Biostatistics (2008), **9**, 1, *pp.* 172–186
doi:10.1093/biostatistics/kxm023
Advance Access publication on July 11, 2007

An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown

RICHARD D. RILEY*

Centre for Medical Statistics and Health Evaluation, Faculty of Medicine, University of Liverpool, Shelley's Cottage, Brownlow Street, Liverpool, England L69 3GS
richard.riley@liv.ac.uk

JOHN R. THOMPSON, KEITH R. ABRAMS

Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, Second Floor, Adrian Building, University Road, Leicester, England LE1 7RH

SUMMARY

Multivariate meta-analysis models can be used to synthesize multiple, correlated endpoints such as overall and disease-free survival. A hierarchical framework for multivariate random-effects meta-analysis includes both within-study and between-study correlation. The within-study correlations are assumed known, but they are usually unavailable, which limits the multivariate approach in practice. In this paper, we consider synthesis of 2 correlated endpoints and propose an alternative model for bivariate random-effects meta-analysis (BRMA). This model maintains the individual weighting of each study in the analysis but includes only one overall correlation parameter, ρ, which removes the need to know the within-study correlations. Further, the only data needed to fit the model are those required for a separate univariate random-effects meta-analysis (URMA) of each endpoint, currently the common approach in practice. This makes the alternative model immediately applicable to a wide variety of evidence synthesis situations, including studies of prognosis and surrogate outcomes. We examine the performance of the alternative model through analytic assessment, a realistic simulation study, and application to data sets from the literature. Our results show that, unless $\hat{\rho}$ is very close to 1 or –1, the alternative model produces appropriate pooled estimates with little bias that (i) are very similar to those from a fully hierarchical BRMA model where the within-study correlations are known and (ii) have better statistical properties than those from separate URMAs, especially given missing data. The alternative model is also less prone to estimation at parameter space boundaries than the fully hierarchical model and thus may be preferred even when the within-study correlations are known. It also suitably estimates a function of the pooled estimates and their correlation; however, it only provides an approximate indication of the between-study variation. The alternative model greatly facilitates the utilization of correlation in meta-analysis and should allow an increased application of BRMA in practice.

*To whom correspondence should be addressed.

Keywords: Correlation; Evidence synthesis; Multiple outcomes; Multivariate random-effects meta-analysis; Systematic review.

1. Introduction

Meta-analysis methods combine quantitative data from several studies to produce pooled results that aid evidence-based decision making. Multiple pooled results are required whenever there are multiple outcomes (Berkey *and others*, 1998) or multiple treatment groups (Hasselblad, 1998). For example, in diagnostic studies both sensitivity and specificity are of interest (Reitsma *and others*, 2005; Harbord *and others*, 2006), while in prognostic studies often both overall and disease-free survival are important (Riley *and others*, 2007a). In such situations, the common practice is to perform a univariate meta-analysis for each endpoint independently. This approach is simple but ignores the potential correlation between endpoints. On the other hand, a multivariate meta-analysis jointly synthesizes the endpoints and can incorporate their correlation (Raudenbush *and others*, 1988; Becker *and others*, 2000; Van Houwelingen *and others*, 2002). This can improve efficiency over separate univariate syntheses (Riley *and others*, 2007a) and allows the association between endpoints to be modeled. This facilitates the identification of surrogate outcomes (Daniels and Hughes, 1997) and the production of joint confidence or prediction regions (Reitsma *and others*, 2005).

In multivariate random-effects meta-analysis, both within-study and between-study correlation can be incorporated. The "within-study correlation" indicates the association between endpoint estimates within a study. In some situations, this might be assumed zero, for example, where the 2 endpoints are sensitivity and specificity which are calculated using separate sets of patients (Reitsma *and others*, 2005); however, for structurally dependent endpoints, like overall and disease-free survival, the within-study correlation is likely to be nonzero (Riley *and others*, 2007a). The "between-study correlation" indicates how the underlying true endpoint values are related across studies, perhaps because of differences across studies in patient-level characteristics, such as age, or changes in study-level characteristics, such as the threshold level in diagnostic studies. Both within-study and between-study correlation can influence the meta-analysis results (Riley *and others*, 2007a). The within-study correlation is most influential in the analysis when the within-study variation (i.e. the sampling error of study estimates) is large relative to the between-study variation in the underlying true study values, and the converse is true for the between-study correlation.

The within-study correlations are usually assumed known but in practice they may be difficult to obtain, especially from published information (Becker *and others*, 2000). Calculation of the within-study correlation may also be nontrivial, perhaps requiring bootstrap methods (Daniels and Hughes, 1997), and study authors may be unable to provide the correlation even on request. A number of articles consider the problem of unavailable within-study correlations. Berkey *and others* (1996) assess how their results change for a range of different within-study correlation values, while Nam *and others* (2003) perform sensitivity analyses using a range of different prior distributions for the unknown correlations. Where the multiple endpoints are survival proportions, Dear (1994) suggests a method for retrospectively estimating the within-study correlations. Raudenbush *and others* (1988) suggest using the known correlation from external data as an approximation, while Berrington and Cox (2003) limit the range of possible values for the unknown correlation between multiple relative risks.

Another issue in multivariate random-effects meta-analysis is the estimation of between-study correlation. Riley *and others* (2007b) consider maximum likelihood estimation and show that the between-study correlation is often estimated as 1 or –1, even when the within-study correlations are known. This occurs because the maximum likelihood estimator truncates the between-study covariance matrix on the boundary of its parameter space, and this often occurs when the within-study variation is relatively large or the number of studies is small. However, it is associated with an upward bias in the between-study variance

estimates, which can inflate the mean-square error and standard error of pooled estimates. Thompson *and others* (2005) suggest a reparameterized model that avoids estimating the between-study correlation, at the loss of making strong within-study and between-study correlation assumptions.

In this paper, we consider synthesis of 2 correlated endpoints and propose an alternative model for bivariate random-effects meta-analysis (BRMA) to help alleviate the aforementioned problems associated with the within-study and between-study correlation. Our new model includes only one overall correlation parameter, which removes the need to know the within-study correlations or estimate the between-study correlation, and we show that the alternative model produces appropriate pooled estimates that are superior to those from separate univariate syntheses. In Section 2, we introduce our alternative model in relation to the general BRMA model proposed by Van Houwelingen *and others* (2002). In Section 3, we compare analytically the models and perform a realistic simulation study to examine the statistical properties of their estimates. In Section 4, we then illustrate the benefits and limitations of the alternative model through application to data sets from the literature. Section 5 contains a critical discussion and makes recommendations for practice and for future research priorities.

2. MODELS FOR BRMA

2.1 *The general model for BRMA*

Suppose that 2 endpoints, $j = 1$ or 2, are available from each of $i = 1$ to n studies. Each study supplies summary measures, Y_{ij}, and associated standard errors, s_{ij}, for each endpoint. For instance, for prognostic studies Riley *and others* (2007a) set Y_{i1} and Y_{i2} to be the log-hazard ratio for disease-free survival and overall survival, respectively. Each summary statistic (Y_{ij}) is assumed to be an estimate of a true value (θ_{ij}) in each study, and in a hierarchical structure each θ_{ij} is assumed to be drawn from a distribution with mean, or "pooled," value β_j and between-study variance τ_j^2. If Y_{ij} and θ_{ij} are both assumed normally distributed, then the general BRMA model can be written as (Van Houwelingen *and others*, 2002)

$$\begin{aligned} \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} &\sim N\left(\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}, \boldsymbol{\delta}_i\right), \quad \boldsymbol{\delta}_i = \begin{pmatrix} s_{i1}^2 & s_{i1}s_{i2}\rho_{\mathrm{W}i} \\ s_{i1}s_{i2}\rho_{\mathrm{W}i} & s_{i2}^2 \end{pmatrix}, \\ \begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} &\sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\Omega}\right), \quad \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho_{\mathrm{B}} \\ \tau_1\tau_2\rho_{\mathrm{B}} & \tau_2^2 \end{pmatrix}. \end{aligned} \tag{2.1}$$

In this general model, $\boldsymbol{\delta}_i$ and $\boldsymbol{\Omega}$ are the within-study and between-study covariance matrices, respectively, and the model is equivalent to 2 independent univariate random-effect meta-analyses (URMAs) when the within-study correlations $\rho_{\mathrm{W}i}$ and the between-study correlation ρ_{B} are all zero. In (2.1), it is common to assume that the s_{ij}^2 and the $\rho_{\mathrm{W}i}$ are known (Berkey *and others*, 1998; Van Houwelingen *and others*, 2002). The usually objective from the BRMA is to estimate β_1 and β_2 or some function of these pooled values, for example, $\beta_1 - \beta_2$. However, the estimate of correlation between β_1 and β_2 may also be of importance, for example, when assessing whether one endpoint could be a surrogate for the other (Daniels and Hughes, 1997) or calculating joint confidence regions. The θ_{ij} in (2.1) are often considered nuisance parameters and are rarely of interest. Inference and estimation are thus usually based on the marginal model, which can be written as

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}\mathbf{V}_i\right), \quad \mathbf{V}_i = \begin{pmatrix} s_{i1}^2 + \tau_1^2 & s_{i1}s_{i2}\rho_{\mathrm{W}i} + \tau_1\tau_2\rho_{\mathrm{B}} \\ s_{i1}s_{i2}\rho_{\mathrm{W}i} + \tau_1\tau_2\rho_{\mathrm{B}} & s_{i1}^2 + \tau_1^2 \end{pmatrix}. \tag{2.2}$$

In this paper, we estimate the between-study parameters (τ_1^2, τ_2^2, and ρ_{B}) and the 2 pooled values β_1 and β_2 iteratively using restricted maximum likelihood (REML) in SAS Proc Mixed, as described by

Van Houwelingen *and others* (2002). We also use Cholesky decomposition (Gentle 1998) of $\mathbf{\Omega}$ to ensure that this matrix is estimated to be positive semi-definite and therefore that the between-study correlation estimate, $\hat{\rho}_{\mathrm{B}}$, is in the range $[-1, 1]$.

2.2 *An alternative model for BRMA*

The general BRMA model of (2.1) partitions the observed variation of the Y_{i1} and Y_{i2} into within-study and between-study variation using a fully hierarchical structure. Similarly, (2.1) partitions the observed correlation between the Y_{i1} and the Y_{i2} into within-study and between-study correlation. Given known within-study correlations, one can then, at least in principle, estimate the between-study correlation. However, the within-study correlations are unlikely to be available in practice and so application of the general model may be difficult. Consider now an alternative model for BRMA, where we continue to partition the overall variation but now do not partition the overall correlation. That is, rather than partitioning the overall correlation into within-study and between-study parameters, we propose a "single" parameter, ρ, to model directly the overall correlation. This situation is specified in (2.3), which is essentially a modification of the marginal model in (2.2) to take into account the overall correlation parameter:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \mathbf{\Phi}_i\right), \quad \mathbf{\Phi}_i = \begin{pmatrix} \psi_1^2 + s_{i1}^2 & \rho\sqrt{\left(\psi_1^2 + s_{i1}^2\right)\left(\psi_2^2 + s_{i2}^2\right)} \\ \rho\sqrt{(\psi_1^2 + s_{i1}^2)(\psi_2^2 + s_{i2}^2)} & \psi_2^2 + s_{i2}^2 \end{pmatrix}. \tag{2.3}$$

By modeling the overall correlation directly, we obtain the desirable property that the within-study correlations are not required, unlike in the general model. The only information required to fit (2.3) are the Y_{ij} and the s_{ij}^2, that is, the same information needed to fit a URMA for each endpoint independently. Equation (2.3) thus provides a way to synthesize 2 endpoints and utilize their correlation, even when the within-study correlations are unknown. It can also include studies that provide only one of the 2 endpoints under a missing at random assumption.

Note that the within-study variances (i.e. the s_{ij}^2) are still specified and assumed known in the alternative model in order to preserve the individual weighting of each study. The additional variation beyond sampling error is indicated by ψ_j^2. However, ψ_j^2 is not directly equivalent to τ_j^2, the between-study variance in the general model, although in some circumstances they may be similar (see Sections 3.1 and 3.2). The reason they are different is that by not partitioning the overall correlation, the alternative model does not have a fully hierarchical structure. In Sections 3 and 4, we will examine what effect this has on the pooled estimates from the alternative model and investigate if and how they differ from the pooled estimates from the general model. In this paper, we estimate β_1, β_2, ψ_1^2, ψ_2^2, and ρ in (2.3) using a self-written program in Stata that implements the "maximize" procedure (see supplementary material available at *Biostatistics* online, http://www.biostatistics.oxfordjournals.org). It is not possible to use SAS Proc Mixed due to the nonhierarchical structure of the alternative model. Our Stata program uses the Newton–Raphson procedure to maximize iteratively the restricted log-likelihood, which can be specified as

$$-\frac{1}{2}[(n-k)\log(2\pi) - \log|\mathbf{X}^T\mathbf{X}| + \log|\Phi| + \log|\mathbf{X}^T\Phi^{-1}\mathbf{X}| + (\mathbf{Y}-\mathbf{X}\beta)^T\Phi^{-1}(\mathbf{Y}-\mathbf{X}\beta)], \tag{2.4}$$

where n is the number of studies, $k = 2$ as there are 2 endpoints, $\mathbf{Y}$ is a vector of the Y_{ij}, $\mathbf{X}$ is the design matrix, and $\mathbf{\Phi}$ is a square matrix with diagonal components $\mathbf{\Phi}_i$. Our Stata program ensured that $\mathbf{\Phi}$ is estimated positive definite by modeling ψ_1^2 and ψ_2^2 on the log-scale (see supplementary material available at *Biostatistics* online, http://www.biostatistics.oxfordjournals.org).

We verified our Stata program by also specifying and estimating the general model in this way and found that the REML estimates were practically identical to those obtained from fitting the general model in SAS Proc Mixed.

3. COMPARISON OF THE ALTERNATIVE MODEL WITH THE GENERAL MODEL

3.1 *Analytic comparison of the covariance*

It is important to compare the alternative model directly to the general model as the latter is likely to be more realistic because of its hierarchical structure. We first compare the covariance between Y_{i1} and Y_{i2} from each model, that is, the off-diagonal components of $\mathbf{V}_i$ and $\mathbf{\Phi}_i$ in (2.2) and (2.3):

$$\text{General model:} \quad \text{cov}(Y_{i1}, Y_{i2}) = \rho_{\text{B}}\sqrt{(\tau_1^2\tau_2^2)} + \rho_{\text{W}i}\sqrt{(s_{i1}^2 s_{i2}^2)}, \tag{3.1}$$

$$\text{Alternative model:} \quad \text{cov}(Y_{i1}, Y_{i2}) = \rho\sqrt{(\psi_1^2 + s_{i1}^2)(\psi_2^2 + s_{i2}^2)}. \tag{3.2}$$

As the additional variation beyond sampling error increases so that the s_{ij}^2 become relatively small, $\text{cov}(Y_{i1}, Y_{i2})$ will tend to $\rho_{\text{B}}\tau_1\tau_2$ in the general model and to $\rho\psi_1\psi_2$ in the alternative model. These are likely to be similar because the within-study parameters have little influence and so the overall correlation, ρ, will be based mainly on ρ_{B}, the between-study correlation; similarly, ψ_j^2 is also likely to be very similar to τ_j^2 in this situation. This suggests that the alternative model will closely approximate the general model when the within-study variability is relatively small.

Conversely, as the within-study variability becomes increasingly large, the 2 models are unlikely to be similar. In this situation, $\text{cov}(Y_{i1}, Y_{i2})$ will tend to $\rho_{\text{W}i}s_{i1}s_{i2}$ in the general model and to $\rho s_{i1}s_{i2}$ in the alternative model. These are likely to differ as the $\rho_{\text{W}i}$ in the general model can vary across the $i = 1$ to n studies, whereas the alternative model specifies a common ρ. Further, ρ is an amalgam of the between-study correlation and the within-study correlations as measured by the observed correlation of the Y_{i1} and Y_{i2}; however, each $\rho_{\text{W}i}$ relates to the correlation within a particular study, which is unobservable without the raw study data. Thus, when there is relatively large within-study variability it would seem particularly important to know the within-study correlations.

3.2 *Comparison of the models through simulation*

To compare the statistical properties of estimates from the general and alternative models, we performed a simulation study using a number of realistic scenarios. Each scenario related to a different specification of the general model of (2.1), from which we generated 1000 meta-analysis data sets for subsequent analysis. We chose to generate data from the general model due to its realistic hierarchical structure. The scenarios differed in 4 factors considered important in practice, namely, (a) the number of studies in the meta-analysis, (b) whether complete data were available for both endpoints or whether some data were missing for endpoint $j = 2$ (Rubin, 1976), (c) the size of the within-study variation relative to the between-study variation, and (d) the size of the within-study correlations relative to the between-study correlation. Nine such scenarios are described in Tables 1 and 2 as scenarios (i)–(ix). As with any simulation study, these cannot cover all eventualities, but we would claim that they are sufficiently wide ranging to give useful insights into the comparative performance of the different models.

Each of the 1000 meta-analysis data sets generated in each scenario were analyzed separately by fitting (i) 2 separate URMAs (i.e. as (2.1) but assuming zero within-study and between-study correlation), (ii) the general BRMA model of (2.1), with the within-study correlations known, and (iii) the alternative BRMA

Table 1. *Simulation results for a selection of complete data scenarios*

Model	No. of the 1000 simulations compared	Mean bias			Coverage			Mean-square error			Mean standard error			Mean correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1 - \hat{\beta}_2$	$\hat{\beta}_1$ (%)	$\hat{\beta}_2$ (%)	$\hat{\beta}_1 - \hat{\beta}_2$ (%)	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1 - \hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1 - \hat{\beta}_2$	
Scenario (i): Complete data for $n = 50$ studies; between-study variation large relative to within-study variation; within- and between-study correlations all 0.8														
URMA	1000	0	−0.01	0	96	95	100	0.04	0.04	0.02	0.20	0.20	0.28	0
General BRMA	1000	0	−0.01	0.01	96	96	93	0.04	0.04	0.02	0.19	0.20	0.13	0.76
Alternative BRMA	1000	0	0	0.01	95	96	94	0.04	0.04	0.02	0.19	0.19	0.14	0.74
Scenario (ii): Complete data for $n = 50$ studies; between-study variation similar to the within-study variation; within- and between-study correlations all 0.8														
URMA	1000	0	−0.01	0	96	95	100	0.01	0.01	0.01	0.10	0.11	0.15	0
General BRMA	1000	0	−0.01	0	95	94	94	0.01	0.01	0.01	0.10	0.10	0.07	0.71
Alternative BRMA	1000	0	0	0	95	94	95	0.01	0.01	0.01	0.09	0.10	0.08	0.69
Scenario (iii): Complete data for $n = 5$ studies; between-study variation large relative to within-study variation; within- and between-study correlations all 0.8														
URMA	787	0.02	0.01	0	94	93	100	0.34	0.36	0.15	0.52	0.53	0.75	0
General BRMA	787	0.02	0.01	0.01	95	93	95	0.34	0.36	0.15	0.52	0.52	0.38	0.71
Alternative BRMA	787	0.02	0.01	0.01	95	94	95	0.35	0.36	0.15	0.52	0.53	0.38	0.71
Scenario (iv): Complete data for $n = 5$ studies; between-study variation similar to within-study variation; within- and between-study correlations all 0.8														
URMA	604	−0.01	0.02	0.01	99	93	100	0.09	0.10	0.05	0.28	0.27	0.40	0
General BRMA	604	−0.02	0.02	0.01	98	93	97	0.09	0.09	0.04	0.28	0.27	0.20	0.72
Alternative BRMA	604	−0.01	0.02	0.01	98	93	97	0.09	0.09	0.04	0.28	0.28	0.21	0.71
Scenario (v): Complete data for $n = 50$ studies; between-study variation similar to within-study variation; within-study correlations −0.8 in 25 studies & 0.8 in others; between-study correlation 0.8														
URMA	1000	0	0	0	95	94	99	0.01	0.01	0.02	0.10	0.11	0.15	0
General BRMA	1000	0	0	0	95	96	93	0.01	0.01	0.01	0.09	0.09	0.09	0.52
Alternative BRMA	1000	0	0	0	95	94	96	0.01	0.01	0.01	0.10	0.10	0.12	0.29

NB. The number of simulations compared relates to those which met the comparison criteria specified in Section 3.2. For the coverage, 95% confidence intervals were calculated using a t-distribution with $n - 1$ degrees of freedom. The true value of β_1 and β_2 in the simulations was 0 and 2, respectively. Results are quoted to 2 decimal places as in each scenario the standard error of the values reported (e.g. mean bias) were <0.01.

Table 2. *Simulation results for a selection of missing data scenarios*

Model	No. of the 1000 simulations compared	Mean bias			Coverage			Mean-square error			Mean standard error			Mean correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1 - \hat{\beta}_2$	$\hat{\beta}_1$ (%)	$\hat{\beta}_2$ (%)	$\hat{\beta}_1 - \hat{\beta}_2$ (%)	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1 - \hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1 - \hat{\beta}_2$	
Scenario (vi): $n = 50$ studies; complete data for 25 studies, only endpoint 1 data for others with endpoint 2 missing completely at random; between-study variation similar to within-study variation; within- and between-study correlations all 0.8														
URMA	988	−0.01	−0.01	0	95	95	100	0.01	0.02	0.02	0.10	0.15	0.18	0
General BRMA	988	−0.01	−0.01	0	94	95	94	0.01	0.01	0.01	0.10	0.12	0.09	0.64
Alternative BRMA	988	−0.01	−0.01	0	94	95	96	0.01	0.01	0.01	0.10	0.12	0.10	0.62
Scenario (vii): $n = 10$ studies; complete data for 5 studies, only endpoint 1 data for others with endpoint 2 missing completely at random; between-study variation large relative to within-study variation; within- and between-study correlations all 0.8														
URMA	770	−0.01	0	0	96	94	99	0.15	0.34	0.24	0.39	0.54	0.67	0
General BRMA	770	−0.01	−0.01	0.01	95	94	95	0.15	0.28	0.18	0.39	0.45	0.36	0.69
Alternative BRMA	770	0	−0.01	0.01	95	94	95	0.20	0.26	0.16	0.44	0.48	0.39	0.64
Scenario (viii): $n = 50$ studies; complete data for 25 studies, only endpoint 1 data for others with endpoint 2 missing completely at random; between-study variation similar to within-study variation; within-study correlations −0.8 in 25 studies & 0.8 in others; between-study correlation 0.8														
URMA	997	0	−0.01	0	95	94	98	0.01	0.02	0.02	0.10	0.14	0.18	0
General BRMA	997	0	0	0	95	93	92	0.01	0.02	0.02	0.10	0.12	0.11	0.46
Alternative BRMA	997	0	0	0	95	95	96	0.01	0.02	0.02	0.10	0.14	0.15	0.26
Scenario (ix): $n = 50$ studies; complete data for endpoint 1, but data for endpoint 2 missing if its Y_{i2} was negative (i.e. non-ignorable missing data); between-study variation similar to within-study variation; within- and between-study correlations all 0.8														
URMA	879	0	0.52	−0.52	94	0	1	0.01	0.27	0.29	0.11	0.09	0.14	0
General BRMA	879	−0.02	0.32	−0.34	94	7	7	0.01	0.11	0.13	0.10	0.09	0.10	0.49
Alternative BRMA	879	−0.02	0.32	−0.33	95	12	11	0.01	0.11	0.11	0.11	0.10	0.13	0.70

NB. The number of simulations compared relates to those which met the comparison criteria specified in Section 3.2. For the coverage of $\hat{\beta}_j$, 95 confidence intervals were calculated using a t-distribution with $n_j - 1$ degrees of freedom, where n_j is the number of studies providing data for endpoint j. For the coverage of $\hat{\beta}_1 - \hat{\beta}_2$, 95 confidence intervals were calculated using a t-distribution with n degrees of freedom. The true value of β_1 and β_2 in the simulations was 0 and 2, respectively. Results are quoted to 2 decimal places as in each scenario the standard error of the values reported (e.g. mean bias) were <0.01.

model of (2.3), which does not require the within-study correlations. REML estimation was used to fit each model, and the mean bias, mean standard error, mean-square error, and coverage of pooled estimates $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_1 - \hat{\beta}_2$ were then compared across models. Details of our simulation exercise are shown elsewhere (Riley *and others*, 2007b); we now summarize the key findings.

Estimation issues. When the within-study variation was relatively large, or when the number of studies in the meta-analysis was small (e.g. $n = 5$), the general model often did not converge or estimated the between-study correlation, ρ_B, as 1 or -1, a phenomenon detailed elsewhere (Riley *and others*, 2007b). Also in these situations, the alternative model often did not converge or gave estimates of the overall correlation, ρ, very close to 1 or -1, which were associated with unstable pooled estimates and standard errors. As $\hat{\rho}$ becomes close to 1 or -1, the determinant of $\mathbf{\Phi}_i$ in (2.3) becomes close to zero, causing unstable maximum likelihood solutions; for example, for one of the simulated data sets from scenario (iv), where ρ was estimated as 0.999, $\hat{\beta}_1$ changes rapidly from about -0.2 to -0.45 as $\hat{\rho}$ moves from 0.95 to 1. Interestingly though, a $\hat{\rho}$ very close to 1 or -1 in the alternative model occurred less frequently than a ρ_B equal to 1 or -1 in the general model. For example, in scenario (i), 20 of the 1000 simulations estimated $\hat{\rho}_B$ as 1 in the general model, whereas ρ was always estimated to be less than 0.95 in the alternative model with no evidence of unstable solutions. Thus, even when the within-study correlations are known, those practitioners who would rather estimate a correlation away from 1 or -1 may prefer to fit the alternative model and estimate the overall correlation rather than the between-study correlation (see Section 4.1 for an example).

The problem of non-convergence and unstable estimates when $\hat{\rho}$ is close to 1 or -1 is a clear limitation of the alternative model. Given this, and to aid comparison across models, the results in Tables 1 and 2 only consider those simulated data sets which (a) gave converged estimates for both the general model and the alternative model and (b) gave a value of $-0.95 < \hat{\rho} < 0.95$ in the alternative model. The ± 0.95 limit was chosen because an inspection of results where $\hat{\rho}$ was within this range did not raise concerns about unstable estimates. Hence, Tables 1 and 2 essentially indicate the statistical properties of the alternative model when it does fit adequately; what to do when $|\hat{\rho}| > 0.95$ is discussed in Section 5.

Within-study variance small relative to between-study variance. When the within-study variation is relatively small (scenarios (i) and (iii)), the alternative model and the general model perform well and produce similar pooled estimates, as reasoned in Section 3.1, with small bias relative to the Monte Carlo standard error of the simulations (Table 1). The estimates of ψ_1^2 and ψ_2^2, the additional variation in the alternative model, are also close to the estimates of τ_1^2 and τ_2^2, the between-study variation in the general model. In comparison to 2 independent URMAs, for complete data there are benefits for estimating $\beta_1 - \beta_2$; for example, in scenario (i) the mean standard error (coverage) of $\hat{\beta}_1 - \hat{\beta}_2$ is 0.13 (93%) in the general model, 0.14 (94%) in the alternative model, and 0.28 (100%) in the URMA. For the individual pooled estimates themselves, the complete data scenarios indicate that the alternative model offers no benefit over URMA in terms of their standard error and mean-square error; however, it does provide a good estimate of their correlation which may itself be of interest. For example, in scenario (iii), the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ is 0.71 in both the alternative model and the general model, but it is of course zero in the URMA.

For missing data, there are large benefits for estimating $\beta_1 - \beta_2$ and noticeable benefits for estimating β_2 due to the missing data for endpoint $j = 2$. For example, in scenario (vii), where there is large correlation and data missing completely at random in 5 out of 10 studies for endpoint 2, the alternative model reduces the standard error and mean-square error of $\hat{\beta}_2$ from the URMA by about 11% and 24%, respectively.

Within-study variance similar to between-study variance. When the within-study variances were similar in size to the between-study variances (scenarios (ii), (iv)–(ix)), the general model and the alternative

model again gave similar pooled estimates with little bias (Tables 1 and 2). On the whole, both BRMA models also produced pooled estimates with better statistical properties than those from URMA (Tables 1 and 2), in the same manner as described above. However, unlike the scenarios where the between-study variance was large, the estimates of ψ_1^2 and ψ_2^2 are less similar to the estimates of τ_1^2 and τ_2^2 here, due to the within-study variation being more influential as discussed in Section 3.1.

Discrepant within-study correlations across studies. Scenarios (v) and (viii) involve within-study correlations that vary substantially across studies, being 0.8 in some and -0.8 in others. The alternative model again gives appropriate estimates with little bias in this situation. For example, in scenario (v), the coverage of $\hat{\beta}_1 - \hat{\beta}_2$ is 96% in the alternative model, which is closer to 95%, than the general model (93%) and the URMA (99%). Interestingly, the overall correlation in scenarios (v) and (viii) is much lower here than in other scenarios due to the negative within-study correlations. This causes the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ to be quite small in the alternative model, less than 0.30, and thus there are only very small benefits over URMA. For example, in scenario (viii), where there were data missing completely at random for endpoint 2, the standard error of $\hat{\beta}_1 - \hat{\beta}_2$ is 0.15 in the alternative model and 0.18 in the URMA, but the mean standard error of $\hat{\beta}_2$ is actually the same in both models. This indicates that, as one would expect, the use of the alternative model rather a URMA is more important when the overall correlation is large.

Extreme missing data scenario. Scenario (ix) considers a special case of missing data where, after generating complete data from the general model, we deleted the data for endpoint $j = 2$ if $Y_{i2} < 0$, i.e. non-ignorable, rather than completely random missingness (Rubin, 1976). This is akin to authors or journals not reporting negative results for endpoint $j = 2$. In this situation, the URMA gives estimates of β_2 which are upwardly biased by 0.52 on average (Table 2). However, the alternative model and the general model both "borrow strength" from the data for endpoint $j = 1$ (Riley *and others*, 2007a) and considerably reduce this bias by about 38% to 0.32. Similarly, $\hat{\beta}_1 - \hat{\beta}_2$ is less biased in the BRMA models. In scenarios like this in practice, it is conceivable that, due to the reduction in bias, the alternative model may even lead to different clinical or scientific conclusions than URMA.

Assuming the within-study correlations are all zero. For all the scenarios shown in Tables 1 and 2, we also assessed how well the general model performed when, regardless of their true value, we assumed the within-study correlations to be zero (results available upon request). This approach frequently estimates the between-study correlation to be 1 or –1, especially when the true within-study correlations were far from zero, and gives a large upward bias in the estimates of between-study variance; this in turn greatly increases the standard error and mean-square error of pooled estimates and makes them larger than those from the alternative model on average. Thus, as more suitable estimates are available from the alternative model, simply setting the within-study correlations to zero is not generally recommended unless they truly are zero or close to zero as highlighted in some previous applications (Reitsma *and others*, 2005; Daniels and Hughes, 1997; Korn *and others*, 2005; Thompson *and others*, 2005; Van Houwelingen *and others*, 2002). For example, in scenario (ix), the alternative model reduces the bias of $\hat{\beta}_2$ to 0.32, with coverage 12%, but fitting the general model assuming zero within-study correlation reduces the bias of $\hat{\beta}_2$ to only 0.42 with coverage 1%.

4. APPLICATIONS

We now apply the alternative model to 3 data sets from the medical literature, all of which have previously been considered for BRMA. The first example is a rare case where the within-study correlations are known and so the general and alternative models can be directly compared; the other 2 examples involve nonzero but unavailable within-study correlations, the more common situation in practice.

4.1 *Surrogate outcomes—Daniels and Hughes data*

Daniels and Hughes (1997) assess whether the change in CD4 cell count is a surrogate for time to either development of AIDS or death in drug trials of patients with HIV. They consider between–treatment arm log-hazard ratios of time to onset of AIDS or death (Y_{i1}) and between–treatment arm differences in mean changes in CD4 count (Y_{i2}) from pretreatment baseline to about 6 months. Fifteen relevant trials were identified. Some of the trials involved 3 or 4 treatment arms, but to enable application to BRMA we only consider outcome differences between the control arm and the first treatment arm in the reported data set (Daniels and Hughes, 1997). All 15 studies provided complete data, including the within-study correlations. These were quite small, varying between -0.22 and 0.17 with a mean of -0.08. The within-study variances for endpoint 2 had a mean value of 97 and in some studies were so large that endpoint 2 was akin to being missing; this makes a BRMA particularly appealing to borrow strength in the estimation of β_2 (Riley *and others*, 2007a). The general model estimates the between-study correlation as –1, at the boundary of its parameter space (Table 3), and thus the between-study variances will be inflated (Riley *and others*, 2007b), which likely explains why the standard errors of the pooled estimates are observed larger than for 2 URMAs. The alternative model, however, estimates a more well-defined overall correlation of -0.76 and produces pooled estimates with greater precision than those from URMA; for example, the standard error of $\hat{\beta}_2$ was 4.87 in the alternative model compared to 5.56 in the URMA, a reduction of about 12%. Furthermore, the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ was estimated as -0.70 in the alternative model, which is comparable to the estimate of -0.76 in the general model and more realistic than the estimate of zero from URMA. This estimate of association may itself help facilitate decisions regarding whether CD4 should be used as a surrogate of disease-free survival. This example shows that the alternative model may be considered worthwhile even when the within-study correlations are known.

4.2 *Prognostic studies—Riley data*

A systematic review in neuroblastoma sought to establish the prognostic importance of *MYCN*, a proto-oncogene (Riley *and others*, 2004a). In 17 studies, a log-hazard ratio estimate for "amplified" versus "non-amplified" *MYCN* was available for both disease-free survival (Y_{i1}) and overall survival (Y_{i2}); however, no studies reported the within-study correlations, which are likely to be strongly positive due to the structural relationship between these endpoints. Further, there were 64 studies which provided data for only one of the 2 endpoints. If one assumes the missing endpoints are missing completely at random, then a BRMA is desirable to increase precision and reduce mean-square error. However, the general model cannot be applied unless one makes some assumptions about the within-study correlation values (Riley *and others*, 2007a). The alternative model was thus applied, and the overall correlation was estimated as 0.80 (Table 3), which causes the standard error of pooled estimates to be smaller in the alternative model than in the URMA (Table 3); for example, the standard error of $\hat{\beta}_1$ is 0.113 in the alternative model compared to 0.127 in the URMA (a reduction of 11%). The pooled estimates themselves are very similar; the alternative model has therefore increased our confidence in the meta-analysis results provided that we can assume endpoints were missing completely at random. Whether this assumption holds is open to debate, especially as publication bias may be affecting this data set (Riley, Sutton *and others*, 2004), and so the BRMA results should perhaps be treated with caution here. However, it is important to note that the URMA also makes the missing at random assumption, and simulation scenario (ix) suggests that the alternative model is preferable to 2 URMAs even if the data are not missing at random.

4.3 *Passive smoking studies—Nam data*

Nam *and others* (2003) consider BRMA where the 2 endpoints are the log-odds ratio for developing asthma (Y_{i1}) and the log-odds ratio for developing lower respiratory disease (Y_{i2}), comparing children

Table 3. *Results from the applied examples*

Data set	Model	Pooled value for endpoint $j = 1$ $\hat{\beta}_1$ (SE)	Between-study variance for endpoint $j = 1$ $\hat{\tau}_1^2$	Pooled value for endpoint $j = 2$ $\hat{\beta}_2$ (SE)	Between-study variance for endpoint $j = 2$ $\hat{\tau}_2^2$	Between-study correlation $\hat{\rho}_B$	Overall correlation $\hat{\rho}$	Correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$
Within-study variation large relative to between-study variation for endpoint 1, but relatively small for endpoint 2								
(1) Daniels and Hughes data	URMA	−0.049 (0.070)	0.025	17.300 (5.561)	379.93	—	—	0
	General model	−0.109 (0.075)	0.048	18.314 (5.740)	412.96	−1.0	—	−0.760
	Alternative model	−0.042 (0.063)	—	14.072 (4.871)	—	—	−0.759	−0.700
Within-study variation similar to between-study variation for both endpoints								
(2) Riley Data	URMA	1.478 (0.127)	0.386	1.627 (0.118)	0.374	—	—	0
	General model	NA	NA	NA	NA	NA	NA	NA
	Alternative model	1.474 (0.113)	—	1.649 (0.113)	—	—	0.800	0.452
Within-study variation large relative to between-study variation for both endpoints								
(3) Nam data	URMA	1.260 (0.046)	0.036	1.280 (0.036)	0.019	—	—	0
	General model	NA	NA	NA	NA	NA	NA	NA
	Alternative model	1.184 (0.145)	—	1.661 (0.198)	—	—	0.997	0.467

SE = standard error; NA = model could not be applied due to missing within-study correlations.

exposed and unexposed to passive smoking. Fifty-nine relevant studies were identified of which 8 reported both endpoints but without the within-study correlations. As 51 studies only reported data for 1 endpoint, a BRMA is appealing here in order to borrow strength (Riley *and others*, 2007a). However, the general model is not applicable without making some assumptions about the missing within-study correlations or by performing sensitivity analyses (Nam *and others*, 2003). We applied the alternative model, but the overall correlation was poorly estimated as 0.997 and spurious standard errors were evident for the pooled estimates (Table 3). Further, small changes in the overall correlation caused large changes in the pooled results, and thus the alternative model results should not be used here. This failure is likely due to the within-study variances being relatively large. For instance, the between-study variance for endpoint 2 is estimated as 0.019 in the URMA, whereas the mean within-study variance for endpoint 2 is 0.23, about 12 times larger. In this situation, the alternative model is less appropriate (see Section 3.1) and another approach for dealing with the missing within-study correlations is required; for example, one could perform sensitivity analyses by fitting the general model using either imputed within-study correlation values (Berkey *and others*, 1996) or a range of different prior distributions for the unknown correlations (Nam *and others*, 2003).

5. Discussion

The Campbell Collaboration states that meta-analysts "should not ignore the dependence among study outcomes" and "should use some procedure to deal with dependence" (Becker *and others*, 2004). Multivariate meta-analysis models can facilitate this requirement as they enable the synthesis of multiple, correlated endpoints of interest. However, application to settings involving nonzero within-study correlations is difficult as studies do not usually report the correlation between endpoint estimates; this is also the case for other measures of correlation like the intra-class correlation coefficient in cluster trials (Campbell *and others*, 2004). Unfortunately, without the within-study correlations it is difficult to fit the general multivariate meta-analysis model. For this reason, application of multivariate meta-analysis "may require data imputation, which can be both complex and problematic" (Becker *and others*, 2004). Various articles have considered how to deal with unavailable within-study correlations, such as Berkey *and others* (1996) and Nam *and others* (2003). In this paper, we have introduced an alternative model for BRMA, which provides a further option for synthesizing 2 endpoints when their within-study correlations are unknown. Our new model maintains the individual weighting of each study in the analysis but includes only one overall correlation parameter, which can be considered a hybrid measure of the within-study and between-study correlations. Importantly, this removes the need to know the within-study correlations, and the data required to fit the model are the same as are needed for a separate URMA of each endpoint. Section 4 highlights some areas of potential application; others may include education (Becker *and others*, 2000), psychology (Raudenbush *and others*, 1988), and genetics (Thompson *and others*, 2005).

An important assumption of both the alternative and the general models is that the endpoints are normally distributed. This is commonly used in meta-analysis, for example, where the Y_{ij} are log-odds ratios (Nam *and others*, 2003), mean differences (Berkey *and others*, 1998), and log-event rates (Arends *and others*, 2003). Yet, it is not always appropriate. For example, for synthesis of logit-sensitivity and logit-specificity from diagnostic studies, a bivariate generalized model is preferred as the normality assumption breaks down when the proportions are close to 0 or 1 (Chu and Cole, 2006; Harbord *and others*, 2006). Where the Y_{ij} can be assumed normally distributed, our simulation study shows that when $|\hat{\rho}| < 0.95$ the alternative model produces pooled estimates with little bias that are similar to those obtained from the general BRMA model when the within-study correlations are known. The alternative model also offers benefits over 2 URMAs in a similar manner identified elsewhere for the general model (Riley *and others*,

2007b). That is, the mean-square error is smaller and standard error more appropriate for a contrast of the pooled estimates, for example, $\hat{\beta}_1 - \hat{\beta}_2$. Also, when some data are missing completely at random the alternative model produces, on average, a smaller standard error and mean-square error for the individual pooled estimates themselves. Such a missing data assumption may be hard to justify in practice (Rubin, 1976), but scenario (ix) shows that the alternative model can outperform URMA even for non-ignorable missing data. The benefits of the alternative model increase as the overall correlation increases. Also, an estimate of the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ from the alternative model may be useful for making predictions (Daniels and Hughes, 1997) or calculating joint confidence regions (Reitsma *and others*, 2005).

If practitioners are fortunate to have the within-study correlations available, or if they can be assumed zero (Thompson *and others*, 2005; Arends *and others*, 2003), then we recommend that they still perform a BRMA using the general model as this has a more realistic, hierarchical structure and allows estimation of the between-study variances, τ_1^2 and τ_2^2, whereas the alternative model only provides an approximate indication of these (see Section 3.1). However, even when the within-study correlations are known, the general model may still estimate the between-study correlation as 1 or -1, which leads to upwardly biased between-study variance estimates and thus an increase in the standard error and mean-square error of pooled estimates (Riley *and others*, 2007b). In this situation, one might still apply the alternative model as the overall correlation is often estimated more easily and away from the edge of its parameter space (see Section 4.1). Occasionally, estimation issues also arise for the alternative model (see Section 4.3); in particular, when the overall correlation is very close to 1 or -1 the pooled estimates may be unstable. This occurs most often when the within-study variation is relatively large, in which case a bivariate fixed-effects meta-analysis may be more appropriate (Berkey *and others*, 1995). Indeed, in those scenarios where the alternative model has difficulties, it seems more important to specify the within-study correlations and in which case imputing a range of values within the general model may be the best approach to take (Berkey *and others*, 1996). We restricted our use of the alternative model to when $\hat{\rho}$ was between -0.95 and 0.95, within which we did not observe problems of unstable estimates; this may have been conservative as most problems arose when $|\hat{\rho}| > 0.99$. A cautious suggestion is that practitioners who obtain a high overall correlation, say >0.9 in absolute value, should assess the robustness of pooled results to small changes in $\hat{\rho}$ as a sensitivity analysis.

It is always preferable to explain the between-study variability where possible (Thompson, 1994). The alternative model can be extended to a bivariate meta-regression to include additional study-level covariates (Berkey *and others*, 1998). However, this approach will reduce the between-study variation and make the within-study variation relatively large; this will in turn make estimating the overall correlation difficult. A multivariate meta-regression approach may thus be difficult without the within-study correlations. The ideal way to examine heterogeneity is to obtain and directly model the individual patient data for each study (Lambert *and others*, 2002), and this itself would negate the problem of unavailable within-study correlations. Yet, individual patient data may not always be available (Riley, Look *and others*, 2007) and in practice meta-analysts tend to reduce their available individual patient data to aggregate data for synthesis, suggesting that they are more comfortable with traditional aggregate data techniques (Simmonds *and others*, 2005). For future research, it is important to assess if and how the alternative model may help identify surrogate outcomes (Daniels and Hughes, 1997), make predictions, or calculate predictive regions. Extension to 3 or more correlated endpoints would also be interesting (Berkey *and others*, 1996). The underlying assumptions regarding the multivariate models perhaps also require more critical evaluation; for example, are Y_{i1} and Y_{i2} truly sampled from the same distribution, and can one really assume that the s_{i1}^2 and s_{i2}^2 are known? A Bayesian approach may overcome the latter issue as it accounts for all parameter uncertainty (Nam *and others*, 2003). The impact of dissemination bias within multivariate meta-analysis also warrants attention (Riley, Sutton *and others*, 2004).

Acknowledgments

We would like to thank Alex Sutton and Paul Lambert for helpful discussions regarding bivariate meta-analysis models. We also thank Peter Diggle and the 2 referees, whose comments and suggestions have greatly improved the content of this paper. *Conflicts of Interest:* None declared.

Funding

Department of Health's National Coordinating Centre for Research Capacity Development (RSES C2/PDA/015) to R.R. as a Research Scientist in Evidence Synthesis.

References

Arends, L. R., Voko, Z. and Stijnen, T. (2003). Combining multiple outcome measures in a meta-analysis: an application. *Statistics in Medicine* **22**, 1335–1353.

Becker, B. J., Hedges, L. V. and Pigott, T. D. (2004). Campbell Collaboration Statistical Analysis Policy Brief. *A Campbell Collaboration Resource Document*. Available at http://www.campbellcollaboration.org/ECG/policy_stat.asp.

Becker, B. J., Tinsley, H. E. A. and Brown, S. (2000). *Multivariate Meta-Analysis.* San Diego, CA: Academic Press.

Berkey, C. S., Anderson, J. J. and Hoaglin, D. C. (1996). Multiple-outcome meta-analysis of clinical trials. *Statistics in Medicine* **15**, 537–557.

Berkey, C. S., Antczak-Bouckoms, A., Hoaglin, D. C., Mosteller, F. and Pihlstrom, B. L. (1995). Multiple-outcomes meta-analysis of treatments for periodontal disease. *Journal of Dental Research* **74**, 1030–1039.

Berkey, C. S., Hoaglin, D. C., Antczak-Bouckoms, A., Mosteller, F. and Colditz, G. A. (1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* **17**, 2537–2550.

Berrington, A. and Cox, D. R. (2003). Generalized least squares for the synthesis of correlated information. *Biostatistics* **4**, 423–431.

Campbell, M. K., Elbourne, D. R. and Altman, D. G. (2004). CONSORT statement: extension to cluster randomised trials. *British Medical Journal* **328**, 702–708.

Chu, H. and Cole, S. R. (2006). Bivariate meta-analysis for sensitivity and specificity with sparse data: a generalized linear mixed model approach (letter to the editor). *Journal of Clinical Epidemiology* **59**, 1331–1332.

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

Dear, K. B. (1994). Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* **50**, 989–1002.

Gentle, J. E. (1998). Cholesky factorization. *Numerical Linear Algebra for Applications in Statistics.* Berlin, Germany: Springer.

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P. and Sterne, J. A. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* **8**, 239–251.

Hasselblad, V. (1998). Meta-analysis of multitreatment studies. *Medical Decision Making* **18**, 37–43.

Korn, E. L., Albert, P. S. and Mcshane, L. M. (2005). Assessing surrogates as trial endpoints using mixed models. *Statistics in Medicine* **24**, 163–182.

LAMBERT, P. C., SUTTON, A. J., ABRAMS, K. R. AND JONES, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology* **55**, 86–94.

NAM, I. S., MENGERSEN, K. AND GARTHWAITE, P. (2003). Multivariate meta-analysis. *Statistics in Medicine* **22**, 2309–2333.

RAUDENBUSH, S. W., BECKER, B. J. AND KALAIAN, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin* **103**, 111–120.

REITSMA, J. B., GLAS, A. S., RUTJES, A. W., SCHOLTEN, R. J., BOSSUYT, P. M. AND ZWINDERMAN, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* **58**, 982–990.

RILEY, R. D., ABRAMS, K. R., LAMBERT, P. C., SUTTON, A. J. AND THOMPSON, J. R. (2007a). An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* **26**, 78–97.

RILEY, R. D., ABRAMS, K. R., SUTTON, A. J., LAMBERT, P. C. AND THOMPSON, J. R. (2007b). Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* **7**, 3.

RILEY, R. D., HENEY, D., JONES, D. R., SUTTON, A. J., LAMBERT, P. C., ABRAMS, K. R., YOUNG, B., WAILOO, A. J. AND BURCHILL, S. A. (2004). A systematic review of molecular and biological tumor markers in neuroblastoma. *Clinical Cancer Research* **10**, 4–12.

RILEY, R. D., LOOK, M. P. AND SIMMONDS, M. C. (2007). Combining individual patient data and aggregate data in evidence synthesis: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology* **60**, 431–439.

RILEY, R. D., SUTTON, A. J., ABRAMS, K. R. AND LAMBERT, P. C. (2004). Sensitivity analyses allowed more appropriate and reliable meta-analysis conclusions for multiple outcomes when missing data was present. *Journal of Clinical Epidemiology* **57**, 911–924.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

SIMMONDS, M. C., HIGGINS, J. P. T., STEWART, L. A., TIERNEY, J. F., CLARKE, M. J. AND THOMPSON, S. G. (2005). Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials* **2**, 209–217.

THOMPSON, J. R., MINELLI, C., ABRAMS, K. R., TOBIN, M. D. AND RILEY, R. D. (2005). Meta-analysis of genetic studies using Mendelian randomization—a multivariate approach. *Statistics in Medicine* **24**, 2241–2254.

THOMPSON, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* **309**, 1351–1355.

VAN HOUWELINGEN, H. C., ARENDS, L. R. AND STIJNEN, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* **21**, 589–624.

[*Received October 4, 2006; first revision January 31, 2007; second revision March 5, 2007; accepted for publication April 25, 2007*]