Tutorial in Biostatistics

Statistics in Medicine

Received: 11 March 2016, Accepted: 13 September 2016 Published online 16 October 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7141

Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ

Danielle L. Burke*† Joie Ensor and Richard D. Riley

Meta-analysis using individual participant data (IPD) obtains and synthesises the raw, participant-level data from a set of relevant studies. The IPD approach is becoming an increasingly popular tool as an alternative to traditional aggregate data meta-analysis, especially as it avoids reliance on published results and provides an opportunity to investigate individual-level interactions, such as treatment-effect modifiers. There are two statistical approaches for conducting an IPD meta-analysis: one-stage and two-stage. The one-stage approach analyses the IPD from all studies simultaneously, for example, in a hierarchical regression model with random effects. The two-stage approach derives aggregate data (such as effect estimates) in each study separately and then combines these in a traditional meta-analysis model. There have been numerous comparisons of the one-stage and two-stage approaches via theoretical consideration, simulation and empirical examples, yet there remains confusion regarding when each approach should be adopted, and indeed why they may differ.

In this tutorial paper, we outline the key statistical methods for one-stage and two-stage IPD meta-analyses, and provide 10 key reasons why they may produce different summary results. We explain that most differences arise because of different modelling assumptions, rather than the choice of one-stage or two-stage itself. We illustrate the concepts with recently published IPD meta-analyses, summarise key statistical software and provide recommendations for future IPD meta-analyses. © 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: individual patient data; individual participant data; meta-analysis; IPD; one-stage; two-stage

1. Introduction

Statistical methods for meta-analysis and evidence synthesis are increasingly popular tools in medical research, as they synthesise quantitative information across multiple studies to produce evidence-based results. An aggregate data meta-analysis is the most common approach, where summary study results (such as treatment effect estimates and their standard errors) are obtained from study publications or study authors, and then synthesised. This approach is, at least in principle, relatively quick and inexpensive, but it often faces problems such as poor and selective reporting in primary studies, publication bias and low power to detect individual-level interactions such as how a participant-level covariate modifies treatment effect [1,2]. An individual participant data (IPD) meta-analysis can help overcome many of these issues, by obtaining and then synthesising the raw, participant-level data from each study. For example, IPD allows the meta-analyst to standardise the inclusion criteria and analyses across studies, to obtain study results that had not been provided by the trial publications and to check modelling assumptions [3]. An important advantage is being able to model individual-level interactions directly within studies, which has substantially greater power and avoids ecological bias compared with a meta-regression of aggregate data across studies [4,5]. For such reasons, there has been an increase in the number of IPD meta-analyses in the last decade [5,6].

There are two competing statistical approaches for IPD meta-analysis: a two-stage or a one-stage approach [7]. In the two-stage approach, firstly, the IPD from each study are analysed separately in order

Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, U.K.
**Correspondence to: Danielle L Burke, Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, U.K.*
†E-mail: d.burke@keele.ac.uk

to obtain aggregate (summary) data of interest (such as an effect estimate and its confidence interval (CI)); then secondly, these are combined by an appropriate fixed-effect or random effects meta-analysis model. The alternative one-stage IPD meta-analysis approach analyses all the patient-level data from all the trials in a single step, for example, using a hierarchical (random effects) model that accounts for the clustering of patients within studies [8]. The two-stage approach is often preferred [2,9] because in the second stage it uses standard meta-analysis methods that are well documented, for example, in the Cochrane Handbook [10]. However, one-stage methods have also been recommended because they use a more exact likelihood specification [3,11], which avoids the assumptions of within-study normality and known within-study variances, which are especially problematic in meta-analyses with small studies and/or rare events. Yet, one-stage methods are also criticised for being computationally intensive and prone to convergence problems [3,12].

This discrepant advice is causing a dilemma for researchers who are writing grant applications and statistical analysis protocols: do they adopt a one-stage or a two-stage approach? This is especially important as statistical and/or clinical conclusions may depend on the chosen approach. For example, Debray *et al.* found that erythema was a statistically significant predictor of deep vein thrombosis (DVT) in a one-stage analysis ($p=0.03$), but not a two-stage analysis ($p=0.12$) [3]. For this reason, Tierney *et al.* advise, 'It is important, therefore, that the choice of one or two-stage analysis is specified in advance or that results for both approaches are reported' [13].

To aid this process, this article provides a tutorial of the two-stage and one-stage approaches to IPD meta-analysis. In Section 2, we introduce the approaches using statistical notation and provide examples for continuous, binary and time-to-event outcomes, which illustrate how the two approaches often give similar results. Available statistical software is also summarised. Section 3 then outlines 10 key reasons why differences may arise to help users resolve them if they occur in practice. In particular, we highlight that differences are most likely due to the analyst making discrepant modelling assumptions, changing the specification of unknown parameters or using different techniques for model estimation or CI derivation. Real examples are used to illustrate the messages and include application of fixed-effect and random effects models for obtaining summary meta-analysis results for treatment and prognostic effects, treatment-covariate interactions and test accuracy. Section 4 then concludes with some discussion and recommendations for future IPD meta-analyses.

2. An introduction to one-stage and two-stage IPD meta-analysis models

We now introduce the two approaches using statistical notation, starting with the more familiar two-stage approach.

2.1. The two-stage approach

2.1.1. First stage. Let us assume that there are $i=1$ to K trials for the IPD meta-analysis and that a treatment effect is of interest. In the two-stage approach, the first stage involves a separate analysis in each study to derive the K treatment effect estimates and their variances, using an appropriate method chosen by the meta-analyst. For example, estimates for odds ratios (ORs) or risk ratios (and variances for log ORs and log relative risks) can be derived using standard formulae [14] after collapsing the IPD to 2×2 contingency tables. More generally, the estimates and standard errors can be derived by fitting a regression model suitable for the outcome of interest, with a model specification deemed appropriate by the analyst. This enables covariate adjustment if necessary, which is especially important in situations where confounding is a concern. We focus now on utilising familiar regression models for continuous, binary and time-to-event outcomes; however, there are many other modelling options available.

If the outcome is continuous (blood pressure, say) then one may use, for example, maximum likelihood (ML) or restricted ML (REML) estimation to fit an appropriate linear regression in each study separately, such as an analysis of covariance (ANCOVA) model. At baseline (i.e. before randomisation) the j^{th} participant in the i^{th} trial provides their initial (blood pressure) value, which we denote by y_{Bij} (where B indicates baseline). Also, each participant provides their final (blood pressure) value after treatment, which we denote by y_{Fij} (where F indicates final). Also, let x_{ij} be 0/1 for participants in the control/treatment group, respectively. One can then fit the following ANCOVA model to the IPD in each trial separately, where the final score is regressed against the baseline score and the treatment effect:

$$y_{Fij} = \alpha_i + \beta_i y_{Bij} + \theta_i x_{ij} + e_{ij} \quad (1)$$

$$e_{ij} \sim N\left(0, \sigma_i^2\right)$$

In this model, α_i is the intercept (the expected response in the placebo group for those with a zero y_{Bij}), θ_i is the underlying treatment effect (the mean difference in final score between treatment groups, after adjusting for baseline score), β_i denotes the mean change in y_{Fij} for a one-unit increase in y_{Bij}, and σ_i^2 is the residual variance of the responses after accounting for the treatment effect and the baseline values.

For non-continuous outcomes, generalised linear regression models are usually preferred, such as binomial logistic regression for binary outcomes, ordinal logistic regression for ordinal outcomes, multinomial regression for multinomial outcomes and Poisson regression for count outcomes. For example, for a binary outcome (e.g. death by 1 month), one might use ML estimation to fit the following logistic regression model in each trial separately:

$$\ln\left(\frac{E\left(y_{ij}\right)}{1-E\left(y_{ij}\right)}\right) = \ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + \theta_i x_{ij} \tag{2}$$

where y_{ij} is 1 or 0 for participants with or without the outcome, respectively; p_{ij} is the probability of participant j experiencing the event; α_i is the intercept (the expected log odds of the event for the control group); and θ_i denotes the treatment effect (the log OR). Baseline covariates might also be included in equation (2), alongside x_{ij}, in order to increase power or to adjust for baseline confounding. If interest was in relative risks rather than ORs, then one could fit a binomial regression with a log-link or a Poisson regression with robust standard errors [15].

If the outcome is time-to-event (e.g. time to death), a straightforward approach would be to use ML estimation to fit a Cox regression model in each study separately,

$$h_{ij}(t) = h_{0i}(t) exp\left(\theta_i x_{ij}\right) \tag{3}$$

where $h_{ij}(t)$ is the hazard rate over time, t, for participant j; $h_{0i}(t)$ is the baseline hazard (for those in the placebo group); and θ_i denotes the log hazard ratio (i.e. the treatment effect). As before, baseline covariates might also be included in equation (3), alongside x_{ij}, in order to increase power or to adjust for baseline confounding.

Following estimation of an equation such as (1), (2) or (3) in each trial separately, the meta-analyst obtains K treatment effect estimates, $\hat{\theta}_i$, and their variances, Var($\hat{\theta}_i$), ready for the second stage (see succeeding paragraphs). If a different measure is of interest, then the equations should be modified accordingly to allow the measure to be estimated. For example, to examine the interaction between baseline value and treatment effect (a so-called 'treatment–covariate interaction'), equation (1) can be modified to

$$y_{Fij} = \alpha_i + \beta_i y_{Bij} + \theta_i x_{ij} + \lambda_i\left(y_{Bij} x_{ij}\right) + e_{ij} \tag{4}$$

$$e_{ij} \sim N\left(0, \sigma_i^2\right)$$

where the interaction term, λ_i, denotes the mean increase in treatment effect for a one-unit increase in the baseline value. Estimation of equation (4) in each trial then provides the meta-analyst with K treatment–covariate interaction estimates (and their variances) ready for the second stage. In the following subsection, in our description of the second stage, we focus on the synthesis of treatment effects, but the concepts apply equally to meta-analysis of any parameter estimate of interest.

2.1.2. Second stage. In the second stage, the treatment effect estimates, $\hat{\theta}_i$, obtained in the first stage are then combined across trials, assuming that the true treatment effects are either fixed or random across studies. The fixed-effect model assumes that $\hat{\theta}_i$ are all estimates of the same underlying treatment effect in all studies, represented as θ. It can be written generally as [16]

$$\hat{\theta}_i \sim N\left(\theta, Var(\hat{\theta}_i)\right) \tag{5}$$

where the Var($\hat{\theta}_i$) estimates are also taken from the first stage, and usually assumed known. The most common method to estimate θ is the inverse variance method, which provides a weighted average, where the weight of each trial, w_i, is defined as [10]

$$w_i = \frac{1}{var\left(\hat{\theta}_i\right)} \tag{6}$$

and the pooled treatment effect, θ, and its variance are calculated by

$$\hat{\theta} = \frac{\sum_{i=1}^{K} \hat{\theta}_i w_i}{\sum_{i=1}^{K} w_i} \tag{7}$$

$$var\left(\hat{\theta}\right) = \frac{1}{\sum_{i=1}^{K} w_i} \tag{8}$$

These solutions can also be derived using ML solution estimation. As the fixed-effect model assumes that the true treatment effect is the same in all studies, the obtained summary estimate, $\hat{\theta}$, should be interpreted as the best estimate of this common treatment effect.

The random effects model allows for between-study variation, τ^2, in the true treatment effect and makes the assumption that the different studies are estimating different, yet related, treatment effects. The random effects model can be written generally as [16]

$$\hat{\theta}_i \sim N\left(\theta_i, Var\left(\hat{\theta}_i\right)\right) \tag{9}$$

$$\theta_i \sim N\left(\theta, \tau^2\right)$$

where the $\mathrm{Var}(\hat{\theta}_i)$ estimates are again assumed known and u_i denotes a random-effect, which indicates that the treatment effect in the i^{th} trial, θ_i, is assumed normally distributed about an average treatment effect, θ, with between-study variance, τ^2. As the random effects model assumes the true treatment effect varies across studies, the obtained summary estimate, $\hat{\theta}$, should be interpreted as the estimated *average* of the distribution of true treatment effects in the meta-analysis. Equation (9) reduces to equation (5) when τ^2 equals zero.

To obtain meta-analysis results, an inverse variance approach can again be taken but with the weights of each trial now adjusted to incorporate an estimate of τ^2:

$$w_i^* = \frac{1}{var\left(\hat{\theta}_i\right) + \hat{\tau}^2} \tag{10}$$

Then, the estimate of the pooled effect and its variance are calculated using:

$$\frac{\hat{\theta} = \sum_{i=1}^{K} \hat{\theta}_i w_i^*}{\sum_{i=1}^{K} w_i^*} \tag{11}$$

$$var\left(\hat{\theta}\right) = \frac{1}{\sum_{i=1}^{K} w_i^*} \tag{12}$$

Perhaps the most popular method of estimating τ^2 is the non-iterative, non-parametric methods of moments (MoM) estimator of DerSimonian and Laird [17], due to its speed and availability in non-sophisticated statistical packages, such as RevMan [18]. It also avoids the assumption of normally distributed effects as written in equation (9), as it simply uses the Q-statistic and inverse-variance weights. Other non-iterative estimators are also available, which have been shown to improve upon DerSimonian and Laird in some situations [19,20]. Other iterative methods are available to estimate τ^2 and θ in equation (9) [19,21], including ML and REML. Bayesian approaches can also be used, but these are beyond the scope of this paper. Indeed, there is much ongoing debate in the meta-analysis literature about the best method to estimate τ^2 and (subsequently) θ [22,23], and we return to this issue in Section 3.

2.2. The one-stage approach

The one-stage approach analyses all the IPD from all studies simultaneously. The model framework depends on the outcome type and the model specification deemed appropriate by the analyst. In the succeeding discussion, we again focus on specifying typical models for continuous, binary and time-to-event outcomes in regard to a summary treatment effect estimate. For survival data, this means we focus mainly on Cox regression models but recognise that other (flexible) parametric survival model specifications are possible. Furthermore, as discussed in the first stage of the two-stage approach, one-stage regression models can also be specified for alternative outcomes, such as count and ordinal data.

For continuous outcomes, a one-stage IPD meta-analysis can be specified in a linear (mixed) model. For example, assuming a fixed treatment effect, the following one-stage ANCOVA model can be fitted to all studies simultaneously:

$$\begin{aligned} y_{Fij} &= \alpha_i + \beta_i y_{Bij} + \theta x_{ij} + e_{ij} \\ e_{ij} &\sim N\left(0, \sigma_i^2\right) \end{aligned} \tag{13}$$

The parameters are as defined previously, but we emphasise that the subscript, i, denotes that a separate parameter is estimated for each study. For example, α_i denotes that a separate intercept term is estimated for each study; this is sometimes referred to as 'stratifying by trial', and it allows for different mean control group responses per trial and thus accounts for the clustering of participants in trials. Similarly, β_i denotes that each trial has a different adjustment term for baseline values, and σ_i^2 denotes a distinct residual variance per trial. It is also important to note that the assumptions of distinct baseline values and residual variances represent just one possible model. Other options are available, and they will be discussed in this article as possible reasons for differences between the one-stage and two-stage approaches.

Equation (13) could be extended to include a random treatment effect:

$$\begin{aligned} y_{Fij} &= \alpha_i + \beta_i y_{Bij} + \theta_i x_{ij} + e_{ij} \\ \theta_i &= \theta + u_i \\ u_i &\sim N\left(0, \tau^2\right) \\ e_{ij} &\sim N\left(0, \sigma_i^2\right) \end{aligned} \tag{14}$$

Some researchers also prefer to place a random effect on the study intercepts [24]; for example equation (13) is then modified to

$$\begin{aligned} y_{Fij} &= \alpha_i + \beta_i y_{Bij} + \theta x_{ij} + e_{ij} \\ \alpha_i &\sim N\left(\alpha, \delta^2\right) \\ e_{ij} &\sim N\left(0, \sigma_i^2\right) \end{aligned} \tag{15}$$

This is potentially a strong assumption, and will usually be unnecessary, but illustrates the so-called flexibility of the one-stage approach [3]. It is perhaps more realistic and most helpful when the average baseline response α is also of interest, or when needing to reduce (perhaps due to convergence problems) the number of parameters to estimate (that is, rather than estimating K intercepts in equation (13), we now estimate just α and δ in equation (15)). A similar modification can be made to equation (14) to include a random treatment effect and a random study effect, with a suitable correlation structure between the random effects. Other 'flexible' modifications are also possible; for example, assuming each study has a common (rather than separate) baseline adjustment term and a common (rather than separate) residual variance [25]. As mentioned, we return to these issues in Section 3. The aforementioned linear (mixed) models are typically estimated using ML or REML.

For binary outcomes, a generalised linear (mixed) model is required. For example, a one-stage logistic regression meta-analysis, with stratified intercepts and a random treatment effect, can be expressed as

$$\ln\left(\frac{E\left(y_{ij}\right)}{1-E\left(y_{ij}\right)}\right)=\ln\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_i+\theta_i x_{ij} \tag{16}$$

$$\theta_i=\theta+u_i$$

$$u_i \sim N\left(0,\tau^2\right)$$

where model parameters are as defined previously. As for the one-stage general linear model, the distinct intercept term per trial accounts for different baseline (control group) risks and the clustering of participants in each trial. Alternative specifications of equation (16) could be chosen, including a fixed rather than random treatment effect, and a random rather than stratified trial-specific intercept term. Adjustment terms could also be included, and then the analyst must decide whether a common or distinct adjustment term is specified per trial. Estimation is typically performed using ML via a numerical approach such as Gaussian quadrature.

For time-to-event outcomes, a one-stage Cox regression model can be used [9,12,26,27]. However, unlike the continuous and binary outcome frameworks, there is the added complexity of whether to specify unique or proportional baseline hazards for the trials. For example, assuming a random treatment effect, one might specify proportional baseline hazards by adjusting for a study-specific effect:

$$h_{ij}(t)=h_0(t)\exp\left(\alpha_{0i}+\theta_i x_{ij}\right) \tag{17}$$

$$\theta_i=\theta+u_i$$

$$u_i \sim N\left(0,\tau^2\right)$$

Here, $h_0(t)$ is the baseline hazard function in the reference trial (say $i=1$), and α_{0i} is the proportional effect on the baseline hazard function due to the i^{th} trial (with α_{01} constrained to be zero). Alternatively, we can specify unique baseline hazard functions for each trial (i.e. stratify by trial) by

$$h_{ij}(t)=h_{0i}(t)\exp\left(\theta_i x_{ij}\right) \tag{18}$$

$$\theta_i=\theta+u_i$$

$$u_i \sim N\left(0,\tau^2\right)$$

where $h_{0i}(t)$ is the unique baseline hazard function in the i^{th} trial. By avoiding the assumption of proportional baseline hazards, equation (18) makes less assumptions than equation (17). As for linear and logistic models, the flexibility of the one-stage approach also allows the analyst to make other modifications if desired, including a fixed rather than random treatment effect, a random rather than stratified trial-specific intercept term (for equation (18)) and the inclusion of common adjustment terms.

Estimation of such one-stage Cox models can be via ML or REML [9], which typically avoid estimation of the baseline hazard itself [1,28–30]. Crowther *et al.* also show how to fit the models in a Poisson regression framework using ML estimation [26]. If the baseline hazards are themselves of interest (for example, for developing a prognostic model [31,32]), a one-stage parametric survival model might be preferred; for example, Crowther *et al.* use Gauss–Hermite quadrature to estimate a one-stage IPD meta-analysis model akin to equation (18), but using a flexible parametric approach where the baseline cumulative hazard is modelled via restricted cubic splines [33].

2.3. Statistical software

Debray *et al.* provide an excellent overview of statistical software for IPD meta-analysis [1]. They focus mainly on one-stage models and highlight software that fits the necessary generalised linear mixed models, such as SAS [34], Stata [35], R [36] and MLwiN [37]. We additionally wish to highlight the

'ipdforest' module in Stata, which produces a forest plot following a one-stage IPD meta-analysis of a continuous or binary outcome. This gives the one-stage summary result at the bottom of the plot, but with the study-specific estimates as derived from the first stage of a two-stage analysis [38]. Debray *et al.* also do not mention the excellent 'ipdmetan' package [39]; a Stata module that automatically operates the first and second stages of the two-stage approach. This can handle any outcome type (including continuous, binary or time-to-event outcomes) in the first stage and offer a wide variety of estimation options for the second stage, whilst also producing detailed summary results, heterogeneity statistics and a forest plot. Several modules and software, such as 'metan' [40] in Stata, 'meta' and 'metafor' in R [36,41] and RevMan [18] can perform the first stage of a two-stage meta-analysis if the user provides the 2×2 contingency tables. If users are happy to implement the first stage themselves, then there are also a plethora of other modules and software that can implement the second stage directly, including 'metan' [40], 'metaan' [42], 'metareg' [43] and 'mvmeta' [44] in Stata [45], 'meta' and 'metafor' in R [36,41], SAS Proc Mixed [46,47] and RevMan [18]. However, the available estimation methods differ considerably across these.

2.4. One-stage and two-stage approaches often give very similar results

Several authors have investigated the difference between one-stage and two-stage IPD meta-analysis results (for example, see the following [1–3,9,48–50]), either empirically, theoretically or via simulation. Most authors conclude that they give very similar results. For example, for binary outcomes, Stewart *et al.* state,

> *Major benefits of obtaining IPD are accrued prior to analysis and where an IPD review evaluates effectiveness based on sufficient data from randomised controlled trials, one-stage statistical analyses may not add much value to simpler two-stage approaches. Researchers should therefore not be discouraged from undertaking IPD synthesis through lack of advanced statistical support* [2].

Debray *et al.* also investigated any differences in relation to binary outcomes [3] and concluded that generally, the approaches gives similar results, but all have potential estimation challenges. Also, Senn (2010) discussed the theoretical equivalence of one-stage and two-stage likelihood specifications for binary fixed effects meta-analysis without covariates [51]. Similarly, for time-to-event outcomes, Bowden *et al.* conclude that if the aim of a meta-analysis is to estimate the treatment effect under a random effects model, there appears to be only a very small gain in fitting more complex and computationally intensive one-stage models [9]. For continuous outcomes, Mathew and Nordstrom (2010) extend previous theoretical work to show that one-stage and two-stage summary estimates coincide exactly when fixed treatment and fixed intercept terms are used [52–54].

Two recent overviews of IPD meta-analysis also note that one-stage and two-stage approaches usually give very similar results [3,13]. To illustrate this, Tierney *et al.* give a binary outcome example where the effects of anti-platelets on pre-eclampsia in pregnancy from two-stage (relative risk = 0.90, 95% CI = 0.83 to 0.96) and one-stage (relative risk = 0.90, 95% CI = 0.83 to 0.97) analyses were almost identical [2,13].

Figure 1(a) gives a continuous outcome example, from a random effects IPD meta-analysis of 10 randomised trials to evaluate the effect of anti-hypertensive treatment on systolic blood pressure [55]. Again, the summary results are almost identical when using either the one-stage (equation (14)) or two-stage (equation (1) followed by equation (9)) approaches using REML estimation. Figure 1(b) gives a time-to-event example, showing a fixed-effect IPD meta-analysis of the same 10 hypertension trials [26]. Again, the summary treatment effect on cardiovascular disease is very similar when using either a two-stage approach (equation (3) followed by (5)), or a stratified one-stage approach with a unique baseline hazard per study (equation (18)), via ML estimation.

3. Key reasons why meta-analysis results may differ for the one-stage and two-stage approaches

The examples in Section 2.4 echo the wide-spread belief that one-stage and two-stage results are usually very similar. However, differences can arise [53], and sometimes, these may even be large with discrepant statistical or clinical significance [3].

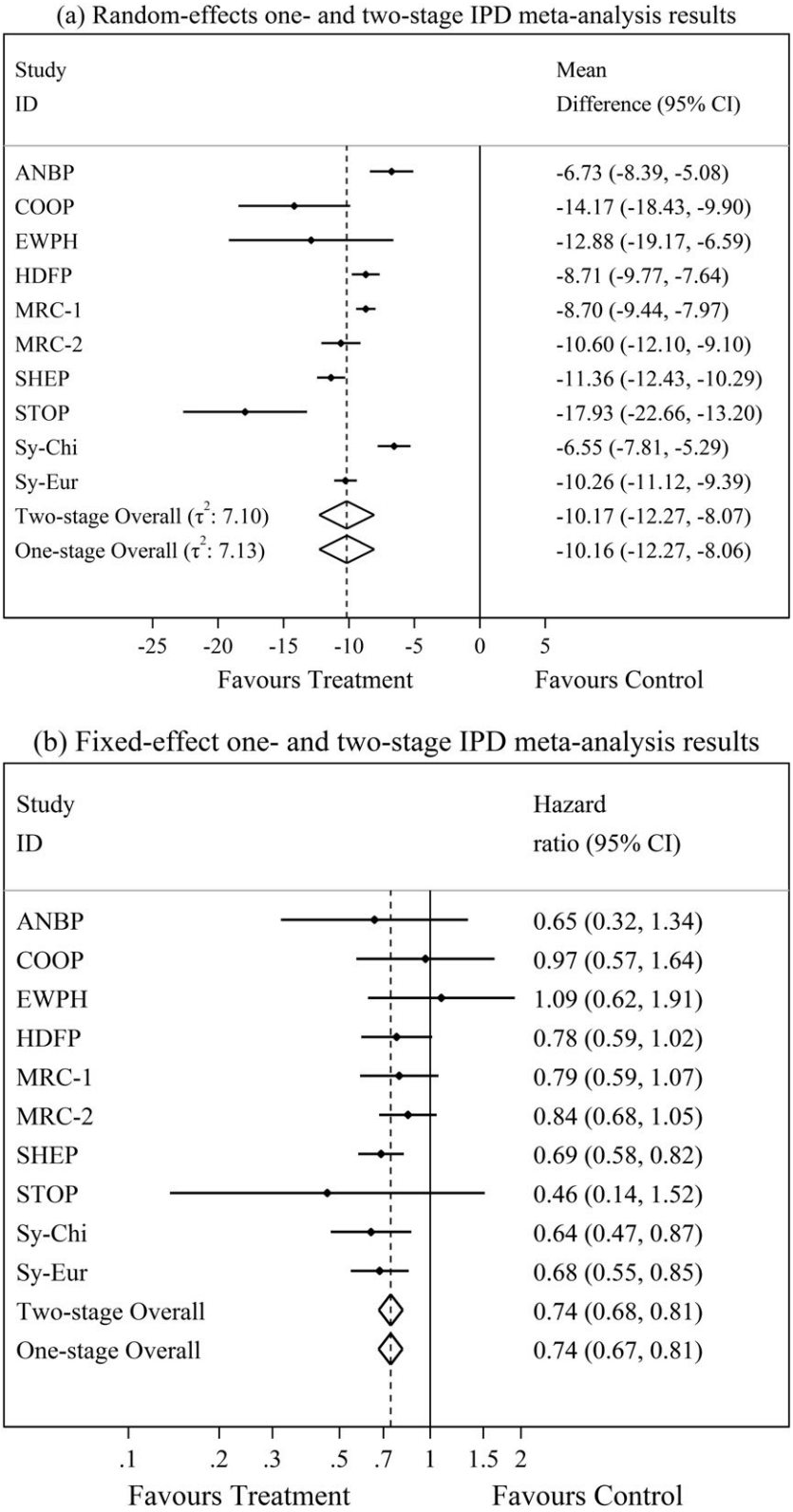

Figure 1. One-stage and two-stage IPD meta-analysis summary treatment effect results for (a) mean difference for final systolic blood pressure adjusting for baseline systolic blood pressure and (b) hazard ratio for cardiovascular disease. IPD, individual participant data; CI, confidence interval.

To aid researchers facing this situation, we now describe 10 key reasons why such differences may arise even when the same IPD are used for both one-stage and two-stage approaches. These relate to different modelling assumptions, parameter specifications and estimation methods. Although some are

perhaps obvious, the majority are subtle, and we suspect that most researchers are unaware of their potential impact. We illustrate each reason with results from previous IPD meta-analyses. In all examples, we ensure that exactly the same IPD are used for both one-stage and two-stage analyses, and thus, any differences cannot be due to discrepant numbers of studies, patients or follow-up times. Some of the issues are inter-related, but we highlight them separately to make them explicit.

3.1. Reason I: exact one-stage likelihood versus approximate two-stage likelihoods

One-stage and two-stage methods may yield different summary results when the second stage of the two-stage method assumes that study treatment effect estimates ($\hat{\theta}_i$) have a normal sampling distribution and that their variances ($Var\left(\hat{\theta}_i\right)$) are known [3,11,56,57]. This first assumption is based on the central limit theorem, and the second assumes that the variance is estimated with reasonable accuracy. These assumptions depend on a combination of the total number of participants in each trial, the number of participants in each treatment group in each trial and the number of events/non-events in each group in each trial. Therefore, both assumptions are unlikely to be appropriate for all outcome types when some or many of the included studies are small (<30 participants). However, if the small trials do not contribute greatly to the pooled effect (i.e. the weight of these trials is low), then this is a less concerning issue. The assumptions are more generally unreliable for binary and time-to-event outcomes when outcomes are rare (or extremely common), as the number of outcomes (and non-outcomes) then drives the study-specific estimates and their variances, not just the total participants. In contrast, a one-stage IPD meta-analysis model, such as a logistic regression model in equation (16), directly models the actual distribution of the IPD (e.g. binomial for logistic regression) and avoids making any assumptions about the distribution of the treatment effect estimates in each study. It thereby provides a more exact likelihood specification from which to make inferences [11,58]. The one-stage approach may additionally, but perhaps more critically, be different from the two-stage approach if there are zero event counts for binary outcomes because the two stage method typically requires a continuity correction (such as +0.5 to cells in the available two-by-two tables [59–61]), whereas this is not necessary with the one-stage logistic regression model [11]. The issue that effect estimates are not normally distributed is possibly less of a concern than the issue of zero cells and subsequent continuity corrections, as the latter actually influence the magnitude of effect estimates and their estimated variances, which may introduce bias that then feeds in to the second stage.

We illustrate this issue with an example from Hamza *et al.* [11] who re-analysed an IPD meta-analysis of nine studies that assessed the operating characteristics of positron emission tomography in the diagnosis of Alzheimer's disease. There were small numbers of patients in each study (range: 19–50) with a total of 254 diseased patients and 218 true positive test results, and the aim was to summarise the sensitivity of the positron emission tomography test. Although sensitivity and specificity are often analysed together, here, we just consider sensitivity alone; indeed, a joint analysis often makes little difference anyway [62,63]. However, three studies contained zero false positives (sensitivity = 100%), and therefore, to perform the two-stage approach, Hamza *et al.* used continuity corrections (+0.5 added to each cell) in the first stage to derive logit-sensitivity estimates and their variances in each study (equation (2)). In contrast, the one-stage approach does not require such assumptions as it models the binomial distribution of the IPD directly (here, a logistic regression model with just an intercept term denoting the logit-sensitivity and an associated random effect). This leads to large differences in the summary logit-sensitivity estimate between the one-stage (ML estimate: 2.20) and two-stage (ML estimate: 1.71) approaches, relating to a summary sensitivity of 0.90 and 0.85, respectively. The estimated τ^2 was also different (0.97 in one-stage analysis, 0.37 in two-stage analysis), which may be especially important for making predictive inferences about the sensitivity in a new population [64]. Riley *et al.* propose how to derive percentage study weights for this example, and these are shown in Figure 2, with clear differences in the one-stage and two-stage weights (Riley *et al.*, *submitted*).

3.2. Reason II: likelihood-based one-stage versus alternative weighting schemes in two-stage

As described in Reason I, stage 2 in a two-stage approach is problematic when outcomes are rare and studies are small and may even be problematic when some studies have unbalanced sample size in the treatment and control group numbers [10,59,60]. To address this, alternative two-stage approaches have been proposed, which use a different weighting scheme to the inverse variance method, such as the Peto method [65] and the Mantel–Haenszel method [66], which also avoid the use of continuity corrections when there are zero cells. The Peto method is considered to work well when intervention effects are small, event risks

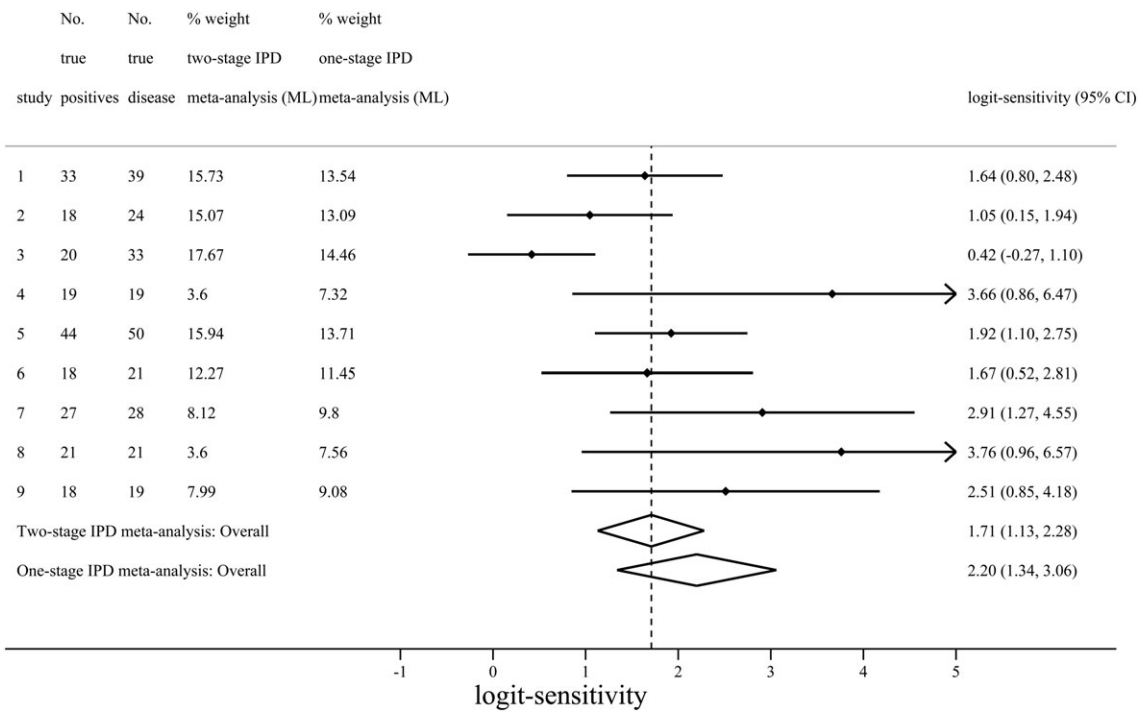

Figure 2. Forest plot of the one-stage and two-stage meta-analysis results for the sensitivity of the PET test for diagnosis of Alzheimer's disease. PET, positron emission tomography; IPD, individual participant data; ML, maximum likelihood.

are <1%, and when there are balanced experimental and control group sizes within trials [10,60]. The Mantel–Haenszel is often preferred when event risk is >1%, and there are unbalanced data [10].

However, in the context of this article, it is important to emphasise that even these alternative two-stage methods may give different results to a one-stage meta-analysis that uses the binomial likelihood, such as a logistic regression. We illustrate this issue with a previous IPD meta-analysis dataset of three trials that investigated whether erythema is a risk factor for deep vein thrombosis (DVT), as shown in Debray *et al.* [3]. In terms of a two-stage approach, the Mantel–Haenszel method might be preferred in this situation because the event rates are not small (10–29%), and sample sizes are not balanced in the control and treatment groups in all three trials. Assuming a fixed treatment effect, the summary results from the Mantel–Haenszel and Peto methods are compared with a one-stage logistic regression model (equation (16) with a fixed treatment effect) in Table I.

Although summary results are qualitatively similar, the summary OR estimates are slightly lower for the two-stage methods compared with the one-stage approach. Furthermore, only the one-stage approach suggests a statistically significant increase in the odds of DVT if there is erythema compared with without. These differences are a consequence of the different weighting schemes employed by the methods.

3.3. Reason III: clustering and choice of specification for the intercept

A two-stage IPD meta-analysis automatically accounts for clustering of patients by trial by analysing the data from each trial separately in the first stage. However, a one-stage IPD meta-analysis models all data

Table I. Summary results from a fixed-effect IPD meta-analysis of the erythema data using two-stage or one-stage approaches.

Approach	Method	Summary OR (95% CI), *p*-value
Two-stage	Mantel–Haenszel	1.277 (0.980 to 1.665), 0.070
Two-stage	Peto	1.285 (0.980 to 1.686), 0.069
One-stage	Logistic regression	1.352 (1.031 to 1.771), 0.029

IPD, individual participant data; OR, odds ratio; CI, confidence interval.

Table II. Fixed-effect one-stage and two-stage approaches to illustrate the effect of clustering in the nicotine gum dataset [68].

Approach	Clustering	Summary OR (95% CI OR)
Two-stage	Accounting for clustering	1.769 (1.257 to 2.488)
One-stage	Ignoring clustering	1.398 (1.020 to 1.916)
One-stage	Accounting for clustering	1.802 (1.290 to 2.517)

OR, odds ratio; CI, confidence interval; τ^2, between-study heterogeneity.

simultaneously, and therefore, the analyst must account for the clustering of patients to avoid misleading effect estimates and conclusions. More specifically, for logistic regression models, one expects downwardly biassed results when ignoring the clustering due to non-collapsibility of the OR [8]. As discussed in Section 2.2, the clustering can be accounted for by stratifying the analysis by trial (i.e. estimating a separate intercept or baseline hazard for each trial) or assuming that the intercept (baseline hazard) is randomly drawn from some distribution. However, there is repeated evidence that some researchers are ignoring the clustering of patients within trials in the one-stage IPD meta-analyses and are therefore analysing data as if it were coming from a single study [7,8,67]. This is another key reason why one-stage and two-stage results may differ.

We illustrate this issue using a previous IPD meta-analysis dataset of two randomised controlled trials (total patients = 1620) that investigated whether nicotine gum increased the odds of smoking cessation [68]. All parameter estimates are obtained using ML for consistency and assume a fixed treatment effect (Table II). The two-stage approach used equation (2) followed by equation (5), and thus, the first-stage automatically assumed a different intercept term per trial. The one-stage approach that accounted for clustering used equation (16) with a separate intercept per trial and a fixed treatment effect; however, the one-stage approach that ignored clustering used the same equation, but with just a single intercept term.

The results highlight the importance of accounting for clustering. The summary OR estimate and CI from the one-stage approach ignoring clustering only just suggest that there is a statistically significant association between nicotine gum use and smoking cessation. However, in contrast, the summary OR estimate from both the two-stage and one-stage approaches that account for clustering are substantially higher, and there is much stronger statistical evidence that nicotine gum is beneficial.

Thus, clustering must be accounted for; yet, even the choice of *how* the clustering is accounted for may cause differences. For example, for continuous outcomes, Matthew and Nordstrom (2010) suggest that a one-stage approach with a random intercept term may be slightly more precise than a two-stage IPD meta-analysis (which has a distinct intercept term per study) [53]. For time-to-event outcomes, there are even more options. In a two-stage IPD meta-analysis, the baseline hazard is uniquely estimated in each trial separately, and thus, there is no assumption about how the shape or magnitude of the baseline hazard is related across trials. In a one-stage meta-analysis, the analyst could also assume this (equation (18)) but, due to the greater flexibility of the approach, could alternatively assume that the baseline hazards are distinct but proportional to each other (as in equation (17)) or assume a random effect (frailty) for the baseline hazard; for example, the 'shared' option within the 'stcox' module within Stata treats the frailties as being gamma distributed (mainly for computational convenience) [35]. In our experience, this decision usually only leads to small differences in the summary treatment effect. For example, in an IPD meta-analysis of 1225 patients from five clinical trials in epilepsy that investigated the effect of treatment on time to remission [29], a one-stage IPD meta-analysis Cox model with proportional baseline hazards (equation (17)) gave a summary hazard ratio of 0.89 (95% CI: 0.77 to 1.03, $p = 0.115$). The summary hazard ratio was 0.90 (95% CI: 0.78 to 1.03, $p = 0.130$) in a Cox model with unique baseline hazard functions for each trial, and 0.90 (95% CI: 0.78 to 1.04, $p = 0.153$) assuming a frailty model. The differences are qualitatively small; nevertheless, changes in the p-value are still apparent and may be more pertinent in other examples.

3.4. Reason IV: choice of specification for any adjustment terms

Similar to the choice of intercept (baseline hazard) specification, in a one-stage IPD meta-analysis, the analyst can also adopt different specifications for any adjustment terms (such as adjustment for baseline score in an ANCOVA, equation (13)). A two-stage approach automatically assumes a different effect of each adjustment factor in each trial, as it analyses each trial separately in the first stage (equation (1)). A

one-stage approach can replicate this by stratifying the effect by study (as in equation (13)), but alternatively, the analyst could assume that the effects were random or even fixed. The fixed assumption is likely an over-simplification, but our experience suggests that it is often adopted in practice, perhaps unknowingly.

The choice of specification of adjustment factors may lead to differences between the one-stage and two-stage approaches, particularly if the effect of the adjustment factor is heterogeneous across the trials. When there are multiple random effects, the specification of their covariance matrix (e.g. unstructured or independent) may also be influential.

We illustrate this using an IPD meta-analysis by Riley *et al.* [55], which used the 10 hypertension trials introduced in Section 2.4. Here, instead of focussing on the treatment effect, the objective was to examine whether smoking is a prognostic factor for high systolic blood pressure at follow-up, after adjusting for baseline blood pressure and treatment group. Utilising a one-stage linear regression model (similar to equation (14) with an additional term for smoking), we compare REML results for the smoking effect when including baseline blood pressure and treatment group as (a) fixed, (b) random with an unstructured covariance structure, (c) random with an independent covariance structure or (d) distinct adjustment terms for each trial. All models assumed a random effect for smoking. The competing two-stage approach used a linear regression model (similar to equation (1) with additional term for smoking and parameters estimated) in each trial separately with smoking status, baseline blood pressure and treatment group, and then a random effects meta-analysis model (equation (9)) using REML to synthesise the smoking effect in the second stage.

Table III shows that the estimate and 95% CI of the summary effect of smoking differs depending on how the adjustment terms were accounted for, although all approaches suggest smoking is a prognostic factor. The one-stage approach with distinct adjustment terms is very similar to the two-stage approach, as expected as the two-stage approach also makes the same assumption. However, the one-stage approach with fixed or random adjustment terms (especially with an unstructured covariance matrix) gives a lower prognostic effect and a wider CI.

3.5. *Reason V: choice of specification for the residual variances*

For continuous outcomes, in the two-stage approach, the residual variances are automatically distinct in each trial, as separate linear models are applied to each trial (equation (1)). In a one-stage IPD meta-analysis, such as in equation (14), the residual variance can also be allowed to vary by trial, such that $e_{ij} \sim N(0, \sigma_i^2)$. However, it is possible to simplify this and make an additional assumption that all trials have the same residual variance, such that $e_{ij} \sim N(0, \sigma^2)$. In our experience, many one-stage analysts are not aware that they even make this assumption, but it can cause summary results to differ to those from the two-stage approach. This is illustrated using the IPD meta-analysis of hypertension trials again (introduced in Section 2.4).

When we assume different residual variances for each trial, the one-stage and two-stage results are almost identical (Table IV). However, the one-stage analysis that assumes the same residual variance in each trial gives a slightly larger summary result, a larger estimate of τ^2 and wider CIs. The larger $\hat{\tau}^2$ is caused by the mis-specified residual variances. Although clinical conclusions are qualitatively the same, this illustrates how the specification of the residual variances can potentially cause differences in meta-analysis estimates. For multivariate responses (e.g. multiple time points), a related issue is

Table III. One-stage and two-stage REML results for the effect of smoking on blood pressure at follow-up after different specifications of the adjustment for baseline blood pressure and treatment group.

Approach	Model specification in regard to adjustment factors	Summary mean difference (smokers versus non-smokers), 95% CI	$\hat{\tau}^2$
Two-stage	Distinct per trial	1.763 (1.146 to 2.380)	0.227
One-stage	Fixed	1.689 (0.951 to 2.426)	0.271
One-stage	Random (correlated*)	1.523 (0.731 to 2.316)	0.510
One-stage	Random (independent*)	1.744 (1.027 to 2.461)	0.235
One-stage	Distinct per trial	1.756 (1.043 to 2.469)	0.229

*Unstructured or independent covariance structure for all included random-effects.
REML, restricted maximum likelihood; CI, confidence interval; τ^2, between-study heterogeneity in the smoking effect.

Table IV. One-stage and two-stage REML results for the effect of hypertension treatment on systolic blood pressure, after different specifications of the residual variances.

Approach	Assumption of residual variances in the trials	Summary mean difference (95% CI)	$\hat{\tau}^2$
One-stage	Same in each trial	−10.34 (−12.55 to −8.13)	8.19
One-stage	Distinct per trial	−10.16 (−12.27 to −8.06)	7.13
Two-stage	Distinct per trial	−10.17 (−12.27 to −8.07)	7.10

REML, restricted maximum likelihood; CI, confidence interval; τ^2, between-study heterogeneity in the treatment effect.

whether the analyses allow for separate or common residual variances for each endpoint, or separate or common correlations between endpoints (Reason IX).

3.6. Reason VI: choice of fixed-effect or random effects for the parameter of interest

Perhaps the most obvious explanation for any differences in one-stage or two-stage approaches is that, for the parameter of interest, the meta-analyst may be assuming a fixed-effect in one approach and random effects in the other. For example, using the hypertension dataset again, we fitted an ANCOVA model to evaluate the effect of hypertension treatment on blood pressure, with a fixed treatment effect in a one-stage approach (equation (13)), and a random treatment effect in a two-stage approach (equation (1) followed by equation (9)). The summary treatment effect in the one-stage approach was −9.31 (95% CI: −9.70 to −8.92) compared with −10.17 (95% CI: −11.99 to −8.35) in the two-stage approach, with the lower effect in the one-stage approach due to the fixed-effect assumption. Sometimes, a fixed-effect assumption may be enforced when using the one-stage approach, due to convergence problems (especially when the outcome is rare [3]) or computational time (especially for time-to-event outcomes with large datasets [12]), thus preventing a direct comparison with a two-stage random effects analysis.

3.7. Reason VII: different estimation method for τ^2

There are various estimation methods for the between-study heterogeneity parameter, τ^2, as mentioned in Section 2. However, some are only available in a two-stage approach and others in only the one-stage approach. Furthermore, some estimation methods may fail to converge because of small numbers of studies in the meta-analysis or large within-study variances [3]. Therefore, differences in one-stage and two-stage meta-analysis results may be a consequence of different estimation methods. For example, it is well known that the ML method tends to underestimate the between-study heterogeneity [69–74] and that REML is preferred in comparison. However, for non-continuous outcomes, one-stage models typically use ML estimation as it is usually the only option. Furthermore, the MoM estimator of DerSimonian and Laird is still the most common estimation method applied in the two-stage approach and is the only option available in RevMan, for example.

We now compare the IPD meta-analysis results for the 10 hypertension trials once more, according to different estimation methods. For the two-stage approach, we fitted model (1) and then adopted a random effects model (9), which was estimated using either MoM, ML or REML. In a one-stage model, ANCOVA model (14) was fitted, and the model parameters are estimated using REML or ML. The results are shown in Table V.

Table V. Summary treatment effect results for the hypertension data to illustrate the differences in summary results according to the estimation method.

Approach	Estimation method	Summary mean difference (95% CI)	$\hat{\tau}^2$
Two-stage	MoM	−9.85 (−11.13 to −8.57)	3.07
Two-stage	ML	−10.10 (−12.03 to −8.16)	5.84
Two-stage	REML	−10.17 (−12.27 to −8.07)	7.10
One-stage	ML	−10.03 (−11.83 to −8.23)	4.94
One-stage	REML	−10.16 (−12.27 to −8.06)	7.13

CI, confidence interval; MoM, method-of-moments; ML, maximum likelihood; REML, restricted maximum likelihood; τ^2, between-study heterogeneity in the treatment effect.

Although summary treatment effect estimates are broadly similar regardless of the estimation approach, it is notable that the estimate of the between-study variance is substantially affected. For example, $\hat{\tau}^2$ is much smaller with the MoM estimator in a two-stage approach ($\hat{\tau}^2 = 3.07$) compared with the estimate from a one-stage approach via REML ($\hat{\tau}^2 = 7.13$). Larger $\hat{\tau}^2$ estimates lead to wider 95% CIs for the summary treatment effect and would have even more impact upon predictive inferences, such as 95% prediction intervals [75]. If the same estimation method is used for a one-stage or two-stage approach, the summary estimates and 95% CI are very similar. The choice of best estimator for $\hat{\tau}^2$ is an ongoing issue in the meta-analysis field [21].

3.8. Reason VIII: derivation of CIs

Not only are there competing estimation methods for fitting the random effects meta-analysis model, but there are also competing methods for the derivation of 95% CIs for the summary effect *post*-estimation. Standard CIs based on the normal distribution (calculated using the summary estimate $\pm 1.96 \times$ s.e. of the summary estimate) are often too narrow, especially because $\hat{\tau}^2$ is often underestimated and no account is taken of the additional uncertainty in $\hat{\tau}^2$ [22,76]. Therefore, alternative methods have been proposed for deriving 95% CIs in the two-stage approach, such as the Hartung–Knapp–Sidik-Jonkman (HKSJ) modification to the variance of the summary estimate [77–81], combined with the t-distribution with $k-1$ (k = number of studies) degrees of freedom (rather than 1.96 from the standard normal distribution). In a one-stage approach, the standard approach is again to use the normal distribution from which large sample inferences are based. However, there are also various methods for small-sample inference based on the t-distribution, also known as denominator-degrees-of-freedom adjustments, including Satterthwaite and Kenward-Roger [82].

Therefore, differences in 95% CIs from one-stage and two-stage approaches may simply be due to the different derivation methods embedded in software packages (e.g. the standard approach is the default in the Stata package 'metan' [40], but the HKSJ is the default in the Stata package 'metareg' [43]). To illustrate this, we return again to the hypertension example. Figure 1 already showed how standard CIs for the summary treatment effect were almost identical for one-stage and two-stage analyses (-12.27 to -8.07). However, the 95% CI is slightly wider (-12.44 to -7.90) in a two-stage approach using the HKSJ variance estimator and t-distribution.

3.9. Reason IX: accounting for correlation amongst parameters

A one-stage approach automatically accounts for all correlation amongst included parameters during model estimation. However, a standard two-stage approach (as outlined in Section 2) does not account for such correlations unless a multivariate meta-analysis model is used to jointly synthesise all parameters, whilst accounting for their within-study, and possibly between-study, correlations [55,83]. This could lead to the standard two-stage approach having reduced efficiency and even a change in the summary estimates. Sometimes, even if desired, the multivariate model may not be estimable, especially when there are many parameters to jointly synthesise, because of the lack of information to estimate the between-study variance–covariance matrix (particularly the unstructured option).

Consider a meta-analysis of longitudinal outcomes, where each patient in a trial provides an outcome value at each of multiple time points during follow-up. These patients responses, and indeed subsequent parameter estimates (such as treatment effects at multiple time points), will be correlated, and ignoring this in the meta-analysis is inadvisable [84,85]. Usually, there are missing data in such situations, such that not all patients provide all time points; furthermore, not all studies will measure the same time points. In a one-stage analysis of all time points, under a missing at random assumption, even patients and studies with missing time points will still contribute towards the estimation of parameters at *all* time points, by utilising the correlation amongst them. However, a standard univariate two-stage analysis will typically take each time point separately, and thus exclude patients or studies at particular time points if they did not provide data.

An example of this was illustrated by Jones *et al.* using an IPD meta-analysis of longitudinal data from five trials that investigated Selegiline versus placebo for the treatment of Alzheimer's disease [84,86]. The outcome of interest over time was the mini-mental state examination, which is a measure of cognitive function. Patients were repeatedly followed up over time at 1, 2, 4, 6, and 12 months; however, the set of time points available in each trial was different, and not all patients provided all time points in the same trial. The one-stage approach applied a linear regression model including a separate

Table VI. ML results of one-stage and two-stage meta-analysis of MMSE longitudinal data: estimates (standard error) of difference between Selegiline and Placebo.

Time point	One-stage model	Standard two-stage approach (each time point assumed independent)	Multivariate two-stage approach (correlation amongst time points accounted for)
1 month	0.31 (0.47)	0.43 (0.54)	0.30 (0.47)
2 months	−0.48 (0.62)	−0.84 (0.97)	−0.47 (0.59)
4 months	0.34 (0.48)	0.75 (0.57)	0.33 (0.47)
6 months	0.20 (0.49)	0.31 (0.50)	0.19 (0.48)
9 months	0.35 (0.53)	0.69 (0.63)	0.34 (0.52)
12 months	−0.02 (0.56)	0.29 (0.66)	−0.03 (0.55)

ML, maximum likelihood; MMSE, mini-mental state examination.

intercept, baseline adjustment term, and residual variance for each trial, and a fixed treatment effect assumed at each distinct time point. This approach accounts for the correlation amongst patient responses with correlated residuals, and handles missing time point data for participants and trials via a missing at random assumption. The two-stage approach firstly analysed each trial separately accounting for correlated patient responses within each trial, to produce a treatment effect at each time point reported; however, in the second stage, each time point was analysed separately (independently), ignoring correlation amongst time points, and thus studies only contributed towards the time points they reported.

There are large differences in the summary treatment effects and their standard errors for the two-stage and one-stage approaches (Table VI). For example, at time point 4 months, the treatment effect is 0.34 (s.e. = 0.48) for the one-stage model, compared with 0.75 (s.e. = 0.57) for the two-stage model. Differences are due to the one-stage approach utilising the correlation amongst time points, which allows the 'borrowing of strength' from the available data to inform the missing data and increases precision [87]. When a multivariate model is used in the second stage of the two-stage approach to account for the correlation, the results become almost identical to those from the one-stage method [84].

3.10. Reason X: ecological bias for treatment covariate interactions

When one wishes to model treatment–covariate interactions, it is important to avoid ecological bias, which occurs when across-study associations (between mean covariate values and treatment effects across trials) do not reflect the true within-study relationships (between individual covariate values and individual response to treatment) [4,88]. For example, when looking at the interaction between sex and treatment effect, across-study associations would allow studies with only males (or only females) to be included, despite containing no within-study differences between males and females. We note that a similar problem would occur if one used data from single arm trials in a meta-analysis to inform the estimate of overall treatment effect [51].

A two-stage approach to the estimation of treatment–covariate interactions automatically avoids ecological bias by fitting a separate model (such as equation (4)) in each trial to obtain within-study interaction estimates, which are then synthesised in the second stage. Thus any across-study associations are ignored. This can be replicated in a one-stage approach by separating the within-study and across-study interaction effects to avoid ecological bias; this is achieved by centring the patient-level covariate about its mean [88–90] (refer to Appendix section). If the within-study and across-study interaction effects are not separated in the model, then the resulting interaction effect is an amalgamation of the within-study and the across-study associations, which does not necessarily reflect the patient-level interaction and may lead to misleading conclusions [89]. Unfortunately, this issue is not well

Table VII. One-stage and two-stage IPD meta-analysis results for treatment–age interactions in the hypertension dataset according to within-trial, across-trial and amalgamated interactions.

Interaction covariate	One-stage			Two-stage
	Amalgamated interaction (95% CI), *p*-value	Within-trial interaction (95% CI), *p*-value	Across-trial interaction (95% CI), *p*-value	Within-trial interaction (95% CI), *p*-value
Age	−0.067 (−0.094 to −0.040), <0.001	−0.050 (−0.116 to 0.017), 0.142	−0.071 (−0.100 to −0.041), <0.001	−0.049 (−0.115 to 0.017), 0.142

IPD, individual participant data; CI, confidence interval.

recognised, and in our experience, most researchers unknowingly fit the amalgamated interaction term in their one-stage models.

The approaches are illustrated by the hypertension dataset again where the outcome is blood pressure and the summary parameter of interest is the patient-level interaction of treatment effect with age. Riley *et al.* fitted one-stage ANCOVA models that either amalgamated the interactions or separated them out [89]. In the two-stage approach, equation (1) was fitted with age and an interaction with age and treatment. A fixed interaction term was assumed for both one-stage and two-stage models, and results obtained using REML are shown in Table VII.

The one-stage approach with the amalgamated interaction estimate suggests a statistically significant association between age and treatment effect (95% CI: −0.094 to −0.040). However, when the one-stage analysis correctly separates the within-trial and across-trial interactions, the summary within-trial interaction is no longer statistically significant (95% CI: −0.116 to 0.017), and gives results almost identical to the two-stage approach. Therefore, the specification of within-trial and across-trial interactions in a one-stage analysis can lead to differences with the two-stage approach and can influence statistical and clinical conclusions.

4. Discussion

We have outlined the key framework for one-stage and two-stage IPD meta-analysis models. Previous authors have compared one-stage and two-stage approaches through theoretical consideration [52–54], simulation [9,49] and empirical examples [2,9,48,49], and found that largely the two approaches produce similar results. Although we agree that this will generally be the case, Section 3 described 10 reasons why one-stage and two-stage results may differ to help the user to understand why differences may arise in their own IPD meta-analyses and illustrated these using a range of examples, including both fixed-effect and random effects models. Although in some examples the differences were small or qualitatively unimportant, in others, the clinical and/or statistical conclusions were affected.

Most differences between one-stage and two-stage approaches occur because of different modelling assumptions, including the specification of the likelihood and included parameters, the choice of fixed or random effects and the utilisation of correlation. Choosing a different estimation procedure may also lead to important differences, especially in the random effects setting. Another way of understanding this is that the different assumptions, parameter specifications and estimation methods lead to different percentage study weights in the one-stage and a two-stage meta-analysis (Riley *et al. submitted*), and therefore, summary results can differ because of a change in weighting. However, when the same assumptions are made in both approaches (and these assumptions are plausible), and the same estimation method is used, the resulting estimates should be very similar, which was demonstrated by the hypertension data example in Section 2.4.

We note, quite importantly, that whilst the one-stage versus two-stage issue is mainly of concern in IPD meta-analysis, one-stage analyses are sometimes possible with published data, and the same similarities and differences apply. For example, for binary outcomes, the analyst can essentially reconstruct the IPD if they extract 2×2 contingency tables from each trial report and then proceed to perform one-stage logistic regression analyses [91] (such as equation (16)), as long as no adjustment for covariates is required.

We focused on summary results and their 95% CIs throughout the article. However, we recognise that other measures may also be of interest such as 95% prediction intervals [92], for which differences may be even more pronounced. We also acknowledge that, although in our experience our 10 reasons are the most likely, the list is not exhaustive, and it may be possible to get differences for other reasons. For example, the meta-analyst may include covariate adjustment for all studies in a one-stage approach, but in the two-stage approach may only include covariate adjustment for a subset of trials. The issue of missing data has also not been explicitly addressed in this paper. In particular, we did not consider differences in one-stage and two-stage results after the imputation of missing participant data, although recognising this is a growing area of interest [1,93–98]. However, in our longitudinal example, we noted that missing outcome data is handled naturally in the one-stage approach under a missing at random assumption, which allows more efficient results by accounting for correlation between time points. Another issue not previously mentioned is how automated selection procedures may lead to differences in one-stage and two-stage results, for example, in regard to identifying the best fitting (non-linear) trend for a continuous predictor in a prognostic model [99]. Abo-Zaid gives an example where the one-stage approach suggests a quadratic trend between age and log-odds of death, whereas the two-stage approach suggests

that a linear trend is preferred in each study and so a pooled linear trend is then obtained [100]. By considering the trends in each study separately, the two-stage approach had lower power than the one-stage approach to identify (at a pre-defined significance level) more complex relationships. However, if a quadratic trend had been forced in each study and then combined, the pooled quadratic trend would have been very similar to that from the one-stage analysis.

So, should researchers choose a one-stage or a two-stage approach? We do not believe there is a 'blanket' answer to this question as it depends on many factors, including the clinical question, the parameter(s) of interest, the desired specification of the model, the desired estimation method, the assumptions willing to be made, the potential for non-convergence and missing data, and the likelihood of small study sizes and rare events. However, we make some specific recommendations in Box 1. In particular, we agree with Debray *et al.* that 'when planning an IPD meta-analysis, the choice and implementation of a one-stage or two-stage method should be pre-specified in the protocol as occasionally they lead to different conclusions' [3]. Moreover, the exact specification of the models should also be pre-specified where possible, for example, in regard to the choice of fixed or random effects, the handling of adjustment factors, how clustering within studies will be accounted for, whether ecological bias will be removed, and (for time-to-event outcomes) the specification of the baseline hazards. Once the analyst has considered the assumptions they want to make and decided that they are plausible in either a one-stage or two-stage framework, then they could adopt either method as they are likely to give very similar results. However, in some situations, the assumptions may be more plausible in the one-stage approach. In particular, for situations of rare outcomes and/or small studies, the one-stage approach is our preferred option, in order to use a more exact likelihood, and to avoid making assumptions about within-study normality and known within-study variances. However, even in this situation, the one-stage approach may suffer from convergence issues, which are difficult to identify in advance. Furthermore, estimation options such as REML may only be available in a two-stage approach, which may help reduce downward bias in estimates of between-study variances in one-stage ML models.

For such reasons, or if the best choice of model specification and/or estimation method are unclear, a sensible strategy is to pre-specify that both one-stage and two-stage analyses will be undertaken, and their results compared with check whether conclusions are the same. If they differ, then the meta-analyst should seek to understand why, and the 10 reasons outlined should help elicit the explanation.

Box 1: Key recommendations for the adoption of one-stage or two-stage IPD meta-analyses.

- State prior to analysis whether a one-stage or two-stage approach will be used for an IPD meta-analysis. If the best choice is unclear, or if estimation difficulties are a concern (e.g. due to rare outcomes), then it is sensible to state that both approaches will be undertaken, and results compared.
- Where both one-stage and two-stage approaches are undertaken, both should be reported in subsequent publications for transparency, and any important differences between them should be explained.
- The estimation method(s) should be specified in advance for both one-stage and two-stage approaches, as the choice can influence the findings and the magnitude of differences between the approaches.
- The model assumptions (e.g. likelihood specification, choice of fixed or random effects) and parameter specifications (e.g. handling of adjustment factors and baseline hazards in each trial) should be pre-specified regardless of whether one-stage or two-stage approaches are used.
- For all outcome types where studies are expected to be small, and in particular, for binary and time-to-event outcomes that are rare (or extremely common), then a one-stage approach is preferred, as it avoids the use of approximate normal sampling distributions, known within-study variances, and continuity corrections that plague the two-stage approach with an inverse variance weighting.
- Any one-stage analysis should account for the clustering of participants with studies. This is best achieved by including a separate intercept (distinct baseline hazard) per trial. One could (if considered plausible) also assume trial intercepts are drawn from some distribution or, for time-to-event outcomes, assume baseline hazards are proportional across trials, but this will usually be unnecessary unless interest lies in the intercept or baseline hazard itself (e.g. for developing a prognostic model).

- In a one-stage approach, it is best to include separate adjustment terms (when included) and separate residual variance terms (for continuous outcomes) for each trial, as this makes less assumptions than a model with common adjustment terms and residual variances. Only when estimation issues arise might it be necessary to move away from this, for which random effects on adjustment terms may be helpful.
- Where random effects models are used, consider methods to derive 95% CIs for the summary effect that account for full uncertainty in the estimated variances in the meta-analysis. For example, for the two-stage approach, the use of methods such as the HKSJ variance estimator and *t*-distribution to estimate CIs might be considered, and for the one-stage approach, the use of Kenward–Roger variance estimator and *t*-distribution may be useful.
- A standard two-stage approach does not automatically account for correlation between the parameters of the regression model estimated in the first stage; this may lead to loss in precision and different summary estimates than the one-stage approach, which automatically accounts for such correlation. This is especially important if there are missing outcome data, for example, missing outcomes at some time points in a meta-analysis of longitudinal data. To account for correlation in a two-stage meta-analysis, a multivariate model is required in the second stage.
- In a one-stage IPD meta-analysis, treatment–covariate interactions should be separated into within-trial and across-trial interactions to avoid ecological bias. This is automatically avoided in a two-stage analysis when within-study interaction estimates are obtained in each trial and then synthesised.

Appendix: Treatment–covariate interactions in a one-stage IPD meta-analysis

A one-stage approach must fit a model that separates the within-study and across-study interaction effects to avoid ecological bias. The one-stage ANCOVA model (equation (14)) can be extended such that

$$Y_{ij} = \alpha_j + \beta_i y_{Bij} + \theta_j \text{treat}_{ij} + \beta_j z_{ij} + \gamma_W \text{treat}_{ij} * \left(z_{ij} - \overline{z_j}\right) + \varepsilon_{ij}$$

$$\theta_j = \alpha + \gamma_A \overline{z_j} + u_j$$

$$\varepsilon_{ij} \sim N\left(0, \sigma_j^2\right)$$

$$u_j \sim N\left(0, \tau^2\right)$$

Here, z_{ij} denotes the covariate of interest, the fixed-effect within-study interaction effect is denoted by γ_W and the across trials interaction is denoted by γ_A (which is also the meta-regression result). The unexplained between-study variance of the treatment effect is given by τ^2, and the covariate is centred about the mean covariate value, $\overline{z_i}$ in each study to ensure that the within-study interaction is separated from the across-study interaction. If the within-study and across-study interaction effects are not separated in the model, then the resulting interaction effect is an amalgamation of the two different interaction effects [89].

Acknowledgements

Richard D Riley was supported by funding from a multivariate meta-analysis grant from the MRC Methodology Research Programme (grant reference number: MR/J013595/1). We would like to thank two anonymous reviewers for their constructive feedback on how to improve the article. Danielle Burke is funded by an NIHR School for Primary Care Research Post-Doctoral Fellowship. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Author contributions

RR and DB developed the research idea. DB undertook all the analyses under the supervision of RR and feedback from JE. DB drafted the paper and revised following comments from RR and JE.

References

1. Debray TP, Moons KG, van Valkenhoef G, Efthimiou O, Hummel N, Groenwold RH, Reitsma JB, GetReal Methods Review Group. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Research Synthesis Methods* 2015; **6**(4):293–309.
2. Stewart GB, Altman DG, Askie LM, Duley L, Simmonds MC, Stewart LA. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PloS One* 2012; **7**(10):e46042.
3. Debray TP, Moons KG, Abo-Zaid GM, Koffijberg H, Riley RD. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PloS One* 2013; **8**(4):e60650.
4. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI, Anti-Lymphocyte Antibody Induction Therapy Study Group. 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**(3):371–387.
5. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010; **340**:c221.
6. Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: a review of current practice. *Contemporary Clinical Trials* 2015; **45**(Pt A):76–83.
7. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials* 2005; **2**(3):209–217.
8. Abo-Zaid G, Guo B, Deeks JJ, Debray TP, Steyerberg EW, Moons KG, Riley RD. Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology* 2013; **66**(8):865–873.
9. Bowden J, Tierney JF, Simmonds M, Copas AJ, Higgins JPT. Individual patient data meta-analysis of time-to-event outcomes: one-stage versus two-stage approaches for estimating the hazard ratio under a random effects model. *Research Synthesis Methods* 2011; **2**(3):150–162.
10. Higgins JPT, Green S (eds). Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration, 2011. Available from: www.cochrane-handbook.org.
11. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology* 2008; **61**(1):41–51.
12. Thompson S, Kaptoge S, White I, Wood A, Perry P, Danesh J, Emerging Risk Factors Collaboration. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *International Journal of Epidemiology* 2010; **39**(5):1345–1359.
13. Tierney JF, Vale C, Riley R, Smith CT, Stewart L, Clarke M, Rovers M. Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. *PLoS Medicine* 2015; **12**(7):e1001855.
14. Bland M. An Introduction to Medical Statistics (4th edn). Oxford University Press: New York, 2015.
15. Zou G. A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* 2004; **159**(7):702–706.
16. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 1991; **10**(11):1665–1677.
17. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
18. The Cochrane Collaboration. Review Manager (RevMan). The Nordic Cochrane Centre: Copenhagen, 5.3. 2014.
19. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Statistical Methods in Medical Research* 2012; **21**(4):409–426.
20. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials* 2007; **28**(2):105–114.
21. Langan D, Higgins JP, Simmonds M. An empirical comparison of heterogeneity variance estimators in 12 894 meta-analyses. *Research Synthesis Methods* 2015; **6**(2):195–205.
22. IntHout J, Ioannidis JP, Borm GF. The Hartung–Knapp–Sidik–Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian–Laird method. *BMC Medical Research Methodology* 2014; **14**:25.
23. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JPT, Langan D, Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods* 2016; **7**(1):55–79.
24. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of Clinical Epidemiology* 2004; **57**(7):683–697.
25. Senn S. The many modes of meta. *Drug Information Journal* 2000; **34**:535–549.
26. Crowther MJ, Riley RD, Staessen JA, Wang J, Gueyffier F, Lambert PC. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Medical Research Methodology* 2012; **12**:34.
27. Tudur Smith C, Williamson PR, Marson AG. An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes. *Journal of Evaluation in Clinical Practice* 2005; **11**(5):468–478.
28. Sargent DJ. A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics* 1998; **54**(4):1486–1497.
29. Tudur Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine* 2005; **24**(9):1307–1319.
30. Michiels S, Baujat B, Mahé C, Sargent DJ, Pignon JP. Random effects survival models gave a better understanding of heterogeneity in individual patient data meta-analyses. *Journal of Clinical Epidemiology* 2005; **58**(3):238–245.
31. 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**(18):3158–3180.

32. Royston P, Parmar MK, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in Medicine* 2004; **23**(6):907–926.
33. Crowther MJ, Look MP, Riley RD. Multilevel mixed effects parametric survival models using adaptive Gauss–Hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Statistics in Medicine* 2014; **33**(22):3844–3858.
34. SAS Institute Inc., Base SAS® 9.4 Procedures Guide. 2011, SAS Institute Inc. Cary, NC.
35. StataCorp, Stata Statistical Software: Release 14. 2015, StataCorp LP. College Station, TX.
36. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2013.
37. Rasbash J, Charlton C, Browne WJ, Healy M, Cameron B. MLwiN. 2.1, 2009, Centre for Multilevel Modelling. University of Bristol.
38. Kontopantelis E, Reeves D. A short guide and a forest plot command (ipdforest) for one-stage meta-analysis. *The Stata Journal* 2013; **13**(3):574–587.
39. Fisher DJ. Two-stage individual participant data meta-analysis and generalized forest plots. *The Stata Journal* 2015; **15**(2):369–396.
40. Harris RJ, Bradburn MJ, Deeks JJ, Harbord RM, Altman DG, Sterne JAC. metan: Fixed- and random-effects meta-analysis. *The Stata Journal* 2008; **8**(1):3–28.
41. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 2010; **36**(3):1–48.
42. Kontopantelis E, Reeves D. metaan: Random-effects meta-analysis. *The Stata Journal* 2010; **10**(3):395–407.
43. Harbord RM, Higgins JP. Meta-regression in Stata. *The Stata Journal* 2008; **8**(4):493–519.
44. White IR. Multivariate random-effects meta-analysis. *The Stata Journal* 2009; **9**(1):40–56.
45. Palmer TM, Sterne JAC (Eds). Meta-Analysis in Stata: An Updated Collection from the Stata Journal (Second edn). Stata Press: College Station, Texas, 2016.
46. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O. SAS® for Mixed Models (Second edn). SAS Institute Inc.: Cary, NC, USA, 2006.
47. 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.
48. Koopman L, van der Heijden GJ, Hoes AW, Grobbee DE, Rovers MM. Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. *International Journal of Technology Assessment in Health Care* 2008; **24**(3):358–361.
49. Tudur Smith C, Williamson PR. A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. *Clinical Trials* 2007; **4**(6):621–630.
50. Steinberg KK, Smith SJ, Stroup DF, Olkin I, Lee NC, Williamson GD, Thacker SB. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *American Journal of Epidemiology* 1997; **145**(10):917–925.
51. Senn S. Hans van Houwelingen and the art of summing up. *Biometrical Journal* 2010; **52**(1):85–94.
52. Mathew T, Nordstrom K. On the equivalence of meta-analysis using literature and using individual patient data. *Biometrics* 1999; **55**(4):1221–1223.
53. Mathew T, Nordstrom K. Comparison of one-step and two-step meta-analysis models using individual patient data. *Biometrical Journal* 2010; **52**(2):271–287.
54. Olkin I, Sampson A. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics* 1998; **54**(1):317–322.
55. Riley RP, MJ, Jackson D, Wardle M, Gueyffier F, Wang J, Staessen JA, White IR. Multivariate meta-analysis using individual participant data. *Research Synthesis Methods* 2015; **6**:157–174.
56. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine* 2010; **29**(29):3046–3067.
57. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2000; **19**(24):3417–3432.
58. Burke DL, Billingham LJ, Girling AJ, Riley RD. Meta-analysis of randomized phase II trials to inform subsequent phase III decisions. *Trials* 2014; **15**:346.
59. 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.
60. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine* 2007; **26**(1):53–77.
61. Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* 2009; **28**(5):721–738.
62. Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Statistical Methods in Medical Research* 2015. doi:10.1177/0962280215592269.
63. Simel DL, Bossuyt PM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. *Journal of Clinical Epidemiology* 2009; **62**(12):1292–1300.
64. Riley RD, Ahmed I, Debray TPA, Willis BH, Noordzij JP, Higgins JPT, Deeks JJ. Summarising and validating test accuracy results across multiple studies for use in clinical practice. *Statistics in Medicine* 2015; **34**(13):2081–2103.
65. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* 1985; **27**(5):335–371.
66. Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Statistics in Medicine* 1990; **9**(3):247–252.
67. 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.

68. Altman DG, Deeks JJ. Meta-analysis, Simpson's paradox, and the number needed to treat. *BMC Medical Research Methodology* 2002; **2**:3.
69. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine* 2007; **26**(9):1964–1981.
70. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**(20):2693–708.
71. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14**(4):395–411.
72. Austin PC. Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *International Journal of Biostatistics* 2010; **6**:1 Article 16. doi:10.2202/1557-4679.1195.
73. Maengseok NL, Y. REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis* 2007; **98**:896–915.
74. Brostrom GH, H. Generalized linear models with clustered data: fixed and random-effects models. *Computational Statistics & Data Analysis* 2011; **55**:3123–3134.
75. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *British Medical Journal* 2011; **342**:964–967.
76. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* 2003; **22**(17):2693–2710.
77. Hartung J. An alternative method for meta-analysis. *Biometrical Journal* 1999; **41**:901–916.
78. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine* 2001; **20**(12):1771–1782.
79. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* 2001; **20**(24):3875–3889.
80. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; **21**(21):3153–3159.
81. Sidik KJ, JN. On constructing confidence intervals for a standardized mean difference in meta-analysis. *Communications in Statistics - Simulation and Computation* 2003; **32**:1191–1203.
82. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**(3):983–997.
83. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Statistics in Medicine* 2011; **30**(20):2481–2498.
84. 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**(1):16–27.
85. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; **172**(4):789–811.
86. Wilcock GK, Birks J, Whitehead A, Evans SJ. The effect of selegiline in the treatment of people with Alzheimer's disease: a meta-analysis of published trials. *International Journal of Geriatric Psychiatry* 2002; **17**(2):175–183.
87. Jackson D, White IR, Price M, Copas J, Riley RD. Borrowing of strength and study weights in multivariate and network meta-analysis. *Statistical Methods in Medical Research* 2015. doi:10.1177/0962280215611702.
88. Fisher DJ, Copas AJ, Tierney JF, Parmar MK. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *Journal of Clinical Epidemiology* 2011; **64**(9):949–967.
89. 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.
90. Simmonds M. Statistical Methodology of Individual Patient Data Meta-analysis. University of Cambridge, 2005.
91. Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Research Synthesis Methods* 2010; **1**(1):2–19.
92. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; **172**(Part 1):137–159.
93. Koopman L, van der Heijden GJ, Grobbee DE, Rovers MM. Comparison of methods of handling missing data in individual patient data meta-analyses: an empirical example on antibiotics in children with acute otitis media. *American Journal of Epidemiology* 2008; **167**(5):540–545.
94. Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. *Statistics in Medicine* 2013; **32**(26):4499–4514.
95. Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine* 2015; **34**(11):1841–1863.
96. White IR, Welton NJ, Wood AM, Ades AE, Higgins JP. Allowing for uncertainty due to missing data in meta-analysis – part 2: hierarchical models. *Statistics in Medicine* 2008; **27**(5):728–745.
97. Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, PROG-IMT Study Group. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine* 2013; **32**(28):4890–4905.
98. Quartagno M, Carpenter JR. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine* 2016; **35**(17):2938–2954.
99. Sauerbrei W, Royston P. A new strategy for meta-analysis of continuous covariates in observational studies. *Statistics in Medicine* 2011; **30**(28):3341–3360.
100. Abo-Zaid, G., Individual patient data meta-analysis of prognostic factor studies. PhD Thesis. 2011, University of Birmingham.

