Statistics and Its Interface Volume 13 (2020) 501–510

Model selection between the fixed-effects model and the random-effects model in meta-analysis

Ke Yang, Hiu-Yee Kwan, Zhiling Yu, and Tiejun Tong*

The common-effect model and the random-effects model are the two most popular models for meta-analysis in the literature. To choose a proper model between them, the Q statistic and the I^2 statistic are commonly used as the criteria. Recently, it is recognized that the fixed-effects model is also essential for meta-analysis, especially when the number of studies is small. With this new model, the existing methods are no longer sufficient for model selection in meta-analysis. In view of the demand, we propose a novel method for model selection between the fixed-effects model and the random-effects model. Specifically, we apply the Akaike information criterion (AIC) to both models and then select the model with a smaller AIC value. A real data example is also presented to illustrate how the new method can be applied. We further propose the generalized AIC (GAIC) to reduce the large variation in the AIC value, and demonstrate its superiority through real data analysis and simulation studies. To the best of our knowledge, this is the first work in meta-analysis for model selection between the fixed-effects model and the random-effects model, and we expect that our new criterion has the potential to be widely applied in meta-analysis and evidence-based medicine.

Keywords and phrases: Akaike information criterion (AIC), Common-effect model, Fixed-effects model, Meta-analysis, Model selection, Random-effects model.

1. INTRODUCTION

The concept of evidence-based approach was first introduced by [11], which aimed to improve the decision-making process based on the scientific evidence. Evidence-based practice has now spread to many different areas including, for example, medicine ([18, 33]), nursing ([13, 30]), education ([9, 35]), and public policy ([34, 31]). It is also well known that the main statistical tool for evidence-based practice is meta-analysis, which was first coined by [16] with the purpose of synthesizing multiple individual studies and producing a summary conclusion for the whole body of research ([12]). For more details of meta-analysis, one may refer to the classic textbooks in the literature, e.g., [4] and [20], and the references therein.

*Corresponding author.

One main benefit of meta-analysis is that the precision of the pooled estimate can be improved and that the results can be generalized to a larger population. In the literature, the most commonly used models for meta-analysis include the common-effect model (CEM, also known as the fixed-effect model) and the random-effects model (REM). For CEM, the effect sizes of different studies are assumed to be the same, whereas the differences between the observed effects are all subject to sampling errors. In the situations when a common effect does not hold, it is believed that the heterogeneity exists among the studies. To account for the heterogeneity in meta-analysis, one often assumes that the study-specific effect sizes are random variables from an underlying distribution, e.g., a normal distribution, and the resulting model is then REM. For more details, one may refer to [5] regarding, for example, the key assumptions and the estimation procedures of the two models.

When conducting a meta-analysis, it is often too restrictive to assume that the study-specific effects are all the same so that CEM may yield misleading results. On the other side, if a meta-analysis includes only few studies, the between-study variance cannot be accurately estimated so that the results from REM will also be unreliable. To improve the moment estimate in [10], a working group of the Cochrane Collaboration recommended to use the Knapp-Hartung method proposed by [19] and [27]. The Knapp-Hartung method considered the uncertainty of the estimation of the heterogeneity with few studies, but unfortunately the wide confidence interval remains to be unsolved. [17] investigated the impact of few studies for the existing methods including the DerSimonian-Laird method and the Knapp-Hartung method. [15] further investigated several meta-analyses with only two studies. They recommended to use a Bayesian approach with a reasonable prior due to the limitations of the currently available frequentist methods.

To conclude, when a meta-analysis includes few studies, neither CEM nor REM may provide accurate meta-analytic results. In view of their limitations, researchers including [32] and [2] have recently revisited the fixed-effects model (FEM) for meta-analysis. FEM was first introduced in the 1990s by [29] and [22]; yet for certain reasons, the model was often overlooked in the previous literature. Unlike CEM, the study-specific effect sizes are not required to be equal in FEM, which makes the meta-analysis more meaningful

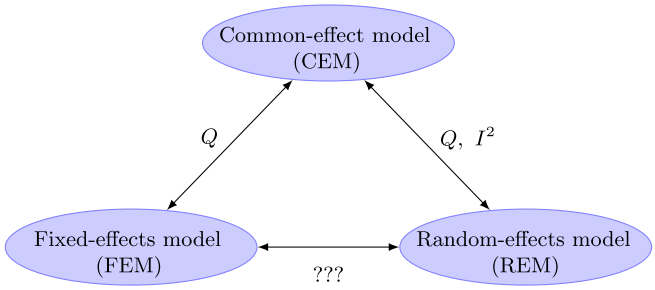

Figure 1. Model selection between the common-effect model (CEM), the fixed-effects model (FEM) and the random-effects model (REM).

when the heterogeneity exists among the studies. On the other side, since the study-specific effect sizes in FEM are assumed to be fixed but not random, we are not obligated to have an estimate of the between-study variance as in REM which can often be unreliable due to few studies. In other words, FEM can be the most appropriate model for meta-analysis with few studies.

In meta-analysis, a model selection is often needed that selects a suitable model between CEM and REM. The distinction between these two models is mainly on whether there exists some heterogeneity among the studies. For this purpose, [10] proposed the Q statistic to test for the existence of heterogeneity among the studies, and [24] proposed the I^2 statistic to measure the degree of the heterogeneity. The Q statistic and the I^2 statistic are nowadays routinely applied in meta-analysis for model selection between CEM and REM. Now with FEM also a candidate model, it is evident that the existing methods for model selection will no longer be sufficient, and so it calls for new methods for model selection.

To further clarify, we summarize the existing methods for model selection between the three models in Figure 1, in which there are 3 different lines for model selection. For model selection between CEM and REM, the Q statistic and the I^2 statistic are commonly used as the criteria (10, 24). For model selection between CEM and FEM, the Q statistic can also be applied as a criterion (21). To the best of our knowledge, however, there is no existing method for model selection between FEM and REM. In order to fill the gap, we propose to apply the Akaike information criterion (AIC) for model selection between FEM and REM. To reduce the large variation in the AIC value, we further propose the generalized AIC (GAIC) and demonstrate the superiority of GAIC through numerical studies and real data analysis.

The remainder of the paper is organized as follows. In Section 2, we review the three models for meta-analysis, introduce their underlying model assumptions, and interpret the meanings of the model parameters. A review of methods is given in Section 3 for model selection between CEM and REM, and in Section 4 for model selection between CEM and FEM. In Section 5, we propose our new method based on AIC for model selection between FEM and REM, followed by a real data analysis that gives a step-by-step instruction. In Section 6, we further propose a model selection criterion based on a generalized AIC (GAIC) and apply it to the same real data example. In Section 7, we conduct simulation studies to compare the performance of AIC and GAIC and suggest the better one for practical use. The paper is concluded in Section 8.

2. STATISTICAL MODELS FOR META-ANALYSIS

Assume that a meta-analysis includes a total of k individual studies. For the ith study, let θ_i be the effect size and y_i be the observed value of θ_i, where $i = 1, \ldots, k$. In this section, we provide a brief review on the three statistical models for meta-analysis.

2.1 Common-effect model

The common-effect model (CEM) is the simplest model for meta-analysis, in which the study-specific effect sizes are assumed to be all the same. In other words, we assume that there is no heterogeneity in the effect sizes among the studies. Let θ_{CEM} be the common effect. Then under the normality assumption for the observed effects, CEM can be formulated as

$$y_i = \theta_{\text{CEM}} + \epsilon_i,\ \epsilon_i \overset{\text{ind}}{\sim} N(0, \sigma_i^2), \tag{1}$$

where ϵ_i are independent normal errors with zero mean and variance $\sigma_i^2 > 0$.

Note that the within-study variances, σ_i^2, can often be estimated with high precision, in particular when the sample sizes are large for each study. For this reason, it is a common practice that the within-study variances are regarded as known for meta-analysis. Under the above assumptions, we can apply the maximum likelihood estimator (MLE) to estimate the common effect and it yields that (22)

$$\hat{\theta}_{\text{CEM}} = \frac{\sum_{i=1}^k w_i y_i}{\sum_{i=1}^k w_i}, \tag{2}$$

where $w_i = 1/\sigma_i^2$ are the inverse-variance weights assigned to each individual study. Moreover, by the normality assumption on the random errors, $\hat{\theta}_{\text{CEM}}$ follows a normal distribution with mean θ_{CEM} and variance $1/\sum_{i=1}^k w_i$.

2.2 Random-effects model

In practice, the effect sizes for different studies can differ and that results in the statistical heterogeneity, or referred to as heterogeneity in short. When the heterogeneity exists, one often assumes that the study-specific effect sizes follow a certain distribution, e.g., a normal distribution. It then

yields a random-effects model (REM) that can be formulated as

$$(3) y_i = \theta_{\text{REM}} + \delta_i + \epsilon_i,\ \delta_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2),\ \epsilon_i \overset{\text{ind}}{\sim} N(0, \sigma_i^2),$$

where δ_i are independent and identically distributed (i.i.d.) deviations of the study-specific effect sizes from the mean effect θ_{REM}, $\tau^2 \geq 0$ is the between-study variance, and ϵ_i and σ_i^2 are defined the same as in model (1). The effect size of the ith study is $\theta_i = \theta_{\text{REM}} + \delta_i$, where $i = 1, \ldots, k$.

By model (3), we have $E(y_i) = \theta_{\text{REM}}$ and $\text{var}(y_i) = \sigma_i^2 + \tau^2$ for any $i = 1, \ldots, k$. In the special case when $\tau^2 = 0$, REM will reduce to CEM. With the above notations, the MLE of θ_{REM} can be represented as

$$(4) \qquad \hat{\theta}_{\text{REM}} = \frac{\sum_{i=1}^k w_i^* y_i}{\sum_{i=1}^k w_i^*},$$

where $w_i^* = 1/(\sigma_i^2 + \tau^2)$ with $i = 1, \ldots, k$ are the study-specific weights. And by the fact that y_i are normally distributed, $\hat{\theta}_{\text{REM}}$ also follows a normal distribution with mean θ_{REM} and variance $1/\sum_{i=1}^k w_i^*$. Finally, noting that the between-study variance may not be known in practice, we need an estimate of τ^2 from the observed data to compute $\hat{\theta}_{\text{REM}}$ and also its confidence interval.

2.3 Fixed-effects model

When the number of studies k is small, it is known that τ^2 cannot be accurately estimated so that the meta-analytic results from REM may not be reliable. On the other side, CEM may yield misleading results when the heterogeneity exists. By contrast, the fixed-effects model (FEM) is a model that fills the gap between CEM and REM (2, 32), in which the effect sizes of the individual studies are assumed to be fixed but unequal. The statistical model for FEM is as follows:

$$(5) \qquad y_i = \theta_i + \epsilon_i,\ \epsilon_i \overset{\text{ind}}{\sim} N(0, \sigma_i^2),$$

where θ_i are the fixed effect sizes, and ϵ_i are independent normal errors with zero mean and variance $\sigma_i^2 > 0$.

By model (5), we have $E(y_i) = \theta_i$ and $\text{var}(y_i) = \sigma_i^2$ for any $i = 1, \ldots, k$. In the special case when θ_i are all equal, FEM will reduce to CEM so that the parameter of interest is the common effect. For the general setting with unequal θ_i, a parameter of interest for FEM is, however, not intuitively known. In [29] and [2], the authors proposed the average effect, $\theta_{\text{FEM}} = k^{-1} \sum_{i=1}^k \theta_i$, as the parameter of interest. They further provided an unbiased estimator of θ_{FEM} as

$$(6) \qquad \hat{\theta}_{\text{FEM}} = \frac{1}{k} \sum_{i=1}^k y_i,$$

which is normally distributed with mean θ_{FEM} and variance $k^{-2} \sum_{i=1}^k \sigma_i^2$. Finally, we note that the main purpose of FEM is to study the statistical inference on the effect sizes of the specific k studies that are given in the meta-analysis, but not on the whole population of the effect sizes.

3. MODEL SELECTION BETWEEN CEM AND REM

Needless to say, most existing methods developed in the literature are for model selection between CEM and REM. In this section, we provide a brief review on the two most widely used methods for this model selection, including the Q statistic and the I^2 statistic.

3.1 The Q statistic

Recall that REM will reduce to CEM when $\tau^2 = 0$, i.e., when there is no heterogeneity among the studies. Hence for model selection between REM and CEM, it can be formulated as a statistical testing problem. Specifically, to test whether there exists the heterogeneity among the studies, we consider the following hypotheses:

$$(7) \qquad H_0 : \tau^2 = 0 \quad \text{versus} \quad H_1 : \tau^2 > 0.$$

For the testing problem (7), [10] proposed the Q statistic as

$$(8) \qquad Q = \sum_{i=1}^k w_i (y_i - \hat{\theta}_{\text{CEM}})^2,$$

where $w_i = 1/\sigma_i^2$ are the optimal weights for CEM as in Section 2.1. It is also worth noting that the origin of the Q statistic can be dated back to [3], [26] and [36], or the well-known Cochran's Q statistic (7).

If the result of the Q test is significant, there is a strong evidence for the presence of the heterogeneity so that we should apply REM for meta-analysis; otherwise, we adopt CEM as the default model. Under the null hypothesis, the Q statistic is often approximated by a chi-square distribution with $k-1$ degrees of freedom. Nevertheless, such an approximation for the null distribution of Q can be less accurate in many practical situations, especially when the number of studies is small and/or the effect sizes are not normally distributed. In view of these limitations, researchers have also considered other approximate null distributions for the Q statistic in the literature; see, for example, the rescaled F-distribution in [36], and the gamma distribution in [28].

3.2 The I^2 statistic

One major criticism on the Q statistic is that it heavily depends on the number of studies, and so may not serve well for model selection between CEM and REM. To be more specific, when the number of studies is sufficiently large, the Q statistic is able to detect any arbitrarily small heterogeneity among the studies. Such a small heterogeneity with statistical significance, however, may not be clinically important. In other words, the statistical significance is not identical to the clinical significance.

To overcome the problem, [24] assumed that $\sigma_i^2 = \sigma^2$ for all $i = 1, \ldots, k$, and then applied the function

$$f(\theta_{\text{REM}}, \tau^2, \sigma^2, k) = \frac{\tau^2}{\tau^2 + \sigma^2}$$

to quantify the extent of the heterogeneity. Noting that $f(\cdot)$ is scale invariant and not affected by the number of studies, it provides an alternative yet better approach for model selection between CEM and REM.

When σ_i^2 are all equal, it can be shown that $\{E(Q) - (k-1)\}/E(Q) = f(\theta_{\text{REM}}, \tau^2, \sigma^2, k)$. This suggests that $\{Q - (k-1)\}/Q$ can be applied to estimate the unknown quantity of $f(\theta_{\text{REM}}, \tau^2, \sigma^2, k)$. To avoid negative values of $Q - (k-1)$, [25] further suggested to estimate the heterogeneity by

$$I^2 = \max\{0, \frac{Q - (k-1)}{Q}\}. \tag{9}$$

Noting that $\{Q - (k-1)\}/Q$ does not explicitly involve σ^2 in its calculation, this method can also be generalized to estimate the heterogeneity for studies with different within-study variances. To conclude, I^2 is not a test statistic but a measure of the degree of the heterogeneity, and it has nowadays been widely applied in meta-analysis for model selection between CEM and REM.

4. MODEL SELECTION BETWEEN CEM AND FEM

With the new recognition of the fixed-effects model (FEM) for meta-analysis with few studies, there is also a demand for developing new model selection methods between FEM and the existing two models including CEM and REM.

In this section, we review the existing method for model selection between CEM and FEM. To start with, we note in Section 3 that the Q statistic was originally proposed for model selection between CEM and REM. In an interesting work by [21], the authors pointed out that the Q statistic can also be applied for model selection between CEM and FEM. Specifically, they formulated the new model selection as the following statistical testing problem:

$$\begin{aligned} & H_0 : \theta_1 = \cdots = \theta_k \\ \text{versus} \quad & H_1 : \theta_i \neq \theta_j \text{ for some } i \neq j. \end{aligned} \tag{10}$$

Note that the null hypothesis in (10) is identical to the null hypothesis in (7), in which they both imply that the study-specific effect sizes are all the same. On the other side, when θ_i are not all equal (no matter whether fixed or random), they will tend to yield a larger Q value than purely generated by random errors. Hence for the testing problem (10), the Q statistic in (8) can still be applied; and consequently, we select FEM for meta-analysis if the null hypothesis is rejected, and otherwise adopt CEM as the default model.

4.1 Power functions

Even though the same null hypothesis and the same test statistic are applied, the alternative hypotheses in (7) and (10) are completely different, and accordingly, they also have different alternative distributions and different power functions.

For the testing problem (7), the statistical power is defined as the probability of detecting the model as REM when the heterogeneity exists. Under the significance level α and the assumption that the within-study variances are all equal to σ^2, [21] derived the power function of the Q statistic as

$$\beta_1 = 1 - \chi^2_{k-1}(c_\alpha \cdot \frac{\sigma^2}{\tau^2 + \sigma^2}), \tag{11}$$

where c_α is the $100(1-\alpha)$ percentile of the chi-square distribution with $k-1$ degrees of freedom, and $\chi^2_{df}(c)$ is the cumulative distribution function of the chi-square random variable with df degrees of freedom.

For the testing problem (10), the statistical power is defined as the probability of detecting the model as FEM when the heterogeneity exists. To compute the power, we let $\lambda = \sum_{i=1}^k w_i(\theta_i - \bar{\theta})^2$, where $\bar{\theta} = \sum_{i=1}^k w_i\theta_i / \sum_{i=1}^k w_i$ is the weighted average of the true effect sizes. Then under the significance level α, [21] also derived that the power function of the Q statistic for testing CEM versus FEM is given by

$$\beta_2 = 1 - \chi^2_{k-1}(c_\alpha; \lambda), \tag{12}$$

where $\chi^2_{df}(c; \lambda)$ is the cumulative distribution function of the noncentral chi-square random variable with df degrees of freedom and noncentrality parameter λ.

4.2 Power comparison

Since the testing problems (7) and (10) can both be used to test whether or not the heterogeneity exists among the studies, we are keen to know which test is more powerful for meta-analysis. For this, we conduct a numerical study to compare the two power functions in (11) and (12) for further insights about the Q statistic.

To make the two tests comparable, we follow the assumption in deriving the power function for the testing problem (7) that the within-study variances are all equal. For simplicity and without loss of generality, we let $\sigma_i^2 = 1$ for all $i = 1, \ldots, k$ for both of the tests. We also let $\tau^2 = 0.5$ and 5, and set the significance level α up to 0.1. Then with the above settings, we compute the power function of the Q statistic for testing CEM versus REM using formula (11), and plot them in Figure 2 and Figure 3 by the red lines with triangles for $k = 2$, 3, 5 and 10, respectively. While for the power function associated with the testing problem (10), we first generate θ_i from $N(0, 0.5)$ and $N(0, 5)$ for $\tau^2 = 0.5$ and 5 respectively so as to maintain the same variability for the

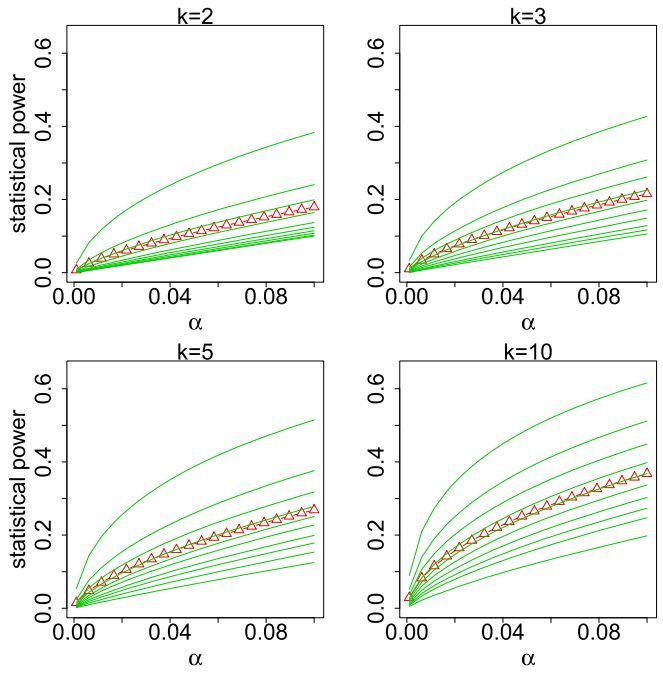

Figure 2. Power comparison between the two testing problems with $\tau^2 = 0.5$. The red lines with triangles represent the power functions of the Q statistic for the testing problem (7), and the green lines represent the power functions of the Q statistic for the testing problem (10).

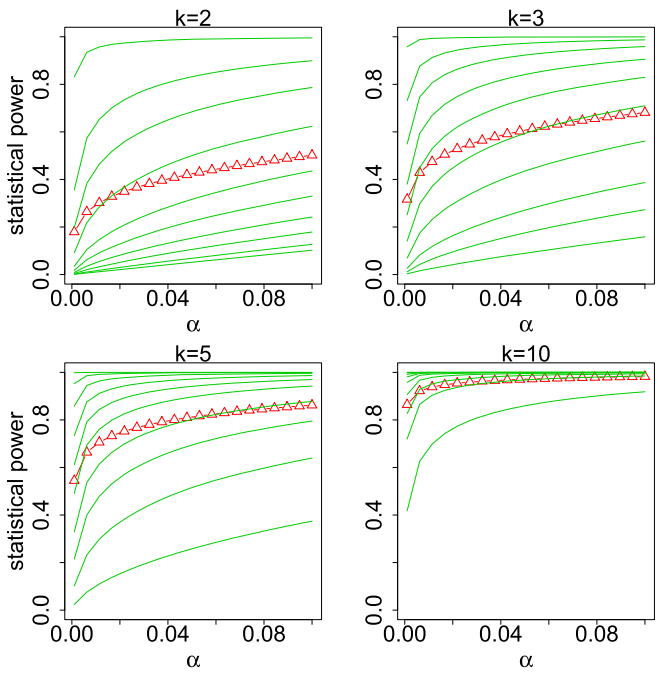

Figure 3. Power comparison between the two testing problems with $\tau^2 = 5$. The red lines with triangles represent the power functions of the Q statistic for the testing problem (7), and the green lines represent the power functions of the Q statistic for the testing problem (10).

between-study variances for both tests, and we then treat them as the fixed quantities so that formula (12) can be applied to compute the power function. Noting also that θ_i are generated with randomness, we repeat the simulation for 10 times and plot their respective power function also in Figure 2 and Figure 3 by the green lines.

From the power curves simulated in Figure 2 and Figure 3, we note that the power functions of the two tests will increase when the number of studies becomes larger, in particular when the between-study variance is also large. Note also that, for a fair comparison between the power functions, the study-specific effect sizes θ_i are generated both from $N(0, \tau^2)$ for FEM and REM. Specifically for FEM, the variation among the generated θ_i can be different in each realization, sometimes large and sometimes small, and that explains why it yields different power functions; while for REM, no matter what the true heterogeneity is, one applies τ^2 to compute the power functions so that it only yields an average power function. In other words, applying the Q statistic to test CEM versus FEM can result in a more accurate power function for a specific meta-analysis. In addition, we note that the between-study variance τ^2 is unknown and needs to be estimated when computing the power function, which can be another problem for applying the Q statistic to test CEM versus REM in meta-analysis with few studies.

5. MODEL SELECTION BETWEEN FEM AND REM

As mentioned earlier, the assumption of a common effect may not be realistic in many studies. On the other side, a meta-analysis with random effects will also not be reliable when the number of studies is small. Recently, [32] and [2] revisited FEM and demonstrated that it can be a good alternative for modeling meta-analysis with few studies. However, as shown in Figure 1, there is no existing method in the literature for model selection between FEM and REM. Note that FEM and REM are both applicable when the heterogeneity exists among the studies. It is thus different from the classical model selection problem in which we have the truth that one model is correct and the other is not. But instead, our aim is to find the better model between FEM and REM so that, for the given summary data, the meta-analytical results will be more meaningful to synthesize the multiple studies for medical decision making.

5.1 A new method for model selection

We propose to apply the Akaike information criterion (AIC) for model selection between FEM and REM. Note that AIC has been widely used for model selection since it was introduced in the 1970s. Unlike statistical hypothesis testing, which is valid only for nested models, AIC

has no such restrictions (6). The main idea of AIC for model selection is to maximize the expected log-likelihood function which can be expressed as $E_{X,\hat{\theta}}(\ln f(X|\hat{\theta})) = E_{\hat{\theta}}(\int f(x|\theta)\ln f(x|\hat{\theta})\mathrm{d}x)$, where X is a random variable with the probability density function $f(x|\theta)$, and $\hat{\theta}$ is the MLE of θ and is independent of X. Further to compute the expected log-likelihood function, [1] defined the AIC value of the model as $\mathrm{AIC} = 2d - 2\sum_{i=1}^{k}\ln f(x_i|\hat{\theta})$, where $x_1,\ldots,x_k$ are the sample values of the random variable X, and d is the number of independent parameters. The AIC value can be easily computed since it does not involve unknown parameters. And under regularity conditions, it was shown that $-\mathrm{AIC}/(2k)$ provides an accurate approximation for the unknown quantity of the expected log-likelihood function $E_{X,\hat{\theta}}(\ln f(X|\hat{\theta}))$.

For model selection between FEM and REM, we propose to compute the AIC values of both models and then select the model with a smaller AIC value. Specifically for FEM in (5), it has a total of k individual distributions with $\theta_1,\ldots,\theta_k$ being the unknown parameters; that is, the number of independent parameters is $d = k$. Noting also that the MLE for each θ_i is given by y_i, the AIC value of FEM then can be specified as

$$\begin{aligned}(13)\qquad \mathrm{AIC}_{\mathrm{FEM}} &= 2k - 2\sum_{i=1}^{k}\ln \\ &\quad \left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left\{-\frac{(y_i-\hat{\theta}_i)^2}{2\sigma_i^2}\right\}\right] \\ &= 2k + \sum_{i=1}^{k}\ln(2\pi\sigma_i^2).\end{aligned}$$

While for REM, we follow the same setting as in the derivation of the I^2 statistic (24) that τ^2 is assumed to be known. Consequently, there is only one parameter, i.e., θ_{REM}, in model (3) that needs to be estimated. This yields the AIC value of REM as

$$\begin{aligned}(14)\qquad \mathrm{AIC}'_{\mathrm{REM}} &= 2 + \sum_{i=1}^{k}\ln\{2\pi\left(\sigma_i^2+\tau^2\right)\} \\ &\quad + \sum_{i=1}^{k}\frac{(y_i-\hat{\theta}_{\mathrm{REM}})^2}{\sigma_i^2+\tau^2},\end{aligned}$$

where $\hat{\theta}_{\mathrm{REM}}$ is the MLE of θ_{REM} which is given in (4). Moreover, for the value of τ^2 in $\hat{\theta}_{\mathrm{REM}}$ and (14), we apply the moment estimate $\hat{\tau}^2_{\mathrm{DL}} = \max\{0, \{Q-(k-1)\}/(\sum_{i=1}^{k}w_i - \sum_{i=1}^{k}w_i^2/\sum_{i=1}^{k}w_i)\}$ proposed by [10]. Letting also $\tilde{\theta}_{\mathrm{REM}} = \{\sum_{i=1}^{k}(\sigma_i^2+\hat{\tau}^2_{\mathrm{DL}})^{-1}y_i\}/\{\sum_{i=1}^{k}(\sigma_i^2+\hat{\tau}^2_{\mathrm{DL}})^{-1}\}$, the AIC value of REM can be rewritten as

$$\begin{aligned}(15)\qquad \mathrm{AIC}_{\mathrm{REM}} &= 2 + \sum_{i=1}^{k}\ln\{2\pi\left(\sigma_i^2+\hat{\tau}^2_{\mathrm{DL}}\right)\} \\ &\quad + \sum_{i=1}^{k}\frac{(y_i-\tilde{\theta}_{\mathrm{REM}})^2}{\sigma_i^2+\hat{\tau}^2_{\mathrm{DL}}}.\end{aligned}$$

Finally, with the numerical values of AIC for both models, we select FEM for meta-analysis if $\mathrm{AIC}_{\mathrm{FEM}}$ in (13) is less than $\mathrm{AIC}_{\mathrm{REM}}$ in (15), and vice versa.

5.2 Real data analysis

To apply the new criterion for model selection, we consider a real data example of systematic review from [14]. The main purpose of the study is to investigate the effect of parental migration on the health of left behind-children and adolescents in low-income and middle-income countries. Among the conducted meta-analyses in their study, we consider the one with the weight-for-age Z scores as the outcomes, where the standardized mean difference (SMD) of the measure is assumed to be normally distributed. And for ease of reference, we also present the observed effect sizes and their respective variances in Table 1.

Table 1. Summary data of the three studies for meta-analysis from [14]

Study	y_i	σ_i^2
Wang et al. (2001)	−0.58	0.0055
Chen et al. (2012)	−0.36	0.0099
Chen et al. (2013)	−0.03	0.0018

By estimator (2) and the observed values in Table 1, the common effect from CEM is given as

$$\begin{aligned}\hat{\theta}_{\mathrm{CEM}} &= \frac{-0.58/0.0055 - 0.36/0.0099 - 0.03/0.0018}{1/0.0055 + 1/0.0099 + 1/0.0018} \\ &= -0.189.\end{aligned}$$

Further by formulas (8) and (9), the Q statistic is

$$\begin{aligned}Q &= \frac{(-0.58+0.189)^2}{0.0055} + \frac{(-0.36+0.189)^2}{0.0099} \\ &\quad + \frac{(-0.03+0.189)^2}{0.0018} \\ &= 44.795,\end{aligned}$$

and the I^2 statistic is

$$I^2 = \frac{44.795-(3-1)}{44.795} = 0.96.$$

Noting that I^2 is as large as 0.96, a common effect is unlikely to be true for the three studies. And more specifically, by Cochrane Handbook for Systematic Reviews of Interventions (23), one would suggest opting out CEM for further consideration.

Now given CEM is no longer considered, as a next step, we apply our AIC in Section 5.1 to perform model selection

between FEM and REM. By (13), the AIC value of FEM is given as

$$\begin{aligned}\mathrm{AIC}_{\mathrm{FEM}} &= 6+\ln(2\pi\times 0.0055)+\ln(2\pi\times 0.0099)\\ &\quad +\ln(2\pi\times 0.0018)\\ &= -4.624.\end{aligned}$$

To compute the AIC value of REM, we have $\sum_{i=1}^{k} w_i = 1/0.0055 + 1/0.0099 + 1/0.0018 = 838.384$ and $\sum_{i=1}^{k} w_i^2 = 1/0.0055^2 + 1/0.0099^2 + 1/0.0018^2 = 351902.9$. They further yields that

$$\hat{\tau}^2_{\mathrm{DL}} = \max\{0, \frac{44.795-(3-1)}{838.384-351902.9/838.384}\} = 0.102,$$

and

$$\begin{aligned}\tilde{\theta}_{\mathrm{REM}} &= \frac{\frac{-0.58}{0.0055+0.102}+\frac{-0.36}{0.0099+0.102}+\frac{-0.03}{0.0018+0.102}}{\frac{1}{0.0055+0.102}+\frac{1}{0.0099+0.102}+\frac{1}{0.0018+0.102}}\\ &= -0.319.\end{aligned}$$

Plugging these estimates into (15), we have the AIC value of REM as

$$\begin{aligned}\mathrm{AIC}_{\mathrm{REM}} &= 2+\ln\{2\pi(0.0055+0.102)\}\\ &\quad +\ln\{2\pi(0.0099+0.102)\}\\ &\quad +\ln\{2\pi(0.0018+0.102)\}\\ &\quad +\frac{(-0.58+0.319)^2}{0.0055+0.102}+\frac{(-0.36+0.319)^2}{0.0099+0.102}\\ &\quad +\frac{(-0.03+0.319)^2}{0.0018+0.102}\\ &= 2.281.\end{aligned}$$

Lastly, since $\mathrm{AIC}_{\mathrm{FEM}} = -4.624$ is less than $\mathrm{AIC}_{\mathrm{REM}} = 2.281$, we select FEM as the final model for meta-analysis with the included studies.

To further compare FEM and REM and show that FEM can be an appropriate model, we present their meta-analytical results using the forest plot in Figure 4. While for reference, the results for CEM are also presented in Figure 4. Firstly, we note that the average effect from FEM and the mean effect from REM are numerically close to each other, even though their estimation formulas in (4) and (6) are rather different. The main reason is that, due to the small number of studies, the between-study variance is much larger than the within-study variances so that the inverse-variance weights $(33.4\%, 32.1\%, 34.5\%)$ for REM are nearly equally weighted. Secondly, we note that FEM and REM are with very different confidence intervals. To be more specific, FEM reports a significant result for the average effect, whereas REM fails to do so due to the unacceptably large standard error of the mean effect. Finally, to verify which model is more suitable, we also note that the first and second studies are both significant, and the third study also shows slight evidence, that the weight-for-age Z scores

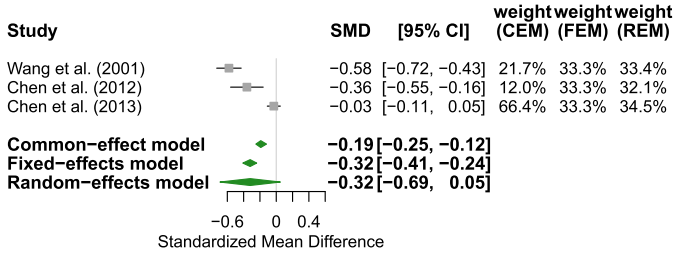

Figure 4. Forest plot of the meta-analysis for the three studies from [14].

of left-behind children and adolescents is smaller than that of children with non-migrating parents. Then following the spirit of meta-analysis, the overall effect of the three studies would also be more likely significant. This coincides more with the meta-analytical result from FEM compared to that from REM, and so it demonstrates that our AIC for model selection is meaningful and can be applied to meta-analysis with few studies.

6. GENERALIZED AIC FOR MODEL SELECTION

Recall that, to compute $\mathrm{AIC}'_{\mathrm{REM}}$ in (14), we have applied the moment estimate $\hat{\tau}^2_{\mathrm{DL}}$ for the unknown τ^2, and that yields the observable value of $\mathrm{AIC}_{\mathrm{REM}}$ in (15). When the number of studies is small, however, the estimate of τ^2 is often unreliable, and so is true for the value of $\mathrm{AIC}_{\mathrm{REM}}$. In this section, we propose a generalized AIC for model selection between FEM and REM that aims to dramatically reduce the dependence of the AIC value on the τ^2 estimate.

6.1 Generalized AIC

To eliminate the influence of randomness, we take the expected values of AIC over the observed effects y_i and define them as the generalized AIC (GAIC) values. For FEM, since $\mathrm{AIC}_{\mathrm{FEM}}$ does not involve y_i, we have

$$\text{(16)}\quad \mathrm{GAIC}_{\mathrm{FEM}} = E(\mathrm{AIC}_{\mathrm{FEM}}) = 2k + \sum_{i=1}^{k}\ln(2\pi\sigma_i^2),$$

which is, in fact, the same as $\mathrm{AIC}_{\mathrm{FEM}}$.

While for REM, we note that $\sum_{i=1}^{k}(y_i-\hat{\theta}_{\mathrm{REM}})^2/(\sigma_i^2+\tau^2)$ follows a chi-square distribution with $k-1$ degrees of freedom (19). Thus by (14), it yields that

$$E(\mathrm{AIC}'_{\mathrm{REM}}) = k+1+\sum_{i=1}^{k}\ln\left\{2\pi\left(\sigma_i^2+\tau^2\right)\right\}.$$

Further, by replacing the remaining τ^2 with $\hat{\tau}^2_{\mathrm{DL}}$, we have the GAIC value for REM as

$$\text{(17)}\quad \mathrm{GAIC}_{\mathrm{REM}} = k+1+\sum_{i=1}^{k}\ln\left\{2\pi\left(\sigma_i^2+\hat{\tau}^2_{\mathrm{DL}}\right)\right\}.$$

Compared to $\mathrm{AIC}_{\mathrm{REM}}$ in (15), it is evident that $\mathrm{GAIC}_{\mathrm{REM}}$ depends less on the τ^2 estimate, and so it provides a more stable estimate for the value of $\mathrm{AIC}'_{\mathrm{REM}}$.

By (16) and (17), both of $\mathrm{GAIC}_{\mathrm{FEM}}$ and $\mathrm{GAIC}_{\mathrm{REM}}$ depend on the number of studies k and the within-study variances σ_i^2. Hence for easy comparison, we further take the difference between them as follows:

$$\mathrm{GAIC}_{\mathrm{FEM}} - \mathrm{GAIC}_{\mathrm{REM}} = k - 1 + \sum_{i=1}^{k} \ln \frac{\sigma_i^2}{\sigma_i^2 + \hat{\tau}_{\mathrm{DL}}^2}. \tag{18}$$

We then select FEM to perform meta-analysis if $\mathrm{GAIC}_{\mathrm{FEM}} - \mathrm{GAIC}_{\mathrm{REM}} < 0$; and otherwise, we apply REM as usual. Also by (18), it suggests that FEM will be preferred when the number of studies is small and/or the estimate of the between-study variance is much larger than the within-study variances.

6.2 Real data analysis

To apply the GAIC method for model selection, we revisit the real data example in Section 4.2. Note that the values of AIC and GAIC are the same for FEM, but not for REM. To check whether the two criteria will select the same model, by (18) we have

$$\begin{aligned} \mathrm{GAIC}_{\mathrm{FEM}} - \mathrm{GAIC}_{\mathrm{REM}} &= 2 + \ln \frac{0.0055}{0.0055 + 0.102} \\ &\quad + \ln \frac{0.0099}{0.0099 + 0.102} \\ &\quad + \ln \frac{0.0018}{0.0018 + 0.102} \\ &= -7.452. \end{aligned}$$

Now since $\mathrm{GAIC}_{\mathrm{FEM}} - \mathrm{GAIC}_{\mathrm{REM}} < 0$, we once again select FEM to perform meta-analysis for the included studies. That is, for this real study, the model selection by GAIC is the same as that by AIC.

It is also interesting to point out that $\mathrm{GAIC}_{\mathrm{FEM}} - \mathrm{GAIC}_{\mathrm{REM}} = -7.452 < -6.905 = \mathrm{AIC}_{\mathrm{FEM}} - \mathrm{AIC}_{\mathrm{REM}}$, which implies that, with the elimination of randomness, GAIC is more inclined to select FEM for meta-analysis. More comparison on the AIC and GAIC values for REM is given in the next section.

7. COMPARISON BETWEEN $\mathrm{AIC}_{\mathrm{REM}}$ AND $\mathrm{GAIC}_{\mathrm{REM}}$

Following the key idea of AIC, for a model selection between FEM and REM, we are required to compare $\mathrm{AIC}_{\mathrm{FEM}}$ in (13) and $\mathrm{AIC}'_{\mathrm{REM}}$ in (14) and then choose the smaller value. As mentioned in Section 5, $\mathrm{AIC}_{\mathrm{FEM}}$ can be readily computed from the sample data, but, by contrast, $\mathrm{AIC}'_{\mathrm{REM}}$ is not computable due to the unknown τ^2. To overcome this problem, we have proposed two estimators, $\mathrm{AIC}_{\mathrm{REM}}$ in (15) and $\mathrm{GAIC}_{\mathrm{REM}}$ in (17), for estimating the unknown $\mathrm{AIC}'_{\mathrm{REM}}$.

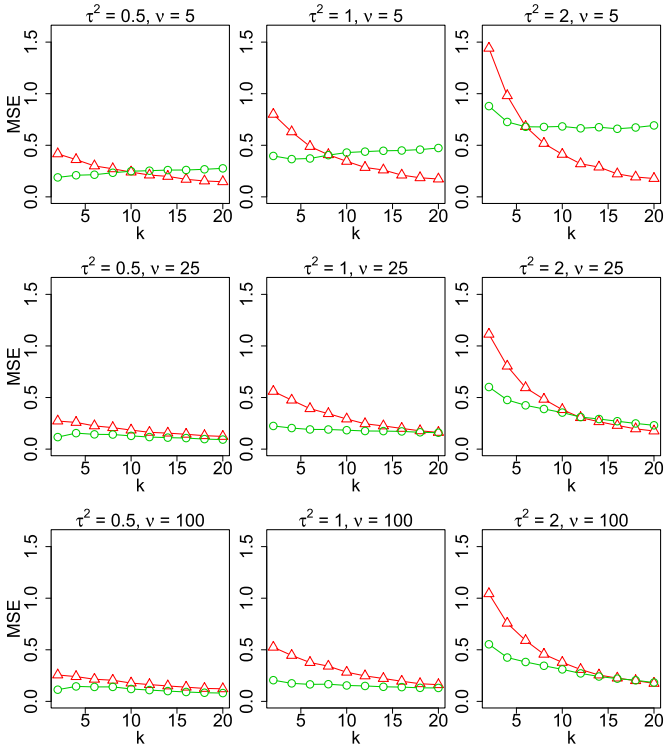

Figure 5. Numerical comparison on the accuracy of AIC_{REM} *and* GAIC_{REM} *to estimate* AIC'_{REM}*, where the red lines with triangles represent the MSE of* AIC_{REM}/k*, and the green lines with circles represent the MSE of* GAIC_{REM}/k.

In this section, we conduct simulation studies to compare their performance and suggest the better one for practical use.

To generate data from model (3), we consider k ranging from 2 to 20, let $\tau^2 = 0.5, 1$ or 2, and set $\theta_{\mathrm{REM}} = 0$ without loss of generality. For the within-study variances, we let σ_i^2 be randomly drawn from a scaled chi-square distribution with ν degrees of freedom, i.e., from χ_ν^2/ν. We further consider three different degrees of freedom, $\nu = 5, 25$ or 100, to represent different levels of heterogeneity in the within-study variances. Then for each combination of (k, τ^2, ν), we generate the data using model (3), and apply (14), (15) and (17) to calculate the values of $\mathrm{AIC}'_{\mathrm{REM}}$, $\mathrm{AIC}_{\mathrm{REM}}$ and $\mathrm{GAIC}_{\mathrm{REM}}$ for each simulation. Finally, with $M = 20{,}000$ repetitions, we compute the mean squared errors (MSE) of $\mathrm{AIC}_{\mathrm{REM}}/k$ and $\mathrm{GAIC}_{\mathrm{REM}}/k$ to evaluate the accuracy of $\mathrm{AIC}_{\mathrm{REM}}$ and $\mathrm{GAIC}_{\mathrm{REM}}$ for estimating $\mathrm{AIC}'_{\mathrm{REM}}$ as follows:

$$\begin{aligned} \mathrm{MSE}(\mathrm{AIC}_{\mathrm{REM}}/k) &= \frac{1}{kM} \sum_{j=1}^{M} (\mathrm{AIC}_{\mathrm{REM}}^{(j)} - \mathrm{AIC}'^{(j)}_{\mathrm{REM}})^2, \\ \mathrm{MSE}(\mathrm{GAIC}_{\mathrm{REM}}/k) &= \frac{1}{kM} \sum_{j=1}^{M} (\mathrm{GAIC}_{\mathrm{REM}}^{(j)} - \mathrm{AIC}'^{(j)}_{\mathrm{REM}})^2. \end{aligned}$$

With MSE as the criterion, it is clear that the smaller the MSE is, the more accurate the estimator is.

As clarified in the beginning of Section 6, when the number of studies is small, τ^2 cannot be estimated reliably that is heavily used in AIC'_{REM}. To visualize the impact of k on the estimation accuracy, we plot the MSE values of AIC_{REM} and GAIC_{REM} along with the number of studies in Figure 5 for comparison. From the results in Figure 5, we can see that GAIC_{REM} always provides a more accurate estimate than AIC_{REM} when the number of studies is small. Such an advantage is getting more evident when the degrees of freedom ν is also large. In particular, when the within-study variances are all the same (or equivalently when $\nu \to \infty$), GAIC_{REM} will perform better than AIC_{REM} in most settings.

According to [8], there are a total of 22,453 meta-analyses in the January 2008 issue of the Cochrane Database of Systematic Reviews, and the median number of studies included in those meta-analyses is only 3 studies (with the interquartile range from 2 to 6). This indicates that our simulation settings with k ranging from 2 to 20 are able to cover the majority of meta-analyses in the literature. In addition, our simulated within-study variances, with ν ranging from 5 to 100, have also covered a wide range of the heterogeneity level for the within-study variances. To conclude, GAIC_{REM} can always be recommended to estimate AIC'_{REM}, as long as the number of studies and/or the heterogeneity level for the within-study variances are/is not extremely large.

8. CONCLUSION

The common-effect model (CEM), the fixed-effects model (FEM) and the random-effects model (REM) consist of the three fundamental models for meta-analysis. When the heterogeneity exists, CEM may not be reasonable due to its restrictive assumption on a common effect for all the studies. On the other side, when there are only few studies, REM will also suffer from the inaccurate estimate of the between-study variance. By contrast, FEM can effectively avoid the limitations on CEM and REM, and thus provides a good compromise between them for meta-analysis with few studies. Methods for model selection between CEM and REM have been well studied in the literature, in which the commonly used statistics include, for example, the Q statistic and the I^2 statistic. It is also noteworthy that the Q statistic can be applied as well to model selection between CEM and FEM. To the best of our knowledge, however, there is no existing method for model selection between FEM and REM up to now.

In this paper, we propose a novel method for model selection between FEM and REM based on the AIC technique. The new method is also applied to a real data example and it shows a reasonable result for model selection. To further reduce the unexpectedly large variation in AIC, we also propose a generalized AIC (GAIC) method for model selection. Through real data analysis and simulation studies, it is evident that the GAIC method performs better than the AIC method in most settings, and so can be routinely recommended for practical use. Specifically, to apply GAIC for model selection, we first compute the difference between GAIC_{FEM} and GAIC_{REM} as

$$\text{GAIC}_{\text{FEM}} - \text{GAIC}_{\text{REM}} = k - 1 + \sum_{i=1}^{k} \ln \frac{\sigma_i^2}{\sigma_i^2 + \hat{\tau}^2_{\text{DL}}},$$

where k is the number of studies, σ_i^2 are the within-study variances, and $\hat{\tau}^2_{\text{DL}} = \max\{0, \{Q - (k-1)\}/(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2 / \sum_{i=1}^k w_i)\}$ is the moment estimate in [10]. We then select FEM to perform meta-analysis if $\text{GAIC}_{\text{FEM}} - \text{GAIC}_{\text{REM}} < 0$; but if not, then REM will be applied. To conclude the paper, we hope to reiterate that this is the first work in meta-analysis for model selection between the fixed-effects model and the random-effects model, and we expect that our proposed GAIC criterion will have the potential to be widely applied in meta-analysis and evidence-based medicine.

There are some future directions related to the current work. Firstly, Bayesian Information Criterion (BIC) can be another potential approach for model selection between FEM and REM, and following its establishment, a comparison on the performance of model selection based on AIC and BIC is also needed. In addition, following our new method, the AIC value of CEM can also be derived as

$$\text{AIC}_{\text{CEM}} = \sum_{i=1}^{k} \left\{ \ln\left(2\pi\sigma_i^2\right) + \frac{\left(y_i - \hat{\theta}_{\text{CEM}}\right)^2}{\sigma_i^2} \right\} + 2.$$

That is, our new method can also be applied for model selection between CEM and REM. Hence as another future work, we will also revisit the model selection between CEM and REM, conduct extensive simulation studies to compare the performance of the Q statistic, the I^2 statistic and the AIC criterion, and make some new recommendations for meta-analysis.

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and two reviewers for their constructive comments that have led to a substantial improvement of the paper. Hiu-Yee Kwan's research was supported by Research Grant Council (No. HKBU-22103017-ECS) and Natural Science Foundation of Guangdong Province (No. 2018A0303130122). Tiejun Tong's research was supported by the Initiation Grant for Faculty Niche Research Areas (No. RC-IG-FNRA/17-18/13) and the Century Club Sponsorship Scheme of Hong Kong Baptist University, General Research Fund (No. HKBU12303918) and National Natural Science Foundation of China (No. 11671338).

Received 11 February 2020

REFERENCES

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723. MR0423716

[2] Bender, R., Friede, T., Koch, A., Kuss, O., Schlattmann, P., Schwarzer, G. and Skipka, G. (2018). Methods for evidence synthesis in the case of very few studies. *Research Synthesis Methods* **9** 382–392.

[3] Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review* **40** 207–227.

[4] Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester: John Wiley & Sons.

[5] Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods* **1** 97–111.

[6] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd Edition*. New York: Springer. MR1919620

[7] Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10** 101–129. MR0067428

[8] Davey, J., Turner, R. M., Clarke, M. J. and Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology* **11** 160.

[9] Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies* **47** 108–121.

[10] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7** 177–188.

[11] Eddy, D. M. (1990). Practice policies – guidelines for methods. *Journal of the American Medical Association* **263** 1839–1841.

[12] Egger, M. and Smith, G. D. (1997). Meta-analysis: potentials and promise. *British Medical Journal* **315** 1371–1374.

[13] Estabrooks, C. A. (1998). Will evidence-based nursing practice make practice perfect? *Canadian Journal of Nursing Research Archive* **30** 15–36.

[14] Fellmeth, G., Rose-Clarke, K., Zhao, C. et al. (2018). Health impacts of parental migration on left-behind children and adolescents: a systematic review and meta-analysis. *The Lancet* **392** 2567–2582.

[15] Friede, T., Röver, C., Wandel, S. and Neuenschwander, B. (2017). Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical Journal* **59** 658–671. MR3672688

[16] Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher* **5** 3–8.

[17] Guolo, A. and Varin, C. (2017). Random-effects meta-analysis: the number of studies matters. *Statistical Methods in Medical Research* **26** 1500–1518. MR3661007

[18] Guyatt, G., Cairns, J., Churchill, D. et al. (1992). Evidence-based medicine: a new approach to teaching the practice of medicine. *Journal of the American Medical Association* **268** 2420–2425.

[19] Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal* **41** 901–916. MR1747520

[20] Hedges, L. V. and Olkin, I. (2014). *Statistical Methods for Meta-Analysis, 2nd Edition*. Orlando: Academic Press. MR0798597

[21] Hedges, L. V. and Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods* **6** 203–217.

[22] Hedges, L. V. and Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods* **3** 486–504.

[23] Higgins, J. P. and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester: John Wiley & Sons.

[24] Higgins, J. P. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21** 1539–1558.

[25] Higgins, J. P., Thompson, S. G., Deeks, J. J. and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal* **327** 557–560.

[26] James, G. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika* **38** 324–329. MR0046616

[27] Knapp, G. and Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* **22** 2693–2710.

[28] Kulinskaya, E., Dollinger, M. B. and Bjørkestøl, K. (2011). On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Research Synthesis Methods* **2** 254–270. MR2733443

[29] Laird, N. M. and Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care* **6** 5–30.

[30] Melnyk, B. M. and Fineout-Overholt, E. (2011). *Evidence-Based Practice in Nursing & Healthcare: A Guide to Best Practice, 2nd Edition*. Philadelphia: Lippincott Williams & Wilkins.

[31] Pawson, R. (2006). *Evidence-Based Policy: A Realist Perspective*. London: SAGE Publications.

[32] Rice, K., Higgins, J. P. and Lumley, T. (2018). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A* **181** 205–227. MR3749516

[33] Rosenberg, W. and Donald, A. (1995). Evidence based medicine: an approach to clinical problem-solving. *British Medical Journal* **310** 1122–1126.

[34] Sanderson, I. (2002). Evaluation, policy learning and evidence-based policy making. *Public Administration* **80** 1–22.

[35] Thomas, G. and Pring, R. (2004). *Evidence-Based Practice in Education*. Berkshire: Open University Press.

[36] Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika* **38** 330–336. MR0046617

Ke Yang
Department of Mathematics
Hong Kong Baptist University
Hong Kong
China
E-mail address: yangke18@life.hkbu.edu.hk

Hiu-Yee Kwan
School of Chinese Medicine
Hong Kong Baptist University
Hong Kong
China
E-mail address: hykwan@hkbu.edu.hk

Zhiling Yu
School of Chinese Medicine
Hong Kong Baptist University
Hong Kong
China
E-mail address: zlyu@hkbu.edu.hk

Tiejun Tong
Department of Mathematics
Hong Kong Baptist University
Hong Kong
China
E-mail address: tongt@hkbu.edu.hk