STATISTICS | RESEARCH ARTICLE

Testing homogeneity of effect sizes in pooling 2x2 contingency tables from multiple studies: a comparison of methods

Osama Almalik¹* and Edwin R. van den Heuvel¹

Abstract: Meta-analysis is a statistical methodology that combines the outcomes of several independent studies. Either a fixed-effects or a random-effects model is used to combine the results of the independent studies. The choice depends on whether the outcomes can be considered homogeneous or not. Several methods have been developed to test the homogeneity of effect sizes. Although many of them have been compared already, they have not been studied in more realistic practical settings where individuals all have their own risk of an event and a complete overview is lacking.

In this article, we investigate the performance of 10 statistical methods to test homogeneity of a binary exposure on a binary clinical outcome, using an extensive simulation study and a real life meta-analysis. We evaluated the Type I error and the statistical power. The fixed-effects regression model for treatment and study interaction was found to be a slightly liberal test, while the $Q$-statistic and the Bliss statistic can be considered conservative tests. The random-effects regression model, the Peto statistic and the $I^2$ perform rather poorly compared to the other methods. All chi-square tests that are based on the calculation of a specific odds ratio and the fixed-effects regression model for treatment and study interaction were found to be slightly liberal tests.

ABOUT THE AUTHOR

Edwin van den Heuvel is a full-time professor at the Department of Mathematics and Computer Sciences in the Eindhoven University of Technology and he is the head of the Statistics group within the Stochastics section. The Statistics group develops and compares data-analytical methods for analyzing and sampling complex structured correlated datasets. It includes parameter estimation, model fitting, latent variable models, mixed models, missing data, statistical process control, survival and reliability theory, time series analysis, and statistical learning methods. One of the central themes is the analysis of high-dimensional temporal datasets and other large datasets. The group actively explores new research lines in data science. This current research article is part of a PhD project of Osama Almalik (the principal author) on meta-analysis for non-standard or complex situations under supervision of van den Heuvel. It is considered beneficial to medical researchers, as well as statisticians, when carrying out meta-analysis for combining effect sizes from cohort studies and clinical trials.

PUBLIC INTEREST STATEMENT

Testing homogeneity of effect sizes (treatment effects, risk factor associations) is an essential part of meta-analysis since it helps us to interpret the consistency of these effects across groups of participants. The testing can be done using various statistical tools, which can result in different conclusions. For this reason we thought it was important to study the performance of the existing statistical tests for homogeneity. Specific comparisons have been done earlier but we also tried to explore the performance of statistical tests that have never been applied to this specific problem, and to adjust existing statistical tests. We studied the performance of 10 different methods to test the homogeneity of effect sizes using a case study and an extensive simulation study. Based on our study, we recommend one specific statistical test for use for meta-analysts when testing the homogeneity since it outperformed the others in most of the settings we studied.
effects logistic regression analysis perform best. Among these chi-square tests, we recommend the Breslow–Day test, but only for convenience purposes, since it is available in most statistical software packages and the other tests are not.

Subjects: Science; Mathematics & Statistics; Statistics & Probability; Statistics; Statistics & Computing; Statistical Theory & Methods;

Keywords: Bliss statistic; Breslow-Day statistic; Cochran’s Q-statistic; generalized linear mixed model; logistic regression; multi-center trials; Peto statistic; Woolf statistic

1. Introduction
Pooling (analysis) results (like odds ratios) from multiple studies is common practice in medical, social and biological sciences. It is the main goal in an aggregate data meta-analysis or in an experimental study where some kind of stratification is applied (like multi-center randomized controlled trials). Researchers would typically be interested mainly in the overall or pooled estimate of a binary exposure or treatment effect, regardless of study differences, since it may better represent an estimate for the full target population than the individual studies. For a binary outcome, this means pooling $2 \times 2$ contingency tables from several studies. A traditional approach for pooling $2 \times 2$ contingency tables is the Cochran–Mantel–Haenszel statistic (Breslow, 1994; Breslow & Day, 1980; Mantel & Haenszel, 1959). However, this approach requires that the odds ratios across studies are homogeneous. This emphasizes the need for a test on homogeneity of the odds ratios or more generally the homogeneity of treatment effects on the binary clinical outcome. This topic is not new and many different approaches have been developed.

A classic test for homogeneity of effect sizes is the $Q$-statistic (Cochran, 1954). It creates a weighted effect size using the squared inverse within-study standard error and evaluates if the variation around this pooled effect is larger than that can be expected from the within-study standard errors. It was initially developed for pooling potency estimates from bioassay experiments (Mzolo, Hendriks, & van den Heuvel, 2013). A similar approach used in this area is the method of Bliss (1952), which uses the $Q$-statistic but in a different way. Related measures that are used in systematic reviews are the $H^2$, $I^2$ and the $R^2$ statistics (Higgins & Thompson, 2002). The $I^2$ statistic is also fully determined by the $Q$-statistic and it was initially developed to quantify the amount of heterogeneity instead of testing for homogeneity. The $H^2$ statistic is just an alternative measure for $I^2$, but the $R^2$ statistic describes the inflation from the random effects to the fixed effects, and it is viewed more as a measure than as a test statistic. However, the three measures of heterogeneity can be used as test statistics using appropriate confidence intervals.

Alternatively, fixed- and random-effects logistic regression models can be applied to test homogeneity. This is performed by including an interaction term between study and treatment, and then applying either a Wald test, a score test, or the Likelihood ratio test for its significance (Agresti, 2002).

A method that studies the heterogeneity of the odds ratio directly is the Breslow-Day test adjusted by Tarone (1985). It is some form of Pearson’s chi-square statistic. Other types of chi-square test statistics, which have been developed for specific situations, are Peto statistic (Yusuf, Peto, Lewis, Collins, & Sleight, 1986), Liang and Self statistics (Liang & Self, 1989) and Zelen statistic (Zelen, 1971). Yusuf et al. (1986) originally presented the Peto method for pooling odds ratios, and accompanied it with a homogeneity test (Yusuf et al., 1986). Liang and Self (1989) focused on developing test statistics in case data are sparse, i.e. some cells in the study-specific contingency tables contain only a few participants. Zelen developed the Zelen statistic to test the homogeneity of the relative risks (Zelen, 1971).

Several of these statistics have been discussed, evaluated and compared on their performance of detecting heterogeneity but, as Sutton and Higgins (2008) point out, there is no overview of the
results (Sutton & Higgins, 2008). The $Q$-statistic has been shown to have low statistical power when the number of studies is small (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003; Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006) and to be conservative for a large number of studies (Higgins & Thompson, 2002). Moreover, the application of the $Q$-statistic to test the homogeneity of treatment effects has come under criticism recently (Hoaglin, 2016). Hoaglin (2016) points out that Cochran did not use the $Q$-statistic itself as a test statistic and that the $Q$-statistic only approaches a chi-square distribution under the null hypothesis asymptotically when within-study sample size increases (since the individual terms will become approximately normally distributed). In addition, the assumptions of normality of effect sizes and that the within-study variances are known do not apply to many meta-analyses. This is especially the case when the effect sizes are risk differences, the logarithm of the risk differences or the logarithm of the odds ratios. Hoaglin’s concerns also apply directly to the widely used $I^2$ as well (Hoaglin, 2016). Jackson (2006) derived a formula to calculate the power of the homogeneity test when using the $Q$-statistic. The author only considered the case when the number of studies is odd and the studies have the same within-variance, and concluded that the homogeneity test based on the $Q$-statistic indeed had low power. Kulinskaya, Dollinger and Bjørkestøl (2011) improved the performance of the $Q$-statistic using two approximations—a Gamma distribution based on expanded first two moments of the $Q$-statistic and a chi-square distribution with fractional degrees of freedom matching the first moment—and recommended using the latter approximation (Kulinskaya et al., 2011). Sánchez-Meca and Marín-Martínez (1997) compared the performance of the $Q$-statistic with the Schmidt and Hunter rule for homogeneity on a number of continuous outcomes (Sánchez-Meca & Marín-Martínez, 1997). The outcomes that were investigated were the standardized mean difference, the point-biserial correlation coefficient and the standardized correlation coefficient. The Schmidt and Hunter rule rejects the homogeneity of treatment effects if the sampling error variance is less than a specific threshold of the total variance (e.g. 75% or 90%). Sánchez-Meca et al. concluded that all tests had low statistical power, the $Q$-statistic resulted in nominal Type 1 errors while the Schmidt and Hunter rule overestimated the Type 1 errors (Sánchez-Meca & Marín-Martínez, 1997). Jones, O’Gorman, Lemke, Woolson and Monte Carlo (1989) examined the performances of the following tests: likelihood ratio test, Pearson’s chi-square test, Breslow–Day test (Jones et al., 1989) and Tarone’s adjustment to it (Tarone, 1985), a conditional score test and Liang and Self’s normal approximation to the score test (Liang & Self, 1989). They found that the likelihood ratio test was too liberal. They recommended using the Breslow–Day statistic when the number of studies is large, and Liang and Self’s normal approximation to the score test when the data are sparse, i.e. when there are few participants in some studies. Similar results were also reached by Paul and Donner (1992). Jones et al. (1989) concluded that all tests considered had low power when the number of studies is small (Jones et al., 1989). This was supported by Gavaghan, Moore, and McQuay (2000) who examined the Peto statistic (Yusuf et al., 1986), the Woolf statistic (Woolf, 1955), the $Q$-statistic (applied to the estimates of the risk difference) (DerSimonian & Laird, 1986), Liang and Self’s normal approximation to the score test, and the Breslow–Day test. These tests were also found to have incorrect Type 1 error rates in case of truly homogeneous sets. The authors concluded that the $Q$-statistic was too liberal and that the Woolf statistic was conservative, and they recommended the Breslow–Day test for routine use (Gavaghan et al., 2000). Recently, the Cochrane Reviews have started including $I^2$ in its reviews (Breslow, 1994). However, the $I^2$ statistic has been shown to be conservative and has a low power for continuous outcomes when the number of studies is small (Huedo-Medina et al., 2006). Bagheri, Ayatollahi and Safari (2011) compared the performance of the random-effects logistic regression model (Bagheri et al., 2011) to Cochran’s $Q$-statistic and the Breslow–Day test, and found that the Breslow–Day test had the highest power. The authors recommended using the random-effects logistic regression model citing practical advantages, though these models are known for their computational difficulties (Wolfinger & O’Connell, 1993).

This overview demonstrates that a lot of work has already been done. However, some measures have never been investigated (e.g. Bliss statistic), and not all measures have been investigated in all settings. More importantly, all studies assumed that the underlying binomial probabilities for
the treatments in the stratified contingency tables would be satisfied and used this in their simulations, while in practical situations this would typically be violated. The aggregate data in a contingency table would be generated by individuals all having a different probability of obtaining the clinical outcome, due to other variables like age and gender. It is unclear whether the above-mentioned statistical methods will remain robust in more realistic practical settings. We use an extensive simulation study trying to account for hidden variations that can occur in practice, and we study the performance of 10 methods, including statistical methods that have not been applied before for this purpose.

This article is organized as follows. In Section 2 the statistical approaches for testing the homogeneity of odds ratios are presented, assuming that there does not exist individual risk. Section 3 describes the simulation model used where individual risk scores are being simulated. In Section 4 we present the simulation results and a case study on tuberculosis to illustrate the methods. The final section provides a discussion of the results and recommendation for use of the methods and possible further research.

2. Statistical models

Before we introduce the methods, we first introduce notation that will be used throughout this section. The binary exposure or treatment will be referred to as just treatment and the clinical outcome is considered binary. Then let $X_{hi}$ be the number of successes for treatment $h$ ($= 0, 1$) in study $i$ ($= 1, 2, ..., m$) out of $n_{hi}$ trials. The control group is indexed by $h = 0$ while the treatment group is indexed by $h = 1$. The random variables $X_{h1}, X_{h2}, ..., X_{hm}$ are assumed to be mutually independent (independent studies). For the methods that test homogeneity of effect sizes, it is assumed that the expected value of $X_{hi}$ is equal to $\mathbb{E}(X_{hi}) = n_{hi} \cdot p_{hi}$, with $p_{hi}$ being the proportion of a success for treatment $h$ in study $i$. All methods for testing homogeneity will assume that the proportion $p_{hi}$ satisfy the following form $p_{hi} = F(\alpha_i + (\beta + \gamma_i) \cdot t_{hi})$, with $F$ being some kind of a function (typically Logistic), $\alpha_i$ an intercept for study $i$, $\beta$ the (mean) treatment effect, $t_{hi}$ a treatment indicator variable for study $i$ with value 1 when $h = 1$ and 0 otherwise, and $\gamma_i$ an interaction effect for study $i$. Then homogeneity means that $H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_m = 0 \quad (1)$ when the parameters $\gamma_i$s are considered fixed or $H_0 : \text{var}(\gamma_i) = 0 \quad (2)$ when they are being assumed random.

Note that in practice the analysis of meta-analysis could also include one or more available covariates, e.g. average age per study or percentage males or females. In this setting, however, we assume that the only available information is the number of successes and the number of units in the treatment and control groups. This way the statistical methods discussed below can be applied to data with the least available information. Many of these methods can be easily generalized to accommodate extra covariates.

2.1. Logistic regression approaches

For fixed-effects logistic regression analysis, the function $F$ is a distribution function of the form $F(x) = \exp(x)/(1 + \exp(x))$. We also need to impose a constraint to obtain identifiability of the model. A natural choice is to take $\gamma_1 = 0$ or $\gamma_m = 0$. Additionally, logistic regression implies that $X_{hi}$ is binomial and that $X_0$ and $X_1$ are independent (this is often not true due to different risks for individuals, but without additional information $X_{hi}$ is assumed binomial).

To test (1) we apply the likelihood ratio test. The likelihood of the full model for $F$ being the logistic distribution function is
\[ l(a_i, \beta, \gamma_i) = \sum_{i=1}^{n_i} (X_i a_i + X_{ij} \beta_i + \log(1 - F(a_i)) + \log(1 - F(a_i + \beta_i))) \]

with \( X_i = X_{i0} + X_{ij} \) being the total number of events and \( \beta_i = \beta + \gamma_i \). The maximum likelihood estimates are denoted by \( \hat{a}_i^f, \hat{\beta}^f \) and \( \hat{\gamma}_i^f \) for the full model and \( \hat{a}_i^0 \) and \( \hat{\beta}^0 \) under the null hypothesis. The likelihood ratio test is now defined by

\[ \text{LRT} = -2 \left( l(\hat{a}_i^0, \hat{\beta}^0, 0) - l(\hat{a}_i^f, \hat{\beta}^f, \hat{\gamma}_i^f) \right) \]

The LRT has an approximate \( \chi^2 \) distribution with \( m - 1 \) degrees of freedom [28]. The LRT statistic can be determined with standard software packages (we applied the GENMOD procedure of SAS).

Instead of a fixed-effects model, the interaction effects \( \gamma_i \) can be taken random for the logistic link function. Under this assumption, it is common to also assume that the intercepts \( a_i \) are random. Thus, we will assume that \( (a_i, \gamma_i)^T \) are bivariate normal with mean \( (\alpha, 0)^T \) and variance-covariance matrix \( \Sigma \) given by

\[ \Sigma = \begin{bmatrix} \sigma^2_\alpha & \rho \sigma_\alpha \sigma_\gamma \\ \rho \sigma_\alpha \sigma_\gamma & \sigma^2_\gamma \end{bmatrix}, \]

with \( \sigma^2_\alpha \) being the standard deviation for the random intercepts, \( \sigma^2_\gamma \) the standard deviation for the random interaction term and \( \rho \) the correlation between \( a_i \) and \( \gamma_i \).

The number of events \( X_{ih} \), conditionally on \( a_i \) and \( \gamma_i \), is considered binomial with parameters \( n_i \) and \( p_{ih} = F(a_i + (\beta + \gamma_i) t_{ih}) \) and \( X_{i0} \) and \( X_{ij} \) are being independent given \( a_i \) and \( \gamma_i \). The log likelihood of the full model can be written as:

\[ l(a_i, \beta, \gamma_i) = \sum_{i=1}^{n_i} \log \left( \frac{n_i}{X_{i0}} \right) \left[ F(z_0 + z_1 x_{ih}) \right] X_{ih} \left[ 1 - F(z_0 + z_1 x_{ih}) \right]^{n_i - X_{ih}} \Phi_2(z_0 - \alpha, z_1 - \beta | \Sigma) dz_0 dz_1 \]

with \( \Phi_2(z_0, z_1 | \Sigma) \) being the bivariate normal density given by

\[ \frac{1}{2 \sigma_{\alpha \gamma} \sqrt{1 - \rho^2}} \exp \left\{ - \frac{1}{2 (1 - \rho^2)} \left[ \frac{z_0^2}{\sigma^2_0} - 2 \rho \frac{z_0 z_1}{\sigma_{\alpha \gamma}} + \frac{z_1^2}{\sigma^2_1} \right] \right\} \]

Homogeneity hypothesis in (2) can be evaluated using a likelihood ratio test again. The LRT becomes

\[ \text{LRT} = -2 \left( l(\hat{a}_i^0, \hat{\beta}^0, 0, \hat{\gamma}_i^0) - l(\hat{a}_i^f, \hat{\beta}^f, 0, \hat{\gamma}_i^f) \right) \]

with \( \hat{a}_i^f, \hat{\beta}^f \) and \( \hat{\gamma}_i^f \) being the maximum likelihood estimators under the full model and \( \hat{a}_i^0, \hat{\beta}^0 \) and \( \hat{\gamma}_i^0 \) being the maximum likelihood estimators under the null hypothesis. Note that the variance-covariance matrix \( \Sigma \) becomes

\[ \Sigma = \begin{bmatrix} \sigma^2_\alpha & 0 \\ 0 & \sigma^2_\gamma \end{bmatrix} \]

when the null hypothesis would be true. The approximate asymptotic distribution of the LRT is a mixture of a chi-square distribution with two degrees of freedom and a degenerate distribution in zero. This means that we should compare the LRT statistic with a chi-square value at significance level \( 2 \alpha \) (McCulloch, Searle, & Neuhaus, 2008; Verbeke & Molenberghs, 2003). The LRT statistic can be determined with standard software packages (we applied the NLMIXED procedure of SAS).

### 2.2. Approaches based on the Q-statistic

The Q-statistic was developed by Cochran (1954) and it can be applied to test the homogeneity of any type of effect size. The Q-statistic is given by

\[ Q = \sum_{i=1}^{n_i} (\hat{\mu}_i - \bar{\mu})^2 / s_{\mu}^2, \]

with \( \hat{\mu}_i \) being the
estimated effect size, $\tilde{\beta}$ a weighted average, and $se^2$ an estimated standard error for $\tilde{\beta}$. For the log odds ratio of study $i$, we would have $\tilde{\beta}_i = \log \left( \frac{X_{1i} (n_{0i} - X_0)}{(n_{1i} - X_{1i}) X_0} \right)$; $\tilde{\beta}$ is the weighted average which is given by

$$\tilde{\beta} = \frac{\sum_{i=1}^{m} (\tilde{\beta}_i / se^2_i)}{\sum_{i=1}^{m} (1 / se^2_i)}$$

with $se^2_i$ being given by

$$se^2_i = \frac{1}{X_{1i}} + \frac{1}{n_{1i} - X_{1i}} + \frac{1}{X_0} + \frac{1}{n_{0i} - X_0}$$

Cochran showed that the $Q$-statistic is approximately $\chi^2$ distributed with $m - 1$ degrees of freedom, although he did not apply or discuss log odds ratios as effect sizes. This was done by Woolf (1955).

Bliss (1952) used the $Q$-statistic in a slightly different way to test the homogeneity hypothesis. Bliss’ test statistic is given by

$$B = (m - 1) + \sqrt{(n - 4)} / (\sqrt{n} - 1) \cdot Q / (n - (m - 1))$$

with $n = (\sum_{i=1}^{m} (\sum_{n=0}^{n} n_{hi} - 2)) / m$.

Note that Bliss’ method was originally developed for testing the equality of potencies obtained from multiple bioassays similar to that of Cochran. In such cases, $n$ is the average of the corresponding degrees of freedom over all bioassays. The corresponding degrees of freedom per bioassay is obtained as a result of applying a parallel line bioassay model, i.e. a linear regression model (Mzolo et al., 2013). In our case we apply a fixed-effects logistic regression per study, so two parameters are estimated per study. Hence, we have chosen the degrees of freedom per study equal to $\sum_{i=1}^{m} (\sum_{n=0}^{n} n_{hi} - 2)$. Under the null hypothesis, the $B$-statistic may be approximated by a $\chi^2$ distributed random variable with $m - 1$ degrees of freedom.

As mentioned in Section 1, Higgins and Thompson (2002) introduced a few measures for quantifying the heterogeneity of treatment effects that are based on the $Q$-statistic (Sutton & Higgins, 2008). The authors suggested a homogeneity test based on a confidence interval of the $I^2$ statistic, given by $I^2 = 100 \cdot (H^2 - 1) / H^2$ with $H^2 = Q / (m - 1)$. A confidence interval for $H$ is given by $\exp (\log (H) \pm z_{\alpha/2} \cdot se_{log(H)})$ with $z_{\alpha/2}$ being the $\alpha/2$ quantile of the standard normal distribution and

$$se_{log(H)} = \left\{ \begin{array}{ll}
(1 / [(log(Q) - log(m - 1)) / [\sqrt{2Q} - \sqrt{2(m - 3)}]]) & \text{if } Q > m \\
(1 / [(2(m - 2))^{-1} [3(m - 2)^2 - 1] / [3(m - 2)^2]]) & \text{if } Q \leq m
\end{array}\right. $$

This confidence interval for $H$ can be transformed to a confidence interval for $I^2$, and the transformed confidence interval can be used to test hypothesis in (1) by verifying whether the value zero lies in the interval or not.

### 2.3. Chi-square approaches

One of the oldest approaches for testing homogeneity of odds ratio is the Breslow–Day test (Breslow & Day, 1980). They used the Cochran–Mantel–Haenszel pooled odds ratio $OR_{CMH}$ calculated by

$$O^r R_{CMH} = \frac{\sum_{i=1}^{m} (X_{1i} \cdot (n_{0i} - X_0) / n_i)}{\sum_{i=1}^{m} (X_{0i} \cdot (n_{1i} - X_{1i}) / n_i)}$$

and assumed the same distributional assumptions for $X_{0i}$ as for the fixed-effects logistic regression analysis. Now define $E_{Xi} = E(X_{1i}|X_{i}, O^r R_{CMH})$ as the expected value of $X_{1i}$ given $X_{i} = X_{0i} + X_{1i}$ and
O^\prime R_{CMH} under the assumption that the null hypothesis in (1) is correct. E_i, can be obtained by solving the following quadratic equation:

\[(O^\prime R_{CMH} - 1)E_i^2 - (\sum X_i n_{ij} + (n_{0i} - X_i)) E_i + X_i n_{ij} O^\prime R_{CMH} = 0 \] (3)

For testing the homogeneity, the Tarone test statistic (Tarone, 1985) is used. This statistic is given by

\[\frac{\sum_{i=1}^{m} \frac{(X_{ji} - E_{ji})^2}{\text{var}(X_{ji}|X_i, O^\prime R_{CMH})} - \sum_{i=1}^{m} \frac{(X_i - E_i)^2}{\text{var}(X_i|X_i, O^\prime R_{CMH})}}{m-1} \] (4)

with \(X_i = \sum_{j=1}^{J} X_{ji}, E_i = \sum_{j=1}^{J} E_{ji}\) and \(\text{var}(X_{ji}|X_i, O^\prime R_{CMH})\) is given by

\[\text{var}(X_{ji}|X_i, O^\prime R_{CMH}) = \left(\frac{1}{E_{ji}^2} + \frac{1}{X_i - E_{ji}} + \frac{1}{n_{0i} - E_{ji}} + \frac{1}{n_{0i} - X_i + E_{ji}}\right)^{-1} \] (5)

Note that we applied the Tarone statistic instead of Breslow–Day statistic. The Breslow–Day statistic is based on the Cochran–Mantel–Haenszel odds ratio estimator. This estimator has been shown to be consistent but not efficient. For this reason Tarone modified the Breslow–Day statistic (Tarone, 1985). If (1) is true, the statistic in (4) approximately follows a \(\chi^2\) distribution with \(m - 1\) degrees of freedom. The Tarone-adjusted Breslow–Day test can be carried out using standard software packages (we applied the procedure FREQ of SAS).

Zelen (1971) also presented a statistic to test the homogeneity of odds ratios (Zelen, 1971). However, it was found to be biased and inconsistent (Halperin et al., 1977). Instead, Halperin et al. (1977) presented a corrected version of the Zelen statistic, which we will use. It also uses an expected value \(E_{ji}\), like Breslow–Day, but now with an alternative estimate of the pooled odds ratio. The odds ratio \(O^\prime R_{CZ}\) is given by \(O^\prime R_{CZ} = \exp(\hat{\beta}^0)\) with \(\hat{\beta}^0\) being the maximum likelihood estimator of the Fixed-effects logistic regression analysis under the null hypothesis of homogeneity. The expected value \(E_{ji} = E(X_{ji}|X_i, O^\prime R_{CZ})\) can be obtained by solving Equation (3) with \(O^\prime R_{CMH}\) being replaced by \(O^\prime R_{CZ}\).

Then the corrected Zelen statistic for testing homogeneity is given by

\[Z = \sum_{i=1}^{m} \frac{(X_{ji} - E_{ji})^2}{\text{var}(X_{ji}|X_i, O^\prime R_{CZ})} \]

with the variance given in (5) using the solution \(E_{ji} = E(X_{ji}|X_i, O^\prime R_{CZ})\). If the null hypothesis in (1) is true, the corrected \(Z\) statistic will follow an asymptotic \(\chi^2\) distribution with \(m - 1\) degrees of freedom (Halperin et al., 1977).

Liang and Self (1989) presented test statistics for sparse data (Liang & Self, 1989). (Recall that sparse data have low cell counts, thus not necessarily low number of studies.) Again, an expected value \(E_{ji}\) is used given \(X_i\) and a pooled odds ratio \(O^\prime R_{CL}\). Here \(O^\prime R_{CL}\) is given by \(O^\prime R_{CL} = \exp(\hat{\beta}^0_{CL})\) with \(\hat{\beta}^0_{CL}\) being the maximum likelihood estimator of the conditional likelihood function given \(X_i\) and \(X_1 = X_2 = \ldots = X_m = 0\). The conditional logistic regression can be obtained using standard software packages. (We applied the procedure LOGISTIC of SAS.) The expected value \(E_{ji} = E(X_{ji}|X_i, O^\prime R_{CL})\) can be obtained by solving Equation (3) with \(O^\prime R_{CMH}\) being replaced by \(O^\prime R_{CL}\). Liang and Self test statistic \(T_2\) is defined by

\[T_2 = \sum_{i=1}^{m} \frac{(X_{ji} - E_{ji})^2}{\text{var}(X_{ji}|X_i, O^\prime R_{CL})} \]
with \( \text{var}(X_i \mid X_i, O^v R_{CL}) \) being given by (5) and \( E_i \) being the described solution of (3) using \( O^v R_{CL} \). The statistic \( T_2 \) is approximately chi-squared distributed with \( m - 1 \) degrees of freedom (Liang & Self, 1989).

Following Liang and Self’s approach, we derived an alternative statistic. Instead of using the conditional logistic regression, we apply a random-effects logistic regression model, assuming that \( \text{var}(\gamma_i) = 0 \). The odds ratio we apply is \( X \), such that

\[
\lambda_i = \frac{a_i}{b_i},
\]

for study \( i \). The test statistic \( T_6 \) is equal to \( T_2 \) but with \( E_{ij} \) the solution of (3) with \( O^v R_{CL} \) replaced by \( O^v R_{CL} \) and the variance given by (4), with \( E_{ij} \) the described solution of (3) using \( O^v R_{CL} \).

Finally, a complete alternative \( \chi^2 \) statistic is the Peto statistic. Yusuf et al. (1986) presented the Peto statistic to test the homogeneity of the odds ratios [10]. The Peto statistic \( P \) is given by

\[
P = \frac{\sum_{i=1}^{m} (O_i - E_i)^2}{\sum_{i=1}^{m} V_i} - \frac{(\sum_{i=1}^{m} (O_i - E_i))^2}{\sum_{i=1}^{m} V_i}
\]

where \( O_i = X_{i1}, E_i = X_i n_{3i} / n_i, n_i = n_{3i} + n_{0i} \) and \( V_i \) defined by

\[
V_i = \frac{X_i \cdot (n_i - X_i) \cdot n_{3i} \cdot n_{0i}}{n_i^2 \cdot (n_i - 1)}
\]

When the null hypothesis in (1) is true, the Peto statistic will have an approximate \( \chi^2 \) distribution with \( m - 1 \) degrees of freedom (Yusuf et al., 1986).

3. Simulation

3.1. Introduction

Although all approaches discussed above assume equal risk probabilities for all participants in one study for one treatment group, this assumption is hardly true in practice. In reality, risk probabilities of individuals will be influenced by individual’s characteristics. Therefore, the data are simulated so that each individual has its unique risk probability. This is done by creating covariates for each individual that affect the risk probability in addition to the influence of the treatment. Since populations across studies vary, the distributions of the covariates are affected by studies, causing already one form of treatment–study interaction. We also added treatment–study interactions independent of the covariates.

3.2. Simulation model and settings

The simulation model to generate \( m \) studies can be described as follows. First, the total number of participants \( n_i \) for study \( i \) is drawn. We used an overdispersed Poisson distribution with parameter \( \lambda \). The value \( \gamma_i \sim \Gamma(a_0, b_0) \) is drawn to make a study-specific parameter \( \lambda_i = \lambda \cdot \exp(0.5 \cdot \gamma_i) \). Then \( n_i \) is drawn from a Poisson distribution \( (\lambda_i) \). As mentioned in Section 1, to make the simulation somewhat more realistic we generated covariates, in this case age and gender. The covariate age \( a_{ij} \) for the \( j \)th participant in study \( i \) is created by

\[
a_{ij} = a_0 + \eta_j + \epsilon_{ij},
\]

with \( a_0 \) being the average age over all participants in all studies, \( \eta_j \sim N(0, \tau_j^2) \) a study effect for age and \( \epsilon_{ij} \sim N\left(0, \tau_{a(i)}^2\right) \) the age effect within study \( i \). The covariate gender \( g_{ij} \) for the \( j \)th participant in study \( i \) is generated with a uniform variable \( z_j \). The \( j \)th participant will be assigned the value \( g_{ij} = 1 \) if \( z_j > 0.5 \) and zero otherwise.

Then we need to generate study-specific treatment effects independent of covariates. We use a bivariate normally distributed variable \( (u_{0i}, u_{1i}) \) such that

\[
\begin{align*}
\text{var}(\epsilon_{ij}) &= \tau_{a(i)}^2, \\
\text{cov}(\epsilon_{ij}, \epsilon_{il}) &= \tau_{a(i)}^2, \\
\text{cov}(\epsilon_{ij}, \epsilon_{ij}) &= \tau_j^2, \\
\text{cov}(\epsilon_{ij}, \epsilon_{il}) &= 0, \\
\text{cov}(\epsilon_{ij}, \epsilon_{ij}) &= 0
\end{align*}
\]
\[
\begin{bmatrix}
\alpha_0 \\
\alpha_1
\end{bmatrix} 
\sim N\left(\begin{bmatrix}
\alpha \\
\alpha + \beta
\end{bmatrix}; \begin{bmatrix}
\sigma_0^2 & \rho \sigma_0 \sigma_1 \\
\rho \sigma_0 \sigma_1 & \sigma_1^2
\end{bmatrix}\right)
\]

The parameter \(\alpha\) represents an intercept to generate the risk of an event for the control group and \(\beta\) is the direct treatment effect. Additionally, the random variable \(\nu_i\) is drawn from a normal distribution, \(\nu_i \sim N(0, \sigma^2)\), to generate heterogeneity across participants in the study due to variables other than age and gender.

A treatment indicator \(\delta_{ij}\) is generated by a uniform \((0, 1)\) distribution. If \(w_{ij}\) is uniformly distributed, then \(\delta_{ij} = 1\) if \(w_{ij} > \rho\) and zero otherwise, with \(\rho\) being the ratio of allocation of control and treatment. The event \(X_{ij}\) is Bernoulli \((p_{ij})\) distributed with \(p_{ij}\) satisfying

\[
\log \left(\frac{p_{ij}}{1 - p_{ij}}\right) = \left(a_0 + \beta_1 \cdot \delta_{ij}\right) \cdot a_0 + \alpha_0 \cdot g_{ij} + u_{ij} + \nu_{ij}
\]

with \(a_0\) being the effect of age on the risk of an event irrespective of treatment, \(\beta_1\) a treatment age interaction effect, \(a_0\) the effect of gender and \(u_{ij} = u_{01}(1 - \delta_{ij}) + u_{11}\delta_{ij}\) the study treatment effect. Finally, the dichotomous response variable \(X_{ij}\) is drawn with a uniform \((0,1)\) distribution, i.e. \(X_{ij} = 1\) if \(y_{ij} > p_{ij}\) and zero otherwise, with \(y_{ij} \sim U(0, 1)\). This procedure is repeated 1,000 times resulting in 1,000 simulation runs for any set of parameters. As per simulation run, it is determined whether the null hypothesis of homogeneity is rejected or not using the approaches in Section 2 which all ignore the individual heterogeneities. The percentage of rejected null hypotheses over the number of simulations is being reported.

Note that the final model in (6) has a study-specific latent variable \(Z_{0i} = a_0\theta_i + u_{0i}\) for the control group, and a study-specific latent variable \(Z_{1i} = (a_0 + \beta_1)\eta_i + u_{1i}\) for the treatment group. The variance–covariance matrix for these latent variables is given by

\[
\begin{bmatrix}
\alpha_0^2 \tau_0^2 + \sigma_0^2 & \alpha_0(a_0 + \beta_1) \tau_0^2 + \rho \sigma_0 \sigma_1 \\
\alpha_0(a_0 + \beta_1) \tau_0^2 + \rho \sigma_0 \sigma_1 & (a_0 + \beta_1)^2 \tau_0^2 + \sigma_1^2
\end{bmatrix}
\]

when \(\beta_1 = 0\), \(\rho = 1\) and \(\sigma_0 = \sigma_1\); the correlation between \(Z_{0i}\) and \(Z_{1i}\) is equal to 1 and there is no treatment–study interaction. In case \(\beta_1 \neq 0\) or \(\rho \neq 1\), there is a treatment–study interaction and thus heterogeneity in treatment effects. The size of the heterogeneity is given by \(\beta_1^2 \tau_0^2 + 2\sigma_0^2(1 - \rho)\), when \(\sigma_0 = \sigma_1\). The larger this value, the higher the statistical power of the methods. Thus, the heterogeneity reduces when \(\rho\) increases to 1. For the simulation, a choice was made to take \(\sigma_0^2 = \sigma_1^2\). Type I error was simulated assuming \(\rho = 1\) and \(\beta_1 = 0\). The different simulation settings that we studied are presented in Table 1.

### 4. Results

#### 4.1. Meta-analysis of antituberculosis therapy

We applied the statistical methods presented in Section 2 to a meta-analysis on antituberculosis therapy (Pasipanodya, Srivastava, & Gumbo, 2012). This meta-analysis consisted of 26 different regimens (plans or regulated courses). In each regimen, slow acetylators (control) and rapid acetylators (treatment) were applied to two groups of patients. The objective was to test whether microbiological failure after therapy, relapse after therapy and acquired drug resistance (ADR) were as likely to occur among slow and rapid acetylators. The authors limited their selection of studies to randomized, controlled clinical trials in which outcomes were reported in tables, graphs or text format. All patients who were considered microbiological failures, relapses or ADR during treatment for slow and rapid acetylators were recorded. For each of the three different outcomes, the authors then calculated risk ratios (RR) with 95% confidence intervals using regimen as unit of analysis. Random-effects models were used to combine regimens when \(I^2 > 50\%\) and the corresponding \(P\)-value < 0.05; otherwise fixed-effects models were used to combine the RR.
For our study we only consider the results of one outcome, namely microbiological failure. The authors performed a test of homogeneity of the relative risks using the $I^2$ and found $I^2 = 39\%$ with a corresponding $P = 0.035$ (Pasipanodya et al., 2012). The authors then applied mixed-effects modeling for this outcome, resulting in a combined RR of 2, with a confidence interval [1.5 ; 2.7] for the RR.

However, four of the 26 regimens contained zero cells, and the selected methods cannot deal with this unless a correction to the number of cells is made (Gart, 1966). We chose to remove these four regimens from our data, and apply all the selected approaches to the remaining datasets.

The test statistic, degrees of freedom and the $P$-value for rejecting the null hypothesis are presented in Table 2. Using 5% level of significance, the null hypothesis is only rejected by the

| Number of studies | $m$ | 5 ; 10 ; 15 ; 20 |
|-------------------|-----|------------------|
| Poisson distribution parameter for a number of participants per study | $\lambda$ | 100 ; 200 |
| Overdispersion on a number of participants | $a_0 = b_0$ | 1 |
| Mean age | $a_0$ | 65 |
| Between-study variance in age | $\tau^2$ | 10 |
| Within-study variance in age | $\tau^2_{a(x)}$ | 15 |
| Heterogeneity in treatment effect | $\rho$ | 1 ; 0.3 ; 0.6 |
| Allocation proportion of treatment | $\rho$ | 0.3 ; 0.5 |
| Overall risk intercept | $\alpha$ | -6 |
| Age effect on risk | $a_a$ | 0.05 |
| Gender effect on risk | $a_g$ | 0.1 |
| Treatment effect | $\beta$ | 2 |
| Age-treatment interaction effect | $\beta_a$ | 0 ; 0.01 |
| Between-study variance on risk for treatment $h$ | $\sigma^2_h$ | 0.2 ; 0.5 |

| Test statistic | DF | $P$-value |
|----------------|----|-----------|
| FE | 33.31 | 21 | 0.124 |
| RE | 5.33 | 2 | 0.035 |
| Q | 29.92 | 21 | 0.094 |
| B | 29.32 | 21 | 0.107 |
| $\rho$ | [0 ; +58.30] | — | >0.05 |
| BD | 32.23 | 21 | 0.055 |
| Z | 32.27 | 21 | 0.055 |
| $T_a$ | 32.24 | 21 | 0.055 |
| $T_R$ | 32.28 | 21 | 0.055 |
| P | 31.73 | 21 | 0.062 |

Note that we could not reject the heterogeneity in treatment effects with $I^2$, as the authors suggested. This did not change when we included the four studies and changed the cell counts with 0.5.

For our study we only consider the results of one outcome, namely microbiological failure. The authors performed a test of homogeneity of the relative risks using the $I^2$ and found $I^2 = 39\%$ with a corresponding $P = 0.035$ (Pasipanodya et al., 2012). The authors then applied mixed-effects modeling for this outcome, resulting in a combined RR of 2, with a confidence interval [1.5 ; 2.7] for the RR.

However, four of the 26 regimens contained zero cells, and the selected methods cannot deal with this unless a correction to the number of cells is made (Gart, 1966). We chose to remove these four regimens from our data, and apply all the selected approaches to the remaining datasets.

The test statistic, degrees of freedom and the $P$-value for rejecting the null hypothesis are presented in Table 2. Using 5% level of significance, the null hypothesis is only rejected by the
random-effects logistic regression model. Additionally, the $P$-values range from 0.035 to 0.124. Thus, the proposed methods do not all seem to agree.

4.2. Simulation results

The full set of results is presented in the supplementary material. A subset of these results is shown in Tables 3–6. Tables 3–5 present a subset of the simulation results where $p_{hi}$ depends on age, gender and other heterogeneous factors ($v_{ij}$). They represent the results for the most realistic practical settings to contrast these results and to be able to compare our results with literature. Table 6 presents a selection of the Type I error and the statistical power results of the test methods under the assumption of the binomial distribution. In such case, the proportion of a success, $p_{hi}$, does not depend on age, gender or any other participant variables ($v_{ij}$). The random-effects logistic regression analysis occasionally showed numerical problems, in particular when heterogeneity between studies was small or not present at all. Note that in around 13.7% of all simulation runs the full random-effects logistic regression model could not converge and the results of these problematic simulation runs are eliminated from the results.

For the realistic practical settings (Table 3), the fixed-effects logistic regression analysis has a slight increased Type I error, when the number of participants within studies is relatively small, irrespective of the number of studies, but this is still acceptable. When sample sizes within studies increase, the Type I error is closer to nominal. The $Q$-statistic, the Bliss statistic and the $I^2$ statistic constantly underestimate the Type I error (Table 3). This is most apparent when only a third of the participants are exposed to the treatment. This gets slightly better when the number of studies increases (supplementary material). The $I^2$ statistic has the smallest Type I error rates of all the methods, except for a few cases where the Peto statistic gives even smaller values. The random-effects logistic regression model mostly underestimates the Type I error rate when $\sigma_0^2$ and $\sigma_1^2$ are small. However as $\sigma_0^2$ and $\sigma_1^2$ increase, it overestimates the Type I error rate (Table 2). This gets even worse when the number of studies increases (supplementary material). The Type I error rates of the chi-square methods using estimates of the odds ratio, i.e. the Breslow–Day statistic, the corrected Zelen statistic, Liang and Self’s $T_2$ and its adjusted version $T_{hi}$, are almost at the nominal value. However, when one-third of the participants is exposed to treatment, the chi-square methods stay just below the nominal level, but this disappears when the number of studies increases. When participants no longer have their individual risk, the Type I error rates (Table 5) are mostly similar to the Type I error rates for the realistic practical settings, except for two cases. The fixed-effects logistic regression no longer gives liberal Type I error rates when there is no treatment effect ($\beta = 0$).

Investigating the power for the realistic practical settings (Tables 4 and 5), it obviously shows that the statistical power increases as the number of studies increases. This improved power for testing the interaction effect can be explained by the reduced sampling error due to increased sample size. Furthermore, the increase in $\sigma_0^2$ and $\sigma_1^2$ results in higher power since they increase the treatment interaction effects. This is due to the term $2\sigma_0^2(1 - \rho)$ in the size of heterogeneity. The statistical power is negatively correlated with $\rho$, with no interaction effect when $\rho = 1$ and $\beta_0 = 0$ (as mentioned earlier). The highest statistical power is obtained by the fixed-effects logistic regression model, followed by the $\chi^2$ statistics (Breslow–Day test, the corrected Zelen test, $T_2$ and $T_{hi}$). The lowest statistical power is obtained by the $I^2$ and the random-effects logistic regression model, respectively. These results hold true irrespective of the number of studies. Although Bliss uses the $Q$-statistic differently from Cochran, the two tests behave almost identically but they are less powerful than the best test statistics. It should be noted that heterogeneity in treatment effects from differences in age effects across studies ($\beta_0 > 0$) is not increasing the power a lot, a result that holds even when the number of studies increases. This is due to the selected settings. The heterogeneity effect in the simulation was only increased with $\beta_0^2 \tau^2 = (0.01)^2 \cdot 10 = 0.001$, which is really small.
Table 3. Type I error rate for \( m = 5, a_n = 0.05, a_y = 0.1 \)

| \( \lambda \) | \( \beta \) | \( \sigma_y^2 \) | \( \rho \) | \( \text{FE} \) | \( \text{RE} \) | \( \text{Q} \) | \( \text{B} \) | \( \rho^2 \) | \( \text{BD} \) | \( Z \) | \( T_2 \) | \( T_8 \) | \( P \) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 0 | 0.2 | 0.3 | 0.060 | 0.024 | 0.028 | 0.027 | 0.020 | 0.045 | 0.044 | 0.045 | 0.045 | 0.042 |
| | | | | 0.5 | 0.060 | 0.033 | 0.039 | 0.037 | 0.025 | 0.054 | 0.054 | 0.054 | 0.054 | 0.050 |
| | | | | 0.5 | 0.060 | 0.059 | 0.020 | 0.015 | 0.010 | 0.040 | 0.040 | 0.040 | 0.040 | 0.036 |
| | | | | 0.5 | 0.064 | 0.054 | 0.030 | 0.024 | 0.014 | 0.050 | 0.050 | 0.050 | 0.050 | 0.046 |
| | 2 | 0.2 | 0.3 | 0.056 | 0.035 | 0.022 | 0.021 | 0.014 | 0.044 | 0.044 | 0.044 | 0.040 | 0.044 | 0.005 |
| | | | | 0.5 | 0.045 | 0.032 | 0.032 | 0.028 | 0.021 | 0.037 | 0.038 | 0.037 | 0.039 | 0.019 |
| | | | | 0.5 | 0.059 | 0.082 | 0.014 | 0.012 | 0.007 | 0.046 | 0.044 | 0.044 | 0.043 | 0.003 |
| | | | | 0.5 | 0.052 | 0.086 | 0.031 | 0.028 | 0.019 | 0.044 | 0.045 | 0.043 | 0.044 | 0.015 |
| 200 | 0 | 0.2 | 0.3 | 0.057 | 0.023 | 0.044 | 0.043 | 0.024 | 0.050 | 0.050 | 0.050 | 0.050 | 0.050 |
| | | | | 0.5 | 0.045 | 0.025 | 0.034 | 0.031 | 0.022 | 0.040 | 0.040 | 0.040 | 0.040 | 0.039 |
| | | | | 0.5 | 0.051 | 0.070 | 0.032 | 0.031 | 0.020 | 0.043 | 0.043 | 0.043 | 0.043 | 0.042 |
| | | | | 0.5 | 0.049 | 0.099 | 0.034 | 0.033 | 0.023 | 0.048 | 0.048 | 0.048 | 0.048 | 0.047 |
| | | | 2 | 0.2 | 0.3 | 0.055 | 0.036 | 0.037 | 0.032 | 0.014 | 0.045 | 0.045 | 0.045 | 0.044 |
| | | | | 0.5 | 0.036 | 0.040 | 0.028 | 0.028 | 0.017 | 0.031 | 0.034 | 0.033 | 0.033 | 0.017 |
| | | | | 0.5 | 0.052 | 0.106 | 0.031 | 0.030 | 0.015 | 0.046 | 0.050 | 0.049 | 0.050 | 0.008 |
| | | | | 0.5 | 0.050 | 0.111 | 0.037 | 0.032 | 0.020 | 0.043 | 0.043 | 0.043 | 0.043 | 0.013 |
| $\beta$ | $\beta_a$ | $\sigma^2_\beta$ | $\rho$ | FE | RE | Q | B | $\beta^2$ | BD | Z | $T^2$ | $T^4$ | P |
|------|----------|----------------|-----|-----|-----|----|----|---------|------|----|------|------|----|
| 0    | 0        | 0.2            | 0.3 | 0.234 | 0.154 | 0.174 | 0.160  | 0.135 | 0.223 | 0.224 | 0.223 | 0.225 | 0.222 |
|      | 0.5      | 0.269          | 0.153 | 0.225 | 0.210 | 0.173 | 0.258 | 0.258  | 0.258 | 0.258 | 0.258 | 0.258 | 0.250 |
|      | 0.5      | 0.501          | 0.367 | 0.359 | 0.344 | 0.299 | 0.481 | 0.481  | 0.481 | 0.481 | 0.481 | 0.481 | 0.460 |
|      | 0.01     | 0.2            | 0.3 | 0.252 | 0.158 | 0.192 | 0.172 | 0.142 | 0.237 | 0.242 | 0.241 | 0.242 | 0.186 |
|      | 0.5      | 0.315          | 0.193 | 0.272 | 0.255 | 0.206 | 0.301 | 0.302  | 0.304 | 0.302 | 0.302 | 0.274 |
|      | 0.5      | 0.520          | 0.423 | 0.417 | 0.401 | 0.347 | 0.498 | 0.500  | 0.500 | 0.500 | 0.500 | 0.500 | 0.400 |
|      | 0.5      | 0.589          | 0.465 | 0.522 | 0.512 | 0.465 | 0.578 | 0.577  | 0.577 | 0.577 | 0.577 | 0.548 |
| 2    | 0        | 0.2            | 0.3 | 0.264 | 0.167 | 0.203 | 0.188 | 0.158 | 0.258 | 0.267 | 0.263 | 0.255 | 0.116 |
|      | 0.5      | 0.323          | 0.210 | 0.284 | 0.275 | 0.234 | 0.315 | 0.318  | 0.315 | 0.316 | 0.316 | 0.212 |
|      | 0.5      | 0.540          | 0.451 | 0.419 | 0.407 | 0.357 | 0.510 | 0.520  | 0.517 | 0.520 | 0.520 | 0.334 |
|      | 0.01     | 0.2            | 0.3 | 0.255 | 0.161 | 0.205 | 0.188 | 0.153 | 0.260 | 0.266 | 0.264 | 0.264 | 0.110 |
|      | 0.5      | 0.303          | 0.190 | 0.276 | 0.260 | 0.221 | 0.301 | 0.306  | 0.304 | 0.307 | 0.307 | 0.173 |
|      | 0.5      | 0.536          | 0.405 | 0.418 | 0.403 | 0.363 | 0.507 | 0.521  | 0.518 | 0.520 | 0.520 | 0.343 |
|      | 0.5      | 0.628          | 0.503 | 0.575 | 0.560 | 0.514 | 0.623 | 0.629  | 0.652 | 0.624 | 0.624 | 0.467 |
| $\beta$ | $p_0$ | $Q_0$ | $p$ | FE | RE | Q | B | $\hat{\beta}$ | BD | Z | $T_2$ | $T_8$ | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0.2 | 0.3 | 0.166 | 0.092 | 0.118 | 0.111 | 0.087 | 0.159 | 0.158 | 0.158 | 0.159 | 0.158 |
| | | | 0.5 | 0.181 | 0.097 | 0.140 | 0.128 | 0.106 | 0.165 | 0.165 | 0.165 | 0.165 | 0.161 |
| | | | 0.5 | 0.352 | 0.245 | 0.229 | 0.213 | 0.175 | 0.337 | 0.337 | 0.337 | 0.336 | 0.330 |
| | | | 0.01 | 0.176 | 0.097 | 0.121 | 0.113 | 0.086 | 0.159 | 0.159 | 0.159 | 0.159 | 0.117 |
| | | | 0.5 | 0.203 | 0.129 | 0.169 | 0.157 | 0.127 | 0.193 | 0.194 | 0.193 | 0.193 | 0.174 |
| | | | 0.5 | 0.351 | 0.288 | 0.246 | 0.235 | 0.195 | 0.336 | 0.337 | 0.337 | 0.337 | 0.255 |
| | | | 0.5 | 0.422 | 0.309 | 0.350 | 0.331 | 0.292 | 0.401 | 0.402 | 0.402 | 0.402 | 0.362 |
| 2 | 0 | 0.2 | 0.3 | 0.179 | 0.093 | 0.118 | 0.106 | 0.082 | 0.164 | 0.167 | 0.166 | 0.165 | 0.047 |
| | | | 0.5 | 0.209 | 0.123 | 0.173 | 0.162 | 0.127 | 0.199 | 0.200 | 0.200 | 0.200 | 0.115 |
| | | | 0.5 | 0.369 | 0.300 | 0.254 | 0.243 | 0.208 | 0.339 | 0.347 | 0.346 | 0.347 | 0.183 |
| | | | 0.01 | 0.180 | 0.093 | 0.121 | 0.110 | 0.083 | 0.174 | 0.178 | 0.177 | 0.177 | 0.046 |
| | | | 0.5 | 0.182 | 0.107 | 0.157 | 0.145 | 0.121 | 0.172 | 0.175 | 0.175 | 0.174 | 0.089 |
| | | | 0.5 | 0.370 | 0.276 | 0.250 | 0.239 | 0.208 | 0.330 | 0.342 | 0.340 | 0.341 | 0.191 |
| | | | 0.5 | 0.442 | 0.312 | 0.380 | 0.367 | 0.324 | 0.427 | 0.435 | 0.432 | 0.432 | 0.255 |
Table 6. Type I error and statistical power for $\alpha - \alpha_g - \sigma^2 = 0, m = 5, l = 100$

| $\rho$ | $\beta$ | $\sigma_\alpha$ | $\rho$ | FE | RE | Q | B | $\bar{\mu}$ | BD | Z | $T_2$ | $T_R$ | $P$ |
|-------|--------|-----------------|-------|-----|----|---|---|---------|-----|---|-------|------|-----|
| 1     | 0.2    | 0.3             | 0.5   | 0.047 | 0.072 | 0.041 | 0.035 | 0.024 | 0.048 | 0.048 | 0.048 | 0.048 | 0.047 |
|       |        |                 | 0.5   | 0.050 | 0.084 | 0.045 | 0.041 | 0.028 | 0.049 | 0.049 | 0.049 | 0.049 | 0.046 |
|       |        |                 | 0.5   | 0.066 | 0.142 | 0.038 | 0.035 | 0.024 | 0.043 | 0.043 | 0.043 | 0.043 | 0.042 |
| 2     | 0.2    | 0.3             | 0.5   | 0.054 | 0.035 | 0.030 | 0.028 | 0.018 | 0.045 | 0.047 | 0.046 | 0.046 | 0.056 |
|       |        |                 | 0.5   | 0.058 | 0.041 | 0.026 | 0.023 | 0.011 | 0.049 | 0.052 | 0.050 | 0.051 | 0.004 |
|       |        |                 | 0.5   | 0.058 | 0.073 | 0.030 | 0.025 | 0.019 | 0.049 | 0.048 | 0.048 | 0.048 | 0.059 |
| 0.3   | 0.2    | 0.3             | 0.5   | 0.596 | 0.471 | 0.583 | 0.570 | 0.531 | 0.595 | 0.595 | 0.595 | 0.595 | 0.588 |
|       |        |                 | 0.5   | 0.656 | 0.510 | 0.646 | 0.632 | 0.585 | 0.651 | 0.651 | 0.651 | 0.651 | 0.645 |
|       |        |                 | 0.5   | 0.828 | 0.774 | 0.821 | 0.807 | 0.785 | 0.828 | 0.829 | 0.829 | 0.828 | 0.821 |
|       |        |                 | 0.5   | 0.876 | 0.791 | 0.869 | 0.864 | 0.839 | 0.874 | 0.874 | 0.874 | 0.874 | 0.869 |
| 2     | 0.2    | 0.3             | 0.5   | 0.443 | 0.252 | 0.390 | 0.370 | 0.320 | 0.432 | 0.437 | 0.437 | 0.437 | 0.422 |
|       |        |                 | 0.5   | 0.425 | 0.299 | 0.341 | 0.330 | 0.272 | 0.401 | 0.411 | 0.408 | 0.410 | 0.211 |
|       |        |                 | 0.5   | 0.726 | 0.587 | 0.659 | 0.641 | 0.596 | 0.716 | 0.716 | 0.716 | 0.717 | 0.699 |
| 0.6   | 0.2    | 0.3             | 0.5   | 0.727 | 0.602 | 0.626 | 0.607 | 0.568 | 0.706 | 0.712 | 0.710 | 0.710 | 0.534 |
|       |        |                 | 0.5   | 0.679 | 0.336 | 0.461 | 0.438 | 0.393 | 0.471 | 0.471 | 0.471 | 0.471 | 0.464 |
|       |        |                 | 0.5   | 0.679 | 0.623 | 0.680 | 0.663 | 0.630 | 0.695 | 0.695 | 0.695 | 0.695 | 0.695 |
|       |        |                 | 0.5   | 0.761 | 0.659 | 0.752 | 0.741 | 0.711 | 0.758 | 0.758 | 0.758 | 0.758 | 0.750 |
| 2     | 0.2    | 0.3             | 0.5   | 0.301 | 0.162 | 0.248 | 0.235 | 0.193 | 0.281 | 0.284 | 0.284 | 0.284 | 0.276 |
|       |        |                 | 0.5   | 0.277 | 0.187 | 0.210 | 0.200 | 0.164 | 0.254 | 0.262 | 0.260 | 0.260 | 0.111 |
|       |        |                 | 0.5   | 0.553 | 0.404 | 0.475 | 0.456 | 0.408 | 0.534 | 0.539 | 0.536 | 0.535 | 0.514 |
|       |        |                 | 0.5   | 0.553 | 0.420 | 0.457 | 0.441 | 0.396 | 0.535 | 0.540 | 0.536 | 0.537 | 0.359 |
Comparing the power results from the setting that participants have a common risk (Table 6) with the power results from the setting that individuals have their own individual risks (Tables 4 and 5) shows a few interesting differences. First, the power is substantially smaller when individuals are heterogeneous in their risk rates, which makes it more difficult to detect treatment heterogeneity across studies. The power values for testing treatment interactions increase slightly when there would be a treatment effect in case individuals have heterogeneous risk rates. However, this is opposite when individuals are not heterogeneous, since the power values for treatment heterogeneity is substantially smaller for $\beta = 2$ than for $\beta = 0$ (Table 6 and supplementary material). With respect to choosing the best methods for detecting treatment effect heterogeneity, no real difference could be detected.

5. Discussion

The purpose of this article was to investigate the performance of statistical methods that may test for the homogeneity of odds ratios in meta-analysis. Ten methods were applied to a real-life meta-analysis study and to practically realistic simulated datasets. The simulated datasets were created using different scenarios based on different numbers of studies, participants within studies, unbalanced number of observations for treatment and control groups, whether or not there is an actual treatment effect, individual risks through covariates and the level of heterogeneity.

The Peto statistic is somewhat unreliable with respect to its Type I error rate, a conclusion that has been reached earlier (Gavaghan et al., 2000). The fixed-effects logistic regression analysis is a slightly liberal test, although the inflated Type I error rate is not too large, and it reduces with larger number of studies. This conclusion has been reached earlier (Paul & Donner, 1992), although we could not demonstrate it for settings where all individuals have the same constant risk rates, the settings used in the literature. In our study the inflation happened when the proportion of success depends on individual covariates and with small numbers of studies. This interesting result merits further investigation to find out the exact analytical cause, but it is beyond the scope of this article.

The fact that the fixed logistic regression has the highest power among all other methods is partly due to this inflated Type I error rate. The I$^2$ statistic can be considered a very conservative test. This conclusion had been reached earlier (Huedo-Medina et al., 2006), but only for continuous responses. It makes the I$^2$ an unsuitable test statistic for homogeneity of odds ratios. The random-effects logistic regression analysis is also conservative, albeit less conservative than the I$^2$ for small numbers of studies, but it is less powerful than I$^2$ for a larger number of studies. This low power combined with the unreliable Type I error rate makes the random-effects model unsuitable for testing homogeneity; see also Bagheri et al. (2011). The Q-statistic and the Bliss statistic can also be considered conservative tests, a conclusion reached earlier (Higgins & Thompson, 2002; Higgins et al., 2003; Huedo-Medina et al., 2006). However, heterogeneity in the individual’s risk rates makes it more apparent. The Type I error rates of the chi-square methods are close to the nominal value, but Paul and Donner (1989) found $T_2$ to have liberal Type I error rates in the setting of a small number of centers and a large number of units. In our study, however, the $T_2$ (and all the $\chi^2$ methods) hardly overestimates the Type I error in such setting, but we did not study large numbers of participants within studies. The Type I error rate of $T_2$ is not affected by heterogeneity of individual’s risk rates. Since the allocation of participants to treatment and control groups seems to affect the statistical power, it should be taken into consideration. Additionally, the heterogeneity in individual’s risk rates reduces the power, in particular when there would be no treatment effects present.

Overall, the chi-square tests based on odds ratios and the fixed-effects logistic regression analysis perform best. We recommend the Breslow–Day test adjusted by Tarone for convenience purposes, since it is available in most statistical software packages and the other tests are not. The Breslow–Day test had been recommended before by Jones et al. (1989), but only when the number of studies is large. We also recommend the Breslow–Day test when the number of studies is small. However, when a treatment effect is suspected to be present and we have only a small number of
This article has its limitations. First, the conclusions presented only hold for dichotomous outcomes. Second, although the simulation method creates data on a unit (subject) level, hence adding more variation, it uses normally distributed latent variables. This assumption may not always be realistic in practice. Finite mixture models for the latent variables can be a potentially useful method. More specifically, the beta-binomial distribution (Skellam, 1948) could be contemplated to perform the interaction test. Another suggestion is to use the methods that estimate between-study variance; see e.g. Brockwell and Gordon (2001) and DerSimonian and Laird (1986) for an overview. The estimated confidence intervals of the between-study variance can be used to test the homogeneity of effect sizes. Finally, Bayesian approaches to test the interaction effects between study and treatment effect are other options for future research.

Supplementary material
Supplemental material for this article can be accessed at https://doi.org/10.1080/25742558.2018.1478698.

Acknowledgments
The authors are very grateful to the reviewer whose comments helped improve the results section and thereby expanding the relevance thereof.

Funding
This work was funded by the Netherlands Organization for Scientific Research [grant number: 023.005.087].

Author details
Osama Almalik1
E-mail: O.Amalik@tue.nl
Edwin R. van den Heuvel1
E-mail: e.r.v.d.heuvel@tue.nl,p.almalik@tue.nl
1 Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, the Netherlands.

Citation information
Cite this article as: Testing homogeneity of effect sizes in pooling 2x2 contingency tables from multiple studies: a comparison of methods, Osama Almalik, Edwin R. van den Heuvel & . Cogent Mathematics & Statistics (2018), 5: 1478698.

References
Agresti, A. (2002). Categorical data analysis (2nd ed.). New York, NY: John Wiley & Sons.
Bagheri, Z., Ayatollahi, S. M. T., & Jafari, P. (2011). Comparison of three tests of homogeneity of odds ratios in multicenter trials with unequal sample sizes within and among centers. BMC Medical Research Methodology, 11, 58.
Breslow, N. E., & Day, N. E. (1980). Statistical methods in cancer research 1980; Volume I—the design and analysis of case-control studies. Vol 32 of IARC Scientific Publications. Lyon: International Agency for Research on Cancer.
Bliss, C. L. (1952). The statistics of Bioassay: With special reference to the vitamins. New York, NY: Academic Press Inc.
Breslow, N. E. (1994). Statistics in epidemiology: The case-control study. Journal of the American Statistical Association, 91(433), 14–28. doi:10.1080/01621459.1996.10476660
Brockwell, S. E., & Gordon, J. R. (2001). A comparison of statistical methods for meta-analysis. Statistics in Medicine, 20(6), 825–840. doi:10.1002/(ISSN)1097-0258
Cochran, W. G. (1954). The combination of estimates from different experiments. Biometrics, 10, 101–129. doi:10.2307/3001666
DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Control Clinical Trials, 7(3), 177–188. doi:10.1016/0197-2456(86)90046-2
Gart, J. J. (1966). Alternative analyses of contingency tables. Journal of the Royal Statistical Society Series B (Methodological), 28(1), 164–179.
Gavaghan, D. J., Moore, R. A., & McQuay, H. J. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. Pain, 85(3), 415–424. doi:10.1016/S0304-3959(99)00302-4
Holperin, M., Ware, J. H., Byar, D. P., Mantel, N., Brown, C. C., Kozul, J., … Green, S. B. (1977). Testing for interaction in an I×J×K contingency table. Biometrika, 64, 271–275.
Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11), 1539–1558. doi:10.1002/sim.1186
Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. British Medical Journal, 327(7414), 557–560. doi:10.1136/bmj.327.7414.557
Hoaglin, D. C. (2016). Misunderstandings about Q and ‘Cochran’s Q test’ in meta-analysis. Statistics in Medicine, 35, 485–495. doi:10.1002/sim.v35.4
Huedo-Medina, T. B., Sánchez-Meca, J., Marin-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index. Psychological Methods, 11(2), 193–206. doi:10.1037/1082-989X.11.2.193
Jackson, D. (2006). The power of the standard test for the presence of heterogeneity in meta-analysis. Statistics in Medicine, 25, 2688–2699. doi:10.1002/sim.1937-0258
Jones, M. P., O’Gorman, T. W., Lemke, J. H., Woolson, R. F., & Monte Carlo, A. (1989). Investigation of homogeneity tests of the odds ratio under various sample size configurations. Biometrics, 45(1), 171–181. doi:10.2307/2532043
Kulinskaya, E., Dollinger, M. B., & Bjerkestøl, K. (2011). Testing for homogeneity in meta-analysis. I. The one parameter case: Standardized mean difference. *Biometrics, 67*, 203–212. doi:10.1111/j.1541-0420.2010.01442.x

Liang, K. Y., & Self, S. G. (1989). Tests of homogeneity of odds ratio when the data are sparse. *Biometrika, 72*, 353–358. doi:10.1093/biomet/72.2.353

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models* (2nd ed.). New Jersey: Wiley.

Mzolo, T., Hendriks, M., & van den Heuvel, E. (2013). A comparison of statistical methods for combining relative bioactivities from parallel line bioassays. *Pharmaceutical Statistics, 12*(6), 375–84.

Pasipanodya, J. G., Srivastava, S., & Gumba, T. (2012). Meta-analysis of clinical studies supports the pharmacokinetic variability hypothesis for acquired drug resistance and failure of antituberculosis therapy. *Clinical Infectious Diseases, 55*(2), 169–177. doi:10.1093/cid/cis353

Paul, S. R., & Donner, A. (1992). Small sample performance of tests of homogeneity of odds ratios in K \( \times \) 2 tables. *Statistics in Medicine, 11*(2), 159–165. doi:10.1002/sim.4780110203

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality and Quantity, 31*, 385–399. doi:10.1023/A:1004298118485

Skellam, J. G. (1946). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, 10*(2), 257–261.

Sutton, A. J., & Higgins, J. P. (2008). Recent developments in meta-analysis. *Statistics in Medicine, 27*(5), 625–650. doi:10.1002/sim.2934

Torone, R. E. (1985). On heterogeneity tests based on efficient scores. *Biometrika, 72*, 91–95. doi:10.1093/biomet/72.1.91

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics, 59*, 254–262. doi:10.1111/1541-0420.00032

Wolfinger, R., & O’Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation, 4*, 233–243. doi:10.1080/00949659308811554

Woof, B. (1955). On estimating the relationship between blood group and disease. *Annual Human Genetics, 19*, 251–253. doi:10.1111/j.1469-1809.1955.tb01348.x

Yusuf, S., Petco, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases, 27*(5), 335–371. doi:10.1016/S0033-0620(85)80003-7

Zelen, M. (1971). The analysis of several 2 × 2 contingency tables. *Biometrika, 58*, 129–137.