ARTICLE TEMPLATE

Statistical inference on mixed one- and two-armed studies in meta-analysis without study-specific variance

Patarawan Sangnawakij[a] and Dankmar Böhning[b]

[a]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand; [b]Mathematical Sciences and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

ARTICLE HISTORY
Compiled October 14, 2021

ABSTRACT
In some meta-analytic data constellations, only the quantity of interest and sample size are available from the published reports. In addition, for some individual studies, this partial information is available for only one of two treatment groups. These are typically excluded from the meta-analysis, whereas in fact, it would be preferable to include such studies. The current paper proposes an approach for estimating the parameter of interest when study-specific variance is not included in the study information and potentially only one arm information is presented. The approach we propose allows the full set of individual studies to be analyzed. The joint likelihoods included missing case modelling is used to estimate the mean difference and variance using a fixed effect model. In simulations, we evaluate the performance of the estimators in terms of bias and standard deviation, and compare the results with those from an existing method but using only studies in which information is available in both treatment arms. The coverage probability is also computed to investigate the efficiency of the confidence intervals. Our estimators derived under the homogeneity model show better performance than the existing method when estimating the mean difference and related variance. They are also useful for estimating the mean difference parameter under several heterogeneity scenarios: baseline heterogeneity but no effect heterogeneity, as well as under baseline heterogeneity jointly with effect heterogeneity across studies. We apply our method to a meta-analysis of clinical study data, and demonstrate its practicality.

KEYWORDS
Mixed information; mean difference; meta-analysis; missing case modelling; missing variance.

1. Introduction

Meta-analysis is a powerful statistical method used to analyze and integrate results obtained from multiple independent, individual studies on the same topic. In recent years, meta-analysis has been developed in theory and applied in areas such as medical science, epidemiology, clinical studies, and social science. The standard meta-analytic inference uses the inverse variance-weighted average method to estimate the overall effect size, such as the mean, risk difference, or risk ratio [**?** **?** **?**]. This approach requires the quantity of interest together with the variance or standard error of the

CONTACT Patarawan Sangnawakij. Email: patarawan.s@gmail.com

effect size for each study, as minimal information. A problem arises in cases in which the study-specific variances or other variability measures, particularly the standard deviation, coefficient of variation, or standard error, are not available [**?**]. As also noted in Batson and Burton [**?**], the associated variance data which are not reported in many publications is a major problem for meta-analysis of continuous outcomes. A standard approach as developed so far is therefore incomplete. A number of approaches for dealing with this issue have been proposed. Marinho et al. [**?**] introduced a method of estimation for the missing standard deviations using simple linear regression of log-transformation, including study-level covariates. Hozo et al. [**?**] proposed a simple method based on median, range, and sample size for estimation of sample variance. Idris and Robertson [**?**] suggested a hot-deck approach to imputing the missing variances. Steven [**?**] introduced a Baysian-based imputation method to deal with the missing values. Chowdhry et al. [**?**] imputed the missing sample variances using gamma meta-regression with covariates. Alternative approaches to the imputation of missing variability for certain studies have been discussed as in Wiebe et al. [**?**] and Philbrook et al. [**?**]. Another common approach often used in applications is excluding studies with missing study-specific variance from the analysis. However, the latter method should be applied with caution, because it can introduce bias in estimation [**?** **?**].

Table 1. Mean of the length of hospital stay in days and number of patients in thoracoscopic and open operations.

		Thoracoscopic		Open	
Study	Authors, year	Size	Mean	Size	Mean
1	Vu, 2008	12	2	24	5
2	Diamond, 2007	12	3.5	24	4
3	Kunisaki, 2014	49	3	13	3
4	Lau, 2013	39	6.95	28	11.96
5	Rahman, 2009	14	2.95	14	2.6
6	Cho, 2012	7	6.1	27	8.1
7	Tolg, 2005	5	6	4	12
8	Fascetti-Leon, 2013	26	5.3	28	9.6
9	Laje, 2015	100	3	188	3.1
10	Kulaylat, 2015	112	3	146	4

Comparison of the means in two treatment groups is often the key goal in practical applications. When comparing two continuous variables, the most commonly used measure is the difference in means. Rukhin [**?**] introduced the common mean of normal observations when sample variances were missing. The prior distribution of the weight of effect size was presented as the geometric mean of the distances between data points. Bayes estimation was used to estimate the mean and within-study variance in the meta-analytic model. Furthermore, a statistical method for estimating the mean difference in meta-analysis was introduced in Sangnawakij et al. [**?**]. They derived the estimators in the fixed effect meta-analysis using maximum likelihood (ML) estimation based on the studies with information on mean and sample size in both treatment arms. Estimation of parameters in the random effects model was reported in Sangnawakij et al. [**?**]. The meta-analytic data that motivated their works were obtained from Adams et al. [**?**], where they had only sample means and sample sizes. An example endpoint of these data is shown in Table 1. However, in fact, there were 21 studies ignored in the analysis, and were not presented in the previous work. From these reports, 14 had

only information on treatment arm, and seven had only information on comparison arm. We present the additional data in Tables 2 and 3. Note that the details of these data will be given in the next subsection.

Table 2. Mean of the length of hospital stay in days and number of patients in thoracoscopic only.

		Thoracoscopic		Open	
Study	Authors, year	Size	Mean	Size	Mean
1	Jesch, 2005	5	7	-	-
2	de Lagausie, 2005	8	4.6	-	-
3	Tanaka, 2013	12	5.8	-	-
4	Rothenberg, 2008	97	2.4	-	-
5	Rothenberg, 2011	75	2.4	-	-
6	Seong, 2013	50	6	-	-
7	Muller, 2012	12	4	-	-
8	Tarrado, 2010	6	3.5	-	-
9	Truitt, 2006	12	2	-	-
10	Zeidan, 2009	6	4.5	-	-
11	Cano, 2006	6	7	-	-
12	Boubnova, 2011	30	8	-	-
13	Albanese, 2007	144	2.8	-	-
14	Johnson, 2011	15	4.2	-	-

Table 3. Mean of the length of hospital stay in days and number of patients in open surgery only.

		Thoracoscopic		Open	
Study	Authors, year	Size	Mean	Size	Mean
1	Tsai, 2008	-	-	105	3
2	Raychaudhuri, 2011	-	-	14	7.2
3	Sueyoshi, 2008	-	-	8	15
4	Aziz, 2004a	-	-	6	3
5	Aziz, 2004b	-	-	3	7
6	Aziz, 2004c	-	-	9	4
7	Ferreira, 2010	-	-	35	11

When using the exclusion method, this could affect the bias and efficiency of the estimators. Therefore, the objective of this study is to develop the mean difference and variance estimation in meta-analysis with one-armed studies and missing variance information. This is, in fact, the more relevant and practical case. We propose an alternative approach in which ML estimation is applied to the full set of individual studies or *mixed one- and two-armed studies* within the meta-analysis, for estimating the related parameters in the homogenous model. Thus, it is clear that the use-all-data approach for deriving the estimators applied here differs from the previous works which focus only on the studies with complete information in both arms. The proposed estimators from the novel method are investigated through simulations, and compared with an existing method obtained from the both arm information. The coverage probability and expected length of the proposed confidence interval for the mean difference are also considered. Finally, a real data example on clinical outcomes is presented to demonstrate our approach.

1.1. Case study

This work is motivated by a meta-analysis using reported data involving 31 studies conducted between 2004 and 2015. These independent studies are concerned with the differences between two surgical methods, open surgery and thoracoscopic or key-hole surgery, for removing asymptomatic congenital lung malformations in young children. This condition is observed in 1 from 2500 live born children in routine antenatal scans [**?** **?**]. The open surgical procedure is the traditional method of excising such lesions. However, this requires a large incision at the side of the chest. The late 1990s, a minimally invasive approach called thoracoscopy or key-hole surgery has become popular as an alternative. This involves making small incision through which a fibre-optic camera and operating instruments can be passed. Although this approach is believed to benefit patients, it is not clear whether this surgery is associated with an increased risk of post operative complications. Several research studies have compared the outcomes of the two methods in asymptomatic cases, based on the number of days that the patient remained in hospital, length of time under surgery, and weight of the patient. Further details can be also found in Adams et al. [**?**]. However, as shown in Tables 1-3, the clinical endpoints of interest are reported only in terms of sample means and number of patients. None of the studies reported sample variance. We point out that these were not clinical trials but rather a collection of study reports as clinical trials with newborns are not available and also ethically problematic. Nevertheless, also these data require attention for a profound analysis within the given circumstances. In the current study, meta-analytic data with lacking study-specific variance information are analyzed using our novel approach, under the assumption of equal variance across all studies.

1.2. Fixed effect model in meta-analysis

Standard meta-analysis is briefly reviewed in this section, starting with the general case. Let $\hat{\theta}_i$, for $i = 1, 2, ..., k$, be the effect size estimates from k independent studies. The random effects model allows the true effect size for each study to vary across studies. This typically takes the form $\hat{\theta}_i = \theta + \xi_i + \varepsilon_i = \theta_i + \varepsilon_i$, where θ is the overall effect size, $\xi_i \sim N(0, \tau^2)$ is a random effect size, $\varepsilon_i \sim N(0, \sigma_i^2)$ is a random sampling error, τ^2 is the between-study variance, and σ_i^2 is the within-study variance. The model treats θ_i, the true effect size of the study i, as a random sample from its effect size distribution. Thus, each study has different effect size, which is known as the heterogeneity case [**?**].

When a single effect measure assumes homogeneity, the results of individual studies differ only by chance [**?**]. In this case, the true effects are equal in all the studies and non-random, $\theta_i = \theta$, for $i = 1, 2, ..., k$. This is a special case of the heterogeneity variance in which $\tau^2 = 0$. Thus, the fixed effect model can be written as

$$\hat{\theta}_i = \theta + \varepsilon_i,$$

where $\hat{\theta}_i \sim N(\theta, \sigma_i^2)$. From the above model, we wish to estimate the overall effect size θ. The most common method uses the weighted average of effect sizes together with the variance estimates. It follows that the unbiased estimator for θ is given by

$$\hat{\theta} = \sum_{i=1}^{k} v_i \hat{\theta}_i,$$

where $v_i = \frac{1/\widehat{Var}(\hat{\theta}_i)}{\sum_{i=1}^{k} 1/\widehat{Var}(\hat{\theta}_i)}$ is the weight of effect size in the study i, and the variance estimate of effect estimate is given by $\widehat{Var}(\hat{\theta}_i) = \hat{\sigma}_i^2$ [**?** **?**].

It is important to note that meta-analyses based on the standard method discussed in this section require the estimated variances and two-armed information to be available. Unfortunately, those variances are sometimes missing in all studies. An alternative method that does not require reported study-specific variance is therefore essential. This will be addressed in the next section.

2. Statistical method

2.1. Notation

In this section, we consider a meta-analysis of k independent studies in which only the sample mean ($\bar{X}_i$) and sample size (n_i) are given for the treatment and control groups. Moreover, some studies report this information for only one of two arms. No information on sample variance or standard deviation is given. The layout of the data considered in this study is shown in Table 4. From the k studies, k_0 studies report sample means and sample sizes in both arms, whereas k_1 studies report one arm information only for treatment and k_2 studies only for control.

Table 4. Layout of meta-analytic data reported from two- and one-armed studies.

Study	Treatment group		Control group	
i	n_i^T	$\bar{X}_i^T$	n_i^C	$\bar{X}_i^C$
1	n_1^T	$\bar{X}_1^T$	n_1^C	$\bar{X}_1^C$
2	n_2^T	$\bar{X}_2^T$	n_2^C	$\bar{X}_2^C$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k_0	$n_{k_0}^T$	$\bar{X}_{k_0}^T$	$n_{k_0}^C$	$\bar{X}_{k_0}^C$
k_0+1	$n_{k_0+1}^T$	$\bar{X}_{k_0+1}^T$		
k_0+2	$n_{k_0+2}^T$	$\bar{X}_{k_0+2}^T$		
$\vdots$	$\vdots$	$\vdots$		
k_0+k_1	$n_{k_0+k_1}^T$	$\bar{X}_{k_0+k_1}^T$		
k_0+k_1+1			$n_{k_0+k_1+1}^C$	$\bar{X}_{k_0+k_1+1}^C$
k_0+k_1+2			$n_{k_0+k_1+2}^C$	$\bar{X}_{k_0+k_1+2}^C$
$\vdots$			$\vdots$	$\vdots$
k			n_k^C	$\bar{X}_k^C$

We assume that the within-study values $X_{ij}^T \sim N(\mu+\delta, \sigma^2)$ and $X_{ij}^C \sim N(\mu, \sigma^2)$, where X_{ij}^T and X_{ij}^C are independent. $\delta = \mu^T - \mu$ is the mean difference between treatment and control groups. The sample means which are effect measure estimates are denoted as $\bar{X}_i^T = \sum_{j=1}^{n_i^T} X_{ij}^T/n_i^T$ and $\bar{X}_i^C = \sum_{j=1}^{n_i^C} X_{ij}^C/n_i^C$. Here, only $\bar{X}_i^T$ and $\bar{X}_i^C$ are observed, but not the within-study values.

Suppose that the true effect size is the same across the studies. The fixed effect models for the means are therefore given by $\bar{X}_i^T = \mu+\delta+\varepsilon_i^T$ and $\bar{X}_i^C = \mu+\varepsilon_i^C$, where $\varepsilon_i^T \sim N(0, \sigma^2 u_i^T)$ and $\varepsilon_i^C \sim N(0, \sigma^2 u_i^C)$ are random errors, being independent, and $u_i^T = 1/n_i^T$ and $u_i^C = 1/n_i^C$ are non-random. From this assumption, it follows that $\bar{X}_i^T \sim N(\mu+\delta, \sigma^2 u_i^T)$ and $\bar{X}_i^C \sim N(\mu, \sigma^2 u_i^C)$. Thus, the probability density functions

of $\bar{X}_i^T$ and $\bar{X}_i^C$ are denoted as

$$f(\bar{x}_i^T;\mu,\delta,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2 u_i^T}}\exp\left(-\frac{(\bar{x}_i^T-(\mu+\delta))^2}{2\sigma^2 u_i^T}\right)$$

and

$$f(\bar{x}_i^C;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2 u_i^C}}\exp\left(-\frac{(\bar{x}_i^C-\mu)^2}{2\sigma^2 u_i^C}\right),$$

respectively. However, it is assumed here that k_1 studies out of k are missing at random in the treatment arm and k_2 studies out of k studies are missing at random in the control arm. Estimation of parameters in the meta-analysis model using typical ML estimation as introduced in Sangnawakij et al. [**?**] cannot be applied, because this method was based on the studies with information in both arms only. The novel method in which the likelihood function depending on the entire information available is therefore discussed in this paper.

2.2. Parameter estimation

According to the data in the light of the information available given in Table 4, we have the following three different likelihood functions. The first likelihood based on the studies with both arm information is given by

$$L_0=\prod_{i=1}^{k_0} f(\bar{x}_i^T;\mu,\delta,\sigma^2)f(\bar{x}_i^C;\mu,\sigma^2).$$

The second likelihood appears for the studies with only information in treatment group:

$$L_1=\prod_{i=k_0+1}^{k_0+k_1} f(\bar{x}_i^T;\mu,\delta,\sigma^2).$$

Finally, we have the likelihood for the studies with only information in control group:

$$L_2=\prod_{i=k_0+k_1+1}^{k} f(\bar{x}_i^C;\mu,\sigma^2).$$

Thus, the joint likelihood is given by $L(\theta)=L_0L_1L_2$, with the log-likelihood

$$\ln L(\theta)=\ln L_0+\ln L_1+\ln L_2, \tag{1}$$

where $\theta = (\mu, \delta, \sigma^2)$ is the parameter involved in the models,

$$\begin{aligned}
\ln L_0 &= -\frac{1}{2\sigma^2}\sum_{i=1}^{k_0}\left(\frac{(\bar{x}_i^T-(\mu+\delta))^2}{u_i^T}+\frac{(\bar{x}_i^C-\mu)^2}{u_i^C}\right)-\sum_{i=1}^{k_0}\ln(2\pi\sigma^2)-\sum_{i=1}^{k_0}\ln\sqrt{u_i^T u_i^C}\\
\ln L_1 &= -\frac{1}{2\sigma^2}\sum_{i=k_0+1}^{k_0+k_1}\frac{(\bar{x}_i^T-(\mu+\delta))^2}{u_i^T}-\frac{k_1\ln(2\pi\sigma^2)}{2}-\sum_{i=k_0+1}^{k_0+k_1}\ln\sqrt{u_i^T}\\
\ln L_2 &= -\frac{1}{2\sigma^2}\sum_{i=k_0+k_1+1}^{k}\frac{(\bar{x}_i^C-\mu)^2}{u_i^C}-\frac{k_2\ln(2\pi\sigma^2)}{2}-\sum_{i=k_0+k_1+1}^{k}\ln\sqrt{u_i^C}.
\end{aligned}$$

Substituting $\ln L_0$, $\ln L_1$, and $\ln L_2$ into Equation (1), we have

$$\begin{aligned}
\ln L(\theta) &= -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{k_0}\frac{(\bar{x}_i^T-(\mu+\delta))^2}{u_i^T}+\sum_{i=1}^{k_0}\frac{(\bar{x}_i^C-\mu)^2}{u_i^C}+\sum_{i=k_0+1}^{k_0+k_1}\frac{(\bar{x}_i^T-(\mu+\delta))^2}{u_i^T}+\right.\\
&\left.\sum_{i=k_0+k_1+1}^{k}\frac{(\bar{x}_i^C-\mu)^2}{u_i^C}\right)-k_0\ln(2\pi\sigma^2)-\frac{k_1\ln(2\pi\sigma^2)}{2}-\frac{k_2\ln(2\pi\sigma^2)}{2}+c, \quad (2)
\end{aligned}$$

where c is the constant. To find the estimator of μ, we take the partial derivative of Equation (2) with respect to μ. This can be written as

$$\begin{aligned}
\frac{\partial}{\partial\mu}\ln L(\theta) &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{k_0}\frac{\bar{x}_i^T-(\mu+\delta)}{u_i^T}+\sum_{i=1}^{k_0}\frac{\bar{x}_i^C-\mu}{u_i^C}+\sum_{i=k_0+1}^{k_0+k_1}\frac{\bar{x}_i^T-(\mu+\delta)}{u_i^T}+\right.\\
&\left.\sum_{i=k_0+k_1+1}^{k}\frac{\bar{x}_i^C-\mu}{u_i^C}\right).
\end{aligned}$$

Setting the derivative of the log-likelihood function equal to zero, the expression for the ML estimator for μ, the mean in control group, is given as

$$\mu = \frac{\sum_{i=1}^{k_0}\left(\frac{\bar{x}_i^T-\delta}{u_i^T}+\frac{\bar{x}_i^C}{u_i^C}\right)+\sum_{i=k_0+1}^{k_0+k_1}\frac{\bar{x}_i^T-\delta}{u_i^T}+\sum_{i=k_0+k_1+1}^{k}\frac{\bar{x}_i^C}{u_i^C}}{\sum_{i=1}^{k_0}\left(\frac{1}{u_i^T}+\frac{1}{u_i^C}\right)+\sum_{i=k_0+1}^{k_0+k_1}\frac{1}{u_i^T}+\sum_{i=k_0+k_1+1}^{k}\frac{1}{u_i^C}}. \quad (3)$$

Similarly, the partial derivatives with respect to δ and σ^2 are taken of Equation (2) and set to zero. Therefore, the estimated parameters for δ and σ^2 are expressed by

$$\delta = \frac{\sum_{i=1}^{k_0}\frac{\bar{x}_i^T-\mu}{u_i^T}+\sum_{i=k_0+1}^{k_0+k_1}\frac{\bar{x}_i^T-\mu}{u_i^T}}{\sum_{i=1}^{k_0}\frac{1}{u_i^T}+\sum_{i=k_0+1}^{k_0+k_1}\frac{1}{u_i^T}} \quad (4)$$

and

$$\sigma^2 = \frac{\sum_{i=1}^{k_0}\left(\frac{(\bar{x}_i^T-(\mu+\delta))^2}{u_i^T}+\frac{(\bar{x}_i^C-\mu)^2}{u_i^C}\right)+\sum_{i=k_0+1}^{k_0+k_1}\frac{(\bar{x}_i^T-(\mu+\delta))^2}{u_i^T}+\sum_{i=k_0+k_1+1}^{k}\frac{(\bar{x}_i^C-\mu)^2}{u_i^C}}{2k_0+k_1+k_2}. \quad (5)$$

As can be seen from Equations (3-5), the ML estimates do not provide closed-form solutions for μ, δ, and σ^2 that can be calculated directly. However, these can be approximated using a numerical method [**?**]. This is based on the following, reliably converging algorithm:

i) choose starting values for μ, δ, and σ^2
ii) given δ and σ^2, use Equation (3) to get a new value for μ
iii) given the new value for μ and current value for σ^2, find a new value for δ using Equation (4)
iv) given the new values for μ and δ, find a new value for σ^2 using Equation (5).

The procedure is iterated until all estimated parameters converge, so $|\mu^{(s+1)} - \mu^{(s)}|$, $|\delta^{(s+1)} - \delta^{(s)}|$, and $|\sigma^{2^{(s+1)}} - \sigma^{2^{(s)}}|$ are less than or equal to ϵ, where s is the iteration and ϵ is the difference of parameter estimates between the sth and $(s+1)$th iteration, or acceptable error, and is a small value. The iterative scheme is similar to the EM algorithm [**?**] in the sense that it increases the likelihood at every iteration, and hence produces monotonically converging estimates. A reasonable starting value for δ could be the no difference setting. this then would produce starting values for μ according to (3) and for σ^2 according to (5). In simulation studies one could take the true values as starting values. The (1-α)100% confidence interval for δ is computed by $\hat{\delta} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\delta})}$, where $\hat{\delta}$ is the ML estimator for δ, $z_{\alpha/2}$ is the $(\alpha/2)100$th percentile of the standard normal distribution, and $\widehat{Var}(\hat{\delta})$ is the estimated variance of $\hat{\delta}$.

In Sangnawakij et al. [**?**] the mean difference and variance estimates in meta-analysis based on k_0 studies with information in both arms were derived using the typical ML estimation. Those estimators are given by

$$\hat{\delta}_B = \frac{\sum_{i=1}^{k_0} \frac{\bar{X}_i^T - \bar{X}_i^C}{u_i^T + u_i^C}}{\sum_{i=1}^{k_0} \frac{1}{u_i^T + u_i^C}} \quad \text{and} \quad \hat{\sigma}_B^2 = \frac{1}{k_0}\sum_{i=1}^{k_0} \frac{((\bar{X}_i^T - \bar{X}_i^C) - \hat{\delta}_B)^2}{u_i^T + u_i^C}$$

with $Var(\hat{\delta}_B) = \sigma_B^2 / \sum_{i=1}^{k_0} 1/(u_i^T + u_i^C)$ and $Var(\hat{\sigma}_B^2) = 2(k_0 - 1)\sigma_B^4/k_0^2$. Note that these estimates are closed-form and do not require any iteration.

To estimate the variance of the estimators derived in the previous section, the approximate method using the inverse of Fisher information is applied. This is related to the Cramer-Rao lower bound, which represents the efficiency of the estimator [**? ?**]. As there are three parameters of interest, the symmetric matrix of Fisher information for $\theta = (\mu, \delta, \sigma^2)$ is given by

$$I(\theta) = -\begin{bmatrix} E\left(\frac{\partial^2}{\partial\mu^2}\ln L(\theta)\right) & E\left(\frac{\partial^2}{\partial\mu\partial\delta}\ln L(\theta)\right) & E\left(\frac{\partial^2}{\partial\mu\partial\sigma^2}\ln L(\theta)\right) \\ E\left(\frac{\partial^2}{\partial\mu\partial\delta}\ln L(\theta)\right) & E\left(\frac{\partial^2}{\partial\delta^2}\ln L(\theta)\right) & E\left(\frac{\partial^2}{\partial\delta\partial\sigma^2}\ln L(\theta)\right) \\ E\left(\frac{\partial^2}{\partial\mu\partial\sigma^2}\ln L(\theta)\right) & E\left(\frac{\partial^2}{\partial\delta\partial\sigma^2}\ln L(\theta)\right) & E\left(\frac{\partial^2}{\partial(\sigma^2)^2}\ln L(\theta)\right) \end{bmatrix}.$$

Note that $\ln L(\theta)$ is obtained from Equation (2). The asymptotic variances and covariances of the proposed estimators, $\hat{\mu}$, $\hat{\delta}$, and $\hat{\sigma}^2$, can be obtained from the inverse

matrix of $I(\theta)$:

$$I^{-1}(\theta) \quad = \quad \begin{bmatrix} Var(\hat{\mu}) & Cov(\hat{\mu},\hat{\delta}) & Cov(\hat{\mu},\hat{\sigma}^2) \\ Cov(\hat{\mu},\hat{\delta}) & Var(\hat{\delta}) & Cov(\hat{\delta},\hat{\sigma}^2) \\ Cov(\hat{\mu},\hat{\sigma}^2) & Cov(\hat{\delta},\hat{\sigma}^2) & Var(\hat{\sigma}^2), \end{bmatrix},$$

where

$$\begin{aligned} Var(\hat{\mu}) &= \frac{\sigma^2}{\sum_{i=1}^{k_0}\frac{1}{u_i^C}+\sum_{i=k_0+k_1+1}^{k}\frac{1}{u_i^C}}, \\ Var(\hat{\delta}) &= \frac{\sigma^2\left(\sum_{i=1}^{k_0}\left(\frac{1}{u_i^T}+\frac{1}{u_i^C}\right)+\sum_{i=k_0+1}^{k_0+k_1}\frac{1}{u_i^T}+\sum_{i=k_0+k_1+1}^{k}\frac{1}{u_i^C}\right)}{\left(\sum_{i=1}^{k_0}\frac{1}{u_i^T}+\sum_{i=k_0+1}^{k_0+k_1}\frac{1}{u_i^T}\right)\left(\sum_{i=1}^{k_0}\frac{1}{u_i^C}+\sum_{i=k_0+k_1+1}^{k}\frac{1}{u_i^C}\right)}, \\ Var(\hat{\sigma}^2) &= \frac{2\sigma^4}{2k_0+k_1+k_2}, \\ Cov(\hat{\mu},\hat{\delta}) &= -Var(\hat{\mu}) \quad \text{and} \quad Cov(\hat{\mu},\hat{\sigma}^2) = Cov(\hat{\delta},\hat{\sigma}^2) = 0. \end{aligned} \tag{6}$$

The standard errors of $\hat{\mu}$, $\hat{\delta}$, and $\hat{\sigma}^2$ are obtained by taking the square roots of the main diagonal elements of $I^{-1}(\theta)$. Hence, $\hat{\delta}$ has the asymptotic normal distribution with mean δ and variance $Var(\hat{\delta})$. To calculate the standard error in interval estimation, we use the estimated inverse Fisher information, where σ^2 is replaced by $\hat{\sigma}^2$.

3. Numerical study

3.1. Simulation

The method based on modelling of mixed one- and two-armed studies presented in this paper was investigated using simulations, via the `R` statistical language [**?**]. The estimators obtained from this approach were also compared with those from the both arm information method. In simulations, $\bar{X}_i^T$ and $\bar{X}_i^C$ were sampled from $N(\mu+\delta, \sigma^2 u_i^T)$ and $N(\mu, \sigma^2 u_i^C)$, respectively, where the parameter values δ and μ were set to 0, 2, and 10, and σ^2 to 2, 4, and 9. The within-study variances were assumed to be dependent on the sample size of the study, so that u_i^T and u_i^C were sampled from a uniform distribution on $(0.02, 0.20)$, corresponding to the study sample size on the interval (5, 50). We set k as 30, 50, and 100, and also considered k_0, k_1, and k_2. The number of missing values m was set at 25%, 50%, and 75% of the number of the studies k. For example, let $k = 100$ and $m = 25\%$, we give $k_0 = 75$, $k_1 = 13$, and $k_2 = 12$. As noted in Section 2, the ML estimators were computed using an algorithm-based method. The algorithm was iterated until the convergence criteria were satisfied for all three parameters: $|\theta^{(s+1)} - \theta^{(s)}| \leq 0.00001$, where θ is the generic parameter (μ, δ, or σ^2). The iterations ended if the estimated value converged to a constant. Each simulation was repeated 10,000 times. The mean of the parameter estimate was computed as follows:

$$\widehat{E(\hat{\theta})} = \frac{1}{10,000}\sum_{i=1}^{10,000}\hat{\theta}_i.$$

The performance of the estimator was judged by the bias and empirical standard error of estimator. These were approximated by

$$\text{Bias}(\hat{\theta}) = \widehat{E(\hat{\theta})} - \theta$$

and

$$\text{SE}(\hat{\theta}) = \sqrt{\frac{1}{10,000} \sum_{i=1}^{10,000} (\hat{\theta}_i - \widehat{E(\hat{\theta})})^2}, \tag{7}$$

where $\hat{\theta}$ is the generic estimator for a parameter θ. If the point estimator has a small bias and standard error, this indicates that the estimated value is, on average, close to the true value. The coverage probability is a measure of calibration defined as the probability that the confidence interval covers the parameter. It can be calculated by

$$\text{CP} = \frac{c(L \leq \theta \leq U)}{10,000},$$

where $c(L \leq \theta \leq U)$ is the number of simulation runs with θ being within the lower and upper limits, L and U. The expected length is estimated by

$$\text{EL} = \frac{1}{10,000} \sum_{i=1}^{10,000} (U_i - L_i).$$

In general, we choose a confidence interval which has a coverage probability greater than or close to the nominal coverage level, and then a short width interval. In this study, the confidence level was 0.95.

The estimated mean difference and the estimated variance from the method presented in this paper are denoted as $\hat{\delta}_P$ and $\hat{\sigma}^2_P$, and those from the both arm method as $\hat{\delta}_B$ and $\hat{\sigma}^2_B$. The key results from the simulations are as follows.

i) From Figure 1, the biases of the mean difference estimators $\hat{\delta}_P$ and $\hat{\delta}_B$ were similar, though that of $\hat{\delta}_P$ was slightly smaller (bias was closer to zero) in many cases. These biases did not depend on the number of missing value m or number of studies k.

ii) The sample variances $\hat{\sigma}^2_P$ and $\hat{\sigma}^2_B$ produced slight underestimates in all cases. However, the bias of $\hat{\sigma}^2_P$ was closer to zero and smaller than that of $\hat{\sigma}^2_B$ in all cases. The bias of $\hat{\sigma}^2_P$ increased as m increased, but decreased if k became large.

iii) From Figure 2, the standard errors of $\hat{\delta}_P$ and $\hat{\sigma}^2_P$ were smaller than that of $\hat{\delta}_B$ and $\hat{\sigma}^2_B$, respectively, particularly when m was large. The standard errors of $\hat{\delta}_P$ and $\hat{\sigma}^2_P$ decreased if k was increased.

iv) The coverage probability of the confidence interval for δ is given in Figure 3. It can be seen that the confidence interval obtained from the proposed method was greater than that of the comparison for all cases, and close to the nominal coverage level at 0.95. However, the coverage probability of the confidence interval from the both arm method was close to 0.95 when $k = 100$ and $m = 25\%$ or 50% only.

v) As can be seen from Figure 3, the expected length of the confidence interval given by the proposed method was slightly shorter than that of the both arm method. Significant differences were observed between the expected lengths of the two interval estimators, when m was greater than 50%.

Table 5. The performance of estimators under baseline heterogeneity using simulations ($\delta = 0$).

Sample size	m	k	Bias of $\hat{\delta}$	Bias of $\hat{\mu}$	$\hat{\sigma}^2$	$SE_1(\hat{\delta})$	$SE_2(\hat{\delta})$	Coverage probability	Expected length
Only baseline heterogeneity: $(\tau^2, \tau^{2C}) = (0, 4)$									
(3, 10)	25%	10	0.0043	0.0001	16.8127	0.8990	0.9413	0.9779	3.6899
		30	0.0040	-0.0056	21.8508	0.3704	0.5654	0.9881	2.2162
		50	0.0020	-0.0004	20.0208	0.3284	0.4381	0.9859	1.7175
		100	0.0025	-0.0007	21.6206	0.2287	0.3166	0.9885	1.2411
	50%	10	-0.0067	-0.0066	19.9004	1.0068	0.9881	0.9698	3.8735
		30	-0.0017	-0.0042	20.6105	0.4939	0.5760	0.9729	2.2578
		50	0.0029	-0.0008	22.7468	0.3518	0.4585	0.9838	1.7972
		100	-0.0012	0.0021	22.1769	0.2744	0.3205	0.9777	1.2564
(5, 50)	25%	10	0.0040	-0.0037	39.6192	0.9860	0.8577	0.9521	3.3620
		30	0.0009	-0.0026	46.5730	0.4434	0.5387	0.9725	2.1116
		50	0.0008	-0.0017	49.0562	0.3792	0.4145	0.9580	1.6248
		100	0.0028	0.0020	44.7031	0.2508	0.3052	0.9728	1.1963
	50%	10	0.0051	0.0037	40.5149	0.7663	0.9301	0.9728	3.6459
		30	0.0011	0.0027	48.8253	0.5009	0.5620	0.9640	2.2029
		50	-0.0023	-0.0015	47.6567	0.4384	0.4355	0.9540	1.7073
		100	-0.0020	0.0040	42.2507	0.2918	0.3153	0.9551	1.2359
(100, 500)	25%	10	0.0001	0.0048	734.0281	0.7891	0.8495	0.9651	3.3300
		30	-0.0074	0.0042	764.4239	0.3478	0.5113	0.9882	2.0041
		50	-0.0040	-0.0003	751.5734	0.2580	0.4056	0.9894	1.5901
		100	-0.0003	0.0040	812.7529	0.2099	0.2900	0.9865	1.1367
	50%	10	0.0029	0.0008	554.5228	0.7844	0.8735	0.9622	3.4240
		30	0.0075	-0.0005	805.0909	0.4451	0.5370	0.9747	2.1052
		50	0.0017	-0.0048	746.5026	0.3588	0.4214	0.9723	1.6520
		100	-0.0025	-0.0014	803.7756	0.2421	0.2997	0.9782	1.1748

Note that $SE_1(\hat{\delta})$ and $SE_2(\hat{\delta})$ are denoted as the empirical standard error and the formula-based standard error of the mean difference estimator ($\hat{\delta}$), respectively.

The simulation work given in the beginning of this section has focused on homogeneity situations. However, in fact, the heterogeneity of effect has become an important issue and often occurred in meta-analysis. To investigate whether the novel method derived in this paper is useful in the heterogeneity case, several scenarios are considered. The scope of simulation is given as follows. We generated the mean in the control group $\bar{X}_i^C$ from a normal distribution $N(\mu_i, \sigma^2 u_i^C)$, for $i = 1, 2, ..., k$, where μ_i were sampled from a normal distribution $N(\mu, \tau^{2C})$, and the true mean in the control group μ was given as 2, and the between-studies variance in the control group τ^{2C} was 2 and 4. We sampled the mean in the treatment group $\bar{X}_i^T$ from an $N(\mu_i + \delta_i, \sigma^2 u_i^T)$, where the random effect δ_i was generated from a normal distribution $N(\delta, \tau^2)$ where the true mean of effect size δ was given as 0 and 20, and the between-studies variance parameter τ^2 was 0 and 4. Thus, these scenarios included $(\tau^2, \tau^{2C}) = (0, 4), (4, 2)$, and $(4, 4)$. The prior setting refers to no effect heterogeneity across studies but baseline heterogeneity exists, and the latter two settings mean there are both effect and baseline heterogeneity. For u_i^C and u_i^T, they were sampled from uniform distributions on (0.1, 0.3), (0.02, 0.2), and (0.002, 0.01), corresponding to the study sample sizes on the intervals (3, 10), (5, 50), and (100, 500), respectively. Furthermore, the number of

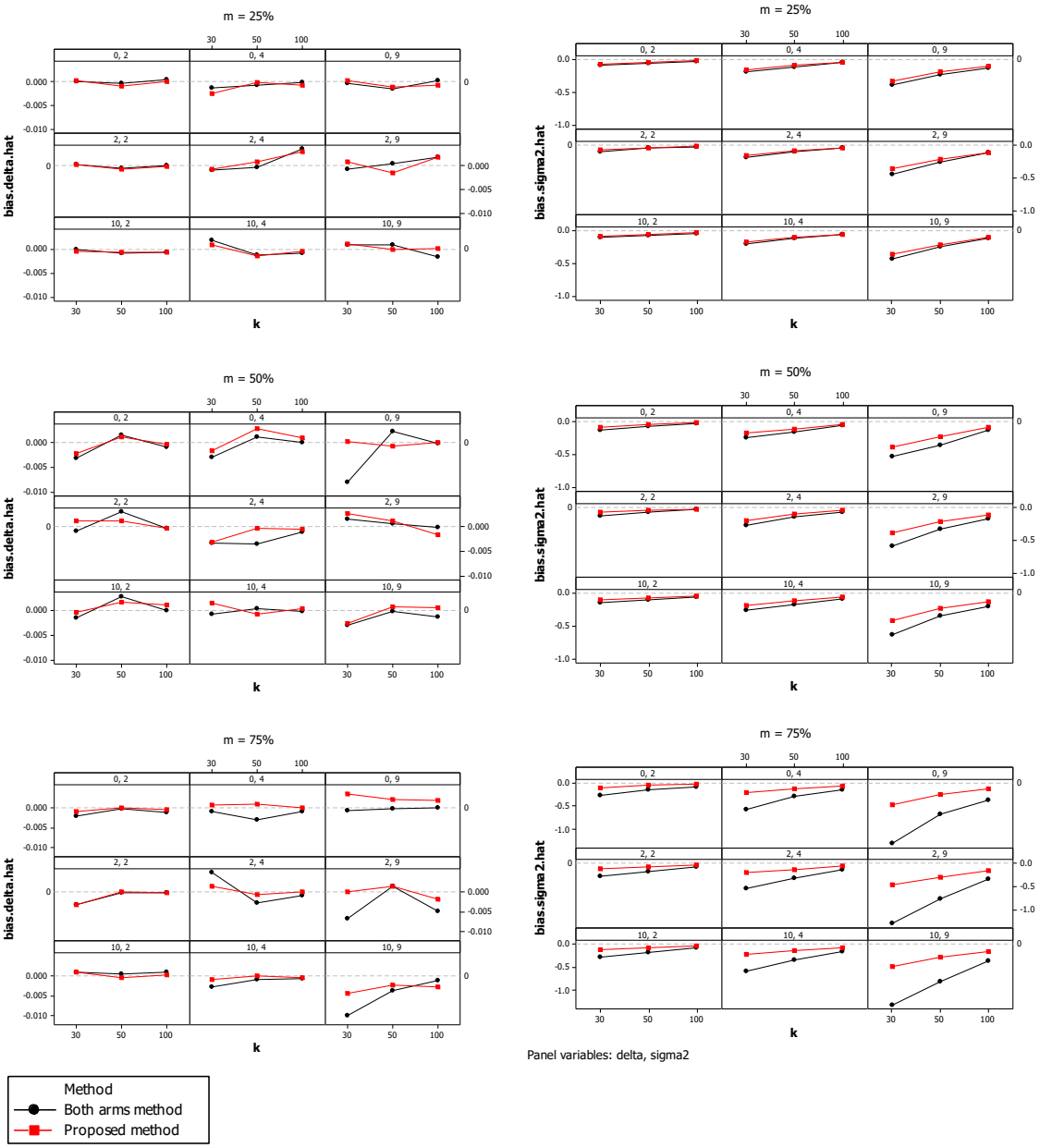

Figure 1. Bias of the mean difference $\hat{\delta}$ (left) and bias of the variance estimator $\hat{\sigma}^2$ (right) using simulations.

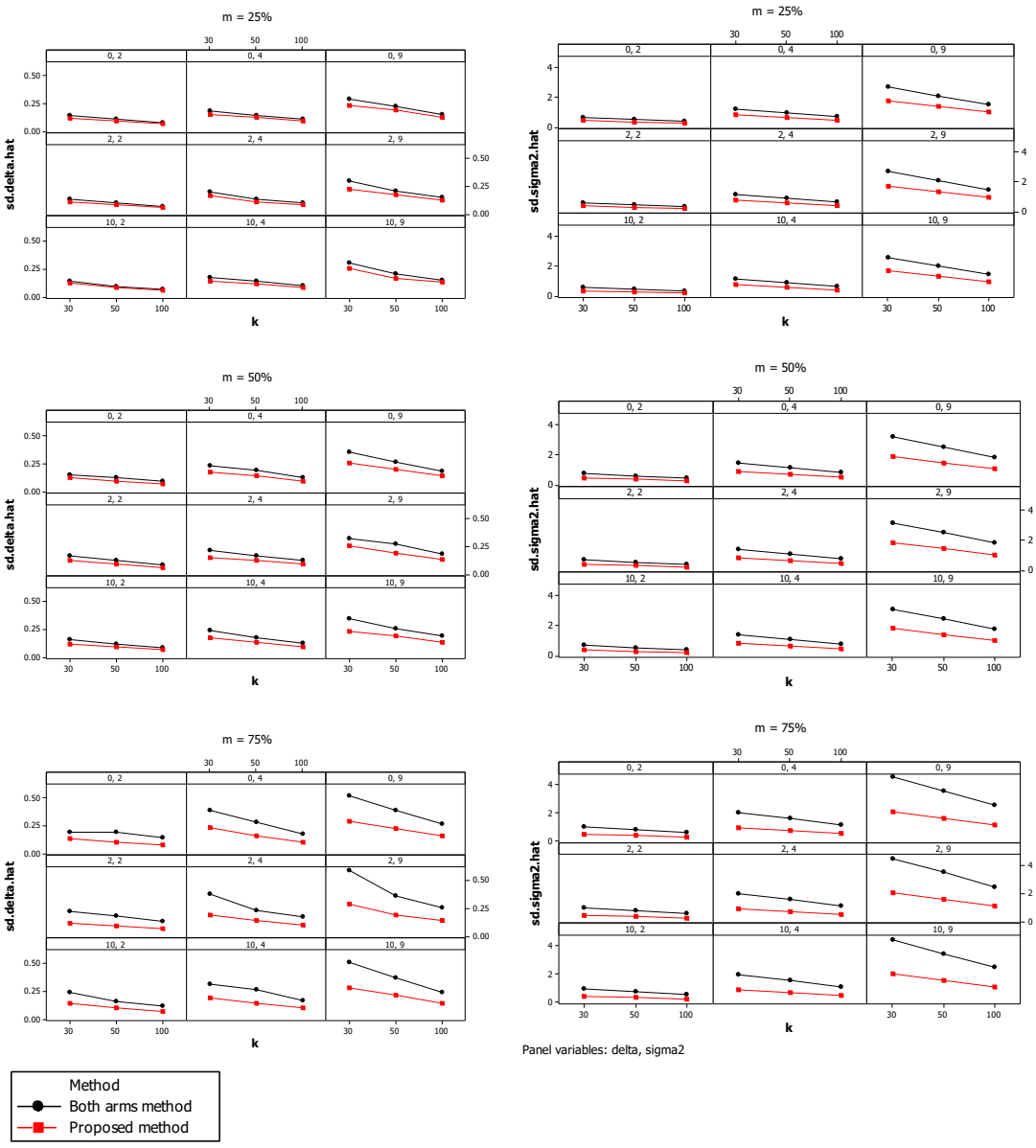

Figure 2. Standard error of the mean difference $\hat{\delta}$ (left) and standard deviation of variance estimator $\hat{\sigma}^2$ (right) using simulations.

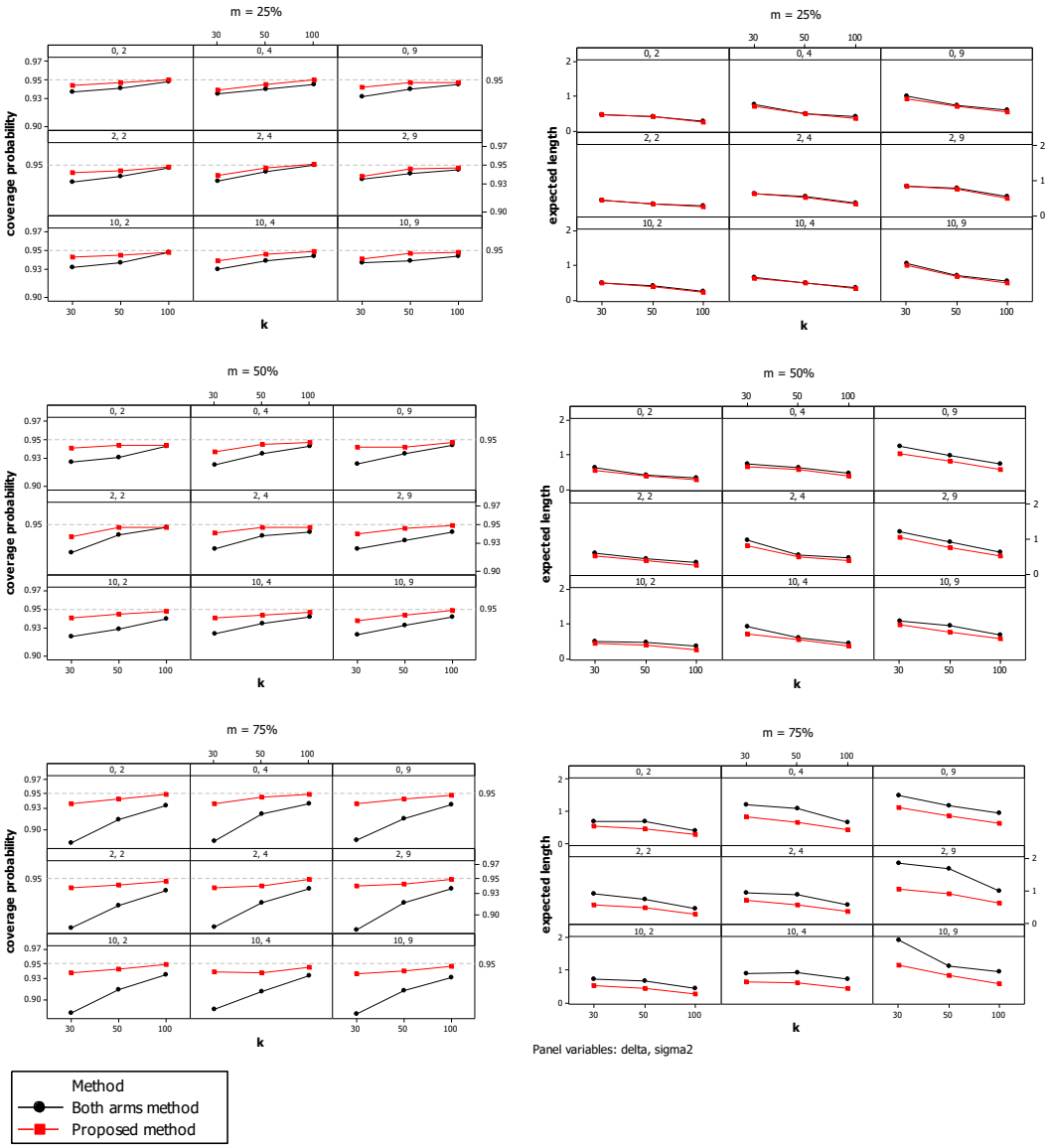

Figure 3. Coverage probability (left) and expected length (right) of the 95% confidence interval for the mean difference (δ) using simulations.

studies (k) and number of missing values (m) were varied, for details see Tables 5 and 6.

Under these heterogeneity settings, or estimating the mean difference parameter δ, we used the homogeneity models given in equations (3-5). So, we are interested here in studying the performance under model mis-specification as there is no effect homogeneity simulated. The performance of estimators was investigated in terms of bias and standard error, the latter was computed from the root of empirical variance found via simulation, see equation (7), and the standard error obtained from the average of the square root of the variance given by formula (6) were estimated. They are denoted as $\mathrm{SE}_1(\hat{\delta})$ and $\mathrm{SE}_2(\hat{\delta})$, respectively. We also estimated the 95% confidence interval for δ using the Wald method, where $\mathrm{SE}_2(\hat{\delta})$ was used in the computation. The main results obtained from the simulations with the parameter setting $\delta = 0$ are given in the following.

i) For the setting of Table 5, the simulation was under the situation on baseline heterogeneity only and it can be seen that the bias of $\hat{\delta}$ was close to zero in all cases in the study. Moreover, the coverage probability of the 95% confidence interval for δ was greater than or equal to the target probability level at 0.95. It can be therefore concluded that the homogeneity modelling was appropriate for the mean difference, the overall effect size of interest. In this situation, the standard error of $\hat{\delta}$ computed based on formula method did not differ much from the empirical standard error in general cases, especially if k was large. Both point and interval estimation based on our approach are then recommended to estimate the mean difference when there is baseline heterogeneity but no effect heterogeneity.

ii) The simulation results under both baseline and effect heterogeneity situation are given in Table 6. Point estimation for δ performed very well, as it had low bias. The standard error of $\hat{\delta}$ was slightly lower than the empirical standard error for large k. Similar to the previous result, the coverage probability of the confidence interval for δ was acceptable, as it was close to the nominal coverage level at 0.95 in many cases. Although the number of missing studies (m) are 25% or 50% of k, the method still works well. Therefore, we recommend to use the proposed methods derived under the homogeneity model for estimating the mean difference if study-specific variance is not included in the study, potentially only one arm information is presented, and baseline and effect heterogeneity is occurred.

iii) For the heterogeneity case, the variance estimate $\hat{\sigma}^2$ faced a strong overestimation in all situations of the study. We point out here that this is because it includes both within-study and across-study variances. More details are given again in the discussion section.

For the results of the simulation study with the parameter setting $\delta = 20$ we refer the reader to Tables A1 and A2 of the appendix section. The results are in line with those presented in the above paragraph.

3.2. Case study (continued)

The meta-analytic data used in this work contained continuous outcomes on treating congenital lung malformation using the thoracoscopic and open operations. The data provided only the means of the quantity of interest and the number of patients or

sample sizes. Some studies reported complete information on both treatments, but many reported information only on one arm. In the analysis, two continuous variables consisted of length of hospital stay (days) with number of the studies based on $(k_0, k_1, k_2) = (10, 14, 7)$ and length of operation (mins) with $(k_0, k_1, k_2) = (11, 12, 2)$ were used. Example of meta-analytic data is given in Tables 1-3. We compared the performance of the treatment by comparing the means between two surgeries, using the proposed method which uses all available information and the both arm information method. The results presented in Table 7 show that more time is spent for operation in thoracoscopic than in the traditional surgery. However, thoracoscopic seems to be better in length of hospital stay. This makes perfect sense because the technique of key-hole surgery is non-invasive, so the patients experience faster recovery.

4. Conclusions

We have seen that the standard meta-analysis using the inverse variance-weighted average method requires an effect measure estimate with an estimate of its variance from each study. This approach can only be applied in a complete case analysis. So that it is a crucial problem in meta-analyses in which all sample variances are missing from the studies. This problem was also addressed by Sangnawakij et al. [**?** **?**]. They presented estimation of the mean difference for two treatments and related variance using ML estimation. However, their method required complete information, sample means and sample sizes, in both treatment arms. It is common sense to include information for all individual studies in the analysis and we have followed this paradigm here. Consequently, a likelihood is considered which includes the entire, available information. This leads to an improved method for missing comparative arm information, as it allows the full set of independent studies in analysis, not only on both arms.

The performance of the estimators obtained from the proposed method was evaluated in terms of bias and standard deviation through simulations, and compared with the performance of the estimators from the both arm information method. It can be concluded that the biases of the mean difference obtained from the proposed method were slightly smaller than those from the both arms method. The biases of variance estimate from the new approach also performed well. Our proposed estimates showed good accuracy with smaller standard deviations than the compared estimators across all cases in the study. Moreover, the coverage probabilities of the confidence interval for the mean difference by the new approach hit the target probability level, and were greater than those of the both arm method, especially in the setting of a large number of single-arm studies. We conclude that all available studies in meta-analysis should be included in the analysis. However, the proposed method is based on certain assumptions. The first assumption is that all studies arise from a population which are measured on a scale with the same standard deviation σ. Secondly, it is assumed that heterogeneity is explained by treatment variation, not any other, in particular unobserved source. The third assumption is that one-armed information behaves in the same way as two-armed information. The last assumption seems plausible in the considered application as the way the reports were collected does not indicate any relationship to the intervention. Within the above described limits, we recommend the method proposed in this study to estimate the mean difference in meta-analysis with no study-specific variance and missing comparative arm information.

In addition, we have included simulation work for heterogeneity across studies. The homogeneity model was used to estimate the parameters in such a case, espe-

cially focussing on evaluating the mean difference (effect size of interest). The results showed that the method was acceptable for the overall effect size, as the estimated mean difference had low bias and the confidence interval for the effect size parameter had reasonable coverage probability in many situations. However, the simulation also showed that the variance parameter experienced potentially strong overestimation. We indicate that this is quite understandable, as it is now a mixture of within- and across-study variation. The variance parameter needs to be considered as a nuisance parameter and cannot be interpreted as a within-study variation term. Hence, Higgins' I^2 statistic, a measure of the magnitude of heterogeneity which is widely used in meta-analysis cannot be used here as it requires separate estimates of the within- and across-study variances, respectively. However, these variance estimates are not available and so I^2 is undefined. We point out here for future work that nonparametric estimation, such as the bootstrap approach, could be used to estimate the variance of the overall estimator, at least for improving interval estimation and providing valid inference for the overall effect even under strong baseline and/or effect heterogeneity.

Appendix A. Additional simulation results

According to the simulation study given in Section 3.1, the performance of estimators generated under the heterogeneity situation with the parameter $\delta = 20$ is presented in Tables A1 and A2. Table A1 shows the simulation results under the situation on baseline heterogeneity only and Table A2 provides the simulation results under both baseline and effect heterogeneity situation.

Table 6. The performance of estimators under baseline and effect heterogeneity using simulations ($\delta = 0$).

Sample size	m	k	Bias of $\hat{\delta}$	Bias of $\hat{\mu}$	$\hat{\sigma}^2$	$\text{SE}_1(\hat{\delta})$	$\text{SE}_2(\hat{\delta})$	Coverage probability	Expected length
Both baseline and effect heterogeneity: $(\tau^2, \tau^{2C}) = (4, 2)$									
(3, 10)	25%	10	0.0079	0.0012	18.7802	1.0049	0.9727	0.9408	3.6325
		30	0.0025	-0.0009	20.8889	0.6066	0.5664	0.9662	2.2477
		50	-0.0005	-0.0011	21.9721	0.4983	0.4489	0.9694	1.7119
		100	0.0005	0.0036	21.9229	0.3807	0.3158	0.9717	1.2407
	50%	10	0.0147	-0.0064	21.5358	1.0844	0.9785	0.9435	3.7462
		30	0.0005	-0.0045	22.0701	0.6697	0.5775	0.9486	2.2587
		50	0.0031	-0.0001	23.5278	0.5367	0.4523	0.9475	1.7245
		100	0.0001	-0.0014	21.2051	0.4188	0.3232	0.9582	1.2922
(5, 50)	25%	10	0.0082	0.0074	43.0258	0.9669	0.9417	0.9358	3.3654
		30	0.0042	-0.0020	44.8737	0.7162	0.5310	0.9452	2.0300
		50	-0.0079	0.0034	47.1877	0.5545	0.4121	0.9466	1.6050
		100	-0.0027	0.0026	54.6754	0.4151	0.2916	0.9486	1.1856
	50%	10	-0.0152	0.0043	40.6055	1.1203	0.8337	0.9212	3.8692
		30	0.0055	-0.0029	41.9702	0.7716	0.5369	0.9305	2.1402
		50	-0.0089	-0.0014	50.1939	0.6389	0.4363	0.9397	1.6650
		100	-0.0054	0.0003	49.1713	0.4699	0.2958	0.9410	1.2378
(100, 500)	25%	10	0.0128	-0.0037	649.0151	0.9347	0.8879	0.9422	3.3251
		30	-0.0015	-0.0038	782.0827	0.6266	0.5155	0.9504	2.0265
		50	-0.0001	0.0033	762.1261	0.4740	0.4028	0.9567	1.5974
		100	0.0027	0.0013	788.5318	0.3658	0.2888	0.9596	1.1385
	50%	10	0.0093	-0.0009	602.2615	0.9816	0.8716	0.9343	3.6274
		30	-0.0068	0.0004	726.5199	0.6694	0.5402	0.9366	2.1024
		50	-0.0002	-0.0025	840.8957	0.5242	0.4240	0.9421	1.6590
		100	0.0019	-0.0004	690.9665	0.3960	0.2935	0.9437	1.1618
Both baseline and effect heterogeneity: $(\tau^2, \tau^{2C}) = (4, 4)$									
(3, 10)	25%	10	-0.0045	0.0037	21.3152	0.8766	0.9821	0.9564	3.8500
		30	0.0083	0.0032	21.8060	0.5221	0.5587	0.9587	2.1900
		50	0.0008	-0.0026	21.5381	0.4032	0.4382	0.9646	1.7177
		100	-0.0014	-0.0008	20.6942	0.2831	0.3138	0.9694	1.2301
	50%	10	-0.0131	0.0053	19.3768	0.9321	0.9548	0.9436	3.7429
		30	-0.0068	0.0108	19.5289	0.5814	0.5882	0.9484	2.3059
		50	0.0033	-0.0005	20.9550	0.4518	0.4576	0.9513	1.7936
		100	-0.0047	0.0026	21.9379	0.3177	0.3204	0.9500	1.2561
(5, 50)	25%	10	0.0121	-0.0013	35.2440	1.0131	0.8611	0.9380	3.3753
		30	-0.0001	-0.0003	42.1298	0.6290	0.5346	0.9401	2.0958
		50	0.0030	0.0022	51.5228	0.4585	0.4147	0.9530	1.6256
		100	-0.0059	0.0002	45.8607	0.3372	0.2969	0.9561	1.1639
	50%	10	0.0129	-0.0031	40.5873	1.0103	0.8734	0.9353	3.4237
		30	-0.0038	0.0006	42.6041	0.5796	0.5504	0.9381	2.1575
		50	-0.0090	0.0050	49.6152	0.5238	0.4054	0.9421	1.5890
		100	0.0061	-0.0030	52.0464	0.3644	0.3071	0.9510	1.2039
(100, 500)	25%	10	-0.0109	-0.0092	638.5015	0.8191	0.8580	0.9433	3.3633
		30	0.0067	-0.0028	699.6881	0.4972	0.5141	0.9525	2.0152
		50	-0.0016	-0.0056	801.1267	0.3914	0.4105	0.9597	1.6090
		100	-0.0015	0.0004	764.8910	0.2720	0.2897	0.9621	1.1358
	50%	10	-0.0002	0.0041	680.5690	0.9240	0.8781	0.9426	3.4421
		30	0.0058	-0.0003	812.0435	0.5487	0.5483	0.9437	2.1494
		50	-0.0003	-0.0003	865.8689	0.4297	0.4212	0.9422	1.6513
		100	0.0035	0.0009	783.5498	0.3075	0.2969	0.9549	1.1637

Note that $\text{SE}_1(\hat{\delta})$ and $\text{SE}_2(\hat{\delta})$ are denoted as the empirical standard error and the formula-based standard error of the mean difference estimator ($\hat{\delta}$), respectively.

Table 7. Mean difference and variance estimates on the meta-analytic data of thoracoscopic and open operations.

Variable/Method	Mean difference ($\hat{\delta}$)	Standard deviation ($\hat{\sigma}$)	95% confidence interval for δ
Length of operation			
Proposed method	31.32	229.71	(7.01, 55.62)
Both arm information method	36.97	131.45	(18.96, 54.99)
Length of stay			
Proposed method	-1.29	14.09	(-2.76, 0.10)
Both arm information method	-1.30	7.49	(-2.34, -0.26)

Table A1. The performance of estimators under baseline heterogeneity using simulations ($\delta = 20$).

Sample size	m	k	Bias of $\hat{\delta}$	Bias of $\hat{\mu}$	$\hat{\sigma}^2$	$\text{SE}_1(\hat{\delta})$	$\text{SE}_2(\hat{\delta})$	Coverage probability	Expected length
Only baseline heterogeneity: $(\tau^2, \tau^{2C}) = (0, 4)$									
(3, 10)	25%	10	-0.0094	0.0138	18.9333	0.7632	0.9530	0.9654	3.7356
		30	0.0002	-0.0013	19.6828	0.5133	0.5674	0.9658	2.2241
		50	-0.0039	-0.0041	20.5569	0.3862	0.4405	0.9677	1.7268
		100	-0.0007	-0.0002	22.6265	0.2963	0.3130	0.9683	1.2270
	50%	10	-0.0043	0.0051	16.8474	0.8357	0.9523	0.9439	3.7330
		30	-0.0006	0.0033	18.9862	0.5612	0.5750	0.9449	2.2539
		50	-0.0012	0.0021	21.0031	0.3944	0.4541	0.9521	1.7799
		100	-0.0006	0.0010	21.8412	0.3128	0.3198	0.9555	1.2534
(5, 50)	25%	10	-0.0071	0.0006	36.9344	0.7330	0.8835	0.9555	3.4633
		30	-0.0014	-0.0040	45.5451	0.4770	0.5419	0.9539	2.1243
		50	-0.0018	0.0041	52.5351	0.4175	0.4101	0.9630	1.6076
		100	-0.0009	0.0009	50.8182	0.2928	0.2965	0.9639	1.1624
	50%	10	-0.0008	0.0022	31.0225	0.8886	0.9190	0.9439	3.6023
		30	-0.0018	0.0033	50.3725	0.5962	0.5284	0.9447	2.0713
		50	-0.0030	-0.0039	42.7111	0.4154	0.4240	0.9451	1.6622
		100	-0.0039	0.0046	54.5578	0.3218	0.3043	0.9468	1.1929
(100, 500)	25%	10	-0.0047	0.0174	690.3803	0.6214	0.8597	0.9871	3.3700
		30	0.0034	-0.0011	750.8532	0.3786	0.5131	0.9879	2.0112
		50	0.0008	0.0023	776.2351	0.2821	0.3051	0.9892	1.5879
		100	-0.0013	0.0008	765.1935	0.2356	0.2904	0.9867	1.1385
	50%	10	0.0039	-0.0073	683.3393	0.8028	0.8594	0.9589	3.3690
		30	0.0003	0.0047	737.8925	0.4272	0.5326	0.9534	2.0879
		50	0.0028	-0.0070	812.0212	0.3945	0.4089	0.9573	1.6030
		100	0.0084	-0.0042	842.9518	0.2857	0.2931	0.9574	1.1489

Note that $\text{SE}_1(\hat{\delta})$ and $\text{SE}_2(\hat{\delta})$ are denoted as the empirical standard error and the formula-based standard error of the mean difference estimator ($\hat{\delta}$), respectively.

Table A2. The performance of estimators under baseline and effect heterogeneity using simulations ($\delta = 20$).

Sample size	m	k	Bias of $\hat{\delta}$	Bias of $\hat{\mu}$	$\hat{\sigma}^2$	$\text{SE}_1(\hat{\delta})$	$\text{SE}_2(\hat{\delta})$	Coverage probability	Expected length
Both baseline and effect heterogeneity: $(\tau^2, \tau^{2C}) = (4, 2)$									
(3, 10)	25%	10	0.0052	0.0022	18.7849	0.9000	0.9578	0.9436	3.7546
		30	-0.0022	0.0034	21.6539	0.5082	0.5629	0.9637	2.2064
		50	0.0045	-0.0051	22.4462	0.3938	0.4374	0.9677	1.7144
		100	-0.0020	0.0011	22.3047	0.2872	0.3135	0.9655	1.2290
	50%	10	0.0051	-0.0196	22.5272	0.9954	0.9332	0.9416	3.6580
		30	-0.0001	-0.0089	24.7230	0.5797	0.5961	0.9498	2.3367
		50	0.0004	-0.0005	20.8363	0.4415	0.4514	0.9529	1.7696
		100	0.0005	0.0020	21.1955	0.3152	0.3208	0.9546	1.2575
(5, 50)	25%	10	0.0030	-0.0012	38.1327	0.9328	0.9732	0.9089	3.8150
		30	-0.0099	0.0029	43.7078	0.6123	0.5285	0.9284	2.0718
		50	-0.0033	0.0041	54.7044	0.4657	0.4176	0.9423	1.6370
		100	0.0041	-0.0001	54.2407	0.3230	0.3072	0.9439	1.2042
	50%	10	-0.0036	0.0111	51.9269	1.1304	0.8025	0.9030	3.1460
		30	0.0101	-0.0001	42.2358	0.6582	0.5630	0.9106	2.2070
		50	-0.0051	0.0036	51.9313	0.5025	0.4189	0.9221	1.6422
		100	0.0062	-0.0027	48.8592	0.3733	0.3112	0.9377	1.2200
(100, 500)	25%	10	0.0144	0.0088	772.2555	0.8633	0.8735	0.9418	3.4240
		30	-0.0047	-0.0020	676.4746	0.4991	0.5067	0.9501	1.9864
		50	-0.0019	-0.0001	772.2814	0.3727	0.4034	0.9646	1.5814
		100	0.0026	0.0022	784.5804	0.2663	0.2896	0.9670	1.1351
	50%	10	-0.0009	-0.0034	712.1790	0.9821	0.8707	0.9158	3.4130
		30	-0.0075	0.0062	755.0461	0.5449	0.5317	0.9483	2.0843
		50	-0.0039	0.0009	845.3486	0.4366	0.4227	0.9497	1.6569
		100	0.0023	-0.0006	773.4007	0.3011	0.2963	0.9500	1.1613
Both baseline and effect heterogeneity: $(\tau^2, \tau^{2C}) = (4, 4)$									
(3, 10)	25%	10	-0.0076	0.0091	20.9216	0.8685	0.9843	0.9600	3.8585
		30	-0.0046	-0.0034	20.4184	0.5111	0.5666	0.9652	2.2210
		50	0.0001	-0.0015	22.9020	0.3913	0.4417	0.9705	1.7313
		100	-0.0044	-0.0006	21.8968	0.2765	0.3129	0.9740	1.2265
	50%	10	0.0085	-0.0033	17.1331	0.9867	0.9611	0.9333	3.7677
		30	0.0008	-0.0010	22.7232	0.5844	0.5752	0.9487	2.2547
		50	-0.0032	0.0017	20.9822	0.4474	0.4538	0.9582	1.7788
		100	0.0025	0.0014	21.5740	0.3140	0.3237	0.9645	1.2691
(5, 50)	25%	10	0.0089	0.0019	71.1829	1.0483	0.9653	0.9246	3.7839
		30	-0.0004	-0.0021	53.3508	0.6062	0.5375	0.9316	2.1070
		50	0.0037	-0.0021	49.8220	0.4437	0.4220	0.9435	1.6543
		100	-0.0032	-0.0010	52.2674	0.3307	0.2961	0.9480	1.1606
	50%	10	-0.0100	0.0106	47.2427	1.0290	1.0211	0.9222	4.0026
		30	-0.0026	0.0030	59.8850	0.6426	0.5636	0.9248	2.2092
		50	0.0042	0.0003	49.3356	0.4986	0.4070	0.9389	1.5956
		100	-0.0010	-0.0005	47.1564	0.3646	0.3200	0.9412	1.2544
(100, 500)	25%	10	-0.0148	0.0025	709.5698	0.8158	0.8636	0.9423	3.3853
		30	-0.0058	0.0026	658.9964	0.4675	0.5018	0.9591	1.9672
		50	-0.0028	-0.0011	774.6089	0.3887	0.4049	0.9554	1.5873
		100	-0.0011	-0.0025	794.8063	0.2644	0.2896	0.9692	1.1351
	50%	10	0.0121	-0.0036	648.1704	0.9271	0.8015	0.9196	3.1417
		30	-0.0008	-0.0033	823.3384	0.5598	0.5331	0.9323	2.0898
		50	-0.0043	0.0053	830.5367	0.4573	0.4212	0.9403	1.6510
		100	0.0046	-0.0016	796.4618	0.3029	0.2948	0.9452	1.1555

Note that $\text{SE}_1(\hat{\delta})$ and $\text{SE}_2(\hat{\delta})$ are denoted as the empirical standard error and the formula-based standard error of the mean difference estimator ($\hat{\delta}$), respectively.