https://doi.org/10.26219/heal.aueb.6502

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES AND TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Expanding the application of the I-squared statistic to characterize heterogeneity in treatment effects within cluster randomized trials, multi-center randomized trials, and meta-analyses using individual patient data

Alexandros Saloustros

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens
July 2023

https://doi.org/10.26219/heal.aueb.6502

https://doi.org/10.26219/heal.aueb.6502

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντά μου, Β. Βασδέκη για την ανεκτίμητη καθοδήγηση, την εμπιστοσύνη και την εμπειρογνωμοσύνη του καθ' όλη τη διάρκεια της παραγωγής αυτής της διατριβής. Οι πολύτιμες συμβουλές και η εποικοδομητική κριτική του συνέβαλαν σημαντικά στη βελτίωση και την ολοκλήρωση αυτής της ερευνητικής εργασίας.

 https://doi.org/10.26219/heal.aueb.6502

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Κάτοχος πτυχίου Ψυχολογίας από το Πάντειο Πανεπιστήμιο και μεταπτυχιακού τίτλου στην Ποσοτική Ψυχολογία από το Πανεπιστήμιο της Ουτρέχτης. Αυτήν την περίοδο, παρακολουθώ πρόγραμμα μεταπτυχιακών σπουδών στην Εφαρμοσμένη Στατιστική στο Οικονομικό Πανεπιστήμιο Αθηνών, ενώ ταυτόχρονα εργάζομαι ως Βιοστατιστικός για μια φαρμακευτική εταιρεία.

https://doi.org/10.26219/heal.aueb.6502

ABSTRACT

Alexander Saloustros

Expanding the application of the I-squared statistic to characterize heterogeneity in treatment effects within cluster randomized trials, multi-center randomized trials, and meta-analyses using individual patient data.

July 2023

The investigation of treatment effect heterogeneity in meta-analyses aims to determine whether treatment effects vary among different studies. In aggregate level data meta-analyses, it is customary to quantify the extent of treatment effect heterogeneity using a well-understood and intuitive concept called the I-squared statistic. Additionally, it is important to examine whether the effect of a treatment differs across clusters in a cluster randomized trial or across centers in a multi-center randomized trial during the analysis stage. When conducting trials that employ cross-over and other randomized designs, where clusters or centers are exposed to both treatment and control conditions, it becomes possible to detect variations in treatment effects. Within this research, I evaluate an alternative I-squared measure, which helps quantify the extent of heterogeneity in treatment effects across clusters or centers in randomized trials. Additionally, I showcase the applicability of this methodology in estimating treatment effect heterogeneity during meta-analysis using individual patient data.

https://doi.org/10.26219/heal.aueb.6502

ΠΕΡΙΛΗΨΗ

Αλέξανδρος Σαλούστρος

Επεκτείνοντας την εφαρμογή του I-τετραγώνου για τη ποσοτικοποίηση της ετερογένειας στη θεραπευτική επίδραση σε κλινικές δοκιμές τυχαιοποιημένων ομάδων (cluster randomized trials), πολυκεντρικά τυχαιοποιημένων κλινικών δοκιμών (multi-center randomized trials) και μετα-αναλύσεις χρησιμοποιώντας ατομικά δεδομένα από ασθενείς.

Ιούλιος 2023

Ο έλεγχος της ομοιογένειας της επίδρασης της θεραπείας στις μετα-αναλύσεις στοχεύει στο να προσδιορίσει εάν η επίδραση της θεραπείας διαφερει σημαντικά μεταξύ διαφορετικών μελετών. Στις μετα-αναλύσεις με συγκεντρωτικά δεδομένα (aggregate data) είναι συνήθες να ποσοτικοποιείται η έκταση της ομοιογένειας της θεραπευτικής επίδρασης χρησιμοποιώντας μια εύκολα κατανοητή έννοια που ονομάζεται I-τετράγωνο. Επιπλέον, είναι σημαντικό να εξεταστεί εάν το αποτέλεσμα μιας αγωγής διαφέρει μεταξύ των συστάδων σε μια τυχαιοποιημένη κλινική δοκιμή ή μεταξύ των κέντρων σε μια τυχαιοποιημένη πολυκεντρική δοκιμή. Όταν διεξάγονται δοκιμές που χρησιμοποιούν διασταυρούμενη (cross over) σχεδίαση και άλλα τυχαιοποιημένα σχέδια, όπου οι συστάδες ή τα κέντρα εκτίθενται και στις δύο συνθήκες αγωγής και ελέγχου, προκύπτει η ανάγκη εξέτασης διαφορών των φαρμακευτικών επίδρασεων . Στα πλαίσια αυτής της έρευνας, αξιολογούμε μια εναλλακτική μεθοδο μέτρησης για το I-τετράγωνο, το οποίο βοηθά στον προσδιορισμό της έκτασης της ετερογενειας στην επίδραση της θεραπείας μεταξύ των συστάδων ή των κέντρων σε τυχαιοποιημένες δοκιμές. Επιπλέον, παρουσιάζουμε την εφαρμοσιμότητα αυτής της μεθοδολογίας στην εκτίμηση της ομοιογένειας του αποτελέσματος της αγωγής για μετα-ανάλυση χρησιμοποιώντας ατομικά δεδομένα ασθενών.

https://doi.org/10.26219/heal.aueb.6502

Table of Contents

Table of Contents

https://doi.org/10.26219/heal.aueb.6502

Tables

 https://doi.org/10.26219/heal.aueb.6502

Figures

 https://doi.org/10.26219/heal.aueb.6502

1. INTRODUCTION

It is widely acknowledged that treatment effects can vary among different populations (Hingorani et al., 2013). This variation in treatment effectiveness is known as treatment effect heterogeneity. It is possible to expect heterogeneity in treatment effects based on various patient characteristics, such as age, severity of disease, and more, as well as across different settings, such as geographical locations or different centers (Brown, 2015). When observing treatment effects across different groups, the observed variation comprises a combination of "true" treatment effect variation and variation caused by sampling error. If there is a need to estimate the extent of variation in the "true" treatment effects and several analysis methods allow for distinguishing between these different sources of variation (Feaster et al., 2011, Fisher et al., 2011; Hemming et al., 2021).

The exploration of treatment effect heterogeneity is frequently conducted in systematic reviews and meta-analyses. This term, often associated with meta-analyses, refers to the potential variations in treatment effects across different studies (Egger et al., 1997). It is not unusual to observe treatment effect heterogeneity across studies, as treatments may have differing impacts on various patient groups or in different settings. Additionally, variations in estimated treatment effects across studies can arise from differences in study conduct, such as subtle discrepancies in intervention delivery methods.

In meta-analyses that employ study-level aggregate data, treatment effect heterogeneity is addressed through the utilization of random effects models. These models incorporate parameters to account for the interaction between treatment effect and study, effectively capturing differences in treatment outcomes across studies. The I-squared statistic is a suggested metric for assessing the magnitude of treatment effect heterogeneity. This statistic offers several advantages, including its independence from the specific treatment effect metric (such as relative risk or odds ratio), and its intuitive and easily comprehensible interpretation (Higgins et al., 2002).

With the increasing availability of individual-level data from studies, one-stage and two-stage approaches can be adopted for meta-analyses. In the one-stage approach, a generalized linear mixed model is utilized, incorporating random effects to account for interactions between studies and treatments (Higgins et al., 2001; Burke et al., 2017). Alternatively, the two-stage approach involves analyzing each individual study to estimate treatment effects and subsequently pooling the results using a random-effects meta-analysis. The estimation of treatment effect heterogeneity through the I-squared statistic occurs naturally in the two-

https://doi.org/10.26219/heal.aueb.6502

stage approach. Nevertheless, at present, there is no endorsed approach available for characterizing the magnitude of treatment effect heterogeneity in the one-stage methodology (Tierney et al., 2015; Kivimäki et al., 2019; Merino et al., 2019).

Treatment effect heterogeneity may also be observed within a randomized trial (Donner et al., 2000). In cluster randomized trials, a common analytical approach involves utilizing a generalized linear mixed model. It's worth noting that the random effects included in this model, often referred to as random intercepts, capture the differences between clusters before treatment (Senn, 2014). This differs from meta-analyses where random effects describe the variations between studies in their response to treatment.

In parallel cluster trials, although it is expected that treatment effects may vary across clusters, the trial design does not permit estimation of this variation. This limitation arises due to the fact that clusters are either fully exposed or unexposed to the intervention, confounding the effects of clusters with any differential treatment effects. However, treatment-by-cluster heterogeneity can be estimated in cluster trials where treatment is crossed with clusters, such as in stepped-wedge and cross-over cluster trials (Campbell, 2019; Arnup, 2017). In the context of a generalized linear mixed model, this is accomplished by integrating a stochastic interaction between clusters and treatments, thereby considering the variability in treatment effects between clusters (Hemming et al., 2018; Hughes et al., 2015).

In individually randomized trials, interest in treatment effect heterogeneity typically revolves around a small number of sub-groups, such as examining whether the treatment works differently for males and females (Donegan et al., 2015). Sub-group analyses are commonly employed to investigate treatment effect heterogeneity in these cases, formally differentiating true variation across groups from sampling variation using interaction tests. However, in pragmatic trials, there may also be interest in exploring whether treatment effects differ across centers in a multicenter randomized trial. Variations in treatment effects across centers could potentially be attributed to differences in fidelity or surgeon skill, indicating subtle variations in intervention delivery. Moreover, in multi-center randomized trials, randomization is often stratified by center, necessitating the inclusion of a random center effect in the analysis, similar to how a random cluster effect is included in cluster randomized trials (Kahan et al., 2013). Therefore, the analytical setup closely resembles that of cluster randomized trials, and in the context of a generalized linear mixed model, modeling center-by-treatment effect heterogeneity can be achieved by incorporating a random center-by-treatment interaction.

https://doi.org/10.26219/heal.aueb.6502

Thus, describing and modeling treatment effect heterogeneity in systematic reviews and aggregate level data meta-analyses is a common practice. However, when analyzing individual-level data from systematic reviews, no equivalent measure to the I-squared statistic existed. Additionally, treatment effect heterogeneity could be of interest in primary studies, such as multicenter randomized trials and cluster randomized trials. Furthermore, there is a strong parallel between how all these types of data would be analyzed in a generalized linear mixed model framework.

Therefore, in this study, we evaluate an alternative method to calculate the I-square statistic has proposed by Hemming et al. (2021). This metric is derived from the generalized linear mixed model to quantify treatment effect heterogeneity in individual patient data meta-analyses, multicenter individually randomized trials, and cluster randomized trials. I investigate the performance of this metric compared to the conventional I-squared metric in meta-analyses and evaluate its effectiveness across various scenarios, including differences in study size and heterogeneity levels. Additionally, I provide illustrative examples of applying this metric in the context of individual patient data meta-analysis. My analysis assumes a superiority comparison of two treatments, a continuous outcome, cross-sectional designs, and an exchangeable correlation structure.

https://doi.org/10.26219/heal.aueb.6502

2. BACKGROUND

When combining treatment effect estimates from numerous studies with available aggregate data, the calculation of treatment effect heterogeneity (across studies) is achieved through a random effects meta-analysis (Higgins et al., 2002). Two commonly used statistical models for meta-analysis are the fixed-effect model and the random-effects model. A fixed effect meta-analysis assumes that all studies are sample of a single study and they reflect an underlying "true" effect size. A random effects meta-analysis assumes that studies are conducted independently and each reflect a separate "true" effect size (equations i-iii). This results in a between-study variance (τ^2) which needs to be estimated. Commonly have very few studies in meta-analysis Too little data to estimate the between study variation reliably This does not automatically mean a fixed effect model ignoring the variation is appropriate. Random effects analysis is preferable where there are conflicting results or differences in patient characteristics between studies. When between-study variance (τ^2) is zero, the random effects model reduces to a fixed effect model.

We have thoroughly addressed the significance of variance due to its central role in computations and its impact on the divergence between the models. When computing a summary effect, it becomes imperative to allocate greater weight to studies that provide a more precise estimation of the effect.

In the fixed-effect model, the comprehensive study error variance pertaining to each study's observed mean (Y_j) for study j, around the common mean (θ) is solely the within-study error variance. Generally, the variance of any study's observed score Y_j relative to V_j. Hence, the allocation of weight to each study using the inverse variance scheme is straightforward.

$$w_j = \frac{1}{V_j} \qquad \text{(i)}$$

On the contrary, in the random-effects model, where study means are employed to estimate μ (the grand mean), two sources of variance exist, leading to two components within the overall study error variance. First, the divergence between the observed value Y for a study (j) and its true value (θ_j) is due to within-study error variance, Vi. Second, the true value (θ_j) for each study deviates from μ due to between-study variance. The variance of any observed score Y_j about μ becomes V_j plus τ^2. As a result, the weight assigned to each study within

https://doi.org/10.26219/heal.aueb.6502

the inverse variance scheme varies accordingly.

$$w_j = \frac{1}{V_j + \tau^2} \tag{ii}$$

It's important to observe that while the within-study error variance (V_j) is distinct for each study, the between-study variance (τ^2) is consistent across all studies.

After allocating weights to each study, the combined effect is determined through a weighted mean spanning all studies. If W_j represents the weight attributed to study j, Y_j stands for the effect size observed in study i, and k denotes the total number of studies in the analysis, then the computation of the weighted mean (θ) is as follows:

$$\theta = \frac{\sum_{j=1}^{K} W_j Y_j}{\sum_{j=1}^{K} W_j} \tag{iii}$$

Certain texts denote W as the fixed-effect weight and W∗ as the random-effects weight. In this context, we've employed W for both to emphasize the parallel between the two models. In both scenarios, the weight hinges on the reciprocal of the overall study error variance, and the sole distinction between the models lies in the characterization of that variance.

Using a random effects approach does not mean heterogeneity can be forgotten about; instead, the reasons and extent of the heterogeneity should be explored. The magnitude of treatment effect heterogeneity can be described using the I-squared statistic, and its presence can be tested using a statistical test known as the Q-statistic test (Veroniki, 2015). The following sections outline these concepts:

$$Q = \sum_{j=1}^{K} \hat{\sigma}_j^{-2} \left(\hat{\theta}_j - \hat{\theta}\right) \tag{iv}$$

In the equations provided, the variable j represents each individual trial (with a total of K trial). The symbol $\hat{\theta}_j$ denotes the estimated treatment effect for study j, while $\hat{\sigma}_j^{-2}$represents the precision of the estimated treatment effect for that study. The weighted pooled estimate of the treatment effect across all studies is denoted by $\hat{\theta}$. In a random effects meta-analysis, the pooled treatment effect is calculated as $\hat{\theta} = \frac{\sum \hat{\theta}_j w_j}{\sum w_j}$ where $w_j = \frac{1}{\hat{\sigma}_j^{-2} + \hat{\tau}^2}$.

Here, $\hat{\tau}^2$ is the estimated variance of the distribution of the $\widehat{\theta'}_j$, and s2 represents the true

https://doi.org/10.26219/heal.aueb.6502

between-study variation in treatment effects. The parameter τ^2 is typically estimated using methods such as moments or the DerSimonian and Laird approach (Higgins, 2008), although REML approaches are recognized to have better performance properties (Veroniki, 2016). The DerSimonian and Laird approach is a widely used method in meta-analysis for estimating the between-study variability of the treatment effect. This approach assumes a random-effects model, which accounts for both within-study variability and between-study variability. It considers the observed variability across studies and incorporates a weight for each study that reflects its precision. The method estimates the overall treatment effect by combining the study-specific effect estimates, with greater emphasis on studies with smaller standard errors. This way, it provides a more accurate and robust estimate of the true between-study variability, taking into account both the within-study and between-study sources of heterogeneity.

In large samples (i.e., large K), the Q-statistic follows a chi-square distribution with K-1 degrees of freedom. When the variance of the treatment effect for study j does not vary across studies (i.e., σ^2j = σ^2), the measure of between-study heterogeneity, denoted as I^2, provides an intuitive understanding.

$$I^2 = 100\% * \frac{\tau^2}{\tau^2 + \sigma^2} \quad \text{(v)}$$

To clarify, I-squared is calculated as the quotient of the estimated between-study variability of the treatment effect (τ^2) divided by the sum of the estimated between-study variability (τ^2) and the estimated within-study variability (σ^2). The I-squared metric and τ^2 are closely connected: As τ^2 increases, the I-squared value also increases. (Huedo-Medina et al., 2006). Higgins and Thompson (2002)introduced a preliminary categorization of I-squared values to aid in understanding its scale. For instance, approximately 25% (I^2=25), 50% (I^2=50), and 75% (I^2=75) percentages would indicate low, medium, and high heterogeneity levels, respectively.

2.1 Incorporating treatment effect heterogeneity within a one-stage meta-analysis of individual patient data

https://doi.org/10.26219/heal.aueb.6502

To model treatment effect heterogeneity the following model has been proposed (Higgins, 2001) :

$$y_{ijs} = \mu + x_{ijs}\theta + a(C)_j + x_{ijs}a(CT)_j + e_{ijs} \tag{1}$$
$$e_{ij} \sim Norm(0, \sigma_e^2)$$

where:

$$\begin{pmatrix} a(C)_j \\ a(CT)_j \end{pmatrix} \sim \text{Multivariate Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_c^2 & \sigma_{ct} \\ \sigma_{ct} & \tau_{ct}^2 \end{pmatrix}\right)$$

and where y_{ijs} is the outcome for patient $\mathbf{i} = \langle i_1, i_2, \ldots, i_{mj} \rangle$, in trial $\mathbf{j} = \langle j_1, j_2, \ldots, {}_{mj} \rangle$, and in arm $\mathbf{s} = \langle 0{,}1 \rangle$; μ denotes the mean in the control arm (across every tial); y_{ijs} is the patient-level treatment indicator for patient i in tials j and treatment arm s (1: treatment; 0: control); θ is the treatment effect; $a(C)_j$ a random intercept for trial j, with variance σ_{ct}; and $a(CT)_j$ is a random interaction between trial and treatment (with variance τ_{ct}^2; and e_{ijs} a residual error term with variance σ_e^2). To maintain consistency in terminology for later models where a random cluster effect is denoted by $a(C)_j$, we utilize the notation $a(C)_j$ to represent the random study effect, although $a(S)_j$ might have been more intuitive. The presence of the non-zero covariance term σ_{ct} is crucial as it allows flexibility in determining whether there is more variation in the control or treatment arm . It is important to note that τ_{ct}^2 describes the variation between studies in the absence of treatment. In the context of an individual patient data meta-analysis, this is commonly substituted with a fixed effect and referred to as stratifying by study. In Equation (1), $a(C)_j$ can be reformatted as fixed study effects. By employing fixed effects for studies, we establish a parallel between the one-stage approach and the commonly used two-stage approach, a point that will be revisited in the discussion. Furthermore, τ_{ct}^2 represents the variation between studies in their response to treatment, analogous to τ^2 in a meta-analysis. It is also worth noting that K represents the total number of studies, and the total sample size for each study is $M_j = 2m_j$, assuming equal sample sizes in the treatment and control arms within each study, which is often a reasonable assumption. Based on this model, Hemming et al, propose the utilization of the I-squared statistic in the one-stage approach. As a reminder, the I-squared statistic quantifies the between-study variability of the treatment effect relative to the sum of the between-study variability and the average within-study variability. Consequently, the suggested measure of treatment heterogeneity is:

$$I^2 = 100\% * \frac{\tau_{ct}^2}{\tau_{ct}^2 + \frac{2\sigma_e^2}{\overline{m}}} \quad (2)$$

The quantity $\overline{m}$ represents the harmonic mean of the sizes of the trial arms. Similarly to the traditional I-squared, the proposed I-squared is the ratio of the estimated between-trial variability of the treatment effect τ_{ct}^2 to the sum of the between-study variability τ_{ct}^2 and the average within-study variability, estimated by Hemming et al. (2021), as $\frac{2\sigma_e^2}{\overline{m}}$. It is worth mentioning that even if fixed study effects had been employed in equation (1), this formula for I-squared would remain valid.

Similar to the conventional I-squared statistic, the proposed modified I-squared metric also adopts a range of values spanning from 0 to 1. In analogy to the traditional I-squared, a provisional interpretive framework can be outlined as mentioned in section 1.

2.2 Incorporating treatment effect heterogeneity within a one-stage meta-analysis of cluster randomized trials

In the context of a two-arm parallel cluster randomized trial, treatment allocation occurs at the cluster level. To ensure comprehensive coverage, Hemming et al. progressively developed an analysis models from the conventional approach to a model incorporating cluster-by-treatment interactions. The basic analysis model for this straightforward configuration is as follows:

$$y_{ij} = \mu + x_j\theta + a(C)_j + e_{ij} \quad (3)$$

where:

$$a(C)_j \sim Norm(0, \tau_c^2)$$

$$e_{ij} \sim Norm(0, \sigma_e^2)$$

In this scenario, the outcome for individual i (i = 1 ... M_j) within cluster j (j = 1 ... K) is denoted as y_{ij}. The mean value in the control arm, aggregated across all clusters, is

 https://doi.org/10.26219/heal.aueb.6502

represented by μ. The cluster-level treatment indicator x_j is assigned a value of 1 for treatment and 0 for control. The treatment effect is denoted as h. The random intercept $a(C)_j$ accounts for the variation between clusters, with a variance of τ_c^2. The residual error term e_{ij} has a variance of σ_e^2. It is important to note that τ_c^2 differs in interpretation from τ^2 in meta-analysis, as τ_c^2 describes the variation between clusters in the absence of treatment. Additionally, K now represents the total number of clusters, maintaining a one-to-one correspondence with the number of studies. The cluster size is indicated by Mj, which previously represented study size in the one-stage analysis of individual patient data.

In stepped-wedge trials, cluster cross-over trials, and other trials where treatment is combined with clusters, the studies are carried out across multiple time periods (Hussey, 2007; Parienti, 2007). In stepped-wedge trials, exposure to the intervention is partly confounded with time, necessitating adjustment for time effects in analysis models. Cluster cross-over trials involve adjusting for time effects to eliminate any residual effects.

$$y_{js} = \mu + x_{js}\theta + \varphi_s + a(C)_j + x_{js}a(CT)_j + e_{ijs} \tag{4}$$

$$e_{ijs} \sim Norm(0, \sigma_e^2)$$

where:

$$\begin{pmatrix} a(C)_j \\ a(CT)_j \end{pmatrix} \sim \text{Multivariate Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_c^2 & \sigma_{ct} \\ \sigma_{ct} & \tau_{ct}^2 \end{pmatrix}\right)$$

In this setting, , s represents the time interval of measurement, and φ denotes a categorical fixed effect assigned to measurements conducted during time period s, with a value of 0 to ensure identifiability; $a(C)_j$represents the primary cluster effect (a random effect), while $a(CT)_j$ represents the interaction between cluster and treatment (also a random effect) with a non-zero covariance term σ_{ct}. Random cluster effects allow for variation across clusters under control conditions and an additional source of variation across clusters under treatment. It is worth noting that this analysis model is conceptually identical (except for period effects) to the model used in an individual patient data meta-analysis (equation (5)). Therefore, for trials where treatment is crossed with the cluster, the following I-squared measure is proposed by Hemming et al.:

 https://doi.org/10.26219/heal.aueb.6502

$$I^2 = 100\% * \frac{\tau_{ct}^2}{\tau_{ct}^2 + \frac{4\sigma_e^2}{S\overline{m}}} \tag{5}$$

The quantity $\overline{m}$ denotes the harmonic mean of the cluster-period sizes, resulting in S$\overline{m}$ being the average total cluster size. Consequently, I-squared can be understood as the ratio of the between-cluster variability of the treatment effect estimated by the model $\tau_{ct}^2$to the sum of the between-cluster variability τ_{ct}^2 and the average within-cluster variability (approximated as $\frac{4\sigma_e^2}{S\overline{m}}$ by Hemming et al.). It's worth mentioning that the within-cluster variance has been estimated based on the assumption that the cluster size under both the control and intervention conditions is approximately equal. This assumption is likely to hold true in cluster cross-over trials (Robinson, 1991). Much like the conventional I-squared metric, this modified I-squared index also encompasses a range of values between 0 and , wherein values nearing signify minimal heterogeneity, while those approaching 1 suggest increased heterogeneity levels.

2.3 Incorporating treatment effect heterogeneity within a one-stage meta-analysis of multi-centre randomized trials

Next, we examine a multi-centre individually randomized trial. When randomization is stratified by center, the precision of the treatment effect is enhanced by incorporating a random center effect (Kahan, 2013). Additionally, to account for treatment effect heterogeneity, Hemming et al. (2020) introduce a random center by treatment effect and propose the following model:

$$\begin{gathered} y_{ijs} = \mu + x_{ijs}\theta + a(C)_j + x_{ijs}a(CT)_j + e_{ijs} \\ e_{ij} \sim Norm(0, \sigma_e^2) \end{gathered} \tag{6}$$

where:

$$\begin{pmatrix} a(C)_j \\ a(CT)_j \end{pmatrix} \sim Multivariate\ Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_c^2 & \sigma_{ct} \\ \sigma_{ct} & \tau_{ct}^2 \end{pmatrix}\right)$$

 https://doi.org/10.26219/heal.aueb.6502

In this context, y_{ijs} denotes the outcome for individual i (i = 1 ... mj), in center j (j = 1 ... K), and in arm s (s = 0, 1); μ represents the mean in the control arm across all centers; x_{ijs} represents the individual-level treatment indicator for individual i in center j and arm s (1: treatment; 0: control); θ denotes the treatment effect; $a(C)_j$ is a random intercept for center j, with variance τ_c^2; and $a(CT)_j$ is a random interaction between center and treatment, with variance τ_{ct}^2. Additionally, e_{ijs} is a residual error term with variance σ_e^2.

It's important to note that s represents the arm and not the time period. Furthermore, τ_c^2 describes the variation between centers in the absence of treatment, while τ_{ct}^2 represents the variation between centers in their response to treatment (similar to τ^2 in a meta-analysis). K now represents the total number of centers, and the total sample size for each center is $M_j = 2m_j$ (assuming an equal number in the two treatment arms within any center).
As a result, the proposed measure of treatment heterogeneity is:

$$I^2 = 100\% * \frac{\tau_{ct}^2}{\tau_{ct}^2 + \frac{2\sigma_e^2}{\overline{m}}} \qquad (7)$$

In this context, $\overline{m}$ denotes the harmonic mean of the sizes of the study arms. It is important to note that even if fixed study effects were utilized instead of random cluster effects, the formula for I-squared would remain valid. As previously mentioned, the values of this newly proposed I-squared metric span from 0 to 1, and their interpretation parallels that of the conventional I-squared metric.

https://doi.org/10.26219/heal.aueb.6502

3. Simulation Study

We will now examine how the proposed I-squared statistic from a one-stage approach compares to the conventional I-squared statistic based on a two-stage approach for meta-analysis. The aim is to ensure that the proposed I-squared is intuitive, strongly correlates with the conventional I-squared, and maintains these properties across various scenarios. To achieve this, we conduct comparisons across multiple sets of simulated individual patient data meta-analyses, considering different scenarios such as small vs. large studies and low vs. high heterogeneity. We will assess the correlation, agreement, and potential bias between the two metrics.

3.1 Process of data generation and simulation scenarios

The data are simulated to represent individual patient level data from multiple two-arm individually randomized trials, resembling the setting of an individual-level data meta-analysis. A linear mixed model (equation (1)) is used to generate the data, incorporating a random study effect and a random study by treatment interaction (Table 1). We consider scenarios with two-arm individually randomized trials, varying the number of trials available in each dataset (10, 50, or 100 trials) and the number of observations per study-arm (10, 50, or 100). The overall treatment effect is assumed to be absent, and the total variance is fixed to 1. Additionally, we explore different variance components (τ_c^2, τ_{ct}^2, and σ_e^2) as detailed in Table 1. The given parameters establish the "true" I-squared value for each scenario, classifying them into low (5% to 20%), medium (60% to 75%), or high (80% to 97%) treatment effect heterogeneity categories (Table 1). It's important to note that actual I-squared values are distinct not only for each heterogeneity scenario but also for every combination of study count and study arm size. Due to this extensive variability, it was not feasible to include the true I-squared values within the confines of this table. Instead, these genuine I-squared values have been presented in Figures 4 to 6 for clarity and comprehensive understanding.

Through numerous trial scenarios, we assess model performance under expected un-problematic situations and more realistic settings where small sample issues might affect performance (McNeish, 2017). For each simulated dataset, we estimate the one-stage and two-stage I-squared statistics derived from their respective analyses. The two-stage approach involves estimating study-specific treatment effects and standard errors, followed by pooling across studies using a random-effects meta-analysis. For the one-stage approach,

 https://doi.org/10.26219/heal.aueb.6502

a linear mixed model is fitted with a random study effect and a random study by treatment interaction (equation 1) using REML estimation methods (unstructured covariance). We then use equation (2) to estimate I-squared.

Performance evaluation includes documenting the correlation and agreement between the two I-squared statistics based on the number of clusters and cluster sizes across nine scenarios determined by study size and the number of studies. We also monitor the number of models that fail to converge and assess absolute bias by comparing the estimated I-squared (for both the one-stage and two-stage approaches) with the true I-squared for each particular scenario. We conduct 1,000 simulations per scenario.

The replication of the simulation study is feasible for both Models 4 and 6. In the case of Model 4, it necessitates the inclusion of an additional fixed effect pertaining to time. The detailed R code to execute this simulation is accessible in Appendix I, providing a comprehensive guide for conducting the replication process. This extension of the study to different models further enhances the comprehensive exploration of the phenomenon under investigation, ensuring the robustness and applicability of the study's findings across varying modeling scenarios.

3.2 Simulation study results

Out of the 27,000 simulated data-sets models failed to converge on 22 (0.081%) occasions. Simulated models failing to converge were excluded for the analysis. Figure 1 and Figure S1 illustrate a strong correlation and agreement between the proposed I-squared and the two-stage I-squared, especially when the number of studies exceeds 50 and there are at least 50 observations per study arm.

Table 1. Summary of scenarios explored in the simulation study (the I-squared statistic was defined by the combination of these parameters).

Study design parameters (considered in factorial combinations)			
Number of studies K	Study size per arm M	Number of arms S	Treatment effect θ
10	10	2	0
50	50	2	0
100	100	2	0

https://doi.org/10.26219/heal.aueb.6502

Variance parameters (considered in combinations as listed)

Study by treatment τ_{ct}^2	Study τ_c^2	Residual σ_e^2	Total Σ
0.2500	0.125	0.6250	1
0.1250	0.125	0.7500	1
0.00125	0.125	0.8725	1

Table 2. Number of models failing to converge per scenario.

N. models no-converging models	High heterogeneity	Moderate heterogeneity	Low heterogeneity
10 per arm, 10 studies	0	7	15
10 per arm, 50 studies	0	0	0
10 per arm, 100 studies	0	0	0
50 per arm, 10 studies	0	0	0
50 per arm, 50 studies	0	0	0
50 per arm, 100 studies	0	0	0
100 per arm, 10 studies	0	0	0
100 per arm, 50 studies	0	0	0
100 per arm, 100 studies	0	0	0

3.2.1 Comparing I-Squared Values: An Examination of Correlation Between One-Step and Two-Step Approaches in Simulation Study Results

Figures 1,2 and 3 illustrate a strong association and consistency between the proposed I-squared and the two-stage I-squared across all three scenarios, especially evident when there are over 50 studies and 50 observations per study arm. However, with only 10 studies or 10 observations per arm, the correlation between the two metrics decreases, particularly in cases of low heterogeneity (Figure 3). In the smallest sample size scenario (10 studies, 10 observations per arm), although the correlation between the two metrics is lower, there is still a moderate correlation across all heterogeneity scenarios, even in the low hetergoreneity scenario (≥0.84).

While agreement between the two metrics was sometimes moderate, it was more noticeable when there were few studies in the dataset or many small studies with low heterogeneity. However, agreement improved significantly with high heterogeneity, regardless of the number or size of the studies. The implications of this divergence are further explored below, by examining the extent of bias in the two metrics.

https://doi.org/10.26219/heal.aueb.6502

Figure 1. Correlation between I-squared one-stage and I-squared two-stage in high heterogeneity scenario: correlation scatterplots and coefficients ($\tau_{ct}^2 = 0.25,\ \sigma_e^2 = 0.65, \theta = 0$).

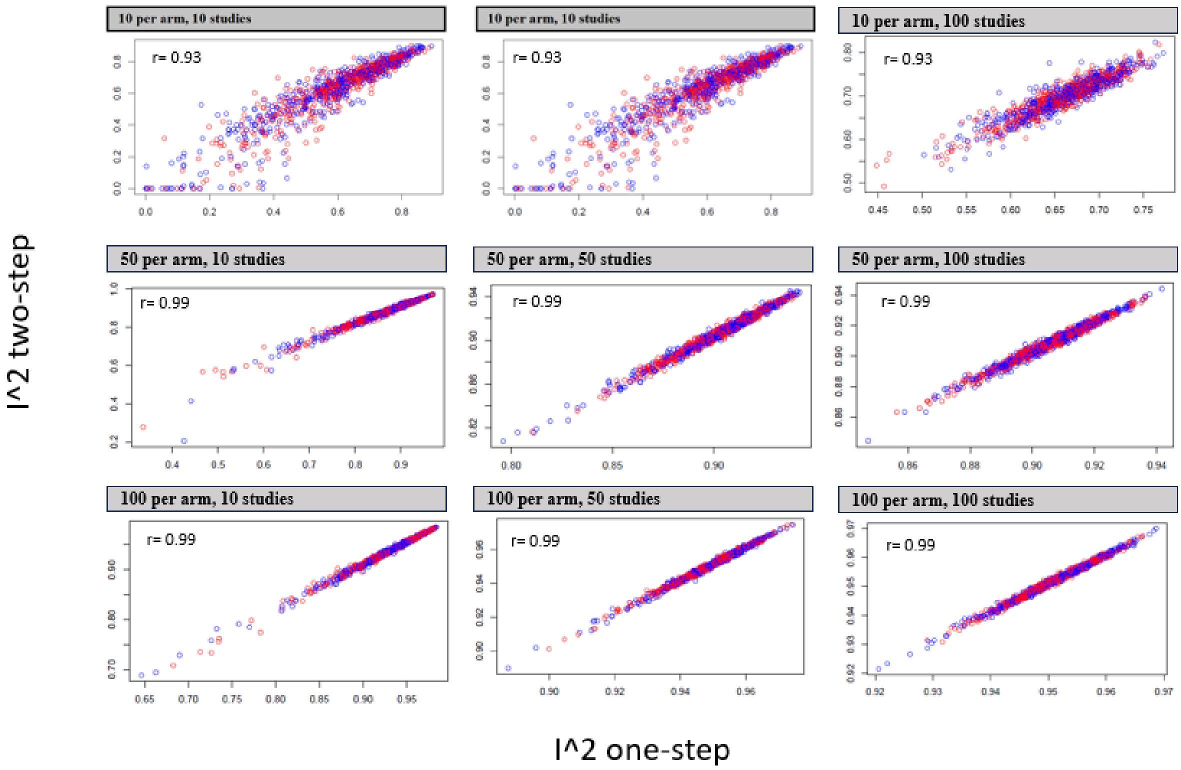


Figure 2. Correlation between I-squared one-stage and I-squared two-stage in moderate heterogeneity scenario: correlation scatterplots and coefficients ($\tau_{ct}^2 = 0.125,\ \sigma_e^2 = 0.75, \theta = 0$).

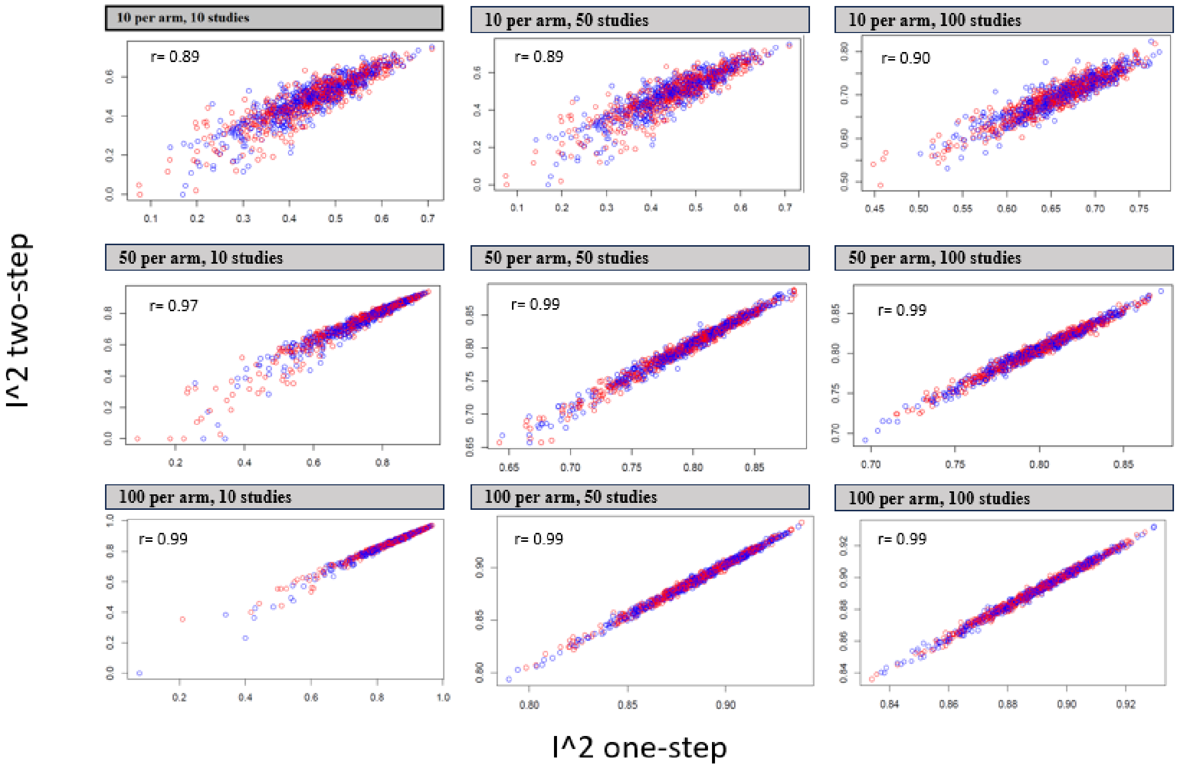
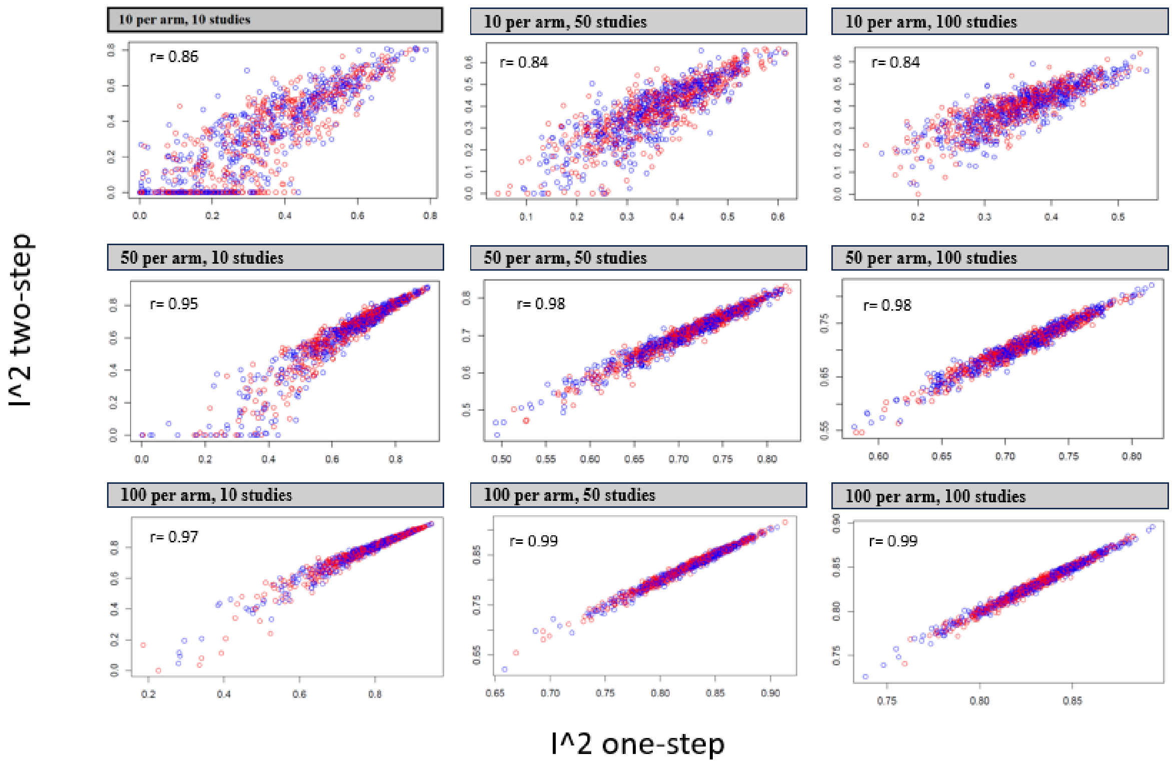


Figure 3. Correlation between I-squared one-stage and I-squared two-stage in low heterogeneity scenario: correlation scatterplots and coefficients ($\tau_{ct}^2 = 0.00125,\ \sigma_e^2 = 0.8725, \theta = 0$).



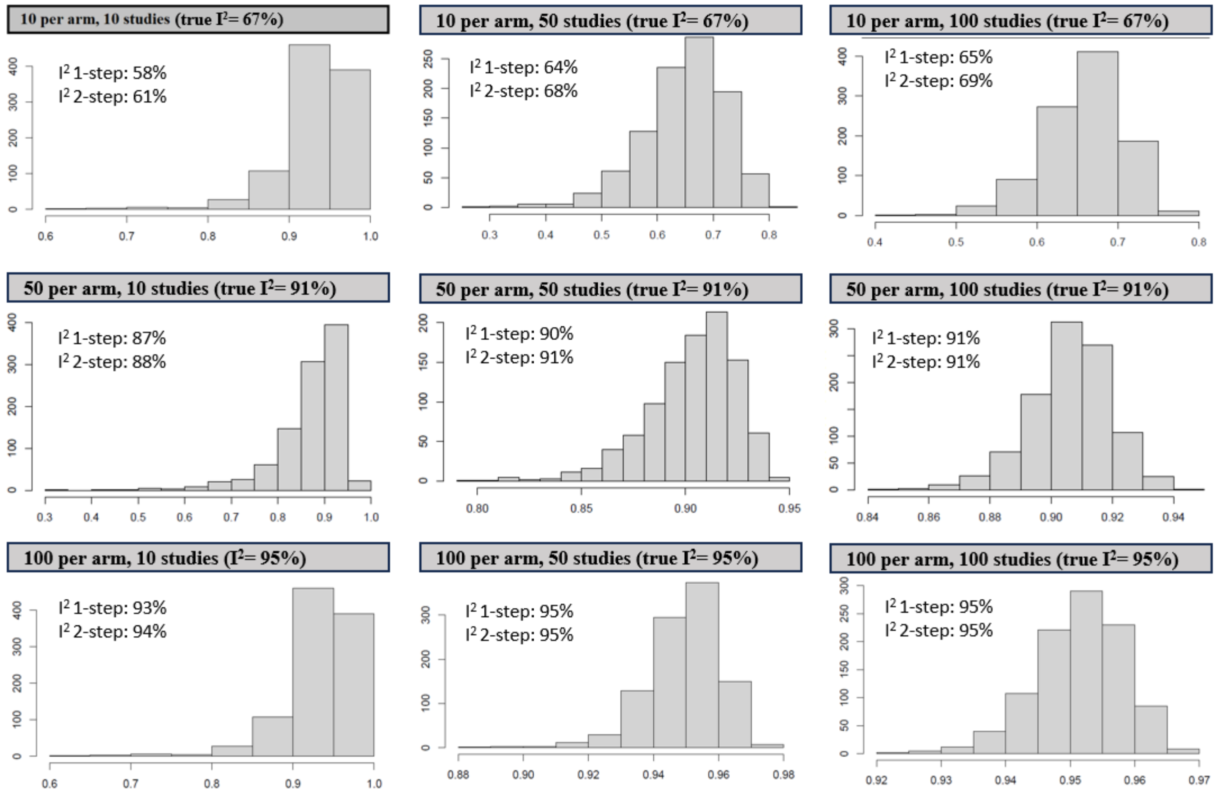
https://doi.org/10.26219/heal.aueb.6502

3.2.2 Evaluating Bias: A Comparative Analysis of I-Squared Values from One-Step and Two-Step Approaches Against True I-Squared Values in Simulation Study Results

Figures 4, 5, and 6 demonstrate the mean I-squared values derived from both the 1-step and 2-step approaches, along with the corresponding true I-squared values specific to each distinct scenario. We estimate bias by comparing the estimated I-squared (for both the one-stage and two-stage approaches) with the true I-squared for each particular scenario. In the majority of scenarios, there was no indication of bias in either of the I-squared metrics to a significant extent (Figures and 5). However, both measures exhibited some degree of bias, especially in low heterogeneity scenarios (Figure 6). For instance, when there were 100 studies and 50 observations per study arm, and low I-squared, there was an observed bias of up to to 56% on a relative scale in both the one-stage and two-stage approaches (Figure 6). While these biases may seem substantial at first glance, their contextual significance diminishes when considering how these metrics are interpreted. Bias tended to be more pronounced when there were only 10 observations per arm compared to 50 or 100 observations per arm, and it was significant at lower levels of heterogeneity than at higher levels. Both measures also tended to overestimate heterogeneity in scenarios with large sample sizes in the low heterogeneity scenario. Higher bias is expected in those scenarios since, based on our previous literature, it has been shown that a high I-squared can either clinically important treatment effect heterogeneity or large sample sizes. Neither of the two I-squared measures consistently exhibited less bias than the other. Moreover, in all scenarios except the unrealistic one (a meta-analysis of 10 studies of size 10), the correlation between the two metrics remained above 0.8.

Figure 4. Histogram illustrating the distribution of I-squared values obtained through the 1-step approach, alongside the means of I-squared values derived from both the 1-step and 2-step approaches and true I-squared values, high heterogeneity scenario ($\tau^2_{ct} = 0.25,\ \sigma^2_e = 0.65, \theta = 0$).

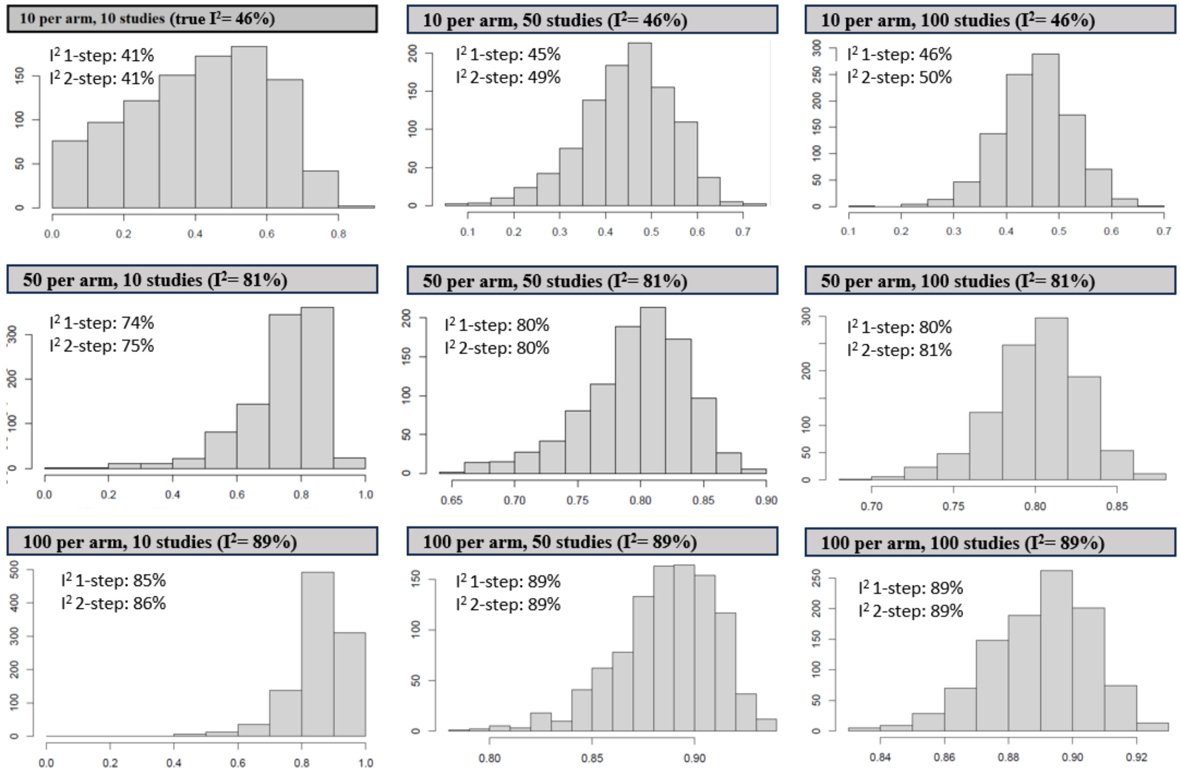
 https://doi.org/10.26219/heal.aueb.6502



*Bias can be calculated by subtracting the true I-squared from the estimated I-squared (for both the one-stage and two-stage approaches) with for each particular scenario.

Figure 5. Histogram illustrating the distribution of I-squared values obtained through the 1-step approach, alongside the means of I-squared values derived from both the 1-step and 2-step approaches and true I-squared values, moderate heterogeneity scenario $\tau_{ct}^2 = 0.125$, $\sigma_e^2 = 0.75, \theta = 0$).

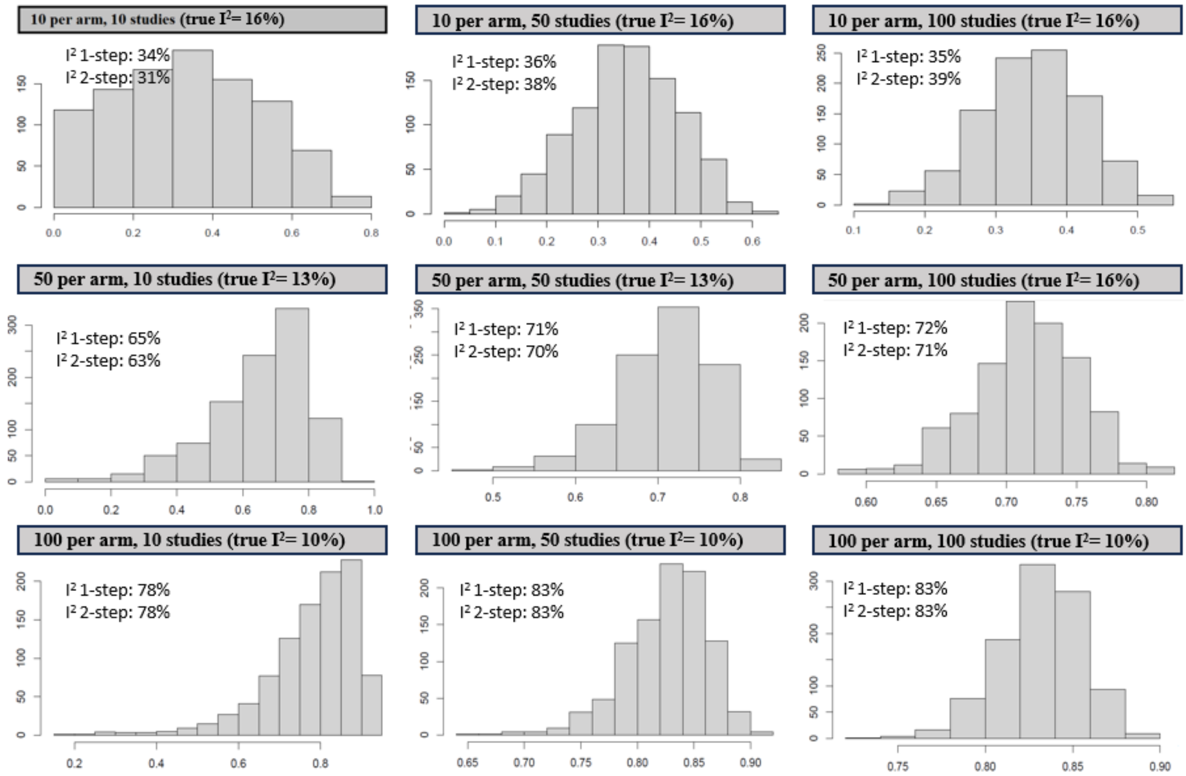
https://doi.org/10.26219/heal.aueb.6502



*Bias can be calculated by subtracting the true I-squared from the estimated I-squared (for both the one-stage and two-stage approaches) with for each particular scenario .

Figure 6. Histogram illustrating the distribution of I-squared values obtained through the 1-step approach, alongside the means of I-squared values derived from both the 1-step and 2-step approaches and true I-squared values, low heterogeneity scenario ($\tau_{ct}^2 = 0.00125$, $\sigma_e^2 = 0.8725, \theta = 0$).

https://doi.org/10.26219/heal.aueb.6502



*Bias can be calculated by subtracting the true I-squared from the estimated I-squared (for both the one-stage and two-stage approaches) with for each particular scenario.

In addition to examining the performance of the proposed methodology for I-squared, a similar simulation study has been conducted to assess treatment effects. The comparison between treatment effects estimated using both the 1-step and 2-step approaches revealed a substantial agreement, as evidenced by a significant correlation coefficient exceeding 0.98 across all three simulated scenarios. For more detailed results and analyses pertaining to the treatment effects simulation study, interested readers are referred to the Appendix section of this thesis.

The simulation study was replicated under varying treatment effects, specifically for θ value of 1 and 3. Remarkably, these altered conditions yielded outcomes that strongly overlapped with the original results. This finding underscores the robustness of the simulation study, indicating that diverse values of theta do not exert any discernible impact on the study's overall outcomes. The findings from the simulation study conducted with a theta value of 1 are detailed in Appendix 3. As the results of the simulation study with theta = 3 align perfectly with those of the other simulation studies, they will not be presented in detail.

 https://doi.org/10.26219/heal.aueb.6502

4. Example: Meta-analysis of simulated individual patient data

I will now demonstrate these concepts through an examples. The example is a (simulated) individual patient data meta-analysis, where the goal is to estimate the proposed I-squared using a one-stage approach and compare it with the conventional two-stage approach.

In this example, we use simulated individual-level data from 10 parallel two-arm individually randomized trials. The data can be accessed for download and use as an illustrative dataset (https://github.com/karlahemming/One-stage-I-squared). The simulated data features a continuous normally distributed outcome, with an average treatment effect of 2 and a residual variance of 1 (σ_e^2) for each of the 10 studies, each having a sample size of 100 per arm. The between-study variance in the absence of treatment is 0.1 (τ_c^2), and the between-study variation in response to treatment is 0.2 (τ_{ct}^2), resulting in a 'true' I-squared value of 91%.

The data have been aggregated to estimate the treatment effect for each study. The data are then pooled across studies using a random effects meta-analysis and the between-study variance (τ_c^2), is estimated using the Dersimonian and Laird approach. This approach yields an estimated I-squared of 86%. The forest plot (Figure 7) displays substantial variability among studies in the estimated treatment effects (mean differences), contributing to the high degree of heterogeneity reflected by the large I-squared value.

Although the individual-level data are available in this simulated example, they have not been directly used in estimating I-squared through the two-stage approach. I now illustrate how we propose to estimate I-squared from a one-stage approach, utilizing the individual-level data directly and fitting the model at equation (1) and equation (2). Employing the individual-level data, we obtain the estimated pooled treatment effect (across studies) of 1.96 (95% CI 1.72 to 2.20), the estimated between-study variability in treatment effects of τ_{ct}^2 =0.12, and the residual variance of σ_e^2=1.02 ($\overline{m}$=100). This yields an estimated I-squared of 87%.

It is worth noting that the estimated value of the treatment effect heterogeneity random effect (0.12) is lower than the true value (0.2) under both approaches, likely due to the downward bias in variance component estimation with few groups. For illustration purposes, we also provide estimates of study-specific treatment effects from a one-stage fixed effects model, where the I-squared is estimated to be 87.10%. Appendix 2 includes the R code for implementing both methods, and the simulated dataset is available as a CSV file.

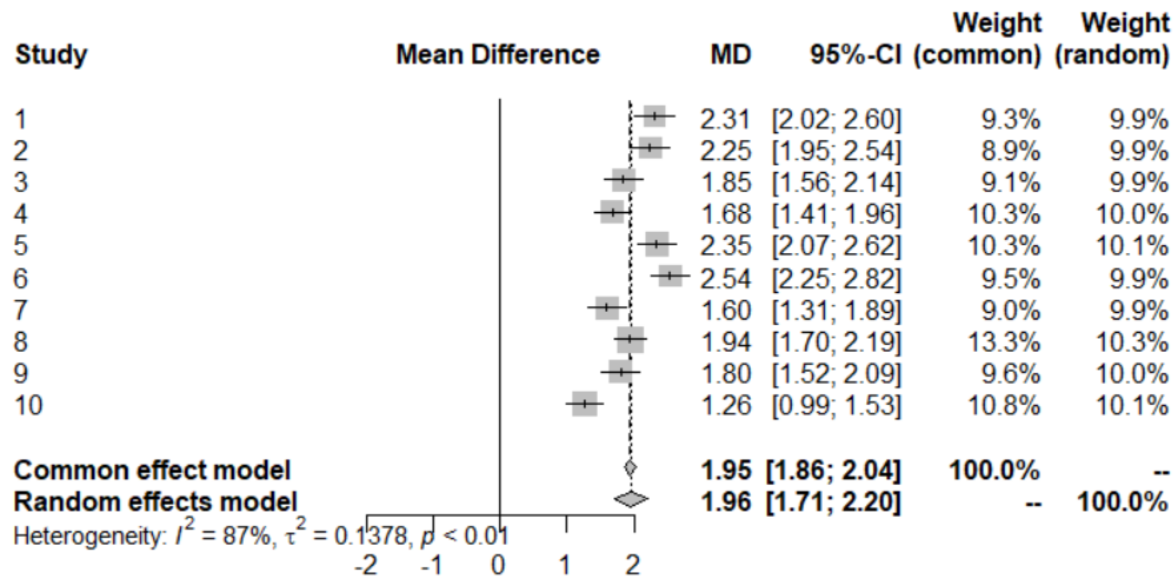
https://doi.org/10.26219/heal.aueb.6502

Table 3. Results of 1-step approach for meta-analysis of simulated individual patient data example

Fixed effects	Estimate	SE	DF	t-value	p-value
(Intercept)	1.27	0.10	1993	13.26	0
Treatment	1.96	0.12	1993	16.41	0
Random effects	**SD**	**Corr**			
(Intercept)	0.28	(Intr)			
Treatment	0.35	-0.06			
Residual	1.01				

Heterogeneity: $I^2 = 86\%$

Figure 7. Results of 2-step approach for meta-analysis of simulated individual patient data example



https://doi.org/10.26219/heal.aueb.6502

5. Discussion

Treatment effect heterogeneity is a common consideration in meta-analysis, but it is less explored in other settings. In meta-analysis, the I-squared statistic is commonly used to quantify treatment effect heterogeneity across studies. We have evaluated a similar concept to measure treatment effect heterogeneity in multi-centre and cluster randomized trials, as well as in individual patient data meta-analyses as proposed by Hemming et al..

High correlation between the two I-squared statistics is observed, especially when there is a high degree of heterogeneity, a large number of studies, and substantial study sizes. We also noticed significant bias in both I-squared metrics, particularly when the number of studies was small and the level of heterogeneity was low. Nevertheless, interpreting the metric broadly as low, medium, or high degrees of heterogeneity is still valid in most cases, as long as researchers remain mindful that the absolute value may deviate from the truth.

The observed bias is likely due to the difficulty in estimating small variance components in small samples, which is evident in both the mixed model estimation of the one-stage approach and potentially under the model estimation of the two-stage approach. Additionally, scenarios with a high sample sizes few large studies or many small studies showed greater bias, indicating challenges in estimating the within-study error term and between-study treatment variance random effect in such cases.

Individual patient data meta-analyses are not as common as aggregate data meta-analyses, partly due to the difficulty in obtaining individual-level data and the added complexity involved. However, the aim of the proposed heterogeneity measure is not only to develop a methodology for individual patient data meta-analyses but also to leverage the similarities between meta-analysis and the analysis of cluster and multi-centre randomized trials, promoting the investigation of treatment effect heterogeneity in these settings. This will become increasingly important as pragmatic randomized trials become more prevalent

https://doi.org/10.26219/heal.aueb.6502

and interventions are evaluated in real-world settings.

The I-squared statistic has appealing properties, mainly its intuitive interpretation. However, it also has some less appealing aspects, particularly in statistical testing, where a low I-squared can indicate no treatment heterogeneity or insufficient evidence to make definitive statements, a high I-squared can indicate either clinically important treatment effect heterogeneity or large sample sizes (Rücker et al., 2008; von Hippel et al., 2015). This is also evidence in our results which indicated that both the one stage and the twos stage I squared methods overestimated heterogeneity in scenarios with high sample sizes.

Quantifying the level of uncertainty in the estimation of I-squared might be helpful, but this feature is primarily related to the construct measured by I-squared, which is sample size-dependent. An alternative method, using τ_{ct}^2, can also create an intuitive measure when used in conjunction with a predictive interval. Furthermore, estimates of treatment effect heterogeneity from mixed effects models are valuable in designing future studies (Hughes et al., 2015).

The proposed estimate for I-squared is based on a mixed effects model with study modeled as a random effect (Morris, 2018). While random effect models are conventional in cluster trials and multi-centre trials, the choice between fixed and random effects depends on the analysis objectives and data characteristics.

Small sample corrections are important for model estimation, especially in individual-level data meta-analyses with a small number of studies, as they might be at risk of inflated type 1 error rates (Kahan, 2013). Therefore, a fixed study effect might be more appropriate in such cases. Nonetheless, Hemming et al. have used random effects in their proposed methodology, considering its conventional usage in cluster trials and multi-centre trials.

Modeling treatment effect heterogeneity is crucial to avoid overprecise confidence intervals and inflated type 1 errors (Thomson et al, 2017). However, large sample theory

https://doi.org/10.26219/heal.aueb.6502

assumptions require a sufficient number of clusters or studies for accurate estimation of variance components using mixed models. Usually, when estimating fixed effects (for continuous outcomes), this implies a requirement for over 40 clusters, centers, or studies in total (Kahan et al., 2016). However, trials and meta-analyses often involve far fewer than 40 clusters, centers, or studies. Moreover, these studies also necessitate accurate estimation of variance components, in addition to fixed effects. Small sample corrections are essential for model estimation, but their impact on the I-squared parameter might be limited (Kenward et al., 1997). Further research is necessary to validate this measure in various model settings, outcome types, and in the context of multiple period designs and time-dependent correlations.

https://doi.org/10.26219/heal.aueb.6502

6. Appendix I

6.1 Appendix I: Two-stage and one-stage analysis code

```
#simulation function, n_studyarm= 100, n_studies = 10, tau2ct= 0.2500, s2e= 0.6250

fiftyten_fun = function(n_studyarm= 100, n_studies = 10, tau2ct= 0.2500, s2e= 0.6250) {

  n_observations= n_studies*2*n_studyarm

  study = rep( c(1:n_studies), each = n_studyarm, times=2)

  treatment = rep( c("control", "treatment"), each=n_observations/2)

  dat <- data.frame(study, treatment)

  theta <- 0

  b0 <- 200 #randomly selected

  sigma <- matrix(c(0.125, 0.1, 0.1, tau2ct), ncol = 2)

  ACTarray <- rmvnorm(n_studies, mean = c(0,0), sigma = sigma, method="chol")

  ACTDF <- data.frame(ACTarray)

  (ACT  = c(rep(0,n_studies),c(ACTDF$X2)))

  (ACT  = rep(ACT, each = n_studyarm))

  (AC  = rep(ACTDF$X1, each = n_studyarm, times=2))

  e_lowhet1 <- rnorm(n_observations, 0,sqrt(s2e))

  (dat$y= with(dat, b0 + theta*(treatment == "treatment") + AC + ACT + e_lowhet1))

  ctrl <- lmeControl(opt='optim')

  tryCatch({

  fit <- lme(y~treatment, random=~1+treatment|study,method="ML",data=dat,  control
=ctrl)

  refit <- VarCorr(fit) #deriving random interaction sd

  var.act <- as.numeric(refit[2,1])
```

https://doi.org/10.26219/heal.aueb.6502

```
  studyarmsize <- (rep( c(n_studyarm), times = n_studies*2)) #calculating study-arm
harmonic mean
  N <- length(studyarmsize)
  invstudyarmsize <- studyarmsize^(-1)
  sumOfInverse  <- sum(invstudyarmsize)
  HM=N/sumOfInverse
  isq <- (var.act )/( (var.act) + ((2*(fit$sigma^2)/HM)) )
  }, error=function(e) rep(NA, 4L))
  tryCatch({
  tablez <- ddply(dat,~study*treatment,summarise,mean=mean(y),sd=sd(y))
  aggmeta2<- tablez %>%
    pivot_wider(names_from = treatment, values_from = c(mean, sd))
  aggmeta2 <- aggmeta2 %>%
    add_column(n.treatment = n_studyarm) %>%
    add_column(n.control = n_studyarm)
  res.flesiss =  metacont(n.treatment, mean.e = mean_treatment, sd.e = sd_treatment,
                          n.control, mean.c = mean_control, sd.c = sd_control,
                          comb.fixed = T, comb.random = T, studlab = study,
                          data = aggmeta2, sm = "MD", overall.hetstat =T)
  }, error=function(e) rep(NA, 4L))
  (list(isq, res.flesiss$I2))
}


set.seed(99)
fiftyten_fun()
```

https://doi.org/10.26219/heal.aueb.6502

```
isq<- NA
tryCatch({
  simzz <- try(replicate(1000, fiftyten_fun(), simplify = FALSE ))
},
error = function(err){
  message("Error in fiftyten_fun() : object 'isq' not found")
  message("Error in fiftyten_fun() : object 'res.flesiss' not found")
  return(NA)
})

fiftyten.isqz <- do.call(rbind.data.frame, simzz)
names(fiftyten.isqz)=c("oneStepISQ","twpStepISQ")
sum(is.na(fiftyten.isqz$oneStepISQ))
sum(is.na(fiftyten.isqz$twpStepISQ))
#output
plot(fiftyten.isqz$oneStepISQ, fiftyten.isqz$twpStepISQ, col=c('red', 'blue'), ylab="I^2 two-
step", xlab="I^2 one-step")

mean(fiftyten.isqz$oneStepISQ, na.rm=TRUE)
mean(fiftyten.isqz$twpStepISQ, na.rm=TRUE)
cor(fiftyten.isqz$oneStepISQ, fiftyten.isqz$twpStepISQ, method = c("pearson"), use =
"complete.obs")
hist(fiftyten.isqz$oneStepISQ)
hist(fiftyten.isqz$twpStepISQ)
```

https://doi.org/10.26219/heal.aueb.6502

6.2 Appendix II: Simulation Study results – treatment effect (θ=0)

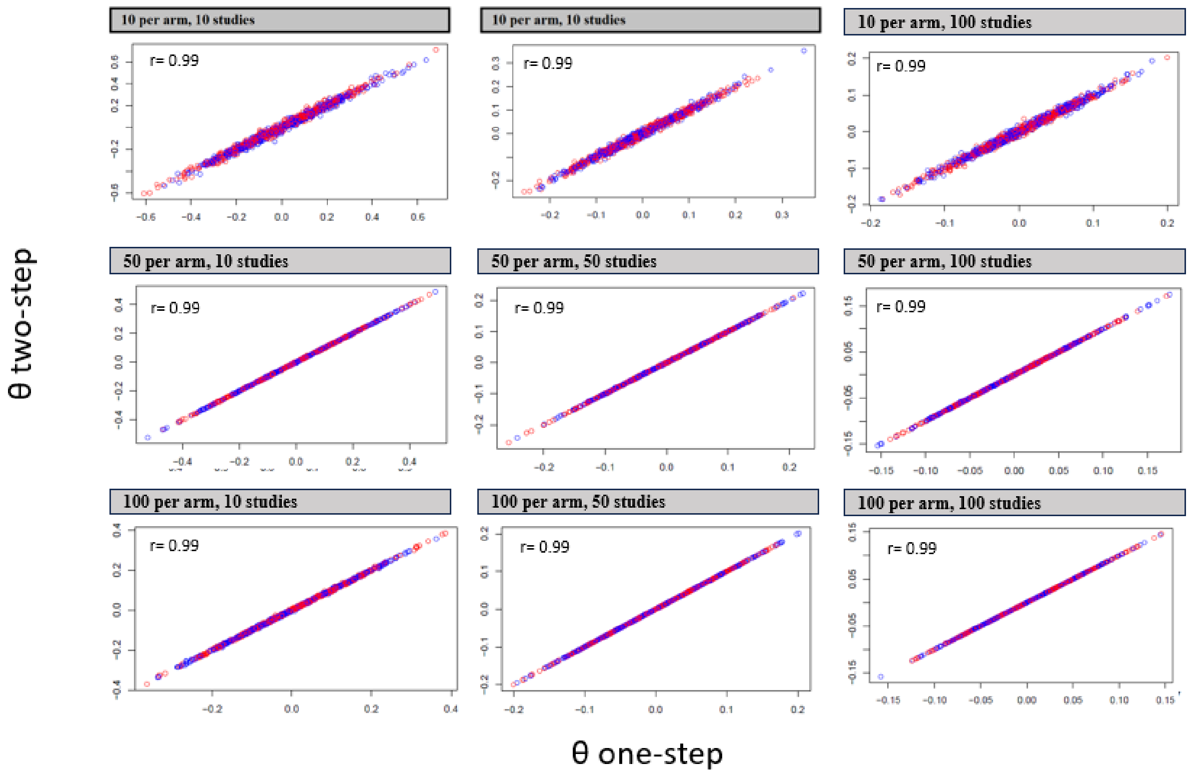
This section of the paper presents the outcomes of our simulation study conducted with a treatment effect value set at theta=0. In particular, our focus here lies on examining whether correlation between the treatment effects derived from the One-Step Approach and those obtained through the Two-Step Approach. Furthermore, we will also assess whether treatment effects obtained from both models exhibit a tight clustering around the predefined value of zero, which serves as a measure of robust convergence.

In the context of our simulation analysis focusing on the examination of correlation between treatment effects derived from the One-Step and Two-Step Approaches, an intriguing trend emerged when the treatment effect was defined as 0. The scatterplots vividly illustrate a remarkably high correlation between the two metrics across all scenarios, with a consistent and substantial correlation coefficient (r) exceeding 0.98 in every case. These results underscore a robust and consistent relationship between the treatment effects produced by the two approaches, further substantiating their comparability.

Furthermore, as can be observed on the density plots, the distributions of thetas from the two approaches overlap almost entirely for all scenarios demonstrating a strikingly high degree of correlation (Appendix II Figures 4,5,6). The distributions of treatment effects demonstrate a closely centered positioning around 0. Treatment effects closely clustered around the predefined value of zero serve as strong indicators of favorable convergence for both modeling approaches. This alignment echoes the robustness of the correlation observed in the scatterplots and reaffirms the reliability and concurrence of the treatment effect estimation derived from these two distinct approaches.

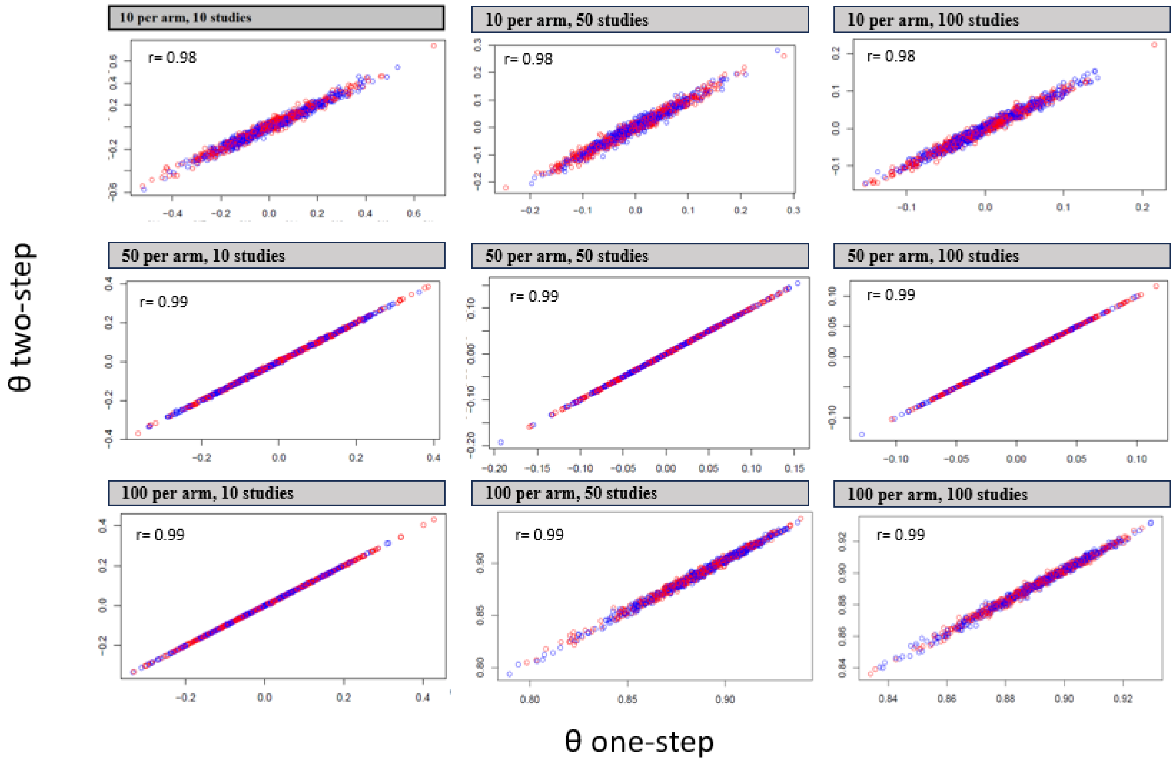
 https://doi.org/10.26219/heal.aueb.6502

Appendix Figure 1. Correlation between θ one-stage and θ two-stage in high heterogeneity scenario ($\tau_{ct}^2 = 0.25,\ \sigma_e^2 = 0.65,\ \theta = 0$).

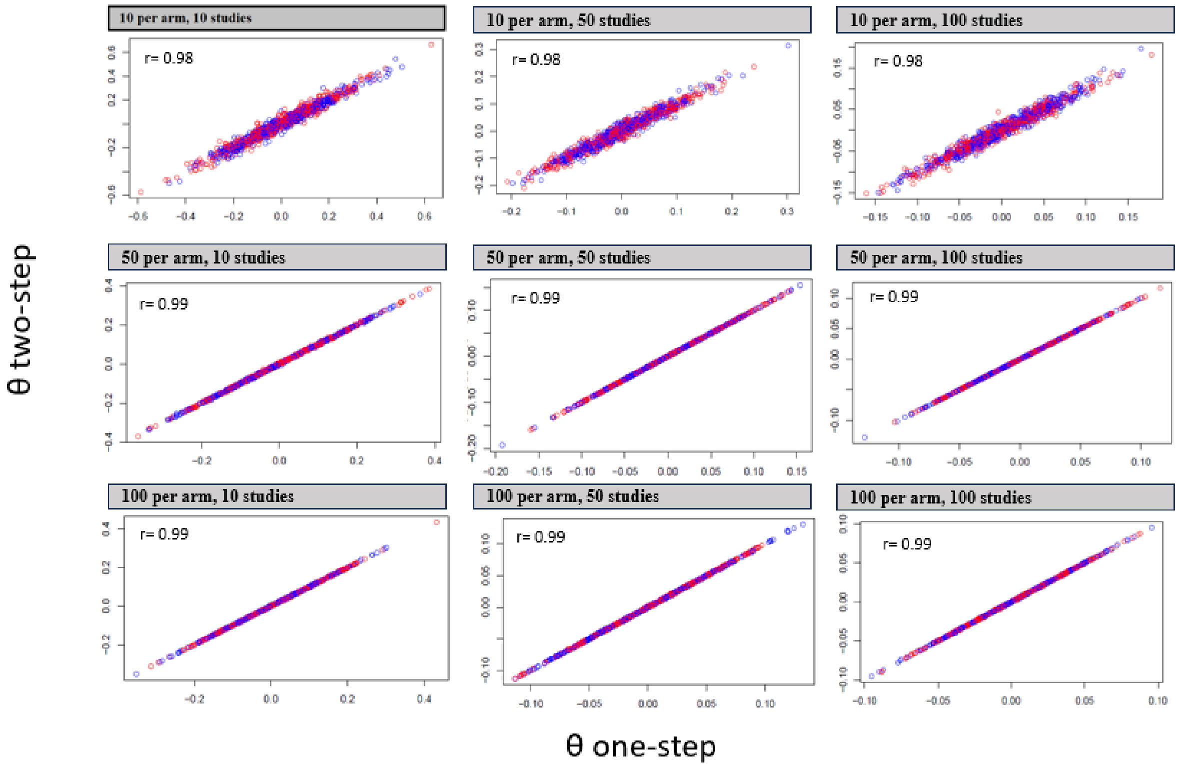


Appendix Figure 2. Correlation between θ one-stage and θ two-stage in moderate heterogeneity scenario ($\tau_{ct}^2 = 0.125,\ \sigma_e^2 = 0.75,\ \theta = 0$).

 https://doi.org/10.26219/heal.aueb.6502

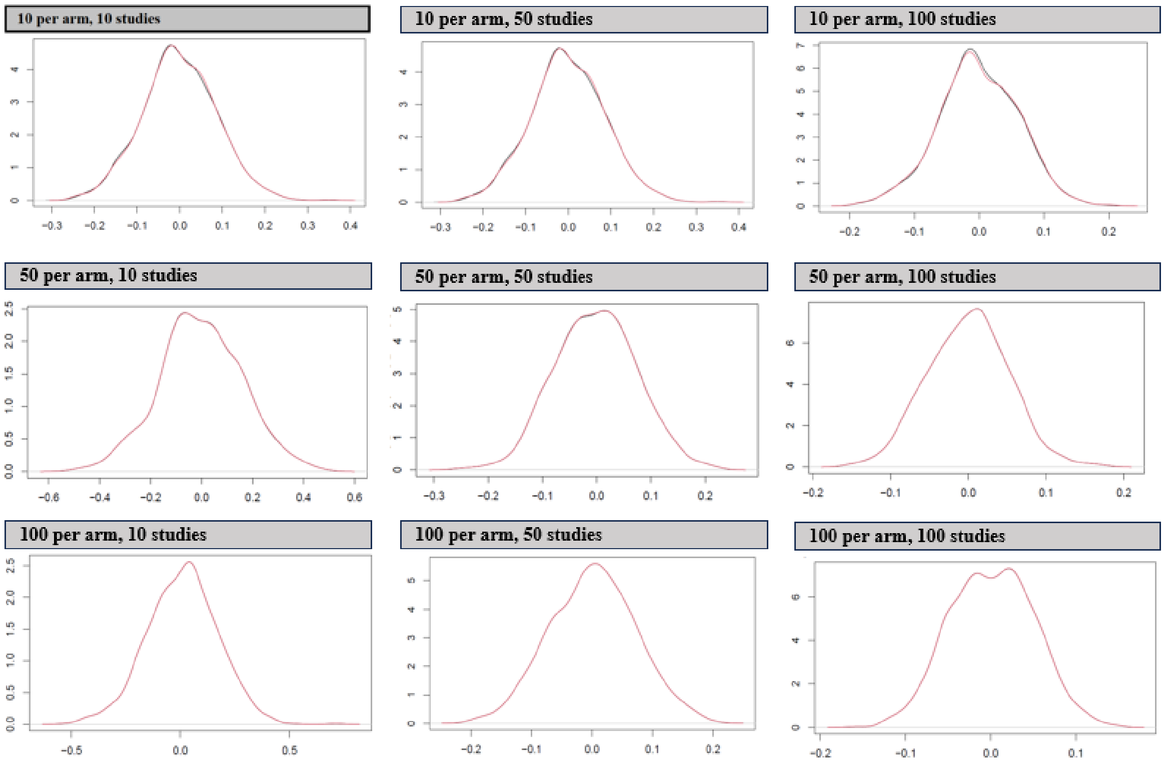


Appendix Figure 3. Correlation between I-squared one-stage and I-squared two-stage in low heterogeneity scenario: correlation scatterplots and coefficients ($\tau_{ct}^2 = 0.00125,\ \sigma_e^2 = 0.8725, \theta = 0$).

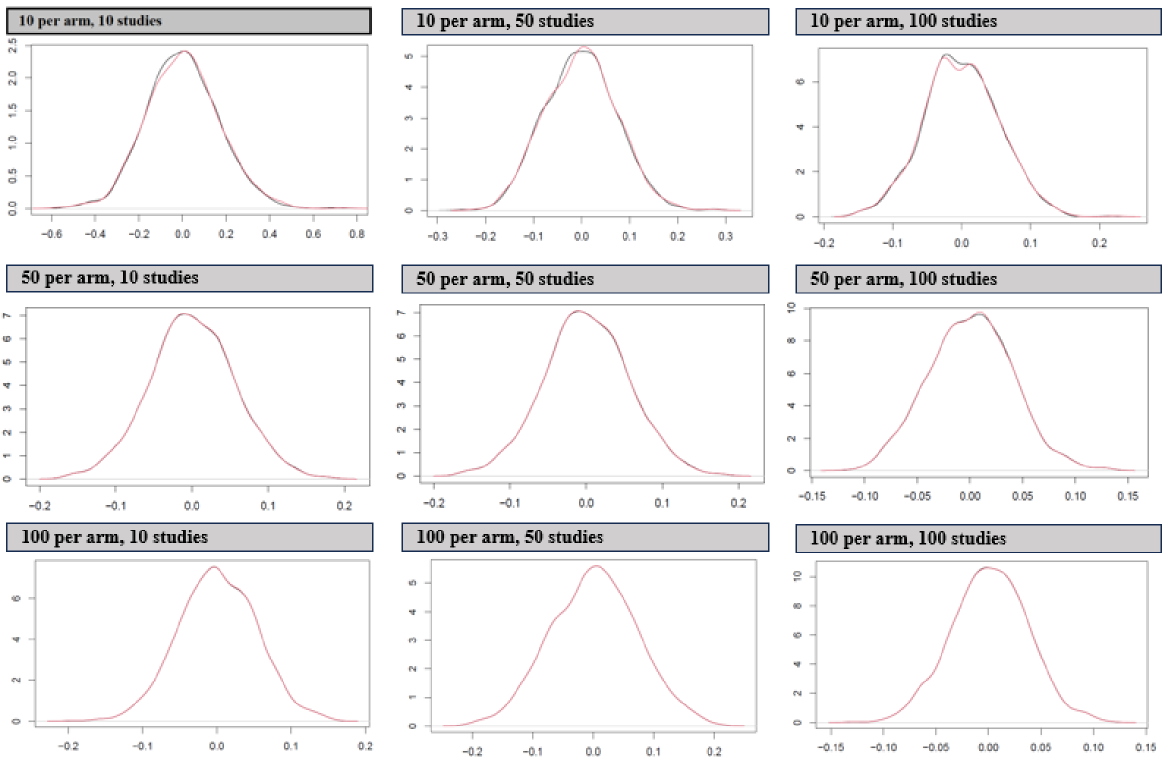


https://doi.org/10.26219/heal.aueb.6502

Appendix Figure 4. Density plots for 1-step approach θ (red) anδ two-step approach θ (black) ($\tau_{ct}^2 = 0.25,\ \sigma_e^2 = 0.65, \theta = 0$).

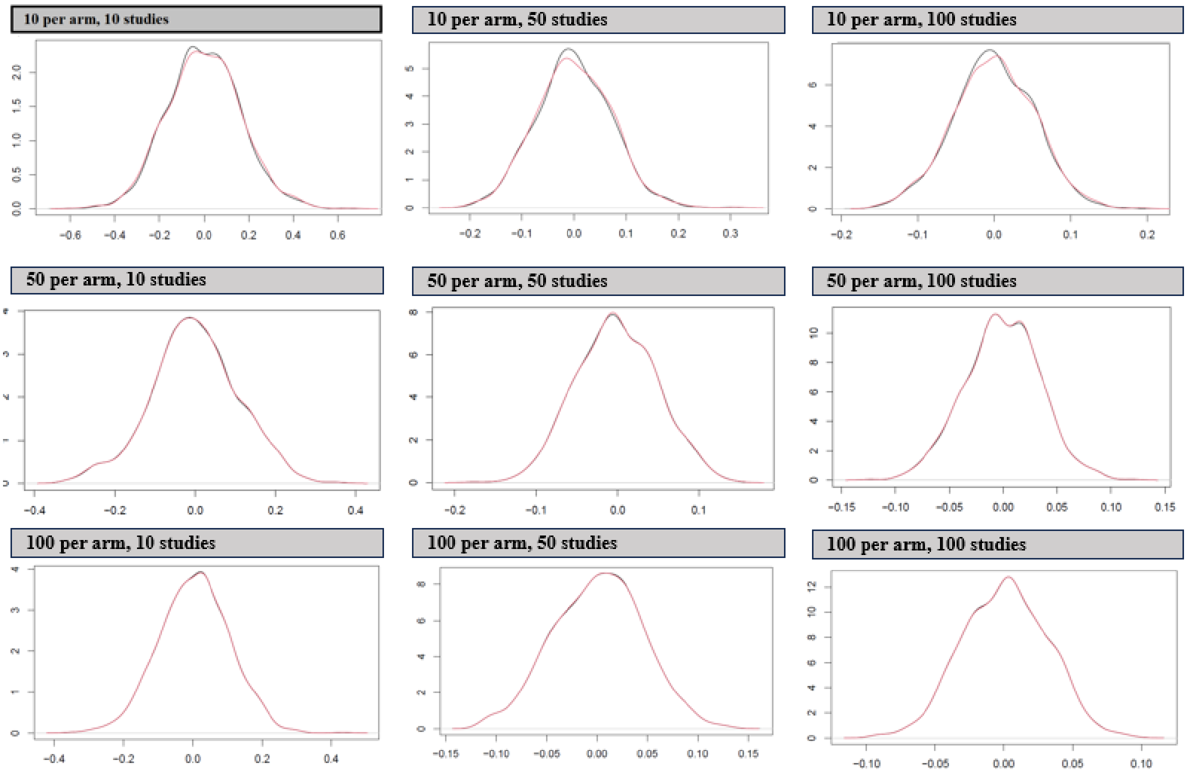


Appendix Figure 5. Density plots for 1-step approach θ (red) and two-step approach θ (black) ($\tau_{ct}^2 = 0.125,\ \sigma_e^2 = 0.75, \theta = 0$).



https://doi.org/10.26219/heal.aueb.6502

Appendix Figure 6. Density plots for 1-step approach θ (red) anδ two-step approach θ (black) ($\tau_{ct}^2 = 0.00125,\ \sigma_e^2 = 0.8725,\ \theta = 0$).



https://doi.org/10.26219/heal.aueb.6502

6.3 Appendix III: Simulation study results, theta= 1

This section aims to present the results of the simulation study under the condition of theta equal to 1 and investigate whether varying theta values have an impact on the simulation study outcomes. In this section, we will conduct a comparative analysis of I-Squared values to assess the correlation between One-Step and Two-Step approaches in Simulation Study Results. Additionally, we will evaluate the bias between the I-Squared values produced by the One-Step and Two-Step approaches when a different value of theta (in this case, 1) is used.

6.3.1 Comparing I-Squared Values: An Examination of Correlation Between One-Step and Two-Step Approaches in Simulation Study Results, theta=1

In the scenario with a treatment effect of theta=1, our analysis yielded results that align perfectly with those obtained in the theta=0 scenario across all examined scenarios. As expected, different values of theta did not appear to have any significant impact on the outcomes. In all cases, the correlation between I-Squared values derived from the One-Step and Two-Step Approaches remained consistently high, exceeding 0.93. These consistent results, regardless of varying treatment effect scenarios, provide strong evidence of the robustness and stability of our findings. They reinforce the notion that the choice of theta value does not substantially affect the comparative performance of the One-Step and Two-Step Approaches in estimating I-Squared values.

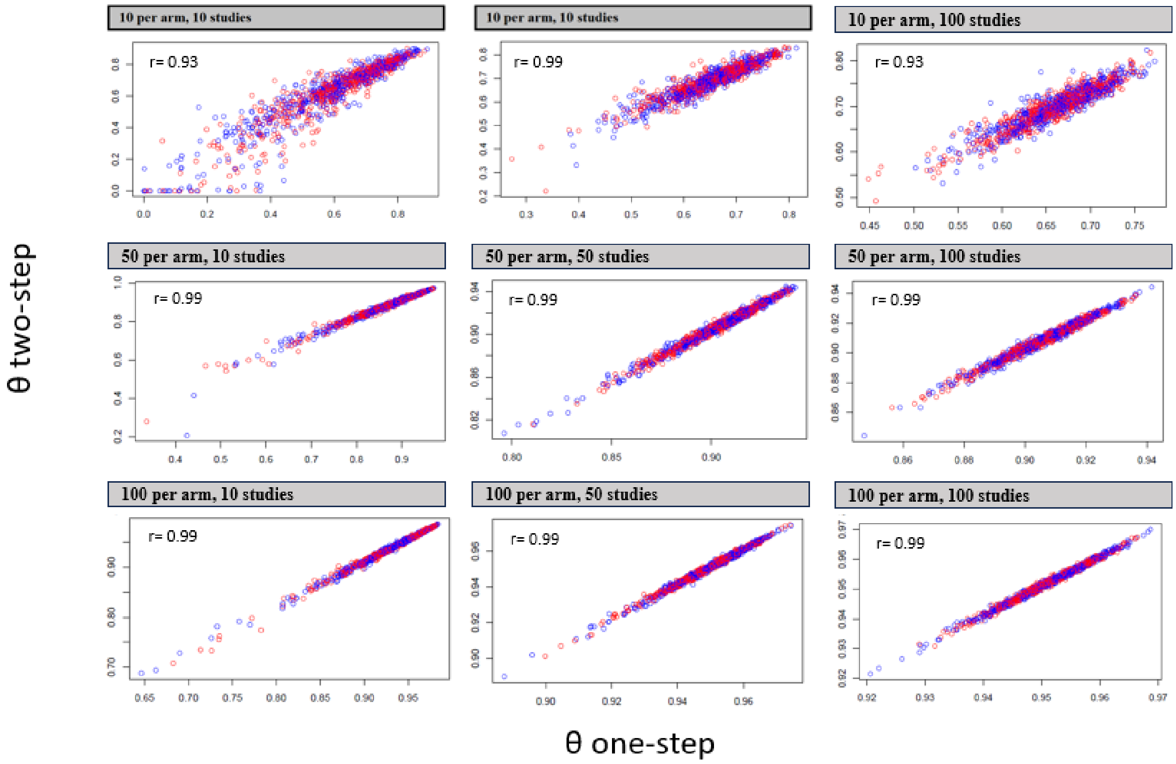
Out of the 27,000 simulated data-sets models failed to converge on 27 (0. 1%) occasions. The majority of instances where non-convergence was observed were within the high heterogeneity scenario. Simulated models failing to converge were excluded for the analysis.

Appendix Table 1. Number of models failing to converge per scenario.

N. models no-converging models	High heterogeneity	Moderate heterogeneity	Low heterogeneity
10 per arm, 10 studies	3	0	0
10 per arm, 50 studies	3	0	0
10 per arm, 100 studies	3	0	0
50 per arm, 10 studies	3	0	0
50 per arm, 50 studies	3	0	0
50 per arm, 100 studies	3	0	0
100 per arm, 10 studies	3	0	0
100 per arm, 50 studies	3	0	0
100 per arm, 100 studies	3	0	0

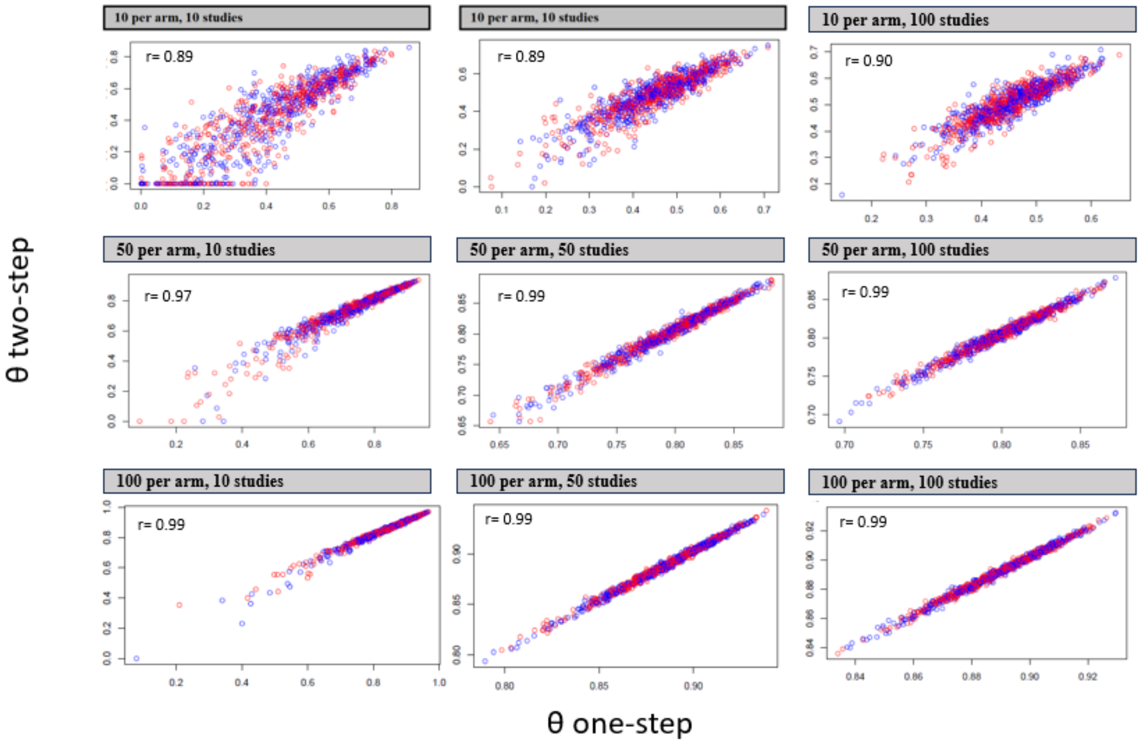
https://doi.org/10.26219/heal.aueb.6502

Appendix Figure 7. Correlation between I-squared one-stage and θ two-stage in high heterogeneity scenario ($\tau_{ct}^2 = 0.25,\ \sigma_e^2 = 0.65,\ \theta = 0$).

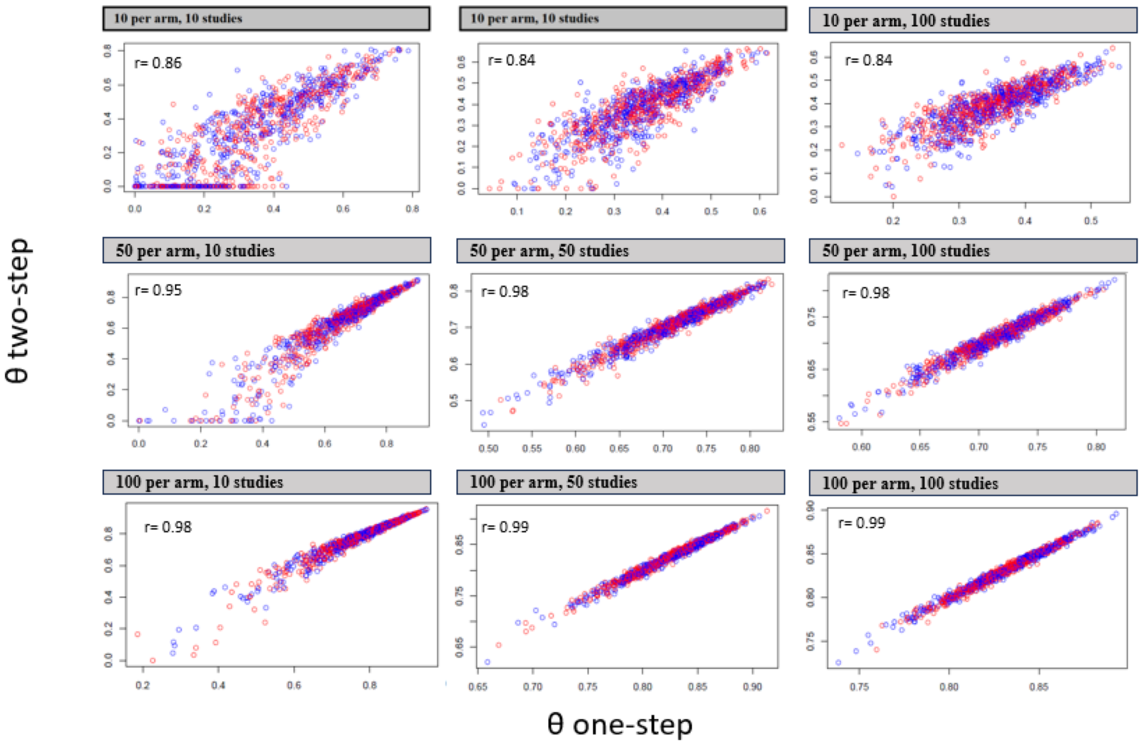


Appendix Figure 8. Correlation between I-squared one-stage and I-squared two-stage in moderate heterogeneity scenario: correlation scatterplots and coefficients ($\tau_{ct}^2 = 0.125$, $\sigma_e^2 = 0.75, \theta = 1$).

https://doi.org/10.26219/heal.aueb.6502



Appendix Figure 9. Correlation between I-squared one-stage and I-squared two-stage in low heterogeneity scenario: correlation scatterplots and coefficients ($\tau_{ct}^2 = 0.00125,\ \sigma_e^2 = 0.8725, \theta = 1$).

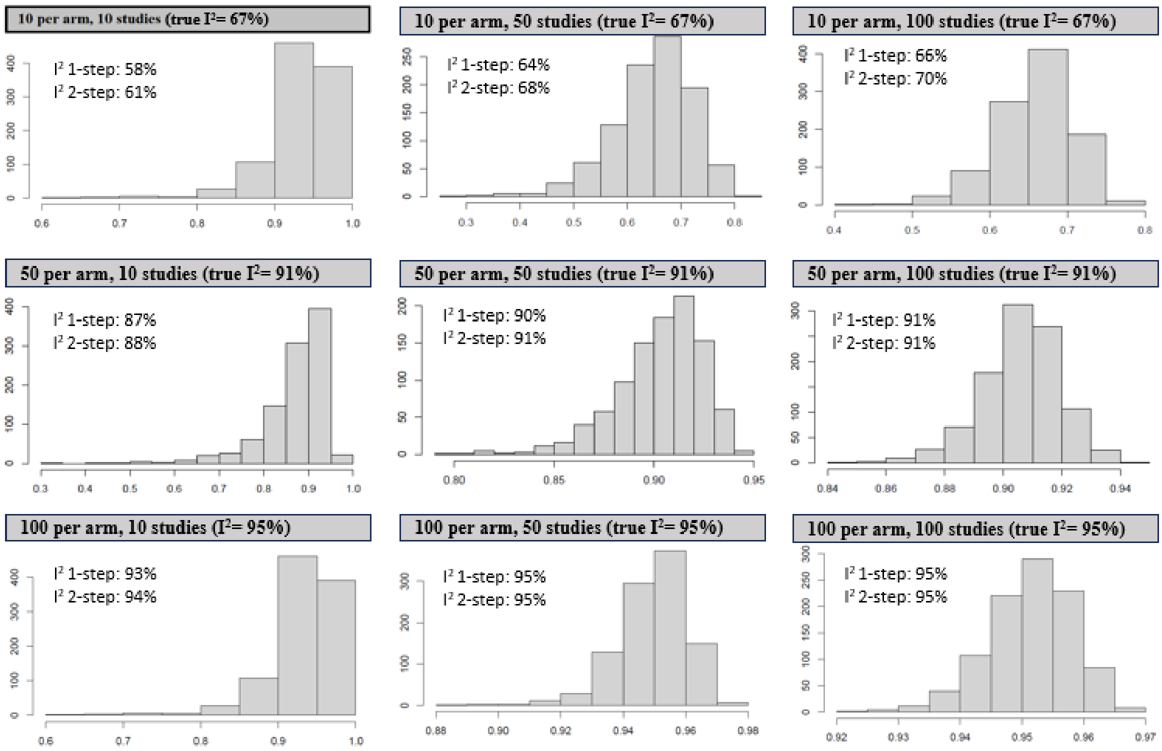


6.3.2 Evaluating Bias: A Comparative Analysis of I-Squared Values from One-Step and Two-Step Approaches Against True I-Squared Values in Simulation Study Results, theta=1

https://doi.org/10.26219/heal.aueb.6502

In the context of the analysis with a treatment effect of theta=1, our findings closely mirrored those of the theta=0 scenario across all examined cases. Just as in the previous simulation, the results remained virtually identical.

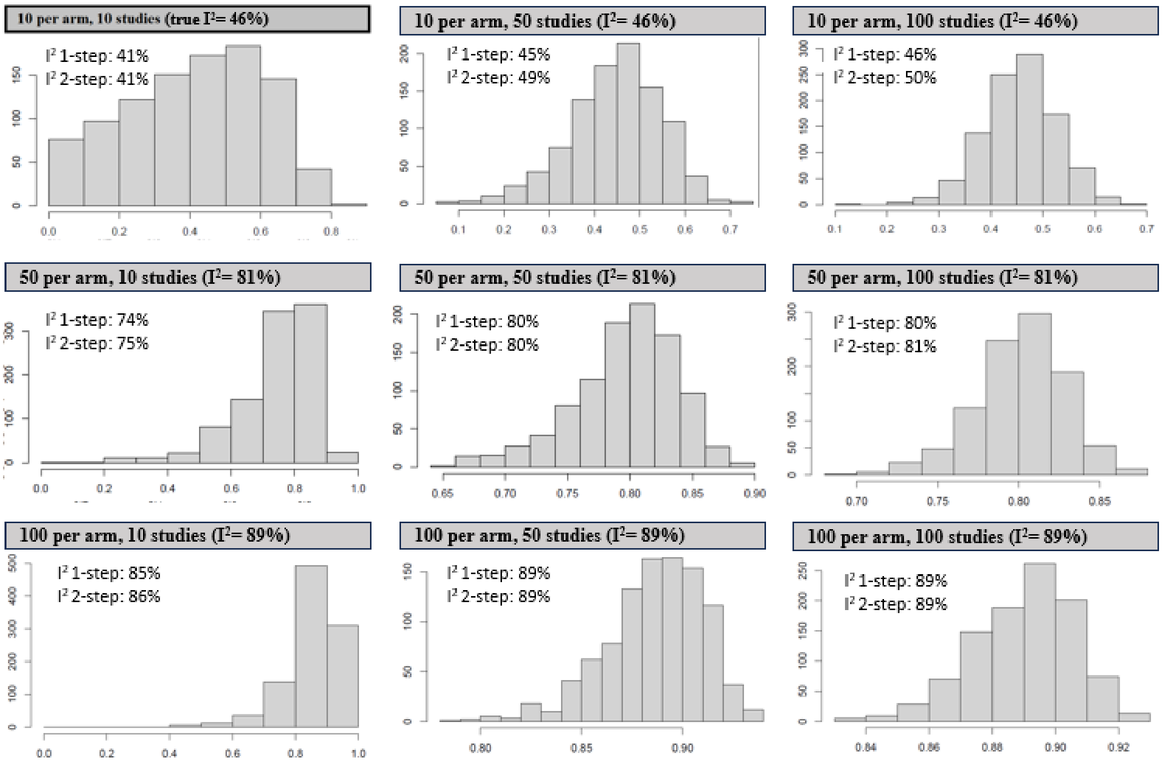
Appendix III Figure 10. Histogram illustrating the distribution of I-squared values obtained through the 1-step approach, alongside the means of I-squared values derived from both the 1-step and 2-step approaches and true I-squared values, high heterogeneity scenario ($\tau_{ct}^2 = 0.25,\ \sigma_e^2 = 0.65, \theta = 1$)



*Bias can be calculated by subtracting the true I-squared from the estimated I-squared (for both the one-stage and two-stage approaches) with for each particular scenario.

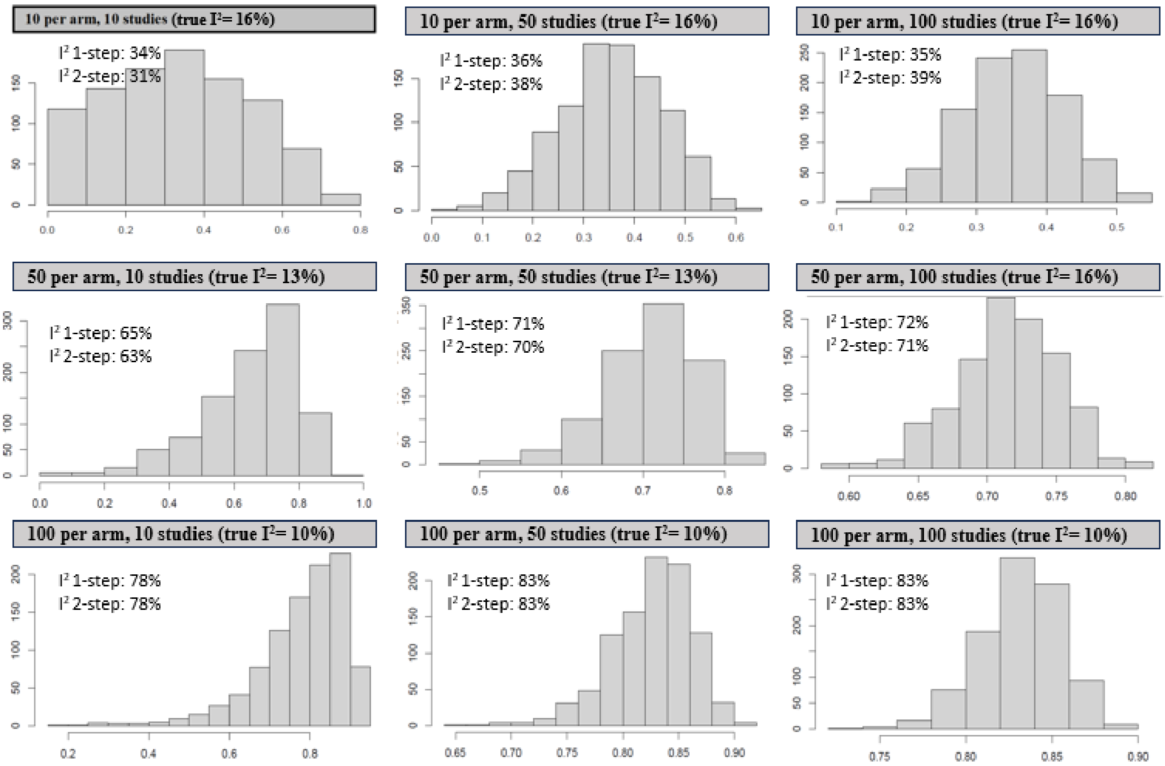
Appendix III Figure 11. Histogram illustrating the distribution of I-squared values obtained through the 1-step approach, alongside the means of I-squared values derived from both the 1-step and 2-step approaches and true I-squared values, moderate heterogeneity scenario $\tau_{ct}^2 = 0.125,\ \sigma_e^2 = 0.75, \theta = 1$)

https://doi.org/10.26219/heal.aueb.6502



*Bias can be calculated by subtracting the true I-squared from the estimated I-squared (for both the one-stage and two-stage approaches) with for each particular scenario.

Appendix Figure 12. Histogram illustrating the distribution of I-squared values obtained through the 1-step approach, alongside the means of I-squared values derived from both the 1-step and 2-step approaches and true I-squared values, low heterogeneity scenario ($\tau_{ct}^2 = 0.00125,\ \sigma_e^2 = 0.8725, \theta = 1$)



https://doi.org/10.26219/heal.aueb.6502

*Bias can be calculated by subtracting the true I-squared from the estimated I-squared (for both the one-stage and two-stage approaches) with for each particular scenario.

 https://doi.org/10.26219/heal.aueb.6502

7. References

Arnup, S. J., McKenzie, J. E., Hemming, K., Pilcher, D., & Forbes, A. B. (2017). Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial. *Trials*, *18*(1), 1-15.

Brown, H., & Prescott, R. (2015). *Applied mixed models in medicine*. John Wiley & Sons.

Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. *Statistics in Medicine*, *36*(5), 855–875. https://doi.org/10.1002/sim.7141

Campbell, M. J., Hemming, K., & Taljaard, M. (2019). The stepped wedge cluster randomised trial: What it is and when it should be used. *Medical Journal of Australia*, *210*(6), 253. https://doi.org/10.5694/mja2.50018

Donegan, S., Williams, L., Dias, S., Tudur-Smith, C., & Welton, N. (2015). Exploring Treatment by Covariate Interactions Using Subgroup Analysis and Meta-Regression in Cochrane Reviews: A Review of Recent Practice. *PLOS ONE*, *10*(6), e0128804. https://doi.org/10.1371/journal.pone.0128804

Donner, A., Klar, N., & Klar, N. S. (2000). *Design and analysis of cluster randomization trials in health research* (Vol. 27). London: Arnold.

Egger, M., Smith, G. D., & Phillips, A. N. (1997). Meta-analysis: Principles and procedures. *BMJ*, *315*(7121), 1533–1537. https://doi.org/10.1136/bmj.315.7121.1533

Feaster, D. J., Mikulich-Gilbertson, S., & Brincks, A. M. (2011). Modeling site effects in the design and analysis of multi-site trials. *The American journal of drug and alcohol abuse*, *37*(5), 383-391.

Fisher, D. J., Copas, A. J., Tierney, J. F., & Parmar, M. K. B. (2011). A critical review of methods for the assessment of patient-level interactions in individual participant data meta-

 https://doi.org/10.26219/heal.aueb.6502

analysis of randomized trials, and guidance for practitioners. *Journal of clinical epidemiology*, *64*(9), 949-967.

Hemming, K., Hughes, J. P., McKenzie, J. E., & Forbes, A. B. (2021). Extending the I-squared statistic to describe treatment effect heterogeneity in cluster, multi-centre randomized trials and individual patient data meta-analysis. *Statistical Methods in Medical Research*, *30*(2), 376–395. https://doi.org/10.1177/0962280220948550

Hemming, K., Taljaard, M., & Forbes, A. (2018). Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. *Statistics in medicine*, *37*(6), 883-898.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in medicine*, *20*(15), 2219-2241.

Hingorani, A. D., Windt, D. A. van der, Riley, R. D., Abrams, K., Moons, K. G. M., Steyerberg, E. W., Schroter, S., Sauerbrei, W., Altman, D. G., & Hemingway, H. (2013). Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ*, *346*, e5793. https://doi.org/10.1136/bmj.e5793

von Hippel, P. T. (2015). The heterogeneity statistic I2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*(1), 35. https://doi.org/10.1186/s12874-015-0024-z

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index?. *Psychological methods*, *11*(2), 193.

Hughes, J. P., Granston, T. S., & Heagerty, P. J. (2015). Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials*, *45*, 55–60. https://doi.org/10.1016/j.cct.2015.07.006

 https://doi.org/10.26219/heal.aueb.6502

Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, *28*(2), 182–191. https://doi.org/10.1016/j.cct.2006.05.007

Kahan, B. C., Forbes, G., Ali, Y., Jairath, V., Bremner, S., Harhay, M. O., ... & Leyrat, C. (2016). Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials*, *17*, 1-8.

Kahan, B. C., & Morris, T. P. (2013). Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects?. *Statistics in medicine*, *32*(7), 1136-1149.

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558

Kivimäki, M., Singh-Manoux, A., Pentti, J., Sabia, S., Nyberg, S. T., Alfredsson, L., Goldberg, M., Knutsson, A., Koskenvuo, M., Koskinen, A., Kouvonen, A., Nordin, M., Oksanen, T., Strandberg, T., Suominen, S. B., Theorell, T., Vahtera, J., Väänänen, A., Virtanen, M., … Jokela, M. (2019). Physical inactivity, cardiometabolic disease, and risk of dementia: An individual-participant meta-analysis. *BMJ*, *365*, l1495. https://doi.org/10.1136/bmj.l1495

Merino, J., Guasch-Ferré, M., Ellervik, C., Dashti, H. S., Sharp, S. J., Wu, P., Overvad, K., Sarnowski, C., Kuokkanen, M., Lemaitre, R. N., Justice, A. E., Ericson, U., Braun, K. V. E., Mahendran, Y., Frazier-Wood, A. C., Sun, D., Chu, A. Y., Tanaka, T., Luan, J., … Florez, J. C. (2019). Quality of dietary fat and genetic risk of type 2 diabetes: Individual participant data meta-analysis. *BMJ*, *366*, l4292. https://doi.org/10.1136/bmj.l4292

Morris, T. P., Fisher, D. J., Kenward, M. G., & Carpenter, J. R. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Statistics in Medicine*, *37*(9), 1419–1438. https://doi.org/10.1002/sim.7589

https://doi.org/10.26219/heal.aueb.6502

Parienti, J.-J., & Kuss, O. (2007). Cluster-crossover design: A method for limiting clusters level effect in community-intervention studies. *Contemporary Clinical Trials*, *28*(3), 316–323. https://doi.org/10.1016/j.cct.2006.10.004

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, *6*(1), 15–32.

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*(1), 79. https://doi.org/10.1186/1471-2288-8-79

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*(1), 79. https://doi.org/10.1186/1471-2288-8-79

Senn, S. (2014). A note regarding 'random effects.' *Statistics in Medicine*, *33*(16), 2876–2877. https://doi.org/10.1002/sim.5965

Tierney, J. F., Vale, C., Riley, R., Smith, C. T., Stewart, L., Clarke, M., & Rovers, M. (2015). Individual Participant Data (IPD) Meta-analyses of Randomised Controlled Trials: Guidance on Their Use. *PLOS Medicine*, *12*(7), e1001855. https://doi.org/10.1371/journal.pmed.1001855

Thompson, J. A., Fielding, K. L., Davey, C., Aiken, A. M., Hargreaves, J. R., & Hayes, R. J. (2017). Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Statistics in Medicine*, *36*(23), 3670–3682. https://doi.org/10.1002/sim.7348

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79. https://doi.org/10.1002/jrsm.1164