mathematics

MDPI

Article

Common Test and Interval Estimation of Risk Ratio for Complex Paired Data Under Dallal's Model

Shuman Sun [1] and Zhiming Li [2,*]

[1] College of Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2] College of Mathematics and System Science, Xinjiang University, Urumqi 830046, China
* Correspondence: zmli@xju.edu.cn

Abstract

Unilateral or bilateral paired data are often encountered when people receive treatment for paired organs or body parts. Compared with using only unilateral or bilateral data alone, the utilization of combined data can provide additional information. With combined unilateral and bilateral data, this article aims to propose three statistical tests and three confidence interval methods of relative risk ratio in a stratified design under Dallal's model. Our simulation results show that the score test has a more robust type I error rate compared to other tests in all cases. Meanwhile, the confidence interval method based on the score test always provides coverage probability close to the nominal level and satisfactory coverage width. Finally, two real examples of otolaryngology and myopathy are applied to illustrate the application of the proposed tests and confidence interval methods.

Keywords: Dallal's model; relative risk ratio; common test; confidence interval

MSC: 92B10

1. Introduction

In clinical medical studies, particularly in ophthalmologic and otology studies, data are often collected in pairs. However, when it is impossible to collect complete bilateral data due to certain special factors of the patients, only the unilateral data can be obtained. So, each subject may contribute measurements from just one side (unilateral) or both sides (bilateral). At the end of treatment, the unilateral data consists of improved or unimproved responses. The bilateral data can be recorded as all improvements, partial improvement, or no improvement.

Many studies have been conducted to analyze the issues of statistical inference for unilateral data [1–3]. For the bilateral data, current studies tend to analyze them in isolation rather than treating the human subject as a unified entity [4,5]. However, Zhang and Ying [6] indicated that this form of data processing is relatively straightforward and simple, but it may fail to conduct proper data analysis when obtaining the bilateral data. This relates to the correlation between paired organs or parts, and ignoring this attribute will lead to misleading inferences. For instance, in ophthalmologic studies, if a person has myopia in their left eye, then the probability that they also have myopia in their right eye is quite high. This indicates that there is a strong correlation between the left and right eyes.

Some probability models have been provided to analyze the bilateral data. Suppose that the probability of a response at the other side was proportional to the prevalence rate of the corresponding group when one organ was given a response. Rosner [7] proposed

Check for updates

Academic Editor: Tiago Dias Domingues

Received: 6 February 2026
Revised: 6 March 2026
Accepted: 9 March 2026
Published: 11 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

an interclass correlation model. However, Rosner's model might lead to a poor fit if the bilateral data come from a population with widely varying group-specific prevalence. Then, Dallal [8] proposed an alternative model for bilateral data. This model assumes that the probability that another organ responds when one organ responds was independent of the probability that the organ responds. Additionally, Donner [9] proposed another model, in which the interclass correlation coefficient was common in each of two groups. Moreover, it was very important to select an appropriate method to explain the interclass correlation within a subject in the statistical inference of bilateral data. Ignoring this inherent data interdependency will undermine the ability to identify the true therapeutic effect accurately. Hence, under these three classical models, there have been research results on the statistical inference of bilateral data in recent years. For asymptotic tests, Tang et al. [10] provided eight statistics to test the homogeneity of risk differences in two groups. The results indicated that the performance of the approximate unconditional score test was superior to that of other tests. Zhuang et al. [11] derived four test statistics of risk ratios for stratified bilateral data. For confidence intervals, Tang et al. [12] studied five asymptotic confidence interval methods for proportion difference. The empirical results showed that Wald-type, score, and bootstrap confidence intervals performed satisfactorily for small to large sample sizes in the sense that their empirical coverage probabilities were close to the pre-specified nominal confidence level. Shen et al. [13] derived the confidence interval methods of risk difference for the stratified bilateral data, including the global Wald-type, complete Wald-type, weighted Wald-type, profile likelihood, and score confidence intervals. The results showed that the score confidence interval induced from the score test statistic performs satisfactorily.

Although the aforementioned research has been applied to unilateral or bilateral data, they may not be applicable to investigations into stratified unilateral and bilateral data. In practice, we often encounter unilateral and bilateral data (Scudamore et al. [14], Petrella et al. [15], and Mona et al. [16]). The statistical inference for unilateral and bilateral data has received some attention in recent years. Under the stratified design, Qiu et al. [17] constructed five confidence intervals of risk differences under Rosner's model. The result pointed out that the score confidence interval method for the common proportion difference behaved well even under small sample sizes. Based on Donner's model, Wang et al. [18] derived three homogeneity tests (the likelihood ratio test, the Wald test, and the score test) for the risk ratio. The results indicated that the score test behaved well in various parameter settings, as the type I error rate was always close to the pre-defined significance level. Moreover, the complete test procedure and interval estimation of the odds ratio were discussed by Hua et al. [19,20]. Under Dallal's model, Liang et al. [21] proposed four CI methods and indicated that the profile likelihood confidence interval outperformed other methods. Sun et al. [22] researched various confidence intervals of risk difference. The result showed that the score method outperformed other methods. Sun et al. [23] proposed six statistical tests and eight sample size methods for risk difference. It indicated that the score test behaved well in various parameter settings.

However, the test procedure and confidence interval methods for the relative risk ratio have not been discussed yet under Dallal's model. Meanwhile, the likelihood functions generated by using relative risk ratio and risk differences are different, which also leads to differences in the theoretical reasoning process of statistical inference methods. Moreover, the relative risk ratio would perform better than the risk difference when the differences between datasets are relatively small [24]. In view of this, this paper aims to investigate the common test and confidence intervals of relative risk ratio for complex paired data under Dallal's model.

2. Data Structure and Dallal's Model

Considering a clinical trial with stratified randomization, the data are divided into a total of J strata and each contains two groups to mitigate potential confounding effects such as age, gender, disease severity, and so on. In the jth stratum ($j = 1, 2, \ldots, J$), N_j subjects contribute unilateral data and N'_j subjects contribute bilateral data. Suppose that $m^{(1)}_{lij}$ is the number of unilateral patients with l responses, and $m^{(2)}_{l'ij}$ is the number of bilateral patients with l' responses in the ith group ($l = 0, 1$, $l' = 0, 1, 2$, $i = 1, 2$) of the jth stratum. For each stratum, denote $m^{(1)}_{ij} = m^{(1)}_{0ij} + m^{(1)}_{1ij}$, $n^{(1)}_{lj} = m^{(1)}_{l1j} + m^{(1)}_{l2j}$, $m^{(2)}_{ij} = m^{(2)}_{0ij} + m^{(2)}_{1ij} + m^{(2)}_{2ij}$, and $n^{(2)}_{l'j} = m^{(2)}_{l'1j} + m^{(2)}_{l'2j}$. The data structure of the jth stratum is summarized in Table 1.

Table 1. Data structure of the stratified unilateral and bilateral data in the jth stratum.

Number of Responses	Unilateral Group (i)		Total	Bilateral Group (i)		Total
	1	2		1	2	
0	$m^{(1)}_{01j}$	$m^{(1)}_{02j}$	$n^{(1)}_{0j}$	$m^{(2)}_{01j}$	$m^{(2)}_{02j}$	$n^{(2)}_{0j}$
1	$m^{(1)}_{11j}$	$m^{(1)}_{12j}$	$n^{(1)}_{1j}$	$m^{(2)}_{11j}$	$m^{(2)}_{12j}$	$n^{(2)}_{1j}$
2	-	-	-	$m^{(2)}_{21j}$	$m^{(2)}_{22j}$	$n^{(2)}_{2j}$
Total	$m^{(1)}_{1j}$	$m^{(1)}_{2j}$	N_j	$m^{(2)}_{1j}$	$m^{(2)}_{2j}$	N'_j

For unilateral data, let $Z^{(1)}_{ijk}$ be the indicator for judging whether the kth patient has a response or not in the ith group of the jth stratum. If there is a response, then $Z^{(1)}_{ijk} = 1$; otherwise, $Z^{(1)}_{ijk} = 0$. For bilateral data, define $Z^{(2)}_{ijkh} = 1$ if the hth organ ($h = 1, 2$) of the kth patient has a response, and $Z^{(2)}_{ijkh} = 0$ otherwise. Under Dallal's model, assume

$$P(Z^{(1)}_{ijk} = 1) = P(Z^{(2)}_{ijkh} = 1) = \pi_{ij},\ P(Z^{(2)}_{ijkh} = 1 | Z^{(2)}_{ijk(3-h)} = 1) = \gamma_{ij},$$

where π_{ij} ($0 \le \pi_{ij} \le 1$) is the response probability, and γ_{ij} ($0 \le \gamma_{ij} \le 1$) is the correlation coefficient of the paired data. If $\gamma_{ij} = \pi_{ij}$, then two organs are completely independent, and $\gamma_{ij} = 1$ if two organs are completely dependent.

Furthermore, the unilateral variable $m^{(1)}_{1ij}$ follows Binomial$(m^{(1)}_{ij}, \pi_{ij})$. And, the bilateral variable $(m^{(2)}_{0ij}, m^{(2)}_{1ij}, m^{(2)}_{1ij})$ follows Multinomial$(m^{(2)}_{ij}; p_{0ij}, p_{1ij}, p_{2ij})$, where $p_{0ij} = \pi_{ij}(\gamma_{ij} - 2) + 1$, $p_{1ij} = 2\pi_{ij}(1 - \gamma_{ij})$, and $p_{2ij} = \pi_{ij}\gamma_{ij}$, with $p_{0ij}, p_{1ij}, p_{2ij} \in [0, 1]$. Hence, the corresponding log-likelihood function for the observed data $\boldsymbol{m_{ij}} = \{m^{(1)}_{0ij}, m^{(1)}_{1ij}, m^{(2)}_{0ij}, m^{(2)}_{1ij}, m^{(2)}_{2ij}\}$ can be obtained as

$$\begin{aligned} l(\boldsymbol{m}_{ij}; \boldsymbol{\pi}, \boldsymbol{\gamma}) = \sum_{j=1}^{J} \sum_{i=1}^{2} [&m^{(1)}_{0ij} \log(1 - \pi_{ij}) + m^{(1)}_{1ij} \log \pi_{ij} + m^{(2)}_{0ij} \log(\pi_{ij}(\gamma_{ij} - 2) + 1) \\ &+ m^{(2)}_{1ij} \log(2\pi_{ij}(1 - \gamma_{ij})) + m^{(2)}_{2ij} \log(\pi_{ij}\gamma_{ij})] + \log C, \end{aligned} \tag{1}$$

where $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$, $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ})^T$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iJ})^T$ ($i = 1, 2$), and $C = \prod_{j=1}^{J} \prod_{i=1}^{2} \frac{m^{(1)}_{ij}! m^{(2)}_{ij}!}{m^{(1)}_{0ij}! m^{(1)}_{1ij}! m^{(2)}_{0ij}! m^{(2)}_{1ij}! m^{(2)}_{2ij}!}$ is a constant.

Let $\delta_j = \pi_{2j}/\pi_{1j}$ ($j = 1, 2, \ldots, J$) be the relative risk ratio between two groups in the jth stratum. If $\delta_1 = \delta_2 = \cdots = \delta_J \triangleq \delta$, we call δ the common relative risk ratio, which means that the differences between the two groups in all strata are the same. Therefore, we

are interested in whether the relative risk ratio is equal to a fixed constant δ_0 at all strata. That is to say, the common test is

$$H_0 : \delta = \delta_0 \quad \text{vs.} \quad H_1 : \delta \neq \delta_0.$$

Next, the parameter estimation under different hypotheses for the common test will be introduced. The MLEs under the alternative hypothesis and null hypothesis are called the unconstrained and constrained MLEs, respectively.

Unconstrained MLEs. Under the alternative hypothesis H_1 and Equation (1), the log-likelihood function can be expressed as

$$\begin{aligned} l_1(\delta, \boldsymbol{\pi}_1, \boldsymbol{\gamma}) = \sum_{j=1}^{J} [& m_{01j}^{(1)} \log(1 - \pi_{1j}) + m_{11j}^{(1)} \log \pi_{1j} + m_{02j}^{(1)} \log(1 - \delta\pi_{1j}) + m_{12j}^{(1)} \log(\delta\pi_{1j}) \\ & + m_{01j}^{(2)} \log(\pi_{1j}(\gamma_{1j} - 2) + 1) + m_{11j}^{(2)} \log(2\pi_{1j}(1 - \gamma_{1j})) + m_{21j}^{(2)} \log(\pi_{1j}\gamma_{1j}) \\ & + m_{02j}^{(2)} \log(\delta\pi_{1j}(\gamma_{2j} - 2) + 1) + m_{12j}^{(2)} \log(2\delta\pi_{1j}(1 - \gamma_{2j})) \\ & + m_{22j}^{(2)} \log(\delta\pi_{1j}\gamma_{2j})] + \log C, \end{aligned} \tag{2}$$

where two types of parameters are defined: δ is the parameter of interest, which is the focus of the common test and confidence interval estimation; π_{1j} and γ_{ij} $(i = 1, 2, j = 1, 2, \ldots, J)$ are nuisance parameters, which are required to fully specify the model but are not of direct scientific interest. Differentiating (2) with respect to $\pi_{1j}, \gamma_{ij}, \delta$ and setting them to 0, there is no close-form solutions. Thus, the Newton–Raphson process and Fisher scoring method are used to estimate all parameters. The detailed process are given by three steps as follows:

(i) Take the initial values $\delta^{(0)} = \pi_{2j}^{(0)} / \pi_{1j}^{(0)}$, and

$$\pi_{ij}^{(0)} = \frac{m_{1ij}^{(1)} + m_{1ij}^{(2)} + 2m_{2ij}^{(2)}}{m_{ij}^{(1)} + 2m_{ij}^{(2)}}, \quad \gamma_{ij}^{(0)} = \frac{4N_j \pi_{ij}^{(0)} (m_{21j}^{(2)} + m_{22j}^{(2)})}{(\sum_{i=1}^{2} (m_{1ij}^{(1)} + m_{1ij}^{(2)} + 2m_{2ij}^{(2)}))^2}.$$

(ii) The $(t+1)$th approximation $\delta^{(t+1)}$ is

$$\delta^{(t+1)} = \delta^{(t)} - \left(\frac{\partial^2 l_1}{\partial \delta^2}\right)^{-1} \frac{\partial l_1}{\partial \delta}\bigg|_{\pi_{1j} = \pi_{1j}^{(t)}, \gamma_{ij} = \gamma_{ij}^{(t)}}.$$

(iii) π_{1j} and γ_{ij} can be updated by the Fisher scoring algorithm

$$\begin{bmatrix} \pi_{1j}^{(t+1)} \\ \gamma_{1j}^{(t+1)} \\ \gamma_{2j}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \pi_{1j}^{(t)} \\ \gamma_{1j}^{(t)} \\ \gamma_{2j}^{(t)} \end{bmatrix} + I(\pi_{1j}^{(t)}, \gamma_{1j}^{(t)}, \gamma_{2j}^{(t)})^{-1} \begin{bmatrix} \frac{\partial l_1}{\partial \pi_{1j}} \\ \frac{\partial l_1}{\partial \gamma_{1j}} \\ \frac{\partial l_1}{\partial \gamma_{2j}} \end{bmatrix} \Bigg|_{\delta = \delta^{(t+1)}},$$

where I is the Fisher information matrix (see Appendix A for more details).

Constrained MLEs. Under the hypothesis $H_0 : \delta = \delta_0$, the log-likelihood function is $l_0(\boldsymbol{\pi}_1, \boldsymbol{\gamma}) \triangleq l_1(\delta_0, \boldsymbol{\pi}_1, \boldsymbol{\gamma})$. The constrained MLEs $\tilde{\pi}_{1j}, \tilde{\gamma}_{ij}$ of π_{1j}, γ_{ij} can be solved by the Fisher scoring algorithm below

$$\begin{bmatrix} \pi_{1j}^{(t+1)} \\ \gamma_{1j}^{(t+1)} \\ \gamma_{2j}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \pi_{1j}^{(t)} \\ \gamma_{1j}^{(t)} \\ \gamma_{2j}^{(t)} \end{bmatrix} + I_0(\pi_{1j}^{(t)}, \gamma_{1j}^{(t)}, \gamma_{2j}^{(t)})^{-1} \begin{bmatrix} \frac{\partial l_0}{\partial \pi_{1j}} \\ \frac{\partial l_0}{\partial \gamma_{1j}} \\ \frac{\partial l_0}{\partial \gamma_{2j}} \end{bmatrix} \Bigg|_{\delta = \delta_0},$$

where I_0 is the Fisher information matrix and can be obtained by $I(\pi_{1j}, \gamma_{1j}, \gamma_{2j})\ |_{\delta=\delta_0}$. Repeat the above step until convergence.

3. Asymptotic Tests

Likelihood ratio test. The likelihood ratio test statistic is given by

$$T_L = 2[l_1(\hat{\delta}, \hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\gamma}}) - l_0(\tilde{\boldsymbol{\pi}}_1, \tilde{\boldsymbol{\gamma}})],$$

where $\hat{\delta}$, $\hat{\boldsymbol{\pi}}_1$, and $\hat{\boldsymbol{\gamma}}$ are the unconstrained MLEs, and $\tilde{\boldsymbol{\pi}}_1$ and $\tilde{\boldsymbol{\gamma}}$ are the constrained MLEs. Under the null hypothesis H_0, T_L is asymptotically distributed as the χ^2 distribution with one degree of freedom. If $T_L > \chi^2_{1,1-\alpha}$, we should reject H_0, where $\chi^2_{1,1-\alpha}$ is the $100(1-\alpha)$ percentile of χ^2 distribution with one degree of freedom.

Wald-type test. Let $\boldsymbol{\beta} = (\delta\ \pi_{11}\ \gamma_{11}\ \gamma_{21}\ \dots\ \pi_{1J}\ \gamma_{1J}\ \gamma_{2J})^T_{1\times(3J+1)}$ and the corresponding unconstrained MLEs $\hat{\boldsymbol{\beta}} = (\hat{\delta}\ \hat{\pi}_{11}\ \hat{\gamma}_{11}\ \hat{\gamma}_{22}\ \dots\ \hat{\pi}_{1J}\ \hat{\gamma}_{1J}\ \hat{\gamma}_{2J})^T_{1\times(3J+1)}$. Suppose $\boldsymbol{C} = (1\,0\cdots0)_{1\times(3J+1)}$ and we have $\hat{\delta} = \boldsymbol{C}\hat{\boldsymbol{\beta}}$. Then, Wald-type statistic T_W of the common test can be expressed as

$$T_W = \frac{(\hat{\delta}-\delta_0)^2}{\mathrm{Var}(\hat{\delta})} = \frac{(\hat{\delta}-\delta_0)^2}{\boldsymbol{C}\mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^T}\Big|_{\delta=\hat{\delta},\boldsymbol{\pi}_1=\hat{\boldsymbol{\pi}}_1,\boldsymbol{\gamma}_1=\hat{\boldsymbol{\gamma}}_1,\boldsymbol{\gamma}_2=\hat{\boldsymbol{\gamma}}_2}.$$

Based on the asymptotic normality of MLEs, $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ is the inverse of Fisher information matrix $\mathcal{I}$ with respect to $\boldsymbol{\beta}$. Thus, it can be rewritten as

$$T_W = \frac{(\hat{\delta}-\delta_0)^2}{\boldsymbol{C}\mathcal{I}^{-1}\boldsymbol{C}^T} = \frac{(\hat{\delta}-\delta_0)^2}{\mathcal{I}^{-1}(1,1)},$$

where $\mathcal{I}^{-1}(1,1)$ is the first main-diagonal element of $\mathcal{I}^{-1}$. The detailed process is provided in Appendix B. The Wald-type test T_W is asymptotically distributed as a chi-square distribution with one degree of freedom under H_0.

Score test. Let $\boldsymbol{U} = (\frac{\partial l_0}{\partial\delta}, \frac{\partial l_0}{\partial\pi_{11}}, \frac{\partial l_0}{\partial\gamma_{11}}, \frac{\partial l_0}{\partial\gamma_{21}}, \dots, \frac{\partial l_0}{\partial\pi_{1J}}, \frac{\partial l_0}{\partial\gamma_{1J}}, \frac{\partial l_0}{\partial\gamma_{2J}})$ under the null hypothesis $H_0 : \delta = \delta_0$; then, the score test is given by

$$T_S = \boldsymbol{U}\mathcal{I}^{-1}\boldsymbol{U}^T\Big|_{\delta=\delta_0,\boldsymbol{\pi}_1=\tilde{\boldsymbol{\pi}}_1,\boldsymbol{\gamma}_1=\tilde{\boldsymbol{\gamma}}_1,\boldsymbol{\gamma}_2=\tilde{\boldsymbol{\gamma}}_2}.$$

The score test can be simplified as

$$T_S = (\sum_{j=1}^{J} e_j)^2\mathcal{I}^{-1}(1,1),$$

where

$$e_j = \frac{m^{(1)}_{12j} + m^{(2)}_{12j} + m^{(2)}_{22j}}{\delta_0} + \frac{m^{(1)}_{02j}\tilde{\pi}_{1j}}{\tilde{\pi}_{1j}\delta_0 - 1} + \frac{m^{(2)}_{02j}\tilde{\pi}_{1j}(\tilde{\gamma}_{2j}-2)}{\tilde{\pi}_{1j}\delta_0(\tilde{\gamma}_{2j}-2)+1},$$

and $\mathcal{I}^{-1}(1,1)$ is the first main-diagonal element of the inverse of $\mathcal{I}$. Moreover, T_S is asymptotically distributed as a chi-square distribution with one degree of freedom.

Remark 1. *The results of these three tests are mainly determined by the differences in the likelihood functions and the corresponding parameter estimates. Under the same model, the likelihood functions for unilateral or bilateral data can be regarded as a special case of the likelihood functions for both unilateral or bilateral data [21–23]. Therefore, these three tests for unilateral or bilateral data are also a special case of likelihood ratio tests for both unilateral and bilateral data.*

4. Confidence Interval Methods

Based on the above three test statistics, we present two confidence interval methods with an iterative algorithm, and the complete Wald-type confidence interval method derived through formula derivation.

Iteration methods. The null hypothesis $H_0 : \delta = \delta_0$ and alternative hypothesis $H_1 : \delta \neq \delta_0$ of the common test for relative risk ratio δ are helpful to construct the confidence intervals. Based on the two tests T_h $(h = L, S)$, let $p_{\delta_0}(\boldsymbol{m})$ be the two-sided p-value associated the null hypothesis, where $\boldsymbol{m}$ is the observed data. We reject the null hypothesis when $p_{\delta_0}(\boldsymbol{m}) \leqslant \alpha$. A conceptually simple way to construct the confidence intervals is to invert the corresponding test. Thus, the $100(1-\alpha)\%$ confidence interval of $\delta(\in (0, +\infty))$ satisfies

$$\mathrm{CI}_h = \left\{\delta : T_h \leq \chi^2_{1,1-\alpha}\right\}, \ \ h = L, S,$$

where $\chi^2_{1,1-\alpha}$ is a chi-square distribution with one degree of freedom. The inequality has no closed-form solutions. According to Xue et al. [25], the bisection method can be used to obtain the lower and upper limits of the confidence interval. For the upper limit, the iteration procedures are carried out as follows:

(i) Take initial value $\delta^{(0)} = \hat{\delta}$, step size $d \leftarrow 0.1$, and flag $f \leftarrow 1$.

(ii) The $(t+1)$th update is $\delta^{(t+1)} = \delta^{(t)} + d \times f$.

(iii) Given $\delta^{(t+1)}$, calculate the constrained MLEs $\tilde{\pi}_1^{(t+1)}$, $\tilde{\gamma}_1^{(t+1)}$ and $\tilde{\gamma}_2^{(t+1)}$.

(iv) If $f \times T_h \leq f \times \chi^2_{1,1-\alpha}$, return to step (ii). Otherwise, $d \leftarrow 0.1 \times d$, $f \leftarrow -f$, and return to step (ii).

(v) Repeat steps (ii)–(iv) until $d < 10^{-5}$.

For the lower limit, set $f \leftarrow -1$. In the step (iv), it can be replaced by $f \times T_h \geq f \times \chi^2_{1,1-\alpha}$.

Complete Wald-type CI. As can be seen from Section 3, the Wald-type test can be simplified as follows

$$T_W = \frac{(\hat{\delta} - \delta_0)^2}{\boldsymbol{C}\mathcal{I}^{-1}\boldsymbol{C}^T} = \frac{(\hat{\delta} - \delta_0)^2}{\mathcal{I}^{-1}(1,1)}.$$

Based on the asymptotic normality of the unconstrained MLEs, the Wald-type statistic asymptotically follows the standard normal distribution. Thus, the $100(1-\alpha)\%$ confidence interval is

$$\mathrm{CI}_W = \left[\hat{\delta} - z_{\alpha/2}\sqrt{\mathcal{I}^{-1}(1,1)}, \hat{\delta} + z_{\alpha/2}\sqrt{\mathcal{I}^{-1}(1,1)}\right].$$

Remark 2. *The iterative CI methods are applicable to all test methods with known asymptotic distributions, such as the CI methods based on the likelihood ratio test and score test. Furthermore, when the test statistic follows an asymptotic distribution and the parameter of interest admits an explicit expression, the confidence limits can be derived directly rather than through iteration.*

5. Monte Carlo Simulation

5.1. The Performance of Three Tests

In this section, we calculate the type I error rates (TIEs) and statistical powers to assess the performance of the proposed testing methods for the common test of relative risk ratio. The parameter settings are summarized in Table 2 for different event rates $\boldsymbol{\pi}_1$ and the probability γ_i $(i = 1, 2)$ of the other organ responding if one organ has improvement. The values of π_i and γ_i $(i = 1, 2)$ need to ensure that the corresponding p_{0ij}, p_{1ij}, and p_{2ij} are all within the range of $[0, 1]$. When determining the sample size, the selected sample size can reflect the impact of small samples, medium samples and large samples on the robustness of the testing method. According to reference [26], the samples can be roughly

classified as follows: (i) when $m < 30$, it is a small sample; (ii) when $30 < m < 200$, it is a medium sample; and (iii) when $m \geq 200$, it is a large sample, where m represents the number of groups. In addition, previous fitting results indicate that statistical properties improve with increasing sample size. The computational speed slows down significantly as the number of strata increases. To balance these factors in validating the proposed method, this study considers configurations of the number of each group $m = (m_{ij}^{(1)}, m_{ij}^{(2)}) = (25,25), (25,50), (50,50), (50,75), (100,100)$ and the stratum $J = 2, 4, 6$.

Table 2. The parameter configuration.

Parameter	Cases	Number of Strata		
		$J = 2$	$J = 4$	$J = 6$
π_1	I	(0.4,0.4)	(0.4,0.4,0.4,0.4)	(0.4,0.4,0.4,0.4,0.4,0.4)
	II	(0.4,0.5)	(0.4,0.5,0.4,0.5)	(0.4,0.5,0.4,0.5,0.4,0.5)
γ_1	a_1	(0.5,0.5)	(0.5,0.5,0.5,0.5)	(0.5,0.5,0.5,0.5,0.5,0.5)
	a_2	(0.5,0.6)	(0.5,0.6,0.5,0.6)	(0.5,0.6,0.5,0.6,0.5,0.6)
γ_2	b_1	(0.5,0.6)	(0.5,0.6,0.5,0.6)	(0.5,0.6,0.5,0.6,0.5,0.6)
	b_2	(0.6,0.7)	(0.6,0.7,0.6,0.7)	(0.6,0.7,0.6,0.7,0.6,0.7)

In the simulation studies, all tests are conducted at the significance level $\alpha = 0.05$. Assuming the relative risk ratios are equal to δ, 10,000 samples are randomly generated under the null hypothesis. The empirical type I error rates are the proportion rejecting the null hypothesis. According to Tang et al. [27], a test is liberal if its empirical TIEs are greater than 0.06, conservative if the TIEs are less than 0.04, and otherwise robust. The results of $J = 4$ are shown in Table 3, and the others ($J = 2, 6$) are provided in Supplementary Material S2. In summary, the empirical TIEs of the likelihood ratio test and the score test are robust, and they are consistently aligned with the nominal level across all simulated scenarios. Moreover, the Wald-type test is liberal for a small sample $m = (25,25), (25,50)$. To further investigate the performance of the proposed tests, we expand the scope of scenarios by randomly selecting parameters instead of adhering to a predetermined set; 1000 parameter settings are randomly generated, and the corresponding TIEs are calculated simultaneously. The results are presented in a set of box plots (Figure 1). The results show that the likelihood ratio test and the Wald-type test are always liberal when the number of each group is small $m = (25, 25)$. As the number of each group increases, these three tests will become more satisfactory. At the same time, all three tests are always robust for $m = (100, 100)$ and $J = 4, 6$. In particular, the score test always has greater robustness than the likelihood ratio test and the Wald-type test under all conditions. Hence, the score test is recommended.

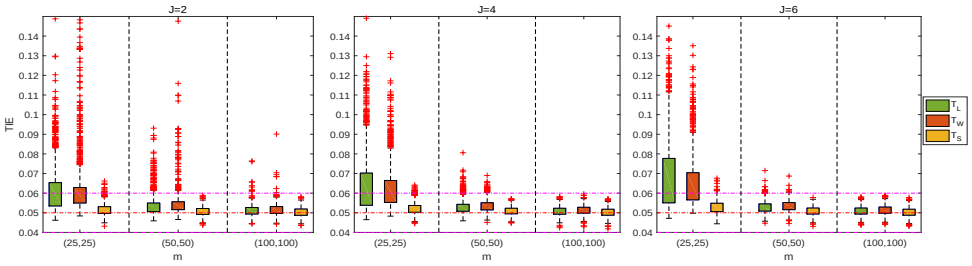


Figure 1. Box plots of empirical TIEs for the likelihood ratio test (T_L), Wald-type test (T_W), and score test (T_S) based on 1000 sets of randomly generated parameter settings when $J = 2, 4, 6$ and $m = (25,25), (50,50), (100,100)$.

Table 3. The empirical type I error rates (%) for testing $H_0 : \delta = \delta_0$ corresponding to nominal α = 5% based on 10,000 replications under $J = 4$.

δ	π_1	γ_1	γ_2	$m = (25, 25)$			$m = (25, 50)$			$m = (50, 50)$			$m = (50, 75)$			$m = (100, 100)$		
				T_L	T_W	T_S	T_L	T_W	T_S	T_L	T_W	T_S	T_L	T_W	T_S	T_L	T_W	T_S
0.5	I	a_1	b_1	5.41	5.92	4.92	5.23	5.52	5.21	5.22	5.54	5.13	5.11	5.36	5.02	5.24	5.62	5.14
			b_2	5.87	**6.03**	5.50	5.46	5.74	5.24	5.23	5.18	5.12	5.08	5.15	5.00	4.95	4.97	4.93
		a_2	b_1	5.77	5.99	5.12	5.46	5.59	5.32	5.15	5.27	4.95	5.07	5.33	5.01	4.83	4.95	4.84
			b_2	5.37	5.65	5.08	4.95	5.19	4.76	5.35	5.67	5.30	5.34	5.43	5.22	4.93	4.99	4.89
	II	a_1	b_1	5.64	5.99	5.26	5.27	5.76	5.14	5.05	5.23	4.98	5.32	5.44	5.19	5.31	5.46	5.24
			b_2	5.42	5.27	5.13	5.11	5.33	5.01	5.09	5.09	4.90	4.72	4.85	4.58	5.16	5.30	5.09
		a_2	b_1	5.18	5.49	4.84	4.67	5.01	4.60	5.13	5.46	4.92	5.51	5.60	5.38	5.26	5.31	5.16
			b_2	5.12	5.72	4.88	5.06	5.32	4.99	5.59	5.65	5.43	5.07	5.33	4.96	5.14	5.39	5.21
0.7	I	a_1	b_1	5.48	5.81	5.28	5.34	5.57	5.25	5.25	5.53	5.18	5.58	5.55	5.53	5.30	5.26	5.26
			b_2	5.00	5.32	4.91	5.05	5.41	4.99	5.02	5.06	4.90	5.47	5.51	5.44	5.21	5.22	5.20
		a_2	b_1	5.32	5.86	5.09	5.22	5.28	5.13	5.03	4.94	5.00	5.10	5.28	5.09	5.28	5.47	5.25
			b_2	5.13	5.48	4.95	5.13	5.43	5.03	5.46	5.53	5.35	4.87	5.04	4.86	5.44	5.50	5.41
	II	a_1	b_1	5.56	**6.16**	5.31	4.89	5.05	4.75	4.98	5.17	4.88	5.39	5.42	5.34	5.21	5.31	5.18
			b_2	4.94	5.01	4.78	5.14	5.32	5.05	5.25	5.15	5.15	5.31	5.47	5.24	5.15	5.23	5.06
		a_2	b_1	5.31	5.70	5.11	5.08	5.18	5.02	5.01	5.11	4.92	5.32	5.56	5.18	4.84	5.23	4.83
			b_2	5.58	5.65	5.39	5.09	5.12	4.99	5.05	5.24	5.03	5.29	5.41	5.18	5.33	5.25	5.27
0.9	I	a_1	b_1	5.10	5.27	4.99	5.46	5.53	5.34	4.99	5.18	4.90	4.80	4.86	4.78	5.22	5.14	5.19
			b_2	5.87	5.99	5.65	4.91	5.11	4.79	5.12	5.35	4.99	5.00	5.21	4.94	5.63	5.50	5.59
		a_2	b_1	5.59	5.74	5.30	5.06	5.14	5.01	5.02	5.11	4.94	4.91	4.90	4.85	5.38	5.38	5.36
			b_2	5.29	5.60	5.18	4.88	5.03	4.75	5.00	5.03	4.87	5.15	5.00	5.12	4.98	5.09	4.91
	II	a_1	b_1	5.60	5.72	5.30	5.39	5.39	5.27	4.94	4.98	4.87	5.16	5.19	5.11	4.90	5.14	4.89
			b_2	5.59	5.85	5.40	5.31	5.24	5.20	4.92	5.01	4.81	4.86	4.87	4.78	4.78	4.73	4.72
		a_2	b_1	5.83	5.96	5.62	5.51	5.47	5.40	5.24	5.37	5.13	5.23	5.30	5.17	4.93	5.05	4.92
			b_2	5.40	5.87	5.19	5.30	5.52	5.22	4.93	5.14	4.83	5.05	5.09	4.98	4.73	4.92	4.71

Note: T_L, T_W, and T_S, respectively, represent the empirical Type I error rates obtained from the likelihood ratio test, Wald-type test, and score test.

Then, we consider the null hypothesis that $\delta = \delta_0 = 1$ across strata, implying that the treatment is not effective in terms of the relative risk ratio. Conversely, 10,000 samples are randomly generated under the alternative hypothesis, selecting $\delta = 0.5, 0.7, 0.9$. The performance of powers is investigated for the proposed test statistics under the same parameter configurations (Table 2). The empirical power is calculated as the proportion of correct rejections of the null hypothesis in 10,000 Monte Carlo repetitions. The results of empirical powers for $J = 2, 4, 6$ are shown in Table 4 and Supplementary Material S2. They reflect that the powers of three test statistics become larger when increasing the number of each group and stratum. The performance of these three tests tends to be similar as sample sizes increase. However, under the same parameter settings, the Wald-type test has greater power than the likelihood ratio test, and the likelihood ratio test has greater power than the score test. It is worth noting that the simulation of empirical powers has significant implications for the sample size determination in practical scenarios. For instance, a sample size of approximately $m = (50, 50)$ (or $m = (25, 25)$) is sufficient to achieve nearly 80% empirical powers when rejecting the null hypothesis that the treatment is not effective, given a true treatment effect of $\delta = 0.7$ and the stratum $J = 4$ (or $J = 2$).

Furthermore, based on the results of empirical powers, it can be observed that the value of empirical powers is significantly influenced by δ. Therefore, we will conduct further research on the relationship between the empirical powers and the parameter δ. Suppose that $\delta = 0.51(0.03)1.1$, i.e., from 0.51 to 1.1 with step size 0.03. Other parameter configurations are $\delta_0 = 1$, $\pi_1 =$ I, $\gamma_1 = a_2$, and $\gamma_2 = b_2$ (Table 2) for strata $J = 2, 4, 6$ and group number $m = (m_{ij}^{(1)}, m_{ij}^{(2)}) = (25, 25), (50, 50), (100, 100)$. For the convenience of representation in the figure, let M1 = (25,25), M2 = (50,50), and M3 = (100,100). The results are presented in Figure 2. It more clearly shows that the empirical powers become larger

as the sample size increases under the same parameter settings. Meanwhile, it indicates that the empirical power decreases as δ approaches 1. The reason is that the rate of type II error will increase when the parameter settings under the alternative hypothesis are closer to the null hypothesis. By analyzing the results for δ values around 0.9, we found that the empirical power of the Wald-type test is significantly greater than those of the likelihood ratio test and the score test. At the same time, the empirical power of the likelihood ratio test is very close to that of the score test.

Table 4. The empirical powers (%) for $J = 4$.

δ	π_1	γ_1	γ_2	$m = (25,25)$			$m = (25,50)$			$m = (50,50)$			$m = (50,75)$			$m = (100,100)$		
				T_L	T_W	T_S	T_L	T_W	T_S	T_L	T_W	T_S	T_L	T_W	T_S	T_L	T_W	T_S
0.5	I	a_1	b_1	99.0	99.6	99.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
			b_2	99.2	99.7	99.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		a_2	b_1	98.8	99.4	98.8	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
			b_2	98.9	99.4	98.9	99.9	99.9	99.9	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0	100.0
	II	a_1	b_1	99.3	99.9	99.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
			b_2	99.1	99.8	99.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		a_2	b_1	99.2	99.7	99.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
			b_2	99.2	99.7	99.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0.7	I	a_1	b_1	82.5	87.2	82.1	95.6	96.9	95.5	98.2	98.7	98.2	99.6	99.8	99.6	100.0	100.0	100.0
			b_2	81.0	85.9	80.5	95.1	96.4	95.0	97.9	98.5	97.8	99.5	99.6	99.5	100.0	100.0	100.0
		a_2	b_1	81.4	86.6	81.0	95.3	96.8	95.2	98.3	98.8	98.2	99.6	99.7	99.6	100.0	100.0	100.0
			b_2	80.3	85.2	80.0	94.5	96.1	94.4	97.9	98.5	97.8	99.6	99.7	99.6	100.0	100.0	100.0
	II	a_1	b_1	89.9	92.9	89.6	98.5	98.9	98.4	99.5	99.7	99.5	99.9	100.0	99.9	100.0	100.0	100.0
			b_2	88.8	92.1	88.6	97.9	98.5	97.9	99.4	99.6	99.4	99.9	99.9	99.9	100.0	100.0	100.0
		a_2	b_1	89.3	92.3	89.1	98.1	98.7	98.1	99.4	99.6	99.4	99.9	100.0	99.9	100.0	100.0	100.0
			b_2	87.9	91.0	87.6	97.8	98.4	97.7	99.3	99.5	99.3	99.9	99.9	99.9	100.0	100.0	100.0
0.9	I	a_1	b_1	16.3	20.8	15.9	23.4	27.7	23.2	27.3	31.8	27.1	33.0	37.7	32.7	46.5	50.6	46.4
			b_2	15.8	20.1	15.5	21.9	26.6	21.7	26.6	30.7	26.2	33.3	37.3	33.1	46.0	49.8	45.8
		a_2	b_1	15.7	20.8	15.3	22.7	27.5	22.4	26.5	31.1	26.2	32.9	37.2	32.6	46.3	50.5	46.1
			b_2	14.6	19.4	14.2	21.6	25.9	21.3	25.3	29.9	25.1	32.8	37.2	32.5	45.6	49.6	45.4
	II	a_1	b_1	19.6	24.4	19.1	28.4	33.0	28.1	33.1	37.3	32.7	41.2	45.3	41.0	56.3	59.8	56.2
			b_2	19.6	24.6	19.0	27.9	31.6	27.5	32.8	36.9	32.5	39.6	43.5	39.3	55.6	58.9	55.4
		a_2	b_1	20.0	24.7	19.4	26.3	31.0	26.0	31.1	35.8	30.8	40.4	44.6	40.2	55.5	58.8	55.3
			b_2	18.4	22.8	17.9	26.5	30.8	26.2	31.1	35.5	30.9	37.7	41.5	37.5	53.1	56.5	52.9

Note: T_L, T_W, and T_S, respectively, represent the empirical powers obtained from the likelihood ratio test, Wald-type test, and score test.

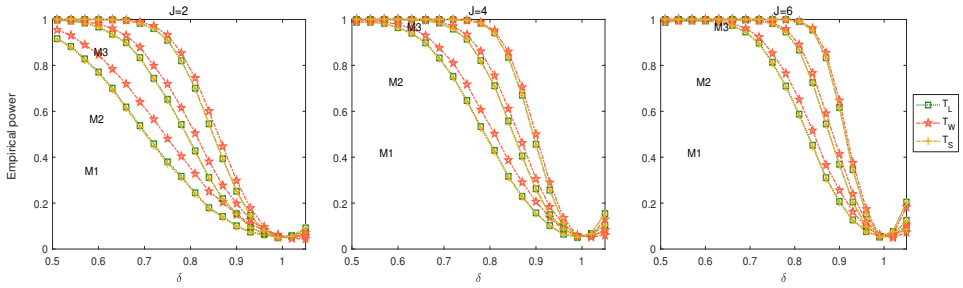


Figure 2. The curve graph of empirical powers for δ, under the parameter settings $\delta_0 = 1$, $\pi_1 = \mathrm{I}$, $\gamma_1 = a_2$, and $\gamma_2 = b_2$. The quantities for each group are selected as M1 = (25, 25), M2 = (50, 50), and M3 = (100, 100).

5.2. Comparison of Confidence Intervals

In this subsection, the performance of four proposed CI methods for the common relative risk ratio is evaluated by using three evaluation indices: the average coverage probability (MCP), average coverage width (MCW), and the ratio of the median non-coverage probability to the non-coverage probability (RMNCP). The MCP represents the proportion of confidence intervals that successfully capture the true parameter value. The MCW is the average of all simulated CI widths. Suppose that the pre-specified confidence level is 0.95. The CI method can be said to be satisfactory, conservative, or liberal if its MCPs are

in the range of (0.94, 0.96), greater than 0.96, or less than 0.94, respectively [28]. Generally, a method with shorter MCW is recommended when MCP is satisfactory. The RMNCP is an index to describe the symmetry of CIs [29]. The interval is classified as satisfactory if it is within $[0.4, 0.6]$. But, it is too medial if the RMNCP is below 0.4 and too distal if it is above 0.6.

In the simulation studies, the parameter settings $\pi_1, \gamma_1, \gamma_2$ are the same as above (Table 2). A total of 10,000 replications are randomly generated under $\delta = 0, 0.1, 0.3$. For $J = 4$, we consider two cases: (i) balanced case, $m = (25, 25), (50, 50), (100, 100)$; (ii) unbalanced case: $m = (25, 50), (50, 75)$. For $m = (50, 50)$, the stratum designs $J = 2, 4, 6, 8$ are studied. As presented in Figure 3, MCPs of these four methods are mostly within the interval [0.94, 0.96] for different sample sizes and strata. Further, CI_W may result in insufficient coverage because most of the MCPs are less than 0.95. CI_L will become satisfactory as the sample size increases. CI_S is always close to the predetermined level 0.95, and it has a better coverage effect compared to other methods. The MCWs of the four methods become shorter, and the differences between them also decrease with the increase in sample size and strata. CI_L, and CI_W have shorter MCWs than CI_S when the sample size and the number of strata are small ($m = (25, 25)$ and $J = 2$). For the RMNCPs, CI_L and CI_S are satisfactory in all cases. Meanwhile, the RMNCPs of CI_L are closer to 0.5. Hence, CI_S is recommended in these cases, and CI_L is more worthy of being considered for use for a smaller sample size and stratum ($m = (25, 25)$ and $J = 2$).

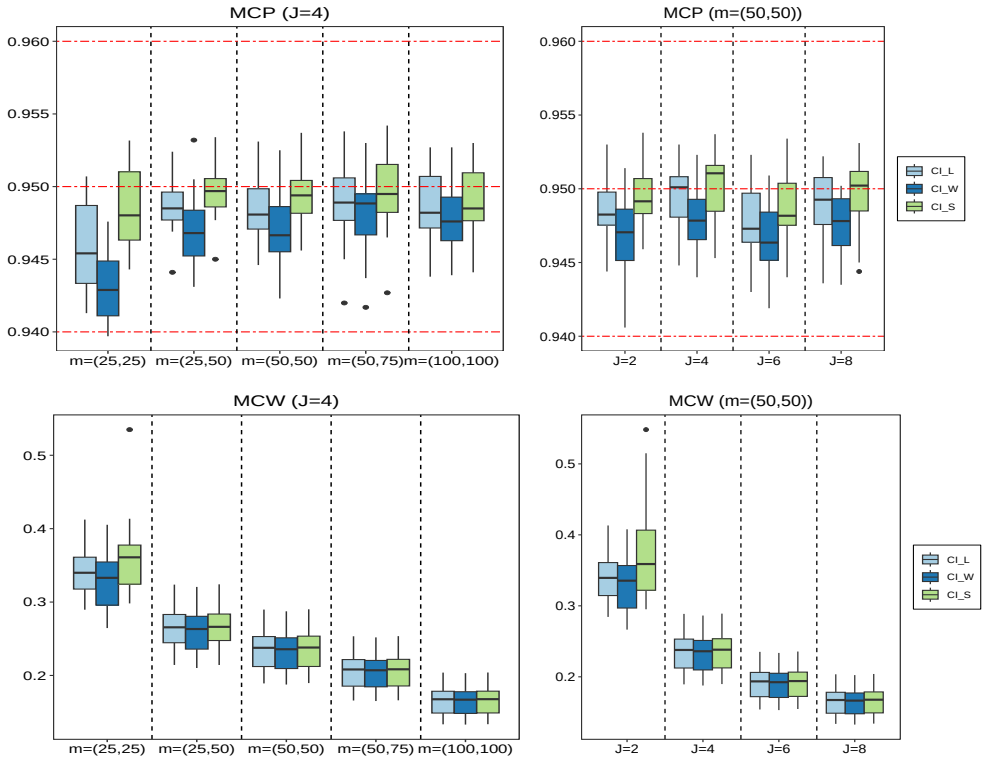
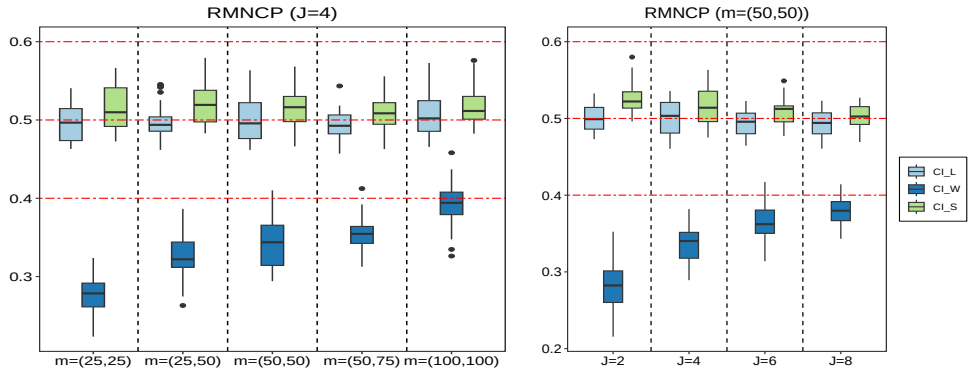


Figure 3. *Cont.*



(**a**) (**b**)

Figure 3. Boxplots of MCPs, MCWs, and RMNCPs: (**a**) $J = 4$ with balance and unbalance cases. (**b**) $m = (50, 50)$ with strata $J = 2, 4, 6, 8$. (Note: the parameter settings are selected from Table 2).

6. Two Real Examples

The first data example comes from a double-blinded randomized clinical trial conducted in the context of acute otitis media with effusion [30]. It can be used to compare the effect of two antibiotics (Cefaclor and Amoxicillin) in the treatment of otitis media with effusion (OME). In total, 203 children, with a total of 293 ears, were randomly assigned to receive a 14-day course of either Amoxicillin or Cefaclor after unilateral or bilateral tympanic membrane perforation. They were divided into three strata ($j = 1, 2, 3$) by age (<2, 2–6, and ≥6 years old), where each stratum is formed by two groups ($i = 1, 2$) with Amoxicillin or Cefaclor. The number of children with OME is listed in Table 5.

Table 5. The number of children after receiving treatment in the age stratums and antibiotic groups.

Age Groups	Responses	Unilateral		Total	Bilateral		Total
		Amoxicillin	Cefaclor		Amoxicillin	Cefaclor	
<2 years	0	10	9	19	11	8	19
	1	2	3	5	2	2	4
	2	-	-	-	2	8	10
2–6 years	0	22	7	29	3	6	9
	1	14	24	38	1	6	7
	2	-	-	-	5	10	15
≥6 years	0	7	8	15	1	0	1
	1	11	11	22	0	1	1
	2	-	-	-	6	3	9

For the first study, the relative risk ratio is $\delta_j = \pi_{2j}/\pi_{1j}$ $(j = 1, 2, 3)$. Suppose that the difference between the two antibiotics is the same at various ages, that is, $\delta_1 = \delta_2 = \delta_3 = \delta$. Then, we assume that $\delta_0 = 1$ under the null hypothesis, which implies that there is no difference in the treatment effects of cefaclor versus amoxicillin across age strata. Through the calculation, we can get the unconstrained MLEs $\hat{\delta} = 1.3451$. The MLEs $\hat{\boldsymbol{\pi}}_1$, $\tilde{\boldsymbol{\pi}}_1$, $\hat{\gamma}_i$, and $\tilde{\gamma}_i$ $(i = 1, 2)$ are summarized in Table 6; for instance, $\hat{\boldsymbol{\pi}}_1 = (0.2583, 0.4887, 0.5433)$.

Table 6. The MLEs of parameters for the otolaryngology study.

Age Groups	Unconstrained MLEs			Constrained MLEs		
	$\hat{\boldsymbol{\pi}}_1$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\tilde{\boldsymbol{\pi}}_1$	$\tilde{\gamma}_1$	$\tilde{\gamma}_2$
<2 years	0.2583	0.7160	0.8566	0.3002	0.7487	0.8475
2–6 years	0.4887	0.8870	0.8052	0.5781	0.9027	0.7627
≥6 years	0.5433	0.9863	0.7792	0.6636	0.9877	0.7666

The otolaryngology section of Table 7 presents the corresponding statistical values and p-values for the first example based on the three proposed tests. The results indicate that all p-values are less than the predefined level $\alpha = 0.05$. Thus, the null hypothesis H_0 is rejected. There exists a significant difference among the age groups for the relative risk ratio. Table 8 provides the four confidence intervals under consideration. The results show that the confidence width of CI_L and CI_S are similar, and they are both smaller than CI_W. In the meantime, the lower limits of the four CI methods are greater than 1. That is to say, the null hypothesis $\delta_1 = \delta_2 = \delta_3 = 0$ should be rejected. This result is consistent with the result of the common test. Therefore, unconstrained MLEs have greater significance for the study of unilateral and bilateral otolaryngology data. Furthermore, based on the unconstrained MLE $\hat{\delta} = 1.3451$, Cefaclor is more effective than Amoxicillin.

Table 7. Statistic values and p-values of common test for two examples.

Test Statistics	Otolaryngology			Myopathy		
	T_L	T_W	T_S	T_L	T_W	T_S
Statistic	5.4845	3.9312	5.3355	2.0679	3.1300	1.9923
p-value	0.0192	0.0474	0.0209	0.1504	0.0769	0.1581

Table 8. The 95% confidence interval estimation of δ with the proposed methods for two examples.

Methods	Otolaryngology		Methods	Myopathy	
	CI	Width		CI	Width
CI_L	(1.0483,1.5649)	0.5166	CI_L	(0.2981,1.2138)	0.9156
CI_W	(1.0040,1.6862)	0.6822	CI_W	(0.0477,1.0487)	1.0010
CI_S	(1.0442,1.5638)	0.5196	CI_S	(0.2333,1.2274)	0.9942

Another example comes from an observational study on myopia patients [21]. Sixty subjects diagnosed with myopia received Orthokeratology (Ortho-k), a non-surgical vision correction technique by wearing overnight specialized contact lenses to reshape the cornea. The benefits of Ortho-k include freedom from glasses during the day, potential myopia control in kids, reversibility if Ortho-k is discontinued, and a non-invasive alternative to surgery. Wearing the lenses on both eyes or a single eye is based on the patient's choice. Corneal refractive therapy (CRT) lenses and vision shaping treatment (VST) lenses are two lens treatment options. The treatment success or improvement is determined by the criterion of axial length growth being $<$0.3 mm [31]. The number of patients with unilateral or bilateral myopia improved eyes is summarized in Table 9. The participants are stratified by gender and grouped by these two lens treatment options. Let $i = 1$ be VST, and $i = 2$ be CRT. We are interested in comparing the effectiveness of treatments between the VST and CRT groups across genders.

Table 9. The patients with myopia improved eyes by receiving either VST or CRT treatment.

Gender Groups	Responses	Unilateral		Total	Bilateral		Total
		VST	CRT		VST	CRT	
Female	0	2	0	2	9	7	16
	1	1	0	1	3	0	3
	2	-	-	-	7	0	7
Male	0	1	0	1	11	6	17
	1	2	0	2	4	2	6
	2	-	-	-	3	2	5

For the second example, the unconstrained MLE of the relative risk ratio is $\hat{\delta} = 0.9375$. Under the null hypothesis $H_0 : \delta = \delta_0$ of the common test, suppose that $\delta_0 = 1$. The MLEs $\hat{\pi}_1$, $\tilde{\pi}_1$, $\hat{\gamma}_i$, and $\tilde{\gamma}_i (i = 1, 2)$ are showed in Table 10. Following the same calculation as above, the testing results in Table 7 show that p-values are greater than $\alpha = 0.05$ under the null hypothesis. Therefore, we have no evidence to reject the null hypothesis H_0. Moreover, Table 8 displays the four confidence interval estimations of δ for the proposed methods. It is found that the upper and lower limits of the four CI methods include 1. Therefore, based on the results of the tests and confidence interval estimations for the relative risk ratio, it can be concluded that there is no significant difference in the effectiveness between the two lens treatment options.

Table 10. The MLEs of parameters for the myoporthosis study.

Gender Groups	Unconstrained MLEs			Constrained MLEs		
	$\hat{\pi}_1$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\tilde{\pi}_1$	$\tilde{\gamma}_1$	$\tilde{\gamma}_2$
Female	0.3922	0.8045	0.9512	0.3430	0.7882	0.9737
Male	0.3619	0.6706	0.5990	0.3145	0.6307	0.6767

7. Conclusions

Under Dallal's model, this article proposes three large-sample tests and three confidence interval methods of relative risk ratio for the stratified unilateral and bilateral data, in order to infer common treatment effects. The Newton–Raphson procedure and Fisher scoring algorithm are applied to calculate the unconstrained and constrained MLEs. The detailed step-by-step process of the research for this article can be found in Appendix C.

Simulation results indicated that three tests would become more satisfactory as the number of each group increases. The sample size was small, with $m = (25, 25)$, and the likelihood ratio test and the Wald-type test were always liberal. The primary reason is that the Wald-type test is more vulnerable because estimates are unstable or convergence is hard to reach. The likelihood ratio test requires fitting two models, while the score test only needs to fit one model. Therefore, the score test demonstrates its stability in terms of calculation when the sample size is small [32]. Hence, the score test was recommended for evaluating treatment effectiveness under a variety of data-generating scenarios. The confidence interval method based on the score test always provided coverage probability close to the nominal level and satisfactory coverage width. This recommendation was based on its three evaluation indices, which demonstrated its robust performance even with small sample sizes. Furthermore, two real-world datasets of acute otitis media and myopic eyes were used to illustrate the proposed tests and confidence interval methods. Through the study of the examples, we found that the common test and the confidence interval method consistently yield similar results. This is not a coincidence, but rather because, on the one hand, the iterative confidence interval method utilizes the results of the common test; on the other hand, both the test and the confidence interval employ the same maximum likelihood estimates.

The contributions made by this work extend beyond the following: (i) Many current studies tend to analyze bilateral data without considering unilateral data. However, unilateral data can be obtained in clinical practice, when a patient has only a single organ affected or undergoing treatment. Our methodologies can be applied not only to the research of bilateral data but also to unidirectional and bilateral data. In our context, the scenario of bilateral data alone constitutes a special case. (ii) In the statistical inference of bilateral data, choosing an appropriate method is of great significance for explaining the interclass correlations within a subject. In the meantime, the relative risk ratio would perform better

than the risk difference for statistical inference when the differences between different datasets are relatively small. Meanwhile, the relative risk ratio has a simpler structure compared to the odds ratio. In this paper, the statistical inference of the relative risk ratio is discussed based on the stratified unilateral and bilateral data. In real data research, we thoroughly combined the results of the common test and confidence interval estimation to analyze the real problem.

Compared with the studies on the risk difference of the stratified unilateral and bilateral data [22,23], we came up with the following two suggestions: (i) When the differences between different datasets are relatively small, we recommend using the relative risk ratio as studied in this article. (ii) The score test always performs satisfactorily in both risk ratio and risk difference studies. In the meantime, the confidence interval method based on the score test is also always recommended.

The proposed methods are asymptotic and can be implemented in the case of a large sample. However, some stratified unilateral and bilateral data may encounter the problem of insufficient sample size in reality. Hence, exact methods will be considered for small-sample cases as an interesting future work. Furthermore, the proposed confidence interval methods involve complex computations. Therefore, the bootstrap confidence intervals can be introduced in our future work.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math14060948/s1, File S1: MATLAB code of common test and interval estimation; File S2: Results of empirical TIEs and empirical powers.

Author Contributions: Conceptualization, investigation, project administration, writing—original draft, S.S.; Writing—review and editing, supervision, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (XK0070922093), the Central Guidance for Local Science and Technology Development Fund (ZYYD2025ZY20), the National Natural Science Foundation of China (12561047), and the Xinjiang Talent Development Fund (XJRC-2025-KJ-PY-KJLJ-108).

Data Availability Statement: The data presented in this study are openly available in [21,30].

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

T_L	Likelihood ratio test
T_W	Wald-type test
T_S	Score test
CI_L	The confidence intervals based on the likelihood ratio test
CI_W	The confidence interval based on the Wald-type test
CI_S	The confidence intervals based on the score test
TIE	Type I error rates
MCP	The average coverage probability
MCW	The average coverage width
$RMNCP$	The ratio of the median non-coverage probability to the non-coverage probability

Appendix A. Information Matrix for Computing MLEs

Differentiating the log-likelihood function (2) with δ, π_{1j}, γ_{1j}, and γ_{2j} are as follows:

$$\frac{\partial l_1}{\partial \delta} = \sum_{j=1}^{J}\left(\frac{m_{12j}^{(1)}+m_{12j}^{(2)}+m_{22j}^{(2)}}{\delta}+\frac{m_{02j}^{(1)}\pi_{1j}}{\pi_{1j}\delta-1}+\frac{m_{02j}^{(2)}\pi_{1j}(\gamma_{2j}-2)}{\pi_{1j}\delta(\gamma_{2j}-2)+1}\right),$$

$$\frac{\partial^2 l_1}{\partial \delta^2} = -\sum_{j=1}^{J}\left(\frac{m_{12j}^{(1)}+m_{12j}^{(2)}+m_{22j}^{(2)}}{\delta^2}+\frac{m_{02j}^{(1)}\pi_{1j}^2}{(\pi_{1j}\delta-1)^2}+\frac{m_{02j}^{(2)}\pi_{1j}^2(\gamma_{2j}-2)^2}{(\pi_{1j}\delta(\gamma_{2j}-2)+1)^2}\right),$$

$$\frac{\partial l_1}{\partial \pi_{1j}} = \frac{m_{1+j}^{(1)}+m_{1+j}^{(2)}+m_{2+j}^{(2)}}{\pi_{1j}}+\frac{m_{01j}^{(1)}}{\pi_{1j}-1}+\frac{m_{02j}^{(1)}\delta}{\pi_{1j}\delta-1}+\frac{m_{01j}^{(2)}(\gamma_{1j}-2)}{\pi_{1j}(\gamma_{1j}-2)+1}+\frac{m_{02j}^{(2)}\delta(\gamma_{2j}-2)}{\pi_{1j}\delta(\gamma_{2j}-2)+1},$$

$$\frac{\partial l_1}{\partial \gamma_{1j}} = \frac{m_{21j}^{(2)}}{\gamma_{1j}}+\frac{m_{11j}^{(2)}}{\gamma_{1j}-1}+\frac{m_{01j}^{(2)}\pi_{1j}}{\pi_{1j}(\gamma_{1j}-2)+1},$$

$$\frac{\partial l_1}{\partial \gamma_{2j}} = \frac{m_{22j}^{(2)}}{\gamma_{2j}}+\frac{m_{12j}^{(2)}}{\gamma_{2j}-1}+\frac{m_{02j}^{(2)}\pi_{1j}\delta}{\pi_{1j}\delta(\gamma_{2j}-2)+1}.$$

The Fisher information matrix I is written as

$$I(\pi_{1j},\gamma_{1j},\gamma_{2j}) = I^j = \begin{bmatrix} I_{11}^j & I_{12}^j & I_{13}^j \\ I_{21}^j & I_{22}^j & I_{23}^j \\ I_{13}^j & I_{23}^j & I_{33}^j \end{bmatrix},$$

where

$$I_{11}^j = \mathrm{E}\left(-\frac{\partial^2 l_1}{\partial \pi_{1j}^2}\right) = \frac{m_{+1j}^{(1)}+m_{+2j}^{(1)}\delta-m_{+1j}^{(2)}(\gamma_{1j}-2)-m_{+2j}^{(2)}(\gamma_{2j}-2)\delta}{\pi_{1j}} - \frac{m_{+1j}^{(1)}}{\pi_{1j}-1}-\frac{m_{+2j}^{(1)}\delta^2}{\pi_{1j}\delta-1}+\frac{m_{+1j}^{(2)}(\gamma_{1j}-2)^2}{\pi_{1j}(\gamma_{1j}-2)+1}+\frac{m_{+2j}^{(2)}(\gamma_{2j}-2)^2\delta^2}{\pi_{1j}(\gamma_{2j}-2)\delta+1},$$

$$I_{12}^j = \mathrm{E}\left(-\frac{\partial^2 l_1}{\partial \pi_{1j}\partial \gamma_{1j}}\right) = -\frac{m_{+1j}^{(2)}}{\pi_{1j}(\gamma_{1j}-2)+1},$$

$$I_{13}^j = \mathrm{E}\left(-\frac{\partial^2 l_1}{\partial \pi_{1j}\partial \gamma_{2j}}\right) = -\frac{m_{+2j}^{(2)}\delta}{\pi_{1j}(\gamma_{2j}-2)\delta+1},$$

$$I_{22}^j = \mathrm{E}\left(-\frac{\partial^2 l_1}{\partial \gamma_{1j}^2}\right) = \frac{m_{+1j}^{(2)}\pi_{1j}}{\gamma_{1j}}-\frac{2m_{+1j}^{(2)}\pi_{1j}}{\gamma_{1j}-1}+\frac{m_{+1j}^{(2)}\pi_{1j}^2}{\pi_{1j}(\gamma_{1j}-2)+1},$$

$$I_{23}^j = \mathrm{E}\left(-\frac{\partial^2 l_1}{\partial \gamma_{1j}\partial \gamma_{2j}}\right) = 0,$$

$$I_{33}^j = \mathrm{E}\left(-\frac{\partial^2 l_1}{\partial \gamma_{2j}^2}\right) = \frac{m_{+2j}^{(2)}\pi_{1j}\delta}{\gamma_{2j}}-\frac{2m_{+2j}^{(2)}\pi_{1j}^2\delta^2}{\gamma_{2j}-1}+\frac{m_{+2j}^{(2)}\pi_{1j}^2}{\pi_{1j}(\gamma_{2j}-2)\delta+1}.$$

Appendix B. Derivation of Wald-Type and Score Statistics

The Fisher information matrix $\mathcal{I}$ is a $(3J+1)\times(3J+1)$ block diagonal matrix, as follows:

$$\mathcal{I} = \begin{bmatrix} A & B \\ B' & C \end{bmatrix},$$

where

$$A = \mathrm{E}\left(-\sum_{j=1}^{J}\frac{\partial^2 l_1}{\partial\delta^2}\right) = \sum_{j=1}^{J}\left(\frac{m_{+2j}^{(1)}\pi_{1j} - m_{+2j}^{(2)}\pi_{1j}(\gamma_{2j}-2)}{\delta} - \frac{m_{+2j}^{(1)}\pi_{1j}^2}{\pi_{1j}\delta - 1} + \frac{m_{+2j}^{(2)}\pi_{1j}^2(\gamma_{2j}-2)^2}{\pi_{1j}\delta(\gamma_{2j}-2)+1}\right),$$

$$B = (b_1, b_2, \ldots, b_J),\ b_j = \left(\mathrm{E}(-\frac{\partial^2 l_1}{\partial\delta\partial\pi_{1_j}}), \mathrm{E}(-\frac{\partial^2 l_1}{\partial\delta\partial\gamma_{1j}}), \mathrm{E}(-\frac{\partial^2 l_1}{\partial\delta\partial\gamma_{2j}})\right),$$

$$\mathrm{E}(-\frac{\partial^2 l_1}{\partial\delta\partial\pi_{1_j}}) = -\frac{m_{+2j}^{(1)}}{\pi_{1j}\delta - 1} - \frac{m_{+2j}^{(2)}(\gamma_{2j}-2)}{\pi_{1j}\delta(\gamma_{2j}-2)+1},$$

$$\mathrm{E}(-\frac{\partial^2 l_1}{\partial\delta\partial\gamma_{1j}}) = 0,\ \mathrm{E}(-\frac{\partial^2 l_1}{\partial\delta\partial\gamma_{2j}}) = -\frac{m_{+2j}^{(2)}\pi_{1j}}{\pi_{1j}\delta(\gamma_{2j}-2)+1},$$

$$C = \mathrm{diag}(I^1, I^2, ..., I^J).$$

Appendix C

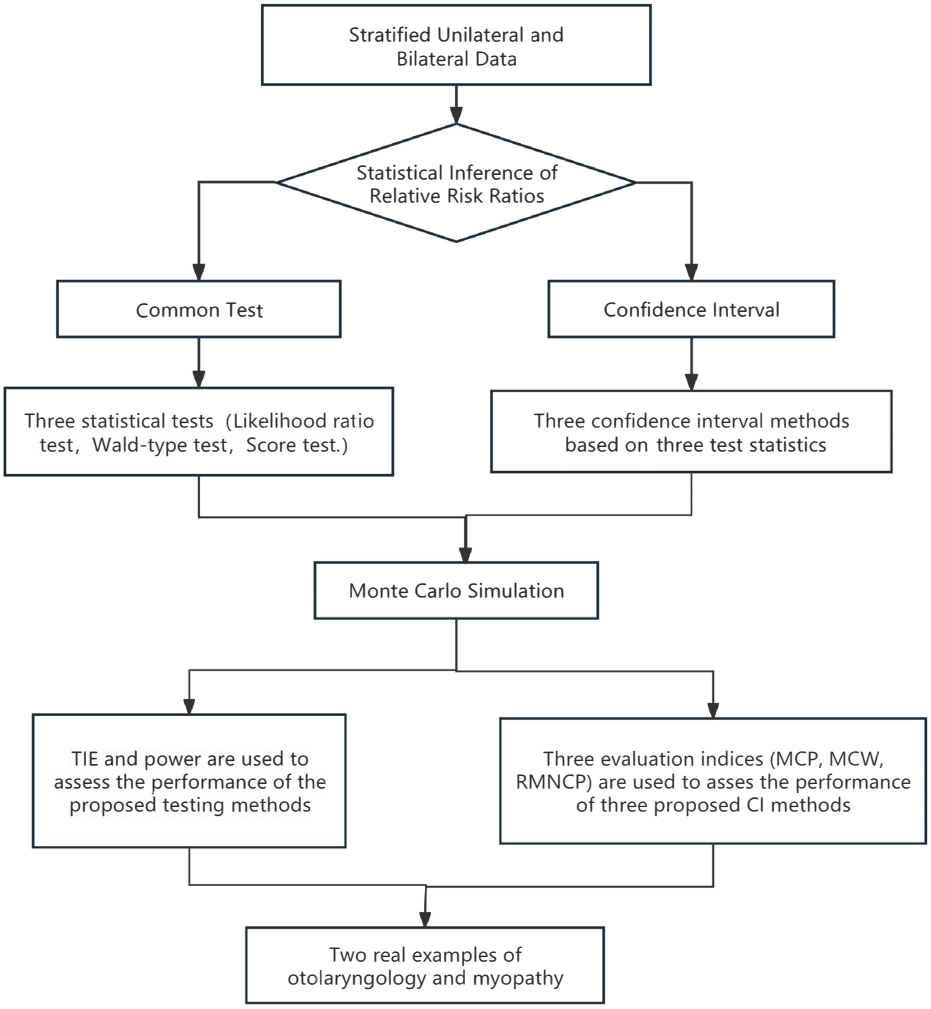


Figure A1. A step-by-step diagram of the analytical process.

References

1. Lui, K.J. A simple test of the homogeneity of risk difference in sparse data: An application to a multicenter study. *Biom. J.* **2005**, *47*, 654–661. [CrossRef] [PubMed]
2. Brown, L.; Li, X.F. Confidence intervals for two sample binomial distribution. *J. Stat. Plan. Inference* **2005**, *130*, 359–375. [CrossRef]
3. Fagerl, M.W.; Lydersen, S.; Laake, P. Recommended confidence intervals for two independent binomial proportions. *Stat. Methods Med. Res.* **2011**, *24*, 224–254. [CrossRef]

4. Gu, J.; Chen, Q.; Zhang, P.; Zhang, T.; Zhou, X.; Zhang, K.; Jiang, T.; Liu, S.; Chen, W.; Zhou, M. Characteristics of vitreoretinal lymphoma in B-scan ultrasonography: A case-control study. *Ophthalmol. Retin.* **2024**, *8*, 264–269. [CrossRef] [PubMed]
5. Bomdica, P.; MacCumber, M.; Abdel-Hadi, S.; Parker, M.; Minaker, S. Surgical outcomes of rhegmatogenous retinal detachment and fellow eye involvement in adolescent and young adult patients. *Ophthalmol. Retin.* **2024**, *8*, 148–154. [CrossRef]
6. Zhang, H.G.; Ying, G.S. Statistical approaches in published ophthalmic clinical science papers: A comparison to statistical practice two decades ago. *Br. J. Ophthalmol.* **2018**, *102*, 1188–1191. [CrossRef]
7. Rosner, B. Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes. *Biometrics* **1982**, *38*, 105–114. [CrossRef]
8. Dallal, G.E. Paired Bernoulli trials. *Biometrics* **1988**, *44*, 253–257. [CrossRef]
9. Donner, A. Statistical methods in opthalmology: An adjusted chi-square approach. *Biometrics* **1989**, *45*, 605–611. [CrossRef]
10. Tang, N.S.; Tang, M.L.; Qiu, S.F. Testing the equality of proportions for correlated otolaryngologic data. *Comput. Stat. Data Anal.* **2008**, *52*, 3719–3729. [CrossRef]
11. Zhuang, T.T.; Tian, G.L.; Ma, C.X. Homogeneity test of ratio of two proportions in stratified bilateral data. *Stat. Biopharm. Res.* **2019**, *11*, 200–209. [CrossRef]
12. Tang, N.S.; Qiu, S.F.; Tang, M.L.; Pei, Y.B. Asymptotic confidence interval construction for proportion difference in medical studies with bilateral data. *Stat. Methods Med. Res.* **2011**, *20*, 233–259. [CrossRef] [PubMed]
13. Shen, X.; Ma, C.X.; Yuen, K.C.; Tian, G.L. Common risk difference test and interval estimation of risk difference for stratified bilateral correlated data. *Stat. Methods Med. Res.* **2018**, *28*, 2418–2438. [CrossRef]
14. Scudamore, I.W.; Dunphy, B.C.; Cooke, I.D. Falloposcopic comparison of unilateral and bilateral proximal tubal ocdusive disease. *Hum. Reprod.* **1994**, *9*, 340–342. [CrossRef]
15. Petrella, M.; Selistre, L.; Serro, P.; Lessi, G.C.; Gonçalves, G.H.; Mattiello, S.M. Kinetics, kinematics, and knee muscle activation during sit to stand transition in unilateral and bilateral knee osteoarthritis. *Gait Posture* **2021**, *86*, 38–44. [CrossRef] [PubMed]
16. Mona, M.; Mustafa, M.; Hadeel, H.; Shehada, R.; Adawi, O.A.; Khzouz, J.; Jaradat, I.; Al-Hussaini, M.; Alnawaiseh, I. The impact of tumor laterality (unilateral vs. bilateral) on presentation and management outcome in patients with retinoblastoma. *J. Clin. Med.* **2024**, *13*, 2146. [CrossRef]
17. Qiu, S.F.; Liu, Q.S.; Ge, Y. Confidence intervals of proportion differences for stratified combined unilateral and bilateral data. *Commun. Stat.-Simul. Comput.* **2021**, *52*, 3839–3862. [CrossRef]
18. Wang, H.P.; Ma, C.X. Homogeneity test of ratios of two proportions in stratified bilateral and unilateral data. *Mathematics* **2024**, *12*, 2719. [CrossRef]
19. Hua, S.C.; Ma, C.X. Testing the homogeneity of odds ratio across strata for combined bilateral and unilateral data. *PLoS ONE* **2024**, *19*, e0307276. [CrossRef]
20. Hua, S.C.; Ma, C.X. Common odds ratio test and interval estimation for stratified bilateral and unilateral data. *Stat. Methods Med. Res.* **2024**, *33*, 1559–1576. [CrossRef]
21. Liang, S.Y.; Fang, K.T.; Huang, X.W.; Xin, Y.J.; Ma, C.X. Homogeneity tests and interval estimations of risk differences for stratified bilateral and unilateral correlated data. *Stat. Pap.* **2024**, *65*, 3499–3543. [CrossRef]
22. Sun, S.M.; Li, Z.M.; Mou, K.Y. Interval estimation of common risk difference for stratified unilateral and bilateral data. *J. Biopharm. Stat.* **2024**, *35*, 85–105. [CrossRef]
23. Sun, S.M.; Li, Z.M.; Jiang, H.J. Homogeneity test and sample size of risk difference for stratified unilateral and bilateral data. *Commun. Stat.-Simul. Comput.* **2024**, *53*, 4209–4232. [CrossRef]
24. Sampson, J.N.; Gail, M.H. Confidence intervals for the difference between two relative risks. *Stat. Methods Med. Res.* **2020**, *29*, 3048–3058. [CrossRef] [PubMed]
25. Xue, Y.Q.; Ma, C.X. Interval estimation of proportion ratios for stratified bilateral correlated binary data. *Stat. Methods Med. Res.* **2019**, *29*, 1987–2014. [CrossRef]
26. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; L. Erlbaum Associates: Mahwah, NJ, USA, 1988.
27. Tang, N.S.; Qiu, S.F. Homogeneity test, sample size determination and interval construction of difference of two proportions in stratified bilateral-sample designs. *J. Stat. Plan. Inference* **2012**, *142*, 1243–1251. [CrossRef]
28. Qiu, S.F.; Guo, L.X.; Zou, G.Y.; Yu, D. Tests for homogeneity of risk differences in stratified design with correlated bilateral data. *J. Appl. Stat.* **2019**, *46*, 2491–2513. [CrossRef]
29. Newcombe, R.G. Measures of location for confidence intervals for proportions. *Commun. Stat. Theory Methods* **2011**, *40*, 1743–1767. [CrossRef]
30. Mandel, E.M.; Bluestone, C.D.; Rockette, H.E.; Blatter, M.M.; Harper, J. Duration of effusion after antibiotic treatment for acute otitis media: Comparison of cefaclor and amoxicillin. *Pediatr. Infect. Dis.* **1982**, *1*, 310–316. [CrossRef]

31. Rose, L.; Schulz, A.; Graham, S. Use baseline axial length measurements in myopic patients to predict the control of myopia with and without atropine 0.01%. *PLoS ONE* **2021**, *16*, e0254061. [CrossRef]
32. Molenberghs, G.; Verbeke, G. Likelihood ratio, score, and wald tests in a constrained parameter space. *Am. Stat.* **2007**, *61*, 22–27. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.