Simulation models for aggregated data meta-analysis: Evaluation of pooling effect sizes and publication biases

Edwin R van den Heuvel¹,², Osama Almalik¹ and Zhuozhao Zhan¹

Abstract
Simulation studies are commonly used to evaluate the performance of newly developed meta-analysis methods. For methodology that is developed for an aggregated data meta-analysis, researchers often resort to simulation of the aggregated data directly, instead of simulating individual participant data from which the aggregated data would be calculated in reality. Clearly, distributional characteristics of the aggregated data statistics may be derived from distributional assumptions of the underlying individual data, but they are often not made explicit in publications. This article provides the distribution of the aggregated data statistics that were derived from a heteroscedastic mixed effects model for continuous individual data and a procedure for directly simulating the aggregated data statistics. We also compare our simulation approach with other simulation approaches used in literature. We describe their theoretical differences and conduct a simulation study for three meta-analysis methods: DerSimonian and Laird method for pooling aggregated study effect sizes and the Trim & Fill and precision-effect test and precision-effect estimate with standard errors method for adjustment of publication bias. We demonstrate that the choice of simulation model for aggregated data may have an impact on (the conclusions of) the performance of the meta-analysis method. We recommend the use of multiple aggregated data simulation models to investigate the sensitivity in the performance of the meta-analysis method. Additionally, we recommend that researchers try to make the individual participant data model explicit and derive from this model the distributional consequences of the aggregated statistics to help select appropriate aggregated data simulation models.

Keywords
DerSimonian and Laird, Trim & Fill, precision-effect test and precision-effect estimate with standard errors, Monte Carlo simulation study, heteroscedastic mixed effects model, meta-analysis

1 Introduction
An aggregated data meta-analysis collects the information \((D_i, S_i, df_i)\) for study \(i\), with \(D_i\) the estimated study effect size of interest (e.g. mean difference, Cohen’s \(d\), and log odds ratio), \(S_i\) is the estimated standard error of the study effect size, and \(df_i\) is the corresponding degrees of freedom for the standard error.² The aggregated data is typically constructed from individual data (although a researcher may not have access to this individual data) and used to determine a pooled effect size and to calculate other aspects related to the quality of the meta-analysis (e.g. measure of heterogeneity, publication bias, and sensitivity analyses).

¹Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, the Netherlands
²Department of Preventive Medicine and Epidemiology, School of Medicine, Boston University, Boston, USA

Corresponding author:
Edwin R van den Heuvel, Department of Mathematics and Computer Science, Eindhoven University of Technology, Den Dolech 2, 5612 AZ, Eindhoven, the Netherlands.
Email: e.r.v.d.heuvel@tue.nl
To investigate the performance of meta-analysis methods, researchers often resort to simulation of the aggregated data \( (D_i, S_i, df_i) \) directly, making certain distributional assumptions.\(^6\) The distribution of \( D_i \) is typically assumed normal and the standard error \( S_i \) is often related to a chi-square distribution.\(^7\) While these distributional assumptions may be plausible, they are not properly supported by a statistical model for the underlying individual data, making it difficult to understand if the simulation model for the aggregated statistics represents the practice appropriately. On the other hand, when individual data is being simulated\(^9\) there is limited discussion on the distributional characteristics of the aggregated data \( (D_i, S_i, df_i) \). Not studying these distributional characteristics, makes a comparison with simulations at the aggregated data level difficult and prohibits a comparison of the observed distribution of the aggregated statistics in practice with the theoretical distribution derived from the underlying individual data. Furthermore, research papers often choose one simulation model and it is typically unknown how well the conclusions of this single simulation model would hold for other choices of simulation model.

This article discusses a (fixed and random effects) heteroscedastic mixed effects model for a continuous outcome at an individual level where two groups (e.g. treatments) are being compared. The model assumes three forms of heterogeneity. The first form is a bivariate random effect on the two mean outcomes across studies, indicating heterogeneity in outcome level between individuals across studies. The second form is a fixed heteroscedastic residual error for the two groups within a study, implying that inter-individual variability within studies can depend on other variables (like treatment). The third form is a random heteroscedastic residual error across studies. The inter-individual variability can change with studies due to the selection of more homogeneous or heterogeneous participants. We believe that this model has not been discussed for meta-analysis purposes so far. Based on this model, we will discuss distributional properties of the aggregated data \( (D_i, S_i, df_i) \), which is useful to be able to simulate study data at an aggregated level directly.

These distributional properties will be compared with similar distributional properties of certain aggregated data simulation models from literature to investigate plausibility of these properties. Secondly, we will conduct a simulation study to investigate the impact of the choice of distributions for the aggregated statistics on some well-known meta-analysis approaches. We selected the very popular pooled estimation method of DerSimonian and Laird for the random effects meta-analysis model,\(^12\) the popular Trim and Fill method\(^13,14\) for adjustments of publication bias in meta-analysis practice,\(^15\) and the precision-effect test and precision-effect estimate with standard errors (PET-PEESE) approach\(^9,10\) also for adjustment of publication bias. The PET-PEESE method has been popular in at least the economics field.\(^11\)

In Section 2, we describe a selection of aggregated data simulation models we have encountered in literature. In Section 3, we propose a statistical model for individual participant data and determine distributional properties of the corresponding aggregated statistics. We also provide a simulation procedure to simulate aggregated data according to this model. We discuss the differences between our approach with the selected models from the literature as well. Section 4 presents a simulation study where the aggregated data simulation models are being compared. We also provide the simulation settings and a selection model for publication bias used in literature. The simulation results are presented in Section 5. Then finally we provide a discussion of the results in Section 6.

## 2 Simulation of aggregated data in literature

In this section, we will provide an overview of a few simulation models that we have encountered in literature to support investigations of the performance of meta-analysis approaches. The simulation models used in literature typically use the well-known random effects model\(^12,25\) for the study effect size \( D_i \) at the aggregated level that is given by

\[
D_i = \theta + U_i + \epsilon_i
\]

with \( \theta \) the mean study effect of interest, \( U_i \sim \mathcal{N}(0, \tau^2) \) a random effect for study \( i \) to accommodate study heterogeneity on the effect size (i.e. \( \theta + U_i \) is the effect size for study \( i \)), \( \epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \) the residual for study \( i \), and \( U_i \) and \( \epsilon_i \) independent. Thus, \( D_i \) is created by drawing \( U_i \) and \( \epsilon_i \) for study \( i \).

Simulations at the aggregated level also have to generate estimates \( S_i^2 \) of the residual variance component \( \sigma_i^2 \). Differences between the aggregated data simulation models in literature are usually differentiated by distributions of the standard deviation \( S_i \). The typical choice of distributions have been the central chi-square for \( S_i^2 \) by Sidik and Jonkman\(^26\) and Brockwell and Gordon,\(^4\) the non-central chi-square for \( S_i^2 \) by Ning et al.,\(^7\) and the gamma distribution for \( S_i \) by Duval and Tweedie.\(^14\) All these aggregated data simulation models assume that \( \sigma_i^2 = \text{VAR}(\epsilon_i | S_i^2) = S_i^2 \), implying that \( S_i^2 \) is drawn first before the study effect \( D_i \) is simulated. Furthermore, the size of a study, which may be represented by the degrees of freedom \( df_i \) of \( S_i^2 \), is not considered as a separate measure by these approaches.
2.1 Central chi-square distributed $S_i^2$

For the simulation of log odds ratios in an aggregated data meta-analysis, Brockwell and Gordon\(^4\) and Sidik and Jonkman\(^26\) simulated the log odds ratio $D_i$ according to model (1). Here $S_i^2$ is selected to be equal to $S_i^2 = 0.25 \chi_i^2^2$, with $\chi_i^2$ a chi-square distributed random variable having only one degrees of freedom (thus effectively choosing $df_i = 1$). They also restricted the value of $S_i^2$ to be within (0.009, 0.6) to conform with a typical distribution of the variances of log odds-ratios in practice.\(^4,26\)

For mean differences, we do not think that we should implement these variance restrictions. In their comparisons of several meta-analysis methods of pooling log odd ratios they did not make use of the $df_i$ in any way. They selected a mean study effect of $\theta = 0.5$ and evaluated 11 different values for the heterogeneity $\tau^2$ in the range of 0 to 0.1.

2.2 Non-central chi-square distributed $S_i^2$

For an investigation of the Trim & Fill method for publication bias, Duval and Tweedie\(^13,14\) simulated the study effect size $S_i^2$. They selected, but we guess that the mean of $S_i^2$ would be equal to $1/3 (= 3 \times 1/9)$ and with a variance equal to $1/27 (= 3 \times (1/9)^2)$, considering the values for precision $S_i^{-1}$ they used in their funnel plots. They assumed that both the mean study effect $\theta$ and heterogeneity $\tau^2$ were equal to zero.

2.3 Gamma distributed $S_i$

For an investigation of the Trim & Fill method for publication bias, Duval and Tweedie\(^13,14\) simulated the study effect size $D_i$ according to model (1). To create the variance $S_i^2$, they draw a normally distributed random variable $Z_i \sim \mathcal{N}(0.25, 0.5)$ and calculated the variance by $S_i^2 = Z_i^2$. The distribution of this variance is equal to a non-central chi-square distribution with non-centrality parameter equal to 1/8 (= $(0.25)^2/0.5$). They did not implement the degrees of freedom $df_i$. The mean effect size was taken equal to $\theta = 0.4$ and the heterogeneity was selected equal to $\tau = 0.5$ and $\tau = 1$.

3 Heteroscedastic mixed effects model

Different data-generating mechanisms in simulation studies for meta-analysis may potentially affect the results when comparing the performances of different meta-analysis methods. In this section, we will describe an individual participant data (IPD) simulation model which allows treatment heterogeneity and heteroscedasticity across treatment arms and across trials. Furthermore, we will show that the distributional properties of the aggregated data $D_i$, $S_i^2$, and $df_i$ for a meta-analysis that we obtained from our heterogeneous and heteroscedastic IPD model deviates from what is normally used in literature.

Let $Y_{ijk}$ be a continuous outcome variable for individual $k = 1, 2, \ldots, n_{ij}$ in treatment group $j = 0, 1$ for study $i = 1, 2, \ldots, m$. Treatment $j = 0$ is the control group and treatment $j = 1$ is the treatment under investigation. Note that we use treatment here, but it could be in principle any binary variable that splits the data in two groups (e.g. sex, symptom, and disease). The statistical model is described as follows:

$$Y_{ijk} = \mu + \beta_j + U_{ij} + \epsilon_{ijk}$$ \hspace{1cm} (2)

with $\mu$ the mean response at the control group, $\beta_0 = 0$ for identifiability purposes, $\beta_1$ the mean treatment effect and the parameter of interest that will play the role of $\theta$ in model (1), $U_{ij}$ the (latent) study heterogeneity for treatment group $j$, and $\epsilon_{ijk}$ a potentially heteroscedastic residual. We will assume that the random elements are normally distributed, such that

$$\epsilon_{ijk} | V_i \sim \mathcal{N}(0, \sigma^2 \exp \{V_i\})$$

$$(U_{0}, U_{1}, V) \sim \mathcal{N}(0, \Sigma)$$ \hspace{1cm} (3)

with $V_i$ a (latent) random heteroscedastic study effect on the residuals in (2), with $\Sigma$ given by

$$\Sigma = \begin{pmatrix}
\tau_0^2 & \rho_{01} \tau_0 \tau_1 & \rho_{02} \tau_0 \tau_2 \\
\rho_{01} \tau_0 \tau_1 & \tau_1^2 & \rho_{12} \tau_1 \tau_2 \\
\rho_{02} \tau_2 \tau_1 & \rho_{12} \tau_1 \tau_2 & \tau_2^2
\end{pmatrix}$$
and $\varepsilon_{ijk}$ and $(U_{10}, U_{1i})$ being independent conditionally on $V_i$. It should be noted that our assumptions of normality are not uncommon for the analysis of clinical trials and observational studies.\(^{18,19}\) Furthermore, heteroscedastic hierarchical models defined by (2) and (3) have been discussed in literature as well,\(^{20,21}\) but not in the context of meta-analysis.

Our model describes study heterogeneity ($U_{ij}$) on the mean response, indicating that the outcome level may vary with studies, but this form of heterogeneity is not the same for both treatment groups. This may not be unrealistic when the treatment intervention would fully cure a disease but the control intervention does not, assuming that variability within a population is larger when it is composed of healthy and diseased individuals. It is this form of heterogeneity that leads to the (well-known) random effects meta-analysis model (1) at the aggregated level. When the heterogeneity would be identical for both treatment groups ($\rho_{\theta 1} = 1$ and $r_2 = \tau_2$), we may only observe heterogeneity at the individual level data for each treatment group, but not necessarily at the aggregated study effect (see Section 3.1).

Furthermore, we also assume that the residual variability, which represents inter-individual variability, is treatment dependent (i.e. $\sigma_0$ and $\sigma_1$ may be different). This is not uncommon in several areas of medicine (e.g. hypertension treatment) where treatment is not just affecting the level of the outcome in a population, but also its variability. Our model also describes a residual heteroscedasticity that is study dependent ($V_i$). This could be related to the selection of participants in a study, leading to a study with either more homogeneous or less homogeneous participants. For instance, a clinical trial with mostly men at a specific age group would show less inter-individual variability than a trial with both sexes and a larger variety in age groups. Combining these different trials in a meta-analysis results in a random residual heteroscedasticity. Here we assume that we do not have any complementary data to investigate or eliminate such heteroscedasticity, implying the necessity of $V_i$ in model (2).

Model (2) can be rewritten into $Y_{ijk} = \mu + \beta_j + U_{ij} + \sigma Z_{ijk} \exp\left\{0.5V_i\right\}$, with $Z_{ijk}$ i.i.d. standard normally distributed residuals (i.e. $Z_{ijk} \sim \mathcal{N}(0, 1)$), which are independent of the random effects $U_{10}$, $U_{1i}$, and $V_i$. This formulation demonstrates immediately that the selected model (2) with assumptions (3) has introduced a correlation between the residuals $\sigma Z_{ijk} \exp\left\{0.5V_i\right\}$ and the random location effects ($U_{10}$, $U_{1i}$). In case we choose $\rho_{\theta 2} = \rho_{12} = 0$, the error structure in (2) and the study treatment heterogeneity $U_{ij}$ will be independent. It also shows that the marginal distribution of $Y_{ijk}$ is not normal anymore, due to the product of a normal and log normal random variable in the residual, unless $\tau_2 = 0$ of course.

Finally, the distribution of $U_{1i} - U_{10}$ and $V_i$ is bivariate normally distributed with mean $\theta$ and a variance–covariance matrix given by

$$
\begin{pmatrix}
\tau_0^2 - 2\rho_{\theta 1}\tau_0 \tau_1 + \tau_1^2 \\
\tau_2(\rho_{12} \tau_1 - \rho_{\theta 2} \tau_0) \\
\tau_2^2
\end{pmatrix}
\begin{pmatrix}
\tau_0^2 - 2\rho_{\theta 1}\tau_0 \tau_1 + \tau_1^2 \\
\tau_2(\rho_{12} \tau_1 - \rho_{\theta 2} \tau_0) \\
\tau_2^2
\end{pmatrix}
$$

The difference $U_{1i} - U_{10}$ is important in Section 3.1 when we determine and discuss the study effect size $D_i$. The correlation coefficient $\rho$ between $U_{1i} - U_{10}$ and $V_i$ is equal to $\rho = (\rho_{12} \tau_1 - \rho_{\theta 2} \tau_0)/\sqrt{\tau_0^2 - 2\rho_{\theta 1}\tau_0 \tau_1 + \tau_1^2}$. We will demonstrate that the study effect size $D_i$ and its corresponding standard error $S_i$ will always be dependent even in the absence of correlation ($\rho = 0$) between $U_{1i} - U_{10}$ and $V_i$ (i.e. $\rho_{\theta 2} \tau_0 = \rho_{12} \tau_1$), unless $\tau_2 = 0$ of course.

### 3.1 Aggregated effects sizes and standard errors

A popular effect size used in meta-analysis calculated for each study is the (raw) mean difference between the two treatment groups,\(^b\) that is,

$$
D_i = \bar{Y}_{ij} - \bar{Y}_{10}
$$

with $\bar{Y}_{ij} = \sum_{k=1}^n Y_{ijk}/n_{ij}$ the average response for study $i$ at treatment $j$. Using model (2), we can rewrite the effect size into

$$
D_i = \beta_j + U_{ij} - U_{10} + \exp\left\{0.5V_i\right\} \left[\sigma_1 \bar{Z}_{ij} - \sigma_0 \bar{Z}_{10}\right]
$$

with $\bar{Z}_{ij} = \sum_{k=1}^n Z_{ijk}/n_{ij}$ the average standardized residuals for treatment $j$ in study $i$. Clearly, model (5) is identical to model (1), with $\theta = \beta_j$, $U_j = U_{ij} - U_{10}$ and $\varepsilon_i = \exp\{0.5V_i\}\{\tau_1 \bar{Z}_{ij} - \tau_0 \bar{Z}_{10}\}$, but the distributional assumptions on normality and independence only agree with our setting when the random heteroscedasticity vanishes (i.e. $\tau_2 = 0$). Under our model (5), the variances $\tau^2$ and $\sigma_i^2$ in model (1) would become equal to $\tau^2 = \tau_0^2 - 2\rho_{\theta 1}\tau_0 \tau_1 + \tau_1^2$ and $\sigma_i^2 = \exp(\tau_2^2/2)\{\sigma_0^2/n_{ij} + \sigma_1^2/n_{1i}\}$, respectively. The difference $U_{1i} - U_{10} \sim \mathcal{N}(0, \tau_0^2 - 2\rho_{\theta 1}\tau_0 \tau_1 + \tau_1^2)$ represents the study heterogeneity of effect sizes, but it disappears when $\rho_{\theta 1} = 1$ and $\tau_0 = \tau_1$.

The conditional distribution of the effect size $D_i$, given the three random effects $U_{10}$, $U_{ij}$, and $V_i$, is given by a normal distribution with mean $\beta_j + U_{ij} - U_{10}$ and variance $\exp\{V_i\}\{\sigma_0^2/n_{ij} + \sigma_1^2/n_{1i}\}$. A larger study will demonstrate a smaller residual variance, since the sample sizes $n_{ij}$ and $n_{1i}$ will reduce the residual variance, but it could potentially be compensated.
by a larger inter-individual variability ($V_i > 0$). However, we do not assume a direct relation between a positive value for $V_i$ and the sample size $n_{i0}$ and $n_{i1}$ (i.e. $P(V_i > 0 | n_{i0}, n_{i1}) = P(V_i > 0)$). The conditional distribution of $D_i$ given only $V_i$ is also given by a normal distribution, but now with mean and variance given by

$$
E(D_i | V_i) = \beta_i + \tau_i^2 \rho \frac{\rho_{i0}}{\rho_{i0} + \rho_{i1}} V_i
$$

$$
\text{VAR}(D_i | V_i) = (1 - \rho_{i0}^2) \tau_i^2 + 2(\rho_{i0} \rho_{i2} - \rho_{i0}) \rho \tau_i + (1 - \rho_{i2}^2) \tau_i^2
$$

$$
+ \exp \{ V_i \} \left[ \frac{\sigma^2_0}{n_{i0}} + \frac{\sigma^2_1}{n_{i1}} \right]
$$

(6)

Finally, the marginal distribution function for $D_i$ is less tractable when $V_i$ is non-degenerate, since the distribution of the product of a lognormal and a normal distributed random variable (i.e. $\exp(0.5V_i) \sigma_i \tilde{Z}_i - \sigma_i \tilde{Z}_i$) is unknown whether they are dependently or independently distributed (as we already mentioned for $Y_{ijk}$). However, we can determine the mean and variance of $D_i$ and they are given by

$$
E(D_i) = \beta_i
$$

$$
\text{VAR}(D_i) = \tau_i^2 - 2 \rho \tau_i + \exp \{ \frac{\tau_i^2}{2} \} \left[ \frac{\sigma^2_0}{n_{i0}} + \frac{\sigma^2_1}{n_{i1}} \right]
$$

(7)

Note that correlation coefficient $\rho = \frac{[\rho_{i2} \tau_i - \rho_{i0} \tau_i]}{\sqrt{\rho_{i0}^2 - 2 \rho \tau_i + \tau_i^2}}$ (which was introduced in the paragraph before Section 3.1) does not play a role in the variance of $D_i$, due to the dependence of $D_i$ on $U_{i0} - U_{i1}$ and $\sigma_i \tilde{Z}_i - \sigma_i \tilde{Z}_i$. Thus the variance is just the sum of the variance of the heterogeneity $U_{i1} - U_{i0}$ and the variance of the residual $\exp(0.5V_i)\sigma_i \tilde{Z}_i - \sigma_i \tilde{Z}_i$.

The estimated standard error $S_i$ for the effect size $D_i$ within one study (thus being unaware of any possible random effect $U_{i1} - U_{i0}$ across studies) can be estimated in different ways. If we do not want to make any assumptions on the equality of residual variances for the two treatment groups in model (2), it is most reasonable to estimate the standard error $S_i$ by

$$
S_i^2 = S_{i0}^2 + S_{i1}^2 / n_{i1},
$$

with $S_{i0}^2 = \sum_{k=1}^{n_i} (Y_{ijk} - \bar{Y}_{ij})^2 / (n_i - 1)$ the sample variance for treatment $j$ in study $i$. Using model (2), we can rewrite this sample variance $S_{i0}^2$ into $S_i^2 \exp(V_i) \chi^2_{n_i-1}$, with $\chi^2_{n_i-1} = \sum_{k=1}^{n_i} (Z_{ijk} - \bar{Z}_j)^2$ a chi-square distributed random variable with $n_i - 1$ degrees of freedom. Thus the variance $S_i^2$ is now distributed as

$$
S_i^2 = \exp \{ V_i \} \left[ \frac{\sigma^2_0 \chi^2_{n_i-1}}{n_{i0}(n_{i0}-1)} + \frac{\sigma^2_1 \chi^2_{n_i-1}}{n_{i1}(n_{i1}-1)} \right]
$$

(8)

with an expectation equal to $E(S_i^2) = \exp(\frac{\tau_i^2}{2}) \left[ \frac{\sigma^2_0}{n_{i0}} + \frac{\sigma^2_1}{n_{i1}} \right]$ and a variance equal to

$$
\text{VAR}(S_i^2) = 2 \exp(2 \frac{\tau_i^2}{2}) \left[ \frac{\sigma^2_0}{n_{i0}^2(n_{i0}-1)} + \frac{\sigma^2_1}{n_{i1}^2(n_{i1}-1)} \right]
$$

$$
+ \exp \left( \frac{\tau_i^2}{2} \right) \left[ \exp \left( \frac{\tau_i^2}{2} \right) - 1 \right] \left[ \frac{\sigma^2_0}{n_{i0}} + \frac{\sigma^2_1}{n_{i1}} \right]^2
$$

(9)

The random variable $\sigma^2_0 \chi^2_{n_{i1}} / (n_{i0}(n_{i0}-1)) + \sigma^2_1 \chi^2_{n_{i1}} / (n_{i1}(n_{i1}-1))$ is approximately distributed as $[\sigma^2_0 / n_{i0} + \sigma^2_1 / n_{i1}] \chi^2_{dfi} / dfi$, with $\chi^2_{dfi}$ a chi-square distributed random variable with $dfi$ degrees of freedom. The number of degrees of freedom is given by Satterthwaite and van den Heuvel

$$
dfi = \frac{\left[ \frac{\sigma^2_0 \chi^2_{n_{i0}}}{n_{i0}^2(n_{i0}-1)} + \frac{\sigma^2_1 \chi^2_{n_{i1}}}{n_{i1}^2(n_{i1}-1)} \right]^2}{\left[ \sigma^2_0 \chi^2_{n_{i0}} / n_{i0}^2(n_{i0}-1)^2 + \sigma^2_1 \chi^2_{n_{i1}} / n_{i1}^2(n_{i1}-1)^2 \right]}
$$

(10)

This number of degrees of freedom is bounded from above by $n_{i0} + n_{i1} - 2$ and from below by $\min\{n_{i0} - 1, n_{i1} - 1\}$. In case the sample sizes are equal ($n_{i0} = n_{i1}$) and the residual variances are equal ($\sigma^2_0 = \sigma^2_1$), the random variable $\chi^2_{dfi}$ becomes chi-square with $dfi = n_{i0} + n_{i1} - 2$ degrees of freedom. The distribution of $S_i^2$ in the general setting (8) using model (2) with assumptions (3) is clearly intractable, but becomes approximately equal to a chi-square distribution when $V_i$ becomes degenerate in zero.

The distribution of $S_i^2$ in (8) seems clearly different from those used in literature, since it is the distribution of a random variable that is a product of a lognormally distributed random variable and a weighted average of two chi-square distributed variables, where the weights depend on the sample sizes. Even if there is no residual heteroscedasticity ($\sigma_0 = \sigma_1$ and $\tau_2 = 0$) and the sample sizes $n_{i0}$ and $n_{i1}$ are equal, $S_i^2$ has a central chi-square distribution with $dfi = n_{i0} + n_{i1} - 2$ degrees of freedom.
and still deviates from the distributions for $S_i^2$ in Section 2. However, we have only studied the distribution function of $S_i^2$ conditionally on the sample sizes $n_0$ and $n_1$. In meta-analysis, these sample sizes are typically known and could therefore be used in the analysis of a meta-analysis (e.g. by calculating appropriate degrees of freedom $d_f$), but for the central chi-square, the non-central chi-square, and gamma distributions in Section 2, these sample sizes are not being mentioned and, in a way, they can be viewed as distribution functions where the sample sizes have been integrated out (i.e. they essentially describe marginal distribution functions of $S_i^2$). However, we do not know if there exist distribution functions for $n_0$ and $n_1$ that would make the marginal distribution of $S_i^2$ in (5) equal to any of the distribution functions in Section 2.

Note that $D_i$ and $S_i^2$ are independent conditionally on $V_i$, due to independence of the random variables $\chi_{i0}^2, \chi_{i1}^2, Z_{i0}, Z_{i1},$ and $U_{i0} - U_{i1}$. Here we have used the assumption of normality to obtain independence between the mean $\bar{Z}_i$ and variance $S_i^2$ (conditionally on $V_i$). Using this independence, (5) and (8), the joint distribution of $D_i$ and $S_i^2$ is now a mixture of the product of the two conditional distribution functions for $D_i|V_i$ and $S_i^2|V_i$. This implies that $D_i$ and $S_i^2$ are never independent, irrespective of the value of $\rho$, unless $V_i$ is degenerated. The covariance between $D_i$ and $S_i^2$ is equal to

$$
\text{COV}(D_i, S_i^2) = \mathbb{E}[(D_i - \beta_1)S_i^2]
$$

$$
= \tau_1^2\{\rho_{12}\tau_1 - \rho_{02}\tau_0\}\mathbb{E}[(V_i \exp\{V_i\}) | \sigma_0^2/n_0 + \sigma_1^2/n_1] + \tau_2^2\{\rho_{01}\tau_1 - \rho_{02}\tau_0\}\exp\{\tau_2^2/2\} \{\sigma_0^2/n_0 + \sigma_1^2/n_1\}
$$

(11)

and depends on the covariance of $U_{i0} - U_{i1}$ and $V_i$ (see Section 3). The covariance in (11) vanishes when $\rho_{12}\tau_1 - \rho_{02}\tau_0 = 0$ even though they remain dependent. This occurs when the heterogeneity and heteroscedasticity are unrelated. Using the variance of $D_i$ in (7) and the variance of $S_i^2$ in (9) the correlation between $D_i$ and $S_i^2$ can be obtained as well. Without heteroscedasticity, we would also obtain a variance $S_i^2$ that is related to a central chi-square distribution when we would change our estimator $S_i^2$ into $S_i^2 = S_p^2[n_0^{-1} + n_1^{-1}]$, with $S_p^2 = \{(n_0 - 1)S_{i0}^2 + (n_1 - 1)S_{i1}^2\}/[n_0 + n_1 - 2]$ the pooled variance within study $i$.

3.2 Procedure for simulating aggregated data

One approach is to simulate individual data via model (2) for all $m$ studies and then calculate all the necessary aggregated data statistics $D_i, S_i^2$, and $d_f$ per study. This requires input values for the sample sizes $m, n_0, n_1$ and all parameters in model (2) and (3). Alternatively, we could simulate the aggregated data statistics directly. Then we only need to set values for the following parameters: sample sizes $m, n_0, n_1$, treatment effect $\beta_1$, variance component $\tau^2 = \tau_0^2 - 2\rho_{01}\tau_1 + \tau_1^2$ for a heterogeneous treatment effect, variance component $\tau_2^2$ for a random heteroscedasticity, correlation coefficient $\rho = \rho_{12}\tau_1 - \rho_{02}\tau_0$ between heterogeneity and heteroscedasticity, and residual variances $\sigma_0^2$ and $\sigma_1^2$. Here we describe the aggregated approach.

Since studies often have different sample sizes we first need to choose a distribution for the sample sizes $n_0$ and $n_1$. Although alternative approaches could be used, we first draw one sample size $n_i$ for study $i$ using a mixture of a Poisson-Gamma distribution (i.e. an overdispersed Poisson). We draw a random variable $\gamma_i$ from a gamma distribution $\Gamma(a, b)$ with shape and scale parameters $a$ and $b$, respectively, (having mean $a/b$ and variance $a/b^2$) and then conditionally on this result we draw a sample size $n_i$ from a Poisson distribution with mean parameter $\lambda_i = \lambda_0 exp\{\gamma_i\}$. Then we split the sample size $n_i$ into $n_0$ and $n_1$ using a Binomial distribution. We assumed that the distribution of $n_0|n_i - \text{Bin}(n_i, p)$, with $p$ a proportion which can take different values to create potential imbalances in sample sizes for the two groups.

Since $D_i$ and $S_i^2$ are independently distributed given $V_i$ and $\chi_{i0}^2, \chi_{i1}^2, V_i$, and $V_i$ are also mutually independent, we now independently draw $\chi_{i0}^2, \chi_{i1}^2$, and $V_i$ for study $i$ using the chi-square distributions with $n_0 - 1$ and $n_1 - 1$ degrees of freedom and the normal distribution $\mathcal{N}(0, \tau_i^2)$, respectively. Then $S_i^2$ can be calculated directly using equation (8) and $D_i$ can be directly drawn from a normal distribution using the mean and variance in (6). Although it may look like we need parameters from the individual participant data model (2), for example, $\rho_{02}$, note that the mean can be rewritten into $\beta_1 + \tau_2^2\tau\rho V_i$ and the variance into $\tau^2(1 - \rho^2) + \exp\{V_i\} \sigma_1^2$, with $\sigma_1^2 = \sigma_0^2/n_0 + \sigma_1^2/n_1$, using only parameters at the aggregated level. The degrees of freedom can be calculated from (10) if this would be needed.

It should be noted that we need two chi-square distributed random variables $\chi_{i0}^2$ and $\chi_{i1}^2$ to be able to obtain the exact distribution of $S_i^2$ given $V_i$ and to calculate the degrees of freedom $d_f$ properly. Although we could argue that one chi-square distributed random variable $\chi_{i0}$ would be appropriate based on Satterthwaite approximation theory, it is not that easy to simulate since we need to know the degrees of freedom $d_f$ in (10). Indeed, this degrees of freedom is based on the data and we do not know its distribution, making a simulation of one chi-square distributed random variable more complicated.
There is another important difference between our aggregated data simulation and the simulations performed in literature. Our simulation would draw $D_i$ independently from $S_i^2$ when conditioned on the random heteroscedastic effect $V_i$. Thus dependency between $D_i$ and $S_i^2$ is indirectly created through the random variable $V_i$ and with a covariance given in (11). In literature, the dependency between $D_i$ and $S_i^2$ is derived directly through $S_i^2$, since the residuals in (1) are typically being drawn from a normal distribution with mean zero and variance $\sigma^2 = S_i^2$. Although $D_i$ and $S_i^2$ are dependent in this way, the covariance between $D_i$ and $S_i^2$ is still equal to zero and therefore differs from our setting. We would also obtain a covariance of zero when either $V_i$ is degenerated (but then $D_i$ and $S_i^2$ are independent) or otherwise when $U_{i1} - U_{i0}$ and $V_i$ are independent. We believe that the assumption $\text{VAR}(e_i|S_i^2) = S_i^2$ in model (1) is highly unlikely in reality, as our IPD model shows, although we are aware that both estimation approaches of DerSimonian and Laird\textsuperscript{12} and Hardy and Thompson\textsuperscript{45} assumed that the standard error $S_i$ was the true standard deviation of the residual $e_i$ in model (1). However, this does not imply that we should also simulate according to this assumption.

4 Simulation study

We will first discuss the simulation settings for our own aggregated data meta-analysis (hereafter referred as the “Mixture” method). Then from these settings we will determine appropriate settings for the simulation models from literature to make the comparison of these simulations with our simulations as fair as possible (see Figure 1). Then after defining the settings, we will discuss the publication bias approach and its settings that were used in the simulation study for the evaluation of the publication bias methods. Our aim is to investigate the influence of different simulation models on the performance of various meta-analysis methods. We will study the pooled estimator of DerSimonian and Laird\textsuperscript{12} and the publication bias approaches PET-PEESE,\textsuperscript{9,10} and Trim & Fill.\textsuperscript{13,14} Details on these approaches can also be found in the Appendix.

4.1 Simulation settings

We selected two levels for the number of studies in the simulated meta-analysis ($m \in \{10, 50\}$), but we chose to fix the parameters for simulating the sample sizes within studies ($\lambda = 100$, $a = b = 1$, $p = 0.5$). This results in the expected sample sizes within each study that are equal to $n_{i0} = n_{i1} = 50$, but they could vary strongly from study to study due to the overdispersed Poisson distribution and vary within study due to a binomial division of the total sample size $n_{i0} + n_{i1}$. We considered no treatment effect and two positive levels of treatment effect ($\beta_i \in \{0, 2, 5\}$). The residual variances for the two treatment groups were kept fixed at the levels $\sigma_{i0}^2 = 100$ and $\sigma_{i1}^2 = 64$, respectively. For the variance of treatment heterogeneity, we selected three levels $\tau^2 \in \{0, 2, 5\}$. The variance for the random heteroscedasticity was selected at two levels, $\tau_i^2 \in \{0, \log(2)\}$, to simulate effect sizes with and without a random heteroscedasticity. Finally, we selected three levels for the correlation coefficient between the heteroscedasticity and heterogeneity $\rho \in \{-0.7, 0, 0.7\}$. Clearly, when either $\tau = 0$ or $\tau_i^2 = 0$, the correlation coefficient $\rho$ would not play any role in the simulation. An overview of the simulation settings for one of the choices of $\beta_i$ can be found in Table 1. Each setting will simulate 1000 meta-analysis studies.

Relating our simulation settings to the parameters of model (1), we obviously obtain that $\theta \in \{0, 2, 5\}$ and $\tau^2 \in \{0, 2, 5\}$, since they are direct translations of $\beta_i$ and $\tau^2$ given above. Thus we could draw a normally distributed random variable $U_i$ having a mean equal to $\theta$ and a variance equal to $\tau^2$ for all the simulation models from the literature. Without random heteroscedasticity ($\tau^2 = 0$), the residual variance in (1) would become equal to $\sigma_i^2 = \sigma_{i0}^2/n_{i0} + \sigma_{i1}^2/n_{i1}$. Using the settings $\sigma_{i0}^2 = 100$, $\sigma_{i1}^2 = 64$, and $n_{i0} = n_{i1} \approx 50$, the variance $\sigma_i^2$ would become on average equal to approximately 3.28, but it will vary across studies due to randomness in $n_{i0}$ and $n_{i1}$. Thus when instead we will draw $\sigma_i^2$ directly from the central chi-square, the non-central chi-square, or the gamma distribution, we need to choose the distributional parameters such that the expected value is close to 3.28. For the central chi-square distribution we use $\sigma_i^2 = 3.28 \chi_i^2$, having a mean of 3.28 and a standard deviation of $3.28\sqrt{2}$. For the non-central chi-square distribution we use $\sigma_i^2 = 3(0.3 + Z_i)^2$, with $Z_i \sim \mathcal{N}(0, 1)$. The expected value is then equal to $3(1 + 0.09) = 3.27$ and the standard deviation is $3\sqrt{2}(1 + 0.18)$. For the gamma distribution, we choose the parameter settings $\Gamma(9, 5)$ for standard error $S_i$, leading to a mean of $9/5 \approx 3.28$ and a standard deviation of $3/5$.

In case of random heteroscedasticity ($\tau^2 = \log(2) > 0$), the residual variance in (6) for the simulation models from the literature would just increase on average with a factor $\sqrt{2} = \exp(0.5 \log(2))$ compared to the non-random heteroscedasticity setting ($\tau_i^2 = 0$). These simulation models do not alter the relation between $D_i$ and $S_i$ when a random heteroscedasticity is introduced. Thus for a setting with random heteroscedasticity, we should just multiply the random variables from the setting without random heteroscedasticity with the factor $\sqrt{2}$ when we use the simulation models with the chi-square distributions and $\sqrt{2}$ when we use simulation model with the gamma distribution. Thus correlation coefficient $\rho$, which affects the correlation between $D_i$ and $S_i$ in our simulation model, does not affect any of the simulation models from the literature.
Table 1. An overview of the simulation settings and their corresponding mean $I^2$ estimated from the simulated data for $\beta_1 = 0$.

| $m$ | $r^2$ | $r_2^2$ | $.I^2$ (5th percentile–95th percentile) | Mixture | $\rho = -0.7$ | $\rho = 0$ | $\rho = 0.7$ |
|-----|-------|---------|------------------------------------------|---------|----------------|-----------|-------------|
| 10  | 0     | 0       | 11.6 (0–47.6)                           | Central | 11.0 (0–46.3)  | –         | –           |
|     |       |         | 11.8 (0–46.8)                           | Non-central | 11.3 (0–48.1) | –         | –           |
|     |       |         | 11.5 (0–47.4)                           | Gamma   | –              | 31.6 (0–68.4) | –           |
|     |       |         |                                           |         | –              | 25.2 (0–63.5) | –           |
|     | log(2)|         | 70.4 (10.0–98.2)                        | Central | 64.3 (0–74.9)  | 65.3 (0–79.6) | 71.4 (0–84.6) |
|     |       |         | 71.4 (12.8–98.5)                        | Non-central | 32.7 (0–70.6) | 57.2 (3.6–84.3) | –           |
|     |       |         | 36.0 (0–72.6)                           | Gamma   | 41.8 (0–74.9)  | 64.2 (21.4–85.7) | 52.2 (0–80.0) |
|     |       |         |                                           |         | 31.5 (0–68.4)  | 41.9 (0–74.8) | –           |
|     |       |         |                                           |         | –              | –           | –           |
| 50  | 0     | 0       | 6.7 (0–26.5)                            | Central | –              | 7.0 (0–27.4)  | –           |
|     |       |         | 6.7 (0–26.2)                            | Non-central | –              | 7.3 (0–27.2)  | –           |
|     |       |         | 6.7 (0–26.7)                            | Gamma   | –              | 36.0 (11.1–54.6) | –           |
|     | log(2)|         | 93.1 (78.2–99.7)                        | Central | –              | 44.1 (21.4–61.8) | –           |
|     |       |         | 92.7 (77.9–99.7)                        | Non-central | –              | 49.4 (29.3–64.7) | 37.1 (13.5–56.0) |
|     |       |         | 44.1 (21.4–61.8)                        | Gamma   | 49.4 (29.3–64.7) | 59.0 (43.2–70.9) | 28.1 (0–49.8) |
|     |       |         |                                           |         | 37.1 (13.5–56.0) | –           | –           |
|     |       |         |                                           |         | 28.1 (0–49.8)  | –           | –           |
|     |       |         |                                           |         | –              | –           | –           |
|     |       |         |                                           |         | –              | –           | –           |
|     |       |         |                                           |         | –              | –           | –           |
| 2   | 0     | 0       | 90.7 (71.3–99.5)                        | Central | –              | 39.8 (15.5–58.5) | –           |
|     |       |         | 90.4 (71.3–99.6)                        | Non-central | 39.8 (15.5–58.5) | 49.4 (29.3–64.7) | 59.0 (43.2–70.9) |
|     |       |         | 39.8 (15.5–58.5)                        | Gamma   | 49.4 (29.3–64.7) | 62.8 (47.0–75.6) | 59.0 (43.2–70.9) |
|     |       |         |                                           |         | 39.8 (15.5–58.5) | 59.0 (43.2–70.9) | 59.0 (43.2–70.9) |
|     | log(2)|         | 97.0 (90.2–99.9)                        | Central | –              | 62.8 (47.0–75.6) | –           |
|     |       |         | 96.9 (89.8–99.9)                        | Non-central | –              | 62.8 (47.0–75.6) | –           |
|     |       |         | 62.8 (47.0–75.6)                        | Gamma   | –              | 62.8 (47.0–75.6) | –           |
|     |       |         |                                           |         | –              | –           | –           |
|     |       |         |                                           |         | –              | –           | –           |
| 5   | 0     | 0       | 95.9 (86.6–99.8)                        | Central | –              | 11.1 (59.7–80.1) | –           |
|     |       |         | 95.7 (86.0–99.8)                        | Non-central | –              | 66.8 (52.4–78.5) | 60.0 (45.3–72.6) |
|     |       |         | 66.8 (52.4–78.5)                        | Gamma   | 71.2 (59.7–80.1) | 71.2 (59.7–80.1) | 49.7 (30.5–65.8) |
|     | log(2)|         |                                           |         | –              | –           | –           |
|     |       |         |                                           |         | –              | –           | –           |
|     |       |         |                                           |         | –              | –           | –           |
|     |       |         |                                           |         | –              | –           | –           |
|     |       |         |                                           |         | –              | –           | –           |

Note: $I^2$ is the proportion of total variation explained by the heterogeneity of effect size.
Figure 1. A scatter plot of $D_i$ against $S_i$ for four different simulation settings (i.e. central chi-square (Section 2.1), gamma (Section 2.3), non-central chi-square (Section 2.2), and the mixture (Section 3.2) of two chi-squares) based on 1000 simulated studies.
(a) $\tau_2^2 = 0$ and $\rho = 0$; (b) $\tau_2^2 = \log(2)$ and $\rho = 0$; (c) $\tau_2^2 = \log(2)$ and $\rho = 0.7$; (d) $\tau_2^2 = \log(2)$ and $\rho = -0.7$.

4.2 Publication bias

The performance of the Trim & Fill and PET-PEESE method in the literature was investigated with the same data-driven publication bias selection approach. For this bias selection approach, studies with a significant effect size are never being excluded from the meta-analysis. For the non-significant studies, a standard uniform distributed random variable is generated $U(0, 1)$ for each study and the study is included in the meta-analysis when the value of the uniform distributed random variable is less than $\pi_{\text{pub}}$.

The level of significance and the parameter $\pi_{\text{pub}}$ are chosen such that the desired publication rate over all studies is $\sim 70\%$. A study is considered significant when the standardized effect size $D_i/S_i$ satisfies $D_i/S_i > z_q$, with $z_q$ the upper $q^{th}$ quantile of the standard normal distribution function. We used this $z$-test because the aggregated data simulation models from the literature did not implement degrees of freedom $df_i$, ruling out the use of a $t$-test. In this way, a comparison of all aggregated data simulation models would be fair. For $\theta = 0$ and $\theta = 2$, we used as significance level of $\alpha = 0.05$, ...
Table 2. Comparison of aggregated data simulation models on bias (×10⁻³) and coverage probability (in %) for the DerSimonian and Laird random effects estimator of the overall treatment effect ($m = 10; \theta = 2; \rho = 0$).

| $\tau^2$ | Central | Non-central | Gamma | Mixture |
|----------|---------|-------------|-------|---------|
|          | Bias (MCSE) | 95% CI | Bias (MCSE) | 95% CI | Bias (MCSE) | 95% CI | Bias (MCSE) | 95% CI |
| 0        | 8.5 (6.6) | 94.1 | 11.1 (6.9) | 94.6 | 26.3 (16.5) | 94.6 | 17.5 (19.5) | 95.0 |
| 2        | −7.0 (19.5) | 89.6 | 22.8 (19.7) | 90.7 | 12.7 (22.8) | 92.4 | 23.3 (24.9) | 94.0 |
| 5        | −14.9 (27.3) | 92.9 | 15.2 (27.4) | 93.1 | 9.2 (29.1) | 94.0 | 27.5 (30.7) | 94.1 |
| $\log(2)$ | 10.2 (7.9) | 94.1 | 13.3 (8.2) | 94.6 | 28.7 (18.0) | 94.6 | −10.9 (19.3) | 94.0 |
| 2        | −6.2 (20.7) | 89.1 | 28.7 (20.9) | 89.9 | 14.6 (24.0) | 92.2 | −12.9 (25.2) | 92.5 |
| 5        | −15.1 (28.4) | 91.9 | 22.2 (28.6) | 92.4 | 10.6 (30.1) | 93.7 | −19.4 (31.2) | 93.1 |

CI: confidence interval; MCSE: Monte Carlo standard error.

but for $\theta = 5$ we used the smaller significance levels of $\alpha = 0.0013$ (three sigma), since some settings had more than 70% significant studies. Thus the publication bias approach depends on the relative performance of other studies. We then pragmatically tuned the parameter $\pi_{pub}$ for the different settings to obtain the target of 70%. Note that $\pi_{pub}$ is different for the different simulation models, due to the differences in generating the variance $S_i^2$.

5 Results

5.1 DerSimonian and Laird

The results of the four aggregated data simulation models are provided in Table 2 for $m = 10$. Here we did not include publication bias, since it is well-known that DerSimonian and Laird’s estimate would become biased. The correlation coefficient $\rho$ for the random heteroscedastic model was set to $\rho = 0$ when we studied heteroscedasticity.

All simulation models show hardly any bias, but there is a small difference in the coverage probability for the 95% confidence intervals of the pooled effect size. The DerSimonian and Laird’s estimator applied on the data simulated by the model with the mixture of chi-square distributed variances shows a coverage probability close to nominal while, under other simulation models, it mostly underestimates the coverage probability when there is heterogeneity. The results with the gamma and mixture of chi-square distributions are close when heteroscedasticity is present. In case the number of studies increases, the difference gets smaller, but remains present (data not shown). Simulations with heteroscedasticity with $\rho = 0$, do not seem to affect the results a lot. The reason is that the heteroscedasticity only increases the residual variance due to the lack of correlation between $D_i$ and $S_i^2$.

5.2 Trim & Fill

For the evaluation of the Trim & Fill method, we implemented the publication bias mechanism. For all aggregated data simulation models, the average number of studies that were included was around 70% (and it ranged from 67% to 73%). Table 3 shows the results on estimation bias and coverage probability for the treatment effect of the Trim & Fill approach under homoscedasticity ($\tau^2 = 0$) for $m = 50$ studies. Results for $m = 10$ are slightly worse (data not shown).

The simulation models show some different results with respect to bias and coverage probability. The simulation models with the central and non-central distribution do not show any real bias when heterogeneity is absent, while the simulation model with the mixture of chi-square distributions shows a small positive bias at treatment effects $\theta = 2$ and $\theta = 5$ (i.e. Trim & Fill does not correct enough). The coverage probability of the pooled effect size for the simulation models with the central and non-central chi-square distribution are conservative, but they are underestimated for the simulation models with the gamma and the mixture of chi-square distributions. In the presence of heterogeneity, all simulation models show real biases that increases with the size of heterogeneity. For the simulation model with the central and non-central chi-square distribution, the absolute bias is highest for treatment effect $\theta = 2$, but for the simulation model with the mixture of chi-square distributions the bias is highest at treatment effect $\theta = 5$. For the simulation model with the gamma distribution, the absolute bias is higher at $\theta = 2$ than at $\theta = 5$ when heterogeneity is limited, but smaller when heterogeneity is at the level of $\tau^2 = 5$. The simulation model with the mixture of chi-square distributions generally provides the highest bias with respect to the other simulation models.

Random heteroscedasticity with $\rho = 0$ does not alter the results observed in Table 3, since it merely increases the residual variance for the simulation models. However, when we select $\rho = -0.7$ or $\rho = 0.7$, the correlation between heterogeneity and heteroscedasticity seems to change the results for our aggregated data simulation model (see Table 4).
A negative correlation increases the bias and lowers the coverage, while a positive correlation eliminates the bias or changes it to a negative bias, but always with highly liberal coverage probabilities. A negative correlation $\rho = -0.7$ introduces a negative correlation between $D_i$ and $S_i$ and diminishes the positive correlation between $D_i$ and $S_i$ that was induced by the publication bias mechanism, masking an asymmetry in the funnel plot that was introduced by the publication bias. A positive correlation $\rho = 0.7$ enhances the publication bias, making the Trim & Fill approach correct stronger. Furthermore, the bias shifted from negative to positive as $\theta$ increases (Table 4).

### 5.3 PET-PEESE

For the evaluation of the PET-PEESE estimator, we also implemented the publication bias mechanism. Since we used the same data as for the Trim & Fill approach, the average number of studies that were included did not change (it ranged from 67% to 73%). Table 5 shows the results on estimation bias and coverage probability for the treatment effect of the PET-PEESE under homoscedasticity ($\tau^2 = 0$) for $m = 50$ studies. Results for $m = 10$ are less severe for the simulation models with the central and non-central chi-square distribution when the treatment effect is restricted ($\theta \in \{0, 2\}$), but in all other cases the results are more extreme for $m = 10$ than for $m = 50$.

For the PET-PEESE, we see biases for the simulation models with the central and non-central chi-square distributions similar to their biases of the Trim & Fill approach, but the coverage probabilities are much smaller when heterogeneity is present. Simulation models using the central and non-central chi-square distributions simulate studies with very small standard errors (i.e. mimicking extremely large studies in a meta-analysis). These extremely large studies, in combination with an independent heterogeneous study effect, have a strong influence on the estimation of the effect size and its standard error in a single meta-analysis. Since the heterogeneity is independent of the standard error, the bias is less affected when averaged out over many meta-analyses, but the confidence interval of the pooled study effect size from a single meta-analysis will not capture the true parameter. The PET-PEESE does not incorporate the heterogeneity in the weights and therefore underestimate the standard error, but the Trim & Fill does account somewhat for the heterogeneity.
Table 5. Comparison of aggregated data simulation models on bias (×10\(^{-3}\)) and coverage probability (in %) for the PET-PEESE estimator (m = 50; \(\tau^2 = 0\)).

| \(\theta\) | \(\tau^2\) | Central Bias (MCSE) 95% CI | Non-central Bias (MCSE) 95% CI | Gamma Bias (MCSE) 95% CI | Mixture Bias (MCSE) 95% CI |
|---|---|---|---|---|---|
| 0 | 0 | 1.2 (2.0) | 98.1 | 1.9 (2.2) | 97.1 | −20.9 (23.4) | 97.2 | 59.5 (34.8) | 97.3 |
| 2 | 212.3 (33.0) | 26.8 | 241.2 (31.4) | 26.7 | 78.7 (35.8) | 91.0 | 187.9 (46.8) | 86.9 |
| 5 | 352.6 (51.6) | 25.7 | 400.4 (49.6) | 25.9 | 208.6 (48.9) | 88.3 | 338.8 (61.8) | 83.1 |
| 2 | 0 | 2.2 (1.1) | 97.7 | 2.5 (1.3) | 95.4 | 14.0 (19.2) | 93.7 | 24.1 (29.1) | 94.2 |
| 2 | 130.7 (31.2) | 26.3 | 157.6 (29.6) | 26.4 | 39.5 (32.1) | 86.1 | 115.0 (43.7) | 84.0 |
| 5 | 382.1 (48.5) | 24.6 | 423.4 (46.5) | 23.5 | 127.2 (44.8) | 81.1 | 223.8 (58.3) | 79.1 |
| 5 | 0 | −5.0 (1.1) | 96.2 | −5.4 (1.4) | 94.3 | −41.5 (12.5) | 94.6 | 68.5 (22.5) | 96.4 |
| 2 | −25.4 (33.2) | 26.8 | 16.1 (30.4) | 28.1 | 62.1 (24.4) | 87.7 | 114.7 (39.2) | 85.1 |
| 5 | 37.0 (50.9) | 27.1 | 114.1 (46.3) | 27.1 | 177.8 (40.5) | 82.4 | 140.0 (57.8) | 78.4 |

PET-PEESE: precision-effect test and precision-effect estimate with standard errors; CI: confidence interval; MCSE: Monte Carlo standard error.

Table 6. Bias (×10\(^{-3}\)) and coverage probability (in %) for the PET-PEESE estimator under the mixture of chi-square distributions with heteroscedasticity (m = 50).

| \(\theta\) | \(\tau^2\) | \(\rho = -0.7\) Bias (MCSE) 95% CI | \(\rho = 0\) Bias (MCSE) 95% CI | \(\rho = 0.7\) Bias (MCSE) 95% CI |
|---|---|---|---|---|
| 0 | 2 | 940.6 (45.1) | 77.9 | 176.3 (42.0) | 87.8 | −571.2 (37.1) | 87.9 |
| 5 | 1501.7 (62.1) | 71.4 | 317.8 (54.8) | 83.7 | −805.3 (46.0) | 83.3 |
| 2 | 2 | 774.7 (41.3) | 69.6 | 99.8 (38.2) | 83.6 | −594.7 (38.3) | 86.9 |
| 5 | 1243.0 (58.6) | 62.3 | 231.6 (51.9) | 78.9 | −887.5 (50.9) | 81.1 |
| 5 | 2 | 754.6 (36.3) | 71.8 | 189.4 (32.3) | 84.7 | −376.3 (33.2) | 86.4 |
| 5 | 1186.3 (55.5) | 63.1 | 250.2 (49.2) | 78.9 | −691.8 (48.7) | 80.1 |

PET-PEESE: precision-effect test and precision-effect estimate with standard errors; CI: confidence interval; MCSE: Monte Carlo standard error.

The simulation model with the gamma distribution and the mixture of chi-square distributions seem to show less bias with the PET-PEESE than with the Trim & Fill. The bias with PET-PEESE is limited and acceptable for the simulation model with the gamma distribution, although the coverage probability is liberal when heterogeneity is present. The coverage probabilities for the simulation model using a mixture of chi-square distributions are similar to the ones obtained by the simulation model with the gamma distribution, while the simulation model with the mixture of chi-square distributions produced larger biases.

Similar to the Trim & Fill method, a random heteroscedasticity with \(\rho = 0\) does not alter the results observed in Table 5 a lot, since it does not change the correlation between \(D_i\) and \(S_i^2\) compared to the setting with homoscedasticity. However, when \(\rho = -0.7\) or \(\rho = 0.7\), we see a pattern in the bias that is similar to the results of the Trim & Fill approach, but now the heteroscedasticity has a stronger effect (see Table 6). A negative correlation \(\rho = -0.7\) reduces the positive correlation between \(D_i\) and \(S_i\) induced by the publication bias mechanism, failing the PET-PEESE to correct for publication bias. A positive correlation \(\rho = 0.7\) enhances the publication bias, making the PET-PEESE over correct.

6 Summary and discussion

We discussed four simulation models for generating aggregated data in a meta-analysis study. All four models used the well-known random effects model\(^{12}\) to generate study effect sizes, but each model used their own distribution function for generating the standard error of the study effect size (a central chi-square, a non-central chi-square, a gamma, and a mixture of chi-square distributions). We showed that the mixture of chi-square distributions would follow naturally from a mixed model for individual participant data (IPD), but the other three distributions could not be formulated directly from such an IPD model. The simulation models with a central chi-square, non-central chi-square, and gamma distribution indirectly created a dependency between the study effect size and its standard error, but without introducing a linear dependency. For the simulation model with the mixture of chi-square distributions, a dependency between the study effect size and its
standard error was created through a random heteroscedasticity. The simulation models using a central and non-central chi-square distribution, simulated studies with very small and relatively large standard errors, mimicking meta-analyses with a large variety of study sizes, while the simulation models using the gamma distribution and the mixture of chi-square distributions typically showed standard errors that belong to a smaller variety in study sizes.

Our simulation study showed that the choice of simulation model affects the conclusion of the applied meta-analysis method. The well-known liberal coverage probabilities of the DerSimonian and Laird method for pooled effect sizes were observed with the simulation models using the central and non-central chi-square distribution, and to a lesser extent with the gamma distribution and the mixture of chi-square distributions. The simulation model using either the gamma distribution or the mixture of chi-square distributions would then lead to the conclusion that the DerSimonian and Laird method had close-to-nominal coverage probabilities.

All simulation models showed some biases for the PET-PEESE approach when heterogeneity is present and the treatment effects are moderate to none. This influence of heterogeneity on the performance of PET-PEESE is in line with literature. However, the results in the literature were obtained with a different simulation model, where the regression equations in (12) and (13) were directly simulated using uniform and normal distributions. The biases reported in the literature are therefore somewhat different from ours, but this strengthens our point that performances of meta-analysis methods studied in the simulation studies are sensitive to choice of simulation model. Indeed, in our own simulation study for the PET-PEESE, we demonstrated that the simulation model with the gamma distribution showed the smallest bias and provided acceptable biases across all settings. Contrary to this, the simulation model using the mixture of chi-square distributions gave biases of the pooled effect size even when heterogeneity was not present. This was not seen by any of the other three simulation models. Finally, PET-PEESE was sensitive to standard errors that belong to a larger variety of study sizes, since it gave very small coverage probabilities for the pooled effect size when the central and non-central chi-square distributions were used in simulation models with study heterogeneity. The simulated meta-analyses typically contained extremely small standard errors that caused an underestimation of the variance of the pooled effect size. To our knowledge, this sensitivity to study sizes has not been presented in literature before.

For the Trim & Fill approach, the simulation model with the gamma distribution showed larger biases in the pooled effect than for the PET-PEESE approach. This is in line with findings that PET-PEESE seems to outperform the Trim & Fill approach. However, the simulation models using the central and non-central chi-square distributions did not show any large discrepancies between Trim & Fill and PET-PEESE, arguing that the Trim & Fill approach may not necessarily be worse than the PET-PEESE approach. On the other hand, the simulation models using the mixture of chi-square distributions showed substantial positive biases with the Trim & Fill approach and moderate negative biases with the PET-PEESE approach when the treatment effect is strong, an observation not mentioned in the literature before. Interestingly though, this simulation model with a mixture of chi-square distributions always showed coverage probabilities very close to nominal for all settings for both the Trim & Fill and the PET-PEESE approach, despite the observed biases, and contrary to the other simulation models which showed (very) liberal coverage probabilities when heterogeneity is present. Indirectly, the variability in coverage probabilities also showed that the simulation model affects the mean squared error, since the coverage probability for the PET-PEESE approach is much smaller than for the Trim & Fill approach when the simulation models with central and non-central chi-square distributions are used with heterogeneous effect sizes. Thus, the mean squared errors are not only affected by the simulation method, but also by the type of simulation model.

When we introduce random heteroscedasticity, an element not studied in literature before, the Trim & Fill and PET-PEESE approach may fail completely in their estimation of the pooled treatment effect. This effect could only be observed with our simulation model, since the simulation models using the central chi-square, non-central chi-square, and the gamma distribution do not have a mechanism to change the joint distribution of the study effect and its standard error. This failure of the publication bias method for meta-analyses with heteroscedastic within-study variances is not unexpected, since the linear correlation between the study effect and its standard error is influenced by the heteroscedasticity. The heteroscedasticity is unrelated to publication biases, and confuses the methods for publication bias.

One of the limitations of our study is that we have only considered one type of study effect (mean difference) for aggregated data meta-analysis that is calculated from one specific heteroscedastic IPD model. Although we do not expect to find markedly different conclusions for other effect sizes, additional investigations may be needed to verify how effect size and simulation models interact. Another limitation is our choice of the Poisson-Gamma distribution for selecting sample sizes of the IPD studies in a meta-analysis that resulted in study sizes of limited variability (despite the use of the overdispersed Poisson distribution). This choice for generating sample sizes (in combination with other settings) gave standard errors of the aggregated measure of effect that was rather similar to the standard errors from the simulation model with the gamma distribution. Investigation of other sample size distributions for the IPD studies may be of interest to see if differences in performance of meta-analysis methods between the simulation model with the mixture of chi-square distributions and the other simulation models can be increased.
Our selected IPD model can be used to generate binary outcomes by either using a threshold on the continuous response or by using a link function that would change the continuous response into a binary response, leading to meta-analysis studies consisting of 2x2 contingency tables.\(^\text{37}\) The cell counts can then be used to create an alternative study effect \(D_i\) (e.g. odds ratio) with its appropriate standard error \(S_i\), but their (joint) distribution would be currently unknown and may be work for future research. Alternatively, cell counts in 2 x 2 contingency tables can also be generated differently\(^\text{33–36}\) using a study-specific effect size at the aggregated level according to the random effects model (5); and with an event probability of the treatment arm based on a logit model. This alternative IPD simulation approach showed that DerSimonian and Laird approach could lead to a biased estimate of the between study variability.\(^\text{37}\) Thus the type of study effect and the different IPD models lead to different ways of simulating aggregated data for studying meta-analysis approaches and consequently could results in a variety of distributions for \((D_i, S_i^2, d_i)\) that may deviate from the choices we studied in this paper (see also the discussion of Jackson and White\(^\text{8}\) on the hidden distributional assumptions for the aggregated statistics used in meta-analyses). It emphasizes the importance of the choice of simulation model for generating aggregated measures of effect.

Our restricted investigation of simulation models already demonstrated that the type of simulation model for generating data for meta-analysis studies can have an influence on the conclusion of how well a particular meta-analysis approach performs. Most publications in literature do not give any or strong arguments for their choice of simulation models and could therefore implicitly bias their results or conclusions. We recommend the use of multiple simulation models when meta-analysis approaches are being studied to provide a more fair view of the performance of the meta-analysis approach. Whenever possible, we additionally recommend to clarify or identify the IPD model underneath the studies of the meta-analysis and describe its relation to the distribution of the aggregated data statistics.

Acknowledgements

We thank the two anonymous reviewers and the editor whose comments/suggestions helped improve this manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors ERvdH and OA disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was supported by the Netherlands Organization for Scientific Research [grant number 023.005.087]. The author ZZ received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Osama Almalik https://orcid.org/0000-0001-6696-2286
Zhuozhao Zhan https://orcid.org/0000-0001-8364-0983

Notes

a. Depending on the meta-analysis approaches or applications, the degrees of freedom \(d_i\) may or may not play a role in the analysis. For instance, the degrees of freedom is used in an investigation of residual heteroscedasticity with Bartlett’s test when studies have (approximately) the same study sizes,\(^\text{1,2}\) but in medical sciences it is frequently ignored.\(^\text{3}\)

b. The raw mean difference is more common to use when all studies measure the outcome on the same scale (e.g. hypertension treatment for blood pressure). In the educational or social sciences, it is common to have scales that may vary with study and it would be more appropriate to divide the mean difference by a standard deviation to create a standardized measure of effect size in the form of Cohen’s \(d\).\(^\text{22}\) We will assume a common scale.

c. Under assumption of homogeneous residual variances within studies \((\sigma_0^2 = \sigma_1^2)\), we may choose the pooled variance \(S_i^2 = [n_0^{-1} + n_1^{-1}][(n_0 - 1)S_0^2 + (n_1 - 1)S_1^2]/(n_0 + n_1 - 2)\), instead of \(S_i^2 = S_{0i}/n_0 + S_{1i}/n_1\). Derivations on this pooled variance using our model (2) with assumptions (3) can be implemented in the same way as the derivations we will apply to \(S_i^2 = S_{0i}/n_0 + S_{1i}^2/n_1\). These derivations are actually somewhat simpler than for our choice of \(S_i^2 = S_{0i}/n_0 + S_{1i}^2/n_1\) and it leads to a central chi-square distribution with \(n_0 + n_1 - 2\) degrees of freedom conditionally on \(V_i\) and when properly scaled.

d. Note that we could exclude values for \(\mu\) in model (2), since this parameter cancels out at an aggregated level when mean differences are calculated.

e. The \(I^2\)'s for other choices of the \(\beta_1\) will be very similar to what is being reported in Table 1.
References

1. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954; 10: 101.
2. Bliss C. *The statistics of bioassay: with special reference to the vitamins*, 1952.
3. Borenstein M. Effect sizes for continuous data. In Cooper H, Hedges L and Valentine J (eds.) *The handbook of research synthesis and meta-analysis*, chapter 12, 2009, pp. 221–235.
4. Brockwell SE and Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med* 2001; 20: 825–840.
5. Chung Y, Rabe-Hesketh S and Choi IH. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat Med* 2013; 32: 4071–4089.
6. Rukhin AL. Estimating heterogeneity variance in meta-analysis. *J R Stat Soc: Ser B (Statistical Methodology)* 2013; 75: 451–469.
7. Ning J, Chen Y and Piao J. Maximum likelihood estimation and EM algorithm of copas-like selection model for publication bias correction. *Bioscience* 2017; 18: 495–504.
8. Jackson D and White IR. When should meta-analysis avoid making hidden normality assumptions? *Biometrical J* 2018; 60: 1040–1058.
9. Stanley TD. Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxf Bull Econ Stat* 2008; 70: 103–127.
10. Stanley TD and Doucouliagos H. Meta-regression approximations to reduce publication selection bias. *Res Synth Methods* 2014; 5: 60–78.
11. Alinaghi N and Reed WR. Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? *Res Synth Methods* 2018; 9: 285–311.
12. DerSimonian R and Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; 7: 177–188.
13. Duval S and Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; 56: 455–463.
14. Duval S and Tweedie R. A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *J Am Stat Assoc* 2000; 95: 89–98.
15. Dadi AF, Wolde HF, Baraki AG et al. Epidemiology of antenatal depression in Africa: a systematic review and meta-analysis. *BMC Pregnancy Childbirth* 2020; 20: 1–13. DOI: 10.1186/s12884-020-02929-5.
16. Liu S and Niu W. Anxiety and depression prevalence in children, adolescents, and young adults with life-limiting conditions. *JAMA Pediatr* 2020; 174: 208.
17. Torous J, Lipschitz J, Ng M et al. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. *J Affect Disord* 2020; 263: 413–419.
18. Moltenberghs G and Kenward M. *Missing data in clinical studies*. 61. Chichester, UK: John Wiley & Sons, 2007.
19. Rosenberg PR. *Observational Studies*. New York: Springer, 2002. DOI: 10.1007/978-1-4757-3692-2.
20. Davidian M and Carroll RJ. Variance function estimation. *J Am Stat Assoc* 1987; 82: 1079–1091.
21. Quintero A and Lesaffre E. Multilevel covariance regression with correlated random effects in the mean and variance structure. *Biometrical J* 2017; 59: 1047–1066.
22. McGaw B and Glass GV. Choice of the metric for effect size in meta-analysis. *Am Educ Res J* 1980; 17: 325–337.
23. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 1946; 2: 110.
24. van den Heuvel et al. *The handbook of research synthesis and meta-analysis*, chapter 12, 2009. pp. 221–235.
25. Hardy RJ and Thompson SG. A likelihood approach to meta-analysis with random effects. *Control Clin Trials* 2002; 23: 619–629.
26. Sidik K and Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med* 2002; 21: 3153–3159.
27. Light RJ and Pillemer DB. *Summing up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press, 1986. DOI: 10.2307/1175260.
28. Schwarzer G et al. meta: an R package for meta-analysis. *R News* 2007; 7: 40–45.
29. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010; 36: 1–48. DOI: 10.18637/jss.v036.i03.
30. Egger M, Smith GD, Schneider M et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315: 629–634.
31. Moreno SG, Sutton AJ, Ades AE et al. Adjusting for publication biases across similar interventions performed well when compared with gold standard data. *J Clin Epidemiol* 2011; 64: 1230–1241.
32. Veroniki AA, Jackson D, Viechtbauer W et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods* 2015; 7: 55–79.
33. Sidik K and Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med* 2007; 26: 1964–1981.
34. Berkey CS, Hoaglin DC, Mosteller F et al. A random-effects regression model for meta-analysis. *Stat Med* 1995; 14: 395–411.
35. Platt RW, Leroux BG and Breslow N. Generalized linear mixed models for meta-analysis. *Stat Med* 1999; 18: 643–654.
36. Knapp G and Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med* 2003; 22: 2693–2710.
37. Almalik O and van den Heuvel ER. Testing homogeneity of effect sizes in pooling 2x2 contingency tables from multiple studies: a comparison of methods. *Cogent Math Stat* 2018; 5: 1478698.
Appendix: Meta-analysis methods

In this section, we briefly describe the three meta-analysis approaches that we studied in our simulation: the DerSimonian and Laird pooled analysis approach and two publication bias adjustment methods: Trim & Fill and PET-PEESE.

DerSimonian and Laird method

DerSimonian and Laird\textsuperscript{12} assumed that the study effect size $D_i$ follows model (1) in combination with the normality assumptions on the random effects and with a known variance $\sigma^2_i$ equal to the observed $S_i^2$. The pooled estimate for $\theta$ is given by the weighted average $\hat{\theta}_{DSL} = \sum_{i=1}^{m} w_i D_i \bigg/ \sum_{i=1}^{m} w_i$, with weight $w_i$ equal to $w_i = [\hat{\theta}^2 + S_i^2]^{-1}$ and $\hat{\theta}^2$ estimator of the variance component for the study effect heterogeneity. They proposed the estimator given by

$$\hat{S}_{DSL}^2 = \max \left\{ 0, \frac{Q - (m - 1)}{\sum_{i=1}^{m} S_i^{-2} - \sum_{i=1}^{m} S_i^{-4} / \sum_{i=1}^{m} S_i^{-2}} \right\}$$

with Cochran’s $Q$-statistic given by $Q = \sum_{i=1}^{m} [S_i^{-2} (D_i - \bar{D})^2]$ and with $\bar{D}$ the weighted average given by $\bar{D} = \sum_{i=1}^{m} [S_i^{-1}] (D_i - \bar{D}) / \sum_{i=1}^{m} S_i^{-2}$. The accompanied standard error $S$ of the pooled estimate $\hat{\theta}_{DSL}$ is given by $S^2 = 1 / \sum_{i=1}^{m} w_i$. DerSimonian and Laird\textsuperscript{12} are not very clear on how to calculate confidence intervals, but based on the work of Cochran,\textsuperscript{1} we assume that the degrees of freedom of $S$ is equal to $m - 1$ and use

$$\hat{\theta}_{DSL} \pm t_{\alpha / 2, m-1} S$$

for the $(1 - \alpha)100\%$ confidence interval, with $t_{\alpha / 2}$ the $(1 - \alpha)$ upper quantile of the $t$-distribution with $d$ degrees of freedom. To obtain DerSimonian and Laird’s pooled estimates $\hat{\theta}_{DSL}$ with its 95% confidence limits we applied the R package “meta”.\textsuperscript{28}

Trim & Fill method

The Trim & fill method has been described in detail by Duval and Tweedie.\textsuperscript{13,14} In short, studies are first ranked based on their distance from the pooled treatment effect estimated by the random effects model (1), that is, ranking distances $|D_i - \bar{\theta}|$. Next, the number of unobserved studies is estimated using for instance estimator $L_0 = [4T_m - m(m + 1)] / [2m - 1]$, where $T_m$ is the Wilcoxon rank-sum test statistic estimated from the ranks of studies with $D_i > \bar{\theta}$ (here we assume that $\hat{\theta}$ is positive and it is more likely that studies with effect sizes below zero are potentially missing). We then trim off the $L_0$ most extreme studies (i.e. studies with positive effect sizes furthest away from zero) and re-estimate the pooled treatment effect $\theta$ without these studies. Then all studies are ranked again, based on their distance to the new pooled estimate, and $L_0$ is recomputed. This procedure is repeated until it stabilizes ($L_0$ does not change anymore) and we obtain a final estimate $\hat{\theta}$ and a final estimate $L_0$ of the number of studies missing. Then we impute $L_0$ studies by mirroring the $L_0$ studies with the highest effect sizes around the final estimate $\hat{\theta}$ and provide each imputed study with the standard error $S_i$ from the mirrored study. After imputation, a final pooled estimate with standard error is provided using the random effects model on all $m + L_0$ studies. We used the function “trimfill” in the R package “metafor,” with the average treatment effect estimated using the random effects model.\textsuperscript{29}

PET-PEESE method

The PET-PEESE method is based on a bias function in case publication bias is present. Under certain conditions, this bias function can be determined explicitly,\textsuperscript{10} but it is a complicated function of the true effect size and approximations are needed. One approximation is based on the Egger’s test,\textsuperscript{30} which investigates the linear relation between the $t$-value $T_i = D_i / S_i$ and the precision $S_i^{-1}$, that is,

$$T_i = \alpha_0 + \alpha_1 S_i^{-1} + e_i$$

with $\alpha_0$ and $\alpha_1$ the intercept and slope parameter, respectively, and with residual $e_i \sim \mathcal{N}(0, \sigma^2)$. An intercept deviating from zero ($\alpha_0 \neq 0$) would indicate publication bias, while the slope in (12) represents the effect size (i.e. $\alpha_1 = \theta$). However, Stanley and Doucouliagos\textsuperscript{10} demonstrated that the OLS estimator $\hat{\alpha}_1$ for (12) would be a biased estimator when it deviates from zero. Thus when null hypothesis $H_0: \alpha_1 = 0$ is rejected, they proposed to use the weighted linear regression between
study effect $D_i$ and variance $S_i^2$, that is,

$$D_i = \gamma_0 + \gamma_1 S_i^2 + e_i$$

with weights $S_i^{-2}$ and report the estimate $\hat{\gamma}_0$ as the overall treatment effect $\theta$. This part is called the Precision Effect Estimate with SE (PEESE). In case null hypothesis $H_0 : \alpha_1 = 0$ is not rejected, the reported estimate is $\hat{\alpha}_1$, which is referred to as the Precision Effect Test (PET). Note that model (13) was suggested earlier by Moreno et al. We used Procedure GLM in SAS to carry out the PET-PEESE method.