Systematic Reviews

METHODOLOGY   **Open Access**

Quantifying heterogeneity in individual participant data meta-analysis with binary outcomes

Bo Chen[1] and Andrea Benedetti[1,2]*

Abstract

Background: In meta-analyses (MA), effect estimates that are pooled together will often be heterogeneous. Determining how substantial heterogeneity is is an important aspect of MA.

Method: We consider how best to quantify heterogeneity in the context of individual participant data meta-analysis (IPD-MA) of binary data. Both two- and one-stage approaches are evaluated via simulation study. We consider conventional I^2 and R^2 statistics estimated via a two-stage approach and R^2 estimated via a one-stage approach. We propose a simulation-based intraclass correlation coefficient (ICC) adapted from Goldstein et al. to estimate the I^2, from the one-stage approach.

Results: Results show that when there is no effect modification, the estimated I^2 from the two-stage model is underestimated, while in the one-stage model, it is overestimated. In the presence of effect modification, the estimated I^2 from the one-stage model has better performance than that from the two-stage model when the prevalence of the outcome is high. The I^2 from the two-stage model is less sensitive to the strength of effect modification when the number of studies is large and prevalence is low.

Conclusions: The simulation-based I^2 based on a one-stage approach has better performance than the conventional I^2 based on a two-stage approach when there is strong effect modification with high prevalence.

Keywords: Individual participant data meta-analysis (IPD-MA), Heterogeneity, Two-stage and one-stage approaches, I^2

Background

Meta-analysis (MA) is a statistical method used to draw an overall conclusion based on the total evidence by reviewing previous research work systematically and pooling effect estimates together [1]. MA is an important tool, widely used, and applied in evidence-based medicine [2].

Individual participant data meta-analyses (IPD-MA), collect line by line participant data from each included study, rather than estimates of the parameter of interest. IPD-MA offer several advantages over aggregate data MA (AD-MA) and are considered the gold standard in meta-analytic techniques [3].

Heterogeneity of effect estimates is an important consideration in both AD-MA and IPD-MA. Heterogeneity exists if the true effects vary across studies more than would be expected by chance alone. The estimated inter-study variance (τ^2) of the parameter of interest is the most direct measure of heterogeneity, but interpretation, particularly deciding what might be a problematic level of heterogeneity, is difficult, despite some practical suggestions [4]. The I^2, originally proposed by Higgins and Thompson, meets three important criteria for any measure of heterogeneity: it monotonically increases with between-study variance; it is not varied by changing the scale; and it is not affected by the number of studies [5]. Importantly, despite some limitations [4], the I^2 remains the most often reported measure of heterogeneity and is easily interpretable, appealing to clinicians.

There are two approaches to analyze the data from IPD-MA: the two-stage approach and the one-stage approach.

*Correspondence: andrea.benedetti@mcgill.ca
[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Purvis Hall, 1020 Pine Avenue West, Montreal, Canada
[2] Respiratory Epidemiology and Clinical Research Unit, McGill University, 2155 Guy St. 4th Floor, Office 412, 24105 Montreal, Canada

In the two-stage approach, each study is analyzed separately, then standard meta-analytic techniques are applied, and heterogeneity may be quantified by usual methods. Alternatively, in the one-stage approach, a mixed model is fit and the data is analyzed altogether, accounting for the correlation that may exist between subjects in the same study and allowing the estimated effect to vary across studies. A review of statistical methods used in IPD-MA of binary outcomes found that most do not report any measure of heterogeneity [6]. While some measures of heterogeneity are easily obtained from a one-stage model, the I^2 is not. Our objective in this work was to consider various approaches to quantifying heterogeneity in IPD-MA of binary outcomes analyzed via the one-stage approach. We propose a method to obtain an I^2 from a one-stage model and evaluate it and other possible measures via simulation study.

Metrics of heterogeneity in IPD-MA with binary outcomes

In this section, we describe various measures of heterogeneity that may be used. Here, we consider that the primary analysis is a one-stage analysis of IPD-MA of dichotomous outcome data. Below, we describe four possible measures of heterogeneity: (1) the conventional I^2 from the corresponding two-stage analysis; (2) the R from the corresponding two-stage analysis; (3) a new metric: the I^2 from the one-stage approach; and (4) the R from the one-stage approach.

Between-study variance (τ^2)

The between-study variance, τ^2, quantifies the heterogeneity in IPD-MA directly. A large value of $\hat{\tau}^2$ indicates that heterogeneity exists among the studies. However, the τ^2 is not ideal, since interpretation is difficult: there is no standard criteria to determine the level of heterogeneity (low, moderate, substantial), because the range is from 0 to ∞ [5, 7]. All other approaches to quantify heterogeneity rely on τ^2.

We might estimate the τ^2 via the two-stage or the one-stage approach. For the two-stage approach, we estimate $\tau^2_{\text{two−stage}}$ via the method described by DerSimonian, Laird, and Whitehead [7–9]:

$$\hat{\tau}^2 = \max\left(\frac{Q - (N - 1)}{\sum_{i=1}^{N} \hat{\omega}_i - \frac{\sum_{i=1}^{N} \hat{\omega}_i^2}{\sum_{i=1}^{N} \hat{\omega}_i}}, 0 \right) \quad (1)$$

where N is the number studies, $\hat{\omega}_i$ is the reciprocal of estimated within-study variance, and Q represents Cochran's heterogeneity statistic [5, 10, 11].

A two-stage analysis proceeds as follows. Consider a MA of a binary outcome in N studies. In the first stage, we fit the logistic regressions in each of the N studies:

$$y_j \sim \text{Bernoulli}(p_j)$$

$$\text{logit}(p_j) = \beta_0 + \beta_1 x_j \quad (2)$$

where p_j is the true response probability for the positive result of the jth individual in this study, β_0 represents the intercept, and x_j indicates their treatment status. This model could be expanded to include effect modifiers.

In the first stage, we obtain $\hat{\beta}_{1i}$ the estimated log odds ratio in study i for $i = 1, 2, .., N$ [12], and the variance of the estimated log odds ratio ($\text{var}(\hat{\beta}_{1i})$) for each one of the N studies.

In the second stage, we consider:

$$\hat{\beta}_{1i} \sim \text{Normal}\left(\beta_1, \tau^2_{\beta_1} + \text{var}(\hat{\beta}_{1i})\right) \quad (3)$$

where $\tau^2_{\beta_1}$ ($\tau^2_{\text{two−stage}}$) represents the respective degree of heterogeneity between studies [12]. Here, we assume the covariance between the parameter estimates (β_{0i} and β_{1i}) are equal to 0, which means that we pool the treatment-outcome associations (β_{1i}) together [12]. This is similar to the classic DerSimonian and Laird random-effects model [8, 13] and allows us to obtain an estimate of the between-study variance $\tau^2_{\text{two−stage}}$ [12], as in Eq. 1.

For the one-stage approach, with binary data, we may estimate the $\tau^2_{\text{one−stage}}$ from a generalized linear mixed model (GLMM) [14–16]. Under the one-stage random-effects model, for each study, a study-specific intercept and treatment effect may be estimated. The study-specific intercept and treatment effects are assumed to come from a bivariate normal distribution [3, 17, 18]. Considering again a MA of a binary outcome in N studies:

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = (\beta_0 + \mu_{0i}) + (\beta_1 + \mu_{1i})x_{ij} \quad (4)$$

$$\begin{bmatrix} \mu_{0i} \\ \mu_{1i} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2 \end{bmatrix} \right) \quad (5)$$

where p_{ij} represents the true response probability for the positive result of the jth individual in the ith study and x_{ij} indicates their treatment status. β_1 is the parameter of interest, which represents the pooled log odds ratio, and τ_1^2 is the between-study variance ($\tau^2_{\text{one−stage}}$).

I^2 statistic

Using a two-stage approach, consider a MA of N studies for the parameter of interest, called θ. Under the assumption that the estimated sampling variances are known and equal for all studies ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_N^2 = \sigma^2 = 1/\omega_i$), Higgins and Thompson [5] defined a measurement function I^2 for quantifying the unexplained heterogeneity, where $E(\theta_i) = \theta, V(\theta_i) = \tau^2, E(\hat{\theta}_i|\theta_i) = \theta_i$, and $V(\hat{\theta}_i|\theta_i) = \sigma^2$.

They proposed to estimate $I^2_{\text{two−stage}}$ as [5]:

$$\hat{I}^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2} \quad (6)$$

where

$$\hat{\sigma}^2 = \frac{N-1}{\sum_{i=1}^{N}\hat{\omega}_i - \frac{\sum_{i=1}^{N}\hat{\omega}_i^2}{\sum_{i=1}^{N}\hat{\omega}_i}}. \tag{7}$$

While the I^2 is usually presented as a percentage varying from 0 to 100%, we present it as a proportion varying from 0 to 1.

In clustered data analyses, the I^2 is very similar to the intraclass correlation coefficient (ICC) [5, 19]. The ICC is the ratio of the between-cluster variance to the total variance in the outcome [20]. It provides a quantitative measure of the amount of heterogeneity across clusters [14]. With binary data, estimating the ICC from a GLMM is possible, though more complicated [14–16]. Several measures have been proposed as estimators of the ICC for binary data [14, 21, 22], though none have been evaluated as measures of inter-study heterogeneity for IPD-MA. Goldstein et al. proposed a simulation-based approach that relies on partitioning the variation in the multilevel model to estimate an ICC for binary outcomes [22]. We propose to adapt this ICC estimator to estimate the I^2 in a one-stage IPD-MA. The algorithm is as follows:

Step 1 Fit a random effects model to the data by using a GLMM, and adjust for possible effect modifiers if desired.

Step 2 Simulate a large number (e.g. $m = 5000$) of values from a normal distribution (e.g. Eq. 5), using the estimated covariance matrix from the multilevel random effect logistic regression fitted in Step 1. We denote these as $\mu_{0,ij}, \mu_{1,ij}$.

Step 3 Using the fitted model from Step 1, we estimate the log odds ratio for each subject ($v_{1,ij} = \hat{\beta}_1 + \mu_{1,ij}$) in the dataset, and then estimate the variance as $v_1 = V(v_{1,ij})$.

Step 4 Estimate p_{ij} by using the fitted model (from Step 1) and simulated random effect values (from Step 2).

- replacing (μ_{0i}, μ_{1i}) by $(\mu_{0,ij}, \mu_{1,ij})$ for each subject
- plugging in the fixed effect estimator from the fitted model and the covariates from the dataset
- taking the inverse logit, we obtain the $\hat{p}_{ij}$ for each individual

Step 5 Using the results from Step 4, it is easy to deduce the variance of the estimated log odds ratio via the Delta method, $v_{2,ij} = \frac{1}{n\hat{p}_{ij}(1-\hat{p}_{ij})}$, where n is the average number of subjects among all of the studies. Finally, we obtain $v_2 = E(v_{2,ij})$.

Step 6 The $I^2_{\text{one-stage}}$ is now estimated as

$$\hat{I}^2 = \frac{v_1}{v_1 + v_2}. \tag{8}$$

R statistic

The R statistic is the square root of the ratio of the variance of the summary statistic from the random-effects model divided by the variance of the summary statistic from the fixed-effects model. It quantifies the inflation of the confidence interval in the presence of inter-study heterogeneity [5]. If the estimated value of R is close to 1, then inference from the random- and fixed-effects models are similar [5]. However, unexplained between-study heterogeneity may exist when the estimate of R is greater than 1. Interpretation of R is difficult, for the same reasons as for τ^2.

For the two-stage approach, we may estimate $R_{\text{two−stage}}$ as

$$\hat{R} = \frac{se(\hat{\beta}_{1R})}{se(\hat{\beta}_{1F})} \tag{9}$$

where $se(\hat{\beta}_{1F})$ is the standard error of the estimated pooled log odds ratio in the fixed-effects model and $se(\hat{\beta}_{1R})$ is the standard error of the estimated pooled log odds ratio in the random-effects model (Eq. 3). For the one-stage approach, we may fit a GLMM with random intercept and fixed slope. The standard error of the estimated β_1 (pooled log odds ratio) from that model with fixed intercept may be denoted as $se(\hat{\beta}_{1F})$. We denote the standard error of the estimated β_1 from model 4 as $se(\hat{\beta}_{1R})$. The estimated $R_{\text{one−stage}}$ from a one-stage model may then be estimated using Eq. 9.

Methods

We use simulations to investigate the performance of (i) conventional I^2 and R^2 based on a two-stage approach and (ii) simulation-based I^2 and R^2 based on a one-stage approach. We generated datasets that consisted of three variables: the binary treatment status, a binary effect modifier, and a binary outcome. Each combination of data generation parameters was used to generate 1000 datasets. We considered 84 distinct data generation scenarios (see Table 1).

Data generation details

The treatment variable is the covariate of primary interest in the IPD-MA; the effect modifier changes the effect of treatment on the outcome when present.

The number of studies and subjects

The number of studies in each dataset was given by N and was set to 15 or 30. The number of the subjects within each study ($n_i, i = 1, ..., N$) followed a log-normal distribution, LN($\sigma_{lg}^2 = 1.5^2, \kappa = 10$), that was truncated at 20 and 2000 (to avoid very small and large studies), and rounded to the nearest integer value.

Table 1 Parameter values for generating datasets

Parameter	Value
The number of studies (N)	15, 30
Prevalence (p_{pre})	0.3, 0.7
True between-study variance (τ^2)	0.5, 1, 1.5
No effect modification (β_w, β_{xw})	(0, 0)
With weak effect modification (β_w, β_{xw})	(1, 1)
With moderate effect modification (β_w, β_{xw})	(1, 3)
With moderate effect modification (β_w, β_{xw})	(2, 1)
With moderate effect modification (β_w, β_{xw})	(2, 3)
With strong effect modification (β_w, β_{xw})	(1, 5)
With strong effect modification (β_w, β_{xw})	(2, 5)

Treatment status

The prevalence of treatment for each study was generated from a uniform distribution, $p_{x_i} \sim U(\theta_{\text{lower}} = 0.4, \theta_{\text{upper}} = 0.6)$. Using these study specific treatment prevalences, we generated n_i random variables from a Bernoulli distribution for each subject, as $x_{ij} \sim \text{Bernoulli}(p_{x_i})$. These x_{ij} represented the treatment status of subject j in study i.

Effect modifier

We generated a binary effect modifier. First, N study-specific effect modifier prevalences were generated as $p_{w_i} \sim \text{Uniform}(\theta_{w_{\text{lower}}} = 0.1, \theta_{w_{\text{upper}}} = 0.9)$. Then, using these probabilities, we obtained the effect modifiers from the Bernoulli distributions, $w_{ij} \sim \text{Bernoulli}(p_{w_i})$.

Outcomes

We generated the outcome y_{ij} based on the generated values of the treatment and effect modifier, as well as the regression coefficients that described the association of each of these with the binary outcome, using the following equation:

$$\text{logit}(p_{ij}) = \beta_0 + \mu_{0,i} + (\beta_1 + \mu_{1,i})x_{ij} + \beta_w w_{ij} + \beta_{xw} x_{ij} w_{ij}. \tag{10}$$

β_0, the fixed intercept, was set based on the given value of prevalence p_{pre}, where $\beta_0 = \log\left(\frac{p_{\text{pre}}}{1-p_{\text{pre}}}\right)$. The prevalence ($p_{\text{pre}}$) was set at 0.3 or 0.7. The random intercepts for individuals within each study were $\mu_{0,i} \sim \text{Normal}(0, \sigma_{\mu_0}^2)$, where $\sigma_{\mu_0}^2$ was given. The true pooled treatment effect was β_1. Furthermore, $\mu_{1,i}$ was the study-specific random effect for the slope, which followed a normal distribution with zero mean and variance τ^2. β_w and $(\beta_w + \beta_{xw})$ were the log odds ratio of the effect modifier in untreated and treated individuals, respectively. The parameter value used to generate the random intercepts ($\sigma_{\mu_0}^2$) was given by 1 and the fixed interested β_1 was given by $\log(1.3)$. The

parameter used to generate the random slopes (τ^2) was set to 0.5, 1, or 1.5.

Using Eq. 10, we obtained p_{ij}. Participant level probabilities of outcome were calculated as $\pi_{ij} = \frac{e^{p_{ij}}}{1-e^{p_{ij}}}$, then y_{ij} was generated from a Bernoulli(π_{ij}) distribution.

Datasets

We contemplated two scenarios, including (i) no effect modification and (ii) effect modification, by varying the data generation parameters (β_w, β_{xw}).

Our rationale was to evaluate each measure of heterogeneity according to the following:

(i) Did the measures of heterogeneity increase with increasing τ^2 in datasets that were generated such that there was no effect modification?

(ii) Did the measures of heterogeneity decrease when the effect modifier and an interaction term between treatment and the effect modifier were included in the model when effect modification was present?

Furthermore, we investigate whether the simulation-based I^2 satisfied the criteria proposed by Higgins et al.: (i) monotonically increasing with increasing between-study variance; (ii) not varied by changing scale; and (iii) not affected by the number of studies [5].

IPD-MA with no effect modification

To generate datasets with no effect modification, we set β_w and β_{xw} to zero.

IDP-MA with effect modification

We varied β_w and β_{xw} to generate datasets with weak or strong effect modification, as presented in Table 1.

Data analysis

For each generated dataset, we considered both two-stage and one-stage approaches to quantifying heterogeneity.

Two-stage approach

In this approach, each study is analyzed separately then pooled together using methods described in "Between-study variance (τ^2)" section [12].

In the first stage, we considered two logistic regression models for each study in the dataset: (i) a crude model ($\text{logit}(p_j) = \beta_0 + \beta_1 x_j$) and (ii) an effect modification model ($\text{logit}(p_j) = \beta_0 + \beta_1 x_j + \beta_2 w_j + \beta_3 x_j w_j$), where p_j was the true response probability for the positive result of the jth individual in this study, β_0 represented the intercept, x_j indicated the treatment status, and w_j was the effect modifier for the jth individual in this study.

When the IPD were generated without effect modification, we fitted the crude model to estimate the pooled treatment effect. When the IPD were generated with effect modification, we considered a crude model and a

model that included the effect modifier, the treatment, and an interaction term between the effect modifier and the treatment to estimate the pooled treatment effect.

In the first stage, we estimated the log odds ratio $\hat{\beta}_{1i}$ ($i = 1, ..., N$) from each study. In the second stage, we pooled these together via the DerSimonian and Laird method and, estimated the between-study variance ($\hat{\tau}^2$) and the pooled treatment effect. We also applied the methods described in "I^2 statistic" and "R statistic" sections to estimate the $I^2_{\text{two−stage}}$ and $R^2_{\text{two−stage}}$ for quantifying the heterogeneity in a two-stage IPD-MA.

One-stage approach
For each generated dataset, we fitted a logistic regression with random intercept and slope for studies, estimated via adaptive Gauss-Hermite quadrature. Similar to the two-stage approach, we considered the following models: (i) a crude model ($\text{logit}(p_{ij}) = (\beta_0 + \mu_{0i}) + (\beta_1 + \mu_{1i})x_{ij}$) and (ii) an effect modification model ($\text{logit}(p_{ij}) = (\beta_0 + \mu_{0i}) + (\beta_1 + \mu_{1i})x_{ij} + \beta_2 w_{ij} + \beta_3 x_{ij} w_{ij}$).

In all models, p_{ij} was the true response probability of disease for the jth individual in the ith study, x_{ij} indicated treatment status, and w_{ij} represented the effect modifier. The random intercept and slope were μ_{0i}, μ_{1i}, such that:

$$\begin{bmatrix} \mu_{0i} \\ \mu_{1i} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2 \end{bmatrix} \right),$$

where $\hat{\tau}_1$ was the estimated between-study variance of the treatment effect. We estimated $I^2_{\text{one−stage}}$ via the simulation-based method, and $R^2_{\text{one−stage}}$ was computed by the ratio of the estimated variance of β_1 under a random slopes model and a fixed slope model with random intercepts, as described in the "I^2 statistic" and "R statistic" sections.

Metrics and performance
We collected I^2, R^2, and τ^2 as estimated from both two-stage and one-stage approaches in each generated dataset for all combinations of data generation parameters. We estimated the median and interquartile range (IQR) from 1000 datasets. If the dataset was generated with effect modification, then the median and IQR of the ratio of the I^2 as estimated from a crude model to that estimated from a model that included the effect modifier and the interaction between the effect modifier and treatment status $\left(\frac{I^2_{\text{emod}}}{I^2_{\text{crude}}} \right)$ was collected. Similar measures were reported for R^2 and τ^2. We collected the ratios because we wanted to investigate the differences in $\hat{I}^2$, $\hat{R}^2$, and $\hat{\tau}^2$ before and after taking effect modification into account. All statistical analysis were carried out in R, version 3.2.3 [23].

Results
With no effect modification
Figure 1 shows the estimated between-study variance $\hat{\tau}^2$ from both the two-stage (dashed line) and the one-stage (dotted line) approaches versus the true between-study variance, τ^2 (solid line). As the true τ^2 increased, the estimated $\hat{\tau}^2$ from both approaches also increased. Compared with the estimated $\hat{\tau}^2$ from a two-stage model, the $\hat{\tau}^2$ from the one-stage model increased more rapidly. The two-stage model always underestimated $\hat{\tau}^2$. On the other hand, the one-stage approach very slightly underestimated $\hat{\tau}^2$ when the true τ^2 was small, and it overestimated $\hat{\tau}^2$ when the true τ^2 was larger than 1.3.

Figure 2 shows the conventional I^2 from the two-stage model (dashed line) and the simulation-based I^2 from one-stage model (dotted line) versus the true between-study variance τ^2. Both measures increased, then leveled off as the true between-study variance increased.

Table 2 presents the median value and IQR of the $\hat{I}^2$ and $\hat{R}^2$ across 1000 datasets from the two-stage and one-stage models for different combinations of data generation parameter values. The median estimated I^2 and R^2 from
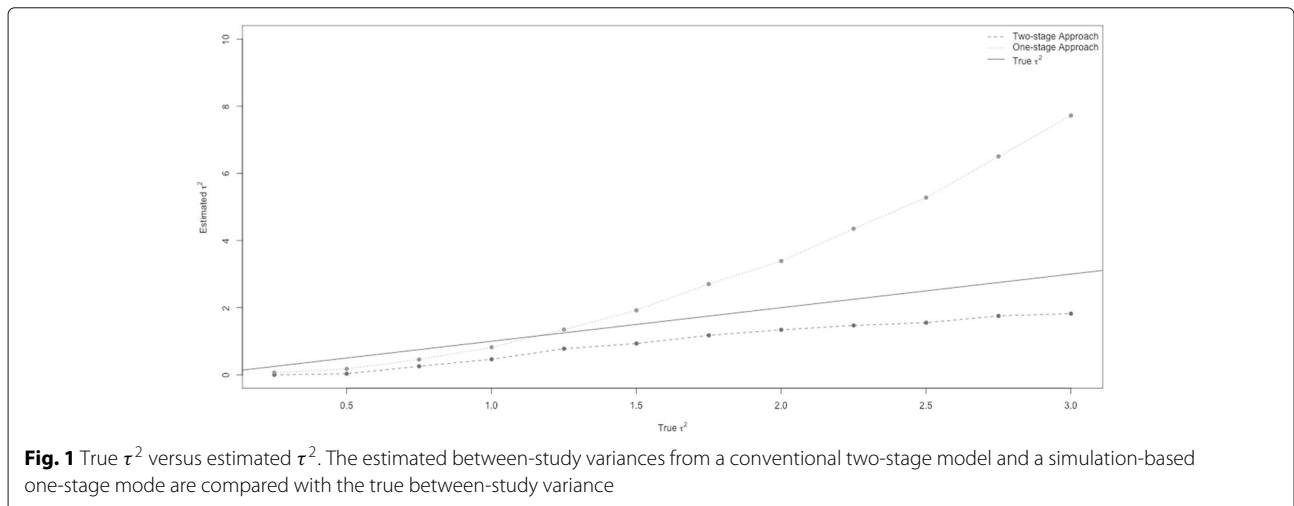
Fig. 1 True τ^2 versus estimated τ^2. The estimated between-study variances from a conventional two-stage model and a simulation-based one-stage mode are compared with the true between-study variance
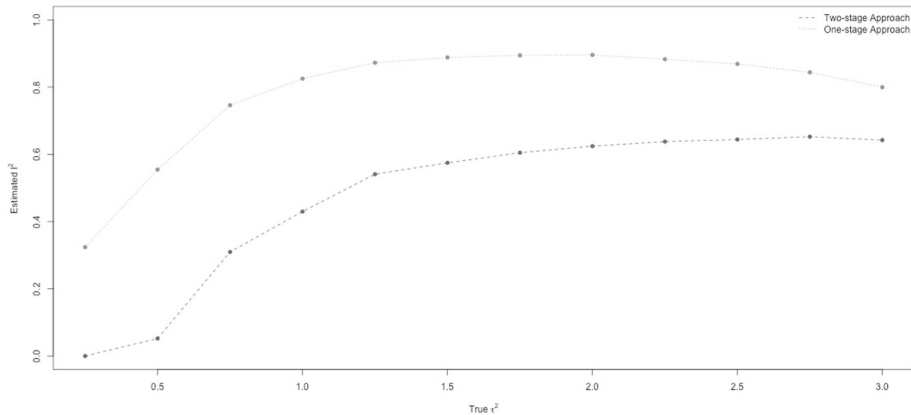
Fig. 2 True τ^2 versus estimated I^2. The estimated I^2 from a conventional two-stage model and a simulation-based one-stage model are compared with the true between-study variance. The dashed line and dotted line represented the estimated I^2 from the two-stage and one-stage models based on its median value across 1000 datasets

both the two-stage and one-stage model increased as the true between-study variance increased. $\hat{I}^2_{\text{two-stage}}$ and $\hat{I}^2_{\text{one-stage}}$ were very similar for $N = 15$ and $N = 30$. However, the R^2 statistic from both approaches slightly increased as the number of studies increased. Varying the

Table 2 Median (IQR) of heterogeneity metrics for the treatment effect when no effect modification was present[a]

τ^2	Prevalence (%)	Number of studies	$I^2_{\text{two-stage}}$	$R^2_{\text{two-stage}}$	$I^2_{\text{one-stage}}$	$R^2_{\text{two-stage}}$
0.5	30	15	0.10 (0.33)	1.22 (0.80)	0.58 (0.41)	1.75 (1.13)
1.0	30	15	0.44 (0.36)	2.30 (1.76)	0.83 (0.15)	3.29 (2.41)
1.5	30	15	0.58 (0.30)	3.22 (2.53)	0.89 (0.09)	5.84 (4.24)
0.5	70	15	0.01 (0.26)	1.00 (0.65)	0.52 (0.47)	1.76 (1.11)
1.0	70	15	0.39 (0.42)	2.12 (1.69)	0.80 (0.18)	3.33 (2.26)
1.5	70	15	0.55 (0.32)	3.07 (2.67)	0.87 (0.11)	5.66 (4.36)
0.5	30	30	0.12 (0.28)	1.29 (0.72)	0.60 (0.32)	1.79 (0.84)
1.0	30	30	0.47 (0.25)	2.44 (1.29)	0.84 (0.09)	3.49 (1.61)
1.5	30	30	0.62 (0.19)	3.54 (1.98)	0.90 (0.05)	6.32 (3.26)
0.5	70	30	0.06 (0.22)	1.14 (0.57)	0.56 (0.34)	1.81 (0.82)
1.0	70	30	0.42 (0.28)	2.30 (1.30)	0.82 (0.11)	3.52 (1.80)
1.5	70	30	0.59 (0.20)	3.37 (1.93)	0.88 (0.07)	6.36 (3.14)

[a]Please note that I^2 is presented here as a proportion varying from 0 to 1, rather than as a percentage

prevalence from 30 to 70% did not affect the estimates of I^2 and R^2 via two- and one-stage models.

Furthermore, $\hat{\tau}^2$ from the two- and one-stage approaches were similar for different prevalence and number of studies (Additional file 1: Table S1).

With effect modification

Table 3 presents the median value and IQR of the ratio of I^2 and R^2 from a model that ignored the effect modifier to one that included the effect modifier and an interaction term between it and the treatment status across 1000 datasets from the two-stage and one-stage approaches with prevalence = 30%. Any measure of heterogeneity should be sensitive to changes in heterogeneity. If we did not account for effect modification when it existed, then heterogeneity might arise due to this effect modification [24]. Hence, if the ratio estimators reported in the Table 3 are less than 1, they indicate good sensitivity of the measure to changing heterogeneity.

When the strength of effect modification was weak, the ratio estimators for $I^2_{\text{two-stage}}$ were well below 1, while the ratio estimators for $I^2_{\text{one-stage}}$ were close to 1. When the strength of effect modification was moderate or strong, we found the ratio estimators for $I^2_{\text{one-stage}}$ were below 1, suggesting the estimated $I^2_{\text{one-stage}}$ was sensitive to changing heterogeneity. When the prevalence increased from 30 to 70%, in the two-stage model, almost all values of the ratio estimator for τ^2 were equal to 1 (Additional file 1: Table S3), as were most values of the ratio estimator for $I^2_{\text{two-stage}}$ (Additional file 1: Table S2). The estimated R^2 from the two-stage model had similar performance. Conversely, most of the ratio estimators for $I^2_{\text{one-stage}}$ were less than 1 when we fixed the prevalence to be 70% and varied other parameter values. However, the convergence rate

Table 3 Sensitivity of heterogeneity measures to accounting for effect modification when prevalence of the outcome was 30%

τ^2	Number of studies	Strength of effect modification[a]	Two-stage approach		One-stage approach	
			$\frac{I^2_{emod}}{I^2_{crude}}$	$\frac{R^2_{emod}}{R^2_{crude}}$	$\frac{I^2_{emod}}{I^2_{crude}}$	$\frac{R^2_{emod}}{R^2_{crude}}$
0.5	15	Weak	0.17 (1.00)	0.89 (0.38)	1.00 (0.41)	0.91 (0.25)
1.0	15	Weak	0.01 (0.57)	0.60 (0.34)	0.99 (0.07)	0.81 (0.25)
1.5	15	Weak	0.06 (0.59)	0.50 (0.30)	0.98 (0.05)	0.75 (0.18)
0.5	15	Moderate	0.02 (1.00)	0.86 (0.39)	0.56 (0.68)	0.86 (0.34)
1.0	15	Moderate	0.08 (0.87)	0.68 (0.46)	0.82 (0.28)	0.98 (0.40)
1.5	15	Moderate	0.34 (0.77)	0.65 (0.39)	0.82 (0.25)	1.06 (0.46)
0.5	15	Strong	0.01 (1.00)	0.82 (0.42)	0.11 (0.23)	0.90 (0.37)
1.0	15	Strong	0.32 (1.00)	0.81 (0.43)	0.20 (0.28)	1.07 (0.57)
1.5	15	Strong	0.42 (0.93)	0.75 (0.46)	0.22 (0.29)	1.31 (0.86)
0.5	30	Weak	0.01 (1.00)	0.78 (0.39)	1.00 (0.27)	0.88 (0.19)
1.0	30	Weak	0.01 (0.36)	0.53 (0.22)	1.00 (0.04)	0.78 (0.16)
1.5	30	Weak	0.16 (0.49)	0.47 (0.22)	0.98 (0.03)	0.74 (0.14)
0.5	30	Moderate	0.01 (1.00)	0.77 (0.40)	0.59 (0.46)	0.86 (0.29)
1.0	30	Moderate	0.01 (0.54)	0.65 (0.30)	0.82 (0.19)	0.99 (0.33)
1.5	30	Moderate	0.19 (0.61)	0.59 (0.30)	0.82 (0.17)	1.07 (0.31)
0.5	30	Strong	0.01 (1.00)	0.79 (0.37)	0.09 (0.15)	0.95 (0.37)
1.0	30	Strong	0.01 (0.68)	0.70 (0.34)	0.16 (0.17)	1.10 (0.46)
1.5	30	Strong	0.33 (0.73)	0.71 (0.33)	0.16 (0.17)	1.23 (0.69)

Median (IQR) was presented

We present the ratios of the measure estimated from a model that ignored the effect modifier to one that included the effect modifier and an interaction term between it and the treatment status

[a]Effect modification was classified as weak when $\beta_w = 1$, $\beta_{xw} = 1$, as moderate when $\beta_w = 1$, $\beta_{xw} = 3$, and as strong when $\beta_w = 2$, $\beta_{xw} = 5$

for one-stage approach decreased as the strength of effect modification became stronger (data not shown).

When the number of studies and prevalence were 30 and 30%, most of ratio estimators for $I^2_{two-stage}$ were equal to 0.01. This occurred because the estimated $\tau^2_{two-stage}$

from the effect modification model was close to zero (Additional file 1: Table S3).

Furthermore, in Additional file 1: Table S3, the ratio estimators for τ^2 in the two-stage model were all less than or equal to 1. However, most of ratio estimators for τ^2 in the one-stage model were larger than 1.

Discussion

IPD-MA are the gold standard of meta-analytic approaches. While the primary objective of most IPD-MA is to estimate pooled treatment effects, quantifying inter-study heterogeneity of those effects is also an important goal. Most statisticians agree that a one-stage approach is the best and most flexible approach to use when analyzing data from IPD-MA. However, how best to quantify inter-study heterogeneity in that case is unclear [3, 5, 12], and most IPD-MA of binary outcomes do not report any measure of heterogeneity [6].

In this work, we considered using usual measures of heterogeneity based on two-stage approaches, as well as novel approaches based on a one-stage model. We evaluated both two-stage and one-stage approaches via simulation studies. In the two-stage approach, we used the usual I^2 and R^2 statistics proposed by Higgins et al. to measure heterogeneity [5]. In the one-stage approach, we adapted a simulation-based ICC proposed by Goldstein et al. to estimate the I^2, as well as considering the R^2 based on the one-stage model.

Our results demonstrated that when there was no effect modification, the estimated τ^2 from the two-stage model was always underestimated. When using a one-stage approach, the estimated τ^2 was underestimated when the true τ^2 was small, but overestimated when the true τ^2 was large. Correspondingly, we may assume that the estimated I^2 from the two-stage model was underestimated, whereas the simulation-based I^2 in the one-stage model was underestimated when inter-study heterogeneity was small and overestimated when it was large. Both the two-stage I^2 and one-stage I^2 increased as the true τ^2 increased.

Including a variable and the interaction of that variable and the treatment of interest when effect modification is present should decrease the estimated between-study heterogeneity. In the presence of weak effect modification, the estimated I^2 from the two-stage model that accounted for the effect modification was less than that from a model that did not. Nevertheless, the estimated I^2 from the one-stage approach that accounted for effect modification quantified heterogeneity well when the strength of effect modification was moderate or strong. The I^2 from the two-stage model was less sensitive to reflect the strength of effect modification when the number of studies was large and prevalence was low. Overall, this suggests that

using the simulation-based I^2 based on one-stage model is preferable.

Differences between measures of heterogeneity in the two-stage and one-stage approaches might be due to real differences in the methods, or because slightly different models were used. In the one-stage approach, we only considered models that fit a random intercept and slope, while the two-stage approaches fit just a random slope. However, these are the approaches most commonly used [6].

Strengths of the work

We have proposed a simulation-based I^2 to use in one-stage IPD-MA of binary outcomes. We have shown that this I^2 satisfies the conditions proposed by Higgins et al., for any measures quantifying heterogeneity, i.e., (i) the measurement function should monotonically increase with increasing between-study variance τ^2 and (ii) not be affected by the number of studies N [5]. Moreover, we have shown that the simnulation-based I^2 is sensitive to changes in heterogeneity.

When the outcome is binary, the within-study variance varies across the studies as between-study variance increases [7]. As a result, the assumption of equal estimated sampling variances across all studies, as in Higgins and Thompson's paper [5], does not hold, and Higgins's I^2 may be biased. For that reason, we would expect the simulation-based I^2 based on the one-stage approach to have better performance than the conventional I^2 based on the two-stage approach.

Using a heterogeneity measure based on the one-stage model is also advantageous, because the one-stage approach allows investigation of patient- and study-level covariates, and the treatment effect can be adjusted for covariates at the participant- or study-level [18]. Moreover, the one-stage model allows MA of dose-response curves (e.g., allowing non-linearity), improves power for interactions [25, 26], and allows better control of effect modification by patient- and study-level covariates than the two-stage approach [3, 17, 27].

While we investigated its performance for binary outcomes, using the ICC as an I^2 for continuous outcomes in the context of a mixed model would be possible, though to our knowledge has not been used like this.

Limitations

There are several limitations in this work. We only considered the ICC estimator proposed by Goldstein to estimate the I^2 in one-stage IPD-MA of binary outcomes. However, there are several other measures that have been proposed as estimators of the ICC for binary data [14, 21, 22]. Wu et al. discussed five different methods to estimate the ICC with binary outcomes: an analysis of variance (ANOVA) estimator, the Fleiss-Cuzick estimator,

the Pearson estimator, an estimator based on generalized estimating equations (GEE), and an estimator from the random intercept logistic model [20]. These could be adapted to estimate I^2 in one-stage IPD-MA. On the other hand, the measure we have proposed is easy to estimate.

Moreover, GLMMs estimated via adaptive quadrature sometimes do not converge in the one-stage model [19]. Indeed, we observed a sometimes important rate of non-convergence when the strength of effect modification was strong and the prevalence was high. Other estimation approaches such as penalized quasi-likelihood (PQL) or Bayesian multilevel models might be interesting to explore [28, 29]. While convergence of PQL is more likely, estimates can be biased with few subjects per cluster, low event rates, or high inter-cluster variability [7, 29, 30].

For the two-stage approaches, we estimated τ^2 via the method of moments estimator of DerSimonian and Laird, despite more recent evidence suggesting that the Paule and Mandel estimator is preferred [31].

Finally, we invsestigated a finite number of data generation parameters. In particular, we considered datasets of 15 or 30 studies, whereas it may have been interesting to consider fewer (e.g., 5).

Conclusion

When a one-stage approach for IDP-MA of binary outcomes is preferred, heterogeneity should be quantified via the model estimated. In that case, we have proposed a simulation-based I^2 that performs as well or better than the conventional I^2 based on a two-stage approach. The R^2 based on the one-stage model also performed adequately but is more difficult to interpret.

Additional file

Additional file 1: Table S1. The Median (IQR) of the estimated τ^2 when no effect modification was present. **Table S2.** Sensitivity of heterogeneity measures to accounting for effect modification when prevalence of the outcome was 70%. Median (IQR) was presented. **Table S3.** Sensitivity of heterogeneity measures $\left(\frac{\tau^2_{emod}}{\tau^2_{crude}}\right)$ to accounting for effect modification. Median (IQR) was presented. (PDF 145 kb)

Availability of data and materials

R code for dataset generation is avaliable by request.

Authors' contributions

AB gave the original idea to conduct this research work. BC implemented this research work under the guidance of AB. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. BMC Med Res Methodol. 2005;5(14):14–14.
2. What Is Meta-analysis? http://www.bandolier.org.uk/painres/download/whatis/Meta-An.pdf. Accessed 5 Mar 2016.
3. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:521–5.
4. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on i2 in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8(1):79.
5. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58.
6. Thomas D, Radji S, Benedetti A. Systematic review of methods for individual patient data meta-analysis with binary outcomes. BMC Med Res Methodol. 2014;14(1):79.
7. Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. Stat Med. 2014;33(16):2701–17.
8. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88.
9. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Stat Med. 1991;10(11):1665–77.
10. Biggerstaff BJ, Jackson D. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. Stat Med. 2008;27(29):6093–110.
11. Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. Stat Med. 2012;31(29):3805–20.
12. Debray TPA, Moons KGM, Abo-Zaid GMA, Koffijberg H, Riley RD. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? pone. 2013;8(4):60650.
13. Jackson D, Bowden R, Baker J. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? J Stat Plan Infer. 2010;48(4):961.
14. Ridout MS, Demetrio CGB, Firth D. Estimating intraclass correlation for binary data. Biometrics. 1999;55(1):137–48.
15. Yelland LN, Salter AB, Ryan P, Laurence CO. Adjusted intraclass correlation coefficients for binary data: methods and estimates from a cluster-randomized trial in primary care. Clin Trials. 2011;8(1):48–58.
16. Thomson A, Hayes R, Cousens S. Measures of between-cluster variability in cluster randomized trials with binary outcomes. Stat Med. 2009;28(12):1739–51.
17. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Stat Med. 2008;27(11):1870–93.
18. Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Stat Med. 2001;20(15):2219–41.
19. Diggle P. Analysis of longitudinal data. Oxford: Oxford University Press; 2002.
20. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp Clin Trials. 2012;33(5):869–80.
21. Browne WJ, Subramanian SV, Jones K, Goldstein H. Variance partitioning in multilevel logistic models that exhibit overdispersion. J R Stat Soc Ser A (Stat Soc). 2005;168(3):599–613.
22. Goldstein H, Browne W, Rasbash J. Partitioning variation in multilevel models. Underst Stat. 2002;1(4):223.
23. The R project for statistical computing. https://www.r-project.org/. Accessed 5 Jan 2016.
24. Gentles SJ, Stacey D, Bennett C, Alshurafa M, Walter SD. Factors explaining the heterogeneity of effects of patient decision aids on knowledge of outcome probabilities: a systematic review sub-analysis. Syst Rev. 2013;2:95.
25. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. J Clin Epidemiol. 2002;55(1):86–94.
26. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol. 2004;57(7):683–97.
27. Riley RD. Commentary: like it and lump it? Meta-analysis using individual participant data. Int J Epidemiol. 2010;39(5):1359–61.
28. Brewlow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993;88:421.
29. Jang W, Lim J. A numerical study of PQL estimation biases in generalized linear mixed models under heterogeneity of random effects. Commun Stat Simul Comput. 2009;38:692–702.
30. Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. J Am Stat Assoc. 1996;91(435):1007.
31. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JP, Langan D, Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Meth. 2016;7(1):55–79.