Research Article

Statistics in Medicine

Received 7 July 2011, Accepted 8 February 2012 Published online 25 April 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5356

A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews

Jamie J. Kirkham,[a*†] Richard D. Riley[b] and Paula R. Williamson[a]

Multivariate meta-analysis allows the joint synthesis of multiple correlated outcomes from randomised trials, and is an alternative to a separate univariate meta-analysis of each outcome independently. Usually not all trials report all outcomes; furthermore, outcome reporting bias (ORB) within trials, where an outcome is measured and analysed but not reported on the basis of the results, may cause a biased set of the evidence to be available for some outcomes, potentially affecting the significance and direction of meta-analysis results. The multivariate approach, however, allows one to 'borrow strength' across correlated outcomes, to potentially reduce the impact of ORB. Assuming ORB missing data mechanisms, we aim to investigate the magnitude of bias in the pooled treatment effect estimates for multiple outcomes using univariate meta-analysis, and to determine whether the 'borrowing of strength' from multivariate meta-analysis can reduce the impact of ORB. A simulation study was conducted for a bivariate fixed effect meta-analysis of two correlated outcomes. The approach is illustrated by application to a Cochrane systematic review. Results show that the 'borrowing of strength' from a multivariate meta-analysis can reduce the impact of ORB on the pooled treatment effect estimates. We also examine the use of the Pearson correlation as a novel approach for dealing with missing within-study correlations, and provide an extension to bivariate random-effects models that reduce ORB in the presence of heterogeneity.

Keywords: correlation; multiple outcomes; multivariate meta-analysis; outcome reporting bias; systematic review

1. Introduction

In the systematic review process, selective outcome reporting is essentially a missing data problem; bias can arise if review outcomes within one or more of the included studies have been measured and analysed but not reported on the basis of the results. When outcome reporting is driven by the significance and/or direction of the effect size (e.g. nonsignificant outcomes with negative results are reported only as $p > 0.05$, or negative results are suppressed altogether regardless of the significance), we refer to this as outcome reporting bias (ORB). For example, in cancer survival studies researchers often examine two outcomes: time to death and recurrence of disease. If only one outcome is found to be significant, then authors may not report the other nonsignificant outcome; similarly, if both outcomes are nonsignificant and/or negative (e.g. treatment is not effective), they may choose to only fully report results for one of the outcomes to save space.

Empirical research provides strong evidence that outcome reporting bias is a significant problem in randomised controlled trials [1, 2]. These reviews concluded that significant results ($p < 0.05$) are more likely to be published, and outcomes that are statistically significant have higher odds of being fully reported than those that are not significant (range of odds ratios: 2.2 to 4.7). This is coherent with the

[a] *Department of Biostatistics, University of Liverpool, Liverpool, L69 3GS, United Kingdom*
[b] *Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, B15 2TT, United Kingdom*
**Correspondence to: Jamie Kirkham, Department of Biostatistics, University of Liverpool, Liverpool, L69 3GS, United Kingdom.*
[†] *E-mail: J.J.Kirkham@liverpool.ac.uk*

findings from a series of semistructured interviews with trialists, which found that in nearly a quarter of trials the direction of the main findings influenced the trialists decision not to analyse and hence report on prespecified outcome data [3]. Furthermore, it was found from the interviews from a randomly selected cohort of *PubMed* trialists that over a quarter (29%) of these trials were found to have displayed ORB [3].

Because of the lack of awareness of ORB, the common practice widely adopted by reviewers is simply to ignore the missing outcome data and to carry out a univariate meta-analysis (analyse each outcome separately) using complete case data, that is just those studies that report the outcome being synthesised [4]. However, given ORB, this approach may lead to biased and overestimated pooled treatment effect estimates for each outcome, because the available study estimates are a biased subset of all the evidence.

If ORB is suspected in a review, and missing data are unobtainable from trials authors, a sensitivity analysis should be considered to determine how robust the meta-analysis conclusions are to ORB. If the sensitivity analyses show that the results are not robust to outcome reporting, the review conclusions may need to be amended. In the recently published Outcome Reporting Bias in Trials (ORBIT) study [5], the maximum bias bound approach was used in a sensitivity analysis [6] to estimate the impact of outcome reporting bias on the review meta-analysis. The limitation of this method is that it does not take into account any extra information that is known about the studies that have not reported the outcome of interest, for example, the sample size and whether a high or low risk of ORB is suspected. Furthermore, the method assumes that larger studies are more likely to be published than smaller studies, so in the case of ORB, the larger studies are more likely to fully report the outcome of interest compared with smaller studies. Finally, the method can only be applied to single outcomes. In practice, multiple outcomes are of interest in the systematic review process, for example overall and disease-free survival as previously mentioned. Such outcomes are often correlated, and as each patient in a study provides data toward each outcome, this results in correlated outcome effect estimates within studies. Evidence synthesis of multiple outcomes can be performed using a multivariate meta-analysis [4, 7, 8] whereby the study effect estimates (which often relate to a treatment effect) for all available outcomes are jointly synthesised, while accounting for their within-study correlation. This has been shown to improve statistical properties of pooled estimates (such as bias and means-square error) compared with a univariate meta-analysis of each outcome separately [8, 9]. This gain is particularly apparent in the presence of missing data [9], that is, where not all studies provide estimates for all outcomes.

The severity of the impact results found in the ORBIT project clearly warrants further research into statistical approaches for adjusting for ORB that affects meta-analysis of multiple outcomes. In this paper we consider multivariate meta-analysis as a potential solution to reducing ORB. Our work is motivated by Jackson *et al.* [8], who provided examples that show multivariate meta-analysis may lead to different pooled estimates than univariate meta-analysis, and Riley [9], who analytically showed that the multivariate approach can use the correlation between outcomes to 'borrow strength' and substantially shift the pooled estimates away from the univariate estimates. The aim of our work is to extend this idea and use the method to adjust for ORB in meta-analysis, and to examine whether the multivariate fixed effect approach reduces ORB and improves the coverage, bias, mean-square error and power of pooled estimates. We also consider a novel approach for dealing with unavailable within-study correlations, a typical problem for multivariate meta-analysis [9], and make extension to random-effects models that additionally account for between-study heterogeneity and correlation.

The remainder of this paper is organised as follows. In Section 2 we provide the detail of the motivating example, where ORB is suspected in a review involving beta-lactam therapy for the treatment of cancer patients with neutropenia. In Section 3 we introduce the standard fixed effect univariate meta-analysis model (UFMA) and then extend this to the most simple multivariate meta-analysis model — the bivariate fixed effect meta-analysis model (BFMA). This section concludes with the detail on how to estimate the pooled estimates, a review of the possible options to deal with unavailable within-study correlation needed for the multivariate meta-analysis approach, and a description of a new method to approximate this correlation when it cannot be obtained from alternative sources. In Section 4 we describe a simulation study to examine how the statistical properties of the pooled effect estimates are affected by (i) the strength of the within-study correlations, (ii) whether or not there is missing data in one or both outcomes, and (iii) whether or not the within-study correlations are known or unknown. The simulation allows us to compare the performance of the multivariate approach against separate univariate fixed effect meta-analyses. In Section 5 we apply the methods to the motivating example. In Section 6, we provide extension to models for bivariate random effects meta-analysis and provide an example of

their application in an ORB situation. We conclude with a discussion of the benefits of the bivariate approach, the challenges with adopting it in practice, the limitations of our work, and future research considerations (Section 7).

2. Motivating example

2.1. Beta-lactam for the treatment of cancer patients with neutropenia

A previous study investigated the impact of selective outcome reporting in a review comparing beta-lactam and beta-lactam–aminoglycoside combination therapy in the treatment of cancer patients with neutropenia [10]. This review has subsequently undergone a substantial update [11] and we present this new data as a motivating example. The review contains multiple intervention comparisons but we only consider here the meta-analysis of 15 trials in which the same beta-lactam was used in each arm. The review did not assess or comment on ORB in the risk of bias table. The two outcomes of primary interest were treatment failure and all-cause mortality. All 15 eligible trials reported on treatment failure but only ten of these trials reported on all-cause mortality. The amount of missing all-cause mortality patient data from the systematic review in these five trials was 41% (1125/2761). Treatment failure was defined in the review to be a composite end-point comprising one or more of the following: death, persistence, recurrence or worsening of clinical signs or symptoms of presenting infection. The inclusion of infection-related mortality indicates that all-cause mortality data had been collected in the five trials missing from the analysis. We attempted to contact the five trialists (via email) not reporting on all-cause mortality who were asked to confirm whether all-cause mortality was measured and analysed and, if so, to provide the reason for not reporting the results. Two trialists replied, one confirmed that the data were collected but not analysed because of the high risk patient population (patients in the study often died of unrelated causes) and the second trialist confirmed that the all-cause mortality rate in the study was low and the results were not significantly different between the treatment arms.

3. Methods

We now describe the univariate and bivariate fixed effect meta-analysis models, and their estimation.

3.1. Univariate fixed effect meta-analysis model

Consider there are $i = 1$ to n studies, and that there are two outcomes of interest for meta-analysis. From each study the treatment effect estimate for each outcome (say X_i for outcome 1, and Y_i for outcome 2) and their within-study variances (s^2_{Xi} and s^2_{Yi}) are sought. However, in practice studies may not report on both outcomes. The UFMA analyses each outcome separately, and assumes that the obtained estimates of the treatment effect from the i^{th} study are normally distributed about a common (fixed) true effect (β_X or β_Y) and variance (s^2_{Xi} or s^2_{Yi}), which are assumed to be known; that is

$$\begin{aligned} Y_i &\sim N\left(\beta_{Y_i}, s^2_{Y_i}\right), \\ X_i &\sim N\left(\beta_{X_i}, s^2_{X_i}\right). \end{aligned} \tag{1}$$

The model is estimated using maximum likelihood estimation, and produces a pooled treatment effect estimate that weighs each study by the inverse of its variance, for example

$$\hat{\beta}_Y = \frac{\sum_{i=1}^{n} \frac{Y_i}{s^2_{Yi}}}{\sum_{i=1}^{n} \frac{1}{s^2_{Yi}}}.$$

3.2. Bivariate fixed effect meta-analysis model

For a BFMA, both outcomes are considered simultaneously, as follows:

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_X \\ \beta_Y \end{pmatrix}, \begin{pmatrix} s^2_{X_i} & \rho_i s_{X_i} s_{Y_i} \\ \rho_i s_{X_i} s_{Y_i} & s^2_{Y_i} \end{pmatrix}\right), \tag{2}$$

where again β_X and β_Y denote the true treatment effects, assumed the same in each study, and s^2_{Xi} and s^2_{Yi} are the within-study variances, which are assumed to be known. Additionally, ρ_i are the within-study correlations, indicating the association between outcome estimates within a study, and these are also assumed known. Non-zero ρ_i typically arise because the same patients provide data for both outcomes, and this induces an association between the two outcome effect estimates. Note that model (2) can accommodate studies providing results for only one of the two outcomes.

Maximum likelihood can again be used to estimate β_X and β_Y and their variances. Riley *et al.* [4] derived the complex analytic solution for these pooled estimates, which shows that each pooled estimate takes account of the study estimates for *both* outcomes, and their variances and correlation. The utilisation of correlation, and thus the amount of 'borrowing of strength' of outcome Y data in the estimation of β_X, and vice versa, increases as the size of within-study correlations increases, and also as the difference in the within-study variances increases between studies [4]. This 'borrowing of strength' is thus particularly large when there are some missing outcomes in some studies, because this is equivalent to arbitrarily setting a within-study variance of infinity to missing outcomes in such studies. The bivariate fixed effect model (2) is estimated using maximum likelihood estimation, for example using the STATA multivariate meta-analysis module 'mvmeta' [12]. This approach can be easily extended to any number of outcomes as required, and to random effects, which we consider in Section 6. We refer to Riley *et al.* [4] where the formulae for the bivariate pooled effect estimates are shown.

3.3. *Estimating the within-study correlation (ρ)*

In model (2), the ρ_i are usually difficult to obtain, because they are rarely reported, or even calculated, for each study, and so an approximation is often required [9, 13]. As mentioned, non-zero within-study correlations arise because the same patients contribute to estimating the effects for each of the multiple outcomes in a study. In this section we discuss three possible ways of obtaining the correlation between these estimators, the third of which has not been considered previously.

3.3.1. Individual patient data. The availability of individual participant data (IPD) would allow us to calculate the within-study correlation between the estimators directly in each study [14]. One problem with this approach is that the IPD are unlikely to be available for all studies and therefore within-study correlations are still not fully available. One solution is to compute the within-study correlation from the IPD of a single study and then use this estimate as a 'likely' within-study correlation estimate for all other studies with unavailable within-study correlations. An average correlation could be used if the IPD is available for more than one study.

3.3.2. Biological reasoning (expert opinion). If no IPD data are available, it may still be possible to approximate the within-study correlation using biological reasoning or expert opinion. An expert in the clinical field could be asked to suggest a plausible within-study correlation between the estimators for use in all studies. For example, in the beta-lactam example, we would hypothesise that the relationship between all-cause mortality and treatment failure has a positive correlation (more deaths would imply more infection-related deaths, which would mean more treatment failures), although the strength of this correlation is more difficult to decipher.

3.3.3. Pearson correlation approach. In this paper we also propose and evaluate a new method, which we term the Pearson correlation method, for when the within-study correlation cannot be obtained from IPD or from expert opinion. Very simply, for a bivariate analysis, a common within-study correlation (ρ) is assumed in each study, and this is estimated by calculating the Pearson correlation between the pairs of available treatment effect estimates in those studies that provide data on both outcomes. This quantity, which we refer to as $\rho_{(\text{Pearson})}$, is then assumed to be the within-study correlation in each study; essentially one is assuming that the within-study correlations are the same in each study and that this correlation will be closely reflected in the observed correlation between paired outcome effects across studies. A previous study suggested a precise estimate of the Pearson correlation could be obtained with five data points [15]. In our example in Section 2, ten studies provide estimates for both outcomes.

4. Simulation study

We now present a simulation study to demonstrate the estimation properties from BFMA when compared with those from two separate UFMA in situations with complete data and nonignorable missing data according to an ORB mechanism.

4.1. Simulation methods

4.1.1. Defining the simulation scenarios. Our simulation study examined 12 scenarios, labelled (i) to (xii) (Table I). Scenarios (i) to (iv) consider complete data, whereas scenarios (v) to (xiii) consider the situation when some data are missing not at random in either a single outcome (scenarios (v) to (vii)) or both outcomes (scenarios (ix) to (xii)). The scenarios also vary in the size of the simulated within study correlation. For example, scenarios (i, v and xi) consider no within-study correlation, whereas scenarios (ii), (vi) and (x) consider high levels of within-study correlation. The number of studies (n) included in the meta-analysis took three different levels, n=5, 10 and 20. These values were chosen to reflect the range of randomised controlled trials included in a cohort of 283 systematic reviews that were assessed in the ORBIT study [5]. In each scenario, 1000 meta-analysis datasets were generated.

4.1.2. Generating the simulated data. For each simulation, two sets (one for outcome X and one for outcome Y) of within-study variances (of length n) were generated from a $0.25 \times \chi^2_1$distribution. Values outside the range [0.009, 0.6] were disregarded and a new value was generated. The two sets of values are then ranked so that the first study has the largest pair of simulated values (s^2_{Xi} and s^2_{Yi}) and so on, until the last study had the smallest pair of within-study variances. New within-study variances were simulated for every meta-analysis in the simulation study. This method for generating within-study variances has been used previously and has been shown to simulate a realistic mixture of study sample sizes [16, 17]. Analytically, it can be shown that if the numbers of patients in the treatment and control arm are the same, and the event rate is equal, this approach will generate sample sizes ranging from approximately 13 to 888. These parameters were therefore chosen as they were realistic choices from the sample sizes obtained in the ORBIT study (median 75; 5th percentile 17, 95th percentile 849). Pairs of X_i and Y_i were simulated directly from model (2); the true overall treatment effects β_X and β_Y were both taken to be zero for all simulation scenarios highlighted in Table I. Non-zero treatment effects were used for the assessment of power; our choices are described later in this section.

In Section 1 we introduced two key types of ORB mechanism. One where outcomes are not reported if they are negative (i.e. against treatment) and not statistically significant and the other where outcomes are not reported if they are negative (regardless of statistical significance). In our simulations we focus on the latter mechanism and consider two particular scenarios:

(a) The first is where missingness was generated by removing any negative treatment effect outcome data for outcome Y only (these are scenarios (v)–(viii)); this situation relates to where outcome one is a primary outcome, so is always reported, but outcome two is a secondary outcome and so only reported if it is positive.
(b) Missingness was generated by removing (in scenarios (ix)–(xii)) any negative values for both outcomes, X and$Y$. This is a more extreme ORB situation than (a), and relates to when perhaps both outcomes are secondary outcomes and only reported if they are positive.

To assess statistical power, non-zero treatment effects were also generated. Data sets were simulated in the same way as described above using 0.1 as a small treatment effect and 0.25 to denote a large treatment effect. Power was assessed using 5 and 10 studies included in the meta-analysis and three different sizes of within-study correlation; no correlation, a correlation of 0.4 and a correlation of 0.8. We used zero correlation because it has been shown that when the within-study correlation is zero, the 'borrowing of strength' does not occur and thus the multivariate meta-analysis approach has no benefit over two separate univariate analyses [4]. As the strength of the correlation increases, the benefits of using the multivariate approach are more profound. Therefore, a within-study correlation set to zero was used for comparison of when the multivariate meta-analysis approach is most beneficial. A correlation of 0.8 was used because this represented a strong correlation; we thought anything more extreme would be rarely observed in practice. The value 0.4 represents a moderate correlation.

Table I. Scenarios used in the simulations based on model 2.

	True effect estimates		True within-study correlation (ρ)	Number of studies (n)	Assumed within-study correlation	Description
Scenario	β_X	β_Y				
Complete data						
(i)	0	0	0	5,10,20	0	No missing study data, zero within- study correlation, treated as known.
(ii)	0	0	0.8	5,10,20	0.8	No missing data, high within-study correlation, treated as known.
(iii)	0	0	–0.8, 0, 0.4,0.6,0.7	5,10,20	0.8	No missing data, varied within-study correlation, treated as unknown, assumed to be high.
(iv)	0	0	–0.8, 0, 0.4,0.6,0.7	10,20	Pearson	No missing data, varied within-study correlation, estimated using Pearson correlation.
Missing not at random in outcome Y						
(v)	0	0	0	5,10,20	0	Nonignorable missing in endpoint Y, zero within-study correlation, treated as known.
(vi)	0	0	0.8	5,10,20	0.8	Nonignorable missing in endpoint Y, high within-study correlation, treated as known.
(vii)	0	0	–0.8, –0.8,0,0.4,0.6,0.7	5,10,20	0.8	Nonignorable missing in endpoint Y, varied within-study correlation, treated as unknown, assumed to be high.
(viii)	0	0	–0.8, 0, 0.4,0.6,0.7	10,20	Pearson	Nonignorable missing in endpoint Y, varied within-study correlation, estimated using Pearson correlation.
Missing not at random in both outcomes (X and Y)						
(ix)	0	0	0	5,10,20	0	Nonignorable missing in both endpoints (X and Y), zero within-study correlation, treated as known.
(x)	0	0	0.8	5,10,20	0.8	Nonignorable missing in both endpoints (X and Y), high within-study correlation, treated as known.
(xi)	0	0	–0.8, 0, 0.4,0.6,0.7	5,10,20	0.8	Nonignorable missing in both endpoints (X and Y), varied within-study correlation, treated as unknown, assumed to be high.
(xii)	0	0	–0.8, 0, 0.4,0.6,0.7	20	Pearson	Nonignorable missing in both endpoints (X and Y), varied within-study correlation, estimated using Pearson correlation.

4.1.3. Models fitted to each generated dataset. To each meta-analysis dataset generated in each simulation scenario, a range of different models were fitted. For each dataset generated, a UFMA was fitted to each outcome separately. A BFMA was then jointly fitted to both outcomes, first using the true within-study correlations and then again for a range of different assumed correlations; the latter was done to simulate the idea that the within-study correlations are often unknown [9], and therefore we wanted to gauge the performance of the BFMA estimates, the further away the assumed within-study correlations were from the truth.

Each of the six scenarios ((i), (ii), (v), (vi), (ix) and (x)) were fitted using the BFMA model assuming that the within-study correlation was known. In other words, the true within-study correlation (ρ) was the same as the fitted within-study correlation, which was assumed to be the same for all studies. The within-study correlations were also assumed known for all simulations assessing power. In scenarios (iii), (vii) and (xi), the fitted within-study correlation was assumed to be high and positive for each study in the meta-analysis when applying the BFMA approach. In scenarios (iv), (viii) and (xii), the within-study correlation for each study was estimated by calculating the Pearson correlation approach ($\rho_{(\text{Pearson})}$) between the pairs of available treatment effect estimates for the two outcomes. Because the Pearson correlation is best estimated given at least five pairs of estimates, this method was only considered when the number of studies included in the meta-analysis was either 10 or 20 in the case of complete data and missing outcome data in outcome Y, and 20 when there was missing data in outcomes X and Y.

4.1.4. Assessment of performance. We assessed performance of the estimates by examining β_X and β_Y estimates in terms of bias, standard error, mean square error (MSE) and coverage. For each of the scenarios, the 1000 BFMA estimates and the corresponding 1000 UFMA estimates were compared by calculating: (a) the average parameter estimates across all the simulations (to estimate bias), (b) the average standard error and MSE of β_X and β_Y (to assess precision) and (c) the coverage of the 95% confidence intervals (CIs) for β_X and β_Y. The coverage was calculated as the percentage of simulated datasets for which the 95% confidence interval for an outcome's treatment effect estimate contained the true effect estimate. Power was calculated as the percentage of simulated datasets for which the 95% confidence interval for an outcome's treatment effect estimate did not contain zero. Power was calculated for β_X and β_Y separately, and for β_X and β_Y jointly (i.e. the percentage of simulated datasets for which both the 95% confidence intervals for β_X and β_Y did not contain zero).

4.2. Simulation results

The results of the simulation studies are provided in a separate web appendix.[‡] Table A.1 provides the simulation results for scenarios with no missing data ((i)–(iv)), Table A.2 the results for scenarios with missing data in outcome Y only ((v)–(viii)) and Table A.3 the results for scenarios with missing data in both outcomes ((ix)–(xii)). As expected, when there is no within-study correlation, irrespective of the missing data mechanism (scenarios (i), (v), (ix)), all performance indicators for the BFMA and the UFMA approach were identical. For all other scenarios, the pooled estimates were more precise using the BFMA approach resulting in smaller standard errors and mean square errors, with largest gains observed for missing data situations. These findings were similar to those found in previous simulation studies of bivariate meta-analysis [8, 18]. We now discuss the new findings that arise from our simulations, in particular focusing on the impact of using the wrong correlations or the Pearson correlation, the power of BFMA versus UFMA, and the comparison of bias in the UFMA and BFMA pooled estimates given ORB. The assumed within-study correlation estimated from the data is provided in Tables A.1–A.3 when the Pearson approach was used.

4.2.1. Complete case scenarios. For all complete data scenarios (i)–(iv), the pooled estimates were approximately unbiased for both BFMA and UFMA (Table A.1). The coverage between the two approaches were comparable except when the within-study correlations are unknown (scenario (iii)). Fitting a BFMA but using an overestimated within-study correlation gives pooled results with a severely affected coverage. For example, in scenario (iii), the coverage for both outcomes can be up to 20% too low for the BFMA approach when the true within-study correlation is smaller than the fitted within-study correlation of 0.8.

[‡]*Supporting information may be found in the online version of this article.*

Applying the BFMA approach using the Pearson method (Table A.1, scenario (iv)) appeared to perform well. Any biases in the pooled estimates were small, the coverage was close to 95% and thus acceptable, and the standard errors were an improvement over the UFMA approach. In terms of power (Table A.4), an improvement was always observed in both parameters using the BFMA approach when the two outcomes were highly correlated ($\rho = 0.8$) compared with no correlation but there were only marginal gains when the correlation was weaker ($\rho = 0.4$). For example, when data were generated for 10 studies with a within-study correlation of 0.8 and true treatment effects of 0.25, the joint power achieved for outcome X and outcome Y using the BFMA approach was 93.2% compared with 77.8% when the correlation was ignored (UFMA approach). For the same scenario using a correlation of 0.4, the joint power achieved for outcome X and outcome Y using the BFMA approach was 79.1% compared with 75.8% when the correlation was ignored (UFMA approach).

4.2.2. ORB situation (a): Missing data in outcome Y when it is negative. For this series of scenarios, the findings are particularly relevant for the beta-lactam example, where there were missing data for one outcome and the within study-correlation was unknown.

When there is missingness in outcome Y only, and the within-study correlations are known (scenario (vi)), applying the BFMA approach leads to substantial reductions in the bias for outcome Y, which are complemented by much improved coverage albeit at the expense of a minor increase in the bias in outcome X (Table A.2). For example, in simulation scenario (vi) (20 studies), for outcome Y, the upward bias in the pooled estimate is 0.179 from the UFMA, but this is reduced to 0.057 in the BFMA. This is at the expense of a very slight increase in bias in the pooled estimate for outcome X, from –0.002 in the UFMA to –0.011 in the BFMA.

If the within-study correlations are unknown (Table A.2, scenario (vii)), applying the BFMA approach can lead to poorer coverage for both outcomes when the assumed correlation is incorrect, and especially when it is far from the truth. For example, consider the situation where there were 10 studies and the true within-correlation was 0.4 (Table A.2). Here, when wrongly assuming a within-study correlation of 0.8, the BFMA approach had poorer coverage for both outcome X (87%) and outcome Y (68%) compared with the UFMA approach where the coverage for outcome X and Y was 95% and 70%, respectively.

Greater bias can also be observed in both outcomes if the estimated within-study correlation is in the wrong direction (i.e. specified as positive when it is truly negative). However, if the estimated within-study correlation is quite close to the true within-study correlation then the advantages of applying the BFMA approach for outcome Y (in terms of coverage and bias) are clear. For example, consider 10 studies and a true within-study correlation of 0.7 (Table A.2, scenario (vii)); because the fitted within-study correlation of 0.8 is close to the truth, the bias in outcome Y is 0.199 in the UFMA but 0.078 in the BFMA, a reduction in bias of over 60%, and the coverage for outcomes X and Y is 93% and 86% in the BFMA and 96% and 73% in the UFMA, respectively.

Using the Pearson correlation approach (Table A.2, scenario (viii)); the BFMA approach performs well in terms of bias reduction in outcome Y and more importantly improves coverage when compared with the UFMA approach. Considering the results from 20 studies, with a true within-study correlation of 0.7, the bias in the pooled estimate for outcome Y is 0.179 for the UFMA approach compared with 0.136 for the BFMA approach using the Pearson approach, a reduction in bias of nearly 25%. Most notably, the coverage for outcome Y is as high as 64% for the BFMA using the Pearson correlation approach, compared with only 36% for the UFMA approach. For outcome X, which has complete data, the coverage is close to 95% and there is little bias in the pooled estimate.

A review of the power analysis (Table A.5) for the joint power between outcome X and Y revealed that the power was improved using the BFMA approach when the correlation was high ($\rho = 0.8$), but the results were marginally worse when the correlations were moderate ($\rho = 0.4$). However, as a trade-off, there were clear advantages in terms of bias reduction and coverage for the BFMA approach when the correlation was moderate ($\rho = 0.4$). For example, for 10 studies where the true treatment effect estimates of both outcomes are set to 0.1, the power for outcome Y (the outcome with missing data) is 59% for BFMA but 65% for UFMA. However, for outcome Y the coverage is 89% and the bias is 0.099 for BFMA compared with 85% (coverage) and 0.120 (bias) for UFMA, an 18% reduction in bias. Thus, the slight gain in power for UFMA is simply a consequence of its upward bias and low coverage for this outcome.

4.2.3. Missing data in both outcomes (X and Y). When there is nonignorable missing data for both outcomes, the simulation results show that, irrespective of whether the within-study correlations are known,

the BFMA approach outperforms the UFMA approach in terms of bias reduction in both outcome X and outcome Y (Table A.4). There is also a tendency for the BFMA approach to have improved coverage, particularly when the estimated within-study correlation is close to the truth; this improvement is also apparent for the Pearson correlation approach (scenario (xii)). For example, in scenario (xi) where there are 10 studies and the true within-study correlation is set to 0.7, there is over a 30% reduction in bias in both outcomes and an improvement in coverage of 8% and 10% for outcome X and Y, respectively. However, the coverage for both outcomes can be slightly worse for the BFMA approach when the true within-study correlation is largely overestimated; this was particularly evident with fewer studies included in the meta-analysis.

The BFMA results in scenario (xii) (Pearson approach) would always be preferred to the UFMA approach (see UFMA results in scenario (xi) for 20 studies), because the bias and coverage are marginally improved. However, an inspection of the estimated within-study correlations using the Pearson approach were largely underestimated compared with the true-within study correlations in the presence of missing data in both outcomes. This is the likely result of having fewer studies reporting both outcomes to estimate the within-study correlation than say when there is no missing data or missing data in just one outcome.

An assessment of power (Table A.6) showed that the BFMA was marginally outperformed by UFMA for both outcome X and outcome Y, especially when the treatment effects differences were small. However, this was offset against reduction in biases, standard errors, and MSEs, and improved coverage for both outcomes. For example, for five studies, when the true within-study correlation was set to 0.4 and the treatment effect estimate was small (0.1) for both outcomes, the UFMA approach produces a power of 33% for outcome X, 33% for outcome Y, and 12% jointly for both outcomes. This compared with 32%, 30% and 12% for outcome X, Y and jointly for outcome X and Y, respectively, for the BFMA approach assuming the correct correlations. Again this is due to the larger upward bias from UFMA and lower coverage. For this particular scenario there is a reduction in bias of above 10% for both outcomes for the BFMA approach compared with the UFMA approach.

5. Application to the review of beta-lactam

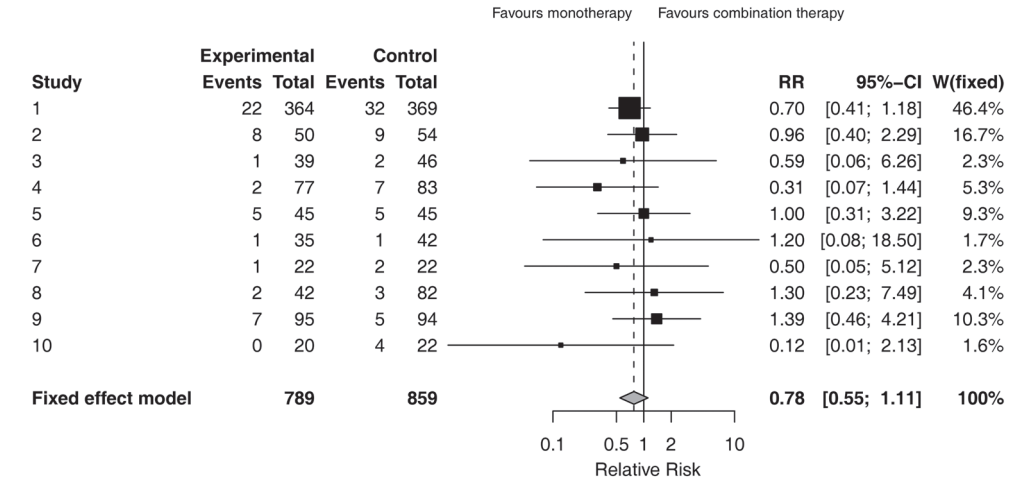
We now apply UFMA model (1) and BFMA model (2) to the beta-lactam example introduced in Section 2. Using the two-by-two table available for each study, we estimated the lnRR and its standard error for both outcomes, with a lnRR > 1 indicating that combination therapy was preferred. These estimates were then utilised in models 1 and 2, with outcome X relating to all-cause mortality and outcome Y relating to treatment failure. For study 10, because of the zero events for one group, 0.5 was added to all cells in the table for this outcome to allow the lnRR and its standard error to be calculated.

On the basis of, the univariate analyses, the I^2 statistic was 0% for the mortality outcome, and 18% for the treatment failure outcome, indicating that the impact of any heterogeneity in treatment effect is likely to be small. We thus considered a fixed effect meta-analysis would be appropriate for each outcome.

Using the UFMA approach, the pooled relative risk estimate (RR) for the 10 studies reporting on all-cause mortality was 0.78 (95% confidence interval [0.55, 1.11]), favouring monotherapy, while the RR for treatment failure based on data from all 15 eligible studies was 1.09 (95% confidence interval [1.01, 1.18]), favouring combination therapy (Figures 1(a) and (b)). Note that the univariate pooled estimates and their CIs are very similar if, rather than using model (1), the Mantel–Haenszel fixed effect approach is preferred for each outcome [19]. Figure 1(b) shows the results for treatment failure, subgrouped by whether both or only one outcome was reported. There is a marked difference between the subgroups in Figure 1(b), raising the suspicion that some of the trials reporting only treatment failure did so because the results were statistically significant whereas the all-cause mortality results were not and were suppressed.

To help reduce this potential ORB problem, the BFMA model (2) was considered. Neither the within-study correlation nor the IPD were available for all studies, so we calculated the Pearson correlation between the ln(relative risk) estimates from the 10 studies that reported both outcomes ($\rho_{(\text{Pearson})} = -0.043$). Although the strength of this correlation was weak, the direction of the correlation was still clinically unexpected (Mical Paul (review author), personal communication), and so we suspected it may be due to chance given the few data points to estimate the correlation coefficient. Data for both outcomes from the same review were also available from a further 28 trials using a different beta-lactam therapy. In the original analysis contained within the systematic review, trials looking at 'same' beta-lactam therapy were separated out from studies looking at 'different' beta-lactam therapies

a



b

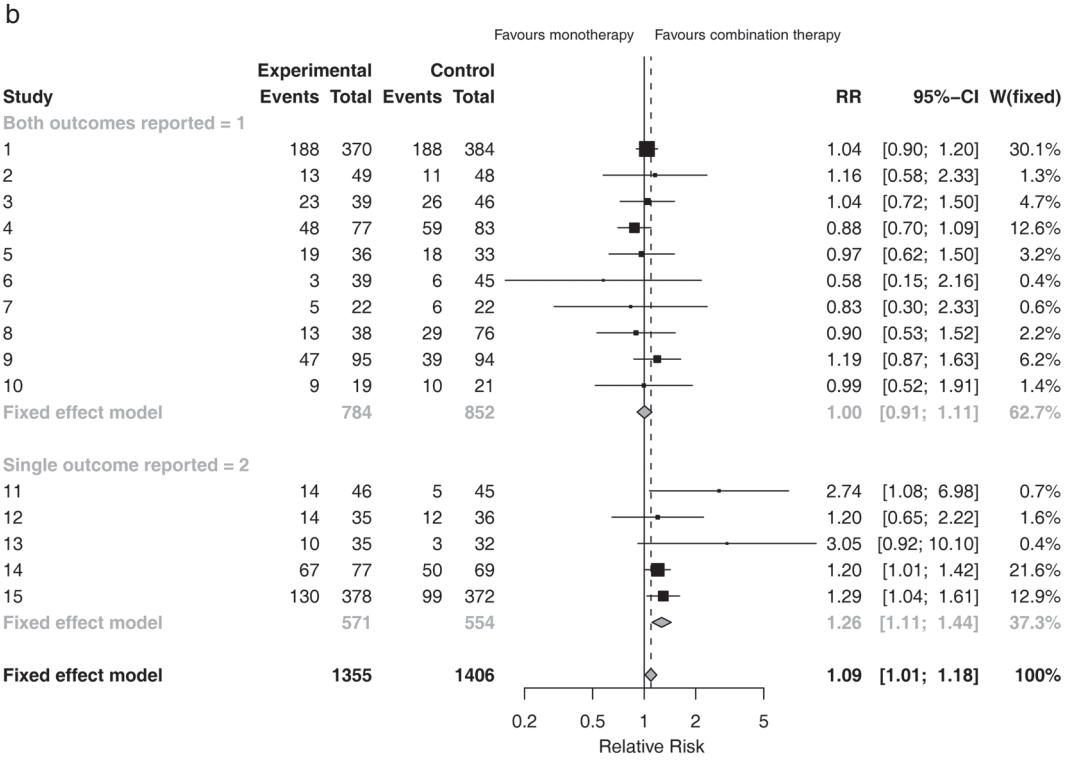


Figure 1. (a) Forest plot of the results for all-cause mortality for the review of beta-lactam therapy in cancer patients with neutropenia. (b) Forest plot of the results for treatment failure for the review of beta-lactam therapy in cancer patients with neutropenia, subgrouped by whether both outcomes (treatment failure and mortality) or only one outcome (treatment failure) was reported.

because this was expected to be a source of heterogeneity. Apart from this issue, we would not expect the correlation between mortality and treatment failure to differ between same and different beta-lactam therapies (Mical Paul, personal communication). Therefore, we additionally used all this data (same and different beta-lactam therapy) to try and obtain a better estimate of the correlation between these two outcomes (as more data were available to estimate this correlation). The Pearson correlation between the ln(relative risk) estimates from the larger combined set of 38 studies was computed as $\rho_{(\text{Pearson})} = 0.213$.

A BFMA was then conducted assuming a value of –0.043 for the within-study correlations, and then again assuming a value of 0.213. Also, using these values, a wider range of positive within-study correlations (0.2, 0.4, 0.6, 0.7 or 0.8) were also used to assess the impact of the assumed strength of correlation in a sensitivity analysis.

The results of each BFMA are shown in Table II. As the strength of the assumed within-study correlation increases, there is a reduction in the standard error of the pooled estimates for both outcomes, leading to increased statistical precision. Of more clinical importance, there is a shift in the pooled estimate for all-cause mortality in the BFMA compared with the UFMA. For example, when the correlation is assumed to be 0.213, the pooled relative risk is 0.82 in the BFMA compared with 0.78 in the UFMA. The UFMA estimate indicates large benefit for monotherapy (22% reduced risk), but the BFMA indicates a smaller difference between the treatment groups (18% reduced risk). As a consequence, the p-value is much larger in the BFMA (p=0.271) than the UFMA ($p = 0.168$), although UFMA and BFMA are both not statistically significant. The shift in pooled estimate, and gain in precision, is due to the all-cause mortality outcome 'borrowing of strength' from the treatment failure outcome, especially in those five trials for which all-cause mortality was not available.

The difference in BFMA and UFMA results for treatment failure is less dramatic, because there was complete data for this outcome. There is little change in the pooled relative risk estimate indicating combination therapy is more effective, and this remains statistically significant regardless of the correlation assumed. There is a small reduction in the standard error as the correlation increases, which leads to a slightly smaller p-value.

The simulations in Section 4 showed that the Pearson correlation approach generally underestimates the true within-study correlations, and so considering correlations > 0.231 (the estimated Pearson correlation) is also important here. The BFMA conclusions are consistent regardless of the size of the positive correlations assumed in the BFMA, although the borrowing of strength for all-cause mortality (in terms of the shift in pooled estimate towards one, reduction in standard error, and increase in p-value) is more dramatic the larger the correlation. For example, assuming a correlation of 0.6, the BFMA pooled estimate for all-cause mortality is 0.91 (p= 0.561), compared with 0.78 ($p = 0.168$) in the UFMA.

6. Extension to bivariate random effects meta-analysis

6.1. *Extensions to account for between-study heterogeneity*

We now extend our ideas to allow for between-study heterogeneity in one or more of the outcomes of interest.

6.2. *Possible model specifications*

The general bivariate random effects meta-analysis model [7] can be written as:

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} b_{Xi} \\ b_{Yi} \end{pmatrix}, \begin{pmatrix} s^2_{X_i} & \rho_{Wi} s_{X_i} s_{Y_i} \\ \rho_{Wi} s_{X_i} s_{Y_i} & s^2_{Y_i} \end{pmatrix}\right)$$
$$\begin{pmatrix} b_{Xi} \\ b_{Yi} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_X \\ \beta_Y \end{pmatrix}, \begin{pmatrix} \tau^2_X & \rho_B \tau_X \tau_Y \\ \rho_B \tau_X \tau_Y & \tau^2_Y \end{pmatrix}\right). \quad (3)$$

In this model the within-study covariance matrices are as in model (2), but now the true treatment effects in each study (b_{Xi} and b_{Yi}) are assumed drawn from a bivariate normal distribution, with mean values β_X and β_Y and between-study variances τ^2_X and τ^2_Y, for outcomes X and Y, respectively. In addition to the within-study correlations (ρ_{Wi}), there is now also a between-study correlation (ρ_{B}) that accounts for any association across studies in the true treatment effects. For example, consider a meta-analysis of studies where the treatment effect on overall survival (time to death) and disease-free survival (time to recurrence or death) are of interest. The within-study correlations arise because the same patients contribute to estimating the effects for both outcomes, and individual time to death is highly correlated with individual recurrence time. The between-study correlation arises because, for example, if a study has a treatment effect for overall survival that is higher than the average (β_X) then the treatment effect for disease-free survival is also likely to be higher than the average (β_Y), because the effect on recurrence has a strong association with the effect on death. Model (3) reduces to two separate univariate random-effects meta-analyses when $\rho_{\mathrm{Wi}} = \rho_{\mathrm{B}} = 0$.

Model (3) has been shown to improve estimation of β_X and β_Y compared with separate univariate analyses [4, 9], and thus can potentially reduce selective outcome reporting in a similar way as we have shown for the fixed effect approach. However, in addition to the problem of unavailable within-study correlations, model (3) also has the difficulty of estimating the between-study correlation. Riley *et al.* [18] showed that when the number of studies is small and/or the within-study variation is large relative

Table II. UFMA and BFMA results for the beta-lactam dataset.

		All-cause mortality (β_X)				Treatment failure (β_Y)			
Model	ρ	Pooled relative risk†	Standard error (Ln RR)	P-value	95% Confidence interval	Pooled relative risk †	Standard error (Ln RR)	P-value	95% Confidence interval
UFMA	0	0.778	0.182	0.168	(0.545, 1.111)	1.095	0.040	0.024	(1.012, 1.184)
BFMA		0.771	0.182	0.152	(0.540, 1.100)	1.094	0.040	0.025	(1.011, 1.183)
	0.042*								
BFMA	0.2	0.818	0.180	0.264	(0.575, 1.164)	1.097	0.040	0.020	(1.015, 1.187)
BFMA	0.213**	0.820	0.180	0.271	(0.577, 1.167)	1.098	0.040	0.020	(1.015, 1.187)
BFMA	0.4	0.861	0.175	0.393	(0.611, 1.213)	1.099	0.040	0.017	(1.017, 1.189)
BFMA	0.6	0.909	0.165	0.561	(0.657, 1.256)	1.101	0.039	0.014	(1.020, 1.189)
BFMA	0.7	0.933	0.158	0.662	(0.685, 1.271)	1.101	0.038	0.012	(1.021, 1.188)
BFMA	0.8	0.959	0.147	0.775	(0.718, 1.280)	1.102	0.037	0.010	(1.024, 1.185)

*Calculated as $\rho_{(\text{Pearson})}$ for the ten trials reporting treatment effect data for both outcomes (using same beta-lactam data)

**Calculated as $\rho_{(\text{Pearson})}$ for the 38 trials reporting treatment effect data for both outcomes (using trials of the same beta-lactam in each arm and trials of different beta-lactam data in each arm)

†A relative risk < 1 favours beta-lactam monotherapy, while a relative risk > 1 favours combination beta-lactam therapy.

to the between-study variation, the between-study correlation is often poorly estimated as +1 or –1. A suggestion to limit both these problems is a bivariate meta-analysis [13] that rather models the 'global' correlation (ρ_G), an amalgam of the within-study and between-study correlations, as follows:

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_X \\ \beta_Y \end{pmatrix}, \begin{pmatrix} s_{X_i}^2 + \tau_X^2 & \rho_G(s_{X_i}s_{Y_i} + \tau_X\tau_Y) \\ \rho_G(s_{X_i}s_{Y_i} + \tau_X\tau_Y) & s_{Y_i}^2 + \tau_Y^2 \end{pmatrix}\right). \quad (4)$$

Model (4) is an approximation to model (3), and conveniently does not require the within-study correlations to be specified nor the between-study correlation to be estimated. Simulation studies show that this model produces unbiased estimates and suitable coverage for β_X and β_Y, as long as ρ_G is not estimated very close to –1 or +1, as then convergence problems arise [13].

The global correlation ρ_G can be viewed as the correlation between the observed X_i and Y_i across those studies providing both, weighted according to the overall variability for each study. Because it uses the observed correlation, it is similar to our concept of using the (albeit unweighted) Pearson correlation to approximate the within-study correlations in a bivariate fixed effect meta-analysis. Thus, we suggest that if convergence of model (4) is difficult and/or ρ_G is poorly estimated close to +1 or –1 (e.g. $\hat{\rho}_G > 0.95$ or < -0.95) [13], then ρ_G could similarly be set to the Pearson correlation to aid estimation.

6.3. An applied example – mutant p53 tumour suppressor gene

Jackson *et al.* [8] considered a bivariate random-effects meta-analysis of six observational studies assessing whether the presence of mutant p53 tumour suppressor gene is a prognostic factor for overall and disease-free survival in patients presenting with squamous cell carcinoma arising from the oropharynx cavity. A random effects model was chosen *a priori* because heterogeneity is usually apparent in meta-analyses of observational studies. All studies provided unadjusted log hazard ratio estimates, comparing hazard rates for mutant p53 patients relative to normal p53 patients, and their standard errors for overall survival, but only three also gave results for disease-free survival (Table III).

Univariate random-effects meta-analysis gives a nonsignificant pooled hazard ratio for overall survival (Table IV). However, the pooled hazard ratio for disease-free survival is 0.45 (95% CI = 0.27 to 0.74), implying that patients with mutant p53 have a decreased hazard of recurrence or death. However, selective outcome reporting is suspected, because the log hazard ratio estimates are all *negative* in the three studies reporting both outcomes, but are all *positive* in those studies reporting only overall survival. Because overall and disease-free survival results are likely to be positively correlated, we are concerned that disease-free survival is only being reported when it suggests mutant p53 is favourable.

A bivariate meta-analysis can utilise the correlation between outcomes to reduce this problem. However, the within-study correlations are unknown in this example but expected to be highly positively correlated because of the aforementioned association between time to recurrence and time to death. Fitting model (3) using restricted maximum likelihood [7] and assuming the within-study correlations are equal to 0.7 gives a pooled hazard ratio of 0.72 (95% CI: 0.32 to 1.66) (Table IV). Compared with the univariate analyses, the pooled result is now closer to 1 and the confidence interval includes this null value. However, the between-study correlation is poorly estimated as +1 in this analysis, and this extreme value may be strongly influencing the results and amount of borrowing strength here [8].

When alternatively fitting model (4) using restricted maximum likelihood [12, 13], the global correlation is also difficult to estimate and the model fails to converge. However, the Pearson correlation is more sensibly estimated as 0.69 across those three studies providing both outcomes. We thus fitted model (4) fixing the global correlation to be 0.69, and the model now converged (Table IV). The estimated pooled

Table III. p53 tumour suppressor gene data.

Study	Ln hazard ratio (standard error) for overall survival	Ln hazard ratio (standard error) for disease free survival
1	–0.18 (0.56)	–0.58 (0.56)
2	0.79 (0.24)	—
3	0.21 (0.66)	—
4	–0.63 (0.29)	–1.02 (0.39)
5	1.01 (0.48)	—
6	–0.64 (0.4)	–0.69 (0.4)

Table IV. Results for the analyses of the p53 tumour suppressor gene dataset.

Analysis method	Pooled hazard ratio (95% CI) for overall survival	Between-study standard deviation for overall survival	Pooled hazard ratio (95% CI) for disease free survival	Between-study standard deviation for disease-free survival	Between-study correlation
Univariate random-effects	1.09 [0.59, 2.02]	0.64	0.45 [0.27, 0.73]	0	—
Bivariate random-effects model (3) assuming within-study correlations = 0.7	1.10 [0.60, 2.02]	0.63	0.72 [0.32, 1.66]	0.46	1
Global correlation model (4)	*	*	*	*	*
Global correlation model (4) assuming global correlation = Pearson correlation = 0.67	1.09 [0.60, 1.99]	0.62	0.57 [0.36, 0.90]	0	—

*Not estimable.

hazard ratio for disease-free survival was closer to one than the univariate meta-analysis suggested (0.57 compared with 0.45) and the confidence interval was also wider, but still did not include the null value of one. Thus, after accounting for the correlation between outcomes to reduce the selective reporting problem, there remains evidence that p53 is a potential prognostic factor for disease-free survival, warranting further research.

7. Discussion

There has been little consideration of how to adjust for ORB in a meta-analysis. One of the challenges is that the underlying ORB mechanism or reason for the missing outcome data is not always known. ORB is an example of a nonignorable missing data mechanism because failure to report outcome data is dependent on the outcome value that would have otherwise been observed.

Our work has examined through simulation the benefits of BFMA over UFMA for estimating pooled (treatment) effects for multiple outcomes in both complete and missing data situations. The UFMA results show that ORB can substantially bias pooled estimates in favour of a treatment being effective, even when it is not. The results from the BFMA approach were encouraging and it was found that the 'borrowing of strength' can reduce the impact of ORB in a meta-analysis. The improvements in terms of the shift in the pooled estimates and thus a reduction in bias (and improved coverage) were found to be more profound as the correlation between the two outcomes increased. This was illustrated in the beta-lactam example, where the BFMA pooled estimate was shifted closer to the null value than the UFMA pooled estimate, and its p-value was substantially higher.

Our simulations showed that when nonignorable missing data were introduced in just one outcome, the BFMA approach introduces a small but noticeable bias in the other outcome. However, we consider the large improvement for the outcome with missing data to be worth the 'trade-off' here; that is, in ORB situations we would accept the slight increased bias for the outcome with no missing data to gain the large benefit reduction in bias in the outcome with missing data when deploying the BFMA approach. Similar trade-offs were found in the analysis of power, particularly when there was missingness in both outcomes and the treatment difference was small. Slight reductions in power were offset against the benefits in bias reduction and improved coverage.

The missing data mechanisms used in the simulations were more severe than perhaps we would expect in practice. In reality, we would not expect all trial authors to suppress negative outcome data. In the beta-lactam example, 5 out of 15 studies had missing outcome data in one outcome. For this reason, we would see our simulation results as a worst case scenario. We also considered a modification of our simulations to generate missingness with the additional criteria that p must be > 0.05 for the outcome to be missing (and it being negative). Our findings were very similar to those shown in the web Appendix, so we do not present these additional results here.

One of the main challenges of the multivariate approach is knowing how to obtain the within-study correlations to measure the dependence between the outcomes for each study when these are unknown [9]. In this article we proposed the Pearson correlation approach to estimate this quantity to be the assumed within-study correlation across all studies. Because the correlation is estimated across studies, it may thus be affected by study-level characteristics that may lead to differences to the true within-study correlation. The approach also assumes that the within-study correlations are the same in each study. Furthermore, if there is heterogeneity in true outcome effects (i.e. when a fixed effect model is not appropriate), then the correlation observed across studies between outcome estimates will be a mixture of within-study correlation and between-study correlation, which Riley *et al.* termed the 'global' correlation [13], as utilised in model (4). We consider that this is a convenient approach when no other suitable within-study correlation is available, and the simulations show this method performed well for a BFMA (generally better than UFMA), although it underestimates the true correlation and so does not perform as well as a BFMA utilising the correct correlations. In our extension to bivariate random-effects meta-analysis (Section 6), the use of the Pearson correlation to approximate the global correlation also helped resolve estimation problems.

The reliability of systematic reviews can be improved if more attention is paid to outcome data missing from the source trial reports. If data are missing, reviewers should be encouraged to contact the trialists to confirm whether the outcome was measured and analysed and, if so, obtain the results. If this approach is not feasible or successful, as often is the case, then rather than do nothing, review authors are encouraged to apply a sensitivity analysis to assess the impact of outcome reporting bias on an individual review. If the results are not robust to outcome reporting bias, the review conclusions may need

to be adjusted. The multivariate meta-analysis approach offers one such sensitivity analysis to adjust for outcome reporting bias when there is missing trial data for many review outcomes. Our recommendation to reviewers would be to use the multivariate meta-analysis approach if one is reasonably confident about the correlation estimates between outcomes (either from actual data, IPD from a single study or the Pearson estimate) or use an alternative univariate adjustment approach, for example the bound for maximum bias [6] if one is not confident about the correlations between outcomes. Where the multivariate approach is desirable but estimates of correlation are imprecise or clinically unexpected (as the negative correlation was in the beta-lactam example), one can consider clinical or biological reasoning to inform the correlation, or consider sensitivity analyses over a range of sensible values.

A clear limitation of our simulation work is that we have only considered fixed effect situations, which was appropriate for the beta-lactam example. Previous work has shown that multivariate random effects meta-analysis models also borrow strength [7, 8], and so they should also reduce ORB problems in a similar manner. This was demonstrated in the mutant p53 example, where heterogeneity was accounted for and the correlation attenuated the pooled effect for disease-free survival. In other examples ORB, like publication bias, may also hide the true extent of heterogeneity, and so multivariate random-effects models may shift the pooled estimates and reveal additional heterogeneity in the presence of ORB. Random-effects models also require a decision on which estimation method to use. For example, in Section 6 we used a restricted maximum likelihood approach as the estimation method but there are other methods available such as maximum likelihood and the method of moments [16], which has been shown to influence the results in some situations [8]. Furthermore, and especially as the number of outcomes increases, the random effects model makes an implicit assumption that the true outcome effects have a linear relationship between studies, which is hard to verify but this may clearly influence the borrowing of strength [8]. Thus, consideration of multivariate meta-analysis methods given ORB and heterogeneity warrants further research, as does the assessment of publication bias within multivariate meta-analysis [20, 21]. One final limitation of our simulations is that by generating the aggregate data we make the assumption that the within-study correlations are the same in each meta-analysis dataset. Other simulations [13, 18, 22, 23] in multivariate meta-analysis have also generated aggregate data and assumed common within-study correlations, which may be plausible, or approximately close, in some situations, but clearly not in others. Further research using simulations at the IPD level and allowing different within-study correlations would thus be informative.

Acknowledgements

We would like to thank Mical Paul for helpful discussions regarding the beta-lactam review. Richard Riley was supported by funding from the MRC Midlands Hub for Trials Methodology Research.

References

1. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, Decullier E, Easterbrook PJ, Von Elm E, Gamble C, Ghersi D, Ioannidis JP, Simes J, Williamson PR. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE* 2008; **3**(8):e3081. DOI: 10.1371/journal.pone.0003081.
2. Dwan K, Altman DG, Cresswell L, Blundell M, Gamble CL, Williamson PR. Comparison of protocols and registry entries to published reports for randomised controlled trials. *Cochrane Database of Systematic Reviews* 2011. DOI: 10.1002/14651858.MR000031.pub2. Issue 1: Art. No.: MR000031.
3. Smyth R, Kirkham JJ, Jacoby A, Altman DG, Gamble C, Williamson PR. Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. *British Medical Journal (Research)* 2011; **342**:c7153. DOI: 10.1136/bmj.c7153.
4. Riley RD, Abrams KR, Lambert PC, Sutton AJ. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**(1):78–97. DOI: 10.1002/sim.2524.
5. Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR. The impact of outcome reporting bias on systematic reviews. *British Medical Journal (Research)* 2010; **340**:c365. DOI: 10.1136/bmj.c365.
6. Williamson PR, Gamble C. Application and investigation of a bound for outcome reporting bias. *Trials* 2007; **8**(9). DOI: 10.1186/1745-6215-8-9.
7. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**(4):589–624. DOI: 10.1002/sim.1040.
8. Jackson D, Riley RD, White I. Multivariate meta-analysis: potential and promise. *Statistics in Medicine* 2011; **30**(20):2481–2498. DOI: 10.1002/sim.4172.
9. Riley RD. Multivariate meta-analysis: the effect of ignoring within study correlation. *Journal of the Royal Statistical Society (Series A)* 2009; **172**(4):789–811. DOI: 10.1111/j.1467-985X.2008.00593.x.

10. Williamson PR, Gamble C. Identification and impact of outcome selection bias in meta-analysis. *Statistics in Medicine* 2005; **24**(10):1547–1561. DOI: 10.1002/sim.2025.
11. Paul M, Schlesinger A, Grozinsky-Glasberg S, Soares-Weiser K, Leibovici L. Beta-lactam versus beta-lactam-aminoglycoside combination therapy in cancer patients with neutropenia. *Cochrane Database of Systematic Reviews* 2003. DOI: 10.1002/14651858.CD003038. Issue 3: Art. No.: CD003038.
12. White IR. Multivariate random-effects meta-analysis. *Stata Journal* 2009; **9**(1):40–56.
13. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* 2008; **9**(1):172–186. DOI: 10.1093/biostatistics/kxm023.
14. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine* 2008; **27**(11):1870–1893. DOI: 10.1002/sim.3165.
15. Abdel-Megeed SM. (1984). Accuracy of Correlation Coefficient with Limited Number of Points. *The Journal of Experimental Education* 1984; **52**(4):188–191.
16. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine* 2010; **29**(12):1282–1297. DOI: 10.1002/sim.3602.
17. Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. *Statistics in Medicine* 2007; **26**(25):4531–4543. DOI: 10.1002/sim.2883.
18. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3. DOI: 10.1186/1471-2288-7-3.
19. Mantel N, Haenszel W. Statistical aspects of the analysis of data from the retrospective analysis of disease. *Journal of the National Cancer Institute* 1959; **22**(4):719–748.
20. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L, Moreno SG. Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *Journal of the Royal Statistical Society (Series A)* 2010; **173**(3):575–91. DOI: 10.1111/j.1467-985X.2009.00629.x.
21. 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):911–24. DOI: 10.1016/j.jclinepi.2004.01.018.
22. 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**(22):2537–2530.
23. Sohn SY. Multivariate meta analysis with potentially correlated marketing study results. *Naval Research Logistics* 2000; **47**(6):500–510.