Research Article

Statistics in Medicine

Received 24 July 2014, Accepted 11 March 2015 Published online 29 April 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6493

# Multivariate meta-analysis of prognostic factor studies with multiple cut-points and/or methods of measurement

**Richard D. Riley,[a*†] Eleni G. Elia,[b] Gemma Malin,[c] Karla Hemming[b] and Malcolm P. Price[b]**

**A prognostic factor is any measure that is associated with the risk of future health outcomes in those with existing disease. Often, the prognostic ability of a factor is evaluated in multiple studies. However, meta-analysis is difficult because primary studies often use different methods of measurement and/or different cut-points to dichotomise continuous factors into 'high' and 'low' groups; selective reporting is also common. We illustrate how multivariate random effects meta-analysis models can accommodate multiple prognostic effect estimates from the same study, relating to multiple cut-points and/or methods of measurement. The models account for within-study and between-study correlations, which utilises more information and reduces the impact of unreported cut-points and/or measurement methods in some studies. The applicability of the approach is improved with individual participant data and by assuming a functional relationship between prognostic effect and cut-point to reduce the number of unknown parameters. The models provide important inferential results for each cut-point and method of measurement, including the summary prognostic effect, the between-study variance and a 95% prediction interval for the prognostic effect in new populations. Two applications are presented. The first reveals that, in a multivariate meta-analysis using published results, the Apgar score is prognostic of neonatal mortality but effect sizes are smaller at most cut-points than previously thought. In the second, a multivariate meta-analysis of two methods of measurement provides weak evidence that microvessel density is prognostic of mortality in lung cancer, even when individual participant data are available so that a continuous prognostic trend is examined (rather than cut-points).** ***Statistics in Medicine*** **Published by John Wiley & Sons Ltd.**

**Keywords:** multivariate meta-analysis; prognostic factors; odds ratios and hazard ratios; cut-points; heterogeneity

## 1. Introduction

A prognostic factor is any measure that, among people with a given health condition, is associated with a subsequent clinical outcome [1,2]. For example, in many cancers, tumour grade at the time of histological diagnosis is a prognostic factor because it is associated with time to disease recurrence or death; those with a higher tumour grade have a worse prognosis. Prognostic factors thus distinguish groups of people with a different average prognosis, and this allows them to be useful for clinical practice and health research. For example, they can help define disease at diagnosis, inform clinical and therapeutic decisions (either directly or as part of multivariable prognostic models), enhance the design and analysis of intervention trials and observational studies (as they are potential confounders) and may even identify targets for new interventions that aim to modify the course of a disease or health condition.

Given their importance, there are often hundreds of studies each year investigating the prognostic value of one or more bespoke factors in each disease field. However, there is often inconsistency in

[a]*Research Institute of Primary Care and Health Sciences, Keele University, Staffordshire, ST5 5BG, U.K.*
[b]*School of Health and Population Sciences, Public Health Building, University of Birmingham, Edgbaston, Birmingham, B15 2TT, U.K.*
[c]*School of Medicine, D Floor Queen's Medical Centre, University of Nottingham, Nottingham, NG7 2UH, U.K.*
[*]*Correspondence to: Richard D. Riley, Research Institute of Primary Care and Health Sciences, Keele University, Staffordshire, ST5 5BG, U.K.*
[†]*E-mail: r.riley@keele.ac.uk*




 

their findings, with some suggesting a particular factor is prognostic and others suggesting the opposite [1, 3, 4]. Meta-analysis is therefore needed to synthesise study findings and summarise the prognostic value of each factor of interest [5]. Unfortunately, this is often problematic as primary studies are prone to poor and selective reporting [6–8] and heterogeneity in, for example, their study populations and type of statistical results [9]. Meta-analyses of prognostic factor studies thus often conclude without strong recommendations [1, 3]. The following is a typical example [10]: 'After 10 years of research, evidence is not sufficient to conclude whether changes in P53 act as markers of outcome in patients with bladder cancer …. That a decade of research on P53 and bladder cancer has not placed us in a better position to draw conclusions relevant to the clinical management of patients is frustrating.'

Two common problems for meta-analysis of a particular factor are between-study differences in its method of measurement and, for continuous factors, the cut-point value used to define 'high' and 'low' (or abnormal and normal) groups. For example, de Azambuja *et al.* [11] perform a meta-analysis of the prognostic ability of Ki-67 in patients with breast cancer and pool 38 unadjusted hazard ratios across studies; however, these related to 20 different cut-points and five different methods of measurement. When pooling such studies, the summary meta-analysis results are difficult to interpret clinically, as they do not relate to a single cut-point or measurement method. Even if studies do report results for multiple cut-points or methods of measurement, meta-analysts usually just take one cut-point and one method of measurement per study and thus lose information about the others. This may be because multiple study results for each cut-point and method of measurement are correlated, and therefore, more advanced statistical methods are necessary to account for this if they are all used in the meta-analysis [12].

In this article, we suggest approaches to meta-analysis of prognostic factor studies when faced with multiple cut-points and/or methods of measurement and missing results in some studies. Firstly, in Section 2, we consider methods for situations where each study provides a single prognostic result for a particular cut-point and method of measurement, but there are between-study differences in the cut-point and method of measurement chosen. We show how a 95% prediction interval best summarises a random-effects meta-analysis in this situation [13], revealing the distribution of a factor's prognostic effect across the different cut-points and measurement methods. Then, in Sections 3 and 4, we consider when each study potentially provides *multiple* prognostic results for each factor, relating to different cut-points and/or methods of measurement. We show how multivariate meta-analysis models can accommodate the correlation between such results [12] and allow summary meta-analysis results to be produced for each cut-point and method of measurement, thereby facilitating clinical interpretation. The multivariate approach handles missing results (e.g. for particular cut-points or methods of measurement) in some studies and utilises correlation to gain more information, which is generally known to improve the statistical properties of meta-analysis results compared with standard (univariate) approaches [14]. We extend the general multivariate model to allow a functional relationship in the prognostic effect size over different values of the cut-point, to improve model convergence and applicability. An application is made to real examples throughout, and Section 6 concludes with some discussion.

## 2. Meta-analysis using one result per study

Let there be $i = 1$ to $k$ studies available for meta-analysis, and let each study provide just one prognostic effect estimate, $y_i$, and its variance, $s_i^2$, for a particular continuous factor of interest when dichotomised at some cut-point and measured using a chosen method of measurement. The $y_i$ will typically be either a log hazard ratio or a log odds ratio estimate. When studies use different cut-points and methods of measurement, a sensible option is to perform a separate meta-analysis for each subset of studies that used the same measurement and cut-point. However, in practice, we rarely see this approach, probably because most subsets contain only a few studies. It is more common to see researchers meta-analyse all studies together and account for potential between-study heterogeneity in prognostic effects using a random-effects model:

$$\begin{aligned} y_i &\sim N\left(\theta_i, s_i^2\right) \\ \theta_i &\sim N\left(\beta, \tau^2\right) \end{aligned} \tag{1}$$

In model (1), the $s_i^2$ is assumed known, which is a common assumption in the meta-analysis field [15], and each study's true prognostic effect, $\theta_i$, is assumed normally distributed about a summary (mean) prognostic effect, $\beta$, with between-study variance, $\tau^2$. The model can be estimated using, for example, restricted maximum likelihood (REML) or methods of moments [16]. The major problem with this approach is

that the summary effect, $\beta$, is difficult to interpret clinically as it does not relate to a particular cut-point or method of measurement. If one adopts this model, we argue that it is better to focus on the range of prognostic effects across studies by calculating a prediction interval for the potential prognostic effect of the factor in a new study [13, 17, 18] by

$$\left[\hat{\beta} - t_{N-2}\sqrt{\hat{\tau}^2 + \mathrm{Var}(\hat{\beta})}, \hat{\beta} + t_{N-2}\sqrt{\hat{\tau}^2 + \mathrm{Var}(\hat{\beta})}\right] \tag{2}$$

where $\mathrm{Var}(\hat{\beta})$ is the variance of $\hat{\beta}$, and $t_{N-2}$ is the $100(1-\alpha/2)$ percentile of the $t$-distribution with $N-2$ degrees of freedom, with $\alpha$ usually chosen as 0.05 to give a 95% prediction interval. A $t$-distribution, rather than a normal distribution, is used to help account for the uncertainty in $\hat{\tau}^2$ [13].

If the entire prediction interval does not include the value of no effect (e.g. a log odds ratio or log hazard ratio of 0), this suggests that the factor is likely to have prognostic value in new populations that use similar cut-points and methods of measurement to those in the included studies. If the interval contains the value of no effect, this indicates the factor may not be prognostic in at least some situations, and the reasons for this could then be explored. For example, if the number of studies is sufficient (e.g. > 10), then the association of the cut-point value ($x_i$) and the prognostic effect can be examined in a meta-regression by

$$\begin{aligned} y_i &\sim N\left(\theta_i, s_i^2\right) \\ \theta_i &\sim N\left(\alpha + \gamma_1 x_i, \tau^2\right) \end{aligned} \tag{3}$$

where $\gamma_1$ gives the expected change in the summary prognostic effect for a 1-unit increase in the cut-point value. After the estimation of model (3), the summary prognostic effect estimate for a particular cut-point is then obtained by $\hat{\alpha} + \hat{\gamma}_1 x_i$. Similarly, covariates could be included for the method of measurement.

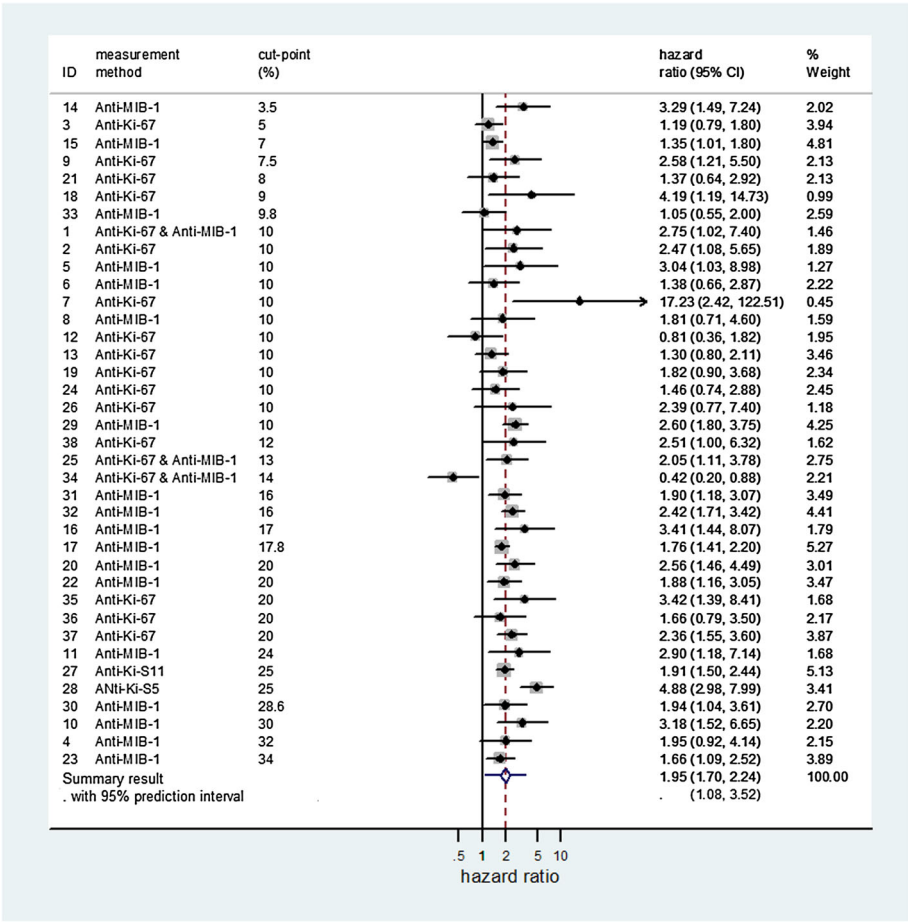


**Figure 1.** Forest plot of the study estimates and meta-analysis results for the prognostic effect of Ki-67 for overall survival in patients with breast cancer.

***Example: Ki-67 as a prognostic factor in breast cancer patients***

The Ki-67 antigen is used to evaluate the proliferative activity of breast cancer, and is recorded on a continuous scale (as a % of tumour cells that are active). De Azambuja *et al.* [11] examine whether Ki-67 is a prognostic factor for overall survival of patients with breast cancer. Across 35 studies identified, 38 unadjusted hazard ratios were obtained for independent groups of patients. The study cut-points ranged from 3.5% to 34%, and five different methods were used to measure Ki-67 activity (Figure 1). Applying model (1) produces a summary log hazard ratio of 0.67 (95% CI: 0.53 to 0.81), corresponding to a summary hazard ratio of 1.95 (95% CI: 1.70 to 2.24) (Figure 1). This reveals that – *on average* across all cut-points, methods of measurement and other heterogeneous factors – Ki-67 has prognostic value, with higher values of activity associated with a higher rate of mortality.

Unsurprisingly, there is a large heterogeneity in the meta-analysis ($I^2 = 52.3\%; \hat{\tau} = 0.28$). A 95% prediction interval for the prognostic effect of Ki-67 in a new population which, using equation (2), is 1.08 to 3.52 (Figure 1). The interval is entirely above 1, suggesting that Ki-67 has prognostic value across all the cut-points and methods of measurement used in the 35 studies included in the meta-analysis. Applying model (3) provided no evidence that either the cut-point ($\hat{\gamma}_1$ =0.008, 95% CI: -0.012 to 0.028, p = 0.43), or the method of measurement (Anti-Ki-67 or Anti-MIB-1; p = 0.77) were associated with the prognostic effect of Ki-67.

## 3. Meta-analysis using multiple cut-point results per study

Model (1) only uses one result per study, but in many studies multiple prognostic results will be available for a particular factor. Consider now that each study in the meta-analysis uses the same method of measurement for a particular factor but may provide *multiple* prognostic effect estimates for a range of different cut-points. To accommodate multiple estimates per study, we use a multivariate meta-analysis model [12, 19] that accounts for *within-study correlation* of the multiple prognostic effect estimates [20] (caused by the same patients contributing to each cut-point estimate) and any *between-study correlation* in the true effects at each cut-point. The model is now detailed in full.

### *3.1. A general model for multivariate meta-analysis of studies with multiple cut-points*

Without loss of generalisability, assume that there is a prognostic effect estimate $y_{ij}$ and its variance, $s^2_{ij}$, for each cut-point of up to $j = 1$ to $T$ different cut-points per study ($i = 1$ to $k$), and let these cut-points be ordered in an increasing value. Further, assume that the within-study covariance between each pair of cut-points (e.g. $cov_{i(1,T)}$ is the within-study covariance between $y_{i1}$ and $y_{iT}$; the estimates for cut-points 1 and $T$, respectively) is known. Section 3.3 discusses how to obtain the within-study covariances, or how to proceed if they are not available. If all studies report all cut-points, the general multivariate normal random-effects meta-analysis model assumes that [12]

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ \vdots \\ \theta_{iT} \end{pmatrix}, \begin{pmatrix} s^2_{i1} & & & \\ cov_{i(1,2)} & s^2_{i2} & & \\ \vdots & \vdots & \ddots & \\ cov_{i(1,T)} & cov_{i(2,T)} & \cdots & s^2_{iT} \end{pmatrix}\right) \tag{4}$$

where

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ \vdots \\ \theta_{iT} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{pmatrix}, \boldsymbol{\Omega}\right)$$

In model (4), $\boldsymbol{\Omega}$ is the between-study variance–covariance matrix for the true log hazard ratios and, if unstructured, is a $T$ by $T$ matrix containing $T$ between-study variances (one for each cut-point, e.g. $\tau^2_1$

for cut-point 1) in the diagonal and (2*T-1)* between-study covariances in the off-diagonals (e.g. one for each pair of cut-points, e.g. $\tau_{1,T}$ for cut-points 1 and $T$) :

$$\mathbf{\Omega} = \begin{pmatrix} \tau_1^2 & & & \\ \tau_{1,2} & \tau_2^2 & & \\ \vdots & \vdots & \ddots & \\ \tau_{1,T} & \tau_{2,T} & \cdots & \tau_T^2 \end{pmatrix}$$

The study estimates ($y_{ij}$), their variances $\left(s_{ij}^2\right)$, and covariances (e.g. $cov_{i(1,T)}$ ) are required to fit model (4). As for model (1), the within-study variances and, additionally here, the within-study covariances are assumed known.

Crucially, the model can accommodate missing results for some cut-points in a study, assuming they are missing at random [21], just as described elsewhere for missing outcomes in a multivariate meta-analysis of multiple outcomes [12, 14]. In other words, the probability that a particular $y_{ij}$ is missing for a cut-point depends solely on the observed $y_{ij}$ for other cut-points and not on the actual value of the missing $y_{ij}$ itself. Interestingly, even when data are not missing at random, this multivariate meta-analysis model has been shown to obtain summary estimates with improved statistical properties compared with univariate meta-analysis [12, 20, 22]. For example, if some cut-points are selectively missing because of their actual value of $y_{ij}$ (such as, always available if the corresponding odds ratio is statistically significant, but often unavailable if non-significant), then the missing data are missing not at random. The multivariate results are less biased than univariate results in this situation; although, the bias is not removed in full [22].

The model can be fitted using, for example, methods of moments [23, 24] or REML, using software such as SAS Proc MIXED [25] or the 'mvmeta' module in STATA [26]. The $\hat{\beta}_j$ terms give the summary (mean) prognostic effect (e.g. log hazard ratio or log odds ratio) at cut-point $j$.

### *3.2. Multivariate meta-analysis assuming a functional relationship between summary prognostic effect and cut-point*

Model (4) is best suited to situations involving a small number of cut-points across studies (e.g. 2 or 3), as otherwise, the number of parameters in the model is potentially large: one has to estimate $T$ summary means, $T$ between-study variances and $(T^2 - T)/2$ between-study covariances (correlations). One could impose a structure to $\mathbf{\Omega}$ to reduce the number of parameters to be estimated. For example, one could assume a common between-study variance at each cut-point and the same between-study correlation for all pairs of cut-point. This is potentially over-simplistic, as the between-correlation is likely to be higher for two neighbouring cut-points than for two cut-points far apart. Adopting an auto-regressive structure for $\mathbf{\Omega}$ may help address this, as done in linear mixed effects models with repeated measurements ordered over time.

Alternatively, one could assume a particular functional form for the relationship between the true prognostic effect and the cut-point value. A similar idea has been proposed in meta-analysis of test accuracy studies reporting multiple thresholds [27, 28] and is closely related to meta-analysis of longitudinal data [29]. For example, a linear relationship could be assumed such that a 1-unit increase in cut-point value, $x_j$, leads to a constant change of $\gamma$ in the prognostic effect

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix} \sim N\left( \begin{pmatrix} \alpha_i + \gamma x_1 \\ \alpha_i + \gamma x_2 \\ \vdots \\ \alpha_i + \gamma x_T \end{pmatrix}, \begin{pmatrix} s_{i1}^2 & & & \\ cov_{i(1,2)} & s_{i2}^2 & & \\ \vdots & \vdots & \ddots & \\ cov_{i(1,T)} & cov_{i(2,T)} & \cdots & s_{iT}^2 \end{pmatrix} \right)$$

$$\alpha_i \sim N\left(\alpha, \tau_\alpha^2\right) \tag{5}$$

where $x_j$ is the $j^{\text{th}}$ cut-point value (where $j = 1$ to $T$, and $T$ is the total number of different cut-points considered across studies, ordered in increasing value). In model (5), $\alpha$ is the average intercept (the summary prognostic effect when the cut-point = 0) and $\tau_\alpha^2$ is the between-study variance in the intercept.

The slope, $\gamma$, gives the summary change in log hazard ratio (odds ratio) for a 1-unit increase in the cut-point value. Extensions to model (5) may specify random slopes, but this may not be practical if some studies only provide results for one or two cut-points.

Following estimation of model (5), for example using REML, summary estimates for the log hazard ratio (odds ratio) at a particular cut-point, $t$, can be obtained by $\hat{\alpha}+\hat{\gamma}x_j$. Confidence and prediction intervals can be obtained as before. SAS code to fit model (5) is given in Supplementary Material 1.

***Nonlinear extensions***

Model (5) has substantially less parameters to estimate than model (4), especially when the number of cut-points is large. However, this computational advantage comes at the expense of assuming a particular functional relationship between prognostic effect size and cut-point value. The linear relationship specified in model (5) may not be appropriate, and rather a nonlinear trend may be preferable, for example using restricted cubic splines or fractional polynomials [30]. Model fit statistics can help identify the best fit, but when the number of studies and cut-points are small, the power to detect the true relationship is likely to be low.

### *3.3. Obtaining within-study correlations*

Models (4) and (5) require the within-study covariances (e.g. $cov_{i(1,T)}$), or equivalently the within-study correlations (e.g. $\rho_{Wi(1,T)} = s_{i1}s_{iT}cov_{i(1,T)}$) between pairs of prognostic effect estimates. These are unlikely to be available from publications [20, 31]. If individual participant data (IPD) are available, non-parametric bootstrapping is a general method that can be used to obtain them as described elsewhere [32, 33]. Where the within-study correlation between two unadjusted odds ratios are of interest, the necessary IPD can be reconstructed if the two by two tables at each cut-point are available (giving the number of patients above and below the cut-point, and the number of events in each group), from which bootstrapping can proceed. If a study's IPD (or subsequent bootstrap samples) produce a two by two table with a zero cell, odds ratios can be calculated by adding a continuity correction to the cells: we suggest the approach of Sweeting *et al.* [34], who add 1/(sample size of the opposite group) to each cell.

For situations where the effect estimates are adjusted odds ratios, unadjusted hazard ratios, or adjusted hazard ratios, it is unlikely that IPD can be recreated from published information to allow within-study correlations (covariances) to be derived via bootstrapping. One could then impute plausible values for the missing within-study correlation. For example, if correlations *are* available for unadjusted but not adjusted results, then one might assume the former is a close approximation for the latter [35]. Alternatively, one could seek clinical opinion, identify within-study correlations from related studies or perform sensitivity analyses across a range of values [20]. In particular, if some studies do provide IPD, then the within-study correlations can be derived in these studies and assumed to be the same in non-IPD studies. A Bayesian approach would also allow a prior distribution to be specified for the missing within-study correlations [33, 36, 37]. Hedges *et al.* [38] propose a robust variance estimation approach for meta-regression with correlated effect sizes but unknown correlations, which they suggest provides accurate results when there are at least 20 studies. Also, Riley *et al.* [39] propose an alternative 'overall correlation' multivariate model that does not require the within-study covariances to be specified, as it includes just one overall correlation term (an amalgamation of the within and between-study correlations), but performs well in terms of estimation of the $\beta_j$s, although it may fail to converge if the between-study heterogeneity is small relative to the within-study variances.

### *3.4. Example: Apgar score as a prognostic factor of neonatal outcomes*

The Apgar score is measured in babies immediately after birth [40]. It ranges from 0 to 10, with lower values considered to be strongly associated with a higher risk of neonatal mortality, morbidity and childhood cerebral palsy. Malin *et al.* [41] performed a systematic review of the prognostic ability of the Apgar score in babies who weigh less than 2500 g in relation to neonatal mortality and identified differences in the cut-points used in each study. Here, we use their data to illustrate the multivariate meta-analysis methods described previously.

*3.4.1. Consideration of two cut-points.* First, consider those 10 studies reporting prognostic results for the two most frequently used cut-points, 3 and 6, where values less than or equal to the cut-point are defined as 'poor'. Five studies presented prognostic results for both cut-points, four studies considered



 

**Table I.** Prognostic effect estimate for the Apgar score at each available cut-point in each study, where the outcome is neonatal mortality.

| | Cut-point 3 | | Cut-point 6 | | | |
|---|---|---|---|---|---|---|
| Study ID | Log odds ratio | SE of log odds ratio | Log odds ratio | SE of log odds ratio | Within-study covariance | Within-study correlation |
| 1 | 2.599 | 0.136 | 2.383 | 0.153 | 0.012 | 0.589 |
| 2 | 1.980 | 0.197 | 2.210 | 0.301 | 0.027 | 0.456 |
| 3 | 2.920 | 0.194 | 2.606 | 0.234 | 0.026 | 0.580 |
| 4 | 3.265 | 0.149 | 2.997 | 0.177 | 0.014 | 0.529 |
| 5 | 2.256 | 0.294 | 1.939 | 0.239 | 0.043 | 0.613 |
| 6 | 1.609 | 0.305 | — | — | — | — |
| 7 | 1.314 | 0.237 | — | — | — | — |
| 8 | 2.311 | 0.421 | — | — | — | — |
| 9 | 0.806 | 0.317 | — | — | — | — |
| 10 | — | — | 2.386 | 0.447 | — | — |

Odds ratios are defined as the odds ratios of death for those with an Apgar score $\leqslant$ cut-point value, divided by the odds of death for those with an Apgar score > cut-point value. SE, standard error.

just cut-point 3 and one study considered just cut-point 6. The unadjusted odds ratio estimates for each cut-point are shown in Table I for each study. In those five studies that provide results for both cut-points, the two estimates have moderately high within-study correlations around +0.5 (calculated using bootstrapping). Multivariate model (4) uses these correlations to gain more information ('borrow strength' [12]), thereby limiting the missing results for some cut-points in some studies, especially for cut-point 6.

Table II compares the results of a separate univariate meta-analysis (model (1)) for each cut-point with those from multivariate meta-analysis model (4). The summary odds ratio for cut-point 3 is very similar for all analyses, between 8.5 and 8.7. However, for cut-point 6 the summary odds ratio is substantially lower in the multivariate analyses. The univariate analysis gives a summary odds ratio of 11.56 (95% CI: 8.35 to 15.99), but multivariate model (4) gives a summary odds ratio of 7.93 (95% CI: 5.17 to 12.16). The between-study correlation is poorly estimated at +1, a common occurrence in multivariate meta-analysis [42]. The 'overall correlation' model of Riley *et al.* [39], which does not require within-study correlations and avoids estimating the between-study correlation, estimates an overall correlation of +0.948 and produces a summary odds ratio of 8.25, again substantially lower than the univariate solution. Indeed, in contrast to univariate results, the multivariate results suggest cut-points 3 and 6 have similar prognostic effects. The multivariate also gives noticeably narrower confidence intervals (Table II). These findings are due to the multivariate meta-analysis, under the missing at random assumption, reducing the impact of missing cut-point results by borrowing strength from other correlated cut-point results that are available.

Using the model (4) estimates, 95% prediction intervals are 1.50 to 50.29, and 2.01 to 31.25 for cut-points 3 and 6, respectively. These suggest that the prognostic effect of the Apgar score will vary greatly in magnitude across populations even when the same cut-point is used, because of unexplained heterogeneous factors. However, the effects are consistently in the same direction, such that lower values of the Apgar score indicate a higher risk of neonatal mortality.

*3.4.2. Consideration of multiple cut-points and a functional relationship.* In the previous example, just two cut-points were considered for illustration. However, there were actually 10 different cut-points considered by a total of 11 studies identified by the review (Supplementary Material 2). One study examined all 10 cut-points, but the other studies just examined one or two cut-points. Only cut-points 3 and 6 were evaluated by more than two studies. As two by two tables were available for all reported cut-points, IPD were recreated and bootstrapping used to obtain the within-study covariance estimates (available on request).

Model (5) was applied to the dataset using REML, and thus a linear relationship estimated between the prognostic effect size and the cut-point value. The estimate of the average intercept ($\hat{\alpha}$) was 2.43 (95% CI: 1.95 to 2.91; $p < 0.001$), which suggests that the odds of death for babies with an Apgar score of zero are 11.36 (= exp(2.43)) times those for babies with an Apgar value greater than zero. The estimate of the slope ($\hat{\gamma}$) was −0.068 (95% CI: −0.11 to −0.025; $p = 0.002$), indicating that the log odds ratio (comparing those below with those above the cut-point) decreases as the cut-point increases. There was

**Table II.** Meta-analysis results for the prognostic effect of the Apgar score at each cut-point, where the outcome is neonatal mortality.

| | Cut-point 3 | | | Cut-point6 | | | | |
|---|---|---|---|---|---|---|---|---|
| Analysis method | Summary log odds ratio (SE) | Summary odds ratio [95% CI] | $\hat{\tau}_1$ | Summary log odds ratio (SE) | Summary odds ratio [95% CI] | $\hat{\tau}_2$ | Between-study correlation = $\hat{\tau}_{12}/\hat{\tau}_1\hat{\tau}_2$ | Overall correlation* |
| Univariate model (1) | 2.14 (0.264) | 8.50 [5.06, 14.27] | 0.75 | 2.45 (0.166) | 11.56 [8.35, 15.99] | 0.319 | — | — |
| Multivariate model (4) | 2.16 (0.246) | 8.69 [5.37, 14.07] | 0.72 | 2.07 (0.218) | 7.93 [5.17, 12.16] | 0.560 | 1 | — |
| Alternative Multivariate model* | 2.14 (0.247) | 8.52 [5.25, 13.81] | 0.72 | 2.11 (0.205) | 8.25 [5.52, 12.34] | 0.502 | — | 0.95 |

*Using the alternative model of Riley *et al.* [39], which does not require within-study correlations

The $\hat{\tau}$ relate to the log odds ratio scale

Odds ratios are defined as the odds of death for those with an Apgar score $\leqslant$ cut-point value, divided by the odds of death for those with an Apgar score > cut-point value. SE, standard error.

substantial between-study heterogeneity in the intercept, with $\hat{\tau}_{\alpha} = 0.67$, a similar value to that observed for cut-points 3 and 6 (the most commonly reported cut-points) in the previous multivariate analyses of Section 3.4.1 (Table II). Allowing for additional heterogeneity in the slope made no difference as this was estimated as zero.

The model estimates allow a summary meta-analysis result for each cut-point, and these are shown in Table III and graphically in Figure 2. However, the linear relationship seems visually inappropriate, as the observed odds ratio estimates at a cut-point of 0 appears to be substantially larger than those for other cut-points. Model (5) was thus refitted using fractional polynomials and a selection procedure [30, 43], which marginally indicated that the cut-point variable should be included as an inverse quadratic term (1/(cut-point)$^2$), rather than linear (AIC of linear = 1353.3, AIC of inverse quadratic = 1352.9). This nonlinear function provides noticeably higher summary odds ratios at the first and last cut-point values than the linear function (Table III and Figure 2). The shape of the relationship is predominately based on the one study (labelled 'A' in Figure 2) that reported all cut-points, which is actually the original Apgar study [40]. However, compared with this single study, the summary curve provides odds ratios closer to

**Table III.** Model (5) meta-analysis results for the prognostic effect of the Apgar score at each cut-point, where the outcome is neonatal mortality.

| | *Model (3) with linear trend* | | *Model (3) with inverse quadratic trend* | | |
|---|---|---|---|---|---|
| Cut-point | Summary odds ratio | 95% CI | Summary odds ratio | 95% CI | 95% prediction interval |
| 0 | 11.36 | 7.06 to 18.30 | 21.30 | 8.07 to 56.17 | 3.33 to 136.40 |
| 1 | 10.61 | 6.68 to 16.86 | 10.40 | 6.50 to 16.63 | 2.15 to 50.41 |
| 2 | 9.91 | 6.30 to 15.59 | 9.11 | 5.88 to 14.10 | 1.90 to 43.60 |
| 3 | 9.26 | 5.92 to 14.47 | 8.69 | 5.63 to 13.42 | 1.82 to 41.57 |
| 4 | 8.65 | 5.54 to 13.49 | 8.51 | 5.51 to 13.13 | 1.78 to 40.69 |
| 5 | 8.07 | 5.17 to 12.62 | 8.41 | 5.44 to 12.99 | 1.76 to 40.23 |
| 6 | 7.54 | 4.80 to 11.86 | 8.35 | 5.40 to 12.90 | 1.75 to 39.95 |
| 7 | 7.04 | 4.44 to 11.18 | 8.31 | 5.38 to 12.85 | 1.74 to 39.77 |
| 8 | 6.58 | 4.09 to 10.58 | 8.29 | 5.36 to 12.81 | 1.73 to 39.65 |
| 9 | 6.14 | 3.76 to 10.05 | 8.27 | 5.35 to 12.78 | 1.73 to 39.57 |

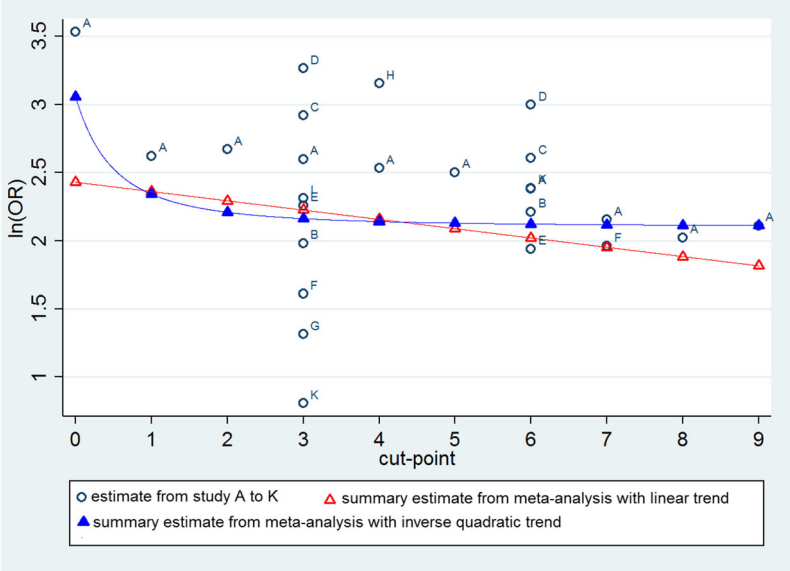


**Figure 2.** Comparison of the individual study estimates and the summary meta-analysis results obtained from model (5), assuming either a linear trend or an inverse quadratic trend, for the prognostic effect of the Apgar score at each cut-point, in relation to neonatal mortality. Confidence intervals around each point are not shown for cosmetic reasons, but are provided for the summary estimates in Table III. Odds ratios are defined as the odds of death for those with an Apgar score $\leqslant$ cut-point value, divided by the odds of death for those with an Apgar score $>$ cut-point value.

1 because it accounts also for the other studies (labelled B to K in Figure 2), which are generally giving estimates closer to 1 than study A. This is especially driven by cut-point 3, the most commonly reported cut-point, whose overall mean across all studies is much lower than its single estimate from study A. In summary, the shape of the relationship across cut-points is driven by study A; however, the location is driven by cut-point 3.

The summary curve and estimated model parameters can be used to estimate a summary odds ratio at each cut-point (Table III). Compared with analysing each cut-point separately, the summary estimates are closer to 1 for most cut-points than the observed estimates might suggest (in other words, the fitted relationship may appear visually a poor fit in Figure 2). However, as described in Section 3.4.1, this is due to other cut-points utilising the correlated information from the available cut-points 3 and 6 results, which are reported more often.

Interestingly, there is very little difference in the summary odds ratio when using cut-points from 3 to 9, but bigger changes occur when using cut-points below 3. The largest summary odds ratio is seen for a cut-point of zero. Given the large heterogeneity, it is important to present 95% prediction intervals for the odds ratio at each cut-point (Table III). All intervals are above 1 and incredibly wide; for example, the 95% prediction intervals at cut-points 3 and 6 are 1.82 to 42 and 1.75 to 40, respectively.

## 4. Meta-analysis with results for multiple measurement methods per study

The models in Section 3 can be applied or extended to deal with multiple results for different methods of measurement per study. A multivariate approach is appropriate because the different methods of measurement are often correlated at the patient-level, which induces correlation amongst the prognostic effects for the measurements. We now briefly outline the model framework, followed by an example.

### *4.1. Multiple methods of measurement per study, but consistent specification of the prognostic factor*

Assume that, for each method of measurement considered in the studies for meta-analysis, there is a consistent specification of the candidate prognostic factor; that is, the same cut-point is used in all studies, or it is always modelled as a linear trend. In this situation, multiple effect estimates arise only because of the multiple methods of measurement, which could be written as $y_{ij}$ where now $j = 1$ to $M$ (rather than $j = 1$ to $T$, as written earlier). Thus, model (4) is now applicable again, with $\hat{\beta}_j$ now giving the summary prognostic effect for the $j^{\text{th}}$ method of measurement.

As before, model (4) requires within-study correlations (covariances) to be known. Given IPD bootstrapping can still be used to obtain them. If IPD are not available, then studies that report multiple measurement results often provide the patient-level correlation between the methods of measurement. This could be used to approximate the within-study correlation of the prognostic effect estimates from the methods. For example, Wei and Higgins provide a formula for the within-study covariance for two unadjusted log odds ratio estimates [31]. This might also be used to approximate the within-study correlation between two adjusted log odds ratio estimates. In other situations (e.g. when dealing with unadjusted or adjusted hazard ratios), one could use the patient-level correlation itself as approximation for the within-study correlation. The alternative 'overall correlation' multivariate model [39] can again be fitted without within-study correlations.

### *4.2. Multiple methods of measurement and cut-points*

If there are *both* multiple methods of measurement and multiple cut-points, then the multivariate models in Section 3 can be extended, with each method of measurement in each study providing a set of $y_{ij}$s for the meta-analysis, thereby enabling a summary mean prognostic effect to be obtained for each method at each cut-point. To keep the number of parameters to a minimum, this is best achieved by extending model (5), and thereby assuming a functional relationship across cut-points for each method of measurement. For example, if there are two methods of measurement, then model (5) can be extended to include a separate intercept and slope for each method. A different between-study variance could also be assumed for each intercept, and then a between-study correlation may also be needed.

### *4.3. Example: Microvessel density as a prognostic factor in non-small-cell carcinoma*

Trivella *et al.* [44] assess whether microvessel density counts (a measure of angiogenesis) are a prognostic factor of mortality in patients with non-small-cell lung carcinoma. IPD were obtained from 16

studies, and the hazard ratio for a 1-unit change in microvessel counts was calculated in each study, after adjusting for age and tumour size. Thus, the prognostic effect is specified consistently across studies. However, two methods of measurement were used by the studies: the Chalkley method and the 'counting all microvessels' method. Three studies used both methods of measurement and so provided two adjusted hazard ratios, one for each method. In the other 13 studies, only one method of measurement was used, and so only one adjusted hazard ratio was available for either the Chalkley method ( three studies) or the all vessels method (10 studies). Multivariate meta-analysis model (4) allows the joint analysis of both methods of measurement to account for their correlation and thereby reduce this missing data problem. Log hazard ratios estimates and their standard errors are shown for each study in Table IV. For those three studies with results for both methods of measurements, within-study correlations were not reported by Trivella *et al.,* but the patient-level correlations were given as 0.55, 0.74 and 0.27, respectively. Here, we take these as an approximation for the missing within-study correlations (Table III), but recognise with the original IPD, one could use bootstrapping to obtain them.

Trivella *et al.* [44] performed a separate univariate meta-analysis (model (1)) for each method of measurement. They concluded that microvessel density was not prognostic when using the all vessels method, and there was only weak evidence that it was prognostic when using the Chalkley method. Importantly, when accounting for within and between-study correlations, the multivariate meta-analysis model (4) reaches the same conclusion (Table V). Summary hazard ratio estimates are very similar, although confidence intervals are slightly narrower. Even when assuming large within-study correlations of +0.9, summary estimates and heterogeneity estimates barely change. Thus, this additional analysis suggests

**Table IV.** Prognostic effect estimate for a 1-unit change in microvessel density, for each method of measurement in each study, where the outcome is mortality.

| | *Chalkley method* | | *All vessels method* | | | |
|---|---|---|---|---|---|---|
| Study ID | Log hazard ratio | SE of log hazard ratio | Log hazard ratio | SE of log hazard ratio | Within-study covariance* | Patient-level correlation |
| 1 | 0.122 | 0.087 | | | | |
| 3 | 0.039 | 0.06 | | | | |
| 4 | | | −0.02 | 0.09 | | |
| 5 | | | 0.058 | 0.063 | | |
| 6 | 0.104 | 0.065 | 0.02 | 0.091 | 0.0033 | 0.55 |
| 7 | 0.039 | 0.038 | | | | |
| 8 | | | 0.239 | 0.039 | | |
| 9 | | | −0.211 | 0.221 | | |
| 10 | | | 0.03 | 0.061 | | |
| 11 | | | −0.01 | 0.02 | | |
| 12 | | | 0.307 | 0.252 | | |
| 13 | 0.02 | 0.066 | 0 | 0.025 | 0.0012 | 0.74 |
| 14 | | | −0.693 | 0.758 | | |
| 15 | | | −0.174 | 0.142 | | |
| 16 | | | −0.02 | 0.025 | | |
| 17 | 0.03 | 0.047 | 0.049 | 0.037 | 0.00048 | 0.27 |

*Assuming within-study correlation is equal to the patient-level correlation.
NB data derived from the reported hazard ratios and CIs in Figures 1 and 2 of Trivella *et al.* [44] SE, standard error.

**Table V.** Meta-analysis results for the prognostic effect of a 1-unit increase in microvessel density for each method of measurement, where the outcome is mortality.

| | *Chalkley method* | | *All vessels method* | | |
|---|---|---|---|---|---|
| Analysis method | Summary hazard ratio (95% CI) | $\hat{\tau}_1$ | Summary hazard ratio (95% CI) | $\hat{\tau}_2$ | Between-study correlation $= \hat{\tau}_{12}/\hat{\tau}_1\hat{\tau}_2$ |
| Univariate model (1) | 1.049 [1.004, 1.096] | $<1 \times 10^{-13}$ | 1.032 [0.973, 1.093] | 0.077 | — |
| Multivariate model (4) | 1.051 [1.007, 1.097] | 0.0025 | 1.030 [0.972, 1.091] | 0.077 | 1 |

that the original meta-analysis results of Trivella *et al.* are robust after accounting for the correlation in the results for the two measurement methods. Finally, we note that the 'overall correlation' model of Riley *et al.* [39] does not converge for this example, most likely due to the miniscule heterogeneity for the Chalkley method causing estimation difficulties for the overall correlation.

## 5. Discussion

Empirical evaluations, across a wide-range of disease fields, have shown that meta-analysis of prognostic factor studies is often limited by heterogeneity and missing results in primary studies [1,45]. In this article, we have suggested meta-analysis approaches that allow more clinically useful results about prognostic factors in the presence of heterogeneity. In particular, we have focused on multivariate meta-analysis methods to examine prognostic value at particular cut-points and for specific methods of measurement. As they utilise more information, multivariate meta-analysis methods are being used to synthesise multiple treatment comparisons [46] and multiple outcomes [47]. Here, under a missing at random assumption, the utilisation of correlation reduces the impact of missing results for particular cut-points and methods of measurement in some studies [14, 22]. Even when data are not missing at random, the multivariate meta-analysis model is likely to obtain more appropriate inferences than current univariate approaches, as the correlation reduces (although does not entirely remove) the impact of selectively missing results [12, 20, 22]. In our examples, the multivariate approach revealed important insight about the prognostic value of the Apgar score and microvessel density. In particular, in the Apgar example, the multivariate approach produced substantially lower summary estimates at some cut-points than previously thought. For example, in Table II the univariate meta-analysis suggests a cut-point of $\leqslant 6$ gives a larger prognostic effect, whereas the multivariate meta-analysis suggests it is slightly higher for cut-point 3. The latter is clinically more intuitive, as lower values are considered to put babies at a higher risk, and so lower cut-points are expected to lead to higher odds ratios.

### Usefulness of prognostic factor effects based on a cut-point

There are a number of clinical applications where a cut-point may be useful for implementing the prognostic factor. For example, in complex health economic models or disease outcome simulation models [48], for parsimony, a population may be divided into two groups defined by a prognostic factor with a cut-point. In randomised trials that incorporate prognostic factors in the randomisation process (e.g. within minimisation or stratification), it may be more convenient to consider dichotomised factors. The actual analysis of trials (or indeed observational studies) may specify, *a priori*, a set of prognostic factors (confounders) to be adjusted for [49]; their inclusion may be based on evidence in previous studies, for which prognostic value at cut-points may only be known. In prognostic models for clinical risk prediction, prognostic factors within the risk equation are sometimes categorised to ease implementation by clinicians and health professionals [50]. In clinical decision making, treatment decisions may be informed by prognostic factor values above a cut-point. For example, the use of drug-eluting stents for the treatment of coronary artery disease was restricted by The National Institute for Health and Clinical Excellence to patients with coronary artery lesions longer than 15 mm [51], a prognostic factor for the probability of restenosis, as patients with such lesions had a worse prognosis and thus were considered to have a greater potential to gain from treatment. In such examples, knowledge of the absolute risk in each of the groups defined by the prognostic factor is clearly important, not just the relative risks of the two groups.

### Linear and nonlinear prognostic effects

Although cut-point specific results can be useful, it is well known that dichotomising continuous factors loses statistical power to detect their true prognostic effect [52]. Therefore, the prognostic ability of a factor is better examined on its continuous, rather than dichotomised, scale and its linear or non-linear relationship with outcome risk examined [53]. Sauerbrei and Royston [54], and Gasparrini *et al.* [55] extensively discuss this approach when IPD are available for meta-analysis. In the microvessel density example, we *were* able to examine the prognostic effect of a 1-unit increase in the factor, for each of two different methods of measurement, as the original meta-analysts used the IPD to analyse the factor on its continuous scale in each study. Analysing prognostic factors on their continuous scale is especially important in risk prediction research, where prognostic models are required to predict absolute outcome risk for individuals. Maintaining a continuous scale improves the range of possible predictions from the

model, and is more likely to lead to a generalisable model than when including factors dichotomised. This is a major reason why IPD is increasingly sought when developing such models from multiple studies [56].

However, without IPD, meta-analysts will predominately have to use reported prognostic results, which are most typically given for two groups defined by a cut-point. In this situation, if researchers still want to examine the effect of a 1-unit increase in a prognostic factor on outcome risk, then meta-analysis approaches for examining dose-response relationship are potentially applicable, as proposed by Greenland and Longnecker [35], and extended by others, for example [57–59]. To apply these methods, some additional knowledge of the factor's underlying distribution is usually needed, as a particular factor value needs to be assigned to all patients within each group defined by a cut-point (e.g. take the mid-point or median) so that the trend across groups can be estimated in each study, to then be pooled in a meta-analysis. The choice of such value can impact upon the results [57].

### Modelling issues and solutions

Researchers wishing to implement our multivariate approaches should recognise potential modelling issues. We showed how to model the functional relationship between prognostic effect size and cut-point value, and this will often be the most useful approach as it substantially reduces the number of parameters to estimate. Another practical issue is the derivation of within-study correlations. These are most easily derived using bootstrapping when IPD are available [60]. For unadjusted effects of dichotomous prognostic factors for a binary outcome, the IPD can be reconstructed from the published two by two table, as discussed in Section 3.3. However, in other situations, the IPD must be obtained directly from the original study authors, which might not be possible. When the IPD are not available, we showed how the multivariate approach might be implemented by using the patient-level correlations [31] or a reparameterised multivariate model that does not require within-study correlations [39]. If the number of studies is very large (>20), then the robust variance approach of Hedges *et al.* [38] would allow the functional relationships to be modelled even when the within-study correlations are unknown. Even obtaining IPD from just a single study can help, as within-study correlations might be assumed exchangeable to other studies [33]. Nevertheless, further methodological research on how to derive within-study correlations from published data is needed, as this would undoubtedly improve the uptake of the multivariate models proposed here.

### Limitations of our work

Our multivariate models with a functional relationship should not be extrapolated outside the range of cut-points available in the studies for meta-analysis. Furthermore, the multivariate models are unlikely to be reliable where most studies just report one cut-point, especially if that cut-point was selected on the basis of optimising the $p$-value [61]. In other words, the multivariate models are likely to perform best when at least one study has a large number of the cut-points of interest, so that the relationship of prognostic effect sizes across cut-points is based primarily on within-study information, rather than between-study information that is more prone to ecological bias and study-level confounding [62].

We note also that, in some situations, there may be a known transformation that maps values from one method of measurement to another method of measurement. Clearly, if one can reliably transform findings from one scale to the other, then this is preferable to our approach and can be used to obtain missing method of measurement in studies that only report a subset of those of interest. However, in many situations, the relationship between competing methods is not known with high precision, or only applicable if the IPD are available. In this situation, our suggested joint multivariate meta-analysis of all methods is then useful to account for their correlation.

Our models dealing with multiple cut-points provide prognostic effect sizes such as odds ratios or hazard ratios. However, they do not provide absolute risks for those above and below a particular cut-point. To derive absolute risks following our model, additional information would be required such as the absolute risk in at least one of the groups defined by a cut-point and the distribution of the factors values, such that the proportion of the population that fall in between each cut-point can be derived. However, absolute risks are typically more important from a prognostic model (which contains multiple prognostic factors in combination) [63], whereas here, the focus is on whether a particular factor has prognostic

value [1]. Furthermore, absolute risks tend to only be applicable to specific populations, whereas relative effects (such as odds ratios) are often more reasonably transportable between populations.

**Is meta-analysis sensible?**

Other researchers have dealt with multiple cut-points and methods of measurement by transforming to a standardised scale [64, 65], under various assumptions. This is often at the expense of clinical interpretability, as the scale then does not translate to a real metric for use. Our approaches rather produce results that translate directly to specific cut-points and methods of measurement. However, the methods do not solve all the problems, as other heterogeneous factors remain unaddressed by our work, such as different adjustment factors [32] and stages of disease across studies. In some situations, heterogeneity may be considered too large to warrant meta-analysis, and researchers should always exercise epidemiological and clinical judgement before pooling. If meta-analysis is deemed sensible, then the pooled result may still be difficult to interpret when there is heterogeneity, and so in Section 2, we proposed that it is better to focus on the range of prognostic effects across studies by calculating a prediction interval. IPD can help reduce heterogeneity in prognostic factor studies, as seen in the microvessel density example where the heterogeneity was zero for both methods of measurement, following extensive data cleaning and standardisation of statistical analysis method and adjustment factors in each study [44]. However, IPD often does not solve all the problems for meta-analysis of prognostic factor studies [66], and in particular, publication bias related issues are a strong concern for this field [6, 7]. Indeed, multivariate meta-analysis may still be important with IPD to reduce the impact of (selectively) missing results [22].

## 6. Conclusion

In conclusion, we have proposed approaches for handling different cut-points and methods of measurement in a meta-analysis of prognostic factor studies. These are especially important when synthesising published prognostic results but are also potentially useful when IPD are available for some or all studies. Many issues remain in this field, and ultimately, a move toward prospectively planned pooled analyses would be preferred [67].

## Acknowledgements

EE and RDR were supported by an MRC Partnership Grant for the PROGnosis RESearch Strategy (PROGRESS) group (grant reference number: G0902393). MJP and RDR were supported by funding from an MRC Methodology Research Grant in Multivariate Meta-analysis (grant reference number: MR/J013595/1). We would like to thank Ian White and Dan Jackson (MRC Biostatistics Unit, Cambridge) for their helpful feedback on earlier versions of this article. We thank two anonymous reviewers whose comments helped considerably improve the article.

## References

1. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, Malats N, Briggs A, Schroter S, Altman DG, Hemingway H. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Medicine* 2013; **10**:e1001380.
2. Armitage P, Gehan EA. Statistical methods for the identification and use of prognostic factors. *International Journal of Cancer* 1974; **13**:16–36.
3. Altman DG. Systematic reviews of evaluations of prognostic variables. *British Medical Journal* 2001; **323**:224–228.
4. Altman DG, Riley RD. An evidence-based approach to prognostic markers. *Nature Clinical Practice Oncology* 2005; **2**:466–472.
5. Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. Prognosis research: toward evidence-based results and a Cochrane methods group. *Journal of Clinical Epidemiology* 2007; **60**:863–865. author reply 865-866.
6. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer* 2007; **43**:2559–2579.
7. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. *Journal of the National Cancer Institute* 2005; **97**:1043–1055.
8. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clinical Chemistry* 2008; **54**:1101–1103.
9. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Jones DR, Heney D, Burchill SA. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *British Journal of Cancer* 2003; **88**:1191–1198.



 

10. Malats N, Bustos A, Nascimento CM, Fernandez F, Rivas M, Puente D, Kogevinas M, Real FX. P53 as a prognostic marker for bladder cancer: a meta-analysis and review. *Lancet Oncology* 2005; **6**:678–686.
11. de Azambuja E, Cardoso F, de Castro G, Jr., Colozza M, Mano MS, Durbecq V, Sotiriou C, Larsimont D, Piccart-Gebhart MJ, Paesmans M. Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients. *British Journal of Cancer* 2007; **96**:1504–1513.
12. Jackson D, Riley RD, White IR. Multivariate meta-analysis: potential and promise. *Statistics in Medicine* 2011; **30**: 2481–2498.
13. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A* 2009; **172**:137–159.
14. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**:78–97.
15. Whitehead A. *Meta-analysis of Controlled Clinical Trials*. Wiley: West Sussex, 2002.
16. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
17. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *British Medical Journal* 2011; **342**: d549.
18. Ades AE, Lu G, Higgins JP. The interpretation of random-effects meta-analysis in decision models. *Medical Decision Making* 2005; **25**:646–654.
19. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589–624.
20. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *JRSS Series A* 2009; **172**(4): 789–811.
21. Little JA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley and Sons: New York, 2002.
22. Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine* 2012; **31**:2179–2195.
23. Jackson D, White IR, Riley RD. A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression. *Biometrical Journal* 2013; **55**:231–245.
24. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine* 2010; **29**:1282–1297.
25. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O. *SAS for Mixed Models* 2nd edn. SAS Institute Inc: Cary, NC, 2006.
26. White IR. Multivariate meta-analysis. *The STATA Journal* 2009; **9**:40–56.
27. Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Medical Research Methodology* 2009; **9**:73.
28. Riley RD, Takwoingi Y, Trikalinos T, Guha A, Biswas A, Ensor J, Morris RK, Deeks JJ. Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model. *Journal of Biometrics and Biostatistics* 2014; **5**:3.
29. Jones AP, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical Trials* 2009; **6**:16–27.
30. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistics Society Series A* 1999; **162**:71–94.
31. Wei Y, Higgins JP. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* 2013; **32**:1191–1205.
32. The Fibrinogen Studies Collaboration. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Statistics in Medicine* 2009; **28**:1218–1237.
33. Bujkiewicz S, Thompson JR, Sutton AJ, Cooper NJ, Harrison MJ, Symmons DP, Abrams KR. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Statistics in Medicine* 2013; **32**:3926–3943.
34. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; **23**:1351–1375.
35. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology* 1992; **135**:1301–1309.
36. Nam IS, Mengersen K, Garthwaite P. Multivariate meta-analysis. *Statistics in Medicine* 2003; **22**:2309–2333.
37. Wei Y, Higgins JP. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* 2013; **32**: 2911–2934.
38. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods* 2010; **1**(1):39–65.
39. 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**:172–186.
40. Apgar V. A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anesthesia & Analgesia* 1953; **32**:260–267.
41. Malin GL, Morris RK, Ahmad S, Riley RD, Khan KS. Does the Apgar score matter? Investigating the relationship between a low score and adverse outcomes from birth to childhood. *Archives of Disease in Childhood - Fetal and Neonatal Edition* 2013; **98**(Suppl 1):A83–A83.
42. 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.
43. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine* 2007; **26**:5512–5528.
44. Trivella M, Pezzella F, Pastorino U, Harris AL, Altman DG. Microvessel density as a prognostic factor in non-small-cell lung carcinoma: a meta-analysis of individual patient data. *Lancet Oncology* 2007; **8**:488–499.



 

45. Sauerbrei W, Holländer N, Riley RD, Altman DG. Evidence-based assessment and application of prognostic markers: the long way from single studies to meta-analysis. *Communications in Statistics* 2006; **35**:1333–1342.
46. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods* 2012; **3**:111–125.
47. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537–2550.
48. Kopec JA, Finès P, Manuel DG, Buckeridge D, Flanagan WM, Oderkirk J, Abrahamowicz M, Harper S, Sharif B, Okhmatovskaia A, Sayre EC, Rahman MM, Wolfson MC. Validation of population-based disease simulation models: a review of concepts and methods. *BMC Public Health* 2010; **10**:710.
49. Senn S. Testing for baseline balance in clinical trials. *Statistics in Medicine* 1994; **13**:1715–1726.
50. Steyerberg EW. *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating*. Springer: New York, 2009.
51. NICE. National Institute for Health and Clinical Excellence. NICE technology appraisal guidance 152: drug-eluting stents for the treatment of coronary artery disease (part review of NICE technology appraisal guidance 71), 2008.
52. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 2006; **25**:127–141.
53. Sauerbrei W. Prognostic factors–confusion caused by bad quality of design, analysis and reporting of many studies. In *Current Research in Head and Neck Cancer. Advances in Otorhinolaryngology*, Bier H (ed.) Karger: Basel, 2005; 184–200.
54. Sauerbrei W, Royston P. A new strategy for meta-analysis of continuous covariates in observational studies. *Statistics in Medicine* 2011; **30**:3341–3360.
55. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine* 2012; **31**:3821–3839.
56. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine* 2013; **32**:3158–3158.
57. Hartemink N, Boshuizen HC, Nagelkerke NJ, Jacobs MA, van Houwelingen HC. Combining risk estimates from observational studies with different exposure cutpoints: a meta-analysis on body mass index and diabetes type 2. *American Journal of Epidemiology* 2006; **163**:1042–1052.
58. Shi JQ, Copas JB. Meta-analysis for trend estimation. *Statistics in Medicine* 2004; **23**:3–19. discussion 159-162.
59. Orsini N, Li R, Wolk A, Khudyakov P, Spiegelman D. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. *American Journal of Epidemiology* 2012; **175**:66–73.
60. Riley RD, Price MJ, Jackson D, Wardle M, Gueyffier F, Wang J, Staessen JA, White IR. Multivariate meta-analysis using individual participant data. *Research Synthesis Methods* 2014. DOI: 10.1002/jrsm.1129.
61. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994; **86**:829–835.
62. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine* 2002; **21**: 371–387.
63. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine* 2013; **10**:e1001381.
64. Look M, van Putten W, Duffy M, Harbeck N, Christensen IJ, Thomssen C, Kates R, Spyratos F, Ferno M, Eppenberger-Castori S, Fred Sweep CG, Ulm K, Peyrat JP, Martin PM, Magdelenat H, Brunner N, Duggan C, Lisboa BW, Bendahl PO, Quillien V, Daver A, Ricolleau G, Meijer-van Gelder M, Manders P, Edward Fiets W, Blankenstein M, Broet P, Romain S, Daxenbichler G, Windbichler G, Cufer T, Borstnar S, Kueng W, Beex L, Klijn J, O'Higgins N, Eppenberger U, Janicke F, Schmitt M, Foekens J, Bendah PO. Pooled analysis of prognostic impact of uPA and PAI-1 in breast cancer patients. *Thrombosis and Haemostasis* 2003; **90**:538–548.
65. Hemingway H, Philipson P, Chen R, Fitzpatrick NK, Damant J, Shipley M, Abrams KR, Moreno S, McAllister KSL, Palmer S, Kaski JC, Timmis AD, Hingorani AD. Evaluating the quality of research into a single prognostic biomarker: a systematic review and meta-analysis of 83 studies of C-reactive protein in stable coronary artery disease. *PLoS Medicine* 2010; **7**(6):e1000286.
66. Abo-Zaid G, Sauerbrei W, Riley RD. Individual participant data meta-analysis of prognostic factor studies: state of the art *BMC Medical Research Methodology* 2012; **12**:56.
67. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology* 1999; **28**:1–9.

## Supporting information

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



 