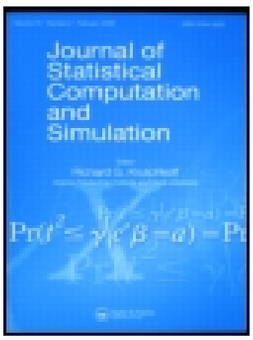

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

A simple test for the difference of means in meta-analysis when study-specific variances are unreported

Patarawan Sangnawakij & Dankmar Böhning

To cite this article: Patarawan Sangnawakij & Dankmar Böhning (2020): A simple test for the difference of means in meta-analysis when study-specific variances are unreported, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2020.1780235

To link to this article: https://doi.org/10.1080/00949655.2020.1780235

Published online: 19 Jun 2020.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gscs20

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2020.1780235

Taylor & Francis
Taylor & Francis Group

Check for updates

A simple test for the difference of means in meta-analysis when study-specific variances are unreported

Patarawan Sangnawakij[a] and Dankmar Böhning [b]

[a]Department of Mathematics and Statistics, Thammasat University, Pathum Thani, Thailand; [b]Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

ABSTRACT

Standard meta-analysis requires the quantity of interest and its estimated variance to be reported for each study. Datasets that lack such variance information pose important challenges to meta-analytic inference. In a study with continuous outcomes, only sample means and sample sizes may be reported in the treatment arm. Classical meta-analytical technique is unable to apply statistical inference to such datasets. In this paper, we propose a statistical tool for testing equal means between two groups in meta-analysis when the variances of the constituent studies are unreported, using pivot inference based on the exact t-distribution and the generalized likelihood ratio. These are considered under a fixed-effect model. In simulations, the type I errors and power probabilities of the proposed tests are investigated as metrics of their performance. The t-test statistic provides type I errors very close to the nominal significance level in all cases and has large power. The generalized likelihood ratio test statistic performs well when the number of studies is moderate-to-large. The performance of our tests surpasses that of the conventional test, which is based on the normal distribution. The difference is especially pronounced when the number of studies is small. The distribution given by our tests is also shown to closely follow the theoretical distribution.

ARTICLE HISTORY
Received 11 November 2019
Accepted 5 June 2020

KEYWORDS
Continuous outcome; likelihood ratio; missing variance; simulation study; t-distribution

1. Background

Meta-analysis is an important statistical tool used to summarize the results from individual, independent studies on the same topic. It is applied in many fields including health science, medicine, psychology, and social science. In particular, meta-analysis of clinical trial results is considered the highest level of evidence and provides a readily accessible synthesis of the evidence on the effectiveness of a given treatment [1]. Both count and continuous outcomes are of interest in meta-analysis. For valid analysis, meta-analysis requires both the quantity of interest or effect size estimate in each study and its estimated standard error. In traditional meta-analysis, the overall estimate of effect size is then computed by the weighted average method [2] . More precisely, suppose that $\hat{\theta}_i$ denotes the effect measure

CONTACT Patarawan Sangnawakija patarawan.s@gmail.com Department of Mathematics and Statistics, Thammasat University, Pathum Thani, Thailand

© 2020 Informa UK Limited, trading as Taylor & Francis Group

Table 1. Meta-analytic data on the length of hospital stay (days) for thoracoscopic and open surgeries.

	Thoracoscopic surgery		Open surgery	
Author, year	n	$\bar{X}$	n	$\bar{X}$
Vu, 2008	12	2	24	5
Diamond, 2007	12	3.5	24	4
Kunisaki, 2014	49	3	13	3
Lau, 2013	39	6.95	28	11.96
Rahman, 2009	14	2.95	14	2.6
Cho, 2012	7	6.1	27	8.1
Tolg, 2005	5	6	4	12
Fascetti-Leon, 2013	26	5.3	28	9.6
Sundararajan, 2007	20	2	9	6
Laje, 2015	100	3	188	3.1
Kulaylat, 2015	112	3	146	4

Note that n is the sample size and $\bar{X}$ is the sample mean.

estimate of the true parameter θ_i from study i, for $i = 1, 2, \ldots, k$, where k is the number of studies. The overall effect size estimate is computed by

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}, \tag{1}$$

where w_i is the weight of the effect size of study i obtained from the inverse of variance of the effect estimate. The weight is computed by $w_i = 1/s_i^2$, depending on whether the fixed or random effects model is used, and s_i^2 denotes the estimated variance of study i.

Although many studies report the basic statistics of mean, range, standard deviation, and coefficient of variation, others omit one or more of these. Published trial results are sometimes statistically incomplete, as the standard errors of the means are missing [3,4]. The present work uses meta-analytic data on the management of antenatally diagnosed congenital lung malformations in young children through two alternative surgical approaches: thoracoscopy and open resection. The key question is whether thoracoscopy is as safe as the traditional method of open surgery. A systematic review from 2007 to 2015 reported that complications arose in 63/404 (16%) of thoracoscopic operations and in 87/483 (18%) of open surgical operations [5]. We emphasize that these studies are not clinical trials, but reports which were found in the literature. Other outcomes used to compare the performance of the two surgeries were length of operation (mins), number of patient days in hospital, number of chest tube days, and weight and age of the child. An example of the length of stay data is shown in Table 1. These data contain only sample means and sample sizes in the two treatment arms, and no sample variance is reported in the published evidence. Statistics such as coefficient of variation, standard error, or confidence interval, which could be used to compute the estimated study-specific variance, are also missing.

Traditional meta-analysis cannot satisfactorily derive the quantitative outcomes, as variability measures are unreported and overall effect estimates cannot be calculated. Under these circumstances, imputation methods have been introduced to estimate the variance. Examples include Philbrook et al., Idris and Robertson, and Chowdhry et al. [6–8]. However, imputation can be applied only in the case that variance is unreported in some of the individual studies. Sangnawakij et al. [9] introduced an approach to the estimation of

the parameter when no estimated variance is available. They proposed the estimators for the overall mean difference and the within-study variance using maximum likelihood estimation and studied the performance of the estimators by simulations. However, the study investigated only parameter estimation for the mean difference. Sidik and Jonkman [10] proposed a method for constructing confidence intervals for the population mean in meta-analysis, based on the normal and t distributions. Two statistics were derived under the assumption that the effect size estimate from study i, denoted as $\bar{X}_i$, had an $N(\mu, \sigma_i^2 + \tau^2)$ for $i = 1, 2, \ldots, k$, where μ is the true mean, σ_i^2 is the true variance of $\bar{X}_i$, and τ^2 is the heterogeneity variance between studies. For this, if $\tau^2 = 0$, homogeneity arises. The fixed effect model is then given by $\bar{X}_i = \mu + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_i^2)$ is the sampling error of $\bar{X}_i$. To estimate the parameter, σ_i^2 was replaced by the sample variance s_i^2, which was known from individual, independent studies. Then, pivotal quantities based on the normal and t distributions used to construct the $(1-\alpha)100\%$ confidence intervals for μ were derived, and given by

$$\frac{\hat{\mu} - \mu}{1/\sqrt{\sum_{i=1}^{n} 1/s_i^2}} \sim N(0, 1)$$

and

$$\frac{\sqrt{\sum_{i=1}^{n} 1/s_i^2}(\hat{\mu} - \mu)}{\sqrt{\left(\sum_{i=1}^{n} (\bar{X}_i - \hat{\mu})^2 / s_i^2\right)/(k-1)}} \sim t_{df=k-1},$$

where $\hat{\mu}$ is the weighted estimate of μ computed using the same method as shown in Equation (1) [10]. Hypothesis testing is an important aspect of statistical inference in many applications, and should therefore be incorporated in a comprehensive method. Park [11] proposed statistics for testing the equality of normal population means when the population size is large related to the sample sizes. However, those papers focused on meta-analysis with the common situation, where the sample variances were assumed to be known. No research has been done on the problem of testing the equality of means when the variances are unreported for all studies in meta-analysis. It is therefore addressed in the current study.

The main objective is to propose an approach to test the difference of means in meta-analysis without study-specific variance information. For clarity, we note that we do not propose to ignore study-specific variance information, but suggest the solutions to use the available information. In this paper, the test statistics are derived using pivot inference based on the t-distribution and the likelihood ratio. We investigate the performance of the test using type I error and power probability and compare the performance of our test with that of a test based on the normal distribution. We apply all methods to meta-analysis of a dataset on open and thoracoscopic operations.

2. Methods

We consider k independent studies in a meta-analysis in which only sample means $\bar{X}_i^T$ and $\bar{X}_i^C$, and sample sizes n_i^T and n_i^C, are available in the treatment (T) and comparison (C) arms, for $i = 1, 2, \ldots, k$. Suppose that the random variable $\bar{X}_i^T$ has $N(\mu^T, \sigma^2/n_i^T)$ and $\bar{X}_i^C$ has $N(\mu^C, \sigma^2/n_i^C)$, where $\bar{X}_i^T$ and $\bar{X}_i^C$ are independent. The study-specific mean difference,

which is also the effect size of interest for study i, is given by $D_i = \bar{X}_i^T - \bar{X}_i^C$. It follows that D_i has $N(\mu, \sigma^2 w_i)$, where $\mu = \mu^T - \mu^C$ is the true mean difference, which is also the parameter of interest, σ^2 is the within-study population variance (assumed to be equal across all studies), and $w_i = 1/n_i^T + 1/n_i^C$ is the constant. In this case, Sangnawakij et al. [9] proposed maximum likelihood (ML) estimators for μ and σ^2. These are given by

$$\hat{\mu}_{ML} = \frac{\sum_{i=1}^{k} D_i/w_i}{\sum_{i=1}^{k} 1/w_i}$$

and

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

The variances of the estimators are given as $Var(\hat{\mu}_{ML}) = \sigma^2/\sum_{i=1}^{k} 1/w_i$ and $Var(\hat{\sigma}^2_{ML}) = (2(k-1)\sigma^4)/k^2$. Note that $\hat{\mu}$ and $\hat{\sigma}^2$ are unbiased and consistent estimators for μ and σ^2, respectively.

Next, the test statistics for the mean difference are constructed. The hypotheses are

$$H_0 : \mu = \mu_0 \quad \text{and} \quad H_1 : \mu \neq \mu_0,$$

where μ_0 is a specified value of the mean difference. A basic approach to deriving the test statistic uses the pivotal quantity, which has a normal distribution. This is given by

$$Z = \frac{\hat{\mu}_{ML} - E(\hat{\mu}_{ML})}{\sqrt{\widehat{Var}(\hat{\mu}_{ML})}} = \frac{\hat{\mu}_{ML} - \mu}{\sqrt{\hat{\sigma}^2_{ML}/\sum_{i=1}^{k} 1/w_i}}. \tag{2}$$

Under H_0, the test statistic Z is an approximate standard normal distribution. In practice, the null hypothesis will be rejected if the observed value of $|Z|$ is greater than $Z_{\alpha/2}$, or the $(\alpha/2)100$th percentile of the standard normal distribution. However, several methods can be used to derive the statistic for test of the mean difference. This study applies the t-test and likelihood ratio test, which we introduce in the following two subsections.

2.1. The proposed t-test

Equation (2) is based on the central limit theorem. This is known to be appropriate when the meta-analysis comprises a large number of studies. In practice, however, the number of studies may be small. In this case, an approach based on a normal approximation may be unreasonable, because the data do not have a normal distribution. To identify an appropriate method for such cases, we investigate inference based on the t-distribution. For meta-analysis of k independent studies, we obtain the following function of the mean difference:

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

$$= \sum_{i=1}^{k} \frac{[(D_i - \hat{\mu}_{ML}) + (\hat{\mu}_{ML} - \mu)]^2}{\sigma^2 w_i}$$

$$= \sum_{i=1}^{k} \left(\frac{D_i - \hat{\mu}_{ML}}{\sigma \sqrt{w_i}} \right)^2 + \frac{(\hat{\mu}_{ML} - \mu)^2}{\sigma^2 / \sum_{i=1}^{k} 1/w_i}. \tag{3}$$

Since D_i has $N(\mu, \sigma^2 w_i)$, the term on the left of Equation (3) has a chi-square distribution with k degrees of freedom. Furthermore, since $\hat{\mu}_{ML}$ has $N(\mu, \sigma^2 / \sum_{i=1}^{k} 1/w_i)$ and $V = \hat{\mu}_{ML} - \mu / \sigma / \sqrt{\sum_{i=1}^{k} 1/w_i}$ has $N(0, 1)$, the second term on the right-hand side has a chi-square distribution with one degree of freedom. Therefore, $U = \sum_{i=1}^{k} ((D_i - \hat{\mu}_{ML}) / \sigma \sqrt{w_i})^2$ is a chi-square distribution with $k-1$ degrees of freedom.

Pivot inference based on the t-distribution is next considered. This is obtained from two variables with normal and chi-square distributions. As we have shown that V and U are random variables whose distributions do not depend on the parameter μ, the pivotal quantity based on the t-distribution can be constructed. It is given by

$$T_{pr} = \frac{V}{\sqrt{U/(k-1)}} = \frac{\sqrt{\sum_{i=1}^{k} 1/w_i}(\hat{\mu}_{ML} - \mu)}{\sqrt{k\hat{\sigma}^2_{ML}/(k-1)}}. \tag{4}$$

Under $H_0 : \mu = \mu_0$, T_{pr} follows a t-distribution with $k-1$ degrees of freedom. The decision rule states that H_0 will be rejected if the observed value of $|T_{pr}|$ is larger than $t_{\alpha/2, df=k-1}$, where $t_{\alpha/2, df=k-1}$ denotes the $(\alpha/2)100$th percentile of the t-distribution with $k-1$ degrees of freedom. It is important to note that T_{pr} differs from the statistic based on t-distribution presented by Sidik and Jonkman [10], which addressed meta-analyses in which the variance estimates are reported.

2.2. The proposed likelihood ratio test

The likelihood ratio (LR) is a statistical method that uses the ratio of maximized likelihoods under the parameter space and null hypothesis. In our study, it is used to construct a test of the mean difference. The procedure is as follows.

Let D_i for $i = 1, 2, \ldots, k$ be a random sample of size k from a population with distribution $N(\mu, \sigma^2 w_i)$. Again, D_i denotes the mean difference of the study i. The probability density function of D_i is then given by

$$f(d_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 w_i}} \exp\left(-\frac{(d_i - \mu)^2}{2\sigma^2 w_i} \right)$$

with the likelihood function $L(\mu, \sigma^2; d_i) = \prod_{i=1}^{k} f(d_i; \mu, \sigma^2)$. We wish to test $H_0 : \theta$ in ω against $H_1 : \theta$ in $\Omega - \omega$, where Ω is the parameter space and ω is the set of unknown parameter values under H_0. Here, we let $\theta = (\mu, \sigma^2)$, so that

$$\Omega = \{\theta; -\infty < \mu < \infty, \sigma^2 > 0\} \quad \text{and} \quad \omega = \{\theta; \mu = \mu_0, \sigma^2 > 0\}.$$

The likelihood function of μ and σ^2 under ω is then given by

$$L(\mu, \sigma^2; d_i) = \frac{1}{(2\pi\sigma^2)^{k/2} \prod_{i=1}^{k} w_i^{1/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \frac{(d_i - \mu_0)^2}{w_i} \right). \tag{5}$$

This yields the ML estimator $\hat{\sigma}_0^2 = 1/k \sum_{i=1}^{k} \frac{(d_i-\mu_0)^2}{w_i}$ for σ^2. The likelihood function under Ω is given as

$$L(\mu, \sigma^2; d_i) = \frac{1}{(2\pi\sigma^2)^{k/2} \prod_{i=1}^{k} w_i^{1/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \frac{(d_i - \mu)^2}{w_i}\right). \tag{6}$$

The ML estimators for μ and σ^2 derived from this function are denoted $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$, respectively. The likelihoods from Equations (5) and (6) are maximized by taking the ML estimators for obtaining the second proposed statistic.

In general, the ratio of the likelihoods, which yields the LR test for $H_0 : \mu = \mu_0$, is given as

$$\Lambda = \frac{\sup_{\theta\in\omega} L(\theta; d_i)}{\sup_{\theta\in\Omega} L(\theta; d_i)}.$$

The test procedure rejects the null hypothesis if the value of Λ is small, or $\Lambda < c$. Usually, a constant c is chosen to specify a value for the type I error associated with the distribution of the test. However, if the ratio is in complex form, identifying the distribution is challenging. The LR statistic can be simplified by applying asymptotic approximation, given by $-2\log\Lambda$. This is the generalized likelihood ratio test statistic. It is straightforward to verify that $-2\log\Lambda$ converges to a chi-square distribution with one degree of freedom. Two likelihood functions derived above yield the generalized LR test:

$$\begin{aligned} -2\log\Lambda &= -2\log\left(\frac{\sum_{i=1}^{k}(d_i - \hat{\mu}_{ML})^2/w_i}{\sum_{i=1}^{k}(d_i - \mu_0)^2/w_i}\right)^{k/2} \\ &= -2\log\left(\frac{\hat{\sigma}^2_{ML}}{\hat{\sigma}^2_0}\right)^{k/2}. \end{aligned} \tag{7}$$

We therefore reject $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu \neq \mu_0$ if $LR_{pr} = -2\log\Lambda$ is larger than $\chi^2_{\alpha,df=1}$, where $\chi^2_{\alpha,df=1}$ denotes the $(\alpha)100$th percentile of the chi-square distribution with one degree of freedom.

3. Simulation study

The performance of the proposed tests using T_{pr} and LR_{pr} was evaluated by type I error and power probability in simulations run under R (https://www.r-project.org/). The Z test statistic shown in Equation (2) provided the benchmark. The two-sided test for $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$, and the one-sided test for $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ were considered. The meta-analysis was assumed to have one true mean difference across all studies (fixed-effect meta-analysis).

3.1. Simulation settings

The mean difference D_i was generated from $N(\mu, \sigma^2 w_i)$. The values of the true mean difference (μ) were set at 0, 2, and 5 and those of the within-study variance (σ^2) at 2, 4, and 9, and the weight w_i was sampled from the uniform distribution on (0.02,0.20). The numbers

Table 2. Average values of the test statistics, type I error, and power rates for the two-sided tests of the mean difference when $\alpha = 0.01$.

			Test statistic			Type I error			Power test ($a = 3$)			Power test ($a = 6$)		
k	μ	σ^2	Z Test	T_{pr}	LR_{pr}	Z Test	T_{pr}	LR_{pr}	Z Test	T_{pr}	LR_{pr}	Z Test	T_{pr}	LR_{pr}
5	0	2	−0.0078	−0.0070	1.3803	0.0778	0.0085	0.0252	0.9668	0.8079	0.0020	0.9960	0.9743	0.0272
		4	0.0037	0.0033	1.4035	0.0839	0.0116	0.0312	0.9630	0.7955	0.0018	0.9969	0.9740	0.0309
		9	0.0125	0.0112	1.3917	0.0800	0.0108	0.0303	0.9679	0.8098	0.0020	0.9970	0.9741	0.0241
	2	2	0.0156	0.0139	1.4222	0.0842	0.0115	0.0314	0.9630	0.7949	0.0023	0.9969	0.9745	0.0265
		4	−0.0212	−0.0189	1.4060	0.0836	0.0102	0.0301	0.9644	0.7995	0.0028	0.9969	0.9714	0.0261
		9	−0.0022	−0.0020	1.3904	0.0826	0.0105	0.0290	0.9667	0.8050	0.0016	0.9970	0.9763	0.0288
	5	2	−0.0155	−0.0139	1.4029	0.0835	0.0097	0.0300	0.9663	0.7949	0.0028	0.9966	0.9743	0.0241
		4	0.0013	0.0012	1.3836	0.0770	0.0093	0.0274	0.9625	0.7981	0.0021	0.9966	0.9755	0.0277
		9	−0.0059	−0.0053	1.4335	0.0887	0.0104	0.0295	0.9658	0.7991	0.0022	0.9971	0.9766	0.0257
10	0	2	-0.0081	−0.0077	1.1511	0.0355	0.0096	0.0172	0.9998	0.9982	0.2278	1	1	0.8686
		4	−0.0028	−0.0026	1.1838	0.0375	0.0085	0.0160	0.9991	0.9974	0.2417	1	1	0.8713
		9	−0.0052	−0.0049	1.1672	0.0353	0.0087	0.0158	0.9991	0.9974	0.2270	1	1	0.8669
	2	2	−0.0003	−0.0002	1.1696	0.0376	0.0109	0.0185	0.9994	0.9978	0.2278	1	1	0.8643
		4	0.0052	0.0050	1.1751	0.0360	0.0096	0.0169	0.9994	0.9974	0.2286	1	1	0.8672
		9	0.0158	0.0150	1.1569	0.0355	0.0099	0.0163	0.9995	0.9972	0.2319	1	1	0.8774
	5	2	−0.0114	−0.0108	1.1856	0.0383	0.0086	0.0156	0.9993	0.9984	0.2319	1	1	0.8715
		4	0.0144	0.0137	1.1710	0.0364	0.0107	0.0177	0.9996	0.9979	0.2331	1	1	0.8710
		9	0.0181	0.0172	1.1836	0.0379	0.0099	0.0187	0.9992	0.9979	0.2232	1	1	0.8690
30	0	2	−0.0013	−0.0013	1.0457	0.0174	0.0098	0.0114	1	1	1	1	1	1
		4	0.0049	0.0049	1.0243	0.0161	0.0093	0.0111	1	1	1	1	1	1
		9	0.0245	0.0241	1.0483	0.0181	0.0113	0.0135	1	1	1	1	1	1
	2	2	0.0153	0.0150	1.0529	0.0172	0.0113	0.0134	1	1	1	1	1	1
		4	−0.0011	−0.0011	1.0392	0.0151	0.0088	0.0109	1	1	1	1	1	1
		9	0.0068	0.0067	1.0275	0.0145	0.0079	0.0101	1	1	1	1	1	1
	5	2	0.0064	0.0063	1.0704	0.0161	0.0096	0.0118	1	1	1	1	1	1
		4	−0.0008	−0.0007	1.0694	0.0172	0.0100	0.0118	1	1	1	1	1	1
		9	−0.0079	−0.0078	1.0326	0.0159	0.0088	0.0111	1	1	1	1	1	1
50	0	2	−0.0282	−0.0279	1.0296	0.0139	0.0100	0.0111	1	1	1	1	1	1
		4	0.0046	0.0045	1.0158	0.0137	0.0090	0.0107	1	1	1	1	1	1
		9	0.0034	0.0033	1.0282	0.0145	0.0105	0.0118	1	1	1	1	1	1
	2	2	0.0095	0.0094	1.0191	0.0141	0.0105	0.0118	1	1	1	1	1	1
		4	0.01700	0.0169	1.0266	0.0135	0.0100	0.0108	1	1	1	1	1	1
		9	−0.0029	−0.0029	1.0367	0.0144	0.0103	0.0116	1	1	1	1	1	1
	5	2	−0.0012	−0.0012	1.0193	0.0137	0.0108	0.0112	1	1	1	1	1	1
		4	0	0	1.0172	0.0120	0.0082	0.0095	1	1	1	1	1	1
		9	−0.0177	−0.0175	1.0278	0.0148	0.0090	0.0104	1	1	1	1	1	1

of studies (k) were set at 5, 10, 30, and 50, reflecting a range from small trials to large. Significance levels (α) of 0.01 and 0.05 were used. Each simulation was run $B = 10{,}000$ times. The average type I error for the test was computed by

$$\hat{\alpha} = \frac{\text{number of reject } H_0 | H_0 \text{ is true}}{B}.$$

To estimate the power of the test, we generated D_i from $N(\mu + aS, \sigma^2 w_i)$ with constants (a) of 3 and 6, reflecting the small-to-large deviation of the mean. Here, S is the standard deviation of D_i sampled from $N(\mu, \sigma^2 w_i)$. The mean of power of the test was estimated by

$$\widehat{1-\beta} = \frac{\text{number of reject } H_0 | H_1 \text{ is true}}{B}.$$

A preferred test that would have a type I error rate close to the nominal significance level and a large power. The main simulation results were as follows.

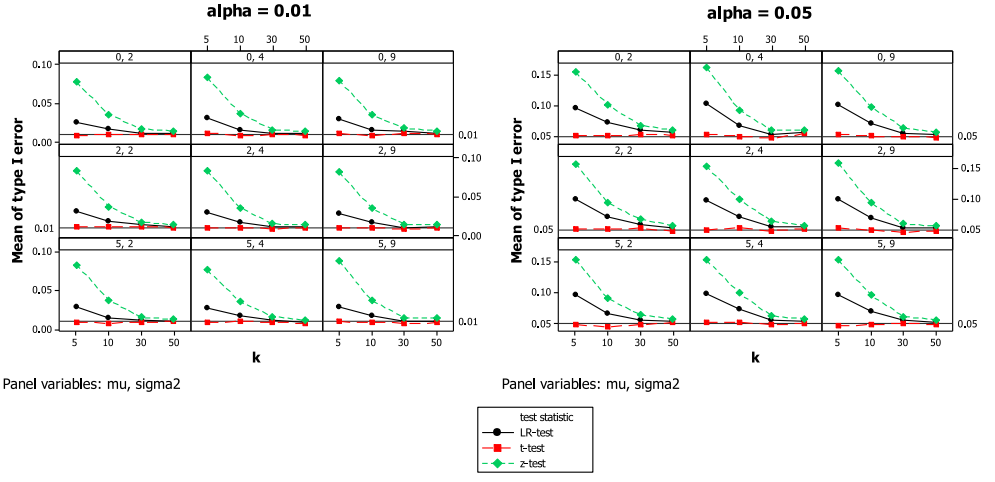

Figure 1. Type I error for the two-sided tests of the mean difference.

3.2. Simulation results

Table 2 compares the proposed and conventional Z statistics for a two-sided test with $\alpha = 0.01$. Type I errors in the three tests decreased to the significant level as k increased. However, they did not depend on μ or σ^2. The type I error of Z test was very much greater than 0.01 when $k \leq 10$ or the number of studies was small. The type I error of T_{pr} was very close to the nominal significance level of 0.01 in all cases in the study. The LR_{pr} also satisfied the type I error criterion, especially when $k \geq 10$. Ranking by closeness of the type I error to the nominal significance level gave T_{pr}, LR_{pr}, and Z. The table also shows the power of the two-sided test. At $\alpha = 0.01$, Z and T_{pr} had high probabilities. LR_{pr} was satisfactory when $k \geq 30$ for small deviations of the mean ($a = 3$), and when $k \geq 10$ for large deviations ($a = 6$). The results of type I error and power of the test for $\alpha = 0.01$ and $\alpha = 0.05$ are shown graphically in Figures 1 and 2. As $\alpha = 0.05$, type I errors of T_{pr} and LR_{pr} were also closer to the nominal significant level than that of Z test. From 2, all tests had powers greater than 0.9.

We investigated the performance of the right-tailed test of the mean difference at $\alpha = 0.01$. Tables 3 shows the results. In this case, only the Z and T_{pr} tests were conducted. The type I errors decreased in both tests as k increased, but did not depend on μ or σ^2. This matched the results from the two-tailed test. As shown in Figures 3 and 4, both tests had high power probabilities, but in T_{pr} the type I errors were closer to 0.01 or 0.05.

4. Application to real data

We applied the three statistics for test of equal means, using data from a meta-analysis in which the study-specific variance was unreported. As noted in the introduction, the studies compared the outcomes from open surgery and thoracoscopy when treating asymptomatic congenital lung malformation. The variables used in this empirical analysis were length of patient stay in hospital (days) included 11 studies, and number of drains left in the chest (days) included 9 studies. In both arms, only the sample mean and sample size were

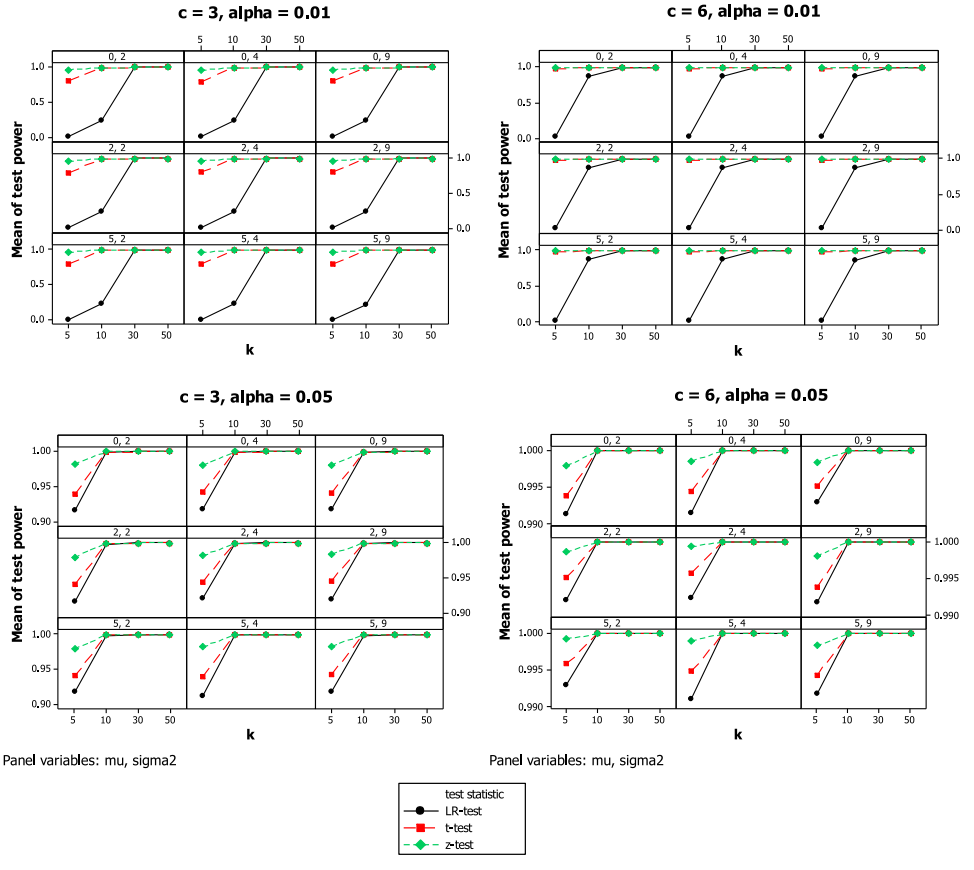

Figure 2. Power of the two-sided tests of the mean difference.

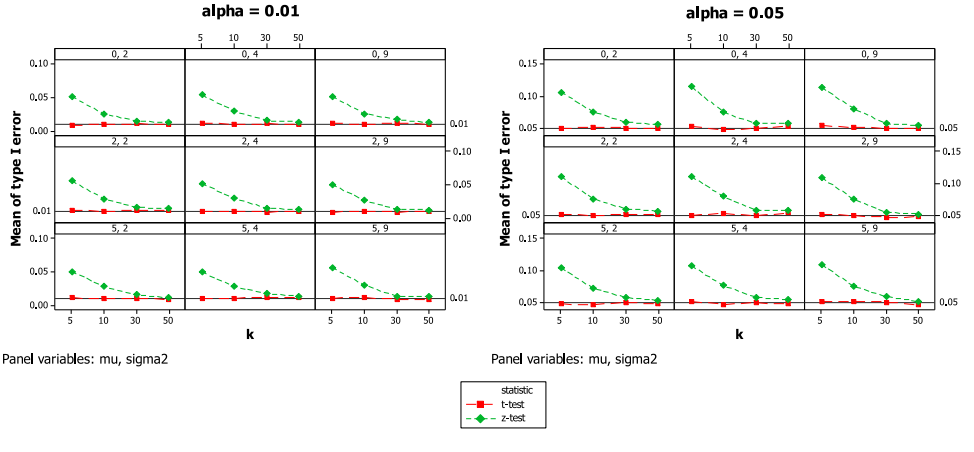

Figure 3. Type I error for the one-sided tests of the mean difference.

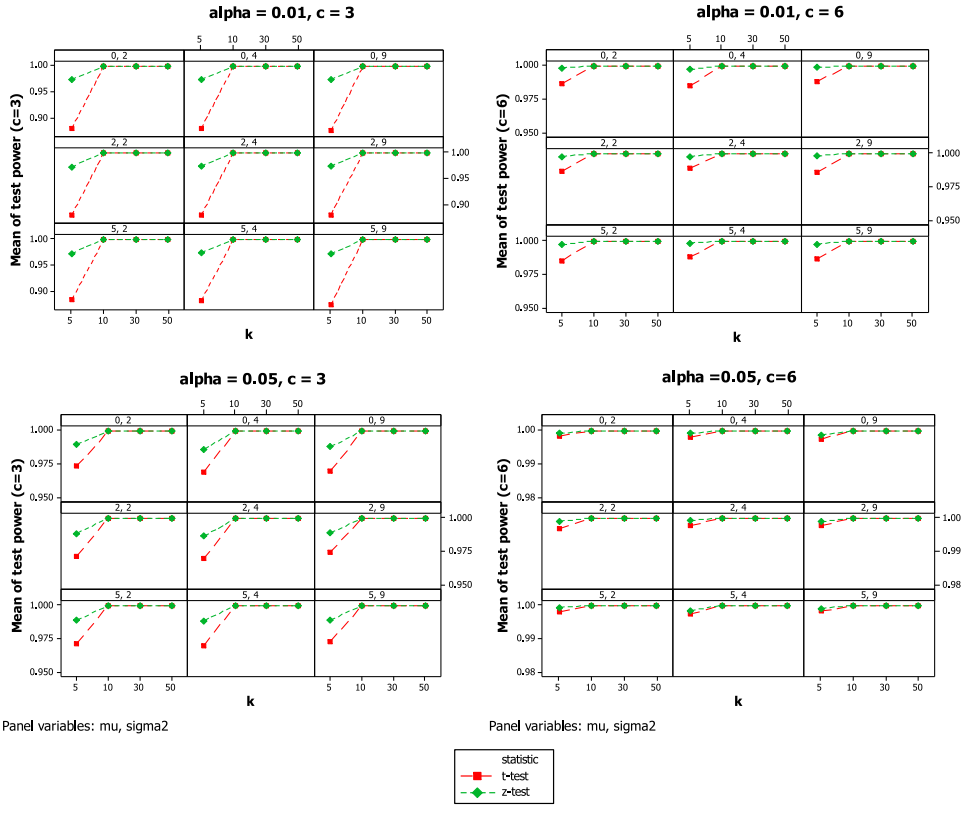

Figure 4. Power of the one-sided tests of the mean difference.

reported for each variable and no sample variance was available. Note that these studies are not clinical trials, but reports which were found in the literature. The analysis used studies which had complete information in both arms.

We computed the overall mean difference and the within-variance estimates. The point estimators for $\mu = \mu^T - \mu^C$ and σ^2 were estimated using ML estimation: $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$. The mean differences for length of stay and number of chest tube days between thoracoscopic (T) and open (C) surgeries were -1.4 and -0.9, with standard deviations 7.420 and 4.793, respectively. *The outcomes from thoracoscopy had greater means, but it was not clear that the difference was significant.* Therefore, this was investigated using hypothesis testing. Firstly, the distribution of the data was analysed using the Anderson-Darling test. It was found that the datasets followed normal distributions with p-values of 0.288 for the mean difference in length of hospital stay and 0.103 for the mean difference in number of chest tube days. The meta-analytic data were there suitable for our approach.

A test of equal means was then conducted using the methods introduced in the previous section. The observed values for testing the hypotheses $H_0 : \mu^T = \mu^C$ and $H_1 : \mu^T \neq \mu^C$ are given in Table 4. For length of stay, the Z test identified a significant difference between the means of the two surgeries at $\alpha = 0.05$. Also, the proposed t-test and LR test showed a significant difference between the means of two surgeries, although their p-values were

Table 3. Average values of the test statistics, type I error, and power rates for the right-tailed tests of the mean difference when $\alpha = 0.01$.

			Test statistic		Type I error		Power test ($a = 3$)		Power test ($a = 6$)	
k	μ	σ^2	Z	T_{pr}	Z	T_{pr}	Z	T_{pr}	Z	T_{pr}
5	0	2	−0.0078	−0.007	0.0512	0.0078	0.9740	0.8784	0.9985	0.9869
		4	0.0037	0.0033	0.0538	0.0111	0.9737	0.8789	0.9971	0.9850
		9	0.0125	0.0112	0.0517	0.0113	0.9738	0.8742	0.9986	0.9884
	2	2	0.0156	0.0139	0.0556	0.0108	0.9722	0.8782	0.9975	0.9867
		4	−0.0212	−0.0189	0.0518	0.0099	0.9735	0.8787	0.9977	0.9888
		9	−0.0022	−0.0020	0.0501	0.0087	0.9731	0.8795	0.9981	0.9855
	5	2	−0.0155	−0.0139	0.0504	0.0105	0.9725	0.8820	0.9971	0.9853
		4	0.0013	0.0012	0.0492	0.0100	0.9745	0.8800	0.9983	0.9878
		9	−0.0059	−0.0053	0.0561	0.0102	0.9708	0.8728	0.9975	0.9862
10	0	2	−0.0081	−0.0077	0.0256	0.0098	0.9992	0.9984	1	1
		4	−0.0028	−0.0026	0.0293	0.0098	0.9996	0.9987	1	1
		9	−0.0052	−0.0049	0.0255	0.0092	0.9995	0.9983	1	1
	2	2	−0.0003	−0.0002	0.0283	0.0099	0.9997	0.9988	1	1
		4	0.0052	0.0050	0.0289	0.0095	0.9998	0.9985	1	1
		9	0.0158	0.0150	0.0263	0.0093	0.9999	0.9986	1	1
	5	2	−0.0114	−0.0108	0.0275	0.0091	0.9998	0.9988	1	1
		4	0.0144	0.0137	0.0275	0.0101	0.9995	0.9986	1	1
		9	0.0181	0.0172	0.0295	0.0106	0.9995	0.9986	1	1
30	0	2	−0.0013	−0.0013	0.0142	0.0103	1	1	1	1
		4	0.0049	0.0049	0.0152	0.0103	1	1	1	1
		9	0.0245	0.0241	0.0165	0.0113	1	1	1	1
	2	2	0.0153	0.0150	0.0153	0.0109	1	1	1	1
		4	−0.0011	−0.0011	0.0147	0.0086	1	1	1	1
		9	0.0068	0.0067	0.0124	0.0087	1	1	1	1
	5	2	0.0064	0.0063	0.0150	0.0097	1	1	1	1
		4	−0.0008	−0.0007	0.0167	0.0114	1	1	1	1
		9	−0.0079	−0.0078	0.0128	0.0085	1	1	1	1
50	0	2	−0.0282	-0.0279	0.0121	0.0095	1	1	1	1
		4	0.0046	0.0045	0.0129	0.0097	1	1	1	1
		9	0.0034	0.0033	0.0119	0.0099	1	1	1	1
	2	2	0.0095	0.0094	0.0139	0.0115	1	1	1	1
		4	0.0170	0.0169	0.0119	0.0094	1	1	1	1
		9	−0.0029	−0.0029	0.0118	0.0092	1	1	1	1
	5	2	−0.0012	−0.0012	0.0106	0.0083	1	1	1	1
		4	0	0	0.0127	0.0103	1	1	1	1
		9	−0.0177	−0.0175	0.0119	0.0086	1	1	1	1

Table 4. The results for test of the mean difference using meta-analytic data on key-hole and open surgeries ($\alpha = 0.05$).

Variable (k)	Method	Observed statistic	p-value
Length of stay (11)	Z test	−2.672	0.008
	T_{pr}	−2.548	0.029
	LR_{pr}	5.502	0.019
	$\hat{\mu}_{ML}$	−1.382	
	$\hat{\sigma}^2_{ML}$	55.063	
Number of chest tube days (9)	Z test	−1.917	0.055
	T_{pr}	−1.808	0.108
	LR_{pr}	3.082	0.079
	$\hat{\mu}_{ML}$	−0.788	
	$\hat{\sigma}^2_{ML}$	22.971	

higher. This was concluded using the forest plot for length of stay in hospital presented in Adams et al. [5], where the confidence limits of the overall difference in means did not covered zero. On the second variable, we found no significant difference in chest tube days between the two groups.

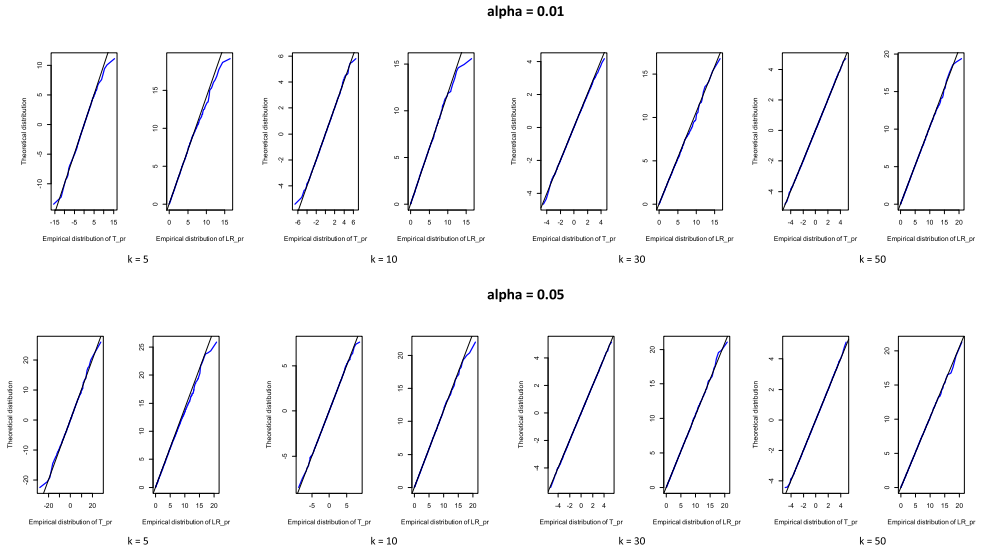

Figure 5. QQ plots for the test statistics when $\mu = 0, \sigma^2 = 2, k = 5, 10, 30$, and 50.

5. Conclusions

One of the main objectives of meta-analysis is to determine the parametric effect size by hypothesis testing. When the effect size estimates and standard errors of the individual studies are available, classical meta-analysis can be used to construct the statistic for hypothesis testing. However, when the variance or standard error of the effect size is unreported, the conventional approach is inappropriate in a study with continuous outcomes, as the weight of the effect size cannot be computed.

We introduce two test statistics to use the available information for test of the difference in means in meta-analysis without study-specific variance information. This is because only the quantity of interest and sample size of the study are reported in the published evidence. Moreover, no other information on uncertainty quantification is available from which the estimated variance can be obtained. In this paper, the test statistics were derived using pivot inference based on the exact t-distribution and generalized LR, assuming a fixed-effect meta-analysis. The performances were compared in terms of type I error and power of the test, using as baseline the Z test statistic based on the normal distribution. This was done through simulation in many situations. The proposed tests were shown to have a good performance. In particular, the t-test provided type I errors very close to the nominal significance level and high power probability. On both criteria, it outperformed the Z test in all cases. The LR test demonstrated high power when the sample sizes were greater than 10, with type I error very close to the nominal significance level.

Finally, we investigated whether the actual distribution given by our tests followed the theoretical distribution. QQ plots were used to explore the behaviour of the tests. The results are shown in Figure 5, which plots the empirical cumulative distribution function (CDF) of the test against the theoretical CDF. The empirical CDFs of the t-test followed the theoretical distribution (t-distribution) in all cases. The empirical CDF of the generalized LR test tracked the theoretical distribution (chi-square distribution) when $k > 10$. This confirms that the exact t-distribution may be used. The t-test is therefore recommended

for test of the mean difference in meta-analyses that do not report the variance of the constituent studies. This method works well for meta-analyses with small numbers of studies. As noted in Seide et al. [12], meta-analyses of small studies are common in practice. A review of the Cochrane Library presented that half of the meta-analyses reported in the Cochrane Library are conducted with two or three studies. Our proposed t-test is therefore importance. For larger studies, the proposed LR test offers as an alternative approach.

Acknowledgments

This study was supported by Thammasat University Research Fund, Contract No. TUGR 2/25/2562. Patarawan Sangnawakij is deeply grateful for receiving this funding. All authors also thank the reviewers for suggestions which will lead to considerable improvements in the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Dankmar Böhning http://orcid.org/0000-0003-0638-7106

References

[1] Garrattini S, Jacobsen JC, Wetterslev J, et al. Evidence-based clinical practice: overview of threats to the validity of evidence and how to minimise them. Eur J Intern Med. 2016;32:13–21.
[2] Borenstein M, Hedges LV, Higgins JPT, et al. Introduction to meta-analysis. Chichester: John Wiley & Sons; 2009.
[3] Ma J, Liu W, Hunter A, et al. Performing meta-analysis with incomplete statistical information in clinical trials. BMC Med Res Methodol. 2008;8 (56):1–11.
[4] Weir CJ, Butcher I, Assi V, et al. Dealing with missing standard deviation and mean values in meta-analysis of continuous outcomes: a systematic review. BMC Med Res Methodol. 2018;18(25):1–14.
[5] Adams S, Jobson M, Sangnawakij P, et al. Does thoracoscopy have advantages over open surgery for asymptomatic congenital lung malformations? an analysis of 1626 resections. J Pediatr Surg. 2017;52(2):247–251.
[6] Philbrook HT, Barrowman N, Garga AX, et al. 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. J Clin Epidemiol. 2007;60(3):228–240.
[7] Idris NRN, Robertson C, The effects of imputing the missing standard deviations on the standard error of meta analysis estimates. Commu Stat - Simul C. 2009;38(3):513–526.
[8] Chowdhry AK, Dworkinb RH, McDermott MP. Meta-analysis with missing study-level sample variance data. Stat Med. 2016;35(17):3021–3022.
[9] Sangnawakij P, Böhning D, Adams S, et al. Statistical methodology for estimating the mean difference in a meta-analysis without study-specific variance information. Stat Med. 2017;36(9):1395–1413.
[10] Sidik K, Jonkman JN, A simple confidence interval for meta-analysis. Stat Med. 2002;21(16): 3153–3159.
[11] Park J, Park D, Testing the equality of a large number of normal population means. Comput Stat Data Anal. 2012;12:1131–1149.
[12] Seide S, Röver C, Friede T. Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. BMC Med Res Methodol. 2019;19(16):1–14.