ARTICLE TEMPLATE

On the exact null-distribution of a test for homogeneity of the risk ratio in meta-analysis of studies with rare events

ARTICLE HISTORY

Compiled August 8, 2020

ABSTRACT

This paper focuses on the test for homogeneity of relative risk in meta-analysis of count outcomes. Meta-analysis of studies with rare events faces particular challenges, since the number of studies are low and the frequency of events may be small in some or all treatment arms. In such a case, the conventional chi-square test for homogeneity becomes undefined and we suggest a new chi-square test which is always defined. However, the chi-square approximation is poor. We therefore introduce methodology to obtain its exact distribution which is based on the product binomial likelihood. The exact p-value is then derived and the performance of the method is investigated using simulations. The results show that the type I error of the proposed method satisfies the nominal significance level in rare events situations. Also, the exact distribution behaves very similar to the simulated distribution. A real data example of a meta-analysis with an extreme form of rare event studies is used to illustrate the new test.

KEYWORDS

Conditional binomial model; count outcomes; exact p-value; null distribution; simulation study

1. Introduction

Meta-analysis is a statistical tool used to analyze and combine the results obtained from many individual, independent studies on the same research topic. The outcome in trials is often an event or a condition, and the studies are designed for comparing the occurrence of that event in two groups, an intervention and a control group. We use here the term *event* in a generic sense and it could mean a death, the occurrence of a complication following a surgery or alike, depending on the setting. In many trials, the count data on the event of interest is often rare. A *rare event* means that event occurrence is very low, so that frequently a small number or no observations of the event of interest are observed in a trial [1]. For example, improved anaesthesia safety has lead to quite rare events of anaesthesia-related incidents, complications, and deaths [2]. When there are no events in treatment or control group (called single-zero study) or both (called double-zero study), simple approaches often used are exclusion of studies or adding a continuity correction, typically 0.5, to all cells of the contingency table for studies with no events. The traditional meta-analysis, commonly referred to the inverse-variance-weighted average method, is then used to estimate the overall effect size parameter, for example the odds ratio (OR) or risk ratio (RR) [3,4]. However, meanwhile a diversity of research works suggest that exclusion of studies and adding continuity corrections can introduce bias in calculation of effect measures in meta-analysis, leading to low performance of the traditional approach. So, both methods should be avoided especially when sample sizes are severely unbalanced or small events are available in studies. This topic is widely discussed in recent literature, for example, Sweeting et al., Bradburn et al., Keus et al., Kuss, Efthimiou, and Jackson et al. [5–10]. The method which is robust to the occurrence of rare events and does not require any continuity correction, called Mantel-Haenszel (MH) [11], has become popular for estimating summary effect measures such as the relative risk or odds ratio. However, as heterogeneity of effect is frequently of interest, it is not clear how testing of homogeneity can be accomplished in the rare events situation. This is discussed in the next section.

1.1. Testing for homogeneity

Let us first outline the conventional approach for testing homogeneity of the risk ratio in meta-analysis. Suppose that X_{ij} is a Poisson random variable denoted as the number of events (deaths or complications, etc.) of study i and group j, for $i = 1, 2, ..., k$ and $j = 0, 1$. Here, k is the number of studies in the meta-analysis, $j = 1$ stands for an intervention and $j = 0$ for a comparison group. The mean and variance of X_{ij} are

$E(X_{ij}) = Var(X_{ij}) = \mu_{ij}T_{ij}$, where μ_{ij} is the incidence rate parameter (event risk) for study i and group j, and T_{ij} is the duration or person-time at risk for study i and group j. The latter is non-random. For each study, the true risk ratio is defined as $RR_i = \mu_{i1}/\mu_{i0}$ and estimated by $\widehat{RR}_i = \frac{X_{i1}/T_{i1}}{X_{i0}/T_{i0}}$. To determine the dispersion of the estimator, we consider the variance of the risk ratio on the log-scale measure, $\log \widehat{RR}_i$. Using the delta method and based on a normal approximation [12], the variance of $\log \widehat{RR}_i$ is given by $\widehat{Var}(\log \widehat{RR}_i) = 1/X_{i1} + 1/X_{i0}$, assuming that X_{i1} and X_{i0} are both positive.

For a meta-analysis of k independent studies, the overall true risk ratio is denoted as RR, the ratio of the event occurrence probability(risks) in the exposed or intervention group to the event probability in the non-exposed or comparison group. If the risk ratios are identical across all studies, homogeneity of effect is present. If there is variability of the risk ratio across studies which frequently occurs, we are in the situation of heterogeneity. To capture this more general situation we allow μ_{i1} and μ_{i0} to have specific distributions across studies, respectively. The true risk ratio is therefore given as $RR = \Delta_1/\Delta_0$, where Δ_1 and Δ_0 are the expected values of the distributions of μ_{i1} and μ_{i0}, respectively. It is important to know about the presence of heterogeneity in the risk ratio and its size as this will likely influence the choice of analysis such as using the fixed effect or random effects model.

A test for homogeneity examines the null hypothesis that all studies are evaluating the same effect [13,14]. For a test of homogeneity of the risk ratios in meta-analysis, the hypotheses are given by

$$H_0 : RR_1 = RR_2 = ... = RR_k = RR = \Delta_1/\Delta_0$$

$$H_1 : \text{ At least one of the studies has different risk ratio.}$$

Here, H_0 means that homogeneity of the risk ratio is present, while H_1 states that heterogeneity of effect has occurred. Note that H_0 neither implies that $\mu_{11} = \cdots = \mu_{k1}$ nor that $\mu_{10} = \cdots = \mu_{k0}$ is valid .

The most commonly used statistic related to these hypotheses is Cochran's Q, which consists of a weighted sum of the square deviations of the observed effect sizes from the overall meta-analytic estimate [15,16]. The test statistic is given by

$$Q_{Hom} = \sum_{i=1}^{k} \frac{(\log \widehat{RR}_i - \log \widehat{RR}_{MH})^2}{\widehat{Var}(\log \widehat{RR}_i)}, \tag{1}$$

where $\widehat{RR}_{MH} = \frac{\sum_{i=1}^{k} X_{i1}T_{i0}/(T_{i1}+T_{i0})}{\sum_{i=1}^{k} X_{i0}T_{i1}/(T_{i1}+T_{i0})}$ is the Mantel-Haenszel (MH) estimator for the summary risk ratio. Under H_0, given some regularity conditions (number of events become large), Q_{Hom} has an asymptotic chi-square distribution with $k-1$ degrees of freedom ($\chi^2_{df.=k-1}$). The p-value, the probability that Q_{Hom} is taking a value equal or larger than the one observed given the null hypothesis is true, is obtained on the basis of a $\chi^2_{df.=k-1}$-distribution.

1.2. Motivating data

This work is motivated by a meta-analysis on the perinatal death in post-term pregnancy of routine and selective inductions, conducted by Crowley [17]. Table 1 shows this dataset. It includes 19 studies, where 11 studies have no events (perinatal deaths) in both treatment arms of induction of labor, and seven studies have zero events in one arm. As noted in many works, small and medium sample sizes as well as rare events are frequently encountered in applications. In these situations, statistics based on normal approximation theory lack of robustness and/or efficiency. The use of the inverse-variance-weighted average method with rare events in many studies can lead to low performance estimator [18–21]. Furthermore and most importantly here, if either number of deaths in routine induction X_{i1} or selective induction X_{i0} is zero, in other words, if at least no event is observed in one arm, the study-specific measure $\widehat{RR}_i$ is *undefined*. Consequently, Q_{Hom} cannot be computed using the type of data as shown in Table 1. This is a crucial problem of meta-analysis of rare count data.

Table 1. Meta-analytic data on the perinatal death in post-term pregnancy for routine and selective inductions.

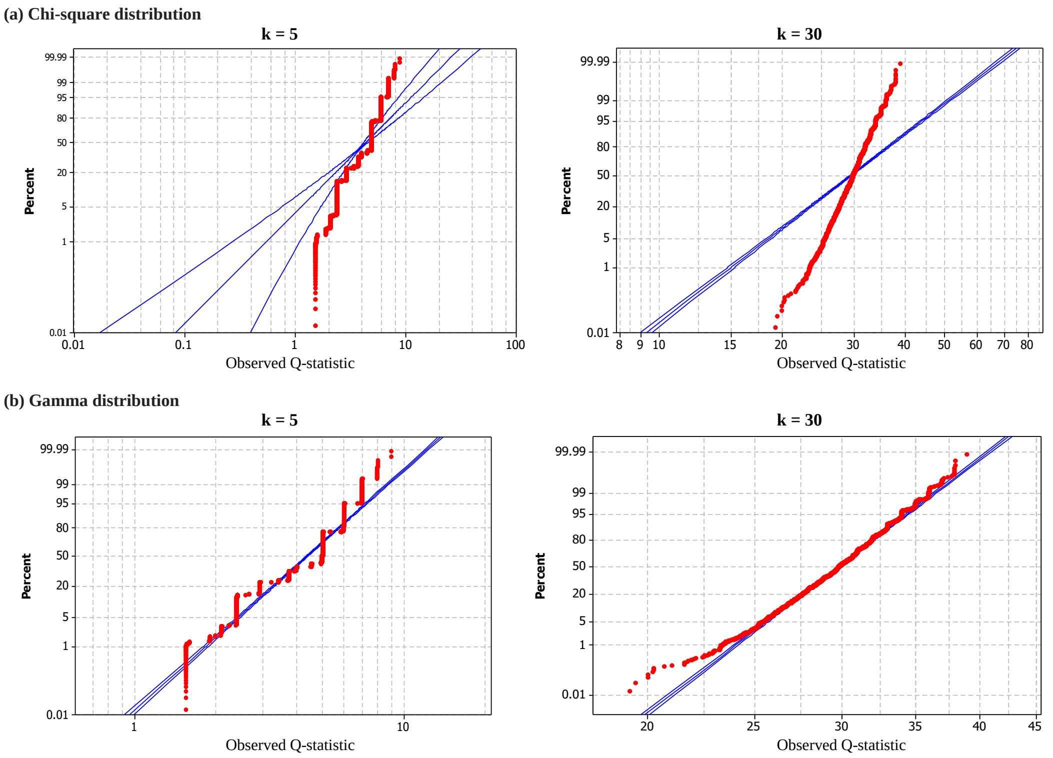
	Routine induction		Selective induction	
Author, year	Number of patients	Number of deaths	Number of patients	Number of deaths
Henry, 1969	57	2	55	0
Cole, 1975	119	0	118	0
Martin, 1978	134	1	131	0
Tylleskar, 1979	55	0	57	0
Breart, 1982	235	0	481	0
Katz, 1983	78	0	78	1
Suikkari, 1983	53	0	66	0
Sande, 1983	90	0	76	0
Cardozo, 1986	207	1	195	0
Augensen, 1987	195	0	214	0
Dyson, 1987	150	1	152	0
Witter, 1987	97	0	103	0
Bergsjo, 1989	94	1	94	0
Egarter, 1989	168	1	188	0
Martin, 1989	10	0	12	0
Heden, 1991	129	0	109	0
Hannah, 1992	1706	2	1701	0
Herabuyta, 1992	51	0	57	0
NICH, 1994	175	0	235	0

1.3. Conditional binomial model

In this section, we introduce an alternative statistic for testing homogeneity for the risk ratio. It is constructed using information on the conditional probability of the number of events in the treatment arm, X_{i1}, given the total number of events $X_i = X_{i1} + X_{i0}$. In this case, X_{i1} conditional on X_i is known to have a binomial distribution with size parameter X_i and probability event parameter $\pi_i = \frac{\mu_{i1}T_{i1}}{\mu_{i0}T_{i0}+\mu_{i1}T_{i1}}$, denoted as $X_{i1}|X_i \sim Bi(X_i, \pi_i)$. The mean and variance of $X_{i1}|X_i$ are $E(X_{i1}|X_i) = X_i\pi_i$ and $Var(X_{i1}|X_i) = X_i\pi_i(1-\pi_i)$, respectively. The likelihood function of $X_{i1}|X_i$ is given by $\prod_{i=1}^{k} \binom{x_i}{x_{i1}} \pi_i^{x_{i1}}(1-\pi_i)^{x_i - x_{i1}}$, where x denotes the observed value of X. Under H_0, homogeneity of risk ratios across trials, the test statistic constructed based on the conditional binomial model is therefore given as

$$Q = \sum_{i=1}^{k} \frac{(x_{i1} - x_i\hat{\pi}_i)^2}{x_i\hat{\pi}_i(1-\hat{\pi}_i)}, \tag{2}$$

where $\hat{\pi}_i = \frac{\widehat{RR}_{MH}(T_{i1}/T_{i0})}{1+\widehat{RR}_{MH}(T_{i1}/T_{i0})}$ is the estimator for π_i and k is the number of studies *excluding double-zero* studies. This simple non-parametric statistic is always defined and has an approximate chi-square distribution with $k-1$ degrees of freedom, and the p-value is given by $p_{\chi^2} = Pr(Q \geq q|H_0)$, where q is the observed value of Q.


(a) Chi-square distribution

(b) Gamma distribution

Figure 1. Probability plots of simulated Q for $k = 5$ and 30, $x_i = 1$ or 2, and $\pi_i = xyz$.

In practice, if the observed value of Q is greater than a critical value, or p_{χ^2} is less than the significance level, H_0 will be rejected. We note that this method is based on a one-step model which avoids the standard practice of estimating effects from the single studies and calculating meta estimators in a second step.

A small simulation to evaluate the performance of Q for rare events situations is shown in Figure 1. It shows the probability plot with respect to a chi-square distribution for the very rare events situation for $k = 5$ and $k = 30$ in the upper two panels. It is clear that the approximation is not satisfactory. Kulinskaya and Dollinger [?] suggest the use of a gamma distribution as approximating distribution. The lower two panels in Figure 1 show the probability plots of the best fitting gamma distribution. Clearly, approximations are substantially improved, but remain lacking fit.

When studies include few or zero events, methods based on a normal approximation can lead to unreliable statistical inference [10,23,24]. Indeed, to reach a satisfying approximation to the $\chi^2_{df.=k-1}$-distribution, X_i has to be *large* which is not the case in the rare events situation. Other parametric distributions, such as the mentioned gamma distribution suggested by Kulinskaya and Dollinger [?] might improve approximations, but, ultimately, the source of the problem of reaching a good approximation remains in the discrete nature of the underlying count data. In this paper, we will derive, for the rare events situation, the exact null-distribution of the homogeneity test statistic based on the one-step model.

The main contribution of this paper is to present an alternative approach to derive the exact distribution of the test statistic constructed using the likelihood of a conditional binomial model. The proposed method to find the exact distribution is based on considering all possible outcomes of cases on the total. Then, the exact p-value is derived. The details of these methods are given in Section 2. In Section 3, we use a real dataset in a meta-analysis of rare events to illustrate our approach. The performance of the test based on the proposed method is assessed by simulations, and compared with that of the bootstrap method and the asymptotic method based on a chi-square distribution. The investigation of the exact distribution and simulation results are given in Section 4. Comprising findings in all sections enable us to reach some important conclusions that we discuss in Section 5.

2. Exact distribution of Q

The exact conditional distribution is derived for the statistic given in equation (2) in the rare events situation. In this section, we define the required notations for clarity and convenience. The non-parametric statistic for test of homogeneity of the risk ratio in meta-analysis is then given by

$$Q = \sum_{i=1}^{k} \frac{(x_{i1} - x_i \pi_i)^2}{x_i \pi_i (1 - \pi_i)} \tag{3}$$

for given probability parameter π_i and total number of events from both treatment arms $x_i = x_{i1} + x_{i0}$. Since $X_{i1} \sim Bi(x_i, \pi_i)$, the probability mass function is given by $Pr(X_{i1} = x_{i1}) = \binom{x_i}{x_{i1}} \pi_i^{x_{i1}} (1 - \pi_i)^{x_i - x_{i1}}$. For all k studies, we need to find the joint likelihood. This can be accomplished by noting that the probability for observing

$\mathbf{X} = \mathbf{x} = (x_{11}, x_{21}, ..., x_{k1})$ is given as the product binomial probability:

$$Pr(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{k} \binom{x_i}{x_{i1}} \pi_i^{x_{i1}} (1 - \pi_i)^{x_i - x_{i1}},$$

where $x_{i1} = 0, 1, 2, ..., x_i$ and $0 \leq \pi_i \leq 1$. It follows that

$$Pr(Q = q) = f(q) = \sum_{\mathbf{x} \in \Omega_q} \prod_{i=1}^{k} \binom{x_i}{x_{i1}} \pi_i^{x_{i1}} (1 - \pi_i)^{x_i - x_{i1}}, \tag{4}$$

where the set $\Omega_q = \{\mathbf{x} | Q = \sum_{i=1}^{k} \frac{(x_{i1} - x_i \pi_i)^2}{x_i \pi_i (1 - \pi_i)} = q\}$, q is the observed value of Q, and $0 \leq f(q) \leq 1$. Therefore, $f(q)$ is denoted as the *exact distribution* of Q presented in equation (3).

To obtain the exact distribution and the p-value of Q, the proposed method uses all feasible vectors $\mathbf{x}$. Algorithms are provided to accomplish these objectives. The procedures for deriving the exact distribution $f(q)$ and computing the exact p-value are given as follows.

Deriving the exact distribution of Q

- Step 1. Consider all possible vectors $\mathbf{x} = (x_{11}, x_{21}, ..., x_{k1})$, which could arise in these k studies with sizes $x_1, x_2, ..., x_k$. The number of these vectors is given by $M = \prod_{i=1}^{k} (x_i + 1)$. Note that $x_1, x_2, ..., x_k$ are considered fixed in this approach.
- Step 2. Compute the observed value of Q for each of these vectors $\mathbf{x}$ from

$$q_{\mathbf{x}} = \sum_{i=1}^{k} \frac{(x_{i1} - x_i \pi_i)^2}{x_i \pi_i (1 - \pi_i)}.$$

- Step 3. Consider each vector of $\mathbf{x}$ and its associated probability

$$f(\mathbf{x}) = f(q_{\mathbf{x}}) = \prod_{i=1}^{k} \binom{x_i}{x_{i1}} \pi_i^{x_{i1}} (1 - \pi_i)^{x_i - x_{i1}},$$

  where $0 \leq f(\mathbf{x}) \leq 1$.

- Step 4. Consider the discrete mass function of Q

$$f(q) = \sum_{\mathbf{x} \in \Omega_q} f(\mathbf{x}), \tag{5}$$

  where $\Omega_q = \{\mathbf{x} | Q = \sum_{i=1}^{k} \frac{(x_{i1} - x_i \pi_i)^2}{x_i \pi_i (1-\pi_i)} = q\}$. This gives the desired distribution $f(q)$, the exact distribution of Q.

Note that in Step 1, a binomial random sample x_{i1} can take any value from zero to the total number of events x_i, hence there are $x_i + 1$ different values. When all binomial random samples in k independent studies are considered, the size of sample space is then given by M, the combinations of $x_i + 1$, for $i = 1, 2, ..., k$. In `R` (https://www.r-project.org/), a simple function suggested for enumerating all possibilities is `expand.grid()`. An example will be given in the next section. In Step 2, the parameter π_i is denoted as the event probability in the intervention group of study i and is given by $\pi_i = \frac{\mu_{i1} T_{i1}}{\mu_{i0} T_{i0} + \mu_{i1} T_{i1}}$. Dividing fractions by $\mu_{i0} T_{i0}$, the probability event can be re-written as $\pi_i = \frac{RR(T_{i1}/T_{i0})}{1 + RR(T_{i1}/T_{i0})}$. We can see that this probability is dependent only on the parameter of interest, the risk ratio. In Step 3, M values of $f(\mathbf{x})$ are obtained and provide the exact distribution of Q in Step 4.

Computing the exact p-value of Q

- Step 1. Compute the observed values of π_i and Q from the real data.
- Step 2. Consider M possible vectors $\mathbf{x} = (x_{11}, x_{21}, ..., x_{k1})$, which could arise in k studies with sizes $x_1, x_2, ..., x_k$.
- Step 3. Compute the observed value of Q for each of these vectors $\mathbf{x}$ from $q_{\mathbf{x}}$.
- Step 4. Calculate the likelihood function $f(\mathbf{x})$.
- Step 5. Compute the p-value of Q from

$$p = \sum_{\mathbf{x} \in \Omega} f(\mathbf{x}), \tag{6}$$

  where $\Omega = \{\mathbf{x} | q_{\mathbf{x}} \geq Q\}$.

Note that in Step 1 the estimator of π_i is given by $\hat{\pi}_i = \frac{\widehat{RR}_{MH}(T_{i1}/T_{i0})}{1+\widehat{RR}_{MH}(T_{i1}/T_{i0})}$. This can be computed even if the studies include zero events. For each of the vectors in Step 2, $q_{\mathbf{x}}$ is computed, and compared to the observed Q obtained from Step 1. If $q_{\mathbf{x}} \geq Q$, this implies that the null hypothesis is rejected. In Step 5, the p-value is then derived. Since it uses the exact distribution in its algorithmic construction, this is noted as the *exact p-value* of Q.

We emphasize that the exact p-value will be used to make a decision in hypothesis testing. If the exact p-value or p given in equation (6) is smaller than the significance level (α), the null hypothesis is rejected. Hence, we have estimated the type I error on the basis of the exact p-value:

$$\hat{\alpha} = Pr(p \leq \alpha | H_0) = 1 - Pr(p > \alpha | H_0),$$

where $Pr(p \leq \alpha | H_0)$ refers to the probability that H_0 is rejected given it is true.

3. Empirical illustrations

In this section, we demonstrate how the exact p-value of Q works using real data. The example on the number of deaths in routine early pregnancy ultrasound and selective induction of labour [17] mentioned in Section 1 are applied. Many studies in this dataset include rare events in both treatment (routine induction) and comparison arms. Q is only defined for $x_i > 0$ in the denominator. The studies with $x_i = 0$ (double-zero studies) are excluded before the analysis, and eight remaining studies of the perinatal death data are used. The Mantel-Haenszel estimate for the risk ratio is computed and given by $\widehat{RR}_{MH} = 0.11$ with the 95% confidence interval of (0.01, 0.88). This shows a decrease of the number of perinatal deaths in the selective induction arm and a high preventive effect of this intervention.

We calculate the Q statistic given in equation (3). To find the exact p-value, the steps discussed in Section 2 are used with `R` package. `R`-code for the use in applications is given in the Appendix. Next, the p-values obtained from two methods (approximate and bootstrap methods) are investigated, and compared to the proposed exact p-value.

Table 2. P-values and observed values of Q statistic for homogeneity test of the risk ratios using the dataset on the perinatal death in post-term pregnancy ($\alpha = 0.05$).

Approach	Exact method	Approximate method	Bootstrap method 1	Bootstrap method 2
Observed Q statistic	9.9822	9.9822	7.9858	7.9915
p-value	0.3604	0.1896	0.3412	0.3427

From the data example, the approximate method uses the Q test statistic based on a chi-square distribution with seven degrees of freedom. The p-value is simply obtained by $p_{\chi^2} = Pr(Q \geq q = 9.9822) = 0.1896$. Then, we compute the p-value using the bootstrap method as a resampling based approach. We design two different methods to obtain the distribution under the null hypothesis and derive the bootstrap p-value. The processes of these methods are given in the following of this section: Algorithm 1 based on recalculation of $\hat{\pi}_i$ for every bootstrap sample and Algorithm 2 based on fixed $\hat{\pi}_i$. Table 2 incorporates the results obtained from four methods. For the bootstrap methods, the observed values of Q and the probability values are estimated on average values of 1,000 samples. All p-values are greater than the significance level at 0.05. It can be concluded that no statistically significant difference in risk ratios is found across studies for this dataset.

From the results in Table 2, we point out that the probability values computed from the methods which do not depend on the chi-square distribution are similar, having values around 0.34 to 0.36. Meanwhile, the p-value based on the approximate method shows the lowest value and differs from the other methods. Therefore, it is of interest to investigate the performance of these methods, in particular, if the latter three give the correct null-distribution. Therefore, these investigation are conducted using simulations in the next section.

Computing the bootstrap p-value (Algorithm 1)

- Step 0. Fix the bootstrap samples $B = 1,000$ (or a similar large number) in order to obtain an accurate estimate of the p-value, and compute the estimates of π_i and Q from the real data.
- Step 1. Draw a sample of size k with replacement from the data, leading to x_i^* and π_i^*.
- Step 2. Sample x_{i1}^* from $Bi(x_i^*, \pi_i^*)$.

- Step 3. Compute π_i^{**} using information obtained from Steps 1 and 2.
- Step 4. Compute a bootstrap sample $Q^{**} = \sum_{i=1}^{k} \frac{(x_{i1}^{*} - x_i^{*}\pi_i^{**})^2}{x_i^{*}\pi_i^{**}(1-\pi_i^{**})}$.
- Step 5. Repeat the procedure in Steps 1 to 4 for B times.
- Step 6. Calculate the bootstrap p-value from $bp_1 = \frac{n(Q^{**} \geq Q)}{B}$, where $n(Q^{**} \geq Q)$ is the number of times that Q^{**} is greater than Q.

Computing the bootstrap p-value (Algorithm 2)

- Step 0. Fix the bootstrap samples $B = 1,000$ in order to obtain an accurate estimate of the p-value, and compute the estimates of π_i and Q from the real data.
- Step 1. Draw a sample of size k with replacement from the data, leading to x_i^{*}.
- Step 2. Sample x_{i1}^{*} from $Bi(x_i^{*}, \pi_i)$.
- Step 3. Compute a bootstrap sample $Q^{*} = \sum_{i=1}^{k} \frac{(x_{i1}^{*} - x_i^{*}\pi_i)^2}{x_i^{*}\pi_i(1-\pi_i)}$.
- Step 4. Repeat the procedure in Steps 1 to 4 for B times.
- Step 5. Calculate the bootstrap p-value from $bp_2 = \frac{n(Q^{*} \geq Q)}{B}$, where $n(Q^{*} \geq Q)$ is the number of times that Q^{*} is greater than Q.

4. Simulation study

To evaluate the performance of the proposed method, simulation studies were conducted to compare the type I error rate of the proposed p-value to that of the approximate p-value (referred to p_{χ^2} hereafter) and the bootstrap p-values (referred to bp_1 and bp_2). The simulations were carried out using `R` [25], and designed to cover scenarios with varying number of studies k = 3, 5, 10, and 15. The total number of events x_i were generated from the discrete uniform distribution on $U[1, 2]$. Hence the binomial size parameter can only take vlues 1 or 2. This seems like an extreme situation but mimics the motivating data example closely. The number of events x_{i1} were sampled from a binomial distribution $Bi(x_i, \pi_i)$, where π_i was 0.3, 0.4, 0.5, and 0.6 reflecting the risk ratios as 0.4, 0.7, 1, and 1.5, respectively. These were set to mimic meta-analysis with small events. The nominal significance level was given by $\alpha = 0.05$. Each case in simulations was repeated R = 5,000 times, and B = 1,000 bootstrap

Table 3. Simulation results for the average p-values and type I errors of the Q statistic for test of homogeneity using four methods ($\alpha = 0.05$).

		Approximate method		Exact method		Bootstrap method 1		Bootstrap method 2	
k	q	p-value	Type I error	p-value	Type I error	p-value	Type I error	p-value	Type I error
3	0.3	0.3383	0.0004	0.3457	0.0328	0.4259	0.0031	0.4258	0.0035
	0.4	0.3199	0.0006	0.3358	0.0384	0.4103	0.0042	0.4101	0.0036
	0.5	0.3159	0.0006	0.3325	0.0466	0.4054	0.0048	0.4052	0.0046
	0.6	0.3194	0.0005	0.3353	0.0424	0.4088	0.0046	0.4083	0.0042
5	0.3	0.3599	0	0.3922	0.0342	0.4134	0.0042	0.4133	0.0038
	0.4	0.3426	0	0.3846	0.0487	0.4090	0.0064	0.4097	0.0050
	0.5	0.3358	0	0.3849	0.0503	0.4179	0.0052	0.4177	0.0058
	0.6	0.3433	0	0.3820	0.0453	0.4083	0.0053	0.4086	0.0054
10	0.3	0.4115	0.0024	0.4744	0.0212	0.4913	0.0124	0.4922	0.0122
	0.4	0.3805	0	0.4498	0.0473	0.4519	0.0425	0.4511	0.0440
	0.5	0.4061	0.0020	0.4997	0.0567	0.5146	0.0328	0.5147	0.0316
	0.6	0.3736	0	0.4348	0.0507	0.4522	0.0463	0.4521	0.0462
15	0.3	0.3907	0	0.4624	0.0095	0.4708	0.0039	0.4706	0.0042
	0.4	0.3866	0	0.4805	0.0225	0.4840	0.0117	0.4840	0.0108
	0.5	0.3936	0	0.4794	0.0275	0.4832	0.0175	0.4833	0.0164
	0.6	0.3871	0	0.4783	0.0237	0.4849	0.0103	0.4849	0.0103

samples were used for each case. The criteria used to evaluate the performance of the approach was the type I error. The latter was estimated by

$$\hat{\alpha} = 1 - \frac{n(\text{p-value} > \alpha | H_0)}{R},$$

where $n(\text{p-value} > \alpha | H_0)$ is the number that the observed p-value is greater than the given significance level under data generated under the null hypothesis. The method that has an average type I error closest to the nominal significance level is preferred. In the other words, the method that can control the type I error rate best outperforms the comparison.

The simulation results are presented in Table 3. The p-values obtained from the approximate method were the smallest in all situations. This matched the result conducted using the real dataset in Section 3. Moreover, this method provided the simulated type I error close to zero, and much lower than the target significance level at $\alpha = 0.05$. It is clear that the approximate method is inappropriate for the rare evens situations. Next, the methods that do not depend on a chi-square distribution were considered. The results showed that two bootstrap methods had similar p-value. Their p-values did not differ much from that of the exact method. These also matched the result in the case study. However, when we considered the evaluation criteria, only type I errors of the exact method were close to the nominal significance level at

$\alpha = 0.05$, especially in small number of studies ($k \leq 10$). Overall, ranking by closeness of the type I error to the nominal significance level gave preference to the exact and then the bootstrap methods. We therefore conclude that our proposed method can satisfactorily control the type I error rates and perform much better than the approximate method in meta-analysis of rare events. It is important to recall that the use of statistics based on an approximate method to test of homogeneity in meta-analysis when studies involve small events must be done with great care.

Then, we investigated the null distribution of the exact test to support the accuracy of the exact approach. The exact distribution of Q was determined using the proposed procedure presented in Section 2. In comparison, we generated the exact null-distribution of Q using simulation based on product-binomial sampling from $10,000$ replications. The distributions of Q considered from many scenarios are shown graphically as in Figures 2 and 3. From the results, our proposed method provided the distribution of Q very similar to the simulated distribution for any setting. We take this as evidence that the null distribution of Q has been correctly determined with the algorithm given in equation (5).

5. Discussion

Testing whether the study results are homogeneous is an important topic in meta-analysis. Evidence of heterogeneity indicates aptness of a random effects approach. We have seen that for rare events analysis the conventional chi-square statistic Q_{Hom} is not defined when studies with no events occur. The alternative test statistic Q which is suggested here and based on the conditional binomial approach is always defined but the approximation of the chi-square distribution poor. This is particular true for the rare events setting considered here and also persists when the number of studies k increase. Kulinskaya and Dollinger [?] suggested to use the gamma distribution to approximate the true distribution of Q. We have found that their suggestion lead to considerable improvement in the approximation if compared to the chi-square distribution. However, for our extreme scenario, also the approximation with the gamma distribution remains unsatisfactory.

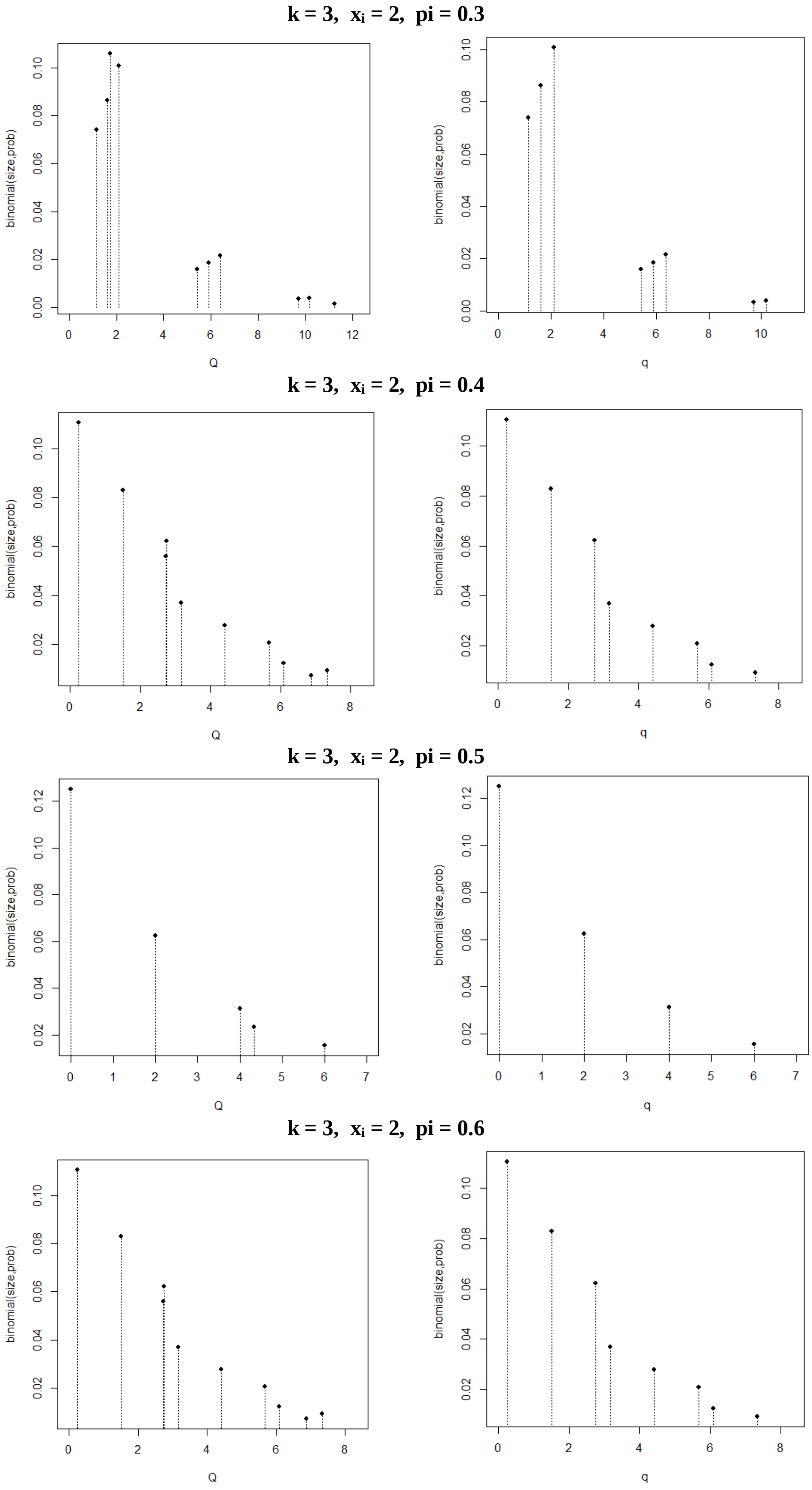


Figure 2. The distributions of Q: simulated distribution (left) and exact distribution (right) for $k = 3$, $x_i = 2$, and vary settings of π_i.

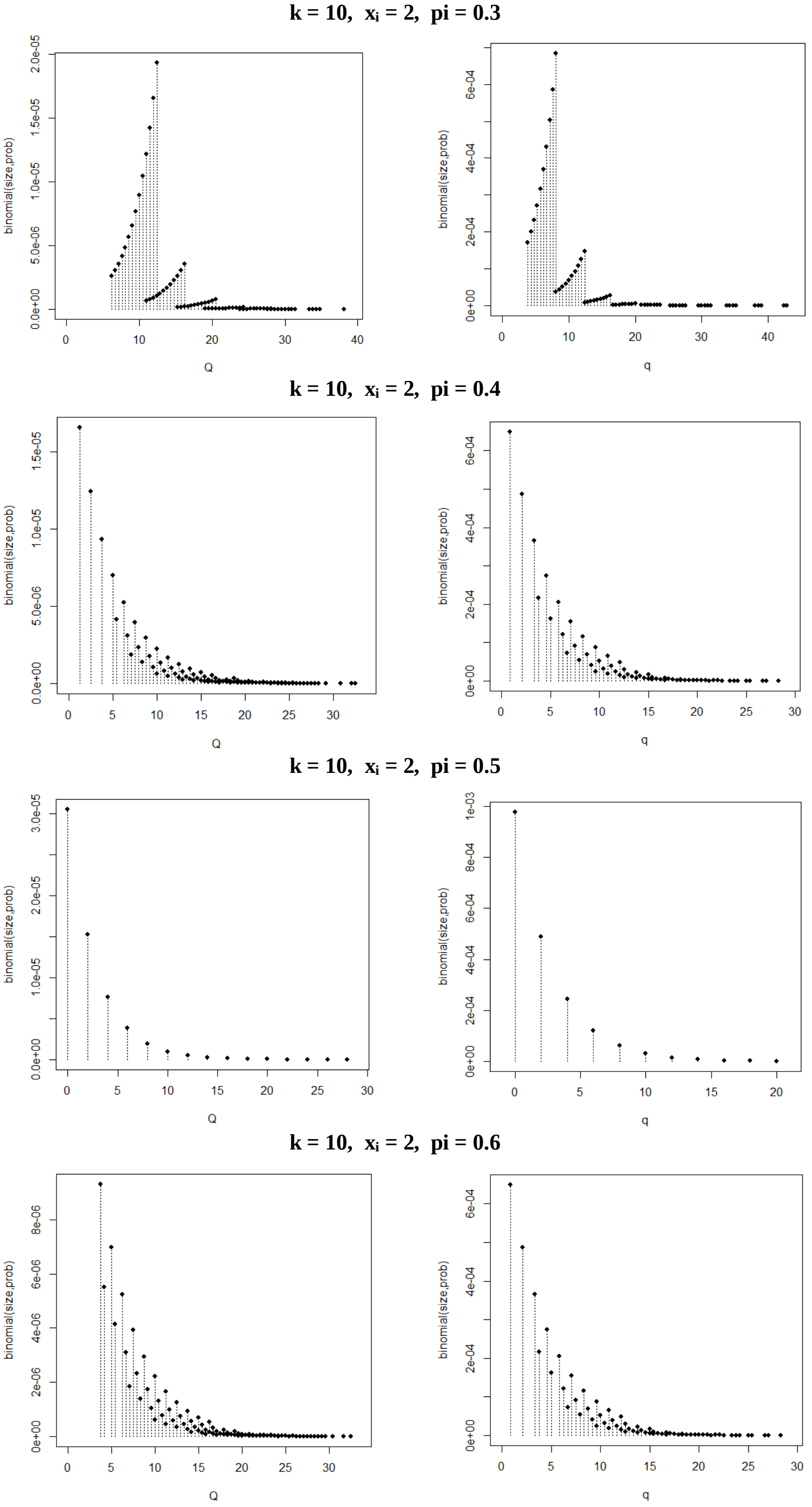


Figure 3. The distributions of Q: simulated distribution (left) and exact distribution (right) for k = 10, x_i = 2, and vary settings of π_i.

It appears that there seems no simple and valid approximation of the null-distribution of Q for the very rare event case. Hence a computational approach seems the best way to find the true null distribution of Q which is suggested in this paper. In practice, the exact distribution is derived using the process given in Section 2 using `R`-code given in the Appendix.

The simulations were used to investigate the performance of the proposed method. It was found that in the cases of rare events the exact p-value of Q can adequately control the type I error rates. The exact p-value had a type I error rate closer to the nominal significance level. Also, it outperformed the p-values obtained from the approximate method based on a chi-square distribution with $k-1$ degrees of freedom and the bootstrap method. Let us discuss this point. We first focus on the Q statistic based on $\chi^2_{df.=k-1}$-distribution. When small number of events occur in a meta-analysis, the observed value computed from Q is also small (smaller than the one obtained from a common events setting and smaller than expected under a valid chi-square distribution), while the degree of freedom used to find the critical value is still from a chi-square with $k-1$ df. In statistical hypothesis testing, this makes it is hard to reject the null hypothesis, especially in meta-analysis of studies with extremely small events. We assume that this is a reason to find a low performance of the conventional test in terms of type I error. For the method proposed in this paper, the exact test uses information from the available data. The p-value is computed based on the exact distribution and does entirely not depend on the degrees of freedom or critical value, which is often muddled up in the analysis of studies with rare events. According to Figures 2 and 3, the exact distribution of Q introduced in this paper was identical to the simulated distribution. This demonstrates that the proposed method provides the correct distribution of Q in the case of rare events. The exact p-value has a simple structure and is easy to compute, given the developed code in `R`.

Appendix

Example of R-code used to compute the exact p-value of Q is given in the following. Here, the perinatal death data noted in Section 3 are applied.

```
######### Computing the exact p-value of Q #########

studyi =c(1,3,6,9,11,13,14,14)

pi1 = c(55,131,78,195,152,94,188,1701)

xi1 = c(0,0,1,0,0,0,0,0)

pi0 = c(57,134,78,207,150,94,168,1706)

xi0 = c(2,1,0,1,1,1,1,2)

xi = xi0+xi1; pi = pi0+pi1; ri = pi1/pi0

data = data.frame(studyi,pi1,xi1,pi0,xi0,pi,xi,ri)

data = data[!(data$xi==0), ]

rrmh = sum(data$xi1*data$pi0/data$pi)/sum(data$xi0*data$pi1/data$pi)

qi = data$ri*rrmh/(1+data$ri*rrmh)

data = data.frame(data,qi)

Q = sum((data$xi1-data$qi*data$xi)^2/(data$xi*data$qi*(1-data$qi)))

approximate.p.value = pchisq(Q,length(data$xi)-1, lower.tail=FALSE )

pi1 = data$pi1; xi1 = data$xi1; pi0 = data$pi0; xi0 = data$xi0

pi = data$pi; xi = data$xi; ri = data$ri; qi = data$qi; k = length(xi)

sam = expand.grid(x1.1=0:xi[1], x1.2=0:xi[2], x1.3=0:xi[3], x1.4=0:xi[4], x1.5=0:xi[5],

      x1.6=0:xi[6], x1.7=0:xi[7], x1.8=0:xi[8] )

pi1 = matrix(pi1, 1, k, dimnames = list(c(), c("p1.1", "p1.2", "p1.3", "p1.4", "p1.5",

      "p1.6", "p1.7", "p1.8")) )

pi0 = matrix(pi0, 1, k, dimnames = list(c(), c("p0.1", "p0.2", "p0.3", "p0.4", "p0.5",

      "p0.6", "p0.7", "p0.8")) )

pi = matrix(pi, 1, k, dimnames = list(c(), c("pi.1", "pi.2", "pi.3", "pi.4", "pi.5",

      "pi.6", "pi.7", "pi.8")) )

ri = matrix(ri, 1, k, dimnames = list(c(), c("ri.1", "ri.2", "ri.3", "ri.4", "ri.5",

      "ri.6", "ri.7", "ri.8")) )

xi = matrix(xi, 1, k, dimnames = list(c(), c("xi.1", "xi.2", "xi.3", "xi.4", "xi.5",

      "xi.6", "xi.7", "xi.8")) )

data.b = data.frame(sam, ri, pi1,pi0,pi,xi )

data.b = transform(data.b, x0.1 = xi.1-x1.1, x0.2 = xi.2-x1.2, x0.3 = xi.3-x1.3, x0.4 = xi.4-x1.4,

     x0.5 = xi.5-x1.5, x0.6 = xi.6-x1.6, x0.7 = xi.7-x1.7, x0.8 = xi.8-x1.8 )

data.b = transform(data.b, b1 = x1.1*p0.1/pi.1, c1 = x0.1*p1.1/pi.1, b2 = x1.2*p0.2/pi.2,

      c2 = x0.2*p1.2/pi.2, b3 = x1.3*p0.3/pi.3, c3 = x0.3*p1.3/pi.3, b4 = x1.4*p0.4/pi.4,

      c4 = x0.4*p1.4/pi.4, b5 = x1.5*p0.5/pi.5, c5 = x0.5*p1.5/pi.5, b6 = x1.6*p0.6/pi.6,

      c6 = x0.6*p1.6/pi.6, b7 = x1.7*p0.7/pi.7, c7 = x0.7*p1.7/pi.7, b8 = x1.8*p0.8/pi.8,

      c8 = x0.8*p1.8/pi.8 )

data.b = transform(data.b, b = b1+b2+b3+b4+b5+b6+b7+b8, c = c1+c2+c3+c4+c5+c6+c7+c8 )

data.b = transform(data.b, rrmh=b/c )

data.b = transform(data.b, qi.1 = data$qi[1], qi.2 = data$qi[2], qi.3 = data$qi[3],

      qi.4= data$qi[4], qi.5= data$qi[5], qi.6= data$qi[6], qi.7= data$qi[7], qi.8 = data$qi[8] )
```

```
data.b = transform(data.b, a1 = (x1.1-qi.1*xi.1)^2/(qi.1*xi.1*(1-qi.1)),
      a2=(x1.2-qi.2*xi.2)^2/(qi.2*xi.2*(1-qi.2)), a3=(x1.3-qi.3*xi.3)^2/(qi.3*xi.3*(1-qi.3)),
      a4=(x1.4-qi.4*xi.4)^2/(qi.4*xi.4*(1-qi.4)), a5=(x1.5-qi.5*xi.5)^2/(qi.5*xi.5*(1-qi.5)),
      a6=(x1.6-qi.6*xi.6)^2/(qi.6*xi.6*(1-qi.6)), a7=(x1.7-qi.7*xi.7)^2/(qi.7*xi.7*(1-qi.7)),
      a8=(x1.8-qi.8*xi.8)^2/(qi.8*xi.8*(1-qi.8)) )
data.b = transform(data.b, q = a1+a2+a3+a4+a5+a6+a7+a8 )
pval = ifelse(data.b$q > Q, 1, 0)
data.b = data.frame(data.b, pval)
data.b = data.b[complete.cases(data.b), ]
data.b = data.b[!(data.b$pval==0), ]
data.b = transform(data.b, pdf = dbinom(x1.1, xi.1, qi.1)*dbinom(x1.2, xi.2, qi.2)*
      dbinom(x1.3, xi.3, qi.3)*dbinom(x1.4, xi.4, qi.4)*dbinom(x1.5, xi.5, qi.5)*
      dbinom(x1.6, xi.6, qi.6)*dbinom(x1.7, xi.7, qi.7)*dbinom(x1.8, xi.8, qi.8) )
exaxt.p.value = sum(data.b$pdf)
```

References

[1] Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. Biom J. 2015;57(4):633–648.

[2] Schiff JH, Welker A, Fohr B, et al. Major incidents and complications in otherwise healthy patients undergoing elective procedures: Results based on 1.37 million anaesthetic procedures. Br J Anaesth. 2014;113(1):109–121.

[3] Schulze R, Holling H, Böhning D, editors. Meta-Analysis: New Developments and Applications in Medical and Social Sciences. Massachusetts: Hogrefe and Huber Publishing; 2003.

[4] Borenstein M, Hedges LV, Higgins JP, et al., editors. Introduction to Meta-Analysis. Chichester: John Wiley & Sons; 2009.

[5] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med. 2004;23(9):1351–1375.

[6] Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. Stat Med. 2007;26(1):53–77.

[7] Keus F, Wetterslev J, Gluud C, et al. Robustness assessments are needed to reduce bias in meta-analyses that include zero-event randomized trials. Am J Gastroenterol. 2009; 104(3):546–551.

[8] Kuss O. Statistical methods for meta-analyses including information from studies without any events - add nothing to nothing and succeed nevertheless. Stat Med. 2015;34(7):1097–

1116.

[9] Efthimiou O. Practical guide to the meta-analysis of rare events. Evidence-Based Mental Health. 2018;21(2):72–76.

[10] Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions? Biom J. 2018;60(6):1040–1058.

[11] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22(4):9–48.

[12] Casella G, Berger RL, editors. Statistical Inference. California: Duxbury Press; 2002.

[13] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–1558.

[14] Almalik O, Heuvel ER. Testing homogeneity of effect sizes in pooling 2 × 2 contingency tables from multiple studies: A comparison of methods. Cogent Math Stat. 2018;5(1):1–18.

[15] Cochran W. Some methods for strengthening the common χ^2 tests. Biometrics. 1954; 10(4):417–451.

[16] Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ. 2003;327:557–560.

[17] Crowley P. Interventions for preventing or improving the outcome of delivery at or beyond term. Cochrane Database Syste Reviews. 2000;2:CD000170.

[18] Sánchez-Meca J, Marán-Martánez F. Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and type I error. Qual Quant. 1997;31(4):385–399.

[19] Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. J Clin Epidemiol. 2001;54(10):1046–1055.

[20] Viechtbauer W. Hypothesis tests for population heterogeneity in meta-analysis. Br J Math Stat Psychol. 2007;60(1):29–60.

[21] Greenland S, Mansournia MA, Altman DG. Sparse data bias: A problem hiding in plain sight. BMJ. 2016;352:1–6.

[22] Böhning D, Holling H, Böhning W, et al. Investigating heterogeneity in meta-analysis of studies with rare events. Metron. 2020;0:1–19.

[23] Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat Med. 2010;29(29):3046–367.

[24] Friede T, Röver C, Wandel S, et al. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. Biom J. 2017;59(4):658–671.

[25] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.