Investigating heterogeneity in meta-analysis of studies with rare events

Estimating the amount of heterogeneity

Dankmar Böhning¹ · Heinz Holling² · Walailuck Böhning² · Patarawan Sangnawakij³

Received: 3 July 2020 / Accepted: 5 May 2021 / Published online: 28 May 2021
© The Author(s) 2021

Abstract

In many meta-analyses, the variable of interest is frequently a count outcome reported in an intervention and a control group. Single- or double-zero studies are often observed in this type of data. Given this setting, the well-known Cochran’s Q statistic for testing homogeneity becomes undefined. In this paper, we propose two statistics for testing homogeneity of the risk ratio, particularly for application in the case of rare events in meta-analysis. The first one is a chi-square type statistic. It is constructed based on information of the conditional probability of the number of events in the treatment group given the total number of events. The second one is a likelihood ratio statistic, derived from the logistic regression models allowing fixed and random effects for the risk ratio. Both proposed statistics are well defined even in the situation of single-zero studies. In a simulation study, the proposed tests show a performance better than the traditional test in terms of type I error and power of the test under common and rare event situations. However, as the performance of the two newly proposed tests is still unsatisfactory in the very rare events setting, we suggest a bootstrap approach that does not rely on asymptotic distributional theory and it is shown that the bootstrap approach performs well in terms of type I error. Furthermore, a number of empirical meta-analyses are used to illustrate the methods.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s40300-021-00211-y.

✉ Dankmar Böhning
d.a.bohning@soton.ac.uk

Heinz Holling
holling@uni-muenster.de

Patarawan Sangnawakij
patarawan.s@gmail.com

¹ Mathematical Sciences and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, UK
² Statistics and Methods, Department of Psychology, University of Münster, 48149 Münster, Germany
³ Department of Mathematics and Statistics, Thammasat University, Pathum Thani 12120, Thailand
1 Introduction

Meta-analysis is a powerful statistical tool for analyzing and combining the results from several studies on the same topic. In the traditional 2-stage approach, meta-analysis of continuous or discrete endpoints requires in the first stage the effect measure estimate \( \hat{\theta}_i \) of the true parameter value \( \theta_i \) to be calculated from study \( i \), together with an estimated standard error. The analysis of these calculated estimates follows in the second stage, either using the fixed effect or random effects model. The fixed effect model allows \( \theta_i \) to have a common true parameter \( \theta \) across all studies. In the random effects model, \( \theta_i \) differs between studies. It is often assumed that across-study variations follow a normal distribution with mean \( \theta \) and between-study variance \( \tau^2 [7,20] \).

In medicine or psychology, counts of events such as the number of health events or deaths are of interest. When these count outcomes are observed, the risk ratio, risk difference, or odds ratio are considered as effect estimates. In many of these meta-analytic datasets, the number of events and number at risk or person-time are also available in an intervention compared with a control group. The risk ratio is often the quantity of interest when considering cohort studies or clinical trials. An overall risk ratio estimate is then computed as a weighted mean of the study-specific risk ratios where the weight for the specific study is obtained from the inverse of variance of the effect estimate, depending on whether the fixed or random effects model is used [5]. To decide on methods for combining studies and for concluding the consistency or inconsistency of findings, a statistical test for heterogeneity is therefore an important tool. As noted in Sánchez-Meca and Marán-Martánez [27], a test whether the effect measures in the studies are homogeneous is necessary when integrating results on a common topic. In other words, heterogeneity evaluation in meta-analysis provides the choice of synthesizing method, especially for interpretation of results. The conventional test for homogeneity is Cochran’s chi-square statistic \( Q \) (see Cochran [9] for count data, Sánchez-Meca and Marán-Martánez [28], Viechtbauer [30], and Kulinskaya et al. [15] for continuous data). However, as we will point out in the following, the \( Q \)-statistic becomes often undefined when the meta-analytic data contain rare events.

In fact, this work is motivated by meta-analytic data on myocardial infection events and cardiovascular deaths as adverse events when taking Rosiglitazone as anti-diabetes therapy in comparison with control. The data originally published by Nissen and Wolski [16] (taken from Böhning et al. [8] and given in Table 1 in the web-supplement) consist of 56 trials with patients numbers and person-times in weeks. They include rare events, including zero counts which frequently occur. The studies with zero events in one arm are called single-zero studies. If zero events occur in both arms, they are called double-zero studies. Unfortunately, in this situation some or many study specific risk ratio estimates become undefined as its associated variance. The problem is often addressed in practice by adding a continuity correction of 0.5 [4,18,26]. However, the use of a continuity correction in meta-analysis with many zero events can introduce bias in estimation and the normal approximation is inappropriate in studies with few events [14,24], and the use of the \( Q \)-statistic (on the basis of the remaining defined risk ratio estimates) becomes questionable.

Given these limitations, it is therefore important to develop a statistical approach that takes into account the characteristics of the rare events cases. The work in Böhning et al. [8] gives an overview on modelling approaches including fixed and random effects models.
which are appropriate for count outcomes and do not involve continuity corrections. Also, papers by Stijnen et al. [24] and by Hamza et al. [12] provide appropriate exact likelihood methods, which do not need a continuity correction, can be used for studies with rare events. A paper related to few studies was proposed by Spittal et al. [23]. They introduce Poisson regression meta-analysis to estimate the incidence risk ratio in the presence of zero events. Their approach as well as the works mentioned above rely on specific forms of modelling the random effects distribution whereas we focus here on a test for the variance of the random effects distribution being zero without making specific assumptions on its distribution.

Hence, in the work presented here, we are interested in addressing the issue of testing homogeneity in meta-analysis with zero-event studies. We will show that the conventional $Q$-statistic performs poorly, even if it can be defined at all. Alternatively, we suggest two novel approaches.

(i) The first approach is based on a conditional likelihood of a Poisson variable. The test using this method is a simple semi-parametric statistic, semi-parametric in the sense that it does not require any assumption of the random effects distribution which is generating the potential heterogeneity. The (weak) parametric assumption is that each study count is assumed to be Poisson with a study-specific parameter.

(ii) The second approach is based on generalized linear mixed models, in particular the conditional logistic regression with a normal random effect. It allows a likelihood ratio test for homogeneity of the risk ratio. This approach is mentioned in Böhning et al. [8] briefly but now fully developed here. The benefit of this approach lies in the fact that eliminates the nuisance intercept parameter and only involves the risk ratio as the parameter of interest.

The remainder of the paper is organized as follows. In Sect. 2, the construction of the proposed statistics is derived, and the distribution of the tests is explained. Section 3 presents a number of different examples from medical meta-analysis, and example of computation for logistic regression is given. In Sect. 4, we investigate the performance of the tests using simulations. Both common and rare events are considered across multiple cases. Sections 5 and 6 present our conclusions and investigation of the asymptotic distribution of the test.

2 Heterogeneity tests

Consider a meta-analysis of $k$ independent studies in which a Poisson distributed variable $X_{ij}$ denotes the number of events for study $i$ and treatment $j$, where $i = 1, 2, \ldots, k$, and $j = 1$ identifies the treatment arm and $j = 0$ the comparison arm. The mean and variance of $X_{ij}$ are given by $\mu_{ij} P_{ij}$, where $\mu_{ij}$ is the true incidence rate for group $j$ in study $i$ and $P_{ij}$ is the person-time for study $i$ in arm $j$, which is non-random. The estimated incidence rate is given as $\hat{\mu}_{ij} = X_{ij} / P_{ij}$. It follows that the estimated variance of $X_{ij}$ is given by $X_{ij}$. For each trial $i$, the true risk ratio is $RR_i = \mu_{i1} / \mu_{i0}$ with its estimate $\hat{RR}_i = X_{i1} P_{i0} / (X_{i0} P_{i1})$. The variance of $\hat{RR}_i$ is generally computed on the log-scale of risk ratio. Using the $\delta-$method, we have $Var(\log \hat{RR}_i) \simeq 1/X_{i1} + 1/X_{i0}$. To estimate the risk ratio under homogeneity $(RR_1 = \cdots = RR_k)$, the Mantel-Haenszel (MH) estimator [29] is then used and given by

$$\hat{RR}_{MH} = \frac{\sum_{i=1}^{k} X_{i1} P_{i0} / P_i}{\sum_{i=1}^{k} X_{i0} P_{i1} / P_i},$$

where $P_i = P_{i1} + P_{i0}$. This is usually applied in Cochran’s $Q$ statistic for testing homogeneity of the risk ratio, reflecting the hypotheses, $H_0 : RR_1 = RR_2 = \cdots = RR_k = RR$ vs $H_1 : \cdots$
not $H_0$. The statistic is given as follows:

$$\chi^2_{HOM} = \sum_{i=1}^{k} \frac{(\log \widehat{RR}_i - \log \widehat{RR}_{MH})^2}{1/X_{i1} + 1/X_{i0}},$$

which is considered to have an approximate chi-square distribution with $k - 1$ degrees of freedom under $H_0$. This is based on the argument that the Poisson distribution converges with increasing parameter $\mu_{ij} P_{ij}$ to a normal distribution and seems to be realistic if $X_{ij}$ is large as it is estimating $\mu_{ij} P_{ij}$. However, we have seen that $\chi^2_{HOM}$ cannot even be calculated if at least one observed $X_{i1}$ or $X_{i0}$ event is zero. A simple way of using this method is to remove studies that have zero events. However, as pointed out in Böhning et al. [8],

... the available test of homogeneity is of unknown behaviour even if infeasible study-specific effect estimates are omitted ...

as asking the degrees of freedom based on the remaining studies after removing zero studies is a random quantity itself. In the following section, an approach based on a conditional likelihood is presented.

### 2.1 The proposed chi-square test

In the following, we will make use of the well known fact that for two independent Poisson variates $V$ and $W$ with means $\mu_V$ and $\mu_W$, respectively, the conditional distribution of $V$ given the sum $V + W = v + w$ is binomial with event parameter $\mu_V / (\mu_V + \mu_W)$ and sample size parameter $v + w$. This is now applied to our situation. Suppose that $X_{i0}$ and $X_{i1}$ are Poisson variables with means $\mu_{i0} P_{i0}$ and $\mu_{i1} P_{i1}$, respectively, and are independent. Thus, $X_i = X_{i0} + X_{i1}$ is the total number of events, which is a Poisson distribution with mean $\mu_{i0} P_{i0} + \mu_{i1} P_{i1}$. Consider the random variable $X_{i1}$ conditional on $X_i = x_i$, where $x_i$ is an observed value. The conditional probability density function is given below

$$P(X_{i1} = x_{i1}|X_i = x_i) = \binom{x_i}{x_{i1}} \left( \frac{\mu_{i1} P_{i1}}{\mu_{i0} P_{i0} + \mu_{i1} P_{i1}} \right)^{x_{i1}} \left( \frac{\mu_{i0} P_{i0}}{\mu_{i0} P_{i0} + \mu_{i1} P_{i1}} \right)^{x_{i0}},$$

where $x_{i1} = 0, 1, 2, \ldots, x_i$. Here, $X_{i1}|X_i$ is a binomial distribution with size parameter $X_i$ and probability event parameter

$$q_i = \frac{\mu_{i1} P_{i1}}{\mu_{i0} P_{i0} + \mu_{i1} P_{i1}} = \frac{RR_i P_{i1}}{1 + RR_i P_{i1} P_{i0}},$$

where $RR_i = \mu_{i1} / \mu_{i0}$ as defined previously. Under $H_0 : RR_1 = RR_2 = \cdots = RR_k = RR$, $q_i$ can be rewritten into the form:

$$q_i = \frac{RR P_{i1}}{1 + RR P_{i1} P_{i0}},$$

which depends only on the true risk ratio $RR$. Note that $q_i$ can easily estimated by replacing $RR$ by its Mantel-Haenszel estimator $RR_{MH}$. The conditional probability based on the binomial distribution together with $q_i$ in (4) is now used to derive the new chi-square statistic
for test of homogeneity. It is given by

$$\chi^2_{pr} = \sum_{i=1}^{k} \frac{(X_{i1} - X_i \hat{q}_i)^2}{X_i \hat{q}_i (1 - \hat{q}_i)},$$

(5)

where \(\hat{q}_i = (\hat{RR}_{MH} P_{i1} / P_{i0})/(1 + \hat{RR}_{MH} P_{i1} / P_{i0})\) is the estimator for \(q_i\). This new form of \(Q\)-statistic has an approximate chi-square distribution with \(k - 1\) degrees of freedom. This approximation underlies the same laws as the approximation of the binomial distribution by the normal. We point out that \(k\) is the number of studies excluding double-zero studies. In applications, the null hypothesis will be rejected if the observed value of \(\chi^2_{pr}\) is greater than \(\chi^2_{1-a,k-1}\), where \(\chi^2_{1-a,k-1}\) is the \((1-a)\)th quantile of the chi-square distribution with \(k - 1\) degrees of freedom. Note that the approximation to the chi-square might be not good although this statistic is always defined. Note that this new definition of a \(\chi^2\)-test on homogeneity would also allow an easy incorporation into heterogeneity measures such as Higgins’ \(I^2\) which would take the form in this setting \(I^2 = \frac{\chi^2_{pr} - (k-1)}{\chi^2_{pr}}\). For more details on Higgins’ \(I^2\) see Higgins and Thompson [13] and Borenstein et al. [5].

### 2.2 The proposed likelihood ratio test

Whereas the approach in the previous section is still in the 2-stage framework, we now make use in a more direct way of the available count data. Böhning et al. [8] pointed out that the functional expression of \(q_i\) can be modelled by a logistic regression. This is taken advantage of in our paper. Suppose that the log-risk ratio is defined by \(\beta_i = \log RR_{i1}\). From (4), \(q_i\) can be rewritten as \(q_i = \frac{\exp(\beta_i + \log(P_{i1} / P_{i0}))}{1 + \exp(\beta_i + \log(P_{i1} / P_{i0}))}\), so that \(1 - q_i = 1/(1 + \exp(\beta_i + \log(P_{i1} / P_{i0})))\).

As the expression \(q_i\) is similar to the standard logistic response function, simple logistic regression analysis can be applied. The logistic regression model is then given as

$$\log \left( \frac{q_i}{1 - q_i} \right) = \beta_i + \log(P_{i1} / P_{i0}),$$

(6)

and \(PM_0\) will denote the logistic regression model with homogeneous effect \(\beta_i = \beta\) for all \(i = 1, 2, \ldots, k\). Note that (6) is a model for the log-odds ratio, but still represents the original risk ratio in which we are interested by taking the exponential of \(\beta\).

It is pointed out here that the model (6) involves only the parameter of interest and includes no further nuisance parameters. This is in contrast to working with the joint Poisson distribution which would require inference on the baseline as well as on the effect parameter (see also Böhning et al. [8]) whereas here inference is reduced to a univariate problem.

The estimate of the unknown parameter \(\beta_i = \beta\) is performed using logistic regression of the binomial count \(X_{i1}\) in \(X_i\) with an offset \(\log(P_{i1} / P_{i0})\). Note that the offset is a covariate with known coefficient that is used in the model to estimate the response. Since \(PM_0\) is under the homogeneous treatment effect, it is denoted as the null model. If the effect estimates differ between trials, the study factor is a potential confounding factor. In that case, it is appropriate to consider \(\beta_i\) as a random effect and it is common to assume its distribution to be normal with mean \(\beta\) and variance \(\sigma^2_\beta\), denoted as \(\beta_i \sim N(\beta, \sigma^2_\beta)\). The random effects model under heterogeneity of the log-risk ratio is denoted as \(PM_1\) or the full model, and its related parameter is fitted using random effects logistic regression. Random effects models have been used for some time in linear regression (see for example Baltagi [2]) and generalized
linear mixed models [25], of which the random effects logistic regression is a special case, have also become more popular.

The likelihood ratio test (LRT) involves a test statistic constructed by the maximised log likelihoods under the null model, $LL_r$, and full model, $LL_f$. Under the general testing $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, where $\theta$ is a generic parameter and $\theta_0$ is a constant, the LRT is defined as $LRT = -2(LL_r - LL_f)$. Under the null model, the distribution of this statistic converges to a chi-square distribution with $\nu$ degrees of freedom, where $\nu$ is the difference between the number of parameters involved in the associated models. Now we have two models, $PM_0$ and $PM_1$, related to the risk ratio. The proposed LRT, used to test $H_0 : \sigma^2 = 0$ and $H_1 : \sigma^2 > 0$, is therefore given by

$$LRT_{pr} = -2(LL_0 - LL_1),$$

(7)

where $LL_0$ and $LL_1$ are the log-likelihoods of $PM_0$ and $PM_1$, respectively. We need to elaborate on this statistics and its distribution. Since in the heterogeneity case $\beta_i$ is a random variable, heterogeneity is assessed by the variance $\sigma^2$. $H_0 : \sigma^2 = 0$ therefore means that there is no variation of the risk ratios between studies, which is equivalent to $H_0 : RR_1 = RR_2 = \cdots = RR_k$. Furthermore, it is important to note that the asymptotic null-distribution of the $LRT_{pr}$ is given by $0.5\chi^2_0 + 0.5\chi^2_1$, a mixture of a one-point mass at zero and a chi-square distribution with one degree of freedom. This is because in this testing $H_0$ (related to $PM_0$) is on the boundary of $H_1$ (related to $PM_1$). The distribution of the test, when testing $H_0 : \sigma^2 = 0$ and $H_1 : \sigma^2 > 0$, has also been discussed elsewhere, for example Böhning et al. [8], Self and Liang [21], Feng and McCulloch [11] or Baey et al. [1] for a more general result. In this case, the theoretical cumulative distribution function (CDF) of the proposed likelihood ratio statistic is derived as follows:

$$P(LRT_{pr} \leq x) = \frac{1}{2} + \frac{1}{2}P(\chi^2_1 \leq x) = 1 - \alpha,$$

(8)

where $x > 0$ and $\alpha$ is the significance level. From the last two terms, we obtain $P(\chi^2_1 \leq x) = 1 - 2\alpha$. In practice, $H_0$ will be rejected if the observed $LRT_{pr}$ is greater than $\chi^2_{1-2\alpha,1}$. Finally, we emphasize again that the considerations on the asymptotic distribution of the likelihood ratio will require large numbers of events which might be critically in our case and need to be investigated in the following.

### 3 Empirical illustrations

The proposed methods were investigated using three real datasets from meta-analytic studies in medical research. The cases involved both common and rare events. The data for all examples discussed in this section are available in the web supplemental material.

#### 3.1 Catheter related bloodstream infection

A meta-analysis on the effect of anti-infective-treated central venous catheters on catheter related bloodstream infection (CRBSI) was obtained from Niël-Weise et al. [17]. The data comprised nine clinical trials comparing the risk of CRBSI in patients with an anti-infective-treated catheter (treatment) and a standard catheter. In this dataset, no trial had zero events. Since patients participated for different observation times, the person-times of central venous catheterisation in days were also given. In the analysis, it was found that the MH risk ratio

\[ \text{MH risk ratio} \]
Table 1  Observed test statistic of homogeneity test for the risk ratio and probability value for the data examples

| Dataset/Method                | Observed test statistic | p-value |
|------------------------------|-------------------------|---------|
| **CRBSI**                    |                         |         |
| Conventional chi-square test | 9.8652                  | 0.2746  |
| Proposed chi-square test     | 10.7747                 | 0.2148  |
| Proposed likelihood ratio test | 0.7056                 | 0.4009  |
| \(-2LL_0\)                  | 36.1388                 |         |
| \(-2LL_1\)                  | 35.4332                 |         |
| **Myocardial infarction**    |                         |         |
| Conventional chi-square test | 16.5896                 | 0.9996  |
| Proposed chi-square test     | 20.2701                 | 0.9960  |
| Proposed likelihood ratio test | 0                     | 1.0000  |
| \(-2LL_0\)                  | 74.9720                 |         |
| \(-2LL_1\)                  | 74.9720                 |         |
| **Cardiovascular mortality** |                         |         |
| Conventional chi-square test | 8.1545                  | 0.9997  |
| Proposed chi-square test     | 8.9231                  | 0.9993  |
| Proposed likelihood ratio test | 0                     | 1.0000  |
| \(-2LL_0\)                  | 45.4570                 |         |
| \(-2LL_1\)                  | 45.4570                 |         |

estimate was 0.6602. The estimated risk ratios using logistic regression with fixed and random effects models were 0.6586 and 0.6218, respectively. This suggests the patients with an anti-infective-treated catheter had lower risk of CRBSI. The variance between the risk ratios was \(\hat{\sigma}^2 = 0.1183\). The \(I^2\) statistic was computed as 17.30%. This leaves the variability of the effect size unclear, making it difficult to decide which model, fixed or random effects, should be used. To make the choice clear, we calculated the statistics \(\chi^2_{HOM}\), \(\chi^2_{pr}\), and \(LRT_{pr}\). In logistic regression analysis, the functions \texttt{glm}, and \texttt{glmer} contained in the \texttt{lme4} package of \texttt{R} were used to fit the models [3]. The formulas were

\[
\text{glm(cbind(x1, xi-x1) } \sim 1 + \text{offset(log(p1/p0)), family = binomial("logit"))}
\]

for the fixed effect model \(PM_0\) and

\[
\text{glmer(cbind(x1, xi-x1) } \sim 1 + \text{offset(log(p1/p0)) + (1|study)},
\text{ family = binomial("logit"))}
\]

for the random effects model \(PM_1\). In this example, we reject \(H_0: RR_1 = RR_2 = \cdots = RR_k\), if the p-value for \(\chi^2_{HOM}\) is less than \(\alpha = 0.05\). \(H_0: \sigma^2 = 0\) is rejected if the p-value from \(LRT_{pr}\) is less than \(2\alpha\). The empirical test statistics and their p-values are presented in Table 1. The results of this meta-analysis showed that the risk ratios of CRBSI did not differ significantly across studies at the 0.05-level. The proposed chi-square test produced the smallest p-value, and the likelihood ratio test the largest.
3.2 Myocardial infarction and cardiovascular mortality

This application used a meta-analytic dataset on myocardial infarction (MI) events and cardiovascular (CV) deaths in the Rosiglitazone and control arms obtained from Böhning et al. [8], where it was originally published by Nissen and Wolski [16]. The data of 56 studies presented the number of events and the person-times in weeks. For MI dataset, it included a large number of rare events, with 46% single-zero and 27% double-zero trials. The MH estimated risk ratio was 1.2782 based upon all studies. Note that this is similar to the estimate calculated by Böhning et al. [8]. Logistic regression with fixed and random effects models produced the same estimated risk ratio of 1.2788 with the variance between the risk ratios $\hat{\sigma}^2_\beta = 0$. As a test of homogeneity, the results are shown in Table 1. At the 0.05 significance level, $H_0$ was not rejected. This confirmed that no significant difference in risk ratios existed between studies for MI. Furthermore, the $I^2$ estimate was zero. Hence, as the 95% confidence interval for the risk ratio is (1.0116, 1.6135), a significant effect homogeneous across trials can be established. The results showed that no significant difference in risk ratios existed between studies for CV.

4 Simulation studies

4.1 Simulation settings

Simulations were undertaken to study the performance of the homogeneity tests for the risk ratio. Two sets of meta-analytic data were considered under common events and rare events. The scope of the simulations was given as follows. For the common events cases, the number of events $X_{i1}$ and $X_{i0}$ were generated from $Po(\mu_{1}P_{i1})$ and $Po(\mu_{0}P_{i0})$, respectively. The incidence rates $(\mu_{1}, \mu_{0})$ were set at (0.1,0.2), (0.4,0.4), and (0.6,0.3) for the population risk ratios $\theta = \mu_{1}/\mu_{0} = 0.5$, 1.0, and, 2.0, respectively. The person-times $P_{i0}$ and $P_{i1}$ were generated from $Po(\eta_i)$, where the mean $\eta_i$ was sampled from $N(100, 1)$, for a balanced design. For the rare events cases, the number of events $X_{i1}$ and $X_{i0}$, and the person-times $P_{i0}$ and $P_{i1}$ were generated from Poisson distributions with the same parameter values as in the previous settings. The event occurrence probabilities were given by $(\mu_{1}, \mu_{0}) = (0.01, 0.02), (0.02,0.02), \text{ and } (0.02,0.01)$ for a small number of events.

In all cases, the number of studies were set as $k = 20, 30, \text{ and } 60$, and the significance levels at $\alpha = 0.01, 0.05, \text{ and } 0.10$. Using R [19], the simulation runs were repeated 10,000 times for each case. If double-zero (DZ) trials occurred they were removed before computing the chi-square tests and the degrees of freedom were adjusted to $k - 1 - d$, where $d$ is the number of DZ studies. If single-zero (SZ) trials occurred, the values were adjusted using a smoothing constant of 0.5. The new likelihood ratio test was estimated using the full dataset based on logistic regression of $X_{i1}$ conditional on $X_{i0} + X_{i1}$.

The performance of the test was evaluated using type I error probability and power of the test. The estimate of type I error is given by $n(\text{reject } H_0|H_0)/10,000$, where $n(\text{reject } H_0|H_0)$ refers to number that the test statistic falls within the critical region if $H_0$ is true. To investigate the power of the test, a number $c_i$ was generated from uniform distributions on (1,2) and (1,3) for common events, and on (1,10) and (1,20) for rare events, reflecting small to large deviations. When used to multiply $\mu_{1}$, this yields $RR_i = c_i\mu_{1}/\mu_{0} = c_i\theta$, so that $H_0$ becomes wrong. In a second approach, we generated a random number $\beta_i$ from $N(\beta, \sigma^2_\beta)$ where $\beta = \log \theta$, the true log-relative risk. It can be seen that the two different simulated
datasets were under the alternative hypothesis, but were investigated under a misspecified alternative for the former approach and a correctly specified alternative for the latter. A second criterion, the average power of the test is determined as \( n(\text{reject } H_0|H_1)/10,000 \). In conclusion, the test statistic that has an appropriate type I error rate with more power of the test is more efficient than the other.

### 4.2 Simulation results

For the common events setting the estimated type I error rates for the proposed tests \((\chi^2_{pr} \text{ and } LRT_{pr})\) and the conventional chi-square test \((\chi^2_{HOM})\) are shown in Table 4 of the web supplement. It was found that \(\chi^2_{pr}\) produced type I errors very close to the target significance level in all cases. The type I error of \(LRT_{pr}\) was close to the significance level when number of studies was greater than 30. Power of all tests when simulated data under the alternative hypothesis using a normal distribution was greater than 0.95. This contrasted with the power of the tests when simulated under the alternative using a uniform distribution given in Table 6 of the supplemental material. In the latter case, the tests showed low power for \(\theta = 0.5\) in \(c_i\theta\) and \(c_i\) ranging uniformly in the small interval although the newly proposed tests showed in all cases larger power than the conventional chi-square test. Therefore, these results suggest that the misspecified alternative has little effect on the relative power performance of these tests in the common events setting. We conclude that the proposed chi-square test is recommended for homogeneity testing in meta-analysis with common events, given its appropriate type I error rate and good test power.

In the rare cases setting, the number of events was small, and some studies had zero values. When SZ or DZ trials were presented, \(LRT_{pr}\) needed no requiring a continuity correction. However, the zero count value needed adjustment for \(\chi^2_{HOM}\) and this was also applied to \(\chi^2_{pr}\) to ensure comparability. Table 5 of the web supplement presents the simulation results for the homogeneity tests. It was found that the values of \(\chi^2_{HOM}\) were very low, while those of \(\chi^2_{pr}\) and \(LRT_{pr}\) in the rare events case did not differ much from those computed in the common events. Only type I errors of \(\chi^2_{pr}\) and \(LRT_{pr}\) were greater than zero, where those of the latter were greater in general. \(LRT_{pr}\) had the highest test power in all cases, while \(\chi^2_{pr}\) performed better than \(\chi^2_{HOM}\). Table 7 in the web-supplement shows the power performance of the three tests under a uniform distribution as alternative. Both new tests perform consistently better than \(\chi^2_{HOM}\). However, in this case \(\chi^2_{pr}\) has consistently larger power than \(LRT_{pr}\). Given the fact that \(\chi^2_{pr}\) and \(LRT_{pr}\) performed similarly in the normal alternative (with slight benefits to the \(LRT_{pr}\)) it appears that \(\chi^2_{pr}\) is a reasonable choice of testing homogeneity.

Evidently, the performance of \(\chi^2_{HOM}\) was unsatisfactory in terms of both type I errors and power of the test. These results suggest that the use of the conventional chi-square statistic should be avoided entirely in the presence of rare events, and applied with caution in the non-rare event situation. When meta-analysis is concerned with rare-event data, a 1-stage approach from the likelihood ratio statistic is possible and recommended. As the results here indicate that using asymptotic distributions for the newly proposed tests might be inappropriate for rare events settings, we suggest to base inference on a bootstrap approach which we outline in the next section.

We have also explored the behavior of the tests for common and rare events in terms of the distribution, using simulation. Figure 1 compares the plots of empirical cumulative
distribution functions (CDFs) of the test statistics with the theoretical CDFs. For common events (upper panel), all test statistics approach the asymptotic reference distribution. For rare events (lower panel of Fig. 1), the empirical CDF of the conventional chi-square test obviously failed to track the theoretical, asymptotic distribution. In contrast, a closer match was found between the empirical and the theoretical distributions for the two proposed tests. This underlines that these novel statistics perform well. Figure 2 illustrates the probability-probability (P-P) of the test statistics. From the P-P plot for rare events (lower panel of Fig. 2), the empirical distribution of the proposed likelihood ratio statistic lies close to the reference line of the distribution $0.5\chi_0^2 + 0.5\chi_1^2$. 

\[ \chi_{HOM}^2, \chi_{pr}^2, \text{and } LRT_{pr} \]
5 Bootstrap p-value

The simulation study for the rare events situation showed that also for the newly proposed test estimated type I errors are below the significance level. To address this shortcoming and improve the power of the test, we suggest to consider a bootstrap correction and outline this here for the p-value. Bootstrap tests are used now for some time (see for example Sinha [22]) and our proposed procedure follows this tradition. In more detail, the bootstrap algorithm below is used to construct an estimate of the bootstrap p-value, relating to the true underlying null-distribution. However, it can also be used to estimate the underlying null-distribution itself. The procedure of parametric bootstrap is given as follows.

Algorithm 1

1. Draw a sample of size \(k\) with replacement from the original sample of \(k\) studies (excluding the DZ studies) leading to \(r^*_i = P_i / P_0\) and \(x^*_i = x^*_i + x^*_i\), for \(i = 1, 2, \ldots, k\).
2. Compute \(q^*_i = r^*_i \hat{RR}_{MH} / (1 + r^*_i \hat{RR}_{MH})\).
3. Sample \(X^*_i\) from a binomial distribution with size \(x^*_i\) and event parameter \(q^*_i\).
4. Compute \(\hat{RR}^*_i = r^*_i \hat{RR}^*_{MH} / (1 + r^*_i \hat{RR}^*_{MH})\).
5. Compute \(q^{**}_i = r^*_i \hat{RR}^{**}_{MH} / (1 + r^*_i \hat{RR}^{**}_{MH})\).
6. Compute a bootstrap sample

\[
Q^* = \sum_{i=1}^{k} \frac{(X^*_i - x^*_i \hat{q}^{**}_i)^2}{x^*_i \hat{q}^{**}_i (1 - \hat{q}^{**}_i)}.
\]

The procedure was repeated \(B = 50,000\) times, so that a bootstrap sample \(Q^*_1, Q^*_2, \ldots, Q^*_B\) is obtained. We then computed the bootstrap p-value by comparing \(Q^*_b\), for \(b = 1, 2, \ldots, B\), to the observed chi-square statistic for measuring heterogeneity

\[
Q = \sum_{i=1}^{k} \frac{(X_i - x_i \hat{q}_i)^2}{x_i \hat{q}_i (1 - \hat{q}_i)}.
\]

The bootstrap p-value is given as \(n(Q^*_b | Q^*_b \geq Q) / B\), \(n(x)\) is defined to be the number of times \(x\) is true.

We applied these concepts to the rare event data examples: myocardial infarction, cardiovascular mortality, catheter related bloodstream infection and an additional data set on perinatal death (the number of deaths induced by routine and selective induction of pregnancies that go beyond term obtained from Crowley [10]). We have added the data set on perinatal deaths to provide a most extreme example of rare event meta-analytic data. All these data are given in the supplemental material.

For the perinatal mortality meta-analysis, we experienced the problem that a full set of \(X^*_i\) generated in step 3 of Algorithm 1 was entirely consisting out of zero counts. Hence the recalculation of \(\hat{RR}^*_{MH}\) was not feasible. In this case, we used a modification of Algorithm 1 as follows.

Algorithm 2

1. Draw a sample of size \(k\) with replacement from the original sample of \(k\) studies (excluding the DZ studies) leading to \(r^*_i = P_i / P_0\) and \(x^*_i = x^*_i + x^*_i\), for \(i = 1, 2, \ldots, k\).
2. Compute \(q^*_i = r^*_i \hat{RR}_{MH} / (1 + r^*_i \hat{RR}_{MH})\).
3. Sample \(X^*_i\) from a binomial distribution with size \(x^*_i\) and event parameter \(q^*_i\).
The $p$-values for both bootstrap algorithms as well as the $p$-value using the chi-square approximation are given in Table 2. From the rare event data, the observed values of the chi-square statistic were different from those of the mean of all bootstrap $Q$ statistics, as the former statistic was often much lower. It can be concluded that the approximate chi-square statistic averted from the null distribution for the rare event cases. It is interesting to note that mean of the bootstrapped $Q$ statistics was close to what would have been expected under a chi-square distribution. Note that Algorithm 1 gets close to the $k - 1$ expectation in the mean as it mimics estimating the risk ratio. As Algorithm 2 is unable to mimic this we see that the expectation is closer to $k$.

Furthermore, bootstrap $p$-values were different from $p$-values using the approximate $X_{k-1}^2$ distribution in all examples. So, the bootstrap value should be used. However, the chi-square statistic can be used in the common event data as the simulation showed. Note that, since the example on perinatal death had many zero studies on the treatment arm leading to all zero samples on $X_{i1}^*$ and $\hat{RR}_{MH}^*$ cannot be computed, only Algorithm 2 was used. In all other cases, the bootstrap $p$-values obtained from these two algorithms were not substantially different.

### 6 Discussion and conclusions

In this paper, we have proposed two statistics for testing of homogeneity as alternatives to the conventional $Q$-statistic which entirely collapses in meta-analysis of rare event studies. Both use information from the distribution of the number of events in an intervention group $X_{i1}$ given the total number of events $X_i$. The advantages of both tests are that they can be applied even if the trials have SZ arms, and that it requires no specialized software in computation. However, when confronted with a DZ event we need to either exclude the trial or add a continuity correction, as in the formula the total number $X_i$ is not allowed to be zero. This places a limitation on the proposed tests as DZ-studies are excluded. However, it is shown in Böhning and Sangnawakij [6] that this is no loss of information with respect to the risk ratio. The difference between the two proposed test statistics lies in the assumption of normality of the random effects distribution whereas this is not needed for the former. Future research will further investigate the impact of this assumption.
The performance of the proposed tests was compared with that of the conventional chi-square test using simulations. For common events, our chi-square test performed overwhelmingly better than the traditional test, producing appropriate type I error rates and demonstrating very high test power in all cases. The performance of the proposed likelihood ratio test was similar to that of the traditional chi-square test. Both tests had type I error lower than the nominal significance level, while retaining satisfactory test power. In the case of rare events, our likelihood ratio test was demonstrated to outperform the competing tests. The conventional chi-square test had very low power. It is therefore not recommended for use in such cases. The newly proposed chi-square test can be always formulated in the rare events setting, but does not reach the significance level. As an alternative method, the bootstrap is recommended for these cases. If there remain concerns of the validity of the approximation for a given rare events data set, we recommend to use the suggested bootstrap approximation of the true null distribution. From these results, and the output of the simulations, we conclude that our tests for homogeneity can be applied in meta-analyses based on count outcomes with rare events.

Finally, we consider the meta-analysis of studies with DZ trials and end the discussion with a cautionary note. When a continuity correction of 0.5 per arm is applied (which we do not recommend but let us assume we want to do this), the study-specific risk ratio is given by $\hat{RR}_i = (0.5/P_{i1})/(0.5/P_{i0}) = P_{i0}/P_{i1}$, which is dependent on the person-times in the arms only. This seems an unreasonable as an estimate of the risk ratio. A question arises how to choose the continuity correction in this case. We therefore introduce the following idea. Let $c_{i1}$ and $c_{i0}$ be the smoothing constants of study $i$ in the treatment and control arms, respectively, with $c_{i1} + c_{i0} = 1$. Suppose that $\hat{RR}_i = 1$ in a DZ trial, as there is no evidence for either side. To satisfy this condition, it can be written as $1 = (c_{i1}/P_{i1})/(c_{i0}/P_{i0})$ or $(1 - c_{i0})/P_{i1} = c_{i0}/P_{i0}$. Thus, the smoothing constant according to the weight of the person-time is given as $c_{i0} = P_{i0}/P_i$ for the control arm, where $P_i = P_{i1} + P_{i0}$. Clearly, $c_i = 0.5$ is applied if the trial is balanced. We suggest this only for settings where the inclusion of DZ trials is required, or otherwise wished to be included. From our perspective the exclusion on inclusion of DZ trials, in contrast to SZ trials, will not change the evidence neither for the effect nor for its heterogeneity.

Acknowledgements All authors would like to thank two anonymous referees for their helpful comments. This work has been partially funded under Grant HO1286/16-1 by the German Research Foundation (DFG).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Baey, C., Cournéde, P.-H., Kuhn, E.: Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. Comput. Stat. Data Anal. 135, 107–122 (2019)
2. Baltagi, B.H.: Econometric Analysis of Panel Data. Wiley, New York (2013)
3. Bates, D., Mächler, M., Bolker, B.M., Walker, S.C.: Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48 (2015)
4. Bhaumik, D.K., Amaty, A., Normand, S.L., Greenhouse, J., Kaizar, E., Neelon, B., Gibbons, R.D.: Meta-analysis of rare binary adverse event data. J. Am. Stat. Assoc. 107, 555–567 (2012)
5. Borenstein, M., Hedges, L.V., Higgins, J.P., Rothstein, H.R.: Introduction to Meta-Analysis. John Wiley & Sons, Chichester (2009)
6. Böhning, D., Sangnawakij, P.: The identity of two meta-analytic likelihoods and the ignorability of double-zero studies. Biostatistics. (2020). https://doi.org/10.1093/biostatistics/kxaa004
7. Böhning, D., Malzahn, U., Schlattmann, P., Danmann, U.-P., Mehnert, W., Holling, H., Schulze, R.: The application of statistical methods of meta-analysis for heterogeneity modelling in medicine and pharmacy, psychology, quality control and assurance. In: Jäger, W., Krebs, H.J. (eds.) Mathematics - Key Technology for the Future, pp. 533–553. Springer, Heidelberg (2003)
8. Böhning, D., Mylona, K., Kimber, A.: Meta-analysis of clinical trials with rare events. Biom. J. 57, 633–648 (2015)
9. Cochran, G.W.: The combination of estimates from different experiments. Biometrics. 10, 101–129 (1954)
10. Crowley, P.: Interventions for preventing or improving the outcome of delivery at or beyond term. Cochrane Database Syst. Rev. (2000). doi: 10.1002/14651858.CD000170
11. Feng, Z., McCulloch, C.E.: Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. Stat. Probabil. Lett. 13, 325–332 (1992)
12. Hamza, T.H., van Houwelingen, H.C., Stijnen, T.: The binomial distribution of meta-analysis was preferred to model within-study variability. J. Clin. Epidemiol. 61, 41–51 (2008)
13. Higgins, J.P.T., Thompson, S.G.: Quantifying heterogeneity in a meta-analysis. Stat Med. 21, 1539–1558 (2002)
14. Jackson, D., White, I.R.: When should meta-analysis avoid making hidden normality assumptions?. Biom. J. 60, 1040–1058 (2018)
15. Kulinskaya, E., Dollinger, M.B., Bjørkestol, K.: Testing for homogeneity in meta-analysis I The one-parameter case: standardized mean difference. Biometrics. 67, 203–212 (2011)
16. Nissen, S.E., Wolski, K.: Rosiglitazone revisited: An updated meta-analysis of risk for myocardial infarction and cardiovascular mortality. Arch. Intern. Med. 170, 1191–1201 (2010)
17. Niel-Weise, B.S., Stijnen, T., van den Broek, P.J.: Anti-infective-treated central venous catheters for total parenteral nutrition or chemotherapy: A systematic review. J. Hosp. Infect. 69, 114–123 (2008)
18. Piaget-Rossel, R., Taffé, P.: Meta-analysis of rare events under the assumption of a homogeneous treatment effect. Biom. J. 61, 1557–1574 (2019)
19. R Core Team: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019). https://www.R-project.org/
20. Schulze, R., Holling, H., Böhning, D.: Meta-Analysis: New Developments and Applications in Medical and Social Sciences. Hogrefe and Huber Publishing, Massachusetts (2003)
21. Self, S.G., Liang, K.Y.: Large sample properties of the maximum likelihood estimator and the likelihood ratio test on the boundary of the parameter space. J. Am. Stat. Assoc. 82, 605–611 (1987)
22. Sinha, S.K.: Bootstrap tests for variance components in generalized linear mixed models. Can. J. Stat. 37, 219–234 (2009)
23. Sánchez-Meca, J., Marán-Martánez, F.: Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and type I error. Qual. Quant. 31, 385–399 (1997)
24. Sánchez-Meca, J., Marán-Martánez, F.: Testing the significance of a common risk difference in meta-analysis. Comput. Stat. Data Anal. 33, 299–313 (2000)
25. Tarone, R.E.: On summary estimators of relative risk. J. Chronic Dis. 34, 463–468 (1981)
26. Viechtbauer, W.: Hypothesis tests for population heterogeneity in meta-analysis. The Br. J. Math. Stat. Psychol. 60, 29–60 (2007)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.