalized Least Squares (GLS) to estimate the parameters in equation (1) [11], the pooled estimate for outcome j can be written analytically by [6]

$$\widehat{\beta}_j(u) = \frac{\sum_{i=1}^{n} \dfrac{Y_{ij}}{s_{ij}^2 + \widehat{\tau}_j^{\,2}(u)}}{\sum_{i=1}^{n} \dfrac{1}{s_{ij}^2 + \widehat{\tau}_j^{\,2}(u)}} = \frac{\sum_{i=1}^{n} w_{ij}(u) Y_{ij}}{\sum_{i=1}^{n} w_{ij}(u)} \tag{2}$$

where $w_{ij}(u) = (s_{ij}^2 + \widehat{\tau}_j^{\,2}(u))^{-1}$ denotes the weighting of study i toward the pooled estimate $\widehat{\beta}_j(u)$, and (u) is used to distinguish that these estimates are from a *univariate* meta-analysis.

The variance of $\widehat{\beta}_j(u)$ can also be estimated using GLS and written analytically it is [6]

$$\text{var}(\widehat{\beta}_j(u)) = \frac{1}{\sum_{i=1}^{n} \dfrac{1}{s_{ij}^2 + \widehat{\tau}_j^2(u)}} = \frac{1}{\sum_{i=1}^{n} w_{ij}(u)} \tag{3}$$

In this random-effects meta-analysis model τ_j^2 also has to be estimated alongside β_j, which is why $\widehat{\tau}_j^2(u)$ is used in equations (2) and (3). This makes the estimation procedure iterative (called Iterative Generalized Least Squares (IGLS)), so that separate estimates of β_j (using equation (2)) and τ_j^2 are obtained at each iteration until a pre-specified convergence criterion (e.g. $<10^{-6}$) is reached between successive iterations for both parameters. For small sample sizes, rather than IGLS, it is often recommended to use Restricted Iterative Generalized Least Squares (RIGLS) estimation, because this provides an *unbiased* estimate of τ_j^2 [12]. One could alternatively use 'method of moments' to estimate τ_j^2, but in practice this obtains very similar estimates to RIGLS [6]. Further details and analytic solutions for $\widehat{\tau}_j^2(u)$ are shown elsewhere [13, 14].

3.3. *Application of two independent URMAs to the Berkey data set*

A URMA was applied separately to each of the two outcomes in the Berkey data set of Table I. This was done using SAS Proc Mixed (as described elsewhere [11]) and restricted

Table II. Univariate (URMA) and bivariate (BRMA) random-effects meta-analysis results for the Berkey data and for data set B (see Table I), using RIGLS estimation.

Outcome	PD		AL					
	$\widehat{\beta}_1$ (s.e.)		$\widehat{\beta}_2$ (s.e.)					$(\widehat{\beta}_1 - \hat{\beta}_2)$
Model	[95% CI]	$\hat{\tau}_1^2$	[95% CI]	$\hat{\tau}_2^2$	$\widehat{\tau}_{12}$	$\hat{\rho}_B$	$\text{corr}(\widehat{\beta}_1, \hat{\beta}_1)$	(s.e.)
Berkey data								
URMA	0.361 (0.0592) [0.196, 0.525]	0.0119	−0.346 (0.0885) [−0.591, −0.100]	0.0331	—	—	—	—
BRMA	0.353 (0.0589) [0.190, 0.517]	0.0117	−0.339 (0.0879) [−0.583, −0.095]	0.0327	0.0119	0.609	0.547	—
Data set B								
URMA	0.378 (0.0662) [0.194, 0.562]	0.0119	−0.346 (0.0885) [−0.591, −0.100]	0.0331	—	—	—	0.724 (0.111)
BRMA	0.378 (0.0662) [0.194, 0.562]	0.0119	−0.325 (0.0877) [−0.569, −0.082]	0.0329	0.0154	0.778	0.531	0.703 (0.0769)

PD = probing depth, AL = attachment level, s.e. = standard error, and CI = confidence interval (calculated using a *t*-distribution with 4 degrees of freedom). The estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ indicate the pooled difference between the two groups in improvement (surgical group minus non-surgical group) for outcome PD and AL, respectively; thus positive pooled estimates indicate that the surgical group produces a better patient outcome. The results for the hypothetical data set B are for illustrative purposes only.
N.B. The above 'Berkey data' results differ slightly to those published by Berkey *et al.* [4] as Berkey *et al.* also include an additional 'year of publication' covariate as they fitted a bivariate meta-regression, and they also did not use RIGLS estimation.

maximum likelihood estimation, which is equivalent to RIGLS for normally distributed responses as we assume here [12]. The pooled estimates indicate that the surgical procedure is better than the non-surgical procedure for probing depth ($\widehat{\beta}_1(u) = 0.361$, 95 per cent CI = 0.196 to 0.525), but that the non-surgical procedure is better than the surgical procedure for attachment level ($\widehat{\beta}_2(u) = -0.346$, 95 per cent CI = -0.591 to -0.100) (Table II). There was also evidence that some between-study heterogeneity exists for each outcome ($\widehat{\tau}^2_1(u) = 0.0119$ and $\widehat{\tau}^2_2(u) = 0.0331$).

4. BIVARIATE RANDOM-EFFECTS META-ANALYSIS (BRMA)

4.1. Specification of the BRMA model

Rather than applying two independent URMAs, one could apply a single BRMA model in order to estimate β_1 and β_2, as follows:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}, \boldsymbol{\delta}_i \right), \quad \boldsymbol{\delta}_i = \begin{pmatrix} s^2_{i1} & \lambda_i \\ \lambda_i & s^2_{i2} \end{pmatrix}$$
$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\Omega} \right), \quad \boldsymbol{\Omega} = \begin{pmatrix} \tau^2_1 & \tau_{12} \\ \tau_{12} & \tau^2_2 \end{pmatrix} \tag{4}$$

This model is a general framework for BRMA using summary statistics, as proposed elsewhere [11]; $\boldsymbol{\delta}_i$ and $\boldsymbol{\Omega}$ are the within-study and the between-study covariance matrices, respectively. The BRMA model differs from two independent URMAs by the inclusion of the within-study covariances (i.e. the λ_is) and also the between-study covariance (τ_{12}). As for a URMA the s^2_{ij}s, and now also the λ_is, are assumed known but τ_{12} must be estimated alongside τ^2_1, τ^2_2, β_1 and β_2. The model reverts to two independent URMAs when $\tau_{12} = \lambda_i = 0$, i.e. there is no within- or between-study correlation.

4.2. Estimation of the BRMA model

A discussion on how to fit and estimate this BRMA model using SAS Proc Mixed has been provided elsewhere [11]. As for URMA, one can estimate the unknown parameters in the BRMA model (i.e. β_1, β_2, τ^2_1, τ^2_2, τ_{12} in equation (4)) by using RIGLS, which now iterates between estimating the pooled values (i.e. β_1 and β_2) and $\boldsymbol{\Omega}$ (i.e. τ^2_1, τ^2_2, and τ_{12}) until the estimates for each parameter have converged (e.g. to $< 10^{-6}$). In Appendix A we provide the BRMA analytic solutions at each iteration for the pooled estimates, $\widehat{\beta}_1(b)$ and $\widehat{\beta}_2(b)$, and their variance when there is complete data, i.e. where Y_{i1}, s^2_{i1}, Y_{i2}, s^2_{i2}, and λ_i are available for each study (N.B. (b) is used here to distinguish that these estimates are from a *bivariate* meta-analysis). Although these analytic solutions are algebraically complex, an important principle is that they each involve all the parameters from both outcomes. Of course, if $\widehat{\tau}_{12}(b) = 0$

and $\lambda_i = 0$ then all the BRMA solutions revert to the URMA solutions presented in Section 3, i.e. $\widehat{\beta}_j(b) = \widehat{\beta}_j(u)$ and similarly $\widehat{\tau}^2_j(b) = \widehat{\tau}^2_j(u)$.

4.3. Application of BRMA to the Berkey data set

The BRMA of equation (4) was applied to the Berkey data using RIGLS and SAS Proc Mixed [11]. The results show that there is a reasonably strong between-study correlation across outcomes ($\hat{\rho}_B(b) = \hat{\tau}_{12}(b)/\widehat{\tau}_1(b)\,\widehat{\tau}_2(b) = 0.61$), and the BRMA pooled estimates for both outcomes are slightly closer to zero than those from the URMAs (Table II; e.g. $\widehat{\beta}_1(b) = 0.353$ and $\widehat{\beta}_1(u) = 0.361$). The standard error of the pooled estimates has also decreased very slightly in the BRMA, most likely relating to the fact that the $\hat{\tau}^2_j(b)$s are slightly smaller than the $\hat{\tau}^2_j(u)$s and that the BRMA additionally incorporates the positive within- and between-study correlations.

Although these BRMA results do not change the overall conclusions from the URMAs in Section 3.3, the findings do illustrate that for a given meta-analysis data set there may be differences between BRMA and URMA results in practice. The difficulty facing meta-analysts in this situation is to understand and coherently explain why these differences arise, and indeed they may have to decide which is more appropriate for aiding evidence-based recommendations, the URMA or the BRMA results? To aid this process we now explore the main reason why URMA and BRMA results can differ.

5. WHY AND HOW BRMA CAN 'BORROW STRENGTH' ACROSS OUTCOMES?

In this section, without loss of generality, we will primarily focus on the pooled estimate for outcome $j = 1$ and discuss why and how $\widehat{\beta}_1(b)$ may differ from $\widehat{\beta}_1(u)$.

5.1. The 'borrowing of strength' framework and its impact

$\widehat{\beta}_1(b)$ not only incorporates the data for outcome $j = 1$ (i.e. the Y_{i1}s, s^2_{i1}s and $\widehat{\tau}^2_1(b)$) but it also incorporates the $j = 2$ outcome data (i.e. the Y_{i2}s, s^2_{i2}s, and $\widehat{\tau}^2_2(b)$) through the covariance between studies (i.e. $\widehat{\tau}_{12}(b)$) and the covariance within each study (i.e. the λ_is) (Appendix A). This is clearly not true for $\widehat{\beta}_1(u)$ as this only takes into account the $j = 1$ outcome (the Y_{i1}s, the s^2_{i1}s and $\widehat{\tau}^2_1(u)$) and it does not include any correlation parameters (see equation (2)). Essentially this means that both URMA and BRMA use the data for outcome $j = 1$, but only a BRMA can in addition 'borrow strength' from the related $j = 2$ outcome data by utilizing the within- and between-study correlations.

This 'borrowing of strength' means that the weighting of each study toward the pooled estimates may be different in the BRMA than in the URMA [15]. It also enables the standard error of the pooled estimates to be potentially smaller in a BRMA compared to a URMA. For example, Riley [15] shows that for the simple situation when $\widehat{\tau}^2_j(u) = \widehat{\tau}^2_j(b)$ the variance of $\widehat{\beta}_1(b)$ is always less than or equal to the variance of $\widehat{\beta}_1(u)$. This finding is true whether the within- or between-study correlations are negative or positive, and it is

important as smaller standard errors allow more certainty in the estimation and thus may allow stronger conclusions for practice. Also, in the simple situation when τ_1^2, τ_2^2, and τ_{12} are known, the BRMA *versus* URMA debate becomes similar to that for 'seemingly unrelated regression' [16], an econometric term denoting the joint, rather than the separate, analysis of a number of correlated linear regressions, and this approach has also been shown to produce more efficient parameter estimates [16–18]. However, for random-effects meta-analysis, the between-study parameters will usually not be known, and $\widehat{\tau}_j^2(u)$ and $\widehat{\tau}_j^2(b)$ may be different as seen in the Berkey results (Table II); clearly their difference will also affect how the standard error of the pooled estimates differs between URMA and BRMA. Assessments of simulated *complete data* by Sohn [19] and Berkey *et al.* [4] both indicate that, on average, the reduction in standard error by using multivariate meta-analysis over separate univariate meta-analyses is negligible for the individual pooled estimates. However, Berkey *et al.* indicate that their simulation results may not generalize for all types of complete data and further investigation is thus required, especially for missing data which is discussed in Section 6.

5.2. Situations when there is no 'borrowing of strength' in a BRMA

If the within- and between-study correlations are all zero then there will be no 'borrowing of strength' and a BRMA will be identical to two independent URMAs. Similarly, if the s_{i1}^2s are the *same* for all studies and the within-study covariance (λ_i) is the *same* for all studies, then $\widehat{\beta}_1(b)$ will be *exactly* the same as $\widehat{\beta}_1(u)$; this issue has been reported by Nam *et al.* [20] and is demonstrated analytically elsewhere [15]. This is perhaps unintuitive because the BRMA and URMA $j=1$ parameter estimates will be identical in this situation, even if there are large within- and between-study correlation values and even if the s_{i2}^2s are very different to one another. To demonstrate this we have created a modified version of the Berkey data set, where our only change was to set all the s_{i1}^2 values to 0.01 and all the λ_is to zero (Table I, 'data set B'). The BRMA results for outcome $j=1$ for this data set are *identical* to those from a URMA, even though there is still a strong between-study correlation across outcomes ($\hat{\rho}_B(b)=0.78$, see Table II). This is because the value of s_{i1}^2 is the same for all 5 studies and so are the λ_is. However, as the s_{i2}^2s are not the same for each study the BRMA results for $j=2$ do 'borrow strength' from the $j=1$ data, which causes the standard error of $\widehat{\beta}_2(b)$ to be slightly smaller than the standard error of $\widehat{\beta}_2(u)$.

5.3. Situations where the 'borrowing of strength' is greatest in a BRMA

In practice, it is unlikely that the s_{i1}^2s will be the same for all i and thus a BRMA is likely to 'borrow strength' in most situations, though the degree to which this will modify estimates will clearly depend on the data set of interest. Indeed it has been shown elsewhere that the larger the between study correlation ($\hat{\rho}_B(b)$), the larger the differences between the s_{i1}^2s, and the larger the differences between the λ_is then the more 'borrowing of strength' can take place [15]. These last two points naturally indicate that a BRMA will perhaps be most valuable when there are some missing outcomes across studies, and so we now consider this situation further.

6. ASSESSMENT AND APPLICATION OF BRMA WHEN THERE IS MISSING DATA

6.1. Missing summary statistics across studies

A useful property of BRMA is that it can incorporate those studies where only one of the two outcomes is known (e.g. study 1 can be included even if only one of Y_{11} and Y_{12} is available with its standard error) [21]. This is not possible in a URMA, where a study will contribute no information if the single outcome of interest is not available (e.g. if Y_{11} was not available then study 1 would not be included in a URMA for outcome $j=1$). This is important because those extracting multiple summary statistics are unlikely to obtain all of them from every study [20]. For example, consider that Y_{11} and s_{11}^2 are missing for study 1 but Y_{12} and s_{12}^2 are available, and all other $(n-1)$ studies have complete data for both outcomes. In this situation:

(i) outcome $j=1$ for study 1 contributes no information toward $\widehat{\beta}_1(u)$ or $\widehat{\beta}_1(b)$.

(ii) in the URMA, outcome $j=2$ for study 1 only contributes toward $\widehat{\beta}_2(u)$ and not $\widehat{\beta}_1(u)$.

(iii) in the BRMA, outcome $j=2$ for study 1 contributes to both $\widehat{\beta}_2(b)$ and $\widehat{\beta}_1(b)$.

Thus, although outcome $j=1$ is missing for study 1, $\widehat{\beta}_1(b)$ 'borrows strength' from outcome $j=2$ of study 1. Riley [15] shows that the BRMA analytic solutions when Y_{11} is missing are equivalent to the limit of the complete data solutions (Appendix A) as s_{11}^2 tends to *infinity*; thus essentially one can consider the differences between s_{11}^2 and other known s_{i1}^2s to be extremely large, which allows great scope for the BRMA to 'borrow strength' as discussed in Section 5.3. This will be further illustrated by a real missing data example in Section 6.4.

6.2. The need to assume outcomes are 'missing at random'

An important caveat to using BRMA given missing data is that the approach is only applicable when the missing summary statistics are *missing at random* [22]. This is a necessary property for using all types of mixed models in the presence of missing data [21], but it may be particularly difficult to justify in the context of BRMA. Meta-analysis is an area where missing summary statistics are often unavailable due to publication bias, within-study selective reporting and other forms of dissemination bias [23], and in these situations the missing data may be *not missing at random*. Of course, the application of two independent URMAs also assumes that any missing summary statistics are missing at random, and both BRMA and URMA pooled estimates are potentially biased if some summary statistics are not missing at random. There are currently a wide variety of methods (e.g. the Trim and Fill method [24]) available to help measure the potential impact of dissemination bias on URMA results. Riley *et al.* [25] have shown how such methods can be implemented within a BRMA framework, but further ways of assessing dissemination bias in multivariate meta-analysis are needed to allow the approach to be more generally useful in practice.

6.3. The problem of missing within-study correlation values

BRMA is only possible when some studies provide both outcomes, otherwise one cannot estimate the between-study correlation. In such studies one needs to know the within-study

covariance (λ_i) or alternatively the within-study correlation ($\rho_{Wi} = \lambda_i / s_{i1} s_{i2}$). Unfortunately, it is unlikely that λ_i will be available from study publications in most situations [26] and this may prevent BRMA being readily used in practice [20]. Indeed, meta-analysts often have trouble just extracting the Y_{ij}s and the s_{ij}s [8], and it will inevitably be even harder to obtain the λ_is as well [3]. The Berkey data set is a rare example where λ_i is available for all studies.

What to do when λ_i is unavailable is on-going research [20, 26, 27] but there are some situations where one may plausibly assume λ_i is zero [11, 28]. For example, Thompson *et al.* [29] apply BRMA models to genetic studies of coronary heart disease that use Mendelian randomization, with the bivariate outcome of interest the genotype-disease association (where Y_{i1} is the log-odds ratio of disease given genotype in study i) and the genotype–phenotype association (where Y_{i2} is the mean change in phenotype given genotype in study i). The authors assume the within-study correlation between these outcomes was zero for each study because the difference in phenotype is often measured in a subset of the total number of subjects and the log-odds ratio of disease given genotype is based on aggregate statistics for that study [29]. However, such a re-parameterization may be more difficult to justify in other contexts.

6.4. A missing data example: joint synthesis of overall and disease-free survival

To highlight the issues in Sections 6.1–6.3, we will now consider again the systematic review and meta-analysis of marker MYCN that was mentioned briefly in Section 1. The review aimed to ascertain the overall evidence regarding the prognostic importance of MYCN in neuroblastoma, and for meta-analysis the authors sought a log-hazard ratio estimate with standard error for both disease-free survival (Y_{i1} with standard error s_{i1}) and overall survival (Y_{i2} with standard error s_{i2}) from each of the 81 studies identified [2, 25]. However, only 17 of the studies provided a log-hazard ratio estimate with standard error for both outcomes, whilst 39 provided just overall survival and the other 25 just provided disease-free survival (Table III). In this situation a BRMA provides an opportunity to 'borrow strength' across outcomes in order to limit the missing data problem (Section 6.1). For example, the BRMA could use the log-hazard ratio for overall survival to 'borrow strength' for disease-free survival when the latter was missing, and *vice versa*.

Although a BRMA is highly desirable, a problem for the approach is that λ_i was not available from any of the 17 studies providing both overall survival and disease-free survival. Furthermore, the within-study correlation is likely to be highly positive, and thus λ_i cannot be assumed zero, because a patient's time of a recurrence of disease is likely to be associated with their time of death. There is also an inherent structural relationship between these outcomes, as 'disease-free survival' is usually defined as the time to either recurrence of disease *or* death. Another issue for applying either BRMA, or indeed two independent URMAs, to this MYCN data set is that the missing summary statistics may be *not missing at random* (Section 6.2). For example, within-study selective reporting may be causing one of the outcomes to be unavailable in some studies [30, 31], and indeed there is some evidence suggesting publication bias is a problem for this data set [25].

Whilst acknowledging these problems, the MYCN data set provides an opportunity to *illustrate* the potential advantages of BRMA when there are missing summary statistics, so we will now make two assumptions: (i) the within-study correlation is 0.8 in those 17 studies providing both outcomes; and, (ii) the missing summary statistics are missing at random in

Table III. The 42 disease-free survival (DFS, $j=1$) and 56 overall survival (OS, $j=2$) estimates of the $\log_e$(hazard ratio) (Y_{ij}) and its standard error (s_{ij}) for marker MYCN from a systematic review in neuroblastoma [25].

Studies providing both outcomes			Studies providing just DFS			Studies providing just OS		
Study ID	DFS Y_{i1} (s_{i1})	OS Y_{i2} (s_{i2})	Study ID	DFS Y_{i1} (s_{i1})	OS Y_{i2} (s_{i2})	Study ID	DFS Y_{i1} (s_{i1})	OS Y_{i2} (s_{i2})
1	−0.11 (0.67)	−0.14 (0.81)	18	0.25 (0.29)	NA	43	NA	−0.84 (0.85)
2	0.30 (0.26)	0.43 (0.81)	19	0.29 (0.59)	NA	44	NA	0.05 (0.40)
3	0.41 (0.82)	0.67 (0.29)	20	0.52 (0.41)	NA	45	NA	0.73 (0.71)
4	0.47 (0.53)	0.70 (0.56)	21	0.55 (0.38)	NA	46	NA	0.76 (0.20)
5	0.76 (0.49)	0.71 (0.63)	22	0.84 (0.26)	NA	47	NA	0.91 (0.66)
6	1.06 (0.54)	1.32 (0.51)	23	0.93 (0.32)	NA	48	NA	0.93 (0.27)
7	1.46 (0.41)	1.38 (0.37)	24	1.18 (0.57)	NA	49	NA	0.96 (0.47)
8	1.64 (0.64)	1.51 (0.48)	25	1.34 (0.51)	NA	50	NA	1.05 (0.86)
9	1.64 (0.64)	1.54 (0.52)	26	1.43 (0.37)	NA	51	NA	1.16 (1.18)
10	1.64 (0.51)	1.82 (0.71)	27	1.44 (1.17)	NA	52	NA	1.22 (0.22)
11	1.70 (0.39)	1.83 (0.47)	28	1.45 (0.57)	NA	53	NA	1.26 (0.49)
12	1.85 (0.66)	2.08 (0.67)	29	1.52 (0.35)	NA	54	NA	1.26 (0.38)
13	1.90 (0.46)	2.59 (1.04)	30	1.60 (0.49)	NA	55	NA	1.27 (1.28)
14	1.90 (0.88)	2.75 (1.10)	31	1.62 (0.42)	NA	56	NA	1.31 (0.82)
15	2.19 (0.42)	2.90 (1.10)	32	1.77 (0.46)	NA	57	NA	1.52 (0.46)
16	2.95 (1.08)	2.99 (0.51)	33	1.90 (0.58)	NA	58	NA	1.54 (0.55)
17	5.70 (1.73)	5.70 (1.73)	34	1.92 (0.34)	NA	59	NA	1.55 (0.70)
			35	2.04 (0.62)	NA	60	NA	1.63 (0.83)
			36	2.19 (0.35)	NA	61	NA	1.67 (1.13)
			37	2.37 (1.00)	NA	62	NA	1.72 (0.67)
			38	2.39 (0.73)	NA	63	NA	1.74 (0.45)
			39	2.50 (0.76)	NA	64	NA	1.75 (0.72)
			40	2.56 (0.55)	NA	65	NA	1.75 (0.64)
			41	2.98 (0.58)	NA	66	NA	1.87 (0.57)
			42	3.29 (0.50)	NA	67	NA	2.07 (0.69)
						68	NA	2.13 (0.83)
						69	NA	2.19 (0.12)
						70	NA	2.25 (0.87)
						71	NA	2.31 (0.50)
						72	NA	2.33 (0.88)
						73	NA	2.36 (0.57)
						74	NA	2.37 (0.72)
						75	NA	2.63 (0.75)
						76	NA	2.66 (0.68)
						77	NA	2.77 (1.10)
						78	NA	2.80 (0.52)
						79	NA	3.33 (0.71)
						80	NA	3.54 (0.91)
						81	NA	5.04 (1.10)

those 64 studies providing only one outcome. These assumptions enable us to demonstrate the differences between URMA and BRMA results given missing summary statistics, but the results in Table IV *are for illustration only* as the assumptions cannot be verified.

Table IV. Pooled disease-free ($\widehat{\beta}_1$) and overall survival ($\widehat{\beta}_2$) log-hazard ratios from the univariate (URMA) and bivariate (BRMA) random-effects meta-analyses of the MYCN data set (Table III).

	Disease-free survival		Overall-survival			
Model	$\widehat{\beta}_1$ (s.e.) [95% CI]	$\widehat{\tau}_1^2$	$\widehat{\beta}_2$ (s.e.) [95% CI]	$\widehat{\tau}_2^2$	$\widehat{\rho}_B$	$(\widehat{\beta}_1 - \widehat{\beta}_2)$ (s.e.)
URMA	1.478 (0.127) [1.223, 1.734]	0.386	1.627 (0.118) [1.391, 1.863]	0.374	—	−0.149 (0.173)
BRMA assuming $\rho_{Wi} = 0.8$ in all 17 studies providing both outcomes	1.477 (0.111) [1.252, 1.702]	0.382	1.642 (0.108) [1.425, 1.858]	0.378	0.777	−0.164 (0.116)

ρ_{Wi} is the within-study correlation in study i, and $\widehat{\rho}_B$ is the between-study correlation. The results are for illustration only as the BRMA is subject to two unproven assumptions (see Section 6.4), and both URMA and BRMA results may also be subject to problems of dissemination bias [25]. s.e. = standard error, and CI = confidence interval (calculated using a t-distribution with 41 and 55 degrees of freedom for DFS and OS, respectively).

Firstly, two independent URMAs were applied and their results indicate that patients with high levels of marker MYCN are associated with a substantially increased risk of death (56 overall survival studies: pooled log-hazard ratio $\widehat{\beta}_2(u) = 1.63$, 95 per cent CI 1.39 to 1.87), and also risk of either death or recurrence of disease (42 disease-free survival studies: pooled log-hazard ratio $\widehat{\beta}_1(u) = 1.48$, 95 per cent CI 1.22 to 1.73). The BRMA model was then applied and the pooled estimates obtained were very similar to those from the URMAs (Table IV). However, alongside the large within-study correlations, the BRMA also estimates a large between-study correlation ($\widehat{\rho}_B(b) = 0.78$). The BRMA model utilizes this correlation to 'borrow strength' and, conditional on the assumptions (i) and (ii) above, it is able to obtain more precise pooled estimates than the URMAs (Table IV). The standard error of $\widehat{\beta}_1(b)$ is 0.111 whilst the standard error of $\widehat{\beta}_1(u)$ is 0.127 (a reduction of 12.6 per cent); similarly, the standard error of $\widehat{\beta}_2(b)$ is 0.108 whilst the standard error of $\widehat{\beta}_2(u)$ is 0.118 (a reduction of 8.5 per cent), and this is despite $\widehat{\tau}_2^2(b)$ being slightly larger than $\widehat{\tau}_2^2(u)$.

The reduction in the standard error of $\widehat{\beta}_j(b)$ over $\widehat{\beta}_j(u)$ is considerably more in this missing data example than in the complete data examples of Table II, where the largest reduction in standard error was only 0.0008 (for $\widehat{\beta}_2(b)$ in the results for data set B, relating to a reduction of 0.9 per cent). This indicates that where one is solely interested in the individual pooled estimates, the benefits of BRMA over URMA are perhaps only likely to be small given complete data but will become more marked given some missing data. This concurs with the

discussion in Sections 5.2, 5.3 and 6.1, and also findings from complete data investigations elsewhere [28].

7. EXTENSIONS

7.1. Using BRMA to estimate the pooled difference between outcomes

In some situations the difference between pooled estimates (i.e. $(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ from the BRMA or $(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$ from the URMAs) may also be of interest, especially if one wanted some overall score across outcomes or wanted to assess the hypothesis that $\beta_1 = \beta_2$. For example, an estimate of $(\beta_1 - \beta_2)$ is often of interest in the calculation of incremental net monetary benefit in cost effectiveness analyses [32]. Now, by definition the $\text{var}(a-b) = \text{var}(a) + \text{var}(b) - 2\,\text{cov}(a,b)$. If one only performs two independent URMAs, $\text{cov}(\widehat{\beta}_1(u), \widehat{\beta}_2(u))$ will not be available and is essentially assumed zero, making $\text{var}(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$ and thus the coverage of $(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$ potentially misleading. However, $\text{cov}(\widehat{\beta}_1(b), \widehat{\beta}_2(b))$ is available from the BRMA (see Appendix A) and thus $\text{var}(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ will be more appropriate, as will the coverage of $(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$. These issues are true for both complete and missing data situations.

Where $\text{cov}(\widehat{\beta}_1(b), \widehat{\beta}_2(b))$ is positive, $\text{var}(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ is likely to be much smaller than $\text{var}(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$. Consider, again for illustrative purposes, $(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$ and $(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ for the MYCN data (Table IV). The standard error of $(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ is 0.116, whilst the standard error of $(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$ equals 0.173. Interestingly, this reduction in standard error (of 32.9 per cent) is far greater here than for the individual pooled estimates themselves (of 12.6 and 8.5 per cent). Where $\text{cov}(\widehat{\beta}_1(b), \widehat{\beta}_2(b))$ is positive this is likely to be generally true because the standard error of $\widehat{\beta}_j(b)$ is reduced only by the 'borrowing of strength' between outcomes (see Section 5.1) whilst the standard error of $(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ is reduced by both the 'borrowing of strength' framework and also the incorporation of $\text{cov}(\widehat{\beta}_1(b), \widehat{\beta}_2(b))$. This means that for complete data, although little difference may generally exist between $\text{var}(\widehat{\beta}_j(b))$ and $\text{var}(\widehat{\beta}_j(u))$, there may still be large differences between $\text{var}(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ and $\text{var}(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$. This can be seen in the results for data set B (Table II) as the standard error of $(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ is 30.7 per cent smaller, even though the estimates and standard errors of the individual $j = 1$ parameters are equivalent for URMA and BRMA.

Where $\text{cov}(\widehat{\beta}_1(b), \widehat{\beta}_2(b))$ is negative, the $\text{var}(\widehat{\beta}_1(b) - \widehat{\beta}_2(b))$ is likely to be larger than $\text{var}(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$. This situation is particularly important because it means that the $\text{var}(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$ will be underestimated and this may lead to too much confidence being placed in the value of $(\widehat{\beta}_1(u) - \widehat{\beta}_2(u))$ for practice.

7.2. *Bivariate meta-regression*

Meta-regression is the term used to denote a meta-analysis model that seeks to reduce the between-study heterogeneity (i.e. the τ_js) by incorporating additional covariates alongside β_j, and where heterogeneity exists it is advisable to, wherever possible, explain what is causing it [9]. Indeed, by explaining the between-study heterogeneity (and thus reducing the estimates of τ_j^2) this may itself reduce the standard errors of the pooled estimates. Berkey *et al.* [4] and Van Houwelingen *et al.* [11] have previously shown how to perform bivariate meta-regression and applied it to complete data. To illustrate the application of the approach to missing data, consider the MYCN data again under the two assumptions specified in Section 6.4. Large between-study heterogeneity exists across MYCN studies (see Table III) and one possible reason for this may be related to study characteristics (such as treatment) varying over time. One way to assess this is to fit a bivariate meta-regression model that extends equation (4) by including a covariate for the 'year of study publication' [4], as follows:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}, \boldsymbol{\delta}_i \right), \quad \boldsymbol{\delta}_i = \begin{pmatrix} s_{i1}^2 & \lambda_i \\ \lambda_i & s_{i2}^2 \end{pmatrix}$$
$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} X_i, \boldsymbol{\Omega} \right), \quad \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} \tag{5}$$

The term ξ_j in equation (5) is the average change in β_j between two studies published one year apart, and to aid estimation using RIGLS we centred X_i at 1995. The pooled values (i.e. the β_js) in equation (5) thus relate specifically to a study published in 1995. We found that the pooled estimates, their standard error and the between-study estimates obtained from equation (5) were actually very similar to the original BRMA results for MYCN presented in Table IV. This was because the 'year of study publication' covariate did not seem important for either DFS ($\widehat{\xi}_1(b) = -0.0114$, 95 per cent CI: -0.0706 to 0.0478, $p = 0.93$) or OS ($\widehat{\xi}_2(b) = -0.0027$, 95 per cent CI: -0.0605 to 0.0552, $p = 0.70$); thus the pooled estimates for a study published in 1995 are very similar to those from studies published in other years. This finding may be explained by the fact that, of the 81 studies in the meta-analysis, none were published before 1985, only 10 were published from 1985 to 1989, and 71 were published from 1990 onwards. Hence, most of the prognostic studies included have been reported following the improved method for staging and treatment of neuroblastoma that has improved survival for children with this disease over the last 15–20 years [33].

Clearly other factors must be causing the between-study heterogeneity for MCYN; these are likely to include the cut-off level, the method of marker measurement, and patients' stage of disease and age as these varied considerably across studies [8]. However, it was difficult for us to assess these factors in either a univariate or bivariate meta-regression as the relevant information was often not available from the study publications [8]; for example, 11 of the studies did not report the cut-off level used. Also, where patient level characteristics are potentially causing heterogeneity, Lambert *et al.* [34] have shown that IPD is generally required for a univariate meta-regression to be appropriate, as otherwise the statistical power to detect any relationships is very low. It seems highly plausible that this will also be true for

bivariate meta-regression, and the lack of IPD for the MYCN studies prevented us assessing patient characteristics such as age and stage of disease further here.

8. DISCUSSION

In this paper we have demonstrated the benefits and limitations of BRMA for synthesizing two correlated outcomes from healthcare studies, using both complete and missing data examples from the literature. The work presented should therefore help practitioners understand when, how and why BRMA can differ from two independent URMAs, and this should thereby help facilitate the use of BRMA in practice, something that is currently lacking in the medical literature. The SAS Proc Mixed programs used to fit the models in this paper are all available on request from the first author.

8.1. Recommendations

We have shown that one of the main benefits of BRMA, for both complete and missing data, is in the estimation of the pooled difference between outcomes as the model allows the incorporation of $\text{cov}(\widehat{\beta}_1(b), \widehat{\beta}_2(b))$ (see Section 7.1). If one only performs an independent URMA for each outcome then $\text{cov}(\widehat{\beta}_1(u), \widehat{\beta}_2(u))$ will not be available. The need to assess the pooled difference between outcomes may actually be rare in practice, and in applications where it is relevant the individual studies themselves should report the difference between outcomes with standard error. If they do then, rather than applying a BRMA of the two outcomes, one could alternatively perform a URMA of these outcome difference estimates. However, Abrams *et al.* [35] indicate that individual studies often report the standard errors of the individual outcomes but not the standard error of the outcome difference, and so this alternative approach may not be possible.

We have also highlighted that BRMA is potentially beneficial if one is only interested in the individual outcome estimates themselves, as the model allows the 'borrowing of strength' across outcomes and thus may produce an increased precision of results compared to a URMA of each outcome independently. Although such benefits are likely to be marginal for complete data (e.g. see Table II), they may be much more apparent in applications with missing data (e.g. see the MYCN example in Table IV). Previous empirical comparisons concur that little additional benefit exists in the multivariate approach for estimating the individual parameters themselves when there is complete data [4, 19]. Further such work is needed, both for other types of complete data settings and, perhaps most appropriately, for a variety of missing data situations to assess how and under what conditions the pooled estimates, their standard error, and also the between-study variance estimates differ between models. Based on the current evidence in this and other papers we recommend that, where the individual pooled estimates are of interest, two independent URMAs are sufficient if there is complete data, but a BRMA should be preferred if there is some data missing at random across studies.

8.2. Missing information across studies

Given missing data, it may be necessary to contact the original study authors and clarify whether the outcomes unavailable were truly missing at random, and thus whether BRMA (or indeed two URMAs) is appropriate. Sometimes outcomes are 'missing by design' in the

sense that studies did not intend to collect or analyse some outcomes in the first place [36]. Outcomes 'missing by design' are more likely to be missing at random than those outcomes that were measured but not reported, as 'significant' outcome results are often more likely to be published [30]. Where the type of missing data cannot be verified, we recommend assessing the sensitivity of URMA and BRMA results to the potential impact of dissemination bias, as done for MYCN elsewhere [25].

Study authors may also help provide any unavailable within-study correlation values, which are needed to fit the BRMA model. This process may unfortunately be time-consuming and the information required may often not be available. For situations other than where the within-study correlation can be assumed zero [28], only a few articles have considered how to limit the problem of unavailable within-study correlation. One approach is suggested by Nam *et al.* [20] who consider a Bayesian approach to BRMA and perform sensitivity analyses for a range of different prior distributions for the unknown within-study correlations. Similar sensitivity analyses for the MYCN data have also been performed [15]. In another example where survival proportions at various follow-up times are of interest, Dear [26] reports an iterative method for retrospectively estimating the within-study correlation in individual studies that only report the estimate of the proportion surviving and its standard error for each time-point. One other possible solution arises where IPD are available for some studies, as in these the IPD could be used to estimate the within-study correlation directly, and the average of these values could then perhaps be used as a proxy for the missing within-study correlations in the other non-IPD studies. Raudenbush *et al.* [37] use a similar approach to this, as they approximate unknown within-study correlations using the correlation observed in other available data. A similar option would be to use the correlations obtained from the IPD to form a prior distribution for the missing within-study correlations in a Bayesian context [35]. Unfortunately it is often difficult to obtain IPD in practice, usually due to time and financial constraints, but there is an encouraging drive to make IPD more commonly available for meta-analysis [8]. Of course, by imputing approximate values it is debatable whether the within-study correlations can then be assumed *known*, as is done in the BRMA of equation (4). The impact of this assumption needs further investigation, but a Bayesian approach to BRMA would also allow the uncertainty of the within-study correlations and variances to be incorporated.

8.3. Fixed versus random-effects

Some authors are against random-effects approaches to meta-analysis because, where heterogeneity exists, those studies with a large s_{ij}^2 have relatively more weighting in a random-effects model than in a fixed-effects model, and they consider this to be philosophically wrong [7]. If preferred, a bivariate fixed-effects model (i.e. where τ_1^2, τ_2^2, and $\tau_{12} = 0$ in equation (4)) is also possible [38, 39]. Although there is no between-study correlation here, this approach will still allow the within-study correlations to be utilized and, except where the s_{ij}^2s are the same (see Section 5.3), one will always obtain individual pooled estimates with larger precision than in two separate fixed-effects models, with gains again greatest where there is missing data (see Section 5.3).

Where random-effects models are deemed appropriate, we consider that the multivariate approach is a natural and sensible framework to utilize the correlation available between multiple outcomes of interest. Of course, if there is between-study heterogeneity then one

should also attempt to explain its cause if possible [9], and thus we advise that where URMA is preferred an extension to a univariate meta-regression should be considered [40], and similarly a BRMA should be extended to a bivariate meta-regression where appropriate. In Section 7.2 we showed how to extend a BRMA to a bivariate meta-regression, which again will be particularly important over separate univariate meta-regressions where there are outcomes missing at random. However, our MYCN example also highlighted the reasons why IPD will generally be required to properly assess heterogeneity [34], especially when prognostic studies are of interest [41].

8.4. Further extensions

Although we have focused on *two* potentially correlated outcomes in this paper, the benefits and limitations identified for BRMA are likely to generalize to higher order meta-analysis models, where three or more correlated outcomes are to be synthesized; for example, trivariate models have been used elsewhere to jointly synthesize three correlated outcomes [28, 38], and other higher order meta-analysis models have been applied [42]. As for BRMA, the main general benefit of the multivariate approach over URMA is likely to be in the differences between the pooled estimates, whilst the benefits for the individual pooled estimates will again become more marked given missing data. Of course, the missing at random assumption may be even harder to justify when there are two or more missing outcomes. Also, trivariate and other higher order models will inevitably require studies for which three or more within-study correlations are available. This will be extremely unlikely unless IPD are available, stressing again why the problem of unknown within-study correlations is a pressing research issue for multivariate meta-analysis.

Finally, in this paper we have only focused on multiple *outcomes*, but multivariate meta-analysis can also be applied to multiple *treatment groups* [39, 43], or even to a combination of multiple outcomes across multiple treatment groups [42]. Multivariate meta-analysis can therefore play an important role in evidence-based clinical decision-making and we thus encourage practitioners to consider the appropriate use of the approach in practice.

APPENDIX A: ANALYTIC SOLUTIONS FOR THE BRMA

The following results are taken from Riley [15]. At each iteration of the RIGLS procedure, the pooled estimates from the BRMA of equation (4) are found by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} (\boldsymbol{\Omega} + \boldsymbol{\delta}_i)^{-1}\right)^{-1} \left(\sum_{i=1}^{n} (\boldsymbol{\Omega} + \boldsymbol{\delta}_i)^{-1}\mathbf{Y}_i\right) \quad \text{and} \quad \text{cov}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} (\boldsymbol{\Omega} + \boldsymbol{\delta}_i)^{-1}\right)^{-1}$$

This RIGLS solution can be further expressed as

$$\hat{\beta}_1 = \frac{\left(\begin{array}{l} \sum_{i=1}^{n}\left[\dfrac{Y_{i1}}{(\widehat{\tau}_1^2+s_{i1}^2)(\widehat{\tau}_2^2+s_{i2}^2)-(\widehat{\tau}_{12}+\lambda_i)^2}\left[\sum_{k=1}^{n}\dfrac{(\widehat{\tau}_2^2+s_{i2}^2)(\widehat{\tau}_1^2+s_{k1}^2)-(\widehat{\tau}_{12}+\lambda_i)(\widehat{\tau}_{12}+\lambda_k)}{(\widehat{\tau}_1^2+s_{k1}^2)(\widehat{\tau}_2^2+s_{k2}^2)-(\widehat{\tau}_{12}+\lambda_k)^2}\right]\right] \\ +\sum_{i=1}^{n}\left[\dfrac{Y_{i2}}{(\widehat{\tau}_1^2+s_{i1}^2)(\widehat{\tau}_2^2+s_{i2}^2)-(\widehat{\tau}_{12}+\lambda_i)^2}\left[\sum_{k=1}^{n}\dfrac{(\widehat{\tau}_{12}(s_{i1}^2-s_{k1}^2)+\lambda_k(\widehat{\tau}_1^2+s_{i1}^2)-\lambda_i(\widehat{\tau}_1^2+s_{k1}^2))}{(\widehat{\tau}_1^2+s_{k1}^2)(\widehat{\tau}_2^2+s_{k2}^2)-(\widehat{\tau}_{12}+\lambda_k)^2}\right]\right] \end{array}\right)}{\sum_{i=1}^{n}\dfrac{\widehat{\tau}_1^2+s_{i1}^2}{(\widehat{\tau}_1^2+s_{i1}^2)(\widehat{\tau}_2^2+s_{i2}^2)-(\widehat{\tau}_{12}+\lambda_i)^2}\sum_{i=1}^{n}\dfrac{\widehat{\tau}_2^2+s_{i2}^2}{(\widehat{\tau}_1^2+s_{i1}^2)(\widehat{\tau}_2^2+s_{i2}^2)-(\widehat{\tau}_{12}+\lambda_i)^2}-\left(\sum_{i=1}^{n}\dfrac{(\widehat{\tau}_{12}+\lambda_i)}{(\widehat{\tau}_1^2+s_{i1}^2)(\widehat{\tau}_2^2+s_{i2}^2)-(\widehat{\tau}_{12}+\lambda_i)^2}\right)^2} \quad \text{(A1)}$$

$$\hat{\beta}_2 = \frac{\left(\begin{array}{l} \sum_{i=1}^{n} \left[\dfrac{Y_{i2}}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \left[\sum_{k=1}^{n} \dfrac{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{k2}^2)-(\hat{\tau}_{12}+\lambda_i)(\hat{\tau}_{12}+\lambda_k)}{(\hat{\tau}_1^2+s_{k1}^2)(\hat{\tau}_2^2+s_{k2}^2)-(\hat{\tau}_{12}+\lambda_k)^2} \right] \right] \\ +\sum_{i=1}^{n} \left[\dfrac{Y_{i1}}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \left[\sum_{k=1}^{n} \dfrac{(\hat{\tau}_{12}(s_{i2}^2-s_{k2}^2)+\lambda_k(\hat{\tau}_2^2+s_{i2}^2)-\lambda_i(\hat{\tau}_2^2+s_{k2}^2))}{(\hat{\tau}_1^2+s_{k1}^2)(\hat{\tau}_2^2+s_{k2}^2)-(\hat{\tau}_{12}+\lambda_k)^2} \right] \right] \end{array} \right)}{\sum_{i=1}^{n} \dfrac{\hat{\tau}_1^2+s_{i1}^2}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \sum_{i=1}^{n} \dfrac{\hat{\tau}_2^2+s_{i2}^2}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} - \left(\sum_{i=1}^{n} \dfrac{(\hat{\tau}_{12}+\lambda_i)}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \right)^2} \quad \text{(A2)}$$

$$\mathrm{var}(\hat{\beta}_1) = \left(\frac{\left[\sum_{i=1}^{n} \dfrac{(\hat{\tau}_1^2+s_{i1}^2)}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \right]}{\sum_{i=1}^{n} \dfrac{\hat{\tau}_1^2+s_{i1}^2}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \sum_{i=1}^{n} \dfrac{\hat{\tau}_2^2+s_{i2}^2}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} - \left(\sum_{i=1}^{n} \dfrac{\hat{\tau}_{12}+\lambda_i}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \right)^2} \right) \quad \text{(A3)}$$

$$\mathrm{var}(\hat{\beta}_2) = \left(\frac{\left[\sum_{i=1}^{n} \dfrac{(\hat{\tau}_2^2+s_{i2}^2)}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \right]}{\sum_{i=1}^{n} \dfrac{\hat{\tau}_1^2+s_{i1}^2}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \sum_{i=1}^{n} \dfrac{\hat{\tau}_2^2+s_{i2}^2}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} - \left(\sum_{i=1}^{n} \dfrac{\hat{\tau}_{12}+\lambda_i}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \right)^2} \right) \quad \text{(A4)}$$

$$\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) = \left(\frac{\left[\sum_{i=1}^{n} \dfrac{(\hat{\tau}_{12}+\lambda_i)}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \right]}{\sum_{i=1}^{n} \dfrac{\hat{\tau}_1^2+s_{i1}^2}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \sum_{i=1}^{n} \dfrac{\hat{\tau}_2^2+s_{i2}^2}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} - \left(\sum_{i=1}^{n} \dfrac{\hat{\tau}_{12}+\lambda_i}{(\hat{\tau}_1^2+s_{i1}^2)(\hat{\tau}_2^2+s_{i2}^2)-(\hat{\tau}_{12}+\lambda_i)^2} \right)^2} \right) \quad \text{(A5)}$$

In equations (A1) and (A2) $k = 1, \ldots, n$ represents the n studies, and subscript k is needed to distinguish the summation from 1 to n within the summation for $i = 1$ to n. The values of $\hat{\tau}_1^2$, $\hat{\tau}_2^2$, and $\hat{\tau}_{12}$ are their values from the previous iteration. In the main text these BRMA estimates are denoted $\hat{\beta}_j(b)$, $\mathrm{var}(\hat{\beta}_j(b))$, $\mathrm{cov}(\hat{\beta}_1(b), \hat{\beta}_2(b))$, $\hat{\tau}_j^2(b)$, and $\hat{\tau}_{12}(b)$ to distinguish them from the alternative URMA solutions.

ACKNOWLEDGEMENTS

We would like to thank the NHS HTA Programme, who funded the systematic review of prognostic markers in neuroblastoma, and also the Department of Health, who funded Richard Riley on a Training Fellowship and currently a Post-Doc Fellowship in Evidence Synthesis that has allowed him to undertake this work. We also thank the three reviewers whose helpful comments greatly improved this paper.

REFERENCES

1. Egger M, Davey Smith G, Altman DG. *Systematic Reviews in Health Care*: *Meta-Analysis in Context*. BMJ Publishing Group: London, 2001.
2. Riley RD, Heney D, Jones DR *et al.* A systematic review of molecular and biological tumor markers in neuroblastoma. *Clinical Cancer Research* 2004; **10**:4–12.
3. Gleser L, Olkin I. Stochastically dependent effect sizes. In *The Handbook of Research Synthesis*, Cooper H, Hedges L (eds). Russell Sage Foundation: New York, 1994; 339–355.
4. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537–2550.

5. Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. *Journal of Urology* 2003; **169**:1975–1982.
6. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
7. Peto R. Why do we need systematic overviews of randomized trials? *Statistics in Medicine* 1987; **6**:233–244.
8. Riley RD, Abrams KR, Sutton AJ *et al*. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *British Journal of Cancer* 2003; **88**:1191–1198.
9. Thompson SG. Why and how sources of heterogeneity should be investigated? In *Systematic Reviews in Health Care*: *Meta-Analysis in Context*, Egger M, Davey Smith G, Altman DG (eds). BMJ Publishing Group: London, 2001; 157–175.
10. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**:619–629.
11. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589–624.
12. Goldstein H. Restricted unbiased iterative generalized least-squares estimation. *Biometrika* 1989; **76**:622–623.
13. Whitehead A. *Meta-Analysis of Controlled Clinical Trials*. Wiley: West Sussex, 2002.
14. Goldstein H. *Multilevel Statistical Models*. Edward Arnold: London, 1995.
15. Riley RD. Evidence synthesis of prognostic marker studies. *Ph.D. Thesis*, University of Leicester, 2005.
16. Zellner A. An efficient method for estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association* 1962; **57**:500–509.
17. Mehta JS, Swamy PAVB. Further evidence on the relative efficiences of Zellner's seemingly unrelated regressions estimator. *Journal of the American Statistical Association* 1976; **71**:634–639.
18. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics* 2004; **13**:461–475.
19. Sohn SY. Multivariate meta-analysis with potentially correlated marketing study results. *Naval Research Logistics* 2000; **47**:500–510.
20. Nam IS, Mengersen K, Garthwaite P. Multivariate meta-analysis. *Statistics in Medicine* 2003; **22**:2309–2333.
21. Brown H, Prescott R. *Applied Mixed Models in Medicine*. Wiley: Chichester, 1999.
22. Little JA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 2002.
23. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal* 2001; **323**:101–105.
24. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; **56**:455–463.
25. Riley RD, Sutton AJ, Abrams KR, Lambert PC. Sensitivity analyses allowed more appropriate and reliable meta-analysis conclusions for multiple outcomes when missing data was present. *Journal of Clinical Epidemiology* 2004; **57**(9):911–924.
26. Dear KB. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* 1994; **50**:989–1002.
27. Berrington A, Cox DR. Generalized least squares for the synthesis of correlated information. *Biostatistics* 2003; **4**:423–431.
28. Arends LR, Voko Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. *Statistics in Medicine* 2003; **22**:1335–1353.
29. Thompson JR, Minelli C, Abrams KR, Tobin MD, Riley RD. Meta-analysis of genetic studies using Mendelian randomisation—a multivariate approach. *Statistics in Medicine* 2005; **24**:2241–2254.
30. Hahn S, Williamson PR, Hutton JL, Garner P, Flynn EV. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Statistics in Medicine* 2000; **19**:3325–3336.
31. Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Applied Statistics* 2000; **49**:359–370.
32. Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making* 1998; **18**:S68–S80.
33. Brodeur GM, Seeger RC, Barrett A *et al*. International criteria for diagnosis, staging, and response to treatment in patients with neuroblastoma. *Journal of Clinical Oncology* 1988; **6**:1874–1881.
34. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology* 2002; **55**:86–94.
35. Abrams KR, Lambert PC, Sanso B, Shaw C, Marteau TM. Meta-analysis of heterogeneously reported study results—a Bayesian approach. In *Meta-Analysis in Medicine and Health Policy*, Stangl D, Berry D (eds). Marcel Dekker: New York, 2000; 29–63.
36. Brown CH, Indurkhya A, Kellam SK. Power calculations for data missing by design: applications to a follow up study of lead exposure and attention. *Journal of the American Statistical Association* 2000; **95**:383–385.
37. Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. *Psychological Bulletin* 1988; **103**:111–120.
38. Berkey CS, Anderson JJ, Hoaglin DC. Multiple-outcome meta-analysis of clinical trials. *Statistics in Medicine* 1996; **15**:537–557.

39. Hasselblad V. Meta-analysis of multitreatment studies. *Medical Decision Making* 1998; **18**:37–43.
40. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14**:395–411.
41. Altman DG, Riley RD. An evidence-based approach to prognostic markers. *Nature Clinical Practice Oncology* 2005; **2**:466–472.
42. DuMouchel W. Repeated measures meta-analysis. *Bulletin of the International Statistical Institute*. Session 51, Tome LVII, Book 1, 285–288.
43. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**:2273–2284.

