Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

An Alternative Measure for Quantifying the Heterogeneity in Meta-Analysis

Ke Yang[1] | Enxuan Lin[2] | Wangli Xu[3] | Liping Zhu[4] | Tiejun Tong[5]

[1]Department of Statistics and Data Science, Beijing University of Technology, Beijing, China | [2]Department of Biostatistics and Information, Innovent Biologics, Inc., Beijing, China | [3]Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China | [4]Institute of Statistics and Big Data, Renmin University of China, Beijing, China | [5]Department of Mathematics, Hong Kong Baptist University, Hong Kong

Correspondence: Tiejun Tong (tongt@hkbu.edu.hk)

Received: 25 March 2024 | **Revised:** 14 February 2025 | **Accepted:** 31 March 2025

Funding: This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFA1003702), Initiation Grant for Faculty Niche Research Areas of Hong Kong Baptist University (RC-FNRA-IG/23-24/SCI/03), National Natural Science Foundation of China (Grant Nos. 12071305 and 12371294), MOE Project of Key Research Institute of Humanities and Social Sciences (Grant No. 22JJD910001), and General Research Fund of Hong Kong (Grant Nos. HKBU12300123 and HKBU12303421).

Keywords: ANOVA | heterogeneity | intraclass correlation coefficient | meta-analysis | the I^2 statistic | the I_A^2 statistic

ABSTRACT

Quantifying the heterogeneity is an important issue in meta-analysis, and among the existing measures, the I^2 statistic is most commonly used. In this article, we first illustrate with a simple example that the I^2 statistic is heavily dependent on the study sample sizes, mainly because it is used to quantify the heterogeneity between the observed effect sizes. To reduce the influence of sample sizes, we introduce an alternative measure that aims to directly measure the heterogeneity between the study populations involved in the meta-analysis. We further propose a new estimator, namely the I_A^2 statistic, to estimate the newly defined measure of heterogeneity. For practical implementation, the exact formulas of the I_A^2 statistic are also derived under two common scenarios with the effect size as the mean difference (MD) or the standardized mean difference (SMD). Simulations and real data analyses demonstrate that the I_A^2 statistic provides an asymptotically unbiased estimator for the absolute heterogeneity between the study populations, and it is also independent of the study sample sizes as expected. To conclude, our newly defined I_A^2 statistic can be used as a supplemental measure of heterogeneity to monitor the situations where the study effect sizes are indeed similar with little biological difference. In such scenario, the fixed-effect model can be appropriate; nevertheless, when the sample sizes are sufficiently large, the I^2 statistic may still increase to 1 and subsequently suggest the random-effects model for meta-analysis.

1 | Introduction

Meta-analysis is a statistical technique for evidence-based practice, which aims to synthesize multiple studies and produce a summary conclusion for the whole body of research [1]. In the literature, there are two commonly used statistical models for meta-analysis, namely, the fixed-effect model and the random-effects model. Among them, the fixed-effect model assumes that the effect sizes of different studies are all the same, which is somewhat restrictive and may not be realistic in practice. The effect sizes often differ between the studies due to variability in study design, outcome measurement tools, risk of bias, and the participants, interventions and outcomes studied [2], etc. Such diversity in the effect sizes is known as the heterogeneity.

When the heterogeneity exists, the random-effects model ought to be applied for meta-analysis. In such scenarios, it is of great importance to properly quantify the heterogeneity so as to explore the generalizability of the findings from a meta-analysis.

To describe the heterogeneity in detail, we first introduce the random-effects model for meta-analysis. Let $k \geq 2$ be the total number of studies, and y_i be the observed effect sizes from the studies $i = 1, \ldots, k$. For each study with true effect size μ_i, we assume that y_i is normally distributed with mean $\mu_i = E(y_i|\mu_i)$ and variance $\sigma^2_{y_i} = \text{var}(y_i|\mu_i)$. Moreover, to account for the heterogeneity between the studies, we also assume that the individual effect sizes μ_i follow another normal distribution with mean μ and variance $\tau^2 > 0$. Taken together, the random-effects model for meta-analysis can be expressed as

$$y_i = \mu + \delta_i + \epsilon_i, \quad \delta_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2), \quad \epsilon_i \overset{\text{ind}}{\sim} N(0, \sigma^2_{y_i}) \tag{1}$$

where "i.i.d." represents independent and identically distributed, "ind" represents independently distributed, τ^2 is the between-study variance, and $\sigma^2_{y_i}$ are the within-study variances. In addition, the study deviations $\delta_i = \mu_i - \mu$ and the random errors ϵ_i are assumed to be independent of each other. When δ_i are all zero, model (1) reduces to the fixed-effect model and there is no heterogeneity between the studies.

To test the existence of heterogeneity for model (1), Cochran [3] proposed the Q statistic as

$$Q = \sum_{i=1}^{k} w_i \left(y_i - \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \right)^2 \tag{2}$$

where $w_i = 1/\sigma^2_{y_i}$ are the inverse-variance weights. Noting that $\sigma^2_{y_i}$ can often be estimated with high precision, it is a common practice in meta-analysis that the within-study variances are regarded as known. Nevertheless, when used as a measure of heterogeneity, it is often criticized that the value of Q will increase with the number of studies. Another measure for heterogeneity is to apply the between-study variance τ^2, yet it is known to be specific to a particular effect metric, making it impossible to compare across different meta-analyses [4]. To have a fair comparison, Higgins and Thompson [5] and Higgins et al. [6] introduced the I^2 statistic by a two-step procedure, under the assumption that the within-study variances $\sigma^2_{y_i} = \sigma^2_y$ are all the same. They first defined the measure of heterogeneity between the studies as

$$\text{ICC}_{\text{HT}} = \frac{\tau^2}{\text{var}(y_i)} = \frac{\tau^2}{\tau^2 + \sigma^2_y} \tag{3}$$

and then proposed

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \tilde{\sigma}^2_y} = \max\left\{ \frac{Q - (k-1)}{Q}, 0 \right\} \tag{4}$$

to estimate the unknown ICC_{HT}, where $\hat{\tau}^2 = \max\{\{Q - (k-1)\}/(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i), 0\}$ is the DerSimonian-Laird estimator [4] and $\tilde{\sigma}^2_y = (k-1)/(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i)$. When the within-study variances are all the same, $\tilde{\sigma}^2_y$ is identical to the common σ^2_y. Otherwise if they differ, Böhning et al. [7] has showed that $\tilde{\sigma}^2_y$ is asymptotically identical to the harmonic mean $(\sum_{i=1}^k w_i/k)^{-1}$ of the within-study variances. Moreover, the I^2 statistic is also guaranteed to be within the interval $[0, 1)$, which is appealing in that it does not depend on the number of studies and is irrespective of the effect metric.

Thanks to its nice properties, the I^2 statistic is nowadays routinely reported in the forest plots for meta-analyses, and/or used as a criterion for model selection between the fixed-effect model and the random-effects model. In Google Scholar, as of January 2025, the two articles by Higgins and Thompson [5] and Higgins et al. [6] have been cited more than 32,000 and 57,000 times, respectively. Despite of its huge popularity, there were evidences in the literature reporting the limitations of the I^2 statistic. In particular, Rücker et al. [8] found that the I^2 statistic always increases rapidly to 1 when the sample sizes are large, regardless of whether or not the heterogeneity between the studies is clinically important. For other discussions on the I^2 statistic as a measure of heterogeneity, one may refer to, for example, Riley et al. [9], IntHout et al. [10], Borenstein et al. [11], Sangnawakij et al. [12], Holling et al. [13], and the references therein. This motivates us to further explore the characteristics of the I^2 statistic as a measure of heterogeneity for meta-analysis.

To answer this question, we first present a motivating example to demonstrate that the I^2 statistic was defined to quantify the heterogeneity between the observed effect sizes rather than that between the study populations. In view of this, we regard the I^2 statistic as a relative measure of heterogeneity. We further draw a connection between the one-way analysis of variance (ANOVA) and the random-effects meta-analysis, and subsequently introduce an alternative measure for quantifying the heterogeneity in the random-effects model, which is independent of study sample sizes and can serve as an absolute measure of heterogeneity. For details, see Section 3.2 for the defined ICC_{MA} in formula (8). To move forward, the statistical properties of ICC_{MA} are also derived that explore the distinction between our new measure and the existing measures, together with an asymptotically unbiased estimator of the unknown ICC_{MA} based on ANOVA. Lastly and most importantly, we also manage to provide an easy-to-implement estimator, namely the I^2_{A} statistic, to estimate ICC_{MA} based on the Q statistic, in a way similar for I^2 in (4) to estimate ICC_{HT}.

The remainder of the article is organized as follows. In Section 2, we give a motivating example to illustrate that ICC_{HT} heavily depends on the study sample sizes. In Section 3, by drawing a close connection between ANOVA and the random-effects meta-analysis, we introduce an alternative measure for quantifying the heterogeneity between the studies, namely ICC_{MA}, and then provide an ANOVA-based method to estimate this measure. In Sections 4–6, we further derive the easy-to-implement I^2_{A} statistic to estimate the new heterogeneity measure ICC_{MA}, using the Q statistic under three common scenarios with the raw mean, the mean difference, or the standardized mean difference as the effect size, respectively. While for practical implementation, real data analysis and numerical results are also presented for each scenario. Finally, we conclude the article in Section 7 and provide the technical details in the Appendices A–F.

2 | A Motivating Example

In this section, we illustrate how $\mathrm{ICC}_{\mathrm{HT}}$ in (3) varies along with the sample sizes, and so may not be able to serve as a measure of heterogeneity between the study populations. To confirm this claim, we first consider a motivating example of three studies with data generated from normal populations $N(-0.05, 1)$, $N(0, 1)$, and $N(0.05, 1)$, respectively. From the top-left panel of Figure 1, it is evident that the three study populations are largely overlapped. Taken the three study means as a random sample, the between-study variance can be estimated by the sample variance as $\tilde{\tau}^2 = \{(-0.05-0)^2 + (0-0)^2 + (0.05-0)^2\}/2 = 0.0025$.

To explain why $\mathrm{ICC}_{\mathrm{HT}}$ is not a measure of heterogeneity between the study populations, we consider two scenarios to meta-analyze the three studies, with the population means being treated as the effect sizes. The first scenario assumes $n = 400$ patients in each study. By taking the sample means, the sampling distributions of the observed effect sizes are thus $N(-0.05, 0.0025)$, $N(0, 0.0025)$, and $N(0.05, 0.0025)$, respectively, yielding $\sigma_y^2 = 0.0025$ as the common within-study variance. Further by the definition in (3), we have

$$\mathrm{ICC}_{\mathrm{HT}} \approx \frac{0.0025}{0.0025 + 0.0025} = 50\%$$

In the second scenario, we consider $n = 4000$ for each study. This leads to the sampling distributions of the observed effect sizes as $N(-0.05, 0.00025)$, $N(0, 0.00025)$, and $N(0.05, 0.00025)$, respectively. Further by $\sigma_y^2 = 0.00025$, the measure of heterogeneity is

$$\mathrm{ICC}_{\mathrm{HT}} \approx \frac{0.00025}{0.00025 + 0.0025} = 90.9\%$$

Finally, for ease of comparison, we also plot the sampling distributions of the observed effect sizes in Figure 1 for the two hypothetical scenarios with varying study sample sizes.

The above example clearly shows that $\mathrm{ICC}_{\mathrm{HT}}$, defined in (3) by Higgins and Thompson (2002), measures the heterogeneity between the observed effect sizes and thus heavily depends on the study sample sizes. In other words, $\mathrm{ICC}_{\mathrm{HT}}$ is a relative measure of heterogeneity for meta-analysis. Consequently, as a sample estimate of $\mathrm{ICC}_{\mathrm{HT}}$, the I^2 statistic is also heavily dependent on the sample sizes. This coincides with the observations by Rücker et al. [8]. Specifically, in our motivating example, $\mathrm{ICC}_{\mathrm{HT}}$ increases rapidly to about 90% when the sample sizes are 4000, even though it is evident that the three populations are largely overlapped with each other. To summarize, when the study sample sizes n_i are large enough, it will always yield an I^2 value being close to 1. On the other hand, compared with the population variance 1, the differences between the three study means $(-0.05, 0, 0.05)$ may not be clinically important. To support this claim, we note that the Scientific Committee of the European Food Safety Authority have also emphasized the importance of assessing the biological differences [14]. This hence motivates us to introduce an alternative measure that quantifies the heterogeneity between the study populations directly, in a way to reduce the influence of sample sizes.

3 | A New Measure of Heterogeneity and the I^2_{ANOVA} Statistic

To further explore the characteristics of $\mathrm{ICC}_{\mathrm{HT}}$, we also draw in this section an interesting connection between one-way analysis of variance (ANOVA) and meta-analysis. And on the basis of that, a new measure for quantifying the heterogeneity between the study populations will be introduced, and moreover by studying its statistical properties, it is also clarified why it can add new value to meta-analysis. Lastly for completeness, we also provide an asymptotically unbiased estimator, namely the I^2_{ANOVA} statistic, to estimate the new measure of heterogeneity in Section 3.3. Nevertheless, as will be seen, the I^2_{ANOVA} statistic may not be easy to implement for practitioners, which motivates us to further propose a much simpler and more elegant estimator in Sections 4–6 based on the Q statistic. For readers who are not familiar with ANOVA, Section 3.3 can be skipped without affecting the subsequent reading.

3.1 | Connection Between ANOVA and Meta-Analysis

To introduce the one-way ANOVA, we let y_{ij} be the jth observation in the ith population, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$,

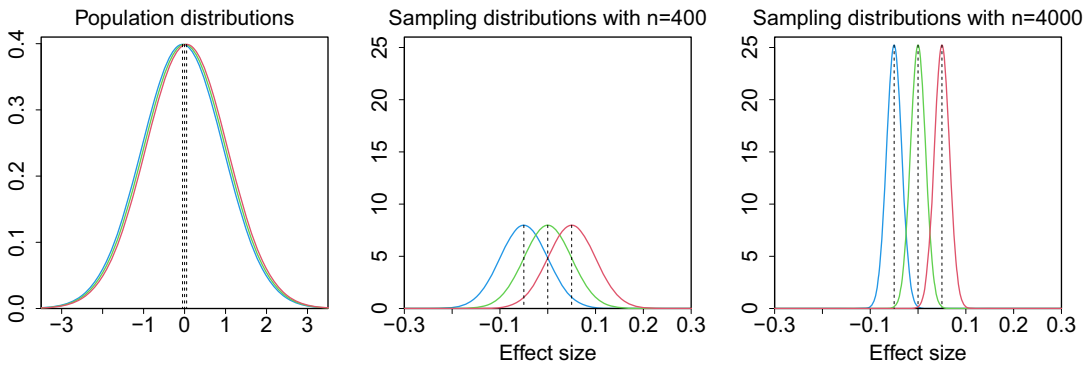

FIGURE 1 | Population distributions of the three studies and the sampling distributions of the observed effect sizes. Left panel: population distributions are $N(-0.05, 1)$ in blue, $N(0, 1)$ in green, and $N(0.05, 1)$ in red, respectively. Middle panel: sampling distributions are $N(-0.05, 0.0025)$, $N(0, 0.0025)$, and $N(0.05, 0.0025)$, respectively. Right panel: sampling distributions are $N(-0.05, 0.00025)$, $N(0, 0.00025)$, and $N(0.05, 0.00025)$, respectively.

where k is the number of studies and n_i are the study sample sizes from each population. The random-effects ANOVA for the observed data is then

$$y_{ij} = \mu + \delta_i + \xi_{ij}, \quad \delta_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2), \quad \xi_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \tag{5}$$

where μ is the grand mean, δ_i are the treatment effects, and ξ_{ij} are the random errors. We further assume that δ_i are i.i.d. normal random variables with mean 0 and variance $\tau^2 \geq 0$, ξ_{ij} are i.i.d. normal random errors with mean 0 and variance $\sigma^2 > 0$, and that δ_i and ξ_{ij} are independent of each other. In addition, we refer to $\mu_i = \mu + \delta_i$ as the individual means, τ^2 as the between-study variance, σ^2 as the common error variance for all k populations, and $\tau^2 + \sigma^2$ as the total variance of each observation.

To draw a close connection between ANOVA and meta-analysis, we consider a hypothetical scenario in which the experimenter first computed the sample mean and its variance for each population, namely $y_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\hat{\sigma}^2_{y_i} = \sum_{j=1}^{n_i}(y_{ij} - y_i)^2/\{n_i(n_i - 1)\}$ for $i = 1, \ldots, k$, and then reported these summary data rather than the raw data to the public. In practice, there are reasons why one must do so, including, for example, due to the privacy protection for which the individual patient data cannot be released. Under such a scenario, if some researchers want to re-analyze the experiment using only the publicly available data, it then yields a new random-effects model as

$$y_i = \mu + \delta_i + \epsilon_i, \quad \delta_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2), \quad \epsilon_i \overset{\text{ind}}{\sim} N(0, \sigma^2/n_i) \tag{6}$$

where y_i are the sample means, μ and δ_i are the same as defined in model (5), and $\epsilon_i = \sum_{j=1}^{n_i} \xi_{ij}/n_i$ are independent random errors with mean 0 and variance σ^2/n_i, where $i = 1, \ldots, k$. Now from the point of view of meta-analysis, if we treat y_i as the reported effect sizes and $\hat{\sigma}^2_{y_i}$ as the within-study variances representing σ^2/n_i, then model (6) is essentially the same as the random-effects model in (1). This interesting connection shows that, when the ANOVA model with raw data only releases the summary data to the public, it will then yield a meta-analysis model with summary data.

For ease of comparison, we also summarize some key components in Table 1 for both the ANOVA model in (5) and the meta-analysis model in (6). For the meta-analysis model, under the assumption that the within-study variances, i.e., σ^2/n_i, are all equal, Higgins and Thompson [5] interpreted the measure of heterogeneity as the proportion of total variance that is "between the studies". More specifically, by the last column of Table 1, they introduced the measure of heterogeneity for meta-analysis as in (3), where $\sigma^2_y = \sigma^2/n_i$ is the common within-study variance for the observed effect sizes. This clearly explains why ICC_{HT} will be heavily dependent on the study sample sizes. When the sample sizes go to infinity, the within-study variances will converge to zero so that ICC_{HT} will increase to 1, as having been observed in Rücker et al. [8]. This also coincides with our motivating example in Section 2 that, when the sample size varies from 400 to 4000, their measure of heterogeneity will increase from 50% to about 90%, regardless of whether or not the heterogeneity between the studies is clinically important.

TABLE 1 | Connection between the ANOVA model in (5) and the meta-analysis model in (6), where $y_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\epsilon_i = \sum_{j=1}^{n_i} \xi_{ij}/n_i$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$.

	ANOVA	Meta-analysis
Model	$y_{ij} = \mu + \delta_i + \xi_{ij}$	$y_i = \mu + \delta_i + \epsilon_i$
Between-study variance	τ^2	τ^2
Error (or within-study) variance	σ^2	σ^2/n_i
Total variance	$\text{var}(y_{ij}) = \tau^2 + \sigma^2$	$\text{var}(y_i) = \tau^2 + \sigma^2/n_i$

For the ANOVA model, it is well known that the intraclass correlation coefficient (ICC) is the most commonly used measure of heterogeneity [15–18], which interprets the proportion of total variance that is "between populations". More specifically, by Table 1, ICC can be expressed as

$$\text{ICC} = \frac{\tau^2}{\text{var}(y_{ij})} = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{7}$$

As shown in the hypothetical scenario, the ANOVA model in (5) and the meta-analysis model in (6) are, in fact, modeling the same populations, even though one uses the raw data and the other uses the summary data. In the special case when the mean value is taken as the effect size, it is known that the sample mean is a sufficient and complete statistic for the normal mean; in other words, the raw data and the summary data contain exactly the same information regarding the effect size. With this insight, we expect that the measures of heterogeneity between the study populations for the two models should also be the same, regardless of whether the raw data or the summary data are being used.

3.2 | An Intrinsic Measure of Heterogeneity

Inspired by the intrinsic connection between ANOVA and meta-analysis, we now follow the same assumption as in ANOVA that the population variances $n_i\sigma^2_{y_i}$ are all equal. For ease of presentation, we also denote the common study population variance as σ^2_{pop}. Then by following ICC in (7) for ANOVA, we propose the following measure of heterogeneity for meta-analysis:

$$\text{ICC}_{\text{MA}} = \frac{\tau^2}{\text{var}(y_{ij})} = \frac{\tau^2}{\tau^2 + \sigma^2_{\text{pop}}} \tag{8}$$

Note that the range of ICC_{MA} is always within the interval $[0, 1)$. Regarding the rationale of ICC_{MA} for meta-analysis, one may also refer to the proposed measure in Sangnawakij et al. [12]. And as mentioned in Section 2, a common population variance can be a more reasonable assumption for meta-analysis compared to a common within-study variance for all studies, in a way to mitigate the impact caused by the study sample sizes.

To further study the properties of ICC_{MA} and explain why it can serve as an absolute measure of heterogeneity for meta-analysis, we first present the three statistical properties of ICC_{HT} as follows.

i *Monotonicity*. $\mathrm{ICC}_{\mathrm{HT}}$ is a monotonically increasing function of the ratio τ^2/σ_y^2. When the common within-study variance σ_y^2 is fixed, $\mathrm{ICC}_{\mathrm{HT}}$ will solely increase with the between-study variance τ^2. This property was referred to as the "dependence on the extent of heterogeneity" by Higgins and Thompson [5].

ii *Location and scale invariance*. $\mathrm{ICC}_{\mathrm{HT}}$ is not affected by the location and scale of the effect sizes. This property was referred to as the "scale invariance" by Higgins and Thompson [5].

iii *Study size invariance*. $\mathrm{ICC}_{\mathrm{HT}}$ is not affected by the total number of studies k. This property was referred to as the "size invariance" by Higgins and Thompson [5].

Thanks to the above properties, the I^2 statistic is nowadays the most popular measure for quantifying the heterogeneity in meta-analysis, compared to other existing measures including Q and τ^2. Nevertheless, we do wish to point out that the "size invariance" in their property (iii) only represents the study size invariance but not includes the sample size invariance. As shown in the motivating example and also from the historical evidence in the literature, $\mathrm{ICC}_{\mathrm{HT}}$ does suffer from a heavy dependence on the study sample sizes.

While for the new measure of heterogeneity in (8), we show in Appendix A that $\mathrm{ICC}_{\mathrm{MA}}$ shares the following four properties:

i′ *Monotonicity*. $\mathrm{ICC}_{\mathrm{MA}}$ is a monotonically increasing function of the ratio $\tau^2/\sigma^2_{\mathrm{pop}}$. When the common population variance σ^2_{pop} is fixed, $\mathrm{ICC}_{\mathrm{MA}}$ will solely increase with the between-study variance τ^2.

ii′ *Location and scale invariance*. $\mathrm{ICC}_{\mathrm{MA}}$ is not affected by the location and scale of the effect sizes.

iii′ *Study size invariance*. $\mathrm{ICC}_{\mathrm{MA}}$ is not affected by the total number of studies k.

iv′ *Sample size invariance*. $\mathrm{ICC}_{\mathrm{MA}}$ is not affected by the sample size n_i of each study.

Note that the first three properties for $\mathrm{ICC}_{\mathrm{MA}}$ are essentially the same as those for $\mathrm{ICC}_{\mathrm{HT}}$. While for the importance of property (iv′), let us illustrate again using the motivating example in Section 2. Under the assumption of a common population variance, the term σ^2_{pop} remains constant at 1 no matter how the sample sizes vary. Further by (8), the value of $\mathrm{ICC}_{\mathrm{MA}}$ under each scenario will always be $0.0025/(0.0025+1) \approx 0.25\%$, indicating that the three study populations are indeed highly overlapped with a small amount of heterogeneity. To conclude, it is because of the sample size invariance in property (iv′) that distinguishes our new $\mathrm{ICC}_{\mathrm{MA}}$ from the existing $\mathrm{ICC}_{\mathrm{HT}}$, which also perfectly explains why $\mathrm{ICC}_{\mathrm{MA}}$ can serve as a new measure for quantifying the heterogeneity between the study populations. Due to its sample size invariance, we regard $\mathrm{ICC}_{\mathrm{MA}}$ as an absolute measure of heterogeneity.

3.3 | Estimation of $\mathrm{ICC}_{\mathrm{MA}}$ Based on ANOVA

In ANOVA, there has been extensive and well-established research on the estimation of ICC. For easy reference, we have also provided a brief review in Appendix B. Among the existing methods, it is known that the ANOVA estimator is the most widely used thanks to its straightforwardness and effectiveness. Inspired by this, we also propose an ANOVA-based estimator for $\mathrm{ICC}_{\mathrm{MA}}$ in the framework of meta-analysis.

Following the random-effects ANOVA in (5), the total variance of the observations is given by $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$, which can be divided into two components as the sum of squares between the populations and the error sum of squares within the populations. Based on this variance partitioning, Cochran [19] derived the method of moments estimators of τ^2 and σ^2, and then by plugging them into formula (7), it yields the well-known ANOVA estimator for the unknown ICC. In parallel, following the random-effects model for meta-analysis in (1), we first assume that $\hat{\sigma}^2_{y_i}$ are the estimated within-study variance from each study, as also mentioned in Section 3.1. We further define the mean square between the populations (MSB) as

$$\mathrm{MSB}_{\mathrm{MA}} = \frac{1}{k-1}\sum_{i=1}^{k}\left\{n_i\left(y_i-\bar{y}\right)^2\right\} \tag{9}$$

where $\bar{y} = \sum_{i=1}^{k}(n_i y_i)/\sum_{i=1}^{k} n_i$, and the mean square within the populations (MSW) as

$$\mathrm{MSW}_{\mathrm{MA}} = \frac{1}{\sum_{i=1}^{k}\left(n_i-1\right)}\sum_{i=1}^{k}\left\{n_i\left(n_i-1\right)\hat{\sigma}^2_{y_i}\right\} \tag{10}$$

Moreover, let

$$\tilde{n} = \frac{1}{k-1}\left(\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} n_i^2 \Big/ \sum_{i=1}^{k} n_i\right) \tag{11}$$

be the adjusted mean sample size [20] that accounts for the variation of the sample sizes from different studies. Then by the same method for estimating ICC, our new estimator for $\mathrm{ICC}_{\mathrm{MA}}$ is given as

$$I^2_{\mathrm{ANOVA}} = \max\left\{\frac{\mathrm{MSB}_{\mathrm{MA}} - \mathrm{MSW}_{\mathrm{MA}}}{\mathrm{MSB}_{\mathrm{MA}} + (\tilde{n}-1)\mathrm{MSW}_{\mathrm{MA}}}, 0\right\} \tag{12}$$

Similar to the I^2 statistic in (4), the maximum operation is taken to avoid a negative estimate. For more details on the derivation of I^2_{ANOVA}, one may refer to Appendix C.

Up to now, we have used the generic notation y_i as the observed effect size, together with its standard error $\hat{\sigma}_{y_i}$ and the sample size n_i. Note that this is the simplest scenario, in which the effect size is represented by the mean y_i from each study with only one arm. In addition to the mean, two other commonly used effect sizes for continuous outcomes are the mean difference (MD) and the standardized mean difference (SMD), which are applicable to meta-analysis of studies with two arms. In the next two paragraphs, we show that the I^2_{ANOVA} statistic in (12) can be directly generalized to handle these two scenarios.

For a meta-analysis of MD, the summary statistics for the ith study often consist of the observed MD y_i, the sample sizes n_i^T and n_i^C, and the standard errors $\hat{\sigma}_{y_i^T}$ and $\hat{\sigma}_{y_i^C}$ associated with the treatment and control groups. With these notations, the mean square within the populations can be computed as

$$\mathrm{MSW}_{\mathrm{MA}} = \frac{\sum_{i=1}^{k}\left\{n_i^T\left(n_i^T-1\right)\hat{\sigma}^2_{y_i^T} + n_i^C\left(n_i^C-1\right)\hat{\sigma}^2_{y_i^C}\right\}}{\sum_{i=1}^{k}\left(n_i^T+n_i^C\right)-2k}$$

In addition, by defining the effective sample size as $n_i = 1/(1/n_i^T + 1/n_i^C)$ for each study, the formulas (9) and (11) can also be directly followed to calculate $\mathrm{MSB}_{\mathrm{MA}}$ and the adjusted mean sample size $\tilde{n}$. Lastly by (12), we can derive the I^2_{ANOVA} statistic as the estimated measure of heterogeneity. For more details about the meta-analysis of MD including the statistical models and the underlying assumptions, one may refer to Appendix D.

For a meta-analysis of SMD, the summary statistics will instead report the observed SMD y_i for each study, together with the sample sizes n_i^T and n_i^C, and the standard errors $\hat{\sigma}_{y_i^T}$ and $\hat{\sigma}_{y_i^C}$. Then to compute the I^2_{ANOVA} statistic by (12), we note that the same procedure as that for MD can still be followed to determine the values of $\mathrm{MSB}_{\mathrm{MA}}$ and $\tilde{n}$. And moreover, we can also set $\mathrm{MSW}_{\mathrm{MA}}$ directly to 1 since the observed effect sizes are already standardized. For a comprehensive understanding of the model specifications for meta-analysis of SMD, one may refer to Appendix E.

4 | The I^2_{A} Statistic for the Mean

Recall that to estimate $\mathrm{ICC}_{\mathrm{HT}}$, Higgins and Thompson [5] proposed the easy-to-implement I^2 statistic based on the Q statistic. Nevertheless, for our new measure of heterogeneity $\mathrm{ICC}_{\mathrm{MA}}$, the ANOVA-based estimator in (12) is somewhat complicated and may not be familiar for meta-analysts. This motivates us to further propose a new estimator of $\mathrm{ICC}_{\mathrm{MA}}$, referred to as I^2_{A}, which turns out to have a similar form as the I^2 statistic. More specifically, we will present the I^2_{A} statistic in Sections 4 to 6 for the three effect sizes including the mean, MD and SMD, respectively, followed by real data analyses and simulation studies that compare the numerical performance of the I^2, I^2_{ANOVA} and I^2_{A} statistics.

By (3), $\mathrm{ICC}_{\mathrm{HT}}$ is defined based on the assumption of a common within-study variance. When this assumption does not hold, as pointed out in the literature, the common within-study variance σ_y^2 can be replaced by $\tilde{\sigma}_y^2$ in (4) as an average of the k within-study variances. Now for $\mathrm{ICC}_{\mathrm{MA}}$ in (8), we have $\sigma^2_{\mathrm{pop}} = n_i\sigma^2_{y_i}$, and consequently, $w_i = 1/\sigma^2_{y_i} = n_i/\sigma^2_{\mathrm{pop}}$. Then by letting $\tilde{n} = (\sum_{i=1}^k n_i - \sum_{i=1}^k n_i^2/\sum_{i=1}^k n_i)/(k-1)$ be the adjusted mean sample size as in (11), an identity between $\tilde{\sigma}_y^2$ and σ^2_{pop} can be established as follows:

$$\tilde{\sigma}_y^2 = \frac{k-1}{\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i} = \frac{1}{\tilde{n}}\sigma^2_{\mathrm{pop}} \tag{13}$$

In addition, since $E\{Q/(k-1)-1\} = \tau^2/\tilde{\sigma}_y^2$ by Higgins and Thompson [5], it follows that

$$E\left(\frac{Q-(k-1)}{(k-1)\tilde{n}}\right) = \frac{\tau^2}{\sigma^2_{\mathrm{pop}}}$$

which leads to a method of moments estimator of $\tau^2/\sigma^2_{\mathrm{pop}}$ as $\{Q-(k-1)\}/\{(k-1)\tilde{n}\}$. Lastly, by noting that $\mathrm{ICC}_{\mathrm{MA}} = (\tau^2/\sigma^2_{\mathrm{pop}})/(\tau^2/\sigma^2_{\mathrm{pop}}+1)$, our plug-in estimator for $\mathrm{ICC}_{\mathrm{MA}}$ is given as

$$I^2_{\mathrm{A}} = \max\left\{\frac{Q-(k-1)}{Q+(k-1)(\tilde{n}-1)}, 0\right\} \tag{14}$$

where, as usual, the maximum operation is kept to avoid a negative estimate. By comparing (4) and (14), we note that the difference between the I^2 and I^2_{A} statistics is purely on the term $(k-1)(\tilde{n}-1)$, which is a function of the study sample sizes and the number of studies. In the special case when $\tilde{n}=1$, the two statistics will be exactly the same.

For more insights on how the estimated heterogeneity is adjusted from the relative measure to the absolute measure by the study sample sizes through $(k-1)(\tilde{n}-1)$, we summarize below a few interesting findings with the proofs in Appendix F.

a. First, we have $(k-1)(\tilde{n}-1)\geq 0$, where the equality holds only when $n_i = 1$ for all k studies. Consequently, it yields that
$$I^2_{\mathrm{A}} \leq I^2$$
under all the settings of meta-analysis with at least 2 studies.

b. For the balanced design where all sample sizes are equal to n, we have $\tilde{n} = n$ and moreover $(k-1)(\tilde{n}-1) = (k-1)(n-1)$. When $k\to\infty$ and $n\to\infty$, we can further show that $Q/\{(k-1)(n-1)\} \to \tau^2/\sigma^2_{\mathrm{pop}} = O(1)$, indicating that the two terms in the denominator of (14) are of the same asymptotic order. Moreover by Slutsky's theorem,
$$I^2_{\mathrm{A}} \to \frac{\tau^2/\sigma^2_{\mathrm{pop}}}{\tau^2/\sigma^2_{\mathrm{pop}}+1} = \mathrm{ICC}_{\mathrm{MA}} < 1$$
On the other hand, noting that $Q/(k-1) = O(n)$ for any fixed $k\geq 2$, we have $I^2 = 1 - O(1/n) \to 1$ as $n\to\infty$. Taken together, it clearly explains why I^2 may asymptotically increase to 1, whereas our new I^2_{A} will not.

c. For the unbalanced design, it can also be shown that $(k-1)(\tilde{n}-1)$ is an increasing function of n_i given that all other sample sizes are fixed, and moreover, $Q/\{(k-1)(\tilde{n}-1)\} \to \tau^2/\sigma^2_{\mathrm{pop}}$ when $k\to\infty$ and all $n_i\to\infty$. Consequently, we still have $I^2_{\mathrm{A}} \to \mathrm{ICC}_{\mathrm{MA}} < 1$ as that for the balanced case. In contrast, without the adjustment term $(k-1)(\tilde{n}-1)$ that is the same order of Q, the I^2 statistic will again, as observed in the literature, rapidly increase to 1 as the sample sizes are large.

d. Lastly, we can also express the I^2_{A} statistic in (14) as
$$I^2_{\mathrm{A}} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \tilde{n}\tilde{\sigma}_y^2} \tag{15}$$
where $\hat{\tau}^2$ is the DerSimonian-Laird estimator as mentioned in Section 1. When the assumption of a common population variance holds, $\tilde{n}\tilde{\sigma}_y^2$ is exactly σ^2_{pop} by (13). Otherwise, we can still apply $\tilde{n}\tilde{\sigma}_y^2$ to compute the I^2_{A} statistic by (15). And similarly as $\tilde{\sigma}_y^2$ explained in Böhning et al. [7], $\tilde{n}\tilde{\sigma}_y^2$ is asymptotically equivalent to the adjusted mean sample size multiplied by the harmonic mean of the within-study variances.

4.1 | Real Data Analysis

This section applies a real data example to illustrate the I^2_{A} statistic, and compare it with the I^2_{ANOVA} and I^2 statistics, for quantifying the heterogeneity between the studies. Specifically, we revisit a previous meta-analysis conducted by Jeong et al. [21], which investigated the stem cell-based therapy as a novel approach for the stroke treatment. Among various measures of efficacy and safety, we focus on the point difference in the National Institutes of Health Stroke Scale as the outcome. The summary data for a total of $k = 10$ studies are presented in Table 2.

Treating $\hat{\sigma}^2_{y_i}$ in Table 2 as the true values of $\sigma^2_{y_i}$, we have $\sum_{i=1}^{10} w_i = 7.68$ and $\sum_{i=1}^{10} w_i y_i = -43.39$. This leads to Cochran's Q statistic in (2) as $Q = 106.26$. Moreover, by formula (4),

$$I^2 = \max\left\{\frac{106.26 - (10-1)}{106.26}, 0\right\} = 0.92$$

In addition, the adjusted mean sample size can be computed as $\tilde{n} = 8.97$. Then by formula (14), it yields that

$$I^2_{\mathrm{A}} = \max\left\{\frac{106.26 - (10-1)}{106.26 + (10-1)(8.97-1)}, 0\right\} = 0.55$$

TABLE 2 | The summary data of the ten studies for the meta-analysis from Jeong et al. (2014), where y_i are the observed effect sizes and n_i are the study sample sizes.

Study	y_i	n_i	$\hat{\sigma}^2_{y_i}$
Wang (2013)	−3.10	8	1.81
Prasad (2012)	−6.30	11	3.16
Moniche (2012)	−9.40	10	0.53
Friedrich (2012)	−14.20	20	3.04
Honmou (2011)	−7.00	12	1.40
Savitz (2011)	−9.00	10	1.60
Battistella (2011)	−3.40	6	2.41
Suarez (2009)	−2.20	5	1.15
Savitz (2005)	−1.40	5	0.97
Bang (2005)	−2.00	5	1.06

Lastly, to also include the I^2_{ANOVA} statistic for additional comparison, we have $\bar{y} = \sum_{i=1}^{10} n_i y_i / \sum_{i=1}^{10} n_i = -7.55$, $\mathrm{MSB}_{\mathrm{MA}} = 189.83$, and $\mathrm{MSW}_{\mathrm{MA}} = 25.81$. Consequently, by formula (12),

$$I^2_{\mathrm{ANOVA}} = \max\left\{\frac{189.83 - 25.81}{189.83 + (8.97-1)\times 25.81}, 0\right\} = 0.41$$

To conclude, unlike the I^2 statistic that is very close to 1, the values of I^2_{A} and I^2_{ANOVA} are close to each other and they both indicate a moderate heterogeneity for the ten studies.

To further compare the I^2_{A} statistic and the I^2 statistic, as a common practice we assume that the ten studies are all normally distributed. Then by the reported means and variances, we plot their respective population distributions and the sampling distributions of the observed effect sizes in Figure 2 for visualization. From the figure, it is evident that the ten studies are not very heterogeneous since most of the study populations are largely overlapped in the range roughly from −15 to 5, corresponding to a measure of 0.55 for the I^2_{A} statistic. In contrast, the sampling distributions of the observed effect sizes are less overlapped with each other, indicating a much higher heterogeneity at 0.92 by the I^2 statistic.

4.2 | Numerical Results

To compare the numerical performance of the three statistics, we now conduct simulations based on the random-effects model (6) with $\mu = 0$ and $\sigma^2 = 100$. For the between-study variance, we consider $\tau^2 = 9$ or 90 that corresponds to $\mathrm{ICC}_{\mathrm{MA}}$ as $9/(9+100) = 0.083$ or $90/(90+100) = 0.474$, respectively. Let also $k = 3$ or 10 to represent the small or large number of studies included in the meta-analysis. For the sample size of each study, we consider the unbalanced design with the sample size of the ith study being $i * n$, where $i = 1, \ldots, k$ and the common n ranges from 10 to 90. With each of the above settings, we then generate the raw data from model (6) and report the summary data y_i and $\hat{\sigma}^2_{y_i}$ for the k studies. Finally with $M = 10,000$ repetitions, we present the boxplots of the I^2_{A}, I^2_{ANOVA} and I^2 statistics, together with their mean values, in Figure 3. From the figure, it is evident that the

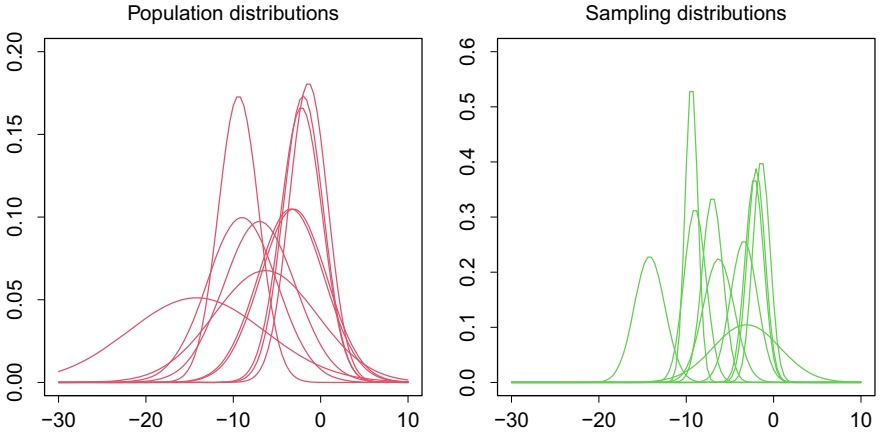

FIGURE 2 | Population distributions of the ten studies and the sampling distributions of the observed effect sizes from Jeong et al. (2014). For each study, the population distribution is assumed to be normal with mean y_i and variance $n_i\hat{\sigma}^2_{y_i}$. The sampling distribution of the effect size is assumed to be normal with mean y_i and variance $\hat{\sigma}^2_{y_i}$.

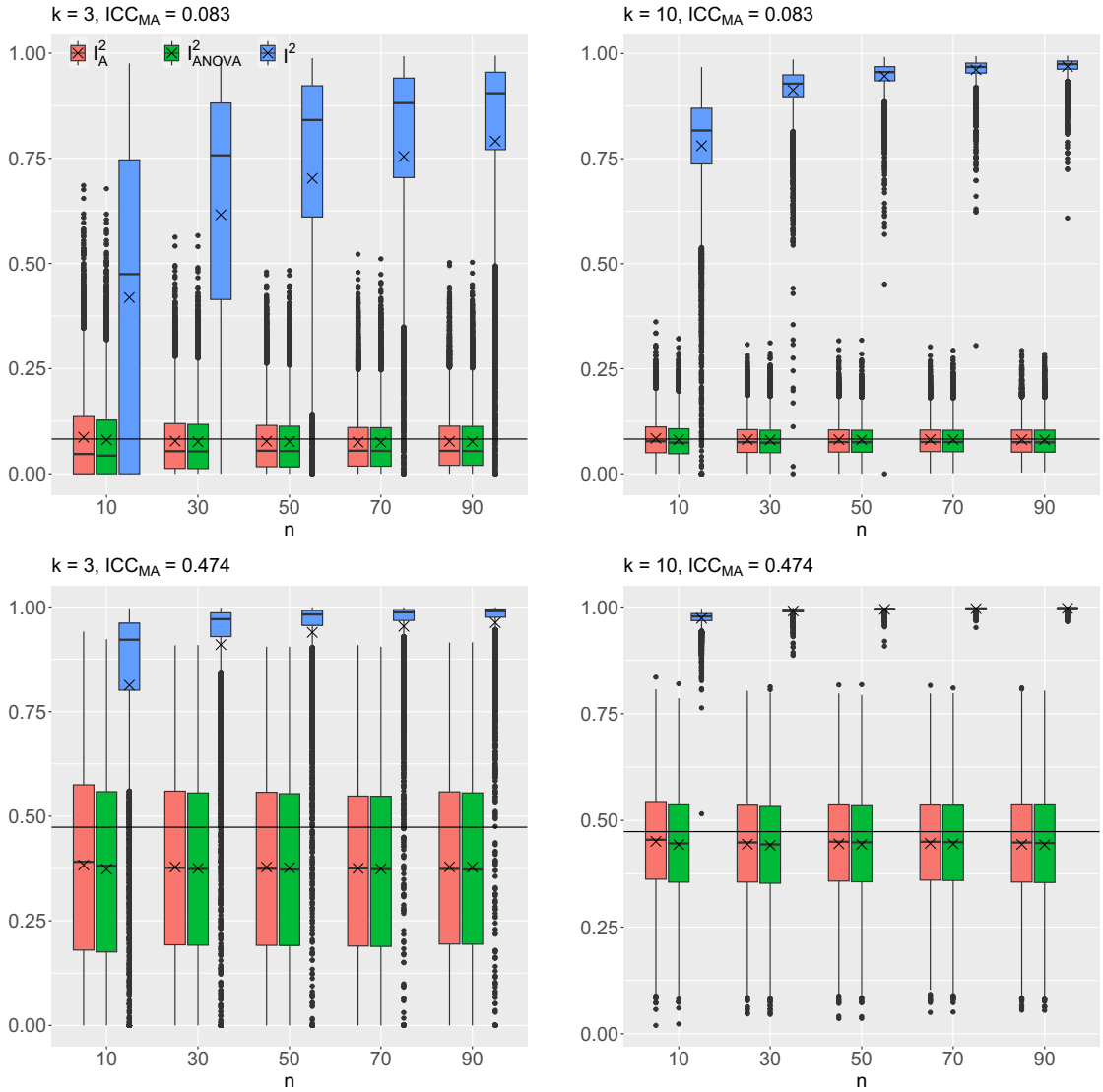

FIGURE 3 | Boxplots of the three statistics for the mean with 10,000 repetitions. The red boxes represent the I^2_{A} statistic, the green boxes represent the I^2_{ANOVA} statistic, and the blue boxes represent the I^2 statistic. The crosses on each box are the mean values of the 10000 repetitions. The solid lines stand for the absolute heterogeneity ICC_{MA}.

I^2 statistic has an increasing trend with the sample size n. This is consistent with what was observed in Rücker et al. [8] that the I^2 statistic always increases rapidly to 1 when the sample sizes are large. In contrast, with each solid line representing the heterogeneity ICC_{MA} between the study populations, we note that I^2_{A} and I^2_{ANOVA} are not influenced by the sample size and also provide comparable estimates for ICC_{MA} in terms of both bias and variance. And more interestingly, they are able to perform even better when the number of studies k is larger.

Our next simulation is to examine the scenario where the common population variance assumption does not hold. For the ith study, we generate the population variance in each replication from a gamma distribution with shape parameter 25 and scale parameter 4, or equivalently, with mean 100 and variance 400. All other settings remain the same as in the previous simulation. Note that for this case, the true value of ICC_{MA} will vary across replications because of the randomness in the population variance. Lastly, we present in Figure F1 the simulation results together with an ICC_{MA} value using $\sigma^2_{\text{pop}} = 100$, which, as can be seen, are similar to those for the common population variance.

5 | The I^2_{A} Statistic for the Mean Difference

In addition to the mean considered in Section 4, two other commonly used effect sizes for continuous outcomes are the mean difference (MD) and the standardized mean difference (SMD). Following this, we will describe the I^2_{A} statistic for MD in this section and then for SMD in Section 6.

For a meta-analysis of MD, each study has two treatment arms including the treatment group and the control group. The summary statistics for each study then consist of the observed MD y_i, the sample sizes n_i^T and n_i^C, and the standard errors $\hat{\sigma}_{y_i^T}$ and $\hat{\sigma}_{y_i^C}$ associated with the treatment and control groups. Given these summary statistics, the estimated variance of the

mean difference y_i is $\hat{\sigma}^2_{y_i} = \hat{\sigma}^2_{y_i^T} + \hat{\sigma}^2_{y_i^C}$, which is also treated as the true within-study variance of y_i as $\sigma^2_{y_i} = \sigma^2_{y_i^T} + \sigma^2_{y_i^C}$. In what follows, we describe the derivation procedure for the I^2_A statistic in the meta-analysis of MD, which extends from that for the meta-analysis of the mean. Following Section 4, we let $\sigma^2_{y_i} = (1/n_i^T + 1/n_i^C)\sigma^2_{\text{pop}}$, and moreover define the effective sample size for each study as $n_i = 1/(1/n_i^T + 1/n_i^C)$. This leads to the inverse-variance weights as $w_i = 1/\sigma^2_{y_i} = n_i/\sigma^2_{\text{pop}}$, and consequently, the I^2_A statistic for the meta-analysis of MD can again be expressed as

$$I^2_A = \max\left\{\frac{Q-(k-1)}{Q+(k-1)(\tilde{n}-1)}, 0\right\} \tag{16}$$

In other words, the newly derived I^2_A shares the same expression as in (14) but with a different definition of n_i. Moreover, the expression in (15) also applies to the I^2_A statistic for MD. Just as property (d) applies to the I^2_A statistic for the mean, the I^2_A statistic for MD is also applicable and interpretable when the population variances differ. For more details on the model assumptions for MD, refer to Appendix D.

5.1 | Real Data Analysis

To exemplify the I^2_A statistic for MD, we revisit a meta-analysis conducted in Avery et al. [22]. This study explores the effect of interventions to taper long term opioid treatment for chronic non-cancer pain. Among the several interventions, we consider the effect of acupuncture. For each study, the observed effect size is the mean difference of reduced opioid dose. For easy reference, we provide the summary data for the three studies in Table 3.

TABLE 3 | The summary data of the three studies for the meta-analysis from Avery et al. (2022).

Study	y_i^T	n_i^T	$\hat{\sigma}_{y_i^T}$	y_i^C	n_i^C	$\hat{\sigma}_{y_i^C}$
Jackson (2021)	−34	9	10.43	−66	6	12.78
Zheng (2019)	−13.6	48	3.23	−8.8	60	3.14
Zheng (2008)	−25.7	17	7.59	−10.9	18	2.80

By Table 3, the estimated effect sizes y_i for the three studies are $(32.0, -4.8, -14.8)$ and the within-study variances of y_i are $(272.14, 20.29, 65.48)$, yielding that $\sum_{i=1}^3 w_i = 0.07$ and $\sum_{i=1}^3 w_i y_i = 0.35$. Moreover, Cochran's Q statistic is given as $Q = 6.50$. Further by formula (4), we have

$$I^2 = \max\left\{\frac{6.50-(3-1)}{6.50}, 0\right\} = 0.69$$

To compute the I^2_A statistic, the effective sample sizes n_i for the three studies can be derived as 3.60, 26.67 and 8.74, respectively. This leads to the adjusted mean sample size as $\tilde{n} = 9.24$, and moreover by formula (14),

$$I^2_A = \max\left\{\frac{6.50-(3-1)}{6.50+(3-1)(9.24-1)}, 0\right\} = 0.20$$

Lastly, noting that $\bar{y} = -3.65$, $\text{MSB}_{\text{MA}} = 2848.76$ and $\text{MSW}_{\text{MA}} = 586.93$, we apply formula (12) and it yields that

$$I^2_{\text{ANOVA}} = \max\left\{\frac{2848.76-586.93}{2848.76+(9.24-1)\times 586.93}, 0\right\} = 0.29$$

To conclude, it is again evident that the values of I^2_A and I^2_{ANOVA} are close to each other, and both of them are significantly different from the value of I^2.

To further compare the I^2_A, I^2_{ANOVA} and I^2 statistics, we also plot the population distributions for the three studies and the sampling distributions of the observed effect sizes in Figure 4 for visualization. We note that two of the populations are largely overlapped with little heterogeneity, whereas the third population is moderately deviated. Given this, we conclude that the heterogeneity between the three studies may not be substantial overall, if measured by the I^2_A statistic. In contrast, the I^2 statistic concludes a very substantial heterogeneity between the sampling distributions of the observed effect sizes.

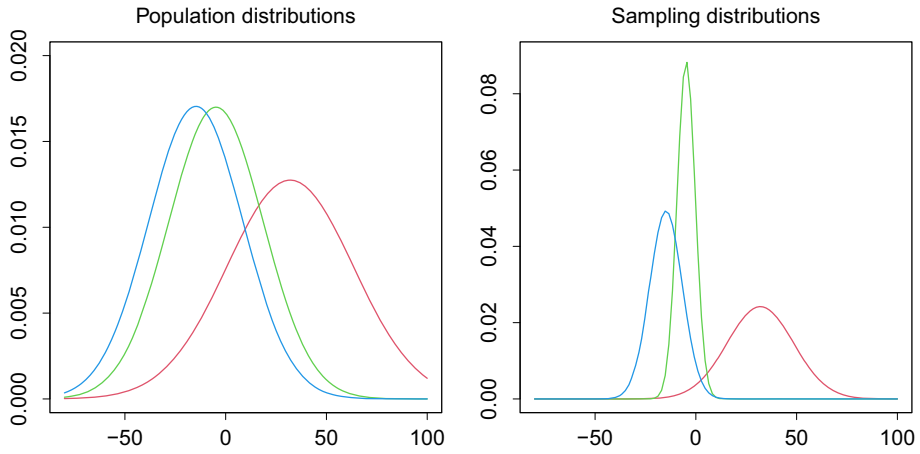

FIGURE 4 | Population distributions of the three studies and the sampling distributions of the observed effect sizes with blue for Zheng (2008), green for Zheng (2019), and red for Jackson (2021). For each study, the population distribution is assumed to be normal with mean $y_i^T - y_i^C$ and variance $\{n_i^T(n_i^T-1)\hat{\sigma}^2_{y_i^T} + n_i^C(n_i^C-1)\hat{\sigma}^2_{y_i^C}\}/(n_i^T+n_i^C-2)$. The sampling distribution of the effect size is assumed to be normal with mean $y_i^T - y_i^C$ and variance $\hat{\sigma}^2_{y_i^T} + \hat{\sigma}^2_{y_i^C}$.

5.2 | Numerical Results

To numerically compare the I^2_{A}, I^2_{ANOVA} and I^2 statistics, we generate the data from two-arm studies as follows:

$$\begin{aligned} y^T_{ij} &= \mu^T + \delta^T_i + \xi^T_{ij}, \quad j = 1, \ldots, n^T_i \\ y^C_{ij'} &= \mu^C + \delta^C_i + \xi^C_{ij'}, \quad j' = 1, \ldots, n^C_i \end{aligned} \tag{17}$$

where ξ^T_{ij} and $\xi^C_{ij'}$ are i.i.d. normal random errors with mean 0 and common variance σ^2. For a more detailed description of model (17), one may refer to Appendix D.

Without loss of generality, we set $\mu^T = \mu^C = 0$ and $\sigma^2 = 1$. We also generate δ^T_i and δ^C_i independently from $N(0, 0.045)$ or $N(0, 0.45)$. With the observed effect sizes being $\sum_{j=1}^{n^T} y^T_{ij}/n^T_i - \sum_{j'=1}^{n^C} y^C_{ij'}/n^C_i$, the between-study variance is $\tau^2 = 0.09$ or 0.9, yielding an $\mathrm{ICC}_{\mathrm{MA}}$ value of 0.083 or 0.474, respectively. For other settings, we consider $k = 3$ or 10 to represent a small or large number of studies within the meta-analysis, and the sample sizes of both treatment arms, n^T_i and n^C_i, to be identical. We further let the sample sizes for both arms of the ith study be $i * n$, where i ranges from 1 to k, and n varies from 10 to 90. Then for each simulation setting, we proceed to generate the raw data and compute the summary statistics, including y^T_i, y^C_i, $\hat{\sigma}^2_{y^T_i}$ and $\hat{\sigma}^2_{y^C_i}$, for each of the k studies. Finally with $M = 10,000$ repetitions, we present the boxplots of the I^2_{A}, I^2_{ANOVA} and I^2 statistics and also visualize their mean values in Figure 5. Based on the numerical results, it is clear again that the I^2 statistic monotonically increases with the sample size n, whereas the I^2_{A} and I^2_{ANOVA} statistics are not affected by the sample size. Moreover, the two new statistics also yield similar estimates for $\mathrm{ICC}_{\mathrm{MA}}$ in most settings, as well as provide a better performance when the number of studies k increases.

6 | The I^2_{A} Statistic for the Standardized Mean Difference

In addition to the mean difference (MD), another commonly used effect size for continuous outcomes in two-arm studies is the

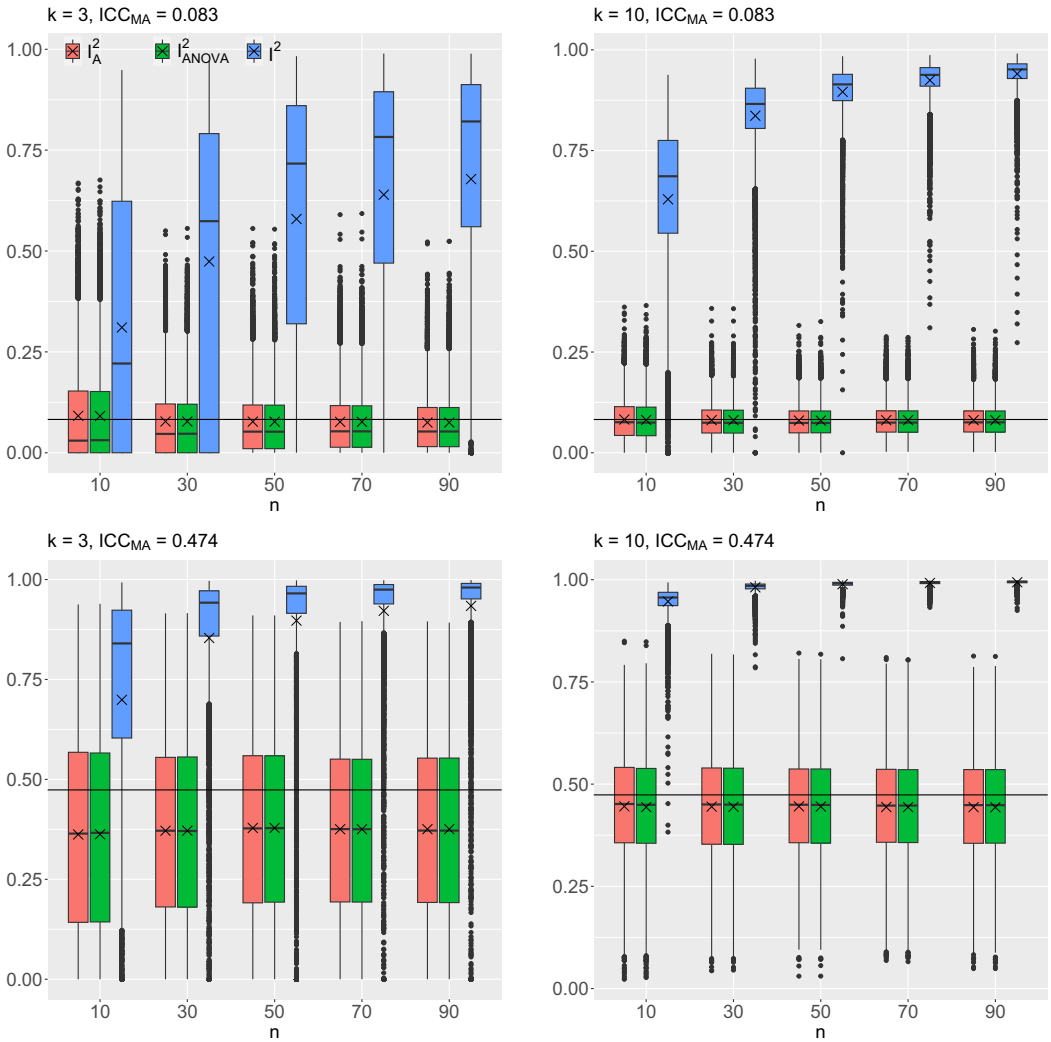

FIGURE 5 | Boxplots of the three statistics for the mean difference with 10,000 repetitions. The red boxes represent the I^2_{A} statistic, the green boxes represent the I^2_{ANOVA} statistic, and the blue boxes represent the I^2 statistic. The crosses on each box are the mean values of the 10000 repetitions. The solid lines stand for the absolute heterogeneity $\mathrm{ICC}_{\mathrm{MA}}$.

standardized mean difference (SMD). SMD is particularly useful when the assumption of equal population variances across different studies cannot be made. In such cases, the mean difference in each study is standardized to a uniform scale, ensuring comparability for the subsequent meta-analysis. Consequently, the estimated standardized mean difference y_i can be viewed as the observed mean difference of two population arms, both with a variance of 1, indicating $\sigma^2_{\text{pop}} = 1$. For a comprehensive understanding of the model specifications, one may refer to Appendix E. Lastly, to estimate ICC_{MA} for SMD, we use the DerSimonian-Laird estimator to estimate τ^2 and set σ^2_{pop} to 1 in formula (8), yielding the I^2_{A} statistic as

$$I^2_{\text{A}} = \max\left\{\frac{Q-(k-1)}{Q+(k-1)(\tilde{w}-1)}, 0\right\} \tag{18}$$

where $\tilde{w} = (\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i)/(k-1)$ with w_i being the inverse-variance weights. Similar to $\tilde{n}$ in (11), $\tilde{w}$ may also be referred to as the adjusted mean inverse-variance weight.

6.1 | Real Data Analysis

To assess the performance of the I^2_{A} statistic in quantifying the heterogeneity for SMD, we revisit the real data example presented in Section 5.1. With the summary data provided in Table 3, we first compute the estimated SMD and its corresponding variance for each study. Two commonly used statistics for estimating SMD are Cohen's d [23] and Hedges' g [24]. For a detailed guide on computing Cohen's d and Hedges' g, one may refer to Lin and Aloe [25]. In this section, we use Hedges' g that derives an unbiased estimate for SMD.

Following the formulas provided by Lin and Aloe [25], we can derive the estimated SMDs for the three studies as $(0.96, -0.20, -0.62)$, and the within-study variances of y_i as $(0.31, 0.04, 0.12)$. Further by $\sum_{i=1}^3 w_i = 38.12$ and $\sum_{i=1}^3 w_i y_i = -7.43$, Cochran's Q statistic can be computed as $Q = 5.83$. Thus by formula (4),

$$I^2 = \max\left\{\frac{5.83-(3-1)}{5.83}, 0\right\} = 0.66$$

Noting also that $\sum_{i=1}^3 w_i^2 = 784.06$ and $\tilde{w} = 8.78$, by formula (18) we have

$$I^2_{\text{A}} = \max\left\{\frac{5.83-(3-1)}{5.83+(3-1)(8.78-1)}, 0\right\} = 0.18$$

Lastly, to compute the I^2_{ANOVA} statistic, we first derive the effective sample sizes n_i for the three studies as 3.60, 26.67, and 8.74, respectively, Moreover, we have $\bar{y} = -0.19$, $\text{MSB}_{\text{MA}} = 3.19$, and $\text{MSW}_{\text{MA}} = 1$. Then by formula (12),

$$I^2_{\text{ANOVA}} = \max\left\{\frac{3.19-1}{3.19+(9.24-1)\times 1}, 0\right\} = 0.19$$

To further compare the three statistics, we plot the scaled population distributions for the three studies and the sampling distributions of the observed effect sizes in Figure 6. Specifically, with SMDs as the effect sizes, all the scaled populations have a common variance of 1. Moreover, we apply the estimated SMDs as the population means. Compared to Figure 4, the three scaled populations in Figure 6 get more close to each other, resulting in smaller values for the I^2_{A} and I^2_{ANOVA} statistics. On the other hand, a measure of 0.66 for the I^2 statistic indicates a large heterogeneity between the observed effect sizes.

6.2 | Numerical Results

To compare the I^2_{A}, I^2_{ANOVA} and I^2 statistics for SMD, we generate the data from the following two-arm studies:

$$\begin{aligned} y^T_{ij} &= \sigma_i(\mu^T + \delta^T_i + \xi^T_{ij}), \quad j = 1, \ldots, n^T_i \\ y^C_{ij'} &= \sigma_i(\mu^C + \delta^C_i + \xi^C_{ij'}), \quad j' = 1, \ldots, n^C_i \end{aligned} \tag{19}$$

where ξ^T_{ij} and $\xi^C_{ij'}$ are i.i.d. normal random errors with mean 0 and variance 1. Compared with model (17), this new model contains an additional parameter σ_i, which is used to rescale each

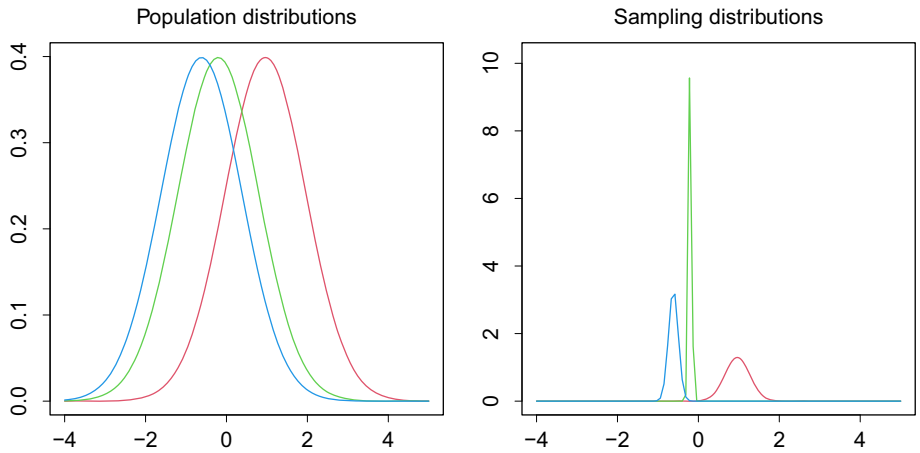

FIGURE 6 | Population distributions of the three scaled studies and the sampling distributions of the observed effect sizes with blue for Zheng (2008), green for Zheng (2019), and red for Jackson (2021). For each study, the population distribution is assumed to be normal with mean SMD and variance 1. The sampling distribution of the effect size is assumed to be normal with mean SMD and the variance is assumed to be the within-study variance.

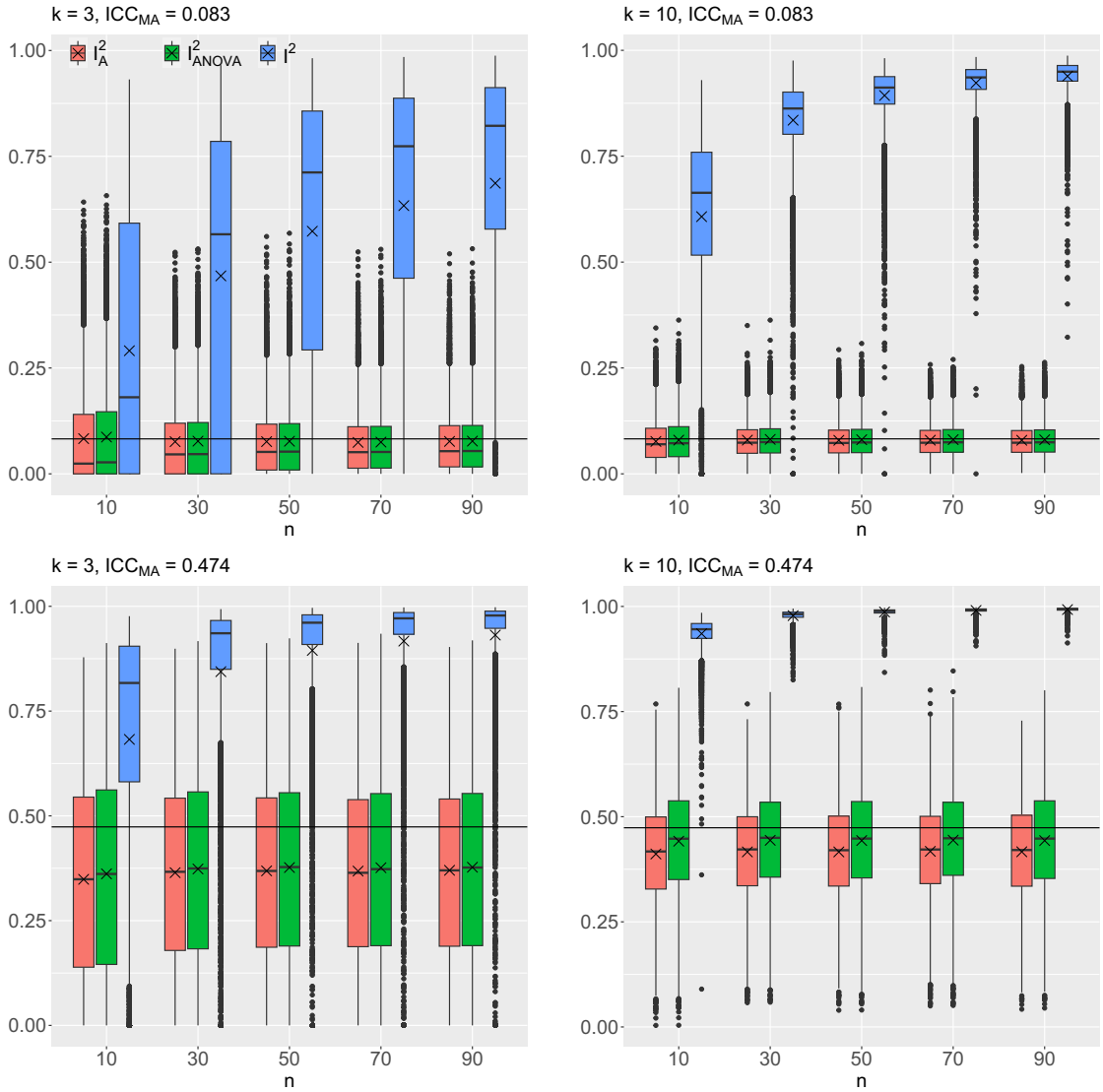

FIGURE 7 | Boxplots of the three statistics for the standardized mean difference with 10,000 repetitions. The red boxes represent the I^2_{A} statistic, the green boxes represent the I^2_{ANOVA} statistic, and the blue boxes represent the I^2 statistic. The crosses on each box are the mean values of the 10000 repetitions. The solid lines stand for the absolute heterogeneity $\mathrm{ICC}_{\mathrm{MA}}$.

study. For a more detailed description of model (19), refer to Appendix E.

In this simulation, we let σ_i follow a uniform distribution Unif(0.5, 1.5), which yields unequal population variances for the k studies and thus SMD ought to be applied rather than MD. The other settings are kept the same as those in Section 6.2. Then for each simulation setting, we proceed to generate the raw data and compute the summary statistics, including y_i^T, y_i^C, $\hat{\sigma}^2_{y_i^T}$ and $\hat{\sigma}^2_{y_i^C}$, for each of the k studies. Finally with $M = 10,000$ repetitions, we present the boxplots and the mean values of the I^2_{A}, I^2_{ANOVA} and I^2 statistics in Figure 7.

From Figure 7, it is evident that the I^2 statistic is always monotonically increasing with the sample size n, which is consistent with the simulation results in Sections 4.2 and 5.2. In contrast, the I^2_{A} and I^2_{ANOVA} statistics can always provide a good measure for the quantify of heterogeneity between the study populations, no matter whether the study sample sizes are large or not. For SMD, I^2_{ANOVA} provides a more accurate estimate for $\mathrm{ICC}_{\mathrm{MA}}$ with large heterogeneity compared with I^2_{A}.

7 | Conclusion and Discussion

Quantifying the heterogeneity is an important issue in meta-analysis for decision making. The presence of heterogeneity affects the extent to which generalizable conclusions can be formed and determines whether the random-effects model or the fixed-effect model should be used. The Q statistic is commonly used to test for the existence of the heterogeneity. However, as mentioned in the Cochrane Handbook for Systematic Reviews of Interventions [2], this test may have low power when the number of studies is small. Some also argue that the heterogeneity always exists, whether detectable by statistical

tests or not. Thus, as a way to remedy, the I^2 statistic was further introduced to measure the extent of heterogeneity as

$$I^2 = \max\left\{\frac{Q-(k-1)}{Q}, 0\right\}$$

Nowadays, both the Q statistic and the I^2 statistic are routinely reported in the forest plot in meta-analysis, and the choice between the random-effects model and the fixed-effect model often relies on these two statistics. More specifically, if the p-value of the Q statistic is less than 0.1 and the I^2 statistic exceeds 0.5, the random-effects model is preferred for meta-analysis; otherwise, the fixed-effect model will be chosen [26–28]. It is noted, however, that these two statistics are highly correlated since the I^2 statistic is a monotonically increasing function of the Q statistic. Additionally, the p-value based on the Q statistic only indicates whether there is a statistical significance [29], but not reflect regarding the biological difference between the studies. Even if heterogeneity is not statistically detected, it may still be clinically meaningful. Therefore, a random-effects model is often more appropriate, and the inclusion of prediction intervals is recommended in the practice of meta-analysis.

In this article, we have introduced a new measure, denoted as $\mathrm{ICC}_{\mathrm{MA}}$, to quantify the between-study heterogeneity for meta-analysis. To explore the distinction between $\mathrm{ICC}_{\mathrm{HT}}$ and $\mathrm{ICC}_{\mathrm{MA}}$, we have also drawn an interesting connection between ANOVA and meta-analysis, and learned that the essence of $\mathrm{ICC}_{\mathrm{HT}}$ is to quantify the heterogeneity between the observed effect sizes. As demonstrated by the motivating example in Section 2, the sampling distributions of the observed effect sizes may exhibit a significant dependency on the sample sizes, and they will asymptotically converge to their true effect sizes. Accordingly, with large sample sizes, the observed effect sizes will also yield an increased $\mathrm{ICC}_{\mathrm{HT}}$ close to one, no matter whether the underlying heterogeneity between the study populations is truly large or not. As an important alternative, our newly defined $\mathrm{ICC}_{\mathrm{MA}}$ is proposed to directly quantify the heterogeneity between the study populations. More specifically, we have systematically studied the statistical properties of $\mathrm{ICC}_{\mathrm{MA}}$, including the monotonicity, the location and scale invariance, the study size invariance, and the sample size invariance. It is the sample size invariance that distinguishes our new absolute measure of heterogeneity from $\mathrm{ICC}_{\mathrm{HT}}$.

Moreover, we have also proposed two new statistics to serve as the estimates of $\mathrm{ICC}_{\mathrm{MA}}$, where the I^2_{ANOVA} statistic in (12) is derived based on ANOVA, and the I^2_{A} statistic in (14) is derived based on the Q statistic as

$$I^2_{\mathrm{A}} = \max\left\{\frac{Q-(k-1)}{Q+(k-1)(\tilde{n}-1)}, 0\right\}$$

where $\tilde{n}$ is the adjusted mean sample size for the k studies. In addition, the I^2_{A} statistic can also be expressed as $I^2_{\mathrm{A}} = \hat{\tau}^2/(\hat{\tau}^2 + \tilde{n}\tilde{\sigma}^2_y)$. When the assumption of a common population variance holds, $\tilde{n}\tilde{\sigma}^2_y$ equals the common population variance σ^2_{pop}; otherwise, it serves as a representative value for σ^2_{pop}. This demonstrates that our I^2_{A} statistic is also widely applicable to scenarios where the population variances differ. For practical use, the exact formulas for the I^2_{A} and I^2_{ANOVA} statistics are also derived under two other common scenarios with the mean difference or the standardized mean difference as the effect size. Simulations and real data analyses demonstrate that they both provide asymptotically unbiased estimators of the absolute heterogeneity between the study populations, and as expected, they also do not depend on the study sample sizes. For most cases, I^2_{A} and I^2_{ANOVA} show similar performance in estimating $\mathrm{ICC}_{\mathrm{MA}}$. However, for meta-analysis of the standardized mean difference with large heterogeneity, I^2_{ANOVA} offers a slightly better estimate of $\mathrm{ICC}_{\mathrm{MA}}$ than I^2_{A}. Given that the Q statistic is commonly reported in meta-analysis and that I^2_{A} can be conveniently calculated from the Q statistic, we recommend using I^2_{A} in practical applications. But, of course, in case a higher accuracy is desired, the more complex I^2_{ANOVA} should be used, particularly for meta-analysis of the standardized mean difference. To conclude, the I^2_{A} statistic can serve as a supplemental measure to monitor the situations where the study effect sizes are indeed similar with little biological difference. In such scenario, the fixed-effect model can be appropriate. Although if the sample sizes are very large, we note that the I^2 statistic may still rapidly increase to 1 showing a large heterogeneity and subsequently a random-effects model will continue to be adopted. In view of this, we are thus confident that the I^2_{A} statistic can add new value to meta-analysis, for example, being included in the forest plot as a supplement to the I^2 statistic.

In addition, as shown in Figures 3, 5, and 7, the I^2_{A} statistic tends to slightly underestimate $\mathrm{ICC}_{\mathrm{MA}}$ when k is small and $\mathrm{ICC}_{\mathrm{MA}}$ is large. This underestimation may be primarily due to the inaccurate estimation of τ^2. While the I^2_{A} statistic has the advantage of being directly expressed using the Q statistic, making it more convenient to use, it implicitly relies on the DerSimonian-Laird (DL) method for estimating τ^2. Although the DL estimator is most commonly used, it does have limitations, and numerous alternative methods have been proposed to further improve it, as summarized in Veroniki et al. [30]. More recently, Kulinskaya et al. [31] and Bakbergenuly et al. [32] further highlighted that the Q statistic may perform poorly in estimating τ^2, partly because it does not account for the uncertainty in the within-study variances. To conclude, future research is warranted to investigate the impact of the τ^2 estimate across various effect sizes on the estimation accuracy of $\mathrm{ICC}_{\mathrm{MA}}$, offering a broader range of options for estimating the measure of heterogeneity.

Lastly, it is worth noting that there are also several interesting directions for future research. First, the current work has presented its primary focus on meta-analysis with continuous outcomes. As a parallel work, it can be equally important for the I^2_{A} statistic to be further extended to meta-analysis with binary outcomes, which are also commonly encountered in clinical studies. However, extending I^2_{A} to meta-analysis with binary outcomes may not be straightforward. Unlike the continuous outcomes, the binary outcomes do not have a direct definition of the population variance for each study, making the task more complex. To illustrate this challenge, we now consider a study with two treatment arms and use the log odds ratio (lnOR) as the effect size. For the treatment group, let a_i represent the number of events, p_i^T the event rate, and n_i^T the total number of participants. In this case, the event number a_i follows the binomial distribution $\mathrm{Bin}(n_i^T, p_i^T)$. Similarly, let b_i, p_i^C and n_i^C be the number of events, the event rate, and the total number of participants in the control group, and moreover $b_i \sim \mathrm{Bin}(n_i^C, p_i^C)$. Then the observed effect size (lnOR) for the study will be calculated

as $y_i = \ln[\{a_i/(n_i^T - a_i)\}/\{b_i/(n_i^C - b_i)\}]$, with the within-study variance estimated by $\sigma^2_{y_i} = 1/a_i + 1/(n_i^T - a_i) + 1/b_i + 1/(n_i^C - b_i)$. For more details, one may refer to Higgins et al. [2]. Given this setup, it may not be straightforward to define a common study population variance σ^2_{pop} based on the within-study variance $\sigma^2_{y_i}$ and the sample sizes n_i^T and n_i^C and so requires further investigation.

Acknowledgments

The authors sincerely thank the Editor, the Associate Editor, and the two reviewers for their constructive comments that have led to a substantial improvement of this article. Liping Zhu and Ke Yang's research was supported in part by the National Key R&D Program of China (2022YFA1003702) and the National Natural Science Foundation of China (12371294). Wangli Xu's research was supported in part by the MOE Project of Key Research Institute of Humanities and Social Sciences (No. 22JJD910001). Tiejun Tong's research was supported in part by the General Research Fund of Hong Kong (HKBU12303421 and HKBU12300123), the Initiation Grant for Faculty Niche Research Areas (RC-FNRA-IG/23-24/SCI/03) of Hong Kong Baptist University, and the National Natural Science Foundation of China (12071305).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

References

1. M. Egger and G. D. Smith, "Meta-Analysis: Potentials and Promise," *British Medical Journal* 315, no. 7119 (1997): 1371–1374.

2. J. P. Higgins, J. Thomas, J. Chandler, et al., *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed. (John Wiley & Sons, 2019).

3. W. G. Cochran, "The Combination of Estimates From Different Experiments," *Biometrics* 10, no. 1 (1954): 101–129, https://doi.org/10.2307/3001666.

4. R. DerSimonian and N. Laird, "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials* 7, no. 3 (1986): 177–188.

5. J. P. Higgins and S. G. Thompson, "Quantifying Heterogeneity in a Meta-Analysis," *Statistics in Medicine* 21, no. 11 (2002): 1539–1558.

6. J. P. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring Inconsistency in Meta-Analyses," *British Medical Journal* 327, no. 7414 (2003): 557–560.

7. D. Böhning, R. Lerdsuwansri, and H. Holling, "Some General Points on the $< i > I < /i>^{<i>2</i>}$-Measure of Heterogeneity in Meta-Analysis," *Metrika* 80, no. 6 (2017): 685–695.

8. G. Rücker, G. Schwarzer, J. R. Carpenter, and M. Schumacher, "Undue Reliance on $< i > I < /i>^{<i>2</i>}$ in Assessing Heterogeneity May Mislead," *BMC Medical Research Methodology* 8, no. 1 (2008): 79.

9. R. D. Riley, J. Ensor, K. I. Snell, et al., "External Validation of Clinical Prediction Models Using Big Datasets From e-Health Records or IPD Meta-Analysis: Opportunities and Challenges," *British Medical Journal* 353, no. 8063 (2016): i3140, https://doi.org/10.1136/bmj.i3140.

10. J. IntHout, J. P. Ioannidis, M. M. Rovers, and J. J. Goeman, "Plea for Routinely Presenting Prediction Intervals in Meta-Analysis," *BMJ Open* 6, no. 7 (2016): e010247.

11. M. Borenstein, J. P. Higgins, L. V. Hedges, and H. R. Rothstein, "Basics of Meta-Analysis: $< i > I < /i>^{<i>2</i>}$ Is Not an Absolute Measure of Heterogeneity," *Research Synthesis Methods* 8, no. 1 (2017): 5–18.

12. P. Sangnawakij, D. Böhning, S. A. Niwitpong, S. Adams, M. Stanton, and H. Holling, "Meta-Analysis Without Study-Specific Variance Information: Heterogeneity Case," *Statistical Methods in Medical Research* 28, no. 1 (2019): 196–210.

13. H. Holling, W. Böhning, E. Masoudi, D. Böhning, and P. Sangnawakij, "Evaluation of a New Version of $< i > I < /i>^{<i>2</i>}$ With Emphasis on Diagnostic Problems," *Communications in Statistics: Simulation and Computation* 49, no. 4 (2020): 942–972.

14. EFSA Scientific Committee, "Statistical Significance and Biological Relevance," *EFSA Journal* 9, no. 9 (2011): 2372.

15. R. A. Fisher, *Statistical Methods for Research Workers* (Oliver & Boyd, 1925).

16. C. A. B. Smith, "On the Estimation of Intraclass Correlation," *Annals of Human Genetics* 21, no. 4 (1957): 363–373.

17. A. Donner, "The Use of Correlation and Regression in the Analysis of Family Resemblance," *American Journal of Epidemiology* 110, no. 3 (1979): 335–342.

18. K. O. McGraw and S. P. Wong, "Forming Inferences About Some Intraclass Correlation Coefficients," *Psychological Methods* 1, no. 1 (1996): 30–46.

19. W. G. Cochran, "The Use of the Analysis of Variance in Enumeration by Sampling," *Journal of the American Statistical Association* 34, no. 207 (1939): 492–510.

20. J. D. Thomas and R. A. Hultquist, "Interval Estimation for the Unbalanced Case of the One-Way Random Effects Model," *Annals of Statistics* 6, no. 3 (1978): 582–587.

21. H. Jeong, H. W. Yim, Y. S. Cho, et al., "Efficacy and Safety of Stem Cell Therapies for Patients With Stroke: A Systematic Review and Single Arm Meta-Analysis," *International Journal of Stem Cells* 7, no. 2 (2014): 63–69.

22. N. Avery, A. G. McNeilage, F. Stanaway, et al., "Efficacy of Interventions to Reduce Long Term Opioid Treatment for Chronic Non-Cancer Pain: Systematic Review and Meta-Analysis," *British Medical Journal* 377 (2022): e066375, https://doi.org/10.1136/bmj-2021-066375.

23. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Routledge, 2013).

24. L. V. Hedges, "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators," *Journal of Educational Statistics* 6, no. 2 (1981): 107–128.

25. L. Lin and A. M. Aloe, "Evaluation of Various Estimators for Standardized Mean Difference in Meta-Analysis," *Statistics in Medicine* 40, no. 2 (2021): 403–426.

26. S. J. Jiang and C. H. Huang, "The Clinical Efficacy of N-Acetylcysteine in the Treatment of ST Segment Elevation Myocardial Infarction A Meta-Analysis and Systematic Review," *International Heart Journal* 62, no. 1 (2021): 142–147, https://doi.org/10.1536/ihj.20-519.

27. M. A. Chinnaratha, M. A. Chuang, R. J. Fraser, R. J. Woodman, and A. J. Wigg, "Percutaneous Thermal Ablation for Primary Hepatocellular Carcinoma: A Systematic Review and Meta-Analysis," *Journal of Gastroenterology and Hepatology* 31, no. 2 (2016): 294–301, https://doi.org/10.1111/jgh.13028.

28. J. Yang, H. P. Wang, L. Zhou, and C. F. Xu, "Effect of Dietary Fiber on Constipation: A Meta Analysis," *World Journal of Gastroenterology* 18, no. 48 (2012): 7378–7383.

29. A. Gelman and H. Stern, "The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant," *American Statistician* 60, no. 4 (2006): 328–331.

30. A. A. Veroniki, D. Jackson, W. Viechtbauer, et al., "Methods to Estimate the Between-Study Variance and Its Uncertainty in Meta-Analysis," *Research Synthesis Methods* 7, no. 1 (2016): 55–79.

31. E. Kulinskaya, D. C. Hoaglin, I. Bakbergenuly, and J. Newman, "A Q Statistic With Constant Weights for Assessing Heterogeneity in Meta-Analysis," *Research Synthesis Methods* 12, no. 6 (2021): 711–730.

32. I. Bakbergenuly, D. C. Hoaglin, and E. Kulinskaya, "On the Statistic With Constant Weights for Standardized Mean Difference," *British Journal of Mathematical and Statistical Psychology* 75, no. 3 (2022): 444–465.

33. S. R. Searle, *Linear Models* (Wiley, 1971).

34. A. Wald, "A Note on the Analysis of Variance With Unequal Class Frequencies," *Annals of Mathematical Statistics* 11, no. 1 (1940): 96–100.

35. S. Karlin, E. C. Cameron, and P. T. Williams, "Sibling and Parent–Offspring Correlation Estimation With Variable Family Size," *Proceedings of the National Academy of Sciences of the United States of America* 78, no. 5 (1981): 2664–2668.

36. A. Donner and J. J. Koval, "The Estimation of Intraclass Correlation in the Analysis of Family Data," *Biometrics* 36, no. 1 (1980): 19–25.

37. A. Donner and J. J. Koval, "The Large Sample Variance of an Intraclass Correlation," *Biometrika* 67, no. 3 (1980): 719–722.

38. A. Donner, "A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model," *International Statistical Review* 54, no. 1 (1986): 67–82.

39. H. Sahai and M. M. Ojeda, *Analysis of Variance for Random Models: Theory, Methods, Applications, and Data Analysis*, vol. 2 (Birkhäuser, 2004).

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Supplementary Data S1**: Supporting Information.

Appendix A

Proof of the Properties of $\mathrm{ICC}_{\mathrm{MA}}$

Proof of "Monotonicity". By the definition in (8), we can rewrite $\mathrm{ICC}_{\mathrm{MA}}$ as

$$\mathrm{ICC}_{\mathrm{MA}} = \frac{1}{1 + \sigma^2_{\mathrm{pop}}/\tau^2}$$

This shows that $\mathrm{ICC}_{\mathrm{MA}}$ is a monotonically increasing function of $\tau^2/\sigma^2_{\mathrm{pop}}$ and so property (i′) holds. □

Proof of "Location and Scale Invariance". To prove the location and scale invariance, for any constants a and $b > 0$, we assume that the newly observed effect sizes are $y'_{ij} = a + by_{ij}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. Let also $\mu'_i = a + b\mu_i$ be the true effect sizes of the new study populations. Then consequently, the between-study variance and the common population variance are given as

$$(\tau^2)' = \mathrm{var}(\mu'_i) = \mathrm{var}(a + b\mu_i) = b^2\tau^2,$$
$$(\sigma^2_{\mathrm{pop}})' = \mathrm{var}(a + by_{ij} | a + b\mu_i) = b^2\sigma^2_{\mathrm{pop}}$$

Further by (8), the measure of heterogeneity between the new studies is

$$\mathrm{ICC}'_{\mathrm{MA}} = \frac{(\tau^2)'}{(\tau^2)' + (\sigma^2_{\mathrm{pop}})'} = \frac{b^2\tau^2}{b^2\tau^2 + b^2\sigma^2_{\mathrm{pop}}} = \frac{\tau^2}{\tau^2 + \sigma^2_{\mathrm{pop}}} = \mathrm{ICC}_{\mathrm{MA}}$$

This verifies the property of location and scale invariance. □

Proof of "Study Size Invariance". To prove the study size invariance, we assume there are a total of k' studies. Then by the random-effects model in (1), since the individual means μ_i are i.i.d. from $N(\mu, \tau^2)$, the between-study variance will remain unchanged as τ^2 regardless of the number of studies. Further by the common population variance assumption, we have $\mathrm{var}(y_{ij}|\mu_i) = \sigma^2_{\mathrm{pop}}$ for all $i = 1, \ldots, k'$ and $j = 1, \ldots, n_i$. This proves the property of study size invariance. □

Proof of "Sample Size Invariance". To prove the sample size invariance, we assume that the new sample sizes are n'_i for each study, and consequently $y'_i = \sum_{j=1}^{n'_i} y_{ij}/n'_i$ are the new effect sizes. Then under the common population variance assumption that $\mathrm{var}(y_{ij}|\mu_i) = \sigma^2_{\mathrm{pop}}$ for all i and j, we have $\sigma^2_{y'_i} = \mathrm{var}(y'_i|\mu_i) = \sigma^2_{\mathrm{pop}}/n'_i$, or equivalently, $n'_i\sigma^2_{y'_i} = \sigma^2_{\mathrm{pop}}$. That is, no matter how the sample sizes vary, the common population variance will always remain unchanged. Finally, noting that τ^2 also remains since the study populations are unaltered, we thus have the property of sample size invariance. □

Appendix B

Methods for Estimating ICC

To estimate ICC from the random-effects ANOVA in (5), we first partition the total variation of the observations into two components as

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} n_i(y_i - \bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - y_i)^2 \tag{B1}$$

where $y_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ are the individual sample means, and $\bar{y} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}/\sum_{i=1}^{k} n_i$ is the grand sample mean. More specifically, the term on the left-hand side of (B1) is the total sum of squares (SST), and the two terms on the right-hand side are the sum of squares between the populations (SSB) and the error sum of squares within the populations (SSW), respectively.

By equating SSB and SSW to their respective expected values, Cochran [19] derived the method of moments estimators of τ^2 and σ^2. Further by plugging these two estimators in formula (7), it yields the ANOVA estimator for the unknown ICC. By Smith [16], the ANOVA estimator is a biased but consistent estimator. Moreover, as the method of moments estimators may take a negative value when $\mathrm{SSB}/k < \mathrm{SSW}/(\sum_{i=1}^{k}(n_i - 1)$, one often truncates the negative value to 0 when it occurs. For the balanced case when the sample sizes are all equal, Searle [33] derived an exact confidence interval for ICC based on the ANOVA table. For the unbalanced case, however, the exact confidence interval from the ANOVA table is not available. As a remedy, Thomas and Hultquist [20] and Donner [17] suggested an adjusted confidence interval in which the common sample size in the balanced case is replaced by the average sample size. They further showed by simulation studies that the adjusted confidence interval performs very well in terms of the coverage probability.

Besides the well-known ANOVA estimator, it is noteworthy that there are also other estimators for ICC in the literature. To name a few, Thomas and Hultquist [20] constructed a confidence interval for ICC based on the unweighted average of the individual sample means $\tilde{y} = \sum_{i=1}^{k} y_i/k$. Observing that $\mathrm{ICC} = (\tau^2/\sigma^2)/(\tau^2/\sigma^2 + 1)$, Wald [34] proposed another estimator for ICC by first estimating τ^2/σ^2, yet as a limitation, there does not exist a closed form for either the point estimator or its confidence interval. As another alternative, by the facts that $\mathrm{cov}(y_{ij}, y_{il}) = \tau^2$ for $j \neq l$ and $\mathrm{var}(y_{ij}) = \tau^2 + \sigma^2$, Karlin et al. [35] proposed to estimate ICC by the Pearson product-moment correlation computed over all the possible pairs of (y_{ij}, y_{il}) for $j \neq l$ with some weighting schemes. In addition, Donner and Koval [36, 37] proposed an iterative algorithm to compute the maximum likelihood estimator (MLE) for ICC directly, and presented its performance by simulations when the number of studies is large. For more estimators of ICC, one may also refer to Donner [38], Sahai and Ojeda [39], and the references therein.

Despite the rich literature on the estimation of ICC, none of the existing estimators is known to be uniformly better than the others in the unbalanced case [39]. In practice, thanks to its simple and elegant form, the ANOVA estimator is frequently treated as the optimal estimator and so is most commonly used for estimating ICC. Lastly, we also note that the ANOVA estimator and the confidence interval suggested by Thomas and

Hultquist [20] and Donner [17] can be readily implemented by the function *ICCest* in the R package 'ICC'.

Appendix C

The Derivation of the I^2_{ANOVA} Statistic

To estimate ICC_{MA}, we begin by presenting the following lemma along with its proof.

Lemma 1. *With model (5) and the summary data $y_i, \hat{\sigma}^2_{y_i}$ for $i = 1, \ldots, k$ in meta-analysis, we have $E(\text{MSB}_{\text{MA}}) = \tilde{n}\tau^2 + \sigma^2_{\text{pop}}$ and $E(\text{MSW}_{\text{MA}}) = \sigma^2_{\text{pop}}$.*

Proof of Lemma 1. Denote by $\sigma^2_{y_i} = \sigma^2/n_i$. With the summary data, y_i are independent normal random variables with mean μ and variances $\tau^2 + \sigma^2_{y_i}$. Then the variance of $\sum_{i=1}^k n_i y_i$ is

$$\text{Var}\left(\sum_{i=1}^k n_i y_i\right) = \sum_{i=1}^k \text{Var}(n_i y_i) = \tau^2 \sum_{i=1}^k n_i^2 + \sum_{i=1}^k n_i^2 \sigma^2_{y_i}$$

Thus,

$$\begin{aligned} E\left(\sum_{i=1}^k n_i y_i\right)^2 &= \text{Var}\left(\sum_{i=1}^k n_i y_i\right) + \left\{E\left(\sum_{i=1}^k n_i y_i\right)\right\}^2 \\ &= \tau^2 \sum_{i=1}^k n_i^2 + \sum_{i=1}^k n_i^2 \sigma^2_{y_i} + \mu^2 \left(\sum_{i=1}^k n_i\right)^2 \end{aligned}$$

Furthermore, it can be derived that

$$\begin{aligned} E\left\{\sum_{i=1}^k n_i (y_i - \bar{y})^2\right\} &= \sum_{i=1}^k n_i E(y_i^2) - \frac{1}{\sum_{i=1}^k n_i} E\left(\sum_{i=1}^k n_i y_i\right)^2 \\ &= \sum_{i=1}^k n_i \left[\text{Var}(y_i) + \{E(y_i)\}^2\right] - \frac{1}{\sum_{i=1}^k n_i} E\left(\sum_{i=1}^k n_i y_i\right)^2 \\ &= \sum_{i=1}^k n_i \left(\tau^2 + \sigma^2_{y_i} + \mu^2\right) - \frac{1}{\sum_{i=1}^k n_i} \\ &\quad \left\{\tau^2 \sum_{i=1}^k n_i^2 + \sum_{i=1}^k n_i^2 \sigma^2_{y_i} + \mu^2 (\sum_{i=1}^k n_i)^2\right\} \\ &= \tau^2 (\sum_{i=1}^k n_i - \frac{\sum_{i=1}^k n_i^2}{\sum_{i=1}^k n_i}) + \sum_{i=1}^k n_i \sigma^2_{y_i} - \frac{\sum_{i=1}^k n_i^2 \sigma^2_{y_i}}{\sum_{i=1}^k n_i} \end{aligned}$$

Since $\sigma^2_{y_i} = \sigma^2_{\text{pop}}/n_i$, and $\tilde{n} = (\sum_{i=1}^k n_i - \sum_{i=1}^k n_i^2 / \sum_{i=1}^k n_i)/(k-1)$,

$$E\left\{\sum_{i=1}^k n_i (y_i - \bar{y})^2\right\} = (k-1)\tilde{n}\tau^2 + (k-1)\sigma^2_{\text{pop}}$$

Thus, $E(\text{MSB}_{\text{MA}}) = \tilde{n}\tau^2 + \sigma^2_{\text{pop}}$.

As for $E(\text{MSW}_{\text{MA}}) = \sigma^2_{\text{pop}}$, it is derived directly by the fact that $E(n_i \hat{\sigma}^2_{y_i}) = \sigma^2_{\text{pop}}$.

With Lemma 1, $E(\text{MSB}_{\text{MA}} - \text{MSW}_{\text{MA}}) = \tilde{n}\tau^2$, and $E\{\text{MSB}_{\text{MA}} + (\tilde{n} - 1)\text{MSW}_{\text{MA}}\} = \tilde{n}(\tau^2 + \sigma^2_{\text{pop}})$. Thus, $\text{ICC}_{\text{MA}} = \tau^2/(\tau^2 + \sigma^2_{\text{pop}})$ can be estimated by $(\text{MSB}_{\text{MA}} - \text{MSW}_{\text{MA}})/\{\text{MSB}_{\text{MA}} + (\tilde{n} - 1)\text{MSW}_{\text{MA}}\}$. Truncating the negative value to zero, the I^2_{ANOVA} statistic in (12) can be derived.

Appendix D

The Statistical Model for the Mean Difference in Meta-Analysis

For meta-analysis of studies with two arms, we start with modeling the individual patient data in a single study. In analogy with model (5), we model the individual observations y^T_{ij} and $y^C_{ij'}$ of the treatment group and the control group for the ith study as

$$\begin{aligned} y^T_{ij} &= \mu^T + \delta^T_i + \xi^T_{ij}, \quad j = 1, \ldots, n^T_i, \\ y^C_{ij'} &= \mu^C + \delta^C_i + \xi^C_{ij'}, \quad j' = 1, \ldots, n^C_i \end{aligned}$$

where the superscript "T" represents the treatment group, and the superscript "C" represents the control group. Similar to the assumptions in model (5), we assume that δ^T_i, $\xi^T_{ij'}$, δ^C_i and $\xi^C_{ij'}$ are independent of each other. For the random errors of different observations in the same study, it is natural to assume they are i.i.d. normal random errors with mean 0 and share a common variance σ^2. Then the true effect size for each study is routinely presented by the mean difference

$$\text{MD}_i = (\mu^T + \delta^T_i) - (\mu^C + \delta^C_i)$$

For each study, the observed mean difference is

$$y^T_i - y^C_i = (\mu^T - \mu^C) + (\delta^T_i - \delta^C_i) + \left(\frac{\sum_{j=1}^{n^T_i} \xi_{ij}}{n^T_i} - \frac{\sum_{j'=1}^{n^C_i} \xi_{ij}}{n^C_i}\right) \tag{D1}$$

where $y^T_i = \sum_{j=1}^{n^T} \xi_{ij}/n^T_i$, and $y^C_i = \sum_{j=1}^{n^C} \xi_{ij}/n^C_i$. Furthermore, let $y_i = y^T_i - y^C_i$, $\mu = \mu^T - \mu^C$, $\delta_i = \delta^T_i - \delta^C_i$, and $\epsilon_i = \sum_{j=1}^{n^T_i} \xi_{ij}/n^T_i - \sum_{j'=1}^{n^C_i} \xi_{ij}/n^C_i$. Regardless of the dependence between δ^T_i and δ^C_i, we simply assume that δ_i are i.i.d. normal random variables with mean 0 and variance $\tau^2 \geq 0$, where τ^2 measures the magnitude of the heterogeneity between the studies. Then model (D1) reduces to

$$y_i = \mu + \delta_i + \epsilon_i \tag{D2}$$

where $\delta_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$ and $\epsilon_i \overset{\text{ind}}{\sim} N(0, (1/n^T_i + 1/n^C_i)\sigma^2)$. We note that model (D2) has the same form as in (6), except for the variance of ϵ_i.

To estimate ICC_{MA} for the mean difference based on ANOVA, we apply the results for the single-arm studies directly. Letting $n_i = 1/(1/n^T_i + 1/n^C_i)$, Lemma 1 in Appendix C also holds that

$$\begin{aligned} E(\text{MSB}_{\text{MA}}) &= \tilde{n}\tau^2 + \sigma^2, \\ E(\text{MSW}_{\text{MA}}) &= \sigma^2 \end{aligned}$$

Together with the notation of $\tilde{n}$, the I^2_{ANOVA} statistic in (12) can be derived.

The I^2_{A} statistic in (16) is derived similar to that for the mean. Furthermore, similar to the expression in formula (15), the I^2_{A} statistic can also be applied and well interpreted when the population variances differ across the studies.

Appendix E

The Statistical Model for the Standardized Mean Difference in Meta-Analysis

For the standardized mean difference, we model the individual observations y^T_{ij} and $y^C_{ij'}$ of the treatment group and the control group for the ith study as

$$\begin{aligned} y^T_{ij} &= \sigma_i(\mu^T + \delta^T_i + \xi^T_{ij}), \quad j = 1, \ldots, n^T_i, \\ y^C_{ij'} &= \sigma_i(\mu^C + \delta^C_i + \xi^C_{ij'}), \quad j' = 1, \ldots, n^C_i \end{aligned}$$

where the superscript "T" represents the treatment group, and the superscript "C" represents the control group. Similar to the assumptions in model (5), we assume that δ^T_i, $\xi^T_{ij'}$, δ^C_i and $\xi^C_{ij'}$ are independent of each other. In (ipdsmd), $\xi^T_{ij'}$ and $\xi^C_{ij'}$ are assumed to be i.i.d. normal random errors with mean 0 and variance 1. Then with different values of σ_i, the

population variances for different studies are σ_i^2, respectively. To eliminate the influence of the scale, SMDs are considered to represent the effect sizes, which is defined by

$$\begin{aligned}\text{SMD}_i &= \{(\sigma_i\mu^T + \sigma_i\delta_i^T) - (\sigma_i\mu^C + \sigma_i\delta_i^C)\}/\sigma_i \\ &= (\mu^T + \delta_i^T) - (\mu^C + \delta_i^C)\end{aligned}$$

For each study, SMD_i is estimated by

$$\frac{y_i^T - y_i^C}{\hat{\sigma}_i} = \frac{\sigma_i}{\hat{\sigma}_i}\left\{(\mu^T - \mu^C) + (\delta_i^T - \delta_i^C) + \left(\frac{\sum_{j=1}^{n_i^T}\xi_{ij}}{n_i^T} - \frac{\sum_{j'=1}^{n_i^C}\xi_{ij}}{n_i^C}\right)\right\} \quad \text{(E1)}$$

where $\hat{\sigma}_i$ is an estimate for σ_i, $y_i^T = \sum_{j=1}^{n^T}\xi_{ij}/n_i^T$, and $y_i^C = \sum_{j=1}^{n^C}\xi_{ij}/n_i^C$. For simplicity, we assume that σ_i can be accurately estimated and thus $\sigma_i/\hat{\sigma}_i = 1$. Furthermore, let $y_i = y_i^T - y_i^C$, $\mu = \mu^T - \mu^C$, $\delta_i = \delta_i^T - \delta_i^C$, and $\epsilon_i = \sum_{j=1}^{n_i^T}\xi_{ij}/n_i^T - \sum_{j'=1}^{n_i^C}\xi_{ij}/n_i^C$. Regardless of the dependence between δ_i^T and δ_i^C, we simply assume that δ_i are i.i.d. normal random variables with mean 0 and variance $\tau^2 \geq 0$, where τ^2 measures the magnitude of the heterogeneity between the studies. Then model (E1) reduces to

$$y_i = \mu + \delta_i + \epsilon_i \quad \text{(E2)}$$

where $\delta_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$ and $\epsilon_i \overset{\text{ind}}{\sim} N(0, 1/n_i^T + 1/n_i^C)$. We note that model (E2) has the same form as in (6), except for the variance of ϵ_i.

To estimate ICC_{MA} for the standardized mean difference based on ANOVA, we also apply the results for the single-arm studies directly. Letting $n_i = 1/(1/n_i^T + 1/n_i^C)$, Lemma 1 in Appendix C also holds that

$$E(\text{MSB}_{\text{MA}}) = \tilde{n}\tau^2 + 1$$

Together with the notation of $\tilde{n}$ and $\text{MSB}_{\text{MA}} = 1$, the I^2_{ANOVA} statistic in (12) can be derived.

Appendix F

Comparison Between the I^2 and I^2_{A} Statistics

Proof of (a). First, $(k-1)(\tilde{n}-1)$ in (14) can be expressed as

$$\begin{aligned}&\sum_{i=1}^k n_i - \frac{\sum_{i=1}^k n_i^2}{\sum_{i=1}^k n_i} - (k-1) \\ &= \frac{(\sum_{i=1}^k n_i)^2 - \sum_{i=1}^k n_i^2 - (k-1)\sum_{i=1}^k n_i}{\sum_{i=1}^k n_i} \\ &= \frac{\{\sum_{i=1}^k (n_i-1)+k\}^2 - \sum_{i=1}^k\{(n_i-1)+1\}^2 - (k-1)}{\{\sum_{i=1}^k (n_i-1)+k\}} \\ &= \frac{\{\sum_{i=1}^k (n_i-1)\}^2 - \sum_{i=1}^k (n_i-1)^2 + (k-1)\sum_{i=1}^k (n_i-1)}{\sum_{i=1}^k n_i}\end{aligned}$$

Since the sample sizes $n_i \geq 1$ for all the studies, we have $\{\sum_{i=1}^k (n_i-1)\}^2 - \sum_{i=1}^k (n_i-1)^2 \geq 0$. Noting also that $k \geq 2$, it further yields that $\sum_{i=1}^k n_i - \sum_{i=1}^k n_i^2/\sum_{i=1}^k n_i - (k-1) \geq 0$, and the equality holds only when $n_i = 1$ for all studies. □

Proof of (b). For the balanced design, the weights are given by $w_i = n/\sigma^2_{\text{pop}}$. Hence,

$$\begin{aligned}\frac{Q}{(k-1)(\tilde{n}-1)} &= \frac{\sum_{i=1}^k w_i(y_i - \sum_{i=1}^k w_i y_i/\sum_{i=1}^k w_i)^2}{(k-1)(n-1)} \\ &= \frac{n}{n-1}\cdot\frac{1}{\sigma^2_{\text{pop}}}\cdot\frac{\sum_{i=1}^k (y_i - \sum_{i=1}^k y_i/k)^2}{k-1}\end{aligned}$$

As $n \to \infty$, y_i converges in distribution to $N(\mu, \tau^2)$. Therefore, $\sum_{i=1}^k (y_i - \sum_{i=1}^k y_i/k)^2/\tau^2$ converges in distribution to $\chi^2(k-1)$. Along with the fact that $\text{Var}\{\sum_{i=1}^k (y_i - \sum_{i=1}^k y_i/k)^2/(k-1)\} \to 0$ as $k \to \infty$, it follows that $Q/\{(k-1)(n-1)\}$ converges in probability to $\tau^2/\sigma^2_{\text{pop}}$ as $k \to \infty$ and $n \to \infty$.

Similarly, for any fixed k, it can be drived that $Q/(k-1) = (n/\sigma^2_{\text{pop}})\sum_{i=1}^k (y_i - \sum_{i=1}^k y_i/k)^2/(k-1) = O(n)$. □

Proof of (c). When all other sample sizes are fixed, we have

$$\begin{aligned}&\frac{\partial\{(k-1)(\tilde{n}-1)\}}{\partial n_i} \\ &= \frac{\partial\left\{n_i + \sum_{j\neq i} n_j - (n_i^2 + \sum_{j\neq i} n_j^2)/(n_i + \sum_{j\neq i} n_j)\right\}}{\partial n_i} \\ &= \frac{2\sum_{j\neq i} n_j^2}{n_i^2 + 2n_i\sum_{j\neq i} n_j + \sum_{j\neq i} n_j^2} > 0\end{aligned}$$

This shows that $(k-1)(\tilde{n}-1)$ is an increasing function of n_i given that all other sample sizes are fixed.

For the unbalanced design, by noting that $w_i = n_i/\sigma^2_{\text{pop}}$, we have

$$\begin{aligned}Q &= \frac{1}{\sigma^2_{\text{pop}}}\sum_{i=1}^k n_i\left(y_i - \frac{\sum_{i=1}^k n_i y_i}{\sum_{i=1}^k n_i}\right)^2 \\ &= \frac{1}{\sigma^2_{\text{pop}}}\left\{\sum_{i=1}^k n_i y_i^2 - \frac{(\sum_{i=1}^k n_i y_i)^2}{\sum_{i=1}^k n_i}\right\}\end{aligned}$$

As $n_i \to \infty$, y_i converges in distribution to $N(\mu, \tau^2)$. With $y_i \sim N(\mu, \tau^2)$, we have $\{(n_i - \sum_{i=1}^k n_i^2)/\sum_{i=1}^k n_i\}^{-1}E\left\{\sum_{i=1}^k n_i y_i^2 - (\sum_{i=1}^k n_i y_i)^2/\sum_{i=1}^k n_i\right\}$ converges to τ^2. Additionally, when $k \to \infty$ and n_i are of the same order, $\{(n_i - \sum_{i=1}^k n_i^2)/\sum_{i=1}^k n_i\}^{-2}\text{Var}\left\{\sum_{i=1}^k n_i y_i^2 - (\sum_{i=1}^k n_i y_i)^2/\sum_{i=1}^k n_i\right\} \to 0$. Thus, as $n \to \infty$ and $k \to \infty$, $Q/\{(k-1)(\tilde{n}-1)\} \to \tau^2/\sigma^2_{pop}$. □

Proof of (d). By (14), we have

$$\begin{aligned}I^2_{\text{A}} &= \max\left\{\frac{Q-(k-1)}{Q+(k-1)(\tilde{n}-1)}, 0\right\} \\ &= \max\left\{\frac{\{Q-(k-1)\}/(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i)}{\{Q-(k-1)\}/\left(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i\right) + \{(k-1)\tilde{n}\}/\left(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i\right)}, 0\right\}\end{aligned}$$

where $\hat{\tau}^2 = \max\{\{Q-(k-1)\}/(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i), 0\}$ and $\tilde{\sigma}_y^2 = (k-1)/(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2/\sum_{i=1}^k w_i)$. Note that the above equality holds for any Q value, and in case $Q < k-1$, the both sides of the last equation are zero and it still holds. □

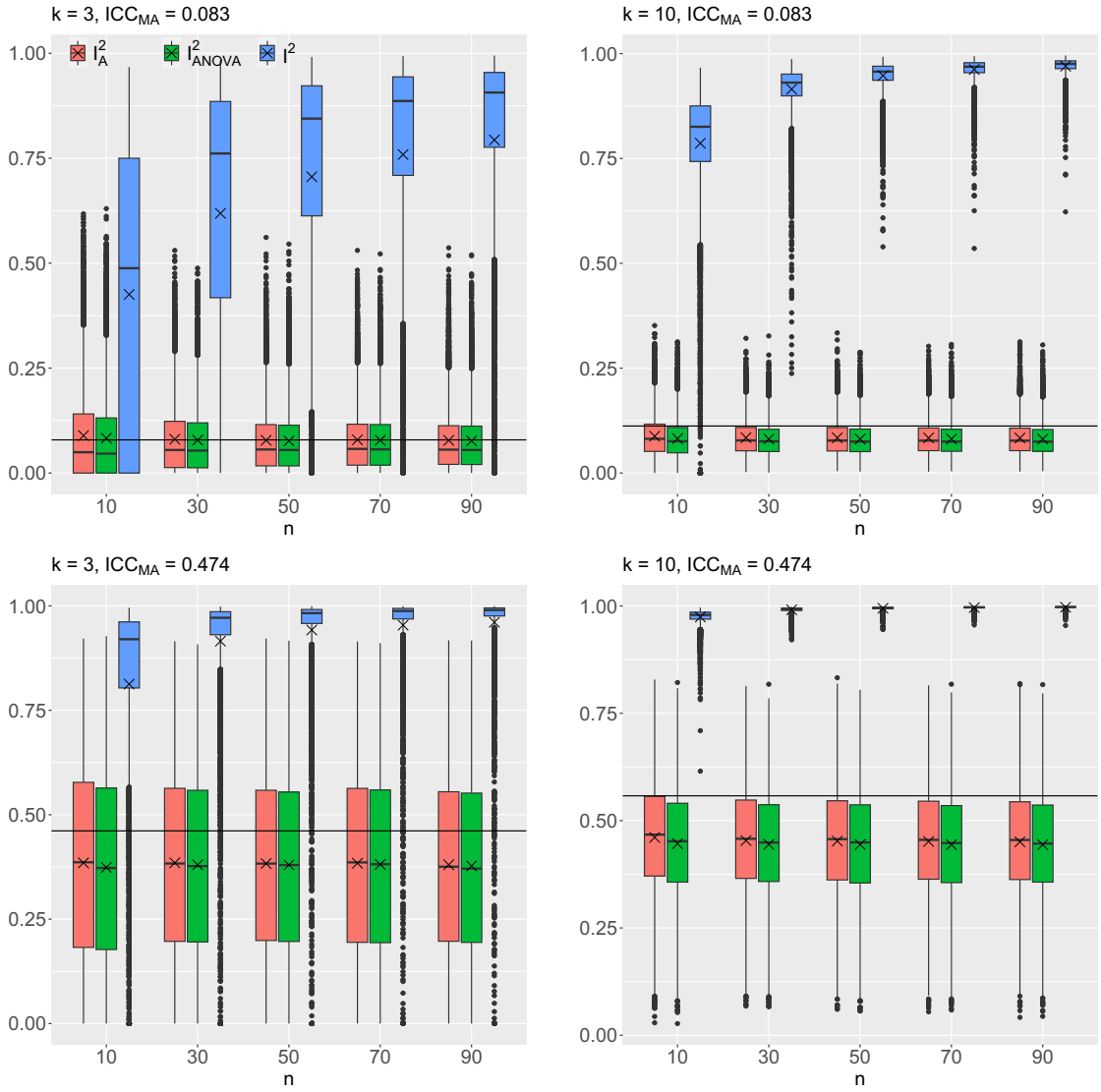

FIGURE F1 | Boxplots of the three statistics for the mean with 10,000 repetitions. The red boxes represent the I^2_{A} statistic, the green boxes represent the I^2_{ANOVA} statistic, and the blue boxes represent the I^2 statistic. The crosses on each box are the mean values of the 10000 repetitions. The solid lines stand for the absolute heterogeneity $\mathrm{ICC}_{\mathrm{MA}}$ with $\sigma^2_{\mathrm{pop}} = 100$.