



# One-stage individual participant data meta-analysis models for continuous and binary outcomes: Comparison of treatment coding options and estimation methods

Richard D. Riley<sup>1</sup> | Amardeep Legha<sup>1</sup> | Dan Jackson<sup>2</sup> | Tim P. Morris<sup>3</sup> |  
Joie Ensor<sup>1</sup> | Kym I.E. Snell<sup>1</sup> | Ian R. White<sup>3</sup> | Danielle L. Burke<sup>1</sup>

<sup>1</sup>Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Keele, UK

<sup>2</sup>Statistical Innovation Group, Advanced Analytics Centre, AstraZeneca, Cambridge, UK

<sup>3</sup>Institute of Clinical Trials and Methodology, MRC Clinical Trials Unit at UCL, London, UK

## Correspondence

Richard D. Riley, Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Keele, Staffordshire ST5 5BG, UK.  
Email: r.riley@keele.ac.uk

## Funding information

Medical Research Council, Grant/Award Number: MC\_UU\_12023/21 and MC\_UU\_12023/29; National Institute for Health Research (NIHR) School for Primary Care Research, Grant/Award Number: Launching fellowships

A one-stage individual participant data (IPD) meta-analysis synthesizes IPD from multiple studies using a general or generalized linear mixed model. This produces summary results (eg, about treatment effect) in a single step, whilst accounting for clustering of participants within studies (via a stratified study intercept, or random study intercepts) and between-study heterogeneity (via random treatment effects). We use simulation to evaluate the performance of restricted maximum likelihood (REML) and maximum likelihood (ML) estimation of one-stage IPD meta-analysis models for synthesizing randomized trials with continuous or binary outcomes. Three key findings are identified. First, for ML or REML estimation of stratified intercept or random intercepts models, a t-distribution based approach generally improves coverage of confidence intervals for the summary treatment effect, compared with a z-based approach. Second, when using ML estimation of a one-stage model with a stratified intercept, the treatment variable should be coded using “study-specific centering” (ie,  $1/0$  minus the study-specific proportion of participants in the treatment group), as this reduces the bias in the between-study variance estimate (compared with  $1/0$  and other coding options). Third, REML estimation reduces downward bias in between-study variance estimates compared with ML estimation, and does not depend on the treatment variable coding; for binary outcomes, this requires REML estimation of the pseudo-likelihood, although this may not be stable in some situations (eg, when data are sparse). Two applied examples are used to illustrate the findings.

## KEYWORDS

estimation methods, individual participant data, IPD, maximum likelihood., meta-analysis, treatment coding

Richard Riley and Amardeep Legha contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

## 1 | INTRODUCTION

An individual participant data (IPD) meta-analysis synthesizes the raw individual-level data from multiple related studies to produce summary results, for example, about the effect of a treatment.<sup>1</sup> A common approach to IPD meta-analysis is a two-stage framework, where the first step analyses the IPD from each study separately to produce aggregate data (such as a treatment effect estimate and its SE), which are then synthesized in the second step using a traditional meta-analysis, such as a random effects model to account for between-study heterogeneity in the (treatment) effect of interest. An alternative approach to IPD meta-analysis is a one-stage framework, in which all studies are analyzed simultaneously using a hierarchical model, such as a generalized linear mixed model or a frailty survival model, to produce summary results in a single step. The one-stage approach has been increasingly used in the past decade.<sup>2</sup>

With a one-stage IPD meta-analysis model, it is essential to account for clustering of participants within studies to avoid misleading conclusions.<sup>3</sup> In particular, in a generalized linear mixed model framework two options to account for this clustering are (i) by using a stratified intercept term in the analysis, which involves estimating a separate intercept for each study; or (ii) by assuming random study intercepts, whereby study intercepts are assumed to be drawn from a distribution (typically a normal distribution). We recently showed through simulation that, when applying a one-stage IPD meta-analysis of randomized controlled trials (RCTs) with a 1:1 treatment:control allocation ratio and a continuous outcome, the meta-analyst can choose either a stratified intercept or random intercepts model when restricted maximum likelihood (REML) is used for estimation.<sup>4</sup> That is, the statistical properties of the estimate of summary treatment effect, the 95% confidence interval for the summary treatment effect, and the estimate of between-study variance of treatment effects are all very similar regardless of whether a stratified intercept or random intercepts model is used. However, when using maximum likelihood (ML) estimation, there was less downward bias in the estimate of between-study variance of the treatment effect when using random intercepts rather than a stratified intercept, due to fewer parameters being estimated.<sup>5,6</sup> Consequently, for ML estimation, the coverage of 95% confidence intervals for the summary treatment effect was better (ie, closer to 95%) when random intercepts rather than a stratified intercept was used.

A recommendation to use random study intercepts, rather than a stratified study intercept, for one-stage IPD meta-analysis models that require ML estimation (eg, for binary outcomes) may be disconcerting to some readers. In particular, the use of random intercepts is often considered inappropriate on philosophical grounds, because it allows across-trial information to inform the control group results, which may compromise randomization within each trial and bias the summary treatment effects. There is the potential for bias in situations when the allocation ratio is associated with the overall mean outcome (risk). In such situations the introduced bias will often be small,<sup>7</sup> but may be substantial in extreme situations. For example, White et al<sup>8</sup> show that when using extreme hypothetical binary outcome data in a network meta-analysis setting, there can be large potential bias in the summary treatment effect when using a random intercept; the summary treatment effect was an odds ratio of 1.35 when the truth was 1. Another issue is that it is usually recommended to allow the random effects on the intercept and treatment effect to be correlated; however, this might then allow the baseline risk to contribute toward the summary treatment effect estimates. As an extreme example, the model could incorporate randomized trials alongside observational studies that only provide information about the control (untreated group); the latter will then contribute (via the correlation) toward the summary treatment effect.

In this article, we aim to build on previous work,<sup>4,5,9-11</sup> and to improve ML estimation of the stratified intercept model so that it is at least comparable to that of the random intercepts model. Specifically, we focus on an IPD meta-analysis of randomized trials, and evaluate whether the coding of the treatment variable is important toward ML estimation properties. Our previous simulations focused on situations where the treatment:control allocation ratio was 1:1 in each study in the meta-analysis, and found a +0.5/−0.5 coding for the treatment variable (instead of 1/0) substantially improved ML estimation performance.<sup>5</sup> Subsequently, we realized that a +0.5/−0.5 coding is the same as using the 1/0 coding minus the proportion in the treatment group (ie, using  $1/0 - 0.5$ ) when the treatment:control allocation ratio is 1:1. This raises the question about whether IPD meta-analysts should always use a +0.5/−0.5 coding of treatment in their one-stage models, or whether the choice should be context specific, especially when the actual allocation ratio is not 1:1. Therefore, in this article our primary aims to assess ML estimation performance of the stratified intercept model when using four treatment coding options:

- the traditional 1/0 coding
- the +0.5/−0.5 coding recommended by Jackson et al<sup>5</sup>
- a coding of 1/0 minus the average proportion of participants in the treatment group in all trials (ie, an “overall centering” approach)

- a coding of 1/0 minus the proportion of participants in the treatment group in that trial (ie, a “study-specific centering” approach)

We evaluate which coding approach gives the smallest bias in the estimates of the summary treatment effect and between-study variance of treatment effects. Two additional objectives are also considered: (a) whether the coverage of 95% confidence intervals for the summary treatment effect are improved by using a t-distribution rather the standard z-based (Wald) approach, and (b) whether REML estimation of the pseudo likelihood leads to better performance than ML estimation of the exact likelihood for one-stage IPD meta-analysis models of binary outcomes.

The structure of this article is as follows. In Section 2, we introduce one-stage IPD meta-analysis models with a stratified intercept, for both continuous and binary outcomes. In Section 3 we evaluate the four treatment coding options in an extensive simulation study, for both continuous and binary outcomes. Section 4 provides real examples and Section 5 concludes with discussion.

## 2 | ONE-STAGE MODEL SPECIFICATIONS AND TREATMENT CODING OPTIONS

In this section we introduce one-stage IPD meta-analysis models for continuous and binary outcomes with either a stratified intercept or random intercepts model, and then define the various treatment coding options.

### 2.1 | Continuous outcomes

Consider that IPD have been obtained from  $i = 1$  to  $K$  RCTs, each of which has a parallel-group design investigating whether a treatment is effective (vs a control or existing treatment) at improving a continuous outcome. The treatment effect then relates to the mean difference (at some follow-up time) in the continuous outcome value between the treatment and control groups. Suppose that there are  $n_i$  participants in trial  $i$ , and that  $Y_{Fij}$  represents the end-of-trial final ( $F$ ) continuous outcome value for participant  $j$  in trial  $i$ . Let the treatment group variable be denoted by  $X_{ij}$ , with coding options (such 1/0 for treatment/control groups) discussed further in Section 2.3.

In this situation, a one-stage IPD meta-analysis with a stratified study intercept (ie, a separate intercept per study to account for within-study clustering of individuals) and assuming between-study heterogeneity of the treatment effect, can be written as follows:

$$\begin{aligned} Y_{Fij} &= \alpha_i + (\theta + u_i)X_{ij} + e_{ij} \\ u_i &\sim N(0, \tau^2) \\ e_{ij} &\sim N(0, \sigma_i^2). \end{aligned} \quad (1)$$

Here the outcome value ( $Y_{Fij}$ ) is assumed normally distributed in each study conditional on the included covariates (here just  $X_{ij}$ , but additional covariates could also be included such as the baseline value of the continuous outcome<sup>12,13</sup>). There are  $K$  distinct intercept terms ( $\alpha_i$ ) and the main parameter of interest is  $\theta$ , which denotes the summary (average) treatment effect from the included studies. The true treatment effect in each study is assumed drawn from a normal distribution with mean  $\theta$  and between-trial variance  $\tau^2$ , and  $\sigma_i^2$  denotes the study-specific residual variance which is assumed normally distributed (this could also be stratified by treatment group, but we do not consider this here). The choice of coding of  $X_{ij}$  (eg, 1/0 or +0.5/−0.5 for treatment/control groups) does not alter the interpretation of  $\theta$ , our key parameter of interest; however, it does change interpretation of the  $\alpha_i$  and may have implications on estimation of  $\tau^2$ , as discussed by Jackson et al.<sup>5</sup> We evaluate this later in our simulations.

Alternatively, a random intercepts model could be specified. For example, allowing for between-study correlation between the random effects of the intercept and the treatment effect, the model can be written as follows:

$$\begin{aligned} Y_{Fij} &= (\alpha + u_{1i}) + (\theta + u_{2i})X_{ij} + e_{ij} \\ \begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} &\sim N \begin{pmatrix} \tau_\alpha^2 & \tau_{12} \\ \tau_{12} & \tau^2 \end{pmatrix} \\ e_{ij} &\sim N(0, \sigma_i^2) \end{aligned} \quad (2)$$

The parameter terms are as defined for model (1), except now the study-specific intercepts are also assumed drawn from a normal distribution, with mean of  $\alpha$  and between trial variance of  $\tau_\alpha^2$ , and the two random effects ( $u_{1i}$  and  $u_{2i}$ ) are allowed to be correlated through the covariance term  $\tau_{12}$ . Allowing for correlation imposes a between-study relationship of control group mean response and treatment effect, which might be viewed as controversial (see Discussion). To avoid this,  $\tau_{12}$  might be set to zero, but then the coding of treatment is potentially crucial (see later).<sup>11</sup>

In a frequentist framework, models (1) and (2) are typically fitted using REML estimation. Following estimation, a 95% confidence interval for  $\theta$  is conventionally derived using a (Wald) z-based method ( $\hat{\theta} \pm (1.96 \times s.e.(\hat{\theta}))$ ), but other options include the Satterthwaite and Kenward-Roger approaches, which replace 1.96 with the critical value of a t-distribution with a particular denominator degrees of freedom.<sup>4,14</sup> In this article, we also consider using  $\hat{\theta} \pm (t_{K-1,0.975} \times s.e.(\hat{\theta}))$ , where  $K$  is the number of studies in the IPD meta-analysis. For brevity, we refer to this as the t-based confidence interval approach.

## 2.2 | Binary outcomes

Now let us consider a binary outcome (eg, dead or alive 1 month after surgery), such that  $Y_{ij}$  is 1 for individuals with an event and 0 for those without an event. We use a logit-link function, such that our one-stage models have a logistic regression modeling framework (as suggested by Simmonds and Higgins<sup>15</sup>) and the treatment effect is measured by a log odds ratio. Again let the treatment group variable be denoted by  $X_{ij}$ , with coding options (such 1/0 for treatment/control groups) discussed further in Section 2.3.

In this situation, the stratified intercept model can be written as,

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \alpha_i + (\theta + u_i)X_{ij}$$

$$u_i \sim N(0, \tau^2), \quad (3)$$

where  $\pi_{ij}$  is the event probability for individual  $j$  in study  $i$ . There are  $K$  distinct intercept terms,  $\alpha_i$ , and the model parameter  $\theta$ , denotes the summary (average) treatment effect (log odds ratio). The true treatment effects are again assumed drawn from a normal distribution with mean  $\theta$ , and between-trial variance  $\tau^2$ . As in models (1) and (2), adjustment for baseline covariates is also possible.

Alternatively specifying a random intercepts model, and allowing for between-study correlation of control group risk and treatment effect, we have<sup>5,11,16</sup>:

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$$

$$\text{logit}(\pi_{ij}) = (\alpha + u_{1i}) + (\theta + u_{2i})X_{ij}$$

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim N \begin{pmatrix} \tau_\alpha^2 & \tau_{12} \\ \tau_{12} & \tau^2 \end{pmatrix}. \quad (4)$$

The parameter terms are as defined for model (3), except now the study-specific intercepts are also assumed drawn from a normal distribution, with mean of  $\alpha$  and between trial variance of  $\tau_\alpha^2$ , and the two random effects ( $u_{1i}$  and  $u_{2i}$ ) are allowed to be correlated through the covariance term  $\tau_{12}$ . As discussed for model (3), setting  $\tau_{12}$  to zero assumes no between-study correlation, but then treatment coding is more important (see below).

Models (3) and (4) are typically fitted using ML estimation, via a numerical integration approach such as (adaptive) Gaussian quadrature. Unfortunately, there is no natural extension from ML to REML estimation for the exact likelihood defined by a GLMM of a binary, ordinal or count outcome, as the model residuals cannot be estimated separately from the main parameters. Thus ML estimation is generally the default frequentist estimation choice for IPD meta-analyses of noncontinuous outcomes, for which downward bias in between-study variance estimates and low coverage of confidence intervals is a strong concern, especially with 10 or fewer studies in the IPD meta-analysis. However, Wolfinger and O'Connell suggest using a pseudo-likelihood approximation of the exact likelihood,<sup>17</sup> where the outcome response variable is transformed to an approximately linear scale. This allows REML to be used for GLMMs of noncontinuous

outcomes, but at the expense of an approximate likelihood. This may be an acceptable trade-off in some situations, to improve between-study variance estimates and confidence interval coverage. We will investigate this in Section 3.

## 2.3 | Coding of treatment

When a treatment variable is entered into a regression model as a covariate, it is typical practice for researchers to code the variable as 1/0 for treatment/control. However, a coding of +0.5/−0.5 has also been used by others, such as Morris et al<sup>9</sup>, Tudur-Smith et al,<sup>18</sup> and Turner et al.<sup>11</sup> For random intercepts models (2) and (4), which allow for between-study correlation in control group risk and treatment effect, the choice of treatment coding should be unimportant, because one can show mathematically a one-to-one correspondence of the model parameters with one coding and the model parameters with another coding, so that the maximized likelihood is the same.<sup>5,11</sup> However, if the between-study correlation is set to zero, then Turner et al suggest a +0.5/−0.5 coding is crucial; in particular, for model (4) this ensures the variance of the log-odds in control group patients is modeled as equal to that in intervention group,<sup>11</sup> as otherwise with a 1/0 treatment/control coding the variation for the intervention group is modeled as greater than or equal to the variance for the control group.

For stratified intercept models (1) and (3), Jackson et al<sup>5</sup> suggest a coding of +0.5/−0.5 improves ML estimation. However, they mainly evaluated situations where the treatment:control allocations were 1:1 in each trial. Therefore, in the following section we address unequal treatment:control allocations, and examine whether the following alternative treatment coding options improve ML estimation even further:

- *overall centering*: a coding of 1/0 minus the average (unweighted across trials) of the proportion of participants in the treatment group in each trial. For example, if there are 10 studies within an IPD meta-analysis, with five of those studies having 70% of participants in the treatment group and the other five having 50% in the treatment group, then the unweighted average proportion treated per trial is 60%. In this situation, the treatment coding is 1/0 - 0.6 in all trials, and thus individuals in the treatment group are coded as +0.4, and those in the control group are coded as −0.6.
- *study-specific centering*: a coding of 1/0 minus the study-specific proportion of participants in the treatment group. For example, for individuals in a trial where 40% of participants are in the treatment group, then an individual in the treatment group will be coded as  $1 - 0.4 = +0.6$ , and an individual in the control group would be coded as  $0 - 0.4 = -0.4$ . However, if another trial had 30% in the treatment group, then individuals in the treatment and control groups would be coded as +0.7 and −0.3, respectively.

## 3 | SIMULATION STUDY FOR CONTINUOUS AND BINARY OUTCOMES

We now use a simulation study to compare the performance of IPD meta-analysis models with stratified intercept or random intercepts, first for continuous outcomes and then for binary outcomes. We have three research questions:

Q1: Does an “overall centering” or “study-specific centering” coding of the treatment variable improve ML estimation of the between-study variance of the treatment effect, over and above the +0.5/−0.5 coding proposed by Jackson et al, for the stratified intercept models (1) and (3) in situations where one or more trials have an unequal treatment:control allocation ratio?

Q2: Does a t-based approach to confidence interval derivation improve coverage compared with a standard z-based (Wald) approach, for both stratified intercept and random intercepts models?

Q3: Does REML estimation of the pseudo-likelihood perform better than ML estimation of the exact likelihood for binary outcome models (3) and (4)?

### 3.1 | Continuous outcome simulation study

In our first simulation, we extend the simulation study of Legha et al for one-stage IPD meta-analysis models of continuous outcomes to the situation when there are varying treatment:control allocation ratios.



### 3.1.1 | Methods

Full details of the simulation methods are provided in Supplementary Material S1. Briefly, we simulated IPD according to model (2), with a 1/0 treatment/control coding and assuming no correlation of the pair of random effects (ie,  $\tau_{12} = 0$ ) for simplicity to avoid a relationship between control group response and treatment effect. A range of different simulation scenarios were considered (see supplementary material S1), each involving varying treatment:control allocation ratios (randomly drawn from a  $U[0.1,0.9]$  distribution), and varying the number of studies, number of participants, and magnitude of between-study variance of treatment effects.

One thousand IPD meta-analysis datasets were generated for each scenario. To each we fitted the random intercepts model (2) used to generate the data; that is, model (2) using a 1/0 treatment coding option, whilst forcing  $\tau_{12}$  to be 0 (its correct value) to avoid estimation issues that often arise when estimating between-study correlations.<sup>19</sup> Then we also fitted the stratified intercept model (1) for each of the four treatment coding options. Both ML and REML estimation were examined, alongside z-based and Satterthwaite confidence intervals. Although REML is the preferred estimation method for one-stage IPD meta-analyses of continuous outcomes, we also considered ML estimation to inform subsequent extension to one-stage IPD meta-analyses of binary outcomes, for which ML estimation is usually the default (see Section 3.2).

Performance was summarized by the bias in the summary treatment effect estimate ( $\hat{\theta}$ ), the coverage of 95% confidence intervals for the summary treatment effect, and the bias in the between-study variance ( $\hat{\tau}^2$ ) of the treatment effects. For the latter, we calculated percentage difference between the estimated and true between-study variance of the treatment effect (ie,  $100 \times (\hat{\tau}^2 - \tau^2)/\tau^2$ ), and report the mean and median of these percentages across the 1000 simulations. Distributions of  $\hat{\tau}^2$  were extremely skewed across each set of 1000 results, and so presentation of median values helps indicate estimation problems in addition to the more formally correct mean bias.

### 3.1.2 | Results when using ML estimation

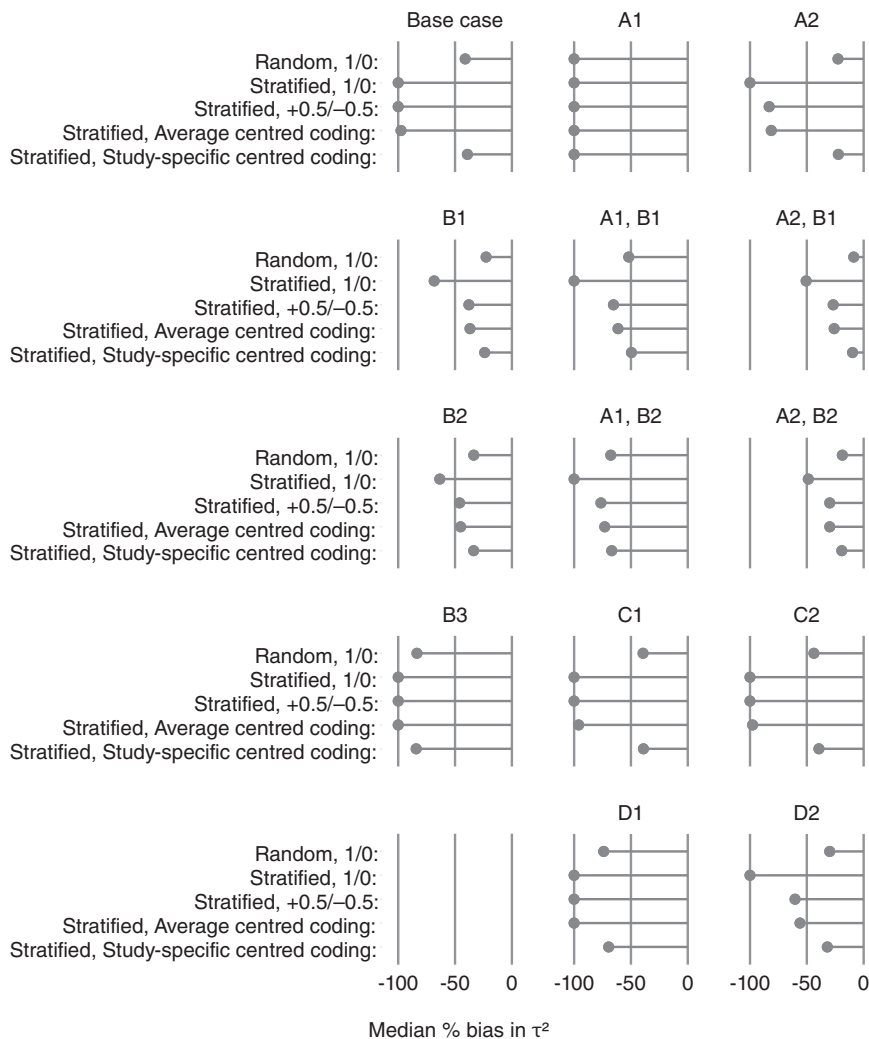
The summary treatment effect estimates were approximately unbiased for all scenarios, modeling approaches, and treatment coding options. However, in most scenarios there was considerable downward bias in the estimated between trial variance ( $\hat{\tau}^2$ ) of the treatment effects (see Figure 1, and also see supplementary material Tables S2(a), (b), and (c)). The bias was largest when using the stratified intercept model (1) with a 1/0 treatment/control coding for the treatment variable, and the bias was least when using the “study-specific centering” coding. For example, under the stratified intercept model and setting B1-A1 (which involves five trials where the number of participants per trial was  $U(30, 1000)$ ), the median downward bias of  $\hat{\tau}^2$  was 100% with 1/0 treatment/control coding, which improved to 65.4%, 61.4%, and 49.5%, with +0.5/−0.5 coding, an “overall centering” coding, and a “study-specific centering” coding, respectively.

Coverage of 95% confidence intervals for the summary treatment effect was also closest to 95% when using the “study-specific centering” approach. For example, in scenario B2 the stratified intercept model had a z-based confidence interval coverage of 79.70%, 84.20%, and 87.10% for the 1/0, +0.5/−0.5 and “study-specific centering” codings, respectively (Table S2(b)).

Crucially, “study-specific centering” of the treatment variable not only improves ML estimation of stratified intercept model (1), but also makes it comparable (in terms of bias and coverage) to ML estimation of the data generating model (3) (ie, the random intercepts model with 1/0 coding). However, despite having the best performance, both approaches still have downward bias in  $\hat{\tau}^2$  and gave z-based confidence interval coverage <95% for most scenarios. This appears worst in situations where the allocation ratio was most unbalanced (eg, see results in Table S2(a) where treatment prevalence was 90% in all studies).

### 3.1.3 | Results when using REML estimation

REML estimation also gives approximately unbiased summary treatment effect estimates for all scenarios, modeling approaches, and treatment coding options. It also reduces the downward bias in the ML estimate of  $\hat{\tau}^2$  for all modelling options (although the downward bias was not removed entirely). Furthermore, unlike for ML estimation, the choice of treatment coding becomes irrelevant when fitting the stratified intercept model (1) using REML estimation. The median (or mean) bias in  $\hat{\tau}^2$  was generally very similar regardless of the coding used (see Supplementary material Table S2[d]), as were the summary treatment effect estimates and their confidence intervals. Therefore, when using REML estimation of



**FIGURE 1** The median percentage bias of the between-trial variance of treatment effects ( $\hat{\tau}^2$ ) for ML estimation of the stratified intercept model (1) and random intercepts model (2)\*, for the continuous outcome simulation scenarios\*\* described in Table S1, allowing for unequal treatment:control allocation ratios in each trial in the IPD meta-analysis from 10% to 90%. Circular points denote the estimated percentage bias, and a horizontal line is drawn from each estimate to the ideal value of 0. IPD, individual participant data; ML, maximum likelihood. \*The random intercepts model refers to the data generating model, which was model (2) but with treatment coded as 1/0 and the between-study correlation assumed zero. \*\*Scenarios are labeled “base case,” “A1,” “A2,” and so on. For explanation of each scenario setting, see Table S1. Briefly, the base case was  $K = 10$ ,  $n_i = 100$ ,  $\theta = -9.66$ ,  $\tau^2 = 7.79$ . Then the other scenarios made changes of: A1:  $K = 5$ ; A2:  $K = 20$ ; B1:  $n_i \sim U(30,1000)$ ; B2:  $K = 10$ ,  $n_i \sim U(30,100)$  for trials 1 to 5,  $n_i \sim U(900,1000)$  for trials 6 to 10; A1,B1:  $K = 5$  and  $n_i \sim U(30,1000)$ ; A2,B1:  $K = 20$  and  $n_i \sim U(30,1000)$ . A1,B2:  $n_i \sim U(30,100)$  for trials 1 and 2,  $n_i \sim U(900,1000)$  for trials 3 to 5. A2,B2:  $K = 20$ ,  $n_i \sim U(30,100)$  for trials 1 to 10,  $n_i \sim U(900,1000)$  for trials 11 to 20; B3:  $n_i \sim U(30,100)$ ; C1: halving variance of intercept; C2: doubling variance of intercept; D1:  $\tau^2 = 3.9$ ; D2:  $\tau^2 = 15.6$

the stratified intercept model (1), “study-specific centering” of the treatment variable does not improve performance over traditional 1/0 coding, and the coding is irrelevant. Results from the stratified intercept model were also comparable to REML estimation of the random intercepts model (2) with 1/0 coding. Coverage of z-based 95% confidence intervals for the summary treatment effect were generally too low, but improved close to 95% when using the Satterthwaite approach (as also shown by Legha et al<sup>4</sup>).

### 3.2 | Binary outcome simulation study

In our second simulation study we extend the simulations of Jackson et al for IPD meta-analysis of binary outcomes,<sup>5</sup> to evaluate the ML estimation performance of random intercepts model (4) with a 1/0 coding and the stratified intercept

**TABLE 1** Simulation study scenarios of Jackson et al<sup>5</sup> that were extended in this article, by allowing for unequal treatment:control allocation ratios in each trial in the IPD meta-analysis

Data generation scenario, as labeled by Jackson et al	$K$	$\tau^2$	Number of participants in the treatment group ( $N$ )	Number of participants in the control group <sup>a</sup>	Baseline log-odds of the event in the control group ( $LO_c$ )
1	10	0.024	$N \sim U(50, 500)$	$N$	$LO_c N(\text{logit}[0.2], 0.3^2)$
3	10	0.168	$N \sim U(50, 500)$	$N$	$LO_c N(\text{logit}[0.2], 0.3^2)$
4	3	0.024	$N \sim U(50, 500)$	$N$	$LO_c N(\text{logit}[0.2], 0.3^2)$
5	5	0.024	$N \sim U(50, 500)$	$N$	$LO_c N(\text{logit}[0.2], 0.3^2)$
6	20	0.024	$N \sim U(50, 500)$	$N$	$LO_c N(\text{logit}[0.2], 0.3^2)$
7	10	0.024	$N \sim U(10, 100)$	$N$	$LO_c N(\text{logit}[0.2], 0.3^2)$

Note: All scenarios were performed with  $\theta = 0$  and  $\theta = \log(2)$ . For further details see Section 6 in Jackson et al.

Abbreviation: IPD, individual participant data.

<sup>a</sup>Shows the number used in the control group of the Jackson et al simulation. However, in our simulations we changed the number in the control group of a trial to ensure the treatment group prevalence was one of the following (chosen at random): 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90%.

model (3) for each of the four treatment coding options. A variety of simulation settings are considered, allowing for unequal treatment:control allocation ratios in each trial within the IPD meta-analysis. We examine the estimate of the between-study variance ( $\hat{\tau}^2$ ) of the treatment effect, the estimate of the summary treatment effect ( $\hat{\theta}$ ), and the coverage of the 95% confidence interval for the summary treatment effect.

### 3.2.1 | Methods

We chose the simulation scenarios from Jackson et al<sup>5</sup> that most closely correspond to those used for our continuous outcome simulations; that is, Jackson et al scenario settings 1, 3, 4, 5, 6, and 7, where the mean risk in the control group was 20%. These are summarized in Table 1, and cover a range of settings that vary in terms of the number of trials, number of participants per trial, and heterogeneity of parameters. We now consider these scenarios when we vary the treatment:control allocation ratio in the simulated trials. The omitted Jackson et al scenarios mainly corresponded to either zero or extremely large between-study heterogeneity, or a different logit data generating mechanism for the control group.

For each scenario, 1000 simulated IPD meta-analysis datasets were created from model (4), with a 1/0 treatment/control coding, but assuming no correlation of the pair of random effects (ie,  $\tau_{12} = 0$ ) for simplicity to avoid a relationship between baseline risk and treatment effect. Within each scenario we allowed for potential unequal treatment:control allocation ratios in each trial, by randomly drawing from a  $U(0.1, 0.9)$  distribution. Each scenario was undertaken assuming the summary odds ratio of 1 for the treatment effect (ie,  $\theta = \ln(1) = 0$ ), and also assuming a summary odds ratio of 2 (ie,  $\theta = \ln(2)$ ). Of note, the chosen magnitude of heterogeneity in setting 1 corresponded to a mean  $I^2$  of about 25% in the meta-analysis datasets simulated.

To each simulated IPD meta-analysis dataset, ML estimation was used to fit the random intercepts model used to generate the data (ie, model (4) with a 1/0 coding option and forcing  $\tau_{12} = 0$  to avoid estimation issues that often arise when estimating between-study correlation<sup>19</sup>), and the stratified intercept model (3) for each of the four treatment coding options. We used the *lme4* R package (version 1.1-17),<sup>20</sup> with ML estimation undertaken using adaptive Gaussian quadrature, with seven quadrature points.

Across all 1000 results obtained for each scenario, the median percentage bias of the between-study variance was calculated, as well as the mean bias of the summary treatment effect, and the coverage of 95% confidence intervals for the summary treatment effect. The latter was derived as the proportion (across the 1000 results) of 95% confidence intervals that contained the true summary effect. Confidence intervals were calculated using the standard z-based method ( $\hat{\theta} \pm (1.96 \times s.e.(\hat{\theta}))$ ) and also by the t-based approach that replaces 1.96 with the critical value of a t-distribution with  $K-1$  degrees of freedom (ie, use  $\hat{\theta} \pm (t_{K-1, 0.975} \times s.e.(\hat{\theta}))$ ).<sup>21</sup> Given the 1000 simulations in each scenario, if coverage is truly 95%, then we would expect to observe an estimated coverage between 93.4% and 96.2%.

Finally, we repeated our simulations using REML estimation of the pseudo likelihood. Given that all treatment coding options gave similar performance for REML estimation of continuous outcomes (Section 3.1.3), we only considered

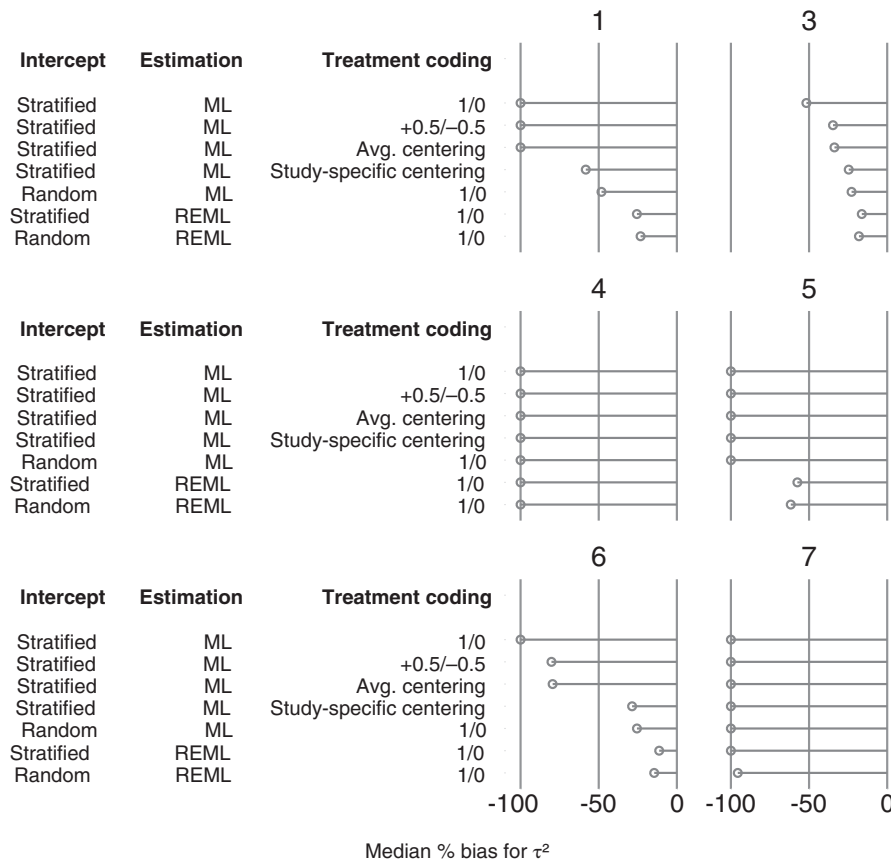


the 1/0 treatment variable coding when fitting the stratified intercept and random intercepts models. REML estimation was undertaken using MLWin, via the *runmlwin* package within Stata.<sup>22</sup> We obtained parameter estimates using REML estimation of the first-order marginal quasi-likelihood linearization of the likelihood (“mql1” option). This linearization approach is noted in the *runmlwin* help file as being the most stable and fastest to converge, and so was deemed sensible for our large simulation study. In practice, more accurate (though potentially less stable) estimation options such as second-order penalized quasi-likelihood linearization (“pql2” option) could be used, with the estimates from the first-order approach used as initial values (see applied examples in Section 4 for further discussion on this).

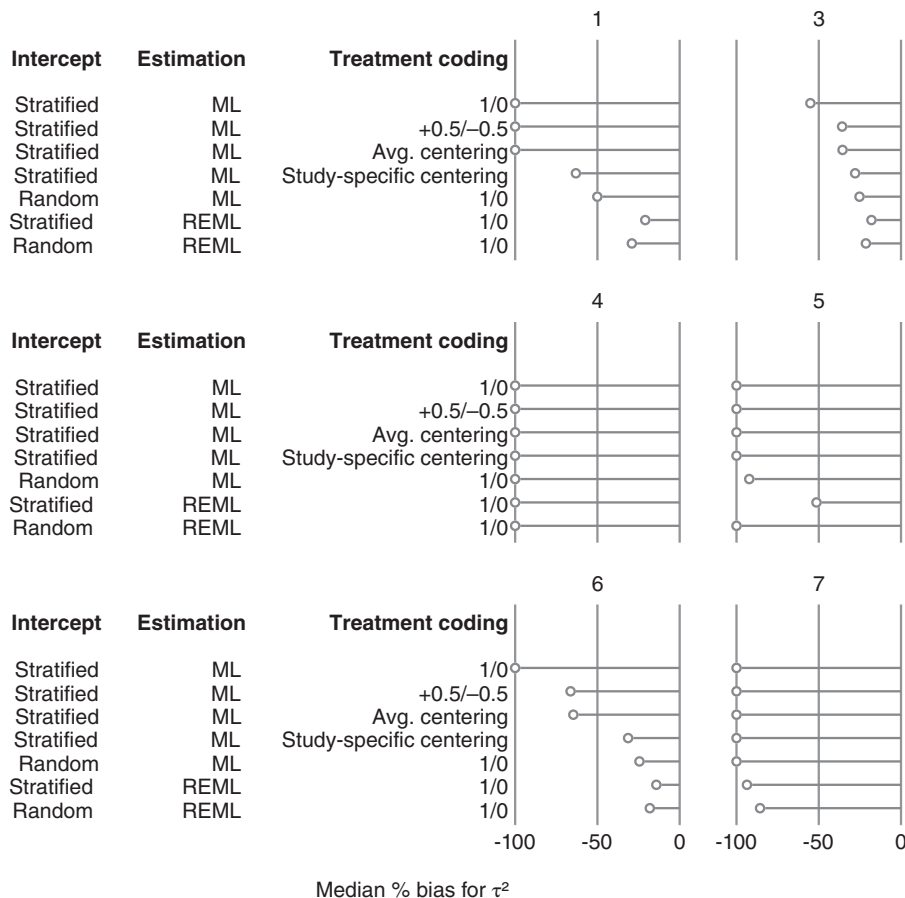
### 3.2.2 | Results when using ML estimation of the exact likelihood and standard z-based confidence intervals

Simulation results for the mean bias of the summary treatment effect estimate ( $\hat{\theta}$ ) across all scenarios for a true treatment effect of  $\theta = 0$  and of  $\theta = 1$  are shown in Supplementary Tables S3 and S4, respectively. The bias of  $\hat{\theta}$  was generally negligible in all cases.

Figures 2 and 3 show the median percentage bias in the between-study variance of the treatment effect ( $\hat{\tau}^2$ ) for the  $\theta = 0$  and  $\theta = \ln(2)$  scenarios, respectively (and also Supplementary Tables S5 and S6 show the mean percentage bias).



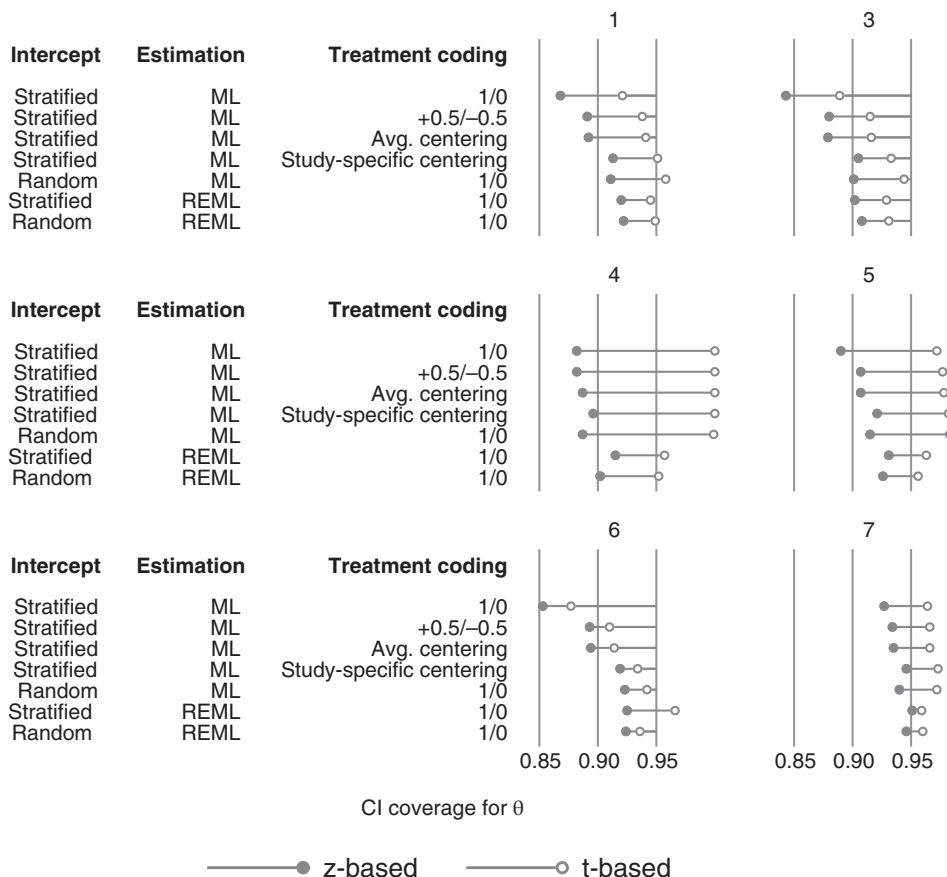
**FIGURE 2** The median percentage bias of the between-trial variance of treatment effects ( $\hat{\tau}^2$ ) for ML estimation (exact likelihood) and REML estimation (pseudo likelihood) of the stratified intercept model (3) and random intercepts model (4)\*, for simulation scenarios\*\* where  $\theta = 0$  and allowing for random treatment prevalences (10%-90%) for each study within the IPD meta-analysis. Circular points denote the estimated percentage bias, and a horizontal line is drawn from each estimate to the ideal value of 0. IPD, individual participant data; ML, maximum likelihood; REML, restricted maximum likelihood. \*The random intercepts model refers to the data generating model, which was model (4) with treatment coded as 1/0 and the between-study correlation assumed zero. \*\*See Table 1 for full details of the scenario corresponding to the number shown. True value for  $\tau^2$  is 0.024, except in setting 3 where  $\tau^2$  equals 0.168. All settings also allow the treatment prevalence for a particular study within a meta-analysis to vary, whereby this is selected from  $U(0,1,0.9)$  and then rounded to the nearest 0.1



**FIGURE 3** The median percentage bias of the between-trial variance of treatment effects ( $\hat{\tau}^2$ ) for ML estimation (exact likelihood) and REML estimation (pseudo likelihood) of the stratified intercept model (3) and random intercepts model (4)\*, for simulation scenarios\*\* where  $\theta = \ln(2)$  and allowing for random treatment prevalences (10%–90%) for each study within the IPD meta-analysis. Circular points denote the estimated percentage bias, and a horizontal line is drawn from each estimate to the ideal value of 0. IPD, individual participant data; ML, maximum likelihood; REML, restricted maximum likelihood. \*The random intercepts model refers to the data generating model, which was model (4) with treatment coded as 1/0 and the between-study correlation assumed zero. \*\*See Table 1 for full details of the scenario corresponding to the number shown. True value for  $\tau^2$  is 0.024, except in setting 3 where  $\tau^2$  equals 0.168. All settings also allow the treatment prevalence for a particular study within a meta-analysis to vary, whereby this is selected from  $U(0.1, 0.9)$  and then rounded to the nearest 0.1 decimal place

As observed for continuous outcomes, the results show that using the stratified intercept model (3) with a 1/0 treatment/control coding gives the most downwardly biased estimates of the between trial variance of treatment effect. Often the median downward bias was 100%, and mean downward bias typically between 40% and 80%. This downward bias was generally reduced when using either a +0.5/−0.5 treatment coding or the “overall centering” treatment coding, and by a similar amount. However, the downward bias was smallest when using a “study-specific centering” coding. For example, under setting 6 (involving 20 trials), the median downward bias of between study variance estimates from stratified intercept model (3) was 100% with a 1/0 treatment/control coding; this was slightly reduced to 80.2% and 79.5% with +0.5/−0.5 coding or “overall centering” coding, but considerably reduced to 28.7% when using the “study-specific centering” coding. The mean downward bias shows a similar pattern; this was 80.8%, 44.5%, 43.3%, and 7.85% when using the 1/0, +0.5/−0.5, “overall centering,” and “study-specific centering,” respectively (Table S5).

For stratified intercept model (3), the reduction in downward median and mean bias of  $\hat{\tau}^2$  when using a “study-specific centering” coding of treatment also improves upon the z-based coverage of 95% confidence intervals for the summary treatment effect, as shown in Figures 4 and 5 (and also supplementary Tables S5 and S6) for the  $\theta = 0$  and  $\theta = \ln(2)$  scenarios, respectively. For example, in setting three of Figure 4, the coverage of z-based confidence intervals from the 1/0 coding is 84%, but this improves to 91% when using the “study-specific centering” coding. Furthermore, the z-based



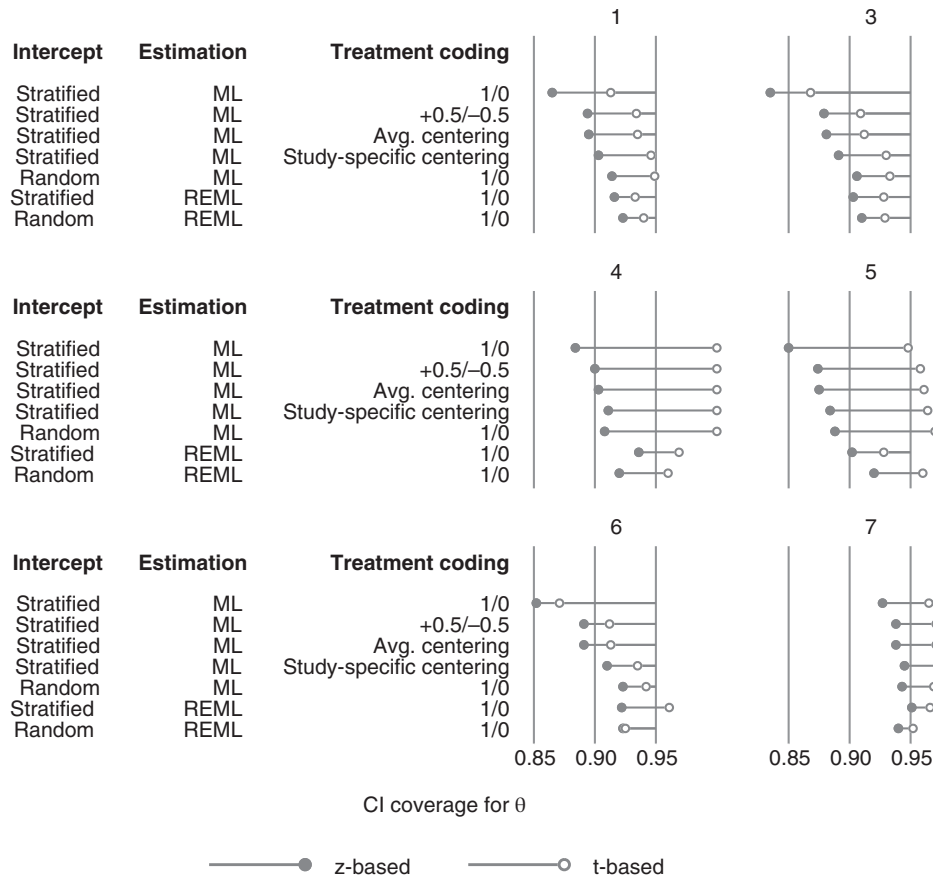
**FIGURE 4** Coverage (proportion) of z-based and t-based 95% confidence intervals for the summary treatment effect for ML estimation (exact likelihood) and REML estimation (pseudo likelihood) of the stratified intercept model (3) and random intercepts model (4)\*, for simulation scenarios\*\* where  $\theta = 0$  and allowing for random treatment prevalences (10%-90%) for each study within the IPD meta-analysis. Circular points denote the estimated coverage, and a horizontal line is drawn from each estimate to the ideal value of 0.95. IPD, individual participant data; ML, maximum likelihood; REML, restricted maximum likelihood. \*The random intercepts model refers to the data generating model, which was model (4) with treatment coded as 1/0 and the between-study correlation assumed zero. \*\* See Table 1 for full details of the scenario corresponding to the number shown. True value for  $\tau^2$  is 0.024, except in setting 3 where  $\tau^2$  equals 0.168. All settings also allow the treatment prevalence for a particular study within a meta-analysis to vary, whereby this is selected from  $U(0.1,0.9)$  and then rounded to the nearest 0.1

coverage is always best (ie, closest to 95.0%) when using the “study-specific centering” coding, and also worst when using the 1/0 coding; the coverage of the +0.5/-0.5 and “overall centering” coding are similar and fall in between the z-based coverage when using the 1/0 coding and “study-specific centering” coding.

Crucially, “study-specific centering” of the treatment variable not only improves ML estimation of stratified intercept model (3), but also makes it comparable to ML estimation of the random intercepts model (4) with 1/0 coding (ie, bias and coverage are very similar). However, despite having the best performance, both approaches still have (often considerable) downward bias in  $\hat{\tau}^2$ , and subsequently z-based coverage is less than 95% for most scenarios. The coverage appears closest to 95% in the scenario 7 setting (see Figure 5); this can be explained by the studies being smaller in this setting, and so the within-study variances dominate the total variability (ie,  $I^2$  is small), and so any downward bias in the between-study variance is less impactful.

### 3.2.3 | Results when using ML estimation and t-based confidence intervals

For all treatment coding options, and for both stratified intercept and random intercepts models, coverage of 95% confidence intervals for the summary treatment effect was generally improved (ie, moved closer to 95%) by using the



**FIGURE 5** Coverage (proportion) of z-based and t-based 95% confidence intervals for the summary treatment effect for ML estimation (exact likelihood) and REML estimation (pseudo likelihood) of the stratified intercept model (3) and random intercepts model (4)\*, for simulation scenarios\*\* where  $\theta = \ln(2)$  and allowing for random treatment prevalences (10%–90%) for each study within the IPD meta-analysis. Circular points denote the estimated coverage, and a horizontal line is drawn from each estimate to the ideal value of 0.95. IPD, individual participant data; ML, maximum likelihood; REML, restricted maximum likelihood. \*The random intercepts model refers to the data generating model, which was model (4) with treatment coded as 1/0 and the between-study correlation assumed zero. \*\*See Table 1 for full details of the scenario corresponding to the number shown. True value for  $\tau^2$  is 0.024, except in setting 3 where  $\tau^2$  equals 0.168. All settings also allow the treatment prevalence for a particular study within a meta-analysis to vary, such that it is selected from  $U(0.1, 0.9)$  and then rounded to the nearest 0.1

t-based approach to deriving confidence intervals based on the t-distribution with  $K-1$  degrees of freedom (see Tables S5 and S6). For example, in Setting 1 when using stratified intercept model (3) with “study-specific centering” of the treatment variable, the coverage was 90.3% and 94.6% for z-based and t-based confidence intervals, respectively. Only in Setting 4, where the number of studies was only 3, is the t-based approach a concern as the coverage is close to 100% and so too high; although arguably this is still preferable to the under-coverage from the corresponding z-based approach.

### 3.2.4 | Results when using REML estimation of the pseudo likelihood

When using REML estimation of the pseudo-likelihood with a 1/0 treatment coding, results are shown in Table S7 for all scenarios. Stratified intercept model (3) and random intercepts model (4) have similar performance in all scenarios, and there is negligible or very small bias in the summary treatment effect estimates.

Compared with ML estimation of the exact likelihood, REML estimation of the pseudo likelihood improved the between-study variance estimate ( $\hat{\tau}^2$ ), for both stratified intercept and random intercepts models. For example, in setting 1 and  $\theta = \ln(2)$ , when using stratified model (3) the median (mean) bias in  $\hat{\tau}^2$  was  $-20.93\%$  ( $10.72\%$ ) using REML

Box 1 A summary of the key findings based on our simulation study results and applied examples

- For ML estimation of a one-stage IPD meta-analysis model with a stratified intercept, a “study-specific centering” coding of the treatment variable reduces downward bias of between-study variances and improves coverage of 95% confidence intervals for the summary (treatment) effect, as compared with other treatment coding options such as 1/0 for treatment/control. Supplementary material S8 also shows this mathematically for a simple case where all studies in the IPD meta-analysis are of the same size.
- REML is better than ML estimation for continuous outcomes. For binary outcomes, the simulations do not suggest an important difference in terms of bias and coverage of confidence intervals for the summary treatment effect when using REML estimation of the pseudo likelihood compared with ML estimation of the exact likelihood. However, in most scenarios REML reduces the downward bias in the between-study variance estimates, which may be important when the focus is on predictive inferences (eg, the predicted treatment effect in a new study<sup>23</sup>). Thus both ML (exact likelihood) and REML (pseudo likelihood) estimation may be important to consider for binary outcomes.
- For either ML or REML estimation, coverage of 95% confidence intervals for the summary treatment is generally too low when using a z-based approach. Improvements are generally made for REML estimation of continuous outcomes by using Satterthwaite or Kenward-Roger approaches, and for ML or pseudo REML estimation of binary outcomes by using  $\hat{\theta} \pm (t_{K-1, 0.975} \times \text{s.e.}(\hat{\theta}))$  where K is the number of studies in the meta-analysis.
- For continuous outcomes, REML estimation is recommended over ML estimation for either stratified intercept or random intercepts models (see work of Legha et al<sup>4</sup>), as it improves estimates of between-study variances (though some downward bias may remain), whilst having negligible bias in the summary treatment effect estimate and does not depend on the treatment coding chosen.
- For binary outcomes, when fitting a stratified intercept model both ML estimation of the exact likelihood (with “study-specific centering” treatment coding) and REML estimation of the pseudo likelihood (with 1/0 treatment coding) give negligible bias in the summary treatment effect estimate, and their coverage of 95% confidence intervals is close to 95% when using the t-based approach (unless the number of studies is less than 5). In addition, REML estimation of the pseudo likelihood often has less downward bias of between-study variance estimates than ML estimation.
- For binary outcomes, REML estimation of the pseudo likelihood may be unstable in sparse data situations, such as when most studies in the IPD meta-analysis are small (in terms of participants or events).
- The decision to use random study intercepts, rather than a stratified study intercept, depends on whether the researcher is willing to borrow information about control group risk across studies and/or assume a between-study relationship of control risk and treatment effect.

estimation (with a 1/0 treatment/control coding) compared with  $-63.08\%$  ( $-19.40\%$ ) when using ML estimation (with a “study-specific centering coding” for treatment).

The coverage of confidence intervals from REML estimation were closest to 95% when using the t-based approach; for example, in setting (1) with  $\theta = \ln(2)$ , the coverage from z-based and t-based confidence intervals was 91.6% and 93.3%, respectively. Indeed, t-based coverage was consistently good in all settings, generally between 93% and 97%, and comparable to that when using ML estimation with “study-specific centering” and the t-based approach.

### 3.3 | Summary of our key findings

A summary of key findings from the simulation study is shown in Box 1.

## 4 | ILLUSTRATION OF KEY FINDINGS IN APPLIED EXAMPLES

We now illustrate the key findings in applied examples, which focus on binary outcomes. Example 2 has data similar to that used in the simulation studies, whilst example 1 considers more sparse data.

Study	Number of women		Number of cardiovascular disease events	
	Control	Treatment	Control	Treatment
1	174	701	0	5
2	14	15	1	0
3	16	15	0	1
4	20	20	1	1
5	26	29	0	1
6	84	84	3	1
7	66	68	0	3

**TABLE 2** Summary of the IPD from seven trials examining the effect of hormone replacement therapy on the incidence of heart disease, as reported by Simmonds and Higgins<sup>15</sup>

Abbreviation: IPD, individual participant data.

**TABLE 3** Results from ML estimation of models (3) and (4) when fitted to the IPD summarized in Table 2

Model	Treatment coding	Summary treatment effect, $\hat{\theta}$	95% CI		Between-study (co)variances
			z-based	t-based	
Stratified intercept model (3)	1/0	0.56	-0.53, 1.64	-0.80, 1.91	$\hat{\tau}^2 = 0$
	“Study-specific centering”	0.65	-0.69, 1.99	-1.02, 2.32	$\hat{\tau}^2 = 0.57$
Random intercepts model (4)	1/0	0.55	-1.10, 2.21	-1.51, 2.62	$\hat{\tau}^2 = 0.74$ $\tau_{12} = -0.81$ $\tau_{\alpha}^2 = 1.16$

Abbreviations: IPD, individual participant data; ML, maximum likelihood.

## 4.1 | Example 1: hormone replacement therapy and incidence of heart disease

Simmonds et al combined IPD from seven trials examining the effect of hormone replacement therapy compared with control on the incidence of heart disease.<sup>15</sup> The binary outcome data are sparse (Table 2), such that the number of events is few in all studies due to the outcome being rare, and some groups have zero events. In this situation, a traditional two-stage IPD meta-analysis (ie, estimating the treatment effect and its variance in each study separately, and then pooling the results in an inverse-variance weighted meta-analysis) is problematic. The treatment effect cannot be estimated in every study unless a continuity correction is applied in those studies with a zero event; further, the assumption in the second stage that study-specific treatment effect estimates are normally distributed with known variances may be inappropriate. A one-stage approach avoids these issues by analyzing the IPD in a single step, for example, using either stratified intercept model (3) or random intercepts model (4), and the results are shown in Table 3.

### 4.1.1 | Stratified intercept model results

One of the studies in the IPD meta-analysis (study 1) had a treatment:control allocation ratio of about 4:1, whereas other studies have close to a 1:1 allocation ratio. Our simulation results showed that in this situation ML estimation of model (3) is improved by using a “study-specific centering” of the treatment variable, as this reduces downward bias in between-study variance estimates compared with a traditional 1/0 coding. The ML estimates in Table 2 reflect this, as  $\hat{\tau}^2 = 0$  when using a 1/0 coding and  $\hat{\tau}^2 = 0.57$  when using “study-specific centering” coding. This led to a noticeably different summary treatment effect of  $\hat{\theta} = 0.65$  (odds ratio of 1.91) when using “study-specific centering” compared with  $\hat{\theta} = 0.56$  (odds ratio of 1.74) when using 1/0 coding. Confidence intervals were also much wider.

Another key finding of the simulation study was that z-based (Wald) confidence intervals are generally too narrow, and t-based confidence intervals are more appropriate. In our example, t-based confidence intervals were also considerably wider. For example, when using the “study-specific centering” coding for ML estimation of model (3), the confidence



interval for the summary odds ratio was 0.36 to 10.15 when using t-based, compared with 0.59 to 6.77 when using the z-based approach.

Although our simulations suggest REML estimation of the pseudo likelihood for model (3) performs well, the scenarios did not cover sparse data akin to that in Table 2. Indeed, when applying REML estimation to this example, parameter estimates were unstable; there were large differences in parameter estimates from first and second-order linearization of the likelihood, and even when changing the coding of treatment (which should not occur for REML; see Section 3.1.3). Therefore, the ML estimates in Table 3 based on the exact likelihood are more reliable for this example. Of note, these ML estimates were very different to those obtained from a traditional two-stage IPD meta-analysis with continuity corrections of +0.5 added to deal with zero cells. The latter gave a summary odds ratio of 1.31 and  $\hat{\tau}^2 = 0$  from REML estimation, which are much lower than the ML estimates from the more exact one-stage model using “study-specific centering.”

#### 4.1.2 | Comparison of results for stratified intercept and random intercepts models

Unlike in the simulation study, the ML estimation results for random intercepts model (4) with 1/0 coding were not comparable to those from stratified intercept model (3) with “study-specific centering” coding (Table 3). The reason is that the simulation did not allow borrowing of information between baseline (control group) risk and treatment effect when generating the IPD. However, in the applied example model (4) estimated a strong negative correlation of  $-0.87$  between the pair of random effects, which had a strong influence on the results. Furthermore, in our simulation study we knew that a normal distribution on the baseline risk was appropriate (as we simulated the IPD from this assumption). However, in this real example we did not know if such an assumption is correct, and it may even compromise randomization in each trial, especially given the sparse events in the included trials. The approach of stratified intercept model (4) avoids making any assumptions about the between-study distribution of baseline risk, or the between-study relationship between baseline risk and treatment effect.

#### 4.2 | Example 2: Diet and lifestyle interventions and health outcomes during pregnancy

In our second example, we used IPD from 36 randomized trials (12 477 women) evaluating the effect of diet and lifestyle interventions compared with control (usual care) on health outcomes during pregnancy.<sup>24</sup> We focused on a subset of 10 trials that recorded the binary outcome of large for gestational age (yes/no). Eight studies had approximately 1:1 treatment:control allocation, and the other two had a 2:1 allocation ratio. The outcome risk in the control group varied, but was about 15% on average, similar to that used in our simulation studies. Although the number of participants was reasonably large in most studies, often there are fewer than 10 events in each group (Table 4), which again raised doubt as to the suitability of a traditional two-stage approach.

ML and REML estimates for one-stage model (3) are shown in Table 5. The findings again mirror those of the simulation study. When using ML estimation for model (3), the estimated between-study variance was much larger when using “study-specific centering” ( $\hat{\tau}^2 = 0.42$ ) rather than 1/0 coding ( $\hat{\tau}^2 = 0.29$ ) of the treatment variable, and this led to wider confidence intervals for the summary treatment effect. REML estimation of the pseudo likelihood was quite stable, with more similar estimates for first- and second-order linearizations of the likelihood. The REML estimates of between-study variances were larger than the ML estimates (Table 5), and this widened confidence intervals for the summary treatment effect. Results for model (4) were again somewhat different to model (3), for the same reasons described in the previous example. For all models, the widest confidence intervals arose when using the t-based rather than z-based approach.

## 5 | DISCUSSION

Our simulation study and applied examples identify key findings for estimation of one-stage IPD meta-analysis models, which are summarized in Box 1. There are three major implications. First, for ML or REML estimation of stratified intercept or random intercepts models, z-based (Wald) confidence intervals for the summary treatment effect are generally too narrow; performance is generally improved by using the Satterthwaite (or Kenward-Roger) approaches for continuous outcomes,<sup>4,14</sup> or a t-based approach with  $K-1$  degrees of freedom for binary outcomes. Second, when using ML

Study	Number of women		Number of babies large for gestational age	
	Control	Treatment	Control	Treatment
1	120	109	23	22
2	37	33	5	7
3	68	72	7	2
4	33	34	1	2
5	143	136	7	2
6	63	134	5	11
7	1095	1104	154	132
8	47	46	27	5
9	65	130	16	22
10	50	51	11	14

**TABLE 4** Summary of the IPD from 10 trials examining the effect of diet and lifestyle interventions on large for gestational age

Abbreviation: IPD, individual participant data.

**TABLE 5** Results from maximum likelihood (ML) and restricted maximum likelihood (REML) estimation of models (3) and (4) when fitted to the individual participant data summarized in Table 4

Model	Estimation method	Treatment coding	Summary			Between-study (co)variances
			treatment effect, $\hat{\theta}$	95% CI:z-based	95% CI:t-based	
Stratified intercept model (3)	ML exact	1/0	-0.43	-0.89, 0.04	-0.96, 0.11	$\hat{\tau}^2 = 0.29$
	ML exact	“Study-specific centering”	-0.40	-0.92, 0.12	-1.00, 0.20	$\hat{\tau}^2 = 0.42$
	REML pseudo	1/0	-0.48	-1.06, 0.09	-1.14, 0.18	$\hat{\tau}^2 = 0.54$
Random intercepts model (4)	ML exact	1/0	-0.38	-0.91, 0.16	-1.00, 0.24	$\hat{\tau}^2 = 0.43$ $\tau_{12} = -0.29$ $\tau_{\alpha}^2 = 0.81$
	REML pseudo	1/0	-0.38	-0.94, 0.18	-1.03, 0.27	$\hat{\tau}^2 = 0.54$ $\tau_{12} = -0.36$ $\tau_{\alpha}^2 = 0.92$

Note: ML estimates-based numerical quadrature of the exact likelihood (with seven quadrature points), and REML estimates based on a second-order penalized quasi-likelihood linearization with the estimates from the first-order marginal quasi-likelihood linearization used as initial values.

estimation of a one-stage model with a stratified intercept, a “study-specific centering” coding of the treatment variable should be chosen, as this reduces the bias in the between-study variance estimate (compared with 1/0 and other coding options). Third, REML estimation reduces downward bias in between-study variance estimates compared with ML estimation, and does not depend on the treatment coding chosen, thus should be used where possible; for IPD meta-analyses of binary outcomes, this requires REML estimation of the pseudo-likelihood, although it may be unstable when data are sparse.

### 5.1 | REML vs ML estimation

Our simulations of continuous outcomes show that REML is better than ML estimation to improve variance estimates, which will not be a surprise to most readers. In particular, REML reduces the bias in  $\hat{\tau}^2$  by adjusting for the total number of parameters being estimated.<sup>6,25,26</sup> However, REML is not an option when using numerical integration of the exact likelihood for a one-stage IPD meta-analysis of a binary outcome, and therefore ML estimation is the most common method

used. In most software packages the default estimation method to fit generalized linear mixed models is ML estimation via a numerical integration method such as quadrature. Therefore, our findings about the importance of “study-specific centering” coding are most relevant for one-stage IPD meta-analyses of binary outcomes, or other generalized linear mixed models or frailty models that apply ML estimation. Indeed, improving ML estimation by centering covariates has a REML essence to it, as both approaches aim to disentangle (ie, make uncorrelated) the estimation of main parameters of interest from other nuisance parameters.

Given that considerable downward bias in between-study variance estimates often occurs using ML estimation (even after “study-specific” treatment coding; see Figures 1 to 3), REML estimation of the pseudo likelihood is appealing for binary outcomes.<sup>17</sup> Our simulations do not suggest an important difference between REML and ML in terms of bias of the summary treatment effect estimate and coverage of t-based confidence intervals for the summary treatment effect. However, in most scenarios REML estimation did reduce the median downward bias in the between-study variance estimates, and therefore we suggest it is the default. However, caution is advised if the data are sparse, such that most studies are small and have few or even zero events. Our simulations did not cover scenarios with sparse data, but previous work suggests that REML estimation of pseudo likelihood is not accurate in such situations and ML estimation of the exact likelihood is preferred.<sup>27</sup> Indeed, REML estimates may be unstable in such situations. Instability is evident when first-order and second-order linearizations of the likelihood lead to very different parameter estimates, or when reparameterizations that should not affect REML (such as centering of covariates) do still change parameter estimates importantly. In our applied example using the trials in Table 2, the data were sparse, and these stability problems were evident when using REML estimation, and so the ML estimates with “study-specific centering” were deemed more reliable. A Bayesian approach could also be considered in such situations, which would retain the exact likelihood during parameter estimation and could be combined with empirically based prior distributions for the between-study variance.<sup>28,29</sup>

Our findings warrant further evaluation in a wider variety of settings than those considered in our simulation, but concur with related work,<sup>30</sup> including Piepho et al<sup>31</sup> who evaluate frequentist network meta-analysis of binary outcomes. They too show that REML estimation of the pseudo-likelihood, and also the use of the h-likelihood, reduce bias in between-study variance estimates and give satisfactory coverage rates, especially when the Kenward-Roger approach is used to derive confidence intervals. They also consider improving ML estimation by various reparameterizations of the exact likelihood that aim to mimic the REML approach for linear mixed models. These reparameterizations also reduce the downward bias in ML estimates, but coverage of summary treatment effects is often too low. Our “study-specific centering of covariates” approach showed suitable coverage when using the t-based approach to confidence intervals, and is potentially simpler to implement in existing software. However, formal comparison of our proposal with those of Piepho et al is needed. Thomas et al also compare the performance of one-stage IPD meta-analyses for binary outcomes,<sup>30</sup> but do not find a “meaningful difference” between results from REML of the pseudo likelihood and ML of the exact likelihood. However, the authors only considered 1:1 treatment:control allocations and implemented a +0.5/−0.5 treatment variable coding, and thus essentially adapted “study-specific centering” in their setting, which ensures ML estimation performs well. Still, their bias in between-study variances were generally lower using REML, akin to our findings. Coverage of confidence intervals was generally much lower than 95%, but were based on a z-based approach.

Comparisons to the traditional two-stage IPD meta-analysis approach are also needed, for which REML is often recommended in the second stage.<sup>32</sup> Although it assumes normality of the between-study variance of treatment effects, REML is quite robust to deviations from this assumption.<sup>33</sup> We suspect that situations where REML estimation of the pseudo-likelihood performs well for a one-stage analysis of binary outcomes, a two-stage approach using REML will also perform well. Thomas et al<sup>30</sup> recommend one-stage rather than two-stage analyses when data in the IPD meta-analysis are sparse. We agree with Langan et al<sup>32</sup> that, especially in IPD meta-analyses of few studies, any heterogeneity variance estimate “should not be used as a reliable gauge for the extent of heterogeneity in a meta-analysis”.

## 5.2 | Implications of our findings

Our findings have important consequences, as they allow researchers to apply one-stage IPD meta-analysis models with a *stratified intercept*, rather than random intercepts, when using either REML or ML estimation. This is important, as the use of random intercepts makes distributional assumptions and potentially compromises within-trial randomization, but this is avoided using a stratified intercept. Previously, we recommended random intercepts for one-stage IPD meta-analysis models fitted using ML estimation,<sup>4</sup> as this approach had better estimates of between-study variances due to reducing

the number of parameters. However, our simulations show that the “study-specific centering” makes the stratified intercept model comparable to the random intercepts model (when no borrowing of information across studies is allowed in control group risk). Such comparable performance is despite the data generating mechanism actually being based on the random intercepts model, and so the simulation set-up might be considered more favorable toward the random intercepts model.

A potential limitation of stratified intercept models is that they may fail to converge when the number of events are rare, and in particular when some studies have a zero event in the control group (although this was not an issue for our first applied example). In that situation, assuming random study intercepts may help, because study-specific intercepts are not estimated directly, and studies rather contribute toward the estimation of the between-study distribution of intercepts (with the caveat of sharing information about control group risk across trials and thus potentially compromising randomization within trials). If between-study correlation is included in such random intercepts models (ie, model (4) is used), then the coding of treatment should not matter. However, if between-study correlation is assumed zero, then a  $+0.5/-0.5$  coding is recommended. An alternative method is the hypergeometric-normal approach of Stijnen et al<sup>34</sup> (referred to as model (7) in Jackson et al<sup>5</sup>), which conditions out the study-specific intercepts (thus avoiding their estimation); in the scenarios of the Jackson et al simulation study, this approach performs well and comparable to using a stratified intercept with “study-specific centering”.

### 5.3 | Extensions

Although we focused on randomized trials, our findings also apply more generally. In particular, any one-stage IPD meta-analysis model fitted using ML estimation should include covariates (treatments, prognostic factors, adjustment factors, and so on) centered by their study-specific means; for example, when synthesizing IPD from observational studies to evaluate risk or prognostic factors for binary or survival outcomes,<sup>35</sup> the included factors and any adjustment covariates should be coded with “study-specific centering”. Indeed, exposure prevalence (eg, the proportion of individuals classed as biomarker positive) is likely to be more varied across included covariates in observational studies (than treatment prevalence in randomized trials), and thus “study-specific centering” will be even more important.

We focused on parallel-group trials, for which our “study-specific centering” approach centers by the proportion in the treatment group; equivalently, we could center around the proportion in the control group. We considered binary variables, but “study-specific centering” should also be used for continuous variables where they are centered by the mean value. Indeed, it generalizes to any covariate: we simply center at the covariate’s mean value in each study. For example, for an ordinal covariate with possible values of 0, 1, 2, and 3, the “study-specific centering” coding is the original value minus the mean value for all individuals in the same study. In our simulations, we assumed a uniform distribution or fixed treatment prevalences across studies. Similarly we generated control groups risks assuming a normal distribution, and did not allow any correlation between baseline risk and treatment effect. Other approaches could be considered for data generation in further work. We also only consider one treatment and one control group per study.

Our simulations did not allow the between-study correlation to be estimated when fitting the random intercepts models (2) or (4), as we forced the correlation to be zero as it was in the data generating model. Further research might also consider how the random intercepts models perform when the correlation is freely estimated, although related simulations have shown that between-study correlations are difficult to estimate reliably, and often estimated values are  $+1$  or  $-1$ .<sup>19</sup>

## 6 | CONCLUSIONS

We recommend one-stage IPD meta-analysis models for continuous or binary outcomes use a stratified intercept. When using ML estimation to fit such models, researchers should use a “study-specific centering” coding of included variables. This will improve estimation of between-study variances and give more appropriate coverage of 95% confidence intervals for the summary (treatment) effects of interest. For continuous outcomes, REML estimation is recommended and then the coding should not be important. For binary outcomes, REML estimation of the pseudo likelihood will often improve

upon ML estimation of the exact likelihood, although it may be unstable when data are sparse. For either ML or REML estimation, confidence intervals should be derived using an approach based on the t-distribution.

## ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor and five anonymous reviewers for their constructive comments that helped us to improve the article upon at the revision stage.

## FUNDING

Danielle Burke and Kym Snell were supported by a National Institute for Health Research (NIHR) School for Primary Care Research Launching Fellowship. Tim Morris and Ian White were supported by the Medical Research Council (grant numbers MC\_UU\_12023/21 and MC\_UU\_12023/29). This article presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. We acknowledge all researchers, research nurses and staff of the participating centers in the trials contributing to the International Weight Management in Pregnancy (i-WIP) IPD meta-analysis and all members of i-WIP Collaborative Group.









## AUTHOR CONTRIBUTIONS

RR developed the research idea, building on discussions with DJ, IW and TM. AL and RR undertook all the simulation analyses, with advice from DB, JE, KS, DJ and TM. RR and DB performed the applied examples. DJ provided code for the binary outcome simulations, which was extended by AL to deal with different types of coding of the treatment variable. DJ produced the mathematical proof in Section 4. AL and RR drafted the article, and RR revised it following comments and corrections from all other authors. TM produced the figures displaying the simulation results.

## DATA AVAILABILITY STATEMENT

The data that support the simulation findings of this study are available from the corresponding author upon reasonable request. The IPD from the two examples can be recreated from the study two by two tables shown in Tables 2 and 4.

## ORCID

Richard D. Riley  <https://orcid.org/0000-0001-8699-0735>  
Amardeep Legha  <https://orcid.org/0000-0001-7389-5384>  
Dan Jackson  <https://orcid.org/0000-0002-4963-8123>  
Tim P. Morris  <https://orcid.org/0000-0001-5850-3610>  
Joie Ensor  <https://orcid.org/0000-0001-7481-0282>  
Kym I.E. Snell  <https://orcid.org/0000-0001-9373-6591>  
Ian R. White  <https://orcid.org/0000-0002-6718-7661>  
Danielle L. Burke  <https://orcid.org/0000-0003-2803-1151>

## REFERENCES

1. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340:c221.
2. Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: a review of current practice. *Contemp Clin Trials*. 2015;45(Pt A):76-83.
3. Abo-Zaid G, Guo B, Deeks JJ, et al. Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol*. 2013;66(8):865-873.
4. Legha A, Riley RD, Ensor J, Snell KIE, Morris TP, Burke DL. Individual participant data meta-analysis of continuous outcomes: a comparison of approaches for specifying and estimating one-stage models. *Stat Med*. 2018;37(29):4404-4420.
5. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med*. 2018;37(7):1059-1085.
6. Kiefer J, Wolfowitz J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann Math Stat*. 1956;27(4):887-906.
7. Senn S, Hans van Houwelingen and the art of summing up. *Biom J*. 2010;52(1):85-94.
8. White IR, Turner RM, Karahalios A, Salanti G. A comparison of arm-based and contrast-based models for network meta-analysis. *Stat Med*. 2019;38(27):5197-5213.
9. Morris TP, Fisher DJ, Kenward MG, Carpenter JR. Meta-analysis of Gaussian individual patient data: two-stage or not two-stage? *Stat Med*. 2018;37(9):1419-1438.



10. Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Stat Med*. 2017;36(5):855-875.
11. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat Med*. 2000;19(24):3417-3432.
12. Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. *BMJ*. 2001;323(7321):1123-1124.
13. Riley RD, Kauser I, Bland M, et al. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. *Stat Med*. 2013;32(16):2747-2766.
14. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*. 1997;53(3):983-997.
15. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. *Stat Methods Med Res*. 2016;25(6):2858-2877.
16. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Stat Med*. 1993;12(24):2273-2284.
17. Wolfinger R, O'Connell M. Generalized linear mixed models: a pseudo-likelihood approach. *J Stat Comput Simulat*. 1993;48:233-243.
18. Tudur-Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Stat Med*. 2005;24(9):1307-1319.
19. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Med Res Methodol*. 2007;7(1):3.
20. Bates D, Mächler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67:1-48.
21. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics*. 1946;2(6):110-114.
22. Leckie G, Charlton C. Runmlwin - A Program to Run the MLwiN Multilevel Modelling Software from within Stata. *J Stat Softw*. 2013;52(11):1-40.
23. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J Royal Stat Soc Ser A*. 2009;172:137-159.
24. Rogozinska E, Marlin N, Jackson L, et al. Effects of antenatal diet and physical activity on maternal and fetal outcomes: individual patient data meta-analysis and health economic evaluation. *Health Technol Assess*. 2017;21(41):1-158.
25. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc*. 1977;72(358):320-338.
26. Brown HK, Kempton RA. The application of REML in clinical trials. *Stat Med*. 1994;13(16):1601-1617.
27. Stroup WW. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Boca Raton, FL: CRC Press; 2012.
28. Rhodes KM, Turner RM, Higgins JP. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol*. 2015;68(1):52-60.
29. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews. *Int J Epidemiol*. 2012;41(3):818-827.
30. Thomas D, Platt R, Benedetti A. A comparison of analytic approaches for individual patient data meta-analyses with binary outcomes. *BMC Med Res Methodol*. 2017;17(1):28.
31. Piepho HP, Madden LV, Roger J, Payne R, Williams ER. Estimating the variance for heterogeneity in arm-based network meta-analysis. *Pharm Stat*. 2018;17(3):264-277.
32. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods*. 2019;10(1):83-98.
33. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a comparison between DerSimonian-Laird and restricted maximum likelihood. *Stat Methods Med Res*. 2012;21(6):657-659.
34. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med*. 2010;29:3046-3067.
35. Abo-Zaid G, Sauerbrei W, Riley RD. Individual participant data meta-analysis of prognostic factor studies: state of the art? *BMC Med Res Methodol*. 2012;12:56.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Riley RD, Legha A, Jackson D, et al. One-stage individual participant data meta-analysis models for continuous and binary outcomes: Comparison of treatment coding options and estimation methods. *Statistics in Medicine*. 2020;1–20. <https://doi.org/10.1002/sim.8555>