

Statistical methodology for estimating the mean difference in a meta-analysis without study-specific variance information

Patarawan Sangnawakij,^a Dankmar Böhning,^{b,*†} 
Stephen Adams,^c Michael Stanton^c and Heinz Holling^d

Statistical inference for analyzing the results from several independent studies on the same quantity of interest has been investigated frequently in recent decades. Typically, any meta-analytic inference requires that the quantity of interest is available from each study together with an estimate of its variability. The current work is motivated by a meta-analysis on comparing two treatments (thoracoscopic and open) of congenital lung malformations in young children. Quantities of interest include continuous end-points such as length of operation or number of chest tube days. As studies only report mean values (and no standard errors or confidence intervals), the question arises how meta-analytic inference can be developed. We suggest two methods to estimate study-specific variances in such a meta-analysis, where only sample means and sample sizes are available in the treatment arms. A general likelihood ratio test is derived for testing equality of variances in two groups. By means of simulation studies, the bias and estimated standard error of the overall mean difference from both methodologies are evaluated and compared with two existing approaches: complete study analysis only and partial variance information. The performance of the test is evaluated in terms of type I error. Additionally, we illustrate these methods in the meta-analysis on comparing thoracoscopic and open surgery for congenital lung malformations and in a meta-analysis on the change in renal function after kidney donation. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: likelihood ratio test; mean difference; meta-analysis

1. Introduction and motivating case study

Meta-analysis is a statistical methodology for the analysis and integration of results from individual, independent studies. In recent decades, meta-analysis developed a crucial role in many fields of science such as medicine and pharmacy, health science, psychology and social science [1–4]. Any meta-analytic inference, as developed so far, requires as minimal information that the quantity of interest is available from each study together with an estimate of its variability. The sample variance information is used to compute the weights of the effect size in each study for achieving the overall estimate [5, 6]. A problem in a meta-analysis of continuous outcomes occurs when incomplete variability measures are not reported or are missing sample variances. In recent years, several approaches to impute variance estimates when lacking those variances for some studies have been considered ([7–9]). For example, Philbrook *et al.* [7] compared changes in renal function after living kidney donation using meta-analysis based on four imputation methods, reported *p*-values in primary studies, reported non-parametric summaries in primary studies, correlation values and multiple imputation, for imputing missing change sample variances.

^aDepartment of Applied Statistics, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand

^bSchool of Mathematics & Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, U.K.

^cDepartment of Paediatric Surgery, Southampton General Hospital, Southampton, SO16 6YD, U.K.

^dStatistics and Quantitative Methods, Faculty of Psychology and Sports Science, University of Münster, Münster, Germany

*Correspondence to: Dankmar Böhning, School of Mathematics & Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, U.K.

†E-mail: d.a.bohning@soton.ac.uk

Table I. Total number of complications in open vs. thoracoscopic surgery.

	Thoracoscopy	Open	Total
Complications	53	87	140
At risk	352	420	772

Chowdhry *et al.* [9] considered a meta-analysis for the difference in means where sample variances are missing for some studies. They used a multiple imputation method employing gamma meta-regression to impute the missing variances. However, meta-analysis, which has only the information of the quantity of interest (with no information on within-study variability for all studies), has not yet been considered as previous work relies on the fact that variance information is available at least for some studies. The question arises how meta-analytic inference can be developed. This is the situation we are dealing with here.

The current work is motivated by difficulties encountered in performing a meta-analysis of two treatments (thoracoscopic and open resections) of asymptomatic lung malformations in young children. Congenital lung malformations are seen on routine antenatal scans of 1 in 2500 live born children, and there is controversy regarding their management [10]. The majority of children's surgeons elect to operate on infants with lesions even if they have not caused any symptoms, because of concern about potential for infection, or the purported small risk of developing a cancer in the abnormal piece of lung later in life [11]. Because the late 1990s, there has been a trend towards surgeons performing these operations using a minimally invasive approach (key-hole surgery) called thoracoscopy [12]. Thoracoscopy involves making several small incisions in the chest through which a fibre-optic camera and operating instruments can be passed, in this manner lesions can be removed with minimal scarring. In contrast, the traditional operation would be performed through a thoracotomy which is a single, large incision on the side of the chest wide enough to see the lung and its attachments. The aim of the operation regardless of the approach (thoracoscopic or open) is to remove the lung lesion including any lung directly involved with it. This often entails removing one of the lobes of the affected lung.

It is currently unclear whether thoracoscopy is associated with an increased risk of operative complications or not. There have been no published reports of mortality secondary to this operation for asymptomatic cases in the past 10 years, although a recent study described a near-death [13]. Other outcomes that are relevant include wound infection, chest infection, bleeding, injury to the trachea (wind-pipe), prolonged air leak into the space around the lung or development of new air leak subsequent to the operation. When the operation is commenced with thoracoscopy and changed to an open technique during the procedure, this is referred to as an open conversion. Typically, studies into this type of surgery report basic demographic details of the study population, conversion rate and complications (Table I). If the interest is in the number of complications as primary endpoint, a standard meta-analysis using count modelling can be performed [14]. In more detail, using a mixed Poisson regression [2, 5, 15] with log-size-at-risk as offset and study as random effect, the results show that total number of complications in thoracoscopic has a significantly decreased risk ratio (0.69) in favour of the thoracoscopic surgery with a p -value of 0.04. Note that the Mantel–Haenszel estimator of the risk ratio [6] with 0.68 is very close to the mixed Poisson model-based estimator. This beneficial aspect of thoracoscopy raises interest in also investigating other clinical endpoints that are of quantitative nature. In contrast to the number of complications where the count structure carries intrinsic information on its variance, this is not that case when we turn to quantitative outcomes where a second independent variance parameter describes the variability. This leads to the problem we would like to address here.

In a meta-analysis [15–17] of 38 studies investigating the performance of open and thoracoscopic surgery, studies report a variety of outcomes including the length of operation, the number of days that drains are left in the chest and number of days in hospital, age and weight of child. As a typical example, available information on mean age of patient (months) is shown in Table II for the two surgical procedures in the 38 studies. Unfortunately the reports in the surgical literature are heterogeneous in the specific type of data presented and generally refer to small patient populations. Quite often papers will present outcomes either for a thoracoscopic series only or an entirely open surgery group only, with no control group for comparison. In order to assess whether the thoracoscopic approach is as safe as the open approach and indeed whether there are benefits to performing this type of surgery using available data, statistical analysis has to be developed to perform a meta-analysis of the data without using reported measures of variance. Hence, we are interested in estimating the mean difference in a meta-analysis when only

Table II. The mean age of child and sample size from thoracoscopic and open operations.

Study	Thoracoscopic		Open	
	<i>n</i>	mean age	<i>n</i>	mean age
Vu 2008	12	6.7	24	3.7
Diamond 2007	12	7.18	24	7.56
Reismann 2012	14	4.8	8	3.5
Kunisaki 2014	49	18	13	1
Lau 2013	39	10.4	28	11.7
Rahman 2009	14	10	14	7
Cho 2012	7	21	27	100
Tolg 2005	5	94	4	39
Jesch 2005	5	0.79	—	—
de Lagausie 2005	8	10	—	—
Tanaka 2013	12	65.6	—	—
Rothenberg 2008	97	45.6	—	—
Rothenberg 2011	75	4.1	—	—
Seong 2013	50	38.4	—	—
Kaneko 2010	7	0.9	—	—
Muller 2012	12	12	—	—
Sundararajan 2007	20	19	—	—
Tarrado 2010	6	8.5	—	—
Truitt 2006	12	10.3	—	—
Zeidan 2009	6	5	—	—
Cano 2006	6	10	—	—
Boubnova 2011	30	3.94	—	—
Tsai 2008	—	—	105	2.5
Calvert 2007	—	—	16	80
Raychaudhuri 2011	—	—	14	8.3
Nagata 2009	—	—	5	4.5
Sueyoshi 2008	—	—	8	0.52
Chow 2007	—	—	6	3.66
Aziz 2004 (Asymp, elective or < 6m)	—	—	6	5
Aziz 2004 (Asymp init but symp > 6mo)	—	—	3	8
Aziz 2004 (Asymp, elective or > 6m)	—	—	9	3
Ferreira 2010	—	—	35	17
Fascetti-Leon 2013	26	—	28	—
Fievet 2012	9	—	2	—
Albanese 2007	144	—	—	—
Johnson 2011	15	—	—	—
Laje 2015	100	1.7	188	1.8
Kulaylat 2015	112	15.6	146	18

sample means and sample sizes of the two groups to be compared are available, but the estimated variances are not. Furthermore, it is possible that population variances between the two groups are homogeneous; however, they might not be identical. Therefore, in this paper, two approaches are considered in order to estimate the overall mean difference in a meta-analysis. A general likelihood ratio test statistic is also derived for testing the equality of variances of the two groups. In any of the two approaches, we assume equality of variances across studies. This seems to be a strong assumption but also reasonable as surgeries take place under fairly standard conditions.

The rest of the paper is organized as follows. In Section 2, we consider the estimator of variance for the equal variance case. This allows variance estimators to be derived for the individual studies involved in the meta-analysis. Under equal variances, the overall mean difference is presented with its associated standard error. In Section 3, we focus on estimation parameters under unequal population variances. Again, variance estimates for the individual studies are developed. Then, the overall mean difference with its associated standard error is derived. A likelihood ratio test for test of equality of variances is introduced in Section 4. Section 5 presents the results of simulation work with focus on mean difference with respect to bias, standard error, coverage probability and expected length. Type I error probability is used to investigate the likelihood ratio test. In Section 6, all estimators are illustrated in two case studies

of comparing open versus thoracoscopic surgery (highlighting the benefits of the latter) and the change in renal function after kidney donation. The paper ends with Section 7 containing discussion and concluding remarks which sets the approach in perspective.

2. Methodology I: estimation of the mean difference for equal variances

We consider k independent studies with only sample means and sample sizes in the treatment group (T) and in the comparison or control group (C) available, respectively, $\bar{X}_i^T$, $\bar{X}_i^C$, n_i^T and n_i^C , where $\bar{X}_i^T = \sum_{j=1}^{n_i^T} X_{ij}^T/n_i^T$ and $\bar{X}_i^C = \sum_{j=1}^{n_i^C} X_{ij}^C/n_i^C$. Only $\bar{X}_i^T$ and $\bar{X}_i^C$ are observed, but not the within-study values X_{ij}^T and X_{ij}^C for $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, n_i$. In addition, no estimate of variances of $\bar{X}_i^T$ and $\bar{X}_i^C$ is available. Here, we are interested in developing statistical inference for the mean difference between two groups as the effect size. The mean difference of the study i is given by

$$D_i = \bar{X}_i^T - \bar{X}_i^C. \quad (1)$$

Assume that X_{ij}^T and X_{ij}^C are independent and normally distributed with means μ^T and μ^C , respectively, and identical variance σ^2 across groups and studies. Thus, it follows that $\bar{X}_i^T$ and $\bar{X}_i^C$ are independent and normally distributed with means μ^T and μ^C , and variances σ^2/n_i^T and σ^2/n_i^C , respectively. These assumptions lead to the distribution of D_i , being also normal with mean $\mu = \mu^T - \mu^C$ and variance $\sigma^2 w_i$, where $w_i = 1/n_i^T + 1/n_i^C$ is non-random. In general, μ and σ^2 are unknown parameters and need to be estimated.

2.1. Maximum likelihood estimation

Under the normal distribution for the effect size D_i as mentioned before, the likelihood function of μ and σ^2 is given by

$$L(\mu, \sigma^2; D_i) = \prod_{i=1}^k \frac{1}{\sqrt{2\pi\sigma^2 w_i}} \exp \left\{ -\frac{(D_i - \mu)^2}{2\sigma^2 w_i} \right\}$$

with associated log-likelihood function

$$\log L(\mu, \sigma^2; D_i) = -\sum_{i=1}^k \frac{(D_i - \mu)^2}{2\sigma^2 w_i} - \frac{k \log(2\pi\sigma^2)}{2} - \log \prod_{i=1}^k \sqrt{w_i}. \quad (2)$$

Taking partial derivatives of (2) with respect to μ and σ^2 , respectively, we yield the maximum likelihood estimators

$$\hat{\mu}_{ML} = \frac{\sum_{i=1}^k D_i/w_i}{\sum_{i=1}^k 1/w_i} \quad \text{and} \quad \hat{\sigma}_{ML}^2 = \frac{1}{k} \sum_{i=1}^k \frac{(D_i - \hat{\mu}_{ML})^2}{w_i}. \quad (3)$$

Furthermore, it is important to investigate the precision of these estimators. We find the variance of estimators in (3) as follows:

$$\text{Var}(\hat{\mu}_{ML}) = \frac{\sigma^2}{\sum_{i=1}^k 1/w_i} \quad \text{and} \quad \text{Var}(\hat{\sigma}_{ML}^2) = \frac{2(k-1)\sigma^4}{k^2},$$

with details given in appendix A as Theorem 1. We also note that $\hat{\mu}_{ML}$ given earlier is an unbiased estimator of μ , while $\hat{\sigma}_{ML}^2$ is an asymptotically unbiased estimator of σ^2 . Both estimators are consistent. The details of proof are given in appendix B as Theorem 2–3.

In this study, the effect size is the mean difference. Hence, we are now able to provide an estimate of the variance of D_i as

$$\widehat{\text{Var}}(D_i) = \hat{\sigma}_{ML}^2 w_i. \quad (4)$$

This allows meta-analytic inference as we have now for each study an effect size estimator with associated estimated variance. This can be seen for mean age in Figure 1.

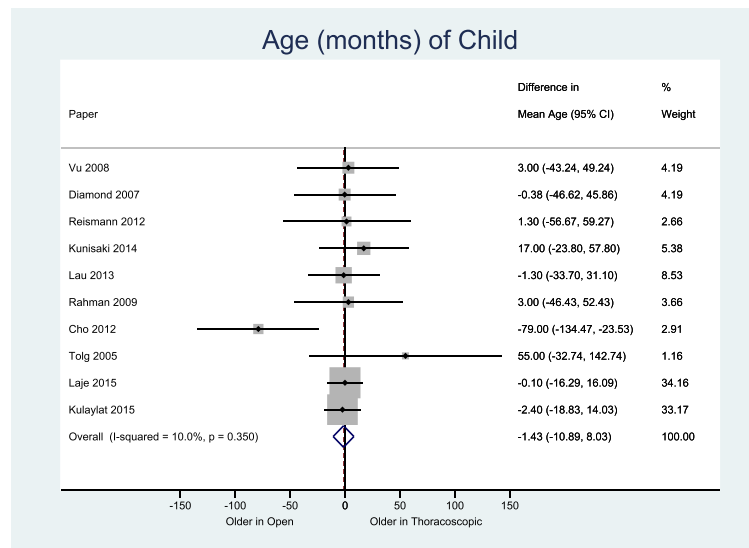


Figure 1. Forest plot for the difference in mean age by Methodology I. [Colour figure can be viewed at wileyonlinelibrary.com]

2.2. Estimation of the overall mean difference

Typically in meta-analysis, summary measures are developed as weighted average of the study-specific outcome measures, here D_i . We first consider the weight of the effect size in the study i , which is computed as the inverse of variance of its effect size [6]. For combining the mean difference, each study's mean difference is given weight

$$v_i = \frac{1/(\hat{\sigma}_{ML}^2 w_i)}{\sum_{i=1}^k 1/(\hat{\sigma}_{ML}^2 w_i)} = \frac{1/w_i}{\sum_{i=1}^k 1/w_i}.$$

Note that v_i only depends on the the sample sizes of the individual studies and is non-random. It follows that the estimator of the overall mean difference is given by

$$D_{\text{overall}} = \sum_{i=1}^k D_i v_i, \quad (5)$$

where D_i is the mean difference of the study i as defined in (1). Note that D_{overall} coincides with $\hat{\mu}_{ML}$ the maximum likelihood estimator as developed in (3).

Moreover, the $(1 - \alpha)100\%$ confidence interval for the overall mean difference can be calculated by

$$D_{\text{overall}} \pm Z_{\alpha/2} \widehat{SE}(D_{\text{overall}}), \quad (6)$$

where $Z_{\alpha/2}$ is $(\alpha/2)100\text{th}$ percentile of the standard normal distribution and $\widehat{SE}(D_{\text{overall}}) = \hat{\sigma}_{ML} / \sqrt{\sum_{i=1}^k 1/w_i}$ is the estimated standard deviation of the overall mean difference. The estimate of the overall mean difference for age together with a 95% confidence interval can be found as the last row in Figure 1.

Additionally, statistical inference for the mean difference is usually used to compare the effect size between two treatment groups. Thus, test of equality of mean is constructed. Under the null hypothesis $H_0 : \xi = 0$, where ξ is the population mean difference between the treatment and control groups, the test statistic is defined as

$$Z_{\text{overall}} = \frac{D_{\text{overall}}}{\sqrt{\hat{\sigma}_{ML}^2 / \sum_{i=1}^k 1/w_i}}, \quad (7)$$

where Z_{overall} has the standard normal distribution. Note that the null hypothesis is rejected, if $|Z_{\text{overall}}| > Z_{\alpha/2}$.

3. Methodology II: estimation of mean difference for unequal variances

The variances in the two groups of interest might not be equal. Therefore, in this section, the mean difference in meta-analysis is considered again, but we allow unequal variances for the two groups.

Assume that X_{ij}^T and X_{ij}^C for $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, n_i$ are independent and normally distributed with means μ^T and μ^C and unknown variances σ^{2T} and σ^{2C} , respectively. The sample means $\bar{X}_i^T$ and $\bar{X}_i^C$ are observed, and are also independent and normally distributed with means μ^T and μ^C , and variances σ^{2T}/n_i^T and σ^{2C}/n_i^C . Because σ^{2T} and σ^{2C} are unknown, we need to estimate them, and this is considered in the following.

3.1. Maximum likelihood estimation

We first consider the maximum likelihood estimator for parameters in the treatment group. Under the normal distribution, the log-likelihood function of μ^T and σ^{2T} is given by

$$\log L(\mu^T, \sigma^{2T}; \bar{X}_i^T) = - \sum_{i=1}^k \frac{(\bar{X}_i^T - \mu^T)^2}{2\sigma^{2T} w_i^T} - \frac{k \log(2\pi\sigma^{2T})}{2} - \log \prod_{i=1}^k \sqrt{w_i^T}, \quad (8)$$

where $w_i^T = 1/n_i^T$. Then taking partial derivatives of (8) with respect to μ^T and σ^{2T} , respectively, the estimators for μ^T and σ^{2T} are

$$\hat{\mu}_{ML}^T = \frac{\sum_{i=1}^k \bar{X}_i^T / w_i^T}{\sum_{i=1}^k 1/w_i^T} \quad \text{and} \quad \hat{\sigma}_{ML}^{2T} = \frac{1}{k} \sum_{i=1}^k \frac{(\bar{X}_i^T - \hat{\mu}_{ML}^T)^2}{w_i^T}. \quad (9)$$

Because the estimators in (9) are of simple nature, the variance estimates are easily derived as

$$\text{Var}(\hat{\mu}_{ML}^T) = \frac{\sigma^{2T}}{\sum_{i=1}^k 1/w_i^T} \quad \text{and} \quad \text{Var}(\hat{\sigma}_{ML}^{2T}) = \frac{2(k-1)\sigma^{4T}}{k^2}.$$

Similarly, we derive the maximum likelihood estimators for μ^C and σ^{2C} of the control group. The estimators are given by

$$\hat{\mu}_{ML}^C = \frac{\sum_{i=1}^k \bar{X}_i^C / w_i^C}{\sum_{i=1}^k 1/w_i^C} \quad \text{and} \quad \hat{\sigma}_{ML}^{2C} = \frac{1}{k} \sum_{i=1}^k \frac{(\bar{X}_i^C - \hat{\mu}_{ML}^C)^2}{w_i^C}, \quad (10)$$

where $w_i^C = 1/n_i^C$. Note that the maximum likelihood estimators given in (9) and (10) are consistent. The estimated means $\hat{\mu}^T$ and $\hat{\mu}^C$ are unbiased estimators for μ^T and μ^C , respectively. Furthermore, the estimated variances $\hat{\sigma}_{ML}^{2T}$ and $\hat{\sigma}_{ML}^{2C}$ are asymptotically unbiased estimators for σ^{2T} and σ^{2C} , respectively. Hence, the estimated variance of the mean difference of the study i in this case is given as

$$\widehat{\text{Var}}(D_i) = \hat{\sigma}_{ML}^{2T} w_i^T + \hat{\sigma}_{ML}^{2C} w_i^C. \quad (11)$$

The associated forest plot of $D_i / \sqrt{\hat{\sigma}_{ML}^{2T} w_i^T + \hat{\sigma}_{ML}^{2C} w_i^C}$ for our case study is presented in Figure 2.

3.2. Estimation of the overall mean difference

Like in the previous section, the weight of effect size of each study is considered in order to estimate the overall effect size. Here, the individual mean differences are weighted as

$$v_i^* = \frac{1/(\hat{\sigma}_{ML}^{2T} w_i^T + \hat{\sigma}_{ML}^{2C} w_i^C)}{\sum_{i=1}^k 1/(\hat{\sigma}_{ML}^{2T} w_i^T + \hat{\sigma}_{ML}^{2C} w_i^C)}.$$

Note that these weights involve the two variance estimates of the two groups and, hence, are random. This is a considerable difference to the previous case of equal variances. Next, we can compute the overall mean difference as follows:

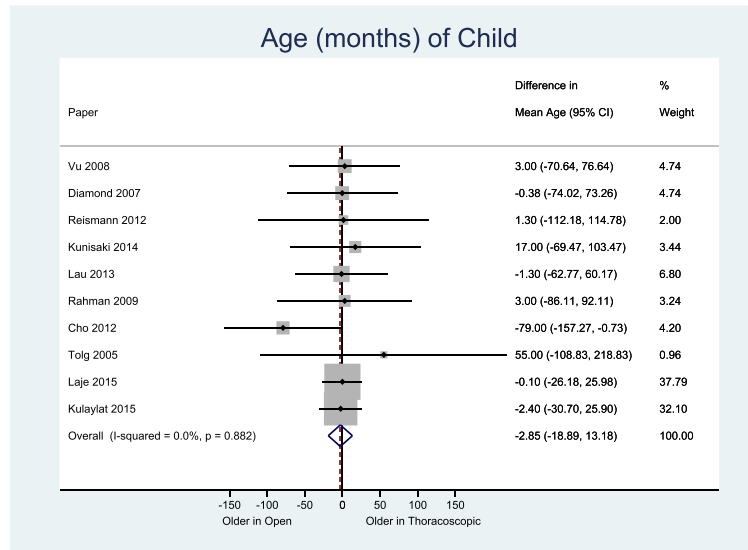


Figure 2. Forest plot for the difference in mean age by Methodology II. [Colour figure can be viewed at wileyonlinelibrary.com]

$$D_{\text{overall}}^* = \sum_{i=1}^k D_i v_i^* \quad (12)$$

with the variance

$$\text{Var}(D_{\text{overall}}^*) = \frac{1}{\sum_{i=1}^k 1/(\sigma^{2T} w_i^T + \sigma^{2C} w_i^C)}.$$

The aforementioned variance is obtained under the assumption that the two sample means are independent. Because we have the variances of the effect size estimates, we can now develop a $(1 - \alpha)100\%$ confidence interval for the overall mean difference. It is given by

$$D_{\text{overall}}^* \pm Z_{\alpha/2} \widehat{SE}(D_{\text{overall}}^*), \quad (13)$$

where $\widehat{SE}(D_{\text{overall}}^*) = \sqrt{1/\sum_{i=1}^k 1/(\hat{\sigma}_{ML}^{2T} w_i^T + \hat{\sigma}_{ML}^{2C} w_i^C)}$ is the estimated standard deviation for the overall mean difference.

For hypothesis testing of the difference between two means, $H_0 : \xi = 0$, the test statistic when the population variances are unequal is

$$Z_{\text{overall}}^* = \frac{D_{\text{overall}}^*}{\sqrt{1/\sum_{i=1}^k 1/(\hat{\sigma}_{ML}^{2T} w_i^T + \hat{\sigma}_{ML}^{2C} w_i^C)}}, \quad (14)$$

where Z_{overall}^* has the standard normal distribution. If $|Z_{\text{overall}}^*| > Z_{\alpha/2}$, the null hypothesis is rejected.

4. Test for equality of variances

Evidently, it is of interest if the two variances can be assumed to be equal or not. This can be accomplished by separately estimating the variances in the two arms and compare the associated likelihoods by means of a likelihood ratio test. As there is no reason to restrict the comparison with those studies where both treatment arms have information, we include all available studies.

Suppose, in generality, we have k_1 studies in the treatment group and k_2 studies in the control group of the same issue. In each group, there are only sample means and sample sizes. Let $\bar{X}_i^T \sim N(\mu^T, \sigma^{2T}/n_i^T)$ for $i = 1, 2, \dots, k_1$ and $\bar{X}_i^C \sim N(\mu^C, \sigma^{2C}/n_i^C)$ for $i = 1, 2, \dots, k_2$. Also, two sample means $\bar{X}_i^T$ and $\bar{X}_i^C$ are independent across groups and studies. We consider a test of the null hypothesis

$$H_0 : \sigma^{2T} = \sigma^{2C} \quad (15)$$

against the alternative hypothesis H_1 that the variances in the treatment and control groups are unequal. In order to construct the likelihood ratio test, the procedure is as follows. From the joint probability density function of $\bar{X}_i^T$ and $\bar{X}_i^C$, the log-likelihood function of $\theta = (\mu^T, \mu^C, \sigma^{2T}, \sigma^{2C})$ is

$$\begin{aligned} \log L(\theta; \bar{X}_i^T, \bar{X}_i^C) = & -\frac{1}{2} \left[\frac{1}{\sigma^{2T}} \sum_{i=1}^{k_1} \left(\frac{\bar{X}_i^T - \mu^T}{\sqrt{w_i^T}} \right)^2 + \frac{1}{\sigma^{2C}} \sum_{i=1}^{k_2} \left(\frac{\bar{X}_i^C - \mu^C}{\sqrt{w_i^C}} \right)^2 \right] \\ & - \frac{k_1 \log \sigma^{2T}}{2} - \frac{k_2 \log \sigma^{2C}}{2} + \text{constant}. \end{aligned}$$

Then, the general likelihood ratio statistic for testing H_0 versus H_1 is defined by

$$\Lambda = (k_1 + k_2) \log \hat{\sigma}_0^2 - k_1 \log \hat{\sigma}_{ML}^{2T} - k_2 \log \hat{\sigma}_{ML}^{2C}, \quad (16)$$

where $\hat{\sigma}_0^2 = (k_1 \hat{\sigma}_{ML}^{2T} + k_2 \hat{\sigma}_{ML}^{2C}) / (k_1 + k_2)$ is the maximum likelihood estimator for the pool variance σ_0^2 under H_0 , and $\hat{\sigma}_{ML}^{2T}$ and $\hat{\sigma}_{ML}^{2C}$ are the likelihood estimators for σ^{2T} and σ^{2C} obtained from (9) and (10), respectively. The details of the proof are presented in appendix C as Theorem 4. Under the null hypothesis, Λ has an approximate chi-square distributed with one degree of freedom as the null-hypothesis lies entirely in the interior of the alternative hypothesis. Therefore, the test rejects H_0 if $\Lambda > \chi_{\alpha, (1)}^2$.

5. Simulation studies

Simulation studies are conducted to evaluate the proposed estimators for the mean difference in meta-analysis and the test for equality of variances using the R statistical language [18]. We simulated the data from a normal distribution under a variety of parameter constellations of mean and variance. The number of studies was varied as $k = 10, 30, 50$ and 100 . The weight of each study, w_i for $i = 1, 2, \dots, k$, was simulated as uniform(0.02, 0.20) in order to produce a reasonable sample size of each study. The setting of the simulation can be distinguished in three cases as follows:

- In the simulation study for Methodology I, the data for the mean difference D_i were generated as independent and identically distributed with $N(\mu, \sigma^2 w_i)$, where $\mu = 0$ and 2 , and $\sigma^2 = 2, 4$ and 9 .
- In the simulation study for Methodology II, we generated the data for the means $\bar{X}_i^T$ and $\bar{X}_i^C$ from $N(\mu^T, \sigma^{2T} w_i^T)$ and $N(\mu^C, \sigma^{2C} w_i^C)$, respectively, where $(\mu^T, \mu^C) = (0, 0)$ and $(4, 2)$, and $(\sigma^{2T}, \sigma^{2C}) = (1, 2), (1, 4)$ and $(2, 9)$. The mean difference was obtained as $D_i = \bar{X}_i^T - \bar{X}_i^C$.
- In the simulation study for the likelihood ratio test, we simulated the means of two treatment groups from the normal distributions as in the second simulation case. The parameters were set to be $(\mu^T, \mu^C) = (0, 0)$ and $(4, 2)$, and $(\sigma^{2T}, \sigma^{2C}) = (\sigma^2, \sigma^2) = (2, 2), (4, 4)$ and $(9, 9)$.

The number of replications was set as 10 000 for each simulation constellation. Bias, estimated standard error, coverage probability and expected length of the estimators for the mean difference were calculated by

$$\begin{aligned} \text{Bias}(\hat{\theta}) &= \hat{E}(\hat{\theta}) - \theta = \frac{1}{10,000} \sum_{i=1}^{10,000} \hat{\theta}_i - \theta, \\ \text{SE}(\hat{\theta}) &= \sqrt{\frac{1}{10,000} \sum_{i=1}^{10,000} (\hat{\theta}_i - \hat{E}(\hat{\theta}))^2}, \\ \text{CP} &= \frac{c(L \leq \theta \leq U)}{10,000}, \end{aligned}$$

and

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

respectively, where $\hat{\theta}$ is the mean difference for parameter θ and $c(L \leq \theta \leq U)$ is the number of simulation runs for θ that lies within the confidence interval. To consider the performance of the likelihood ratio test,

Table III. Bias, standard error, coverage probability and expected length of the estimators for mean difference in the simulation for Methodology I.

<i>k</i>	Distribution of D_i	Point estimator ($D_{overall}$)		Interval estimator	
		Bias	Standard error	Coverage Probability	Expected length
10	N(0, $2w_i$)	0.0013	0.1302	0.9138	0.4837
	N(0, $4w_i$)	0.0001	0.1641	0.9088	0.6050
	N(0, $9w_i$)	−0.0036	0.2534	0.9142	0.9418
	N(2, $2w_i$)	0.0008	0.1202	0.9094	0.4439
	N(2, $4w_i$)	−0.0021	0.1869	0.9069	0.6847
	N(2, $9w_i$)	−0.0012	0.2358	0.9063	0.8658
30	N(0, $2w_i$)	0.0006	0.0724	0.9464	0.2860
	N(0, $4w_i$)	−0.0016	0.0992	0.9403	0.3859
	N(0, $9w_i$)	−0.0004	0.1460	0.9415	0.5734
	N(2, $2w_i$)	−0.0007	0.0762	0.9423	0.2962
	N(2, $4w_i$)	−0.0008	0.1056	0.9428	0.4119
	N(2, $9w_i$)	0.0030	0.1546	0.9428	0.6043
50	N(0, $2w_i$)	0.0005	0.0550	0.9474	0.2155
	N(0, $4w_i$)	0.0018	0.0834	0.9509	0.3344
	N(0, $9w_i$)	0.0015	0.1075	0.9466	0.4249
	N(2, $2w_i$)	0.0005	0.0503	0.9516	0.2008
	N(2, $4w_i$)	0.0007	0.0822	0.9492	0.3264
	N(2, $9w_i$)	0.0002	0.1231	0.9456	0.4802
100	N(0, $2w_i$)	−0.0003	0.0388	0.9499	0.1541
	N(0, $4w_i$)	0.0011	0.0543	0.9509	0.2175
	N(0, $9w_i$)	−0.0003	0.0883	0.9496	0.3474
	N(2, $2w_i$)	0.0000	0.0390	0.9542	0.1557
	N(2, $4w_i$)	0.0009	0.0559	0.9505	0.2217
	N(2, $9w_i$)	0.0007	0.0867	0.9512	0.3488

the type I error probability was used. It was estimated as the proportion of times that $H_0 : \sigma^{2T} = \sigma^{2C}$ is rejected under the assumption that variances between the two treatment groups are equal. Here, the target significant level α was fixed at 0.05. The major findings of the simulation studies can be summarized as follows:

- Table III and Table IV show the performance of the proposed estimators and the confidence intervals for the mean difference from Methodology I and Methodology II, respectively. The results indicate that the bias of estimators tends to be small in general. This means that values of the proposed estimators $D_{overall}$ and $D_{overall}^*$ are very close to the population mean difference. These two estimators have small standard errors, where values of standard error tend to decrease if the number of the study (k) increases.
- For a small number of studies ($k \leq 30$), the coverage probabilities of the proposed confidence intervals for the mean difference shown in equations (5) and (13) are less than the nominal coverage level of 0.95. However, if k increases, the coverage probabilities maintain the target and tend to be greater than 0.95.
- Figure 3 presents type I error probabilities for the general likelihood ratio test. The results show that values of type I error are slightly larger than the target significant level of 0.05 for $k = 10$. If k increases, type I error probabilities become close to 0.05.

In summary, if the number of studies is at least moderately large, the performance of the proposed estimators $D_{overall}$ and $D_{overall}^*$, and corresponding confidence intervals are reasonably well. Moreover, type I error probability of the likelihood ratio test is close to the target significant level.

Additionally, two existing approaches (see also [19, 20]) are used to compare with the method proposed here. We consider three types of scenarios: the within-study variances are available for all studies, the within-study variances are available for some of the studies, and no variance information is available for all studies (our case). For Scenario 1, the number of studies is reduced if there is missing variance information. Thus, only complete studies are used in the analysis. This is done in practice for complete case analysis. The overall mean difference of this scenario can be computed using a meta-analysis based on the conventional fixed effect model, as the sample means and sample variances of the two treatment

Table IV. Bias, standard error, coverage probability and expected length of the estimators for mean difference in the simulation for Methodology II.

<i>k</i>	Distribution of D_i	Point estimator ($D_{overall}^*$)		Interval estimator	
		Bias	Standard error	Coverage Probability	Expected length
10	$N(0, w_i^T + 2w_i^C)$	-0.0024	0.1814	0.9187	0.6696
	$N(0, w_i^T + 4w_i^C)$	-0.0004	0.2010	0.9285	0.7563
	$N(0, 2w_i^T + 9w_i^C)$	-0.0015	0.3455	0.9218	1.2930
	$N(2, w_i^T + 2w_i^C)$	0.0006	0.1771	0.9258	0.6590
	$N(2, w_i^T + 4w_i^C)$	0.0025	0.2391	0.9210	0.8792
	$N(2, 2w_i^T + 9w_i^C)$	0.0010	0.3472	0.9173	1.2855
30	$N(0, w_i^T + 2w_i^C)$	-0.0015	0.1010	0.9453	0.3943
	$N(0, w_i^T + 4w_i^C)$	-0.0017	0.1243	0.9476	0.4872
	$N(0, 2w_i^T + 9w_i^C)$	0.0041	0.1732	0.9450	0.6711
	$N(2, w_i^T + 2w_i^C)$	-0.0014	0.0961	0.9457	0.3777
	$N(2, w_i^T + 4w_i^C)$	-0.0016	0.1131	0.9473	0.4449
	$N(2, 2w_i^T + 9w_i^C)$	-0.0014	0.1851	0.9460	0.7266
50	$N(0, w_i^T + 2w_i^C)$	-0.0011	0.0787	0.9501	0.3092
	$N(0, w_i^T + 4w_i^C)$	0.0003	0.0979	0.9500	0.3875
	$N(0, 2w_i^T + 9w_i^C)$	0.0005	0.1437	0.9496	0.5676
	$N(2, w_i^T + 2w_i^C)$	-0.0003	0.0730	0.9502	0.2874
	$N(2, w_i^T + 4w_i^C)$	0.0017	0.0971	0.9528	0.3851
	$N(2, 2w_i^T + 9w_i^C)$	-0.0014	0.1425	0.9476	0.5630
100	$N(0, w_i^T + 2w_i^C)$	0.0002	0.0539	0.9491	0.2132
	$N(0, w_i^T + 4w_i^C)$	0.0010	0.0651	0.9507	0.2582
	$N(0, 2w_i^T + 9w_i^C)$	0.0001	0.0966	0.9529	0.3852
	$N(2, w_i^T + 2w_i^C)$	-0.0011	0.0536	0.9522	0.2116
	$N(2, w_i^T + 4w_i^C)$	0.0007	0.0656	0.9554	0.2652
	$N(2, 2w_i^T + 9w_i^C)$	-0.0005	0.0960	0.9529	0.3831

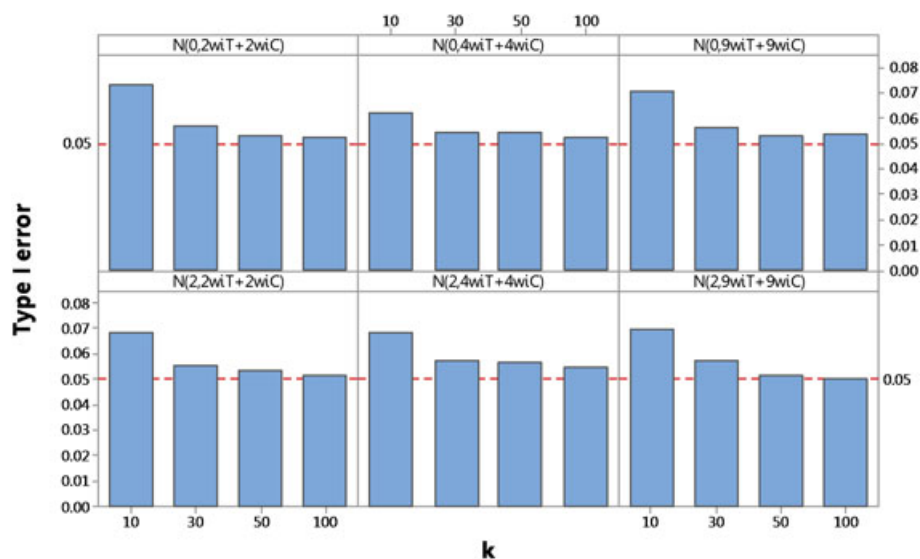


Figure 3. Type I error for the general likelihood ratio test in the simulation for various settings of distribution of D_i and number of study (k) AD: Anderson-Darling. [Colour figure can be viewed at wileyonlinelibrary.com]

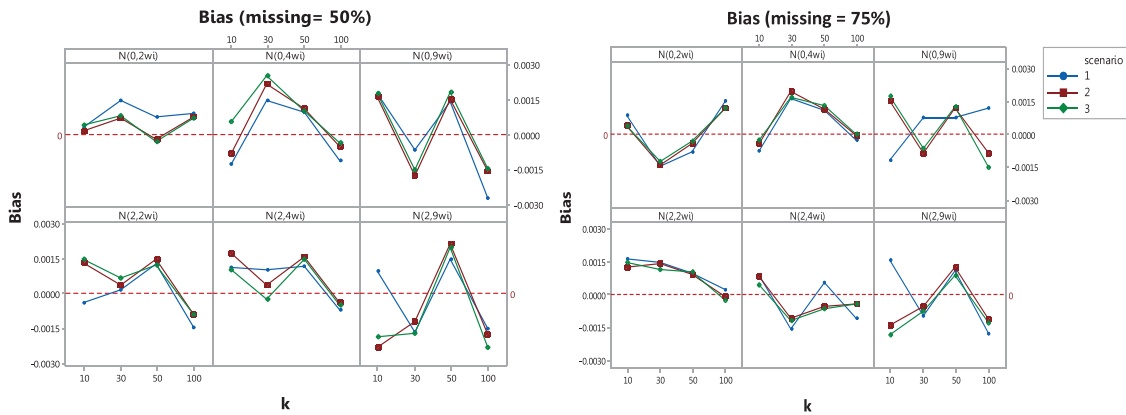


Figure 4. Bias of the overall mean difference from Scenarios 1 and 2 with 50% and 75% studies missing variances and Scenario 3 with no variance information. [Colour figure can be viewed at wileyonlinelibrary.com]

arms in the study i are reported, and the pooled variance estimate can be readily provided as $S_{p,i}^2 = ((n_i^T - 1)S_i^{2T} + (n_i^C - 1)S_i^{2C}) / (n_i^T + n_i^C - 2)$. We note in passing, that also this setting would allow incorporating potential heterogeneity by means of the random effects model, but we will not use this here for better comparability. For the partial information setting in Scenario 2, the Expectation-Maximization (EM) algorithm [21] is utilized in principle to impute the missing variances. The details of this method are as follows. In the E-step, the missing variances are imputed as

$$E(S_i^{2T} | S_i^{2C}) = \frac{\sigma_p^{2(t)}(n_i^T + n_i^C - 2)}{n_i^T - 1} - \frac{(n_i^C - 1)S_i^{2C}}{n_i^T - 1}$$

and

$$E(S_i^{2C} | S_i^{2T}) = \frac{\sigma_p^{2(t)}(n_i^T + n_i^C - 2)}{n_i^C - 1} - \frac{(n_i^T - 1)S_i^{2T}}{n_i^C - 1}.$$

For update on the overall pooled variance σ_p^2 in the M-step (see also [9]), a weighted average $\sigma_p^{2(t+1)} = \sum_{i=1}^k S_{p,i}^2 (n_i^T + n_i^C - 2) / \sum_{i=1}^k (n_i^T + n_i^C - 2)$ of the pooled variances $S_{p,i}^2$ in the i -th trial is determined, where t is the t th iteration. The iterative procedure is terminated if $|\sigma_p^{2(t+1)} - \sigma_p^{2(t)}| < 0.0001$. Then, the overall mean difference is computed using the meta-analysis approach described in the previous scenario.

The simulation studies are performed to evaluate the performance of the overall mean difference under the three scenarios. We simulated $\bar{X}_i^T$ and $\bar{X}_i^C$ from $N(\mu^T, \sigma^{2T} w_i^T)$ and $N(\mu^C, \sigma^{2C} w_i^C)$, respectively, where $(\mu^T, \mu^C) = (0, 0)$ and $(4, 2)$, and $(\sigma^{2T}, \sigma^{2C}) = (\sigma^2, \sigma^2) = (2, 2), (4, 4)$ and $(9, 9)$. We then calculated the mean difference from $D_i = \bar{X}_i^T - \bar{X}_i^C$. For Scenario 1 and Scenario 2, the variances S_i^{2T} and S_i^{2C} are generated from $\chi_{n_i^T-1}^2 \sigma^{2T} / (n_i^T - 1)$ and $\chi_{n_i^C-1}^2 \sigma^{2C} / (n_i^C - 1)$, respectively, where χ^2 is the quantile of the chi-squared distribution. The percentage of studies with missing variances was 50% and 75%. Given these scenarios, the results of the simulation studies were analyzed with respect of bias and standard error. The bias and estimated standard error of the overall mean difference computed from the three scenarios are shown in Figures 4 and 5. It can be concluded that the bias of the estimators is close to the population mean difference for all cases of the studies. The overall mean difference of the proposed method presented in Scenario 3 provides the standard errors slightly smaller than that of Scenario 2. The estimator derived from Scenario 1 gives largest variances.

6. Case studies

6.1. Meta-analysis of thoracoscopic and open resections for treating congenital lung malformation

To illustrate the estimators presented in this paper, we use the meta-analytic data on the two surgeries for treating congenital lung malformation, the thoracoscopic and the open operations, as mentioned in the beginning section. This data set comprises 38 published reports containing the mean of variables of

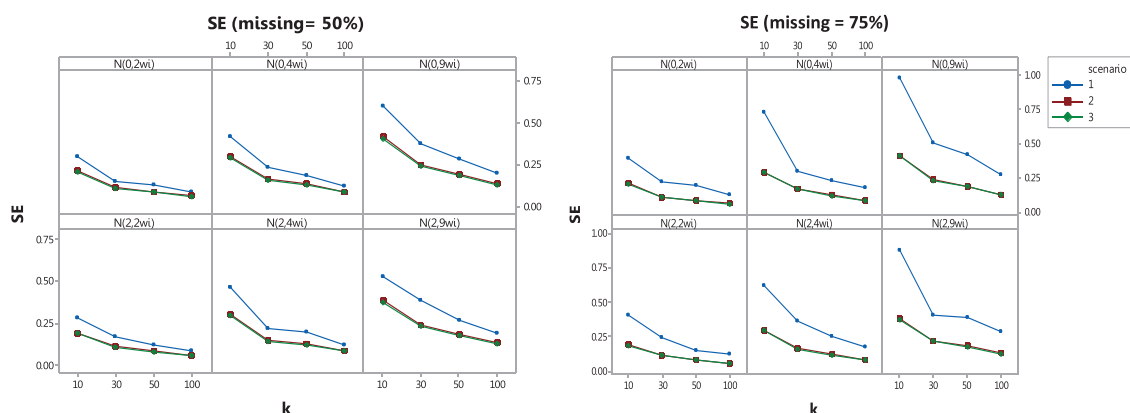


Figure 5. Standard error of the overall mean difference from Scenarios 1 and 2 with 50% and 75% studies missing variances and Scenario 3 with no variance information. [Colour figure can be viewed at [wileyonlinelibrary.com](#)]

Table V. Test for equality of variances for $\alpha = 0.05$.

Variables	k_1	k_2	Λ
Length of operation (min)	23	13	2.62
Length of stay (days)	25	18	9.17
Number of chest tube days (days)	20	11	3.32
Weight of child (kilograms)	15	8	9.47
Age of child (months)	24	20	0.70

Table VI. Results of the study in the clinical trials between the thoracoscopic and the open operations.

Variables	Overall mean difference	Standard Error	95% Confidence interval	Test of equal means
Methodology I:				
Length of operation	36.97	9.19	(18.96, 54.99)	4.03
Length of stay	-1.38	0.52	(-2.40, -0.37)	-2.65
Number of chest tube days	-0.79	0.41	(-1.59, 0.02)	-1.93
Weight of child	0.40	0.31	(-0.21, 1.01)	1.29
Age of child	-1.43	4.83	(-10.89, 8.03)	-0.30
Methodology II:				
Length of operation	36.75	13.14	(11.00, 62.49)	2.80
Length of stay	-1.28	0.90	(-3.05, 0.49)	-1.42
Number of chest tube days	-0.71	0.71	(-2.10, 0.68)	-1.00
Weight of child	0.40	0.75	(-1.07, 1.86)	0.53
Age of child	-2.85	8.18	(-18.89, 13.18)	-0.35

interest such as length of operation, number of chest tube days, length of stay, age, and weight and the number of the patients in each study arm. Some of these reports have only information on one arm (see e.g. Table II).

The test for equal variances presented in Section 4 is conducted prior to further statistical analysis. The results in Table V show that there are two variables, length of stay and weight of child, for which we have to reject the null hypothesis of equal variances between the thoracoscopic and the open study arms.

Next, we use the add-on package METAN of the statistical package STATA14™ [22] to estimate the overall mean difference of variables. The statistics obtained from the Methodology I and the Methodology II are presented in Table VI. The results indicate that the mean of length of operation in the thoracoscopic lung resection is longer than the open procedure. Furthermore, there is evidence that length of stay and number of chest tube days is better for thoracoscopic, significantly so when using Methodology I, although the likelihood ratio test suggests to use Methodology II for length of stay. In

Table VII. The mean change in serum creatine after living kidney donation.

Study	<i>n</i>	mean change in serum creatinine
Johnson <i>et al.</i> 2005	78	23
Mimran <i>et al.</i> 1993	18	28
Sobh <i>et al.</i> 1989	45	27
Kostakis <i>et al.</i> 1997	255	20
Hida <i>et al.</i> 1982	34	29
Rizvi <i>et al.</i> 2005	736	18
Tondo <i>et al.</i> 1998	10	—
Abomelha <i>et al.</i> 1993	70	11
Siebels <i>et al.</i> 2003	122	27
Edgren <i>et al.</i> 1976	46	24
Basseri <i>et al.</i> 1995	87	23
Enger 1973	13	25
Mendoza <i>et al.</i> 1987	152	4
Fourcade <i>et al.</i> 2002	99	23
ter Wee <i>et al.</i> 1994	12	32
O'Donnell <i>et al.</i> 1986	33	14
Laskow <i>et al.</i> 1991	48	31
Miller <i>et al.</i> 1985	40	18
Rodriguez <i>et al.</i> 1985	25	—
Marekovic <i>et al.</i> 1992	50	23
Prandini <i>et al.</i> 1987	32	—
Chen <i>et al.</i> 1992	77	23
D Almeida <i>et al.</i> 1996	110	—
Wiesel <i>et al.</i> 1997	67	27
Najarian <i>et al.</i> 1992	51	20
Toronyi <i>et al.</i> 1998	30	—
Schostak <i>et al.</i> 2004	50	—
Haberal <i>et al.</i> 1998	102	5
Friedlander <i>et al.</i> 1988	17	29
Undurraga <i>et al.</i> 1998	74	12
Eberhardi <i>et al.</i> 1997	29	—
Fehman-Ekholm <i>et al.</i> 2001	348	—
Gonzalez <i>et al.</i> 1989	22	15
Dunn <i>et al.</i> 1986	180	15
Beekman <i>et al.</i> 1994	35	24
Goldfarb <i>et al.</i> 2001	70	18
Williams <i>et al.</i> 1986	24	3
Watnick <i>et al.</i> 1988	29	—
Talseth <i>et al.</i> 1986	68	9
Iglesias-Marquez <i>et al.</i> 2001	20	12

Table VIII. Results of the study for changes in renal function after living kidney donation.

Variables	Overall mean difference	Standard error	95% Confidence interval	Test of equal means
Methodology I:				
Glomerular filtration rate (mL/min)	−21.42	2.51	(−26.35, −16.50)	−8.53
Serum creatinine (μmol/L)	18.21	1.18	(15.89, 20.52)	15.43
Systolic blood pressure (mmHg)	2.65	1.41	(−0.11, 5.42)	1.88
Diastolic blood pressure (mmHg)	3.38	0.73	(1.96, 4.80)	4.63

addition, weight and age of child are not different between the two surgeries which implies that these two covariates will not bias the results for the variables of interest.

Note that the completion of the meta-analysis is straightforward if an estimate for the study-specific mean difference has been obtained. Forest plots for the mean difference of age using Methodology I and Methodology II are shown in Figures 1 and 2, respectively.

6.2. Meta-analysis on changes in renal function after living kidney donation

Philbrook *et al.* [7] presented data that concern changes in continuous outcomes such as glomerular filtration rate (mL/min per 1.73m²), serum creatinine (μmol/L), systolic blood pressure (mmHg) and diastolic blood pressure (mmHg) after living kidney donation. For these data, they reported the mean changes between pre-donation and post-donation ($\bar{X}_i^{post} - \bar{X}_i^{pre}$) and sample sizes ($n_i^{post} = n_i^{pre} = n_i = 1/w_i$) for each study, whereas a variance estimate is often not reported (see e.g. Table VII). Instead of working with pairs of pre-donation and post-donation measurements, we take pairwise differences and base inference on $D_i/\sqrt{\hat{\sigma}^2 w_i}$, $i = 1, 2, \dots, k$ which allows the use of Methodology I for this case study.

The statistics obtained from Methodology I are shown in Table VIII. It can be concluded that glomerular filtration rate decreases after donation. In contrast, serum creatinine and diastolic blood pressure increase after donation. These results do not differ from the findings in meta-analysis presented in the paper of Philbrook *et al.* [7] which uses various imputation methods to cope with missing standard errors in the studies. Finally, our method confirms that there is no change between before and after donation in systolic blood pressure. A forest plot for the change in serum creatinine is shown in Figure 6. We also report probability plots of the standardized quantity $D_i/\sqrt{\hat{\sigma}^2 w_i}$, $i = 1, 2, \dots, k$, for continuous variables in this example (Figure 7). The assumption of a homogeneous normal distribution is supported for serum creatinine, systolic blood pressure, and diastolic blood pressure.

7. Discussion and concluding remarks

Meta-analysis has become a standard instrument in the tool-box for systematically evaluating medical research. In this paper, we focus on estimating the mean difference as the effect size, where only sample means and sample sizes are reported for the study arms, but not variance estimates. Two approaches are introduced depending on whether the variances in the study arms are assumed to be equal or not. The first approach called Methodology I is the estimation of the overall mean difference when the unknown study variances are assumed to be equal. We estimate the pooled variance and then derive the estimated variance of the mean difference in order to provide an overall estimate. The other method called Methodology II is derived under the assumption that the population variances are unequal. For this method, we separately estimate variances of two treatment arms. Then, we find the overall mean difference with an associated variance estimate. Furthermore, a likelihood ratio test is introduced for testing the equality of variances.

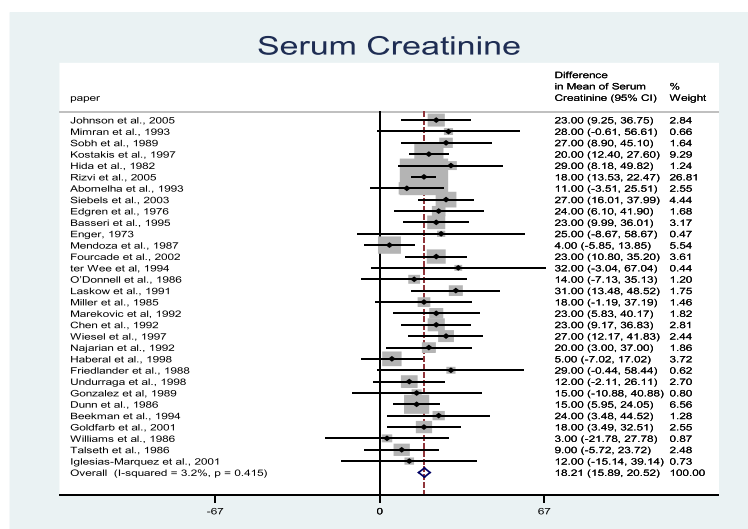


Figure 6. Forest plot for the mean change in Serum Creatinine (μmol/L) after living kidney donation. [Colour figure can be viewed at wileyonlinelibrary.com]

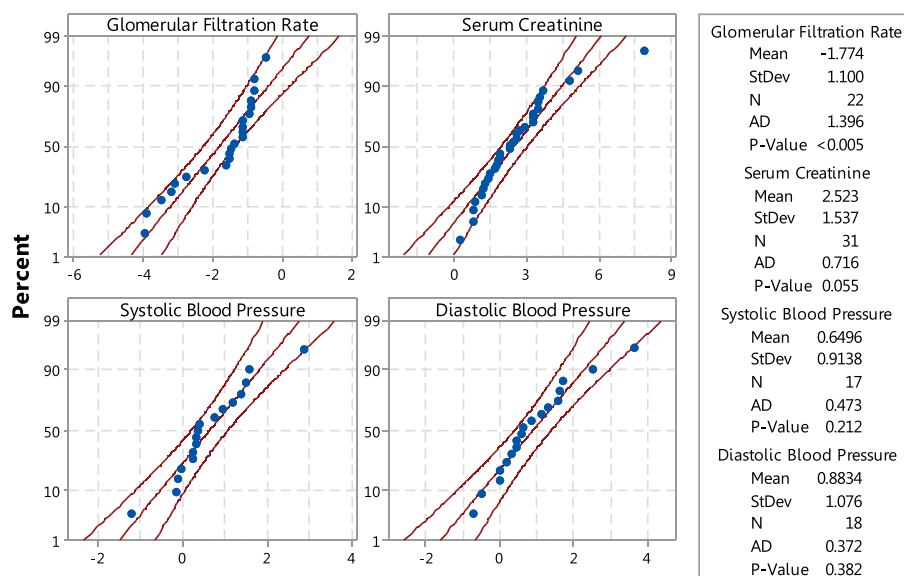


Figure 7. Probability plot based upon $D_i/\sqrt{\hat{\sigma}^2 w_i}$ for Glomerular filtration rate, Serum creatinine, Systolic blood pressure and Diastolic blood pressure. [Colour figure can be viewed at [wileyonlinelibrary.com](#)]

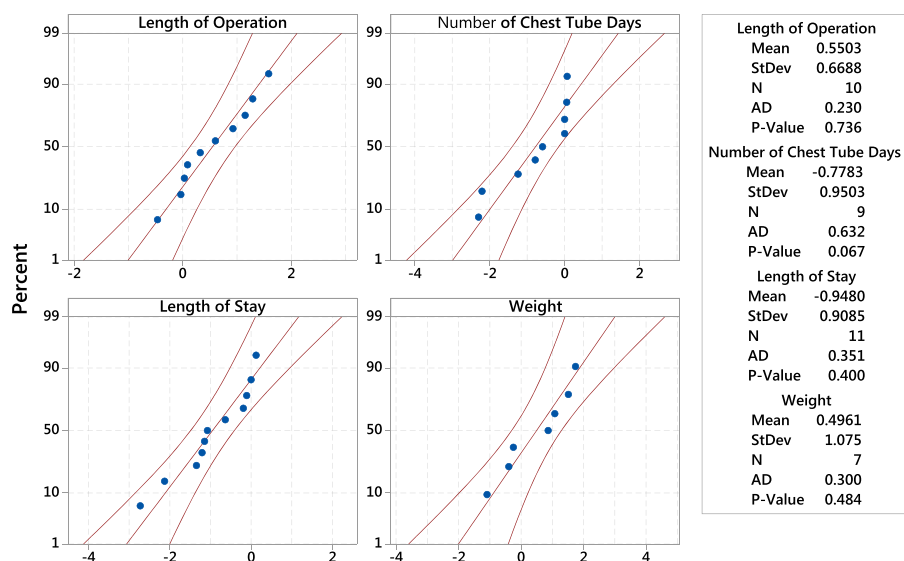


Figure 8. Probability plot based upon $D_i/\sqrt{\hat{\sigma}^2 w_i}$ for Length of operation, Number of chest tube days, Length of stay and Weight of child. [Colour figure can be viewed at [wileyonlinelibrary.com](#)]

Both developed methodologies build on the assumption of equal variances within arms and across studies and no heterogeneity between studies. The latter is a serious restriction but unavoidable given the restricted data available. As we estimate a free variance parameter, we have some freedom in coping with mild forms of heterogeneity. Recall that in a conventional meta-analysis also, only one variance parameter is estimated so that the degree of complexity in the variance structure is identical. In Figure 8, we provide probability plots of the standardized quantity $D_i/\sqrt{\hat{\sigma}^2 w_i}$, $i = 1, 2, \dots, k$, for length of operation, number of chest tube days, length of stay and weight in the case study underlying this work. There is no evidence that assumption of a homogeneous normal distribution is violated. Hence, at least for the case study considered here, the approach seems reasonable.

The performance of the proposed estimators for the difference of means and the test for equality of variances have been investigated in a simulation. The bias of our estimators appears to be small in all cases of the study. The proposed estimators have small standard errors, especially for large number of the

studies. These empirical findings agree with asymptotic results provided in the appendix B. The confidence intervals for the mean difference also provide coverage probabilities close to the nominal coverage level if number of the studies become large ($k > 30$). Complete case analysis and imputation approach for estimating the missing sample variances are not performing as well as the proposed method, in particular the complete case approach suffers under considerable larger variance. As can be seen, although all methods provide the overall mean difference very close to the population but the existing approaches have larger standard errors. For testing of equality of variances, the general likelihood ratio test can control type I error probability as 0.05 for large number of the studies, which also agrees with the asymptotic distributional results mentioned in appendix C. As the proposed approach follows a normal distributional setting, it is easily possible to extend it to include covariate information on study level such as treatment modification or locality. Finally, we have focussed here on performing fixed effects meta-analysis (estimation of one true effect for all studies) when estimated variances are not available in the studies. In further research, random effects modelling will be considered to investigate for unobserved heterogeneity in addition to the before-mentioned inclusion of observed heterogeneity in form of covariate information.

We also mention that R-code for computing the case studies and simulation results as well as more detailed results from simulation work is available as supplementary material.

Appendix

In the following, we are detailing some inferential properties and results for our meta-analytic setting which can be achieved, with appropriate modification and adaptation, from general likelihood inference as provided, for example, in Casella and Berger [23].

Appendix A. Standard errors of estimators

Theorem 1

If $D_i \sim N(\mu, \sigma^2 w_i)$ for $i = 1, 2, \dots, k$ with maximum likelihood estimators of μ and σ^2 as $\hat{\mu}_{ML} = \frac{\sum_{i=1}^k D_i/w_i}{\sum_{i=1}^k 1/w_i}$ and $\hat{\sigma}_{ML}^2 = \frac{1}{k} \sum_{i=1}^k (D_i - \hat{\mu}_{ML})^2/w_i$, respectively, then the standard errors are

1. $SE(\hat{\mu}_{ML}) = \sqrt{\sigma^2 / \sum_{i=1}^k 1/w_i}$ for $\hat{\mu}_{ML}$
2. $SE(\hat{\sigma}_{ML}^2) = \sqrt{2(k-1)\sigma^4/k^2}$ for $\hat{\sigma}_{ML}^2$.

Proof

For the first part, it is easy to see that

$$Var(\hat{\mu}_{ML}) = \frac{Var(\sum_{i=1}^k D_i/w_i)}{(\sum_{i=1}^k 1/w_i)^2} = \frac{\sigma^2}{\sum_{i=1}^k 1/w_i},$$

so it is clear for $SE(\hat{\mu}_{ML}) = \sqrt{Var(\hat{\mu}_{ML})}$. For the second part, we first consider

$$\begin{aligned} \sum_{i=1}^k \frac{(D_i - \mu)^2}{w_i \sigma^2} &= \sum_{i=1}^k \frac{(D_i - \hat{\mu}_{ML} + \hat{\mu}_{ML} - \mu)^2}{w_i \sigma^2} \\ &= \sum_{i=1}^k \left(\frac{D_i - \hat{\mu}_{ML}}{\sigma \sqrt{w_i}} \right)^2 + \sum_{i=1}^k \frac{1}{w_i} \left(\frac{\hat{\mu}_{ML} - \mu}{\sigma} \right)^2. \end{aligned}$$

It can be simply written as

$$\sum_{i=1}^k \left(\frac{D_i - \mu}{\sigma \sqrt{w_i}} \right)^2 = \sum_{i=1}^k \left(\frac{D_i - \hat{\mu}_{ML}}{\sigma \sqrt{w_i}} \right)^2 + \left(\frac{\sqrt{\sum_{i=1}^k (1/w_i)} (\hat{\mu}_{ML} - \mu)}{\sigma} \right)^2.$$

Because $D_i \sim N(\mu, \sigma^2 w_i)$, the term on the left hand side of the aforementioned equation has a chi-square distribution with k degrees of freedom. The second term on the right hand side has a chi-square

distribution with one degree of freedom. Hence, the first term on the right hand side $\sum_{i=1}^k \left(\frac{D_i - \hat{\mu}_{ML}}{\sigma \sqrt{w_i}} \right)^2$ of the aforementioned equation must be a chi-square with $k - 1$ degrees of freedom. Therefore, we achieve that the variance of $\hat{\sigma}_{ML}^2$ is

$$\text{Var}(\hat{\sigma}_{ML}^2) = \frac{1}{k^2} \text{Var} \left(\sum_{i=1}^k \frac{(D_i - \hat{\mu}_{ML})^2}{w_i} \right) = \frac{2(k-1)\sigma^4}{k^2},$$

which ends the proof. $\square$

Appendix B. Properties of estimators

Theorem 2

If $D_i \sim N(\mu, \sigma^2 w_i)$ for $i = 1, 2, \dots, k$ with maximum likelihood estimator of μ as $\hat{\mu}_{ML} = \frac{\sum_{i=1}^k D_i / w_i}{\sum_{i=1}^k 1 / w_i}$, then $\hat{\mu}_{ML}$ is unbiased and consistent for μ .

Proof

To prove the unbiasedness of an estimator, we need to show the identity $E(\hat{\mu}_{ML}) - \mu = 0$. Because $E(D_i) = \mu$, it can be seen that $E(\hat{\mu}_{ML}) = \mu$. Thus, $\hat{\mu}_{ML}$ is an unbiased estimator of μ . To prove the consistency, the sufficient condition is to show that $\lim_{k \rightarrow \infty} E(\hat{\mu}_{ML}) = \mu$ and $\lim_{k \rightarrow \infty} \text{Var}(\hat{\mu}_{ML}) = 0$. As $\text{Var}(\hat{\mu}_{ML}) = \sigma^2 / \sum_{i=1}^k 1 / w_i$, it is easy to see that the sufficient condition of consistency is satisfied. Therefore, $\hat{\mu}_{ML}$ is a consistent estimator of μ . $\square$

Theorem 3

If $D_i \sim N(\mu, \sigma^2 w_i)$ for $i = 1, 2, \dots, k$ with the maximum likelihood estimator of σ^2 as $\hat{\sigma}_{ML}^2 = \frac{1}{k} \sum_{i=1}^k (D_i - \hat{\mu}_{ML})^2 / w_i$, then $\hat{\sigma}_{ML}^2$ is asymptotically unbiased and consistent for σ^2 .

Proof

We find the expectation and variance of $\hat{\sigma}_{ML}^2$, respectively, as

$$E(\hat{\sigma}_{ML}^2) = \frac{1}{k} E \left(\sum_{i=1}^k \frac{(D_i - \hat{\mu}_{ML})^2}{w_i} \right) = \frac{(k-1)\sigma^2}{k}$$

and

$$\text{Var}(\hat{\sigma}_{ML}^2) = \frac{2(k-1)\sigma^4}{k^2}.$$

Here, the bias of $\hat{\sigma}_{ML}^2$ is $-\sigma^2/k$. Taking the limit, we have $\lim_{k \rightarrow \infty} E(\hat{\sigma}_{ML}^2) = \sigma^2$ and $\lim_{k \rightarrow \infty} \text{Var}(\hat{\sigma}_{ML}^2) = 0$. Therefore, $\hat{\sigma}_{ML}^2$ is asymptotically unbiased and consistent for σ^2 . $\square$

Appendix C. General likelihood ratio test

Theorem 4

Suppose that $\bar{X}_i^T \sim N(\mu^T, \sigma^{2T} w_i^T)$, where $w_i^T = 1/n_i^T$ for $i = 1, 2, \dots, k_1$ and $\bar{X}_i^C \sim N(\mu^C, \sigma^{2C} w_i^C)$, where $w_i^C = 1/n_i^C$ for $i = 1, 2, \dots, k_2$. The general likelihood ratio test statistic for the null hypothesis $H_0 : \sigma^{2T} = \sigma^{2C}$ versus the alternative hypothesis $H_1 : \sigma^{2T} \neq \sigma^{2C}$ is given by

$$\Lambda = (k_1 + k_2) \ln \hat{\sigma}_0^2 - k_1 \ln \hat{\sigma}_{ML}^{2T} - k_2 \ln \hat{\sigma}_{ML}^{2C}.$$

Proof

Let $\Omega = \{\theta; -\infty < \mu^T, \mu^C < \infty, \sigma^{2T}, \sigma^{2C} > 0\}$ be the overall parameter space, and $\omega = \{\theta; -\infty < \mu^T, \mu^C < \infty, \sigma^{2T} = \sigma^{2C} = \sigma_0^2 > 0\}$ be the parameter space under H_0 , where $\theta = (\mu^T, \mu^C, \sigma^{2T}, \sigma^{2C})$.

Because $\bar{X}_i^T$ and $\bar{X}_i^C$ are independent, the joint probability density function of $\bar{X}_i^T$ and $\bar{X}_i^C$ is $f(\bar{X}_i^T, \bar{X}_i^C) = f(\bar{X}_i^T)f(\bar{X}_i^C)$. Here, the log-likelihood function of θ is derived as

$$\log L = -\frac{1}{2} \left[\sum_{i=1}^{k_1} \left(\frac{\bar{X}_i^T - \mu^T}{\sigma^T \sqrt{w_i^T}} \right)^2 + \sum_{i=1}^{k_2} \left(\frac{\bar{X}_i^C - \mu^C}{\sigma^C \sqrt{w_i^C}} \right)^2 \right] - \frac{k_1 \log \sigma^{2T}}{2} - \frac{k_2 \log \sigma^{2C}}{2} \\ - \frac{(k_1 + k_2) \log 2\pi}{2} - \log \prod_{i=1}^{k_1} \sqrt{w_i^T} - \log \prod_{i=1}^{k_2} \sqrt{w_i^C},$$

with the maximum likelihood estimators $\hat{\mu}_{ML}^T$, $\hat{\mu}_{ML}^C$, $\hat{\sigma}_{ML}^{2T}$, and $\hat{\sigma}_{ML}^{2C}$. Next, consider the log-likelihood function under H_0 , we have

$$\log L = -\frac{1}{2\sigma_0^2} \left[\sum_{i=1}^{k_1} \left(\frac{\bar{X}_i^T - \mu^T}{\sqrt{w_i^T}} \right)^2 + \sum_{i=1}^{k_2} \left(\frac{\bar{X}_i^C - \mu^C}{\sqrt{w_i^C}} \right)^2 \right] - \frac{(k_1 + k_2) \log \sigma_0^2}{2} \\ - \frac{(k_1 + k_2) \log 2\pi}{2} - \log \prod_{i=1}^{k_1} \sqrt{w_i^T} - \log \prod_{i=1}^{k_2} \sqrt{w_i^C},$$

where the maximum likelihood estimator for σ_0^2 is given by $\hat{\sigma}_0^2 = (k_1 \hat{\sigma}_{ML}^{2T} + k_2 \hat{\sigma}_{ML}^{2C}) / (k_1 + k_2)$. The likelihood ratio test can be obtained from

$$\lambda = \frac{\sup_{\theta \in \omega} L(\theta)}{\sup_{\theta \in \Omega} L(\theta)},$$

[23]. Substituting the maximum likelihood estimators into the likelihood functions, we have

$$\sup_{\theta \in \omega} = \frac{\exp\{-(k_1 + k_2)/2\}}{(2\pi \hat{\sigma}_0^2)^{(k_1 + k_2)/2} \prod_{i=1}^{k_1} \sqrt{w_i^T} \prod_{i=1}^{k_2} \sqrt{w_i^C}}$$

and

$$\sup_{\theta \in \Omega} = \frac{\exp\{-(k_1 + k_2)/2\}}{(\hat{\sigma}_{ML}^{2T})^{k_1/2} (\hat{\sigma}_{ML}^{2C})^{k_2/2} (2\pi)^{(k_1 + k_2)/2} \prod_{i=1}^{k_1} \sqrt{w_i^T} \prod_{i=1}^{k_2} \sqrt{w_i^C}}.$$

Therefore, the likelihood ratio test statistic for testing H_0 versus H_1 is given by

$$\lambda = \frac{(\hat{\sigma}_{ML}^{2T})^{k_1/2} (\hat{\sigma}_{ML}^{2C})^{k_2/2}}{(\hat{\sigma}_0^2)^{(k_1 + k_2)/2}}.$$

It is easy to obtain $\Lambda = -2 \log \lambda$, which is the general likelihood ratio test statistic. The test rejects H_0 if $\Lambda > \chi_{\alpha, (1)}^2$. The proof is therefore complete. $\square$

Acknowledgements

Patarawan Sangnawakij is deeply grateful for receiving funding from Thammasat University (Thailand) to undertake this research. She is also grateful to Southampton Statistical Sciences Research Institute for hosting her visit in the UK. All authors also thank the reviewers for valuable suggestions and comments which lead to considerable improvements in the manuscript.

References

1. Petitti DB. *Meta-Analysis Decision Analysis and Cost-Effectiveness Analysis*. Oxford University Press: New York, 1994.
2. Böhning D, Malzahn U, Schlattmann P, Dammann U P, Mehnert W, Holling H, Schulze R. The application of statistical methods of meta-analysis for heterogeneity modelling in medicine and pharmacy, psychology, quality control and assurance. *Mathematics - Key Technology for the Future* 2003; **VIII**:533–553.
3. Schulze R, Holling H, Böhning D. *Meta-Analysis: New Developments and Applications in Medical and Social Sciences*. Hogrefe & Huber: Göttingen, 2003.

4. Bhning D, Kuhnert R, Rattanasiri S. *Meta-Analysis of Binary Data Using Profile Likelihood*. Chapman & Hall: London, 2008.
5. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Wiley: New York, 2000.
6. Egger M, Smith GD, Aliman DG. *Systematic Reviews in Health Care: Meta-Analysis in Context*. BMJ Publishing Group: London, 2001.
7. Philbrook HT, Barrowmanb N, Garga AX. Imputing variance estimates do not alter the conclusions of a meta-analysis with continuous outcomes: a case study of changes in renal function after living kidney donation. *Journal of Clinical Epidemiology* 2007; **60**(3):228–240.
8. Idris NRN, Robertson C. The effects of imputing the missing standard deviations on the standard error of meta analysis estimates. *Communications in Statistics - Simulation and Computation* 2009; **38**(3):513–526.
9. Chowdhry AK, Dworkinb RH, McDermott MP. Meta-analysis with missing study-level sample variance data. *Statistics in Medicine* 2016; **35**(17):3021–3032.
10. Stocker LJ, Wellesley DG, Stanton MP, Parasuraman R, Howe DT. The increasing incidence of fetal echogenic congenital lung malformations: an observational study. *Prenat Diagn* 2015; **35**(2):148–53.
11. Stanton M, Njere I, Ade-Ajayi N, Patel S, Davenport M. Systematic review and meta-analysis of the postnatal management of congenital cystic lung lesions. *Journal of Pediatric Surgery* 2009; **44**(5):1027–1033.
12. Albanese CT, Sydorak RM, Tsao K, Lee H. Thoracoscopic lobectomy for prenatally diagnosed lung lesions. *Journal of Pediatric Surgery* 2003; **38**(4):553–555.
13. Hall NJ, Chiu PP, Langer JC. Morbidity after elective resection of prenatally diagnosed asymptomatic congenital pulmonary airway malformations. *Pediatr Pulmonology* 2016; **51**(5):525–530.
14. Röver C, Andreas S, Friede T. Evidence synthesis for count distributions based on heterogeneous and incomplete aggregated data. *Biometrical Journal* 2016; **58**(1):170–185.
15. Kulinskaya E, Morgenthaler S, Staudte RG. *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*. John Wiley & Sons: Chichester, 2008.
16. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. John Wiley & Sons: Chichester, 2009.
17. Stangl DK, Berry DA. *Meta-Analysis in Medicine and Health Policy*. Marcel Dekker: Basel, 2000.
18. Venables WN, Smith DM the R Core Team. *Notes on R: A Programming Environment for Data Analysis and Graphics*. <http://cran.r-project.org/> [Accessed on 9 March 2016].
19. Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of Clinical Epidemiology* 2006; **59**:7–10.
20. Idris NRN, Abdullah MH, Tolos SM. Performance of selected imputation techniques for missing variances in meta-analysis. *Journal of Physics: Conference Series* 2013; **435**:1–9.
21. McLachlan G, Krishnan T. *The EM Algorithm and Extensions*. John Wiley & Sons: New Jersey, 2007.
22. Sterne J. *Meta-Analysis in Stata: An Updated Collection from the Stata Journal*. Stata Press: Texas, 2009.
23. Casella G, Berger R L. *Statistical Inference*. Duxbury Press: California, 2002.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.