




# New Confidence Intervals for Relative Risk of Two Correlated Proportions

Natalie DelRocco<sup>1</sup> · Yipeng Wang<sup>1</sup> · Dongyuan Wu<sup>1</sup> · Yuting Yang<sup>1</sup> · Guogen Shan<sup>1,2</sup> 

Received: 15 December 2021 / Revised: 17 March 2022 / Accepted: 23 April 2022  
© The Author(s) under exclusive licence to International Chinese Statistical Association 2022

## Abstract

Biomedical studies, such as clinical trials, often require the comparison of measurements from two correlated tests in which each unit of observation is associated with a binary outcome of interest via relative risk. The associated confidence interval is crucial because it provides an appreciation of the spectrum of possible values, allowing for a more robust interpretation of relative risk. Of the available confidence interval methods for relative risk, the asymptotic score interval is the most widely recommended for practical use. We propose a modified score interval for relative risk and we also extend an existing nonparametric U-statistic-based confidence interval to relative risk. In addition, we theoretically prove that the original asymptotic score interval is equivalent to the constrained maximum likelihood-based interval proposed by Nam and Blackwelder. Two clinically relevant oncology trials are used to demonstrate the real-world performance of our methods. The finite sample properties of the new approaches, the current standard of practice, and other alternatives are studied via extensive simulation studies. We show that, as the strength of correlation increases, when the sample size is not too large the new score-based intervals outperform the existing intervals in terms of coverage probability. Moreover, our results indicate that the new nonparametric interval provides the coverage that most consistently meets or exceeds the nominal coverage probability.

**Keywords** Confidence interval · Continuity correction · Nonparametric method · Paired data · Relative risk · Wilson score interval

---

Natalie DelRocco, Yipeng Wang, Dongyuan Wu and Yuting Yang have contributed equally on this project.

---

✉ Guogen Shan  
gshan@ufl.edu

<sup>1</sup> Department of Biostatistics, University of Florida, Gainesville, FL 32603, USA

<sup>2</sup> Department of Biostatistics, University of Florida, Gainesville, FL, USA

## 1 Introduction

Correlated samples occur in many situations such as pre- and post-studies, cross-over studies, natural pairings (i.e., twin studies), and matched-pairs designs [1]. The degree to which the two samples are correlated must be appropriately accounted for in statistical inference of the interested parameter. In the case of binary data, this design can therefore be parametrized by the probability of observing the primary outcome for each test and the correlation between the two tests (test 1 and test 2). The research question of interest is then investigated by comparing the proportion of responses in the two groups by conducting a statistical hypothesis test for an appropriate, clinically relevant parameter [2–5].

A classic example of such experiments is a study where an outcome is measured on a single sample before and after an exposure of interest. The study by Okely et al. [6] investigates the changes in physical activity among older adults in the 1936 Lothian Birth Cohort due to the public health lockdown due to COVID-19 in Scotland [7]. A total of 137 adults over the age of 75 were surveyed about their physical activity, sleep quality, social activity, and psychological state before and after the national lockdown. The question of whether the proportion of respondents reporting low physical activity after the lockdown compared to before was of interest. As the pre-lockdown and post-lockdown measurements are taken on the same observational units, and the outcome of interest is binary (low physical activity), this is an important experiment comparing correlated binary proportions [8–10].

The most notable parameters in this setting are the risk difference (RD), the relative risk/risk ratio (RR), and the odds ratio [11, 12]. In practice, often the response rates are small, such that a relative measure provides more information about the magnitude of the association. Therefore, RR is often used because it is the ratio of respective binomial proportions from the two tests. In this article, we focus on the RR in the particular situation when correlated binary outcomes are reported.

Of equal importance to the point estimate of RR is the corresponding interval estimate. This is most often represented by the frequentist confidence interval (CI) [7, 13, 14]. The prominent frequentist parametric CI methods for correlated RR which are studied in this paper can be broadly classified according to two factors: (1) Hybrid vs. Nonhybrid intervals and (2) Wald-based vs. Score-based intervals. Hybrid refers to obtaining an interval for each proportion separately, then combining them to obtain a single CI. Wald-based and Score-based intervals refer to the associated test statistic which is inverted [15] as the basis of the method. That is, to obtain a Wald-based confidence interval, one should invert the Wald test for the relative risk. Contrastingly, to obtain a Score-based confidence interval, one inverts the Score test. It is of note that there are likelihood-based methods which are based on inverting the likelihood ratio test (LRT). However, we found that Wald-based and Score-based intervals are more frequently used in practice. Therefore, we do not include likelihood-based intervals in this study but direct the interested reader to the literature [3, 16–18].

The simplest form of the frequentist CI for RR is the Wald asymptotic CI, which is a nonhybrid interval derived on the log scale under the assumption of

joint asymptotic normality of the response rates for test 1 and test 2 [19]. The asymptotic normal approximation is not always appropriate given the discrete nature of binomial data and has previously been shown to yield low coverage probabilities compared to other methods [2, 20–22]. It is possible to apply a continuity correction to the upper and lower limits of the Wald interval which is inversely proportional to the study sample size. However, the continuity-corrected Wald interval has been found to yield poor average coverage probabilities compared to alternatives and is consequently not recommended for use in practice [3].

As an improvement to the traditional Wald interval for a single proportion [23, 24], Wilson [25] proposed a score-based CI for binomial proportions. Employing Wilson's approach in a hybrid manner is the basis for most of the improved CIs for the ratio of two correlated proportions described below. Nam and Blackwelder proposed one of the first improved CIs for RR [18], which was initially solved iteratively and later extended to closed-form confidence limits [26]. This interval improves the Wald interval by employing constrained maximum likelihood estimates of the proportions of interest to develop a Fieller-type CI for RR [18]. This was closely followed by Bonett and Price who combined two Wilson score intervals into a hybrid Fieller-type interval for RR [27]. A continuity-corrected version of the Bonett–Price interval is obtained by applying a constant penalty to the individual upper and lower confidence limits for each individual proportion. However, the continuity-corrected Bonett–Price interval has been shown to be overly conservative [2].

Tang et al. [28] recently extended the score CI to the ratio of two dependent proportions by reparametrizing the multinomial probability model and deriving the corresponding score test statistic. Fagerland et al. [2] showed via simulation studies that the performance of the Nam–Blackwelder interval was nearly identical to Tang's asymptotic score interval in terms of coverage probability. Finally, the MOVER hybrid score CI is another hybrid Fieller-type CI, which differs primarily from the Bonett–Price interval in that it is based on Newcombe's "Square and Add" method for the difference in two proportions [3]. Donner and Zou [29] recently extended the original MOVER interval to the ratio of two correlated proportions with acceptable performance on average coverage probability [2].

Thus far, Tang's Score interval is the most frequently used CI for the ratio of two correlated proportions, and is recommended along with the Bonett–Price interval in simulation studies for general use [2, 30]. The proposed approach of Tang et al. [28] relies on the asymptotic properties of the score test statistic, namely, the asymptotic standard normal distribution of the score test statistic conditional on a given RR. In general, continuity corrections serve to penalize the width of an interval for using a continuous function to approximate a discrete function, which may not be reasonable in small samples [31]. Continuity corrections applied to approximate CIs seek to approach the coverage of exact CIs by imposing more conservative confidence limits [19]. In the case of the normal approximation to the binomial distribution, the approximation with continuity correction is typically superior to that without [15]. Therefore, we propose a new confidence interval which imposes a continuity correction to Tang's existing Score interval, which is a nonhybrid score-based CI for the ratio of two correlated proportions.

It can be advantageous in certain situations not to make assumptions regarding the asymptotic distribution of an estimator, or even to rely on asymptotic sample sizes in practice. In such situations, nonparametric CIs may be considered. Nonparametric CIs for paired binary data do not rely on particular asymptotic distributions (e.g., normal distribution) in the derivation of the estimate or the variance–covariance matrix. Duan et al. [32] developed a nonparametric U-statistic-based CI for the difference of two correlated proportions. To our knowledge, no one has developed a nonparametric CI specifically for the ratio of two correlated proportions. In Sect. 2, we further extend the work of Duan et al. to the situation where the RR is the parameter of interest.

In Sect. 2.1, we explicitly define the statistical notations and the model for correlated two-test experimental designs with binary outcomes. In Sect. 2.2, we present the existing CI methods discussed above. In Sect. 2.3, we introduce two new methods to calculate CIs for the correlated RR. Specifically, in Sect. 2.3.1, we propose a new test based on Tang’s score method in conjunction with continuity correction and provide a closed-form solution for ease of implementation. In Sect. 2.3.2, we extend Duan’s nonparametric U-Statistic-based CI for correlated RD to correlated RR. In Sect. 3, the results of using the CIs in Sects. 2.2 and 2.3 are presented. We apply the CIs to real data from two case studies: one oncology trial used by existing methodological reviews [2] to validate our methods and another recent oncology trial. We additionally compare the proposed CI methods with the existing CIs through extensive simulation studies with regard to coverage probability, interval length, mean squared error, and the proportion of configurations above the nominal coverage level. Finally, we summarize our findings and discuss conclusions in the last Section.

## 2 Methodology

### 2.1 Model and Notation

The general setup for the ratio of correlated proportions is introduced below. A standard two-by-two contingency table layout [19], often arising from a matched-pairs study design, is given in Table 1. Let  $X_1$  and  $X_2$  be the binary outcome from test 1 and test 2 with marginal probabilities of success  $p_1$  and  $p_2$ , respectively. We denote  $X_t = 1$  for event and  $X_t = 2$  for nonevent,  $t = 1, 2$ . Table 1 presents the observed frequency and the corresponding probability  $x_{ij}$  and  $p_{ij}$  for participants

**Table 1** Observed counts of the paired study are shown with the corresponding probabilities in parentheses

Test 1	Test 2		
	Event	Nonevent	Total
Event	$x_{11} (p_{11})$	$x_{12} (p_{12})$	$x_{11} + x_{12} (p_1)$
Nonevent	$x_{21} (p_{21})$	$x_{22} (p_{22})$	$x_{21} + x_{22} (1 - p_1)$
Total	$x_{11} + x_{21} (p_2)$	$x_{12} + x_{22} (1 - p_2)$	$n$

having the outcome  $X_1 = i$  and  $X_2 = j$ , where  $i, j = 1, 2$ . The total sample size is  $n = \sum_{i=1}^2 \sum_{j=1}^2 x_{ij}$ . The parameter of interest is the relative risk between test 1 and test 2:

$$\theta_{RR} = \frac{p_1}{p_2}.$$

We focus on the confidence intervals for  $\theta_{RR}$  in this article.

The data vector  $\mathbf{x} = \{x_{ij} : i, j = 1, 2\}$  follows a multinomial distribution with the probability density function:

$$\Pr(\mathbf{x}|\mathbf{p}, n) = \frac{n!}{x_{11}!x_{12}!x_{21}!x_{22}!} p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}} p_{22}^{x_{22}}, \tag{1}$$

where  $\mathbf{p} = \{p_{ij} \in [0, 1] : i, j = 1, 2\}$  such that  $\sum_{i=1}^2 \sum_{j=1}^2 p_{ij} = 1$ . For a pair of Bernoulli random variables  $(X_1, X_2)$ , the Pearson correlation coefficient  $\rho$  is given by

$$\rho = \frac{p_{11} - p_1 p_2}{\sqrt{p_1(1 - p_1)p_2(1 - p_2)}},$$

where  $p_1, p_2 \neq 0, 1$ . We can re-parameterize  $\mathbf{p}$  into the equivalent parameter set  $\{p_1, p_2, \rho\}$ , but there are restrictions for the values of  $\{p_1, p_2, \rho\}$ , not every arbitrary combination is feasible because the natural range of the probability is between 0 and 1 ( $p_{ij} \in [0, 1]$ ). Using the necessary conditions of pairwise probabilities for  $p_{11}$  [33], we have the inequality,

$$\max(0, p_1 + p_2 - 1) \leq \rho \sqrt{p_1(1 - p_1)p_2(1 - p_2)} + p_1 p_2 \leq \min(p_1, p_2).$$

Given the values of  $\{p_1, p_2\}$ , the upper and lower bounds of  $\rho$  are obtained by solving the above inequality, see Appendix 1 for the detailed formulas for the upper and lower bounds of  $\rho$  ( $L_\rho$  and  $U_\rho$ ).

Conditional on  $n$ , a multinomial distribution is defined by the parameter set  $\{p_1, p_2, \rho\}$ , which satisfies  $p_1, p_2 \neq 0, 1$  and  $\rho$  is in  $[L_\rho, U_\rho]$ . Given the relative risk  $\theta_{RR} = p_1/p_2$  and a fixed  $n$ , the parameter set  $\{p_2, \theta_{RR}, \rho\}$  can also specify a multinomial distribution because  $p_1 = p_2 \theta_{RR}$ . The maximum likelihood method is used to estimate probabilities:  $\hat{p}_{ij} = x_{ij}/n$  for  $i, j = 1, 2$ ,  $\hat{p}_1 = (x_{11} + x_{12})/n$ ,  $\hat{p}_2 = (x_{11} + x_{21})/n$ , and  $\hat{\theta}_{RR} = \hat{p}_1/\hat{p}_2$ .

In the next Sect. 2.2, we present four existing methods to calculate the  $100(1 - \alpha)\%$  CI for  $\theta_{RR}$  (e.g., a 95% CI when  $\alpha = 0.05$ ). In Sect. 2.3, we consider three strengths of continuity correction score intervals and one nonparametric interval.

## 2.2 Existing Interval Estimation Methods

### 2.2.1 Wald CI

The asymptotic Wald CI is constructed assuming that the joint sampling distribution of the two sample proportions  $\hat{p}_1$  and  $\hat{p}_2$  is reasonably approximated by a bivariate normal

distribution when  $n$  is sufficiently large. Under this assumption, using the Delta method [19] to obtain the asymptotic variance of  $\log(\hat{\theta}_{RR})$  gives us the following  $100(1 - \alpha)\%$  Wald CI for  $\theta_{RR}$ :

$$\exp \left\{ \log(\hat{\theta}_{RR}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{x_{12} + x_{21}}{(x_{11} + x_{21})(x_{11} + x_{12})}} \right\}, \tag{2}$$

where  $z_{1-\frac{\alpha}{2}}$  is the upper  $\frac{\alpha}{2}$  quantile of the standard normal distribution. This interval is strictly positive and asymmetric.

### 2.2.2 Bonett–Price CI

The Bonett–Price CI is a hybrid-type CI [27]. Two individual one-sample Wilson score intervals  $(L_1, U_1)$  and  $(L_2, U_2)$  are calculated for  $p_1$  and  $p_2$ . Bonett and Price [27] showed that for the ratio of two proportions with  $100(1 - \alpha)\%$  Wilson CIs  $(L_1, U_1)$  and  $(L_2, U_2)$ , the  $100(1 - \alpha)\%$  CI for  $\theta_{RR}$  is given by

$$\left( \frac{L_1}{U_2}, \frac{U_1}{L_2} \right), \tag{3}$$

where

$$(L_1, U_1) = \frac{2(x_{11} + x_{12}) + \psi^2 \pm \psi \sqrt{\psi^2 + 4(x_{11} + x_{12}) \left(1 - \frac{x_{11} + x_{12}}{n'}\right)}}{2(n' + \psi)}, \tag{4}$$

$$(L_2, U_2) = \frac{2(x_{11} + x_{21}) + \psi^2 \pm \psi \sqrt{\psi^2 + 4(x_{11} + x_{21}) \left(1 - \frac{x_{11} + x_{21}}{n'}\right)}}{2(n' + \psi^2)}, \tag{5}$$

$n' = x_{11} + x_{12} + x_{21}$ , and  $\psi$  is a function of  $z_{1-\alpha/2}$  and  $\mathbf{x}$  [27].

### 2.2.3 MOVER Wilson score CI

The lower and upper limits for the CI of the ratio may be solved for in terms of the lower and upper limits of the individual CIs for  $p_1$  and  $p_2$  by noting that  $p_1 - p_2\theta_{RR} = 0$ . Based on this relationship, Donner and Zou [29] applied the original square and add method to show that the closed form  $100(1 - \alpha)\%$  MOVER confidence limits for  $\theta_{RR}$  are given by  $(L, U)$ :

$$\begin{aligned}
 L &= \frac{\hat{p}_1\hat{p}_2 - C_L - \sqrt{(\hat{p}_1\hat{p}_2 - C_L)^2 - L_1U_2(2\hat{p}_1 - L_1)(2\hat{p}_2 - U_2)}}{U_2(2\hat{p}_2 - U_2)}, \\
 U &= \frac{\hat{p}_1\hat{p}_2 - C_U + \sqrt{(\hat{p}_1\hat{p}_2 - C_U)^2 - U_1L_2(2\hat{p}_1 - U_1)(2\hat{p}_2 - L_2)}}{L_2(2\hat{p}_2 - L_2)},
 \end{aligned} \tag{6}$$

where  $C_L = r(\hat{p}_1 - L_1)(U_2 - \hat{p}_2)$ ,  $C_U = r(U_1 - \hat{p}_1)(\hat{p}_2 - L_2)$ ,  $(L_1, U_1)$  and  $(L_2, U_2)$  are individual  $100(1 - \alpha)\%$  CIs of choice for  $p_1$  and  $p_2$  in Eqs. (4) and (5), and  $r = \widehat{\text{corr}}(\hat{p}_1, \hat{p}_2)$  is an appropriate estimate of the correlation between the two sample proportions (e.g., the Pearson correlation coefficient).

### 2.2.4 Score Asymptotic CI

Under the null hypothesis  $H_0 : \theta_{RR} = \theta_0$ , the score test statistic by Tang et al. [30] is given as

$$S(\theta_0) = \frac{(x_{11} + x_{12}) - (x_{11} + x_{21})\theta_0}{\sqrt{n(1 + \theta_0)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta_0 - 1)}}, \tag{7}$$

where

$$\begin{aligned}
 \tilde{p}_{21} &= (-b + \sqrt{b^2 - 4ac})/2a, \\
 a &= n(1 + \theta_0), \\
 b &= (x_{21} + x_{11})\theta_0^2 - (x_{11} + x_{12} + 2x_{21}), \\
 c &= x_{21}(1 - \theta_0)(x_{11} + x_{21} + x_{12})/n.
 \end{aligned}$$

The score test statistic is asymptotically normal under the null hypothesis. In the previous publications [30], due to symmetry, the lower and upper  $100(1 - \alpha)\%$  confidence limits are found iteratively as the roots of

$$S(\theta) = \pm z_{1-\frac{\alpha}{2}}.$$

### CI by Nam and Blackwelder

Nam and Blackwelder [18] derived the constrained maximum likelihood estimates of  $p_{12}$  and  $p_{21}$  as

$$\hat{p}_{12} = \frac{-(x_{11} + x_{12}) + \theta_0^2(x_{11} + x_{21} + 2x_{12}) + \sqrt{[(x_{11} + x_{12}) - \theta_0^2(x_{11} + x_{21})]^2 + 4\theta_0^2x_{12}x_{21}}}{2n\theta_0(\theta_0 + 1)},$$

and

$$\hat{p}_{21} = \theta_0\hat{p}_{12} - (\theta_0 - 1)\left(1 - \frac{x_{22}}{n}\right).$$

Based on the Fieller-type statistic,

$$T(\theta_0) = \frac{\sqrt{n}[(x_{11} + x_{12}) - \theta_0(x_{11} + x_{21})]}{n\sqrt{\theta_0(\hat{p}_{12} + \hat{p}_{21})}}, \tag{8}$$

Nam and Blackwelder construct a  $100(1 - \alpha)\%$  CI for  $\theta_{RR}$  by solving  $T(\theta) = \pm z_{\alpha/2}$  to obtain the confidence interval.

**Theorem 1** *The Nam–Blackwelder CI based on the test statistic  $T(\theta)$  in Eq. (8) is equivalent to the score interval by Tang et al. [30] based on the test statistic  $S(\theta)$  in Eq. (7).*

**Proof** We show that  $T(\theta) = S(\theta)$  for any observed data in Appendix 2. □

Later, Nam [26] derived the closed-form estimation of the Nam–Blackwelder interval using Ferrari’s formulation. Following his approach, we directly derive the closed-form solutions for the score interval in Appendix 3. The closed-form solutions are computationally less intensive and hence faster than iterative methods. Additionally, the availability of a closed-form solution avoids the common issues that befall root-finding algorithms. This prevents the need to choose an optimization method (or to rely on the default implementation which may not be optimal), specify convergence criteria, diagnose possible failure to converge in extreme contingency tables, etc. The closed form, noniterative solution is also more accessible for clinical researchers who may not have experience with iterative computational algorithms.

### 2.3 Proposed Interval Estimation Methods

#### 2.3.1 Continuity-corrected Score Asymptotic CI

Adding continuity correction to the asymptotic score test statistic in Eq. (7), we have the asymptotic score continuity-corrected (ASCC) test statistic as

$$S_\delta(\theta_0) = \frac{|(x_{11} + x_{12}) - (x_{11} + x_{21})\theta_0| - (\frac{1}{\delta})(\frac{1}{n})(x_{11} + x_{21})}{\sqrt{n(1 + \theta_0)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta_0 - 1)}}, \tag{9}$$

where  $\tilde{p}_{21}$  is obtained as in Sect. 2.2.4, and  $\delta$  is a continuity correction value. For the upper confidence limit  $\theta_U$ , we transform

$$\frac{(x_{11} + x_{21})\theta - \left[ (x_{11} + x_{12}) + \frac{1}{\delta n}(x_{11} + x_{21}) \right]}{\sqrt{n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)}} = z_{1-\frac{\alpha}{2}}$$

into a quartic equation. Then, using the same process in the uncorrected case to calculate  $\theta_U$ , the lower confidence limit  $\theta_L$  is calculated in the same way by transforming



$$\frac{(x_{11} + x_{12}) - \left[ (x_{11} + x_{21})\theta + \frac{1}{\delta n}(x_{11} + x_{21}) \right]}{\sqrt{n(1 + \theta)\hat{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)}} = z_{1-\frac{\alpha}{2}}$$

into a quartic equation. So, the  $100(1 - \alpha)\%$  CI of  $\theta_{RR}$  based on the continuity correction is then obtained as  $(\theta_L, \theta_U)$ .

In the traditional continuity correction interval calculation, the value of  $\delta$  is chosen as 2. In this article, we consider continuity corrections of varying strength. We take  $\delta = 2, 4, 8$  for continuity corrections of high, medium, and low strength, respectively. A smaller value of  $\delta$  corresponds to a stronger continuity correction. Therefore, this is denoted as ‘‘ASCC-H’’ when  $\delta = 2$ , ‘‘ASCC-M’’ when  $\delta = 4$ , and ‘‘ASCC-L’’ when  $\delta = 8$  in the presentation of results.

In addition to being computationally fast and intuitive for a clinical audience due to the explicit representation of the CI rather than a set of iterative equations (see Appendix 3 for closed-form solutions), the incorporation of a continuity correction in the score interval allows the flexibility to be more conservative when necessary. That is, our proposed modification to the asymptotic score interval allows the researcher to easily choose the strength of penalty to the confidence limits guided by the estimates of the effect sizes, correlation between measurements from the two tests, and study sample size. General recommendations are provided in the Sect. 4. Because the score interval is recommended for practical use due to stability and desirable properties in recent simulation studies [2, 30], providing a flexible modification to the score interval which may be easily employed under the conditions discussed below can help researchers maintain nominal coverage.

### 2.3.2 Nonparametric CI

Duan et al. [32] proposed a nonparametric CI for the difference in two correlated proportions derived from the rank-based estimation procedures for correlated area under the curve (AUC) data outlined by Lang [17]. Duan et al. relied on the well-known representation of AUC as a U-Statistic [34] in the context of ROC curve analysis. They derived a general covariance matrix for a study with multiple groups. Following DeLong et al. [35], Duan et al. derive the variance–covariance matrix for  $(\hat{p}_1 \hat{p}_2)^T$  when the components of interest are  $(p_1 p_2)^T$ . The full expression is given in Appendix 4. An estimate of the asymptotic variance of a parameter of choice is then easily obtained via the Delta method [19]. We refer the interested reader to Duan et al. [32] for further details when the parameter of interest is the correlated RD.

We extend this line of work by deriving the asymptotic variance for the U-statistic-based estimate of the ratio of two correlated proportions. The derivation of the estimates of the individual proportions follows the same reasoning as outlined above. Since the parameter of interest is now  $\theta_{RR} = p_1/p_2$ , then in the case of two groups the estimate of interest is calculated as the ratio of respective sample proportions,  $\hat{\theta}_{RR} = \hat{p}_1/\hat{p}_2$ . We then apply the Delta method to obtain the asymptotic variance estimate of  $\theta_{RR}$ . Full derivation details can be found in Appendix 4. The corresponding asymptotic  $100(1 - \alpha)\%$  CI using a conservative  $t$  approximation for the degrees of freedom is then given by

$$\exp \left\{ \log(\hat{\theta}_{RR}) \pm t_{n-1, 1-\alpha/2} \sqrt{\hat{V}_{LRR}} \right\},$$

where  $\hat{V}_{LRR} = \widehat{\text{Var}}[\log(\hat{\theta}_{RR})]$  as derived by the Delta method. We refer this nonparametric interval as the NonP interval.

The U-statistic-based approach to CI construction for correlated RD proposed by Duan et al. is computationally fast with a simple closed-form expression. Additionally, nonparametric CIs are beneficial in studies under which the asymptotic approximations of the previously presented methods do not hold or if the underlying data do not follow a binomial distribution. Our extension of this method by derivation of the CI for the correlated RR makes these desirable properties accessible for studies where the RR is the primary measure of interest.

### 3 Results

We first conduct extensive simulation studies comparing the finite sample properties of the four proposed intervals and another four existing intervals, followed by application of the intervals to two real cancer trials. These 8 intervals are computed from the six methods: (1) the asymptotic Wald CI, (2) the Bonett–Price CI, (3) the CI based on the MOVER hybrid score, (4) Tang’s asymptotic score CI, (5) the proposed asymptotic continuity-corrected score CI, and (6) the extended Duan’s nonparametric CI.

#### 3.1 Simulation Results

All simulations were conducted using R Statistical Software Version 4.0 [36]. Functions from the *ratesci* package in R [37] and a R function from Duan et al. [32] were adapted to obtain existing and proposed confidence intervals. Data were simulated according to the probability model described in Eq. (1). Individual contingency tables for each of  $B = 20,000$  Monte Carlo simulations were randomly generated using the *simstudy* package by R [38]. The simulation was parametrized in terms of sample size  $n$ , correlation  $\rho$  between two tests, relative risk  $\theta_{RR}$ , and  $p_2$ . As opposed to specifying  $p_1$  and  $p_2$ , this allows us to directly display the strength of association, making the results more intuitive. We studied the combinations of the following sets of parameter values:  $n \in \{15, 30, 60, 100, 150, 300, 500\}$ ,  $\theta_{RR} \in \{1, 1.5, 2, 3, 4, 5\}$ ,  $\rho \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ , and  $p_2$  from 0.1 to 0.95 by increments of 0.05.

Given each configuration for  $p_2$  and  $\theta_{RR}$ , we have a specific bound for  $\rho$  (see details in Appendix 1). We remove the following three cases which cause undefined confidence intervals for the existing methods: (1)  $n = 15$  and  $p_2 < 0.3$ ; (2)  $n = 30$  and  $p_2 < 0.2$ ; (3)  $n = 60$  and  $p_2 < 0.1$ . Based on these skip rules, we have 2,235 combinations of parameters in this simulation (referred to the parameter space as  $\Omega$ ). In addition, if the simulated data satisfied  $x_{11} + x_{12} = 0$  and  $x_{11} + x_{21} = 0$

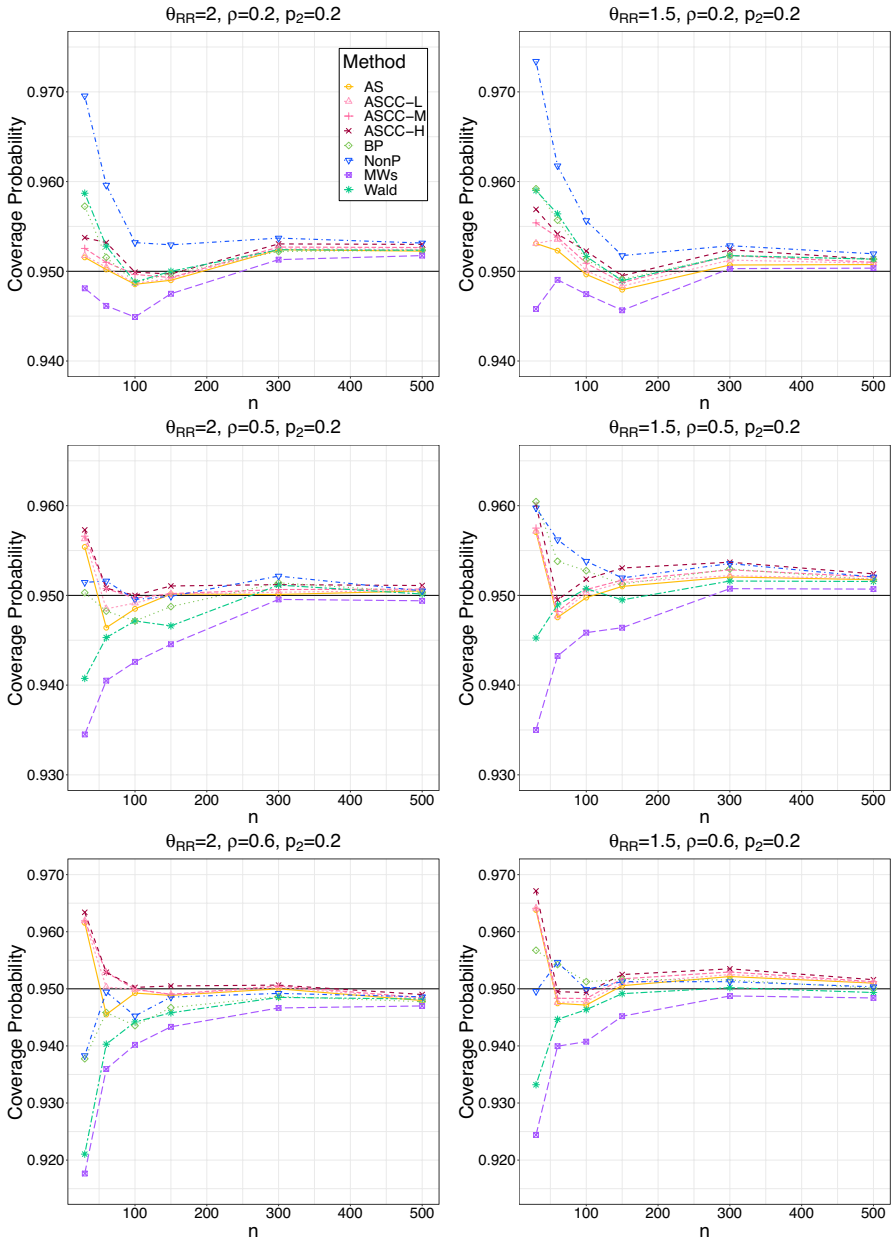
simultaneously, the case was discarded and new data were generated for this combination until we obtained 20,000 cases that can be used.

Figure 1 presents an example of the performance of the Wald interval when  $p_2 = 0.2$  and  $\theta_{RR} = 1.5, 2$ . When  $\theta_{RR} = 1.5$  and  $\rho = 0.2$ , the coverage probability of the score-based intervals starts off above 95% when  $n = 30$  (95.31%, 95.31%, 95.54%, and 95.69% for uncorrected, low, medium, and high corrections, respectively). The Wald interval additionally exceeds nominal coverage on average for  $n = 30$  with coverage probability of 95.90%. However, as the correlation increases, the probability of coverage by the Wald interval decreases for  $\rho = 0.5$  (94.52%) and  $\rho = 0.6$  (93.32%). This low coverage probability by the Wald interval for  $n = 30$  is exacerbated by increased  $\theta_{RR}$  from 1.5 to 2. A similar, if slightly less extreme trend is observed when  $n = 60$ . Interestingly, when  $\theta_{RR} = 1.5$  and  $\rho = 0.6$ , the Wald interval does not reach nominal coverage on average until  $n = 200$ , which is a large sample size in practice. When  $\theta_{RR} = 2$ , this holds true for both  $\rho = 0.6$  and slightly weaker  $\rho = 0.5$ .

As expected, the performance of the Wald asymptotic confidence interval is dubious for small-to-moderate sample sizes. Under weaker correlation between two tests, for  $n = 30, 60, 150$ , we see the general trend of over-conservatism of the Wald interval, which dissipates for larger sample sizes. The coverage probability quickly drops as correlation increases, but recovers as sample size increases.

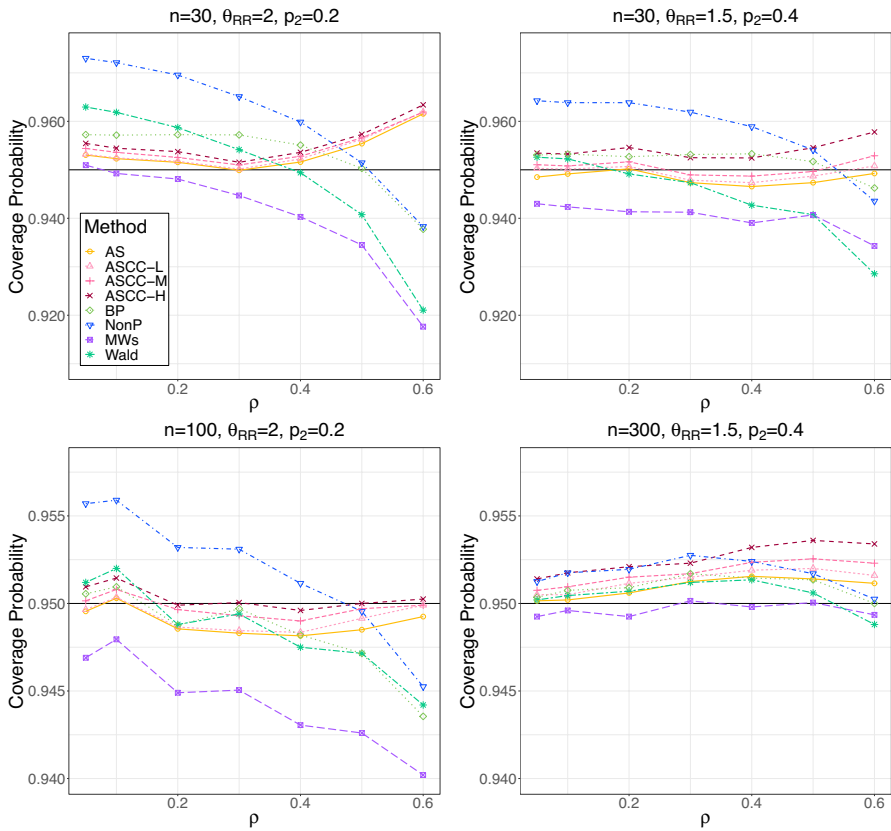
In the situation described above (Fig. 1), when  $n = 30$  and  $\theta_{RR} = 1.5$ , the uncorrected score interval was above nominal coverage on average for all strengths of correlation considered. As the sample size increases, the coverage of the score interval stabilizes close to the nominal 95%. Observing the trend over increasing strength of correlation further highlights this behavior. Figure 2 shows that, when  $\theta_{RR} = 2$ ,  $p_2 = 0.2$ , and  $n = 30$ , the coverage of all score-based intervals increases with increasing correlation, while the coverage of all alternatives decreases. When  $n=100$ , most confidence intervals display a decrease in coverage probability for increasing correlation, while the score-based methods maintain a reasonable level of coverage. In fact, the ASCC-H maintains coverage closest to the nominal level across all values of  $\rho$  considered in Fig. 2. This effect is still present, if somewhat mitigated, when  $\theta_{RR} = 1.5$  and  $p_2 = 0.4$ .

Further comparing the proposed ASCC intervals to the uncorrected score asymptotic interval, we see that, intuitively, when the uncorrected interval already has good coverage then the corrected interval is too conservative. However, the real benefit of the corrected interval presents itself when the score interval has poor coverage. In such scenarios, the continuity-corrected score interval not only brings the score interval closer to nominal coverage, but the closest out of all the confidence interval methods compared. Taking panel 6 of Fig. 1 for example ( $p_2 = 0.2$ ,  $\rho = 0.6$ , and  $\theta_{RR} = 2$ ), Panel 6 is overall the poorest performing scenario in this figure due to the strength of correlation and size of  $\theta_{RR}$  relative to  $p_2$ . The coverage of the uncorrected score interval starts off too conservative when  $n = 30$ , then corrects itself to hover just below the nominal level (between 94.55% when  $n=60$  and 94.79% when  $n=500$ ). Applying the ASCC with a strong correction brings the coverage within 0.01% of nominal coverage while keeping the coverage at or above nominal level for sample sizes between 60 and 300.



**Fig. 1** Coverage probabilities for eight confidence intervals for correlated relative risk by study sample size

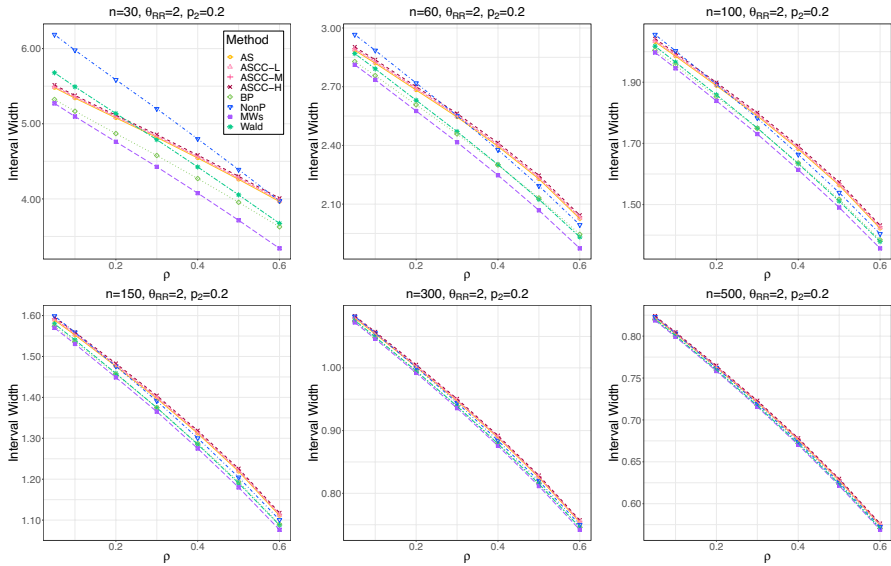
The score asymptotic interval was well behaved in most simulation scenarios included in our investigation. We observed that the score interval tended to be over-conservative for small sample sizes no matter the strength of correlation between



**Fig. 2** Coverage probabilities for eight confidence intervals for correlated relative risk by correlation between test 1 and test 2

tests, which we prefer to the frequency with which the Bonett–Price, MOVER Wilson, and Wald intervals failed to reach nominal coverage for small-to-moderate sample sizes when strong correlation is present. The example represented by Figs. 1 and 2 illustrates that under situations where the asymptotic score interval does not achieve nominal coverage and the alternative intervals also perform poorly (i.e., high correlation and large magnitude of effect), the ASCC can help achieve nominal coverage.

In line with Fagerland et al. [2], we found that the coverage probability of the MOVER Wilson interval is lower than other methods in most of our simulated scenarios. In Fig. 1, when  $n = 30, 60$  and  $\theta_{RR} = 1.5$ , the MOVER Wilson interval does not meet nominal coverage even when  $\rho = 0.2$ , at 94.58% and 94.91%. The coverage only worsens as the strength of correlation increases. The same can be seen in Fig. 2, where the coverage probability of the MOVER Wilson interval only sporadically exceeds the nominal level for various correlation strengths. For the same combination of parameters, the average confidence interval lengths can be seen in Fig. 3.

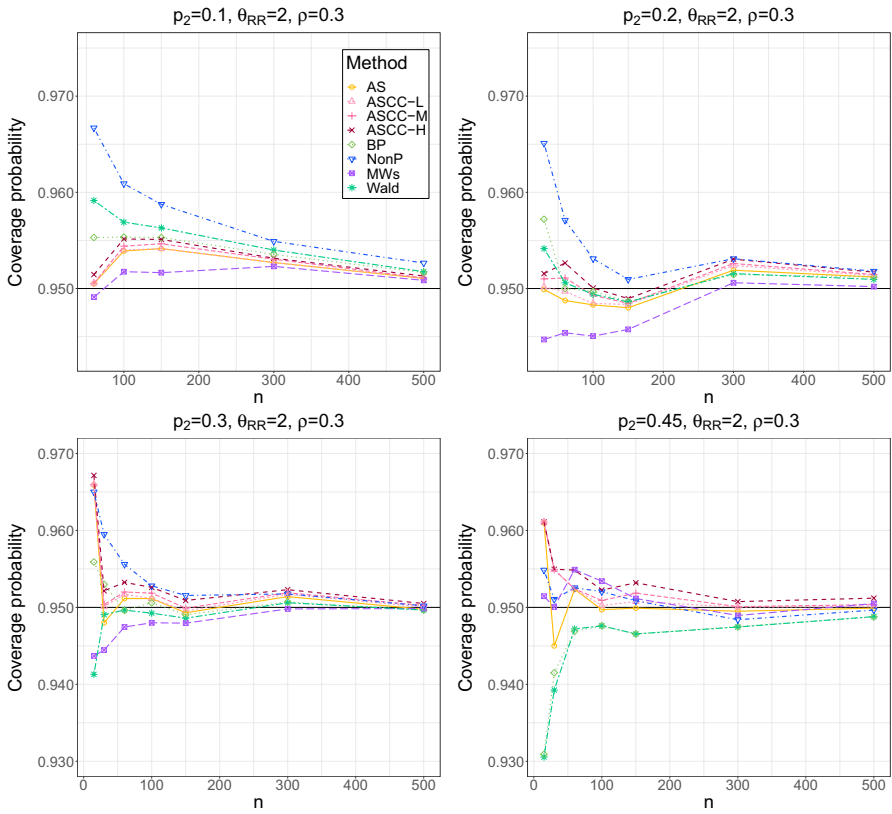


**Fig. 3** Average confidence interval lengths for eight confidence intervals for correlated relative risk by correlation between test 1 and test 2

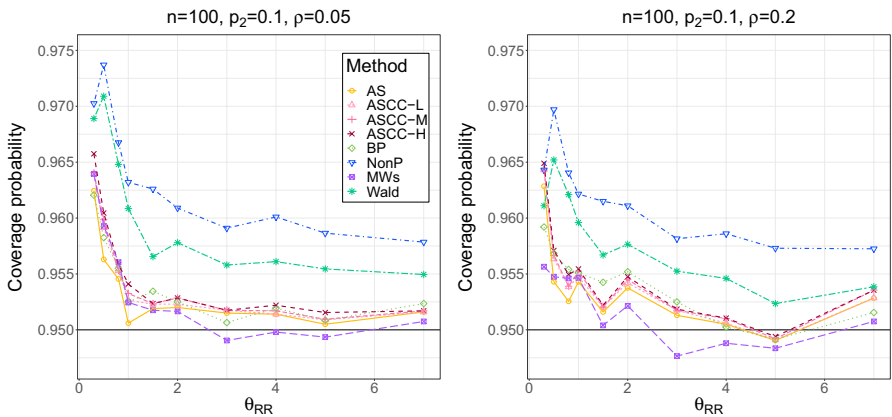
For sample sizes 30, 60, and 100, the MOVER Wilson interval consistently has the shortest interval lengths across all values of  $\rho$  studied. This is at odds with the poor coverage probability described above.

An exception is when  $p_1$  and  $p_2$  take on relatively large values, taking  $\theta_{RR} = 2$  and  $\rho = 0.3$  in Fig. 4. When  $p_2 = 0.1, 0.2$  (and thus  $p_1 = 0.2, 0.4$ , respectively), the MOVER Wilson interval has the lowest coverage probability on average for all sample sizes. However, as  $p_2$  increases to 0.45, the MOVER Wilson interval maintains coverage very close to the nominal level, only dipping to 94.90% when  $n=300$ . Contrastingly, the performance of the Bonett–Price interval when  $n=15$  plummets to 93.09%, recovering to 94.69% when the sample size reaches 60, but never meeting nominal coverage. Thus, in Fig. 4, when  $p_2$  is large causing  $p_1$  to lie closer to the exterior of the parameter space, the Bonett–Price and the Wald intervals have coverage the farthest below nominal while the MOVER Wilson interval maintains satisfactory coverage. The trend of improved performance is seen as  $p_2$  increases in the panels of Fig. 4. This indicates that, when the correlation is moderate, the bias which may affect the MOVER Wilson interval in general may be mediated for large values of the response probabilities.

Figure 5 looks at the performance of the considered confidence intervals under a typical sample size ( $n=100$ ) and probability of success ( $p_2 = 0.1$ ) for a broad range of true relative risk values. Weak correlation ( $\rho = 0.05$ ) and a more moderate correlation ( $\rho = 0.2$ ) are considered. We see similar trends as observed in previous figures, with decreasing coverage for stronger correlation. For very small values of relative risk ( $\theta_{RR} < 1$ ) all confidence intervals tend to be overly conservative. When  $\theta_{RR}$  is 5, most confidence intervals drop just below nominal coverage. In this case, using



**Fig. 4** Coverage probabilities for eight confidence intervals for correlated relative risk by study sample size

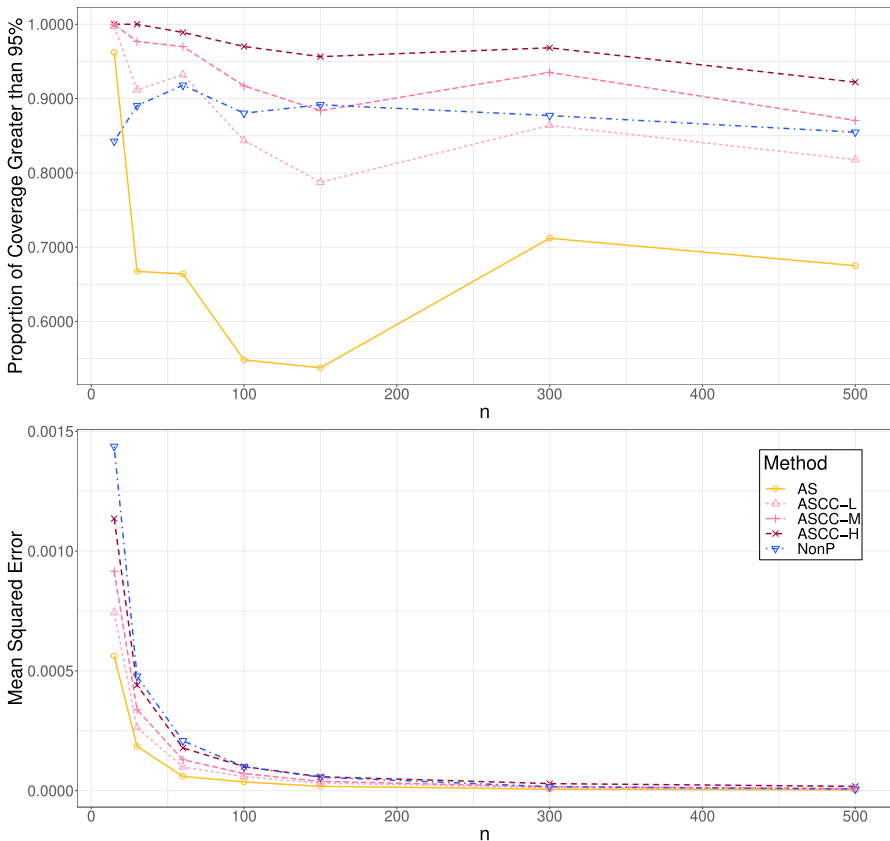


**Fig. 5** Coverage probabilities for eight confidence intervals for correlated relative risk by true relative risk

the nonparametric interval preserves coverage above nominal level. On the other hand, all the ASCC corrections fall below nominal coverage, but attain the closest coverage to 95% of any alternatives (particularly ASCC-H). This is excepting the Wald interval which, as expected under weaker correlations and sufficient sample size, performs well—intermediate coverage to the nonparametric interval and alternatives.

We further compare the proposed four CIs: ASCC-L, ASCC-M, ASCC-H, and the nonparametric interval, with regard to the proportion of scenarios under which each confidence interval method achieves nominal 95% coverage among all the configurations in  $\Omega$  in Fig. 6. That proportion for a given sample size  $n$  is computed as

$$\frac{\sum_{(\theta_{RR}, \rho, p_2) \in \Omega(n)} I\left( CP(\theta_{RR}, \rho, p_2 | n) \geq 95\% \right)}{|\Omega(n)|}$$



**Fig. 6** Mean squared deviation of the average coverage probability from the nominal 95% coverage (MSE) and the proportion of cases with average coverage probability exceeding the nominal 95% for five confidence intervals for correlated relative risk by study sample size



where  $|\Omega(n)|$  is the size of the parameter space given sample size  $n$ , and  $CP(\theta_{RR}, \rho, p_2|n)$  is the coverage probability given the study design parameters:  $\theta_{RR}$ ,  $\rho$ , and  $p_2$ . The score interval is added as reference. Both ASCC-M and ASCC-H have the guaranteed coverage proportion above 80% for all these studies sample sizes, and their proportions of guaranteed coverage probability are much higher than that of the ASCC-L interval. In Fig. 6, we also show the average deviation of simulation-based coverage probability from nominal 95% coverage (or MSE), specifically

$$MSE = \frac{\sum_{(\theta_{RR}, \rho, p_2|n) \in \Omega(n)} [CP(\theta_{RR}, \rho, p_2|n) - 95\%]^2}{|\Omega(n)|}.$$

For MSE, the proposed ASCC-M interval generally has lower MSE than ASCC-H. Based on the results from MSE and the proportion of guaranteed coverage, we would recommend the proposed ASCC-M interval among the three ASCC intervals.

The nonparametric confidence interval typically often has the largest MSE for sample sizes less than  $n = 100$ . This reflects to expected conservatism of nonparametric methods in general, the price paid for making fewer assumptions in the construction of the method. The insights above indicate that, relative to all other confidence interval methods considered, the nonparametric confidence interval’s coverage probability rarely drops below the nominal coverage level on average. This means that the proposed nonparametric confidence interval has a general tendency to be conservative, even when the performance of alternative methods is poor.

### 3.2 Case Studies

#### 3.2.1 Airway Hyper-Responsiveness

The airway hyper-responsiveness (AHR) study [39] was used in the literature to compare CIs for two correlated proportions, including the recent article by Fagerland et al. [2]. Children often experience pulmonary complications following stem cell transplant (SCT). AHR is indicative of the degree of sensitivity of the lungs to foreign stimuli and is associated with unfavorable respiratory symptoms [40]. A prospective pediatric study of 21 participants compared the incidence of AHR before and after stem cell transplant (see data in Table 2). The test for AHR is binary (positive/negative) and is paired data by pre- and post-SCT measurements. We obtain the same confidence intervals for the uncorrected asymptotic score, Wald, Bonett–Price,

**Table 2** Observed counts of pediatric airway hyper-responsiveness (AHR) before and after stem cell transplant (SCT)

	Pre-SCT	Post-SCT		Total
		AHR	No AHR	
AHR		1	1	2
No AHR		7	12	19
Total		8	13	21

and MOVER Wilson intervals as Fagerland et al. [2] in Table 3. Similar to their findings, we find an increased risk of AHR following SCT.

The Wald CI is quite wide compared to alternatives, with a width of 0.94. This is only surpassed in width by the proposed nonparametric method, with a width of 1.08. This is intuitive, as nonparametric methodology tends to be conservative by not making distributional assumptions, which is likely appropriate in a study of this size. The proposed ASCC method widens the original score asymptotic CI from 0.84 to 0.85, 0.87, and 0.89 for continuity corrections of increasing strength, which is particularly important in a study of this size. The MOVER CI has the shortest width at 0.80.

### 3.2.2 Breast Cancer Detection

We additionally apply the methods described above to data from a prospective study of adult women at risk for invasive breast cancer [41]. Dense breast tissue is associated with an increased risk of breast cancer and a greater likelihood of a false-negative mammogram from a screening for early detection of breast cancer [42]. Advanced medical techniques are therefore needed to detect invasive breast cancer in women with dense breast tissue. Sonogram and magnetic resonance imaging (MRI) are traditionally accepted detection technology. However, sonograms are labor intensive and associated with low specificity while gold-standard MRIs are expensive and thus unrealistic as a population screening technology. Increasingly popular alternative detection technologies are digital breast tomosynthesis (DBT) and abbreviated magnetic resonance imaging (AB-MRI).

Comstock et al. [41] sought to compare the detection rates of DBT and AB-MRI to the results of surgical biopsy, the standard of care for breast cancer diagnosis. A total of 1430 women aged 40-75 with dense breast tissue were enrolled in the study between December 2016 and November 2017 (see data in Tables 4 and 5). The primary outcome was detection of breast cancer, defined as either invasive breast cancer or ductal carcinoma in situ (DCIS) which is a cancerous noninvasive lesion. During the course of the study, each participant received DBT

**Table 3** CI lengths for  $\theta_{RR}$  for the AHR pre- and post-SCT study

Method	95% CI		Width
	Lower limit	Upper limit	
Asymptotic score	0.0653	0.9069	0.8416
ASCC-L	0.0628	0.9167	0.8539
ASCC-M	0.0603	0.9265	0.8662
ASCC-H	0.0555	0.9461	0.8906
Wald asymptotic	0.0625	0.9996	0.9371
Bonett–price hybrid Wilson score	0.0677	0.9227	0.8550
MOVER Wilson score	0.0686	0.8695	0.8008
Nonparametric method	0.0551	1.1333	1.0781

**Table 4** Diagnostic accuracy of DBT compared to standard of care

DBT	Invasive cancer or DCIS		Total
	Present	Absent	
Positive	9	36	45
Negative	14	1371	1385
Total	23	1407	1430

**Table 5** Diagnostic accuracy of AB-MRI compared to standard of care

AB-MRI	Invasive cancer or DCIS		Total
	Present	Absent	
Positive	22	187	209
Negative	1	1220	1221
Total	23	1407	1430

**Table 6** CI lengths for  $\theta_{RR}$  for the breast cancer detection study comparing DBT to SOC

Method	95% CI		Width
	Lower limit	Upper limit	
Asymptotic score	1.2795	3.0274	1.7479
ASCC-L	1.2794	3.0275	1.7481
ASCC-M	1.2794	3.0276	1.7482
ASCC-H	1.2792	3.0278	1.7486
Wald asymptotic	1.2717	3.0100	1.7383
Bonett–price hybrid Wilson score	1.2744	3.0037	1.7293
MOVER Wilson score	1.2705	2.9880	1.7175
Nonparametric method	1.2711	3.0116	1.7405

screening and AB-MRI screening for breast cancer. Formal diagnosis with breast cancer was obtained by follow-up within two years of study conclusion via surgical biopsy, the standard of care (SOC) diagnosis method.

Interestingly, when comparing DBT to SOC in Table 6, the nonparametric CI is of intermediate width (1.74) between those of the score intervals (ranging from 1.748 for uncorrected to 1.749 for high correction) and the Wald, Bonett–Price, and MOVER intervals. This is likely owing to the large study sample size. Comparing AB-MRI to SOC presented in Table 7, the nonparametric CI is the longest at 7.231. We see the same increasing trend in the widths of the score-based CIs, from 7.208 uncorrected to 7.209 with high correction. In both DBT and AB-MRI comparisons, the MOVER interval again has the shortest width at 1.717 and 7.147 for DBT and AB-MRI, respectively. This particular case study illustrates that the continuity correction modification of the proposed interval makes only a small, likely not practically significant, difference in large studies where

**Table 7** CI lengths for  $\theta_{RR}$  for the breast cancer detection study comparing AB-MRI to SOC

Method	95% CI		Width
	Lower limit	Upper limit	
Asymptotic score	6.2443	13.4526	7.2083
ASCC-L	6.2443	13.4527	7.2085
ASCC-M	6.2442	13.4528	7.2086
ASCC-H	6.2441	13.4531	7.2090
Wald asymptotic	6.1671	13.3892	7.2220
Bonett–price hybrid Wilson score	6.1815	13.3580	7.1764
MOVER Wilson score	6.1704	13.3173	7.1469
Nonparametric method	6.1643	13.3954	7.2311

the asymptotic approximation assumed by the original score interval is likely reasonable.

The lower limits for comparing AB-MRI to SOC are all above 6, which indicate that AB-MRI is a test that could have a very high false-positive rate. Meanwhile, the lower limits for comparing DBT to SOC are between 1.2 and 1.3 indicating a better performance of DBT as compared to AM-MRI.

## 4 Discussion

As our simulations have shown, imposing a continuity correction to Tang's iterative asymptotic score interval can be quite beneficial in certain situations. In particular, for small sample sizes, the asymptotic score with continuity correction provides the closest average coverage probability to the specified nominal level as correlation increases. Additionally, in both small ( $N=30$ ) and moderate ( $N=60,100$ ) sample sizes, ASCC was shown to be beneficial under increasing strength of correlation when the probability of event is small relative to the true relative risk. Practically, this indicates that if an investigator expects a stronger correlation between the two tests in the study, and expects small probabilities of observing the event but a larger effect size, you will more likely meet nominal coverage when applying ASCC than Tang's original score interval (or any other studied methods, save the nonparametric CI).

When the sample size is large and/or the correlation is low, applying a continuity correction can be conservative compared to the uncorrected asymptotic score interval. This contrasts with the behavior of the Wald, Bonett–Price, and Mover Wilson intervals in such situations, which are over-conservative for weakly correlated samples, but experience unacceptably low average coverage probabilities with increasing strength of correlation. Therefore, in situations with moderate sample and effect sizes, the standard recommendation holds to use the uncorrected asymptotic score interval. However, for larger anticipated effect sizes with strong correlation between tests, we recommend the asymptotic score with continuity correction for general

practice. We developed an R shiny app at the link: <https://dongyuanwu.shinyapps.io/PairedRR/>.

Additional to the improved operating characteristics of the proposed ASCC intervals under the above scenarios are the computational benefits of the proposed method. Provision of a closed form of the uncorrected and continuity-corrected asymptotic score interval decreases computation load relative to the iterative solution search previously required by Tang et al. [30]. Further, closed-form expressions for the confidence limits are more accessible to clinical audience and avoid common optimization challenges in the root-finding process. Regardless of the decision to impose a continuity correction, the use of the closed-form solution is recommended. R function implementations of all available methods can be obtained from the authors upon request.

In agreement with the findings of Duan et al. [32], we illustrate that the U-statistic-based nonparametric method is conservative when the sample size is small. In fact, in the majority of simulation scenarios studied, the nonparametric interval rarely drops below nominal coverage probability. The nonparametric interval frequently meets or exceeds nominal coverage, as shown in the bottom panel of Fig. 6. As a specific example, in Fig. 1, the nonparametric method only shows coverage probabilities consistently below nominal coverage when  $\rho = 0.6$  and  $\theta_{RR} = 2$ , while the performance of most other methods is severely challenged.

Though we recommend the use of the asymptotic score interval (corrected or uncorrected based on the characteristics of the available data) for practical use to achieve coverage probability closer to the nominal level, the practical context of the analysis may warrant the choice of more conservative coverage at the expense of a wider interval. This makes our extension of Duan's nonparametric interval for correlated RD to correlated RR useful. The decision to use the nonparametric confidence interval should in general be motivated by a risk–benefit assessment informed by the real-world consequences of failing to capture the true value of the parameter under study in the confidence interval. This may be particularly desirable in the drug development context, where national regulatory agencies tend to prefer conservative inferential techniques. Note that for both the RD and RR, the intervals are nonparametric in the sense that the derivation of the point estimate and the variance–covariance matrix are conducted without distributional assumptions, as outlined in Sect. 2.3.2. However, the variance of RD and RR are both derived using the Delta method, and hence some asymptotic behavior can be observed in the lines representing the nonparametric interval in the simulation summaries.

In the small sample case where a conservative CI is desired, one could also consider an exact CI for the correlated RR. In contrast with traditional methods, exact CIs do not rely on the asymptotic normal approximation to the binomial distribution to hold reasonably, but instead use the binomial distribution directly to enumerate all cumulative probabilities of interest for interval construction. Thus, exact CIs can be computationally intensive and are often most feasible in small-to-moderate sample size settings. This provides the benefit of improved coverage in small sample settings but sacrifices the simple closed-form expression and ease of computation of our proposed nonparametric interval for the paired RR. According to the authors of the *ExactCIdiff* package in R, even in relatively small sample sizes (such as  $n=100$ ) the computation for a single

exact confidence interval for the difference of two correlated proportions can take an hour to complete on an HP laptop with Intel(R) Core(TM) i5=2520M CPU@2.50 GHz and 8 GB RAM. This puts a time comparison outside of the scope of the current simulation study. Extensive development in this area can be found in the literature [43]. We believe a conservative interval is a good addition to any simulation study for comparison purposes but preferred to include an interval with less computational intensity. For similar reasons, we did not include the Bonett–Price with continuity correction in our simulation study. Fagerland et al. [2] found that the continuity-corrected Bonett–Price interval was so overly conservative, it approached the performance of an exact CI rather than an approximation.

In this paper, improved methodology for calculating confidence intervals for the correlated relative risk is presented in the context of a two-by-two contingency table. A future direction of research is to extend the methods and notations described here to two-way contingency tables of higher dimension. One example would be a study that is stratified by covariate(s) (e.g., gender, race). Estimation in this setting entails providing confidence interval formulas for the stratified correlated relative risk.

We would like to bring attention to another future direction of research which is not exclusive to this paper, but would have greatly added to the quality of the investigation. Though we thoroughly searched, it was a difficult task to find any applied example in which both the experimental design was applicable and the necessary information to construct the 2 by 2 table of interest was published. As a result, we were unable to conduct a resampling-based assessment of confidence interval coverage probabilities in our real data examples. This is a well-known challenge for methodological research related to clinical trial design and analysis. We hope that collaborative movements to make full clinical trial data publicly available where appropriate make this possible in our future lines of research.

### Appendix 1: The Boundary of the Pearson Correlation Coefficient

To obtain the range of the Pearson correlation coefficient  $\rho$  for a pair of Bernoulli random variables  $(X_1, X_2)$ , the following inequality is used,  $\max(0, p_1 + p_2 - 1) \leq p_{11} \leq \min(p_1, p_2)$ . The above inequality can be rewritten as

$$\max(0, p_1 + p_2 - 1) \leq \rho \sqrt{p_1(1 - p_1)p_2(1 - p_2)} + p_1p_2 \leq \min(p_1, p_2).$$

When  $p_1, p_2 \neq 0, 1$ , we can solve the right side of the inequality to obtain the upper bound of  $\rho$ ,

$$\begin{aligned} \rho &\leq \frac{\min(p_1, p_2) - p_1 p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} \\ \rho &\leq \frac{\min\{p_1(1-p_2), p_2(1-p_1)\}}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} \\ \rho &\leq \min \left\{ \left[ \frac{p_1(1-p_2)}{p_2(1-p_1)} \right]^{\frac{1}{2}}, \left[ \frac{p_2(1-p_1)}{p_1(1-p_2)} \right]^{\frac{1}{2}} \right\}. \end{aligned}$$

Moreover, the lower bound of  $\rho$  is obtained by solving the left side of the inequality,

$$\begin{aligned} \rho &\geq \frac{\max(0, p_1 + p_2 - 1) - p_1 p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} \\ \rho &\geq \frac{\max\{-p_1 p_2, -(1-p_1)(1-p_2)\}}{\sqrt{p_1(1-p_1)p_2(1-p_2)}} \\ \rho &\geq \max \left\{ - \left[ \frac{p_1 p_2}{(1-p_1)(1-p_2)} \right]^{\frac{1}{2}}, - \left[ \frac{(1-p_1)(1-p_2)}{p_1 p_2} \right]^{\frac{1}{2}} \right\}. \end{aligned}$$

Appendix 1 shows the derivation of the formulas for  $L_\rho$  and  $U_\rho$

$$\begin{aligned} L_\rho &= \max \left\{ - \left[ \frac{p_1 p_2}{(1-p_1)(1-p_2)} \right]^{\frac{1}{2}}, - \left[ \frac{(1-p_1)(1-p_2)}{p_1 p_2} \right]^{\frac{1}{2}} \right\}; \\ U_\rho &= \min \left\{ \left[ \frac{p_1(1-p_2)}{p_2(1-p_1)} \right]^{\frac{1}{2}}, \left[ \frac{p_2(1-p_1)}{p_1(1-p_2)} \right]^{\frac{1}{2}} \right\}. \end{aligned}$$

The lower bound and the upper bound can be further improved when the ranges of  $p_{12}$ ,  $p_{21}$ , and  $p_{22}$  are utilized. In the R program, we checked all these four cell probabilities to make sure that they are between 0 and 1.

**Appendix 2: Proof of Theorem 1: The Nam–Blackwelder CI Based on the Test Statistic  $T(\theta_0)$  is Equivalent to the Score Interval Based on the Test Statistic  $S(\theta_0)$**

Nam and Blackwelder’s interval is calculated from  $T(\theta) = \pm z_{1-\alpha/2}$ . Given  $\alpha$ , the score asymptotic CI is calculated using  $S(\theta)$ . These two CI methods are equal if  $T(\theta) = S(\theta)$ . To simplify the notations, we denote  $x_{.1} = x_{11} + x_{21}$  and  $x_{1.} = x_{11} + x_{12}$ . These two test statistics are

$$T(\theta) = \frac{x_{1.} - x_{.1}\theta}{\sqrt{n\theta(\hat{p}_{12} + \hat{p}_{21})}}$$

and

$$S(\theta) = \frac{x_{1\cdot} - x_{1\cdot}\theta}{\sqrt{n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)}}$$

where  $\tilde{p}_{21}$ ,  $\hat{p}_{12}$ , and  $\hat{p}_{21}$  are presented in the manuscript. We calculated the difference of denominators of  $T(\theta)$  and  $S(\theta)$ ,

$$\begin{aligned} & n\theta(\hat{p}_{12} + \hat{p}_{21}) - [n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)] \\ &= n\theta(1 + \theta)\hat{p}_{12} - \theta(\theta - 1)(x_{11} + x_{12} + x_{21}) - \left[ \frac{1}{2}\sqrt{(x_{1\cdot} - x_{1\cdot}\theta)^2 + 4x_{12}x_{21}\theta^2} - \frac{1}{2}x_{1\cdot}\theta^2 \right. \\ &\quad \left. + (\theta - 1)(x_{11} + x_{12} + x_{21}) + \frac{1}{2}(x_{1\cdot} + 2x_{21}) \right] \\ &= \frac{1}{2}\left[ \sqrt{(x_{1\cdot} - x_{1\cdot}\theta)^2 + 4x_{12}x_{21}\theta^2} - \sqrt{(x_{1\cdot} - x_{1\cdot}\theta)^2 + 4x_{12}x_{21}\theta^2} \right] \\ &= 0. \end{aligned}$$

Therefore,  $T(\theta) = S(\theta)$ . The Nam–Blackwelder  $100(1 - \alpha)\%$  CI based on the constrained maximum likelihood is exactly the same as the  $100(1 - \alpha)\%$  score asymptotic CI.

### Appendix 3: Closed-Form Solutions for the Score Interval and the Proposed ASCC Intervals

Following the Appendix in Nam [26], we first derive the closed-form estimation of confidence limits for the score asymptotic CI. Then, we present the closed-form formula to calculate exact confidence limits of the continuity-corrected score asymptotic CI. To simplify the notations, we denote  $x_{\cdot 1} = x_{11} + x_{21}$  and  $x_{1\cdot} = x_{11} + x_{12}$ .

Similar to the Appendix in Nam [26], we solve the quartic equation to obtain the exact values of two confidence limits for the Score Asymptotic CI. The following context presents steps to obtain the exact solutions. When  $\theta$  greater than  $x_{1\cdot}/x_{1\cdot}$ , the equation to calculate the upper confidence limit is

$$\frac{x_{1\cdot}\theta - x_{1\cdot}}{\sqrt{n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)}} = z_{1-\frac{\alpha}{2}},$$

where  $\tilde{p}_{21}$  is shown in Sect. 2.2.4. Solving the equation,

$$\begin{aligned} x_{1\cdot}\theta - x_{1\cdot} &= z_{1-\frac{\alpha}{2}}\sqrt{n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)} \\ (x_{1\cdot}\theta - x_{1\cdot})^2 &= z_{1-\frac{\alpha}{2}}^2[n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)] \\ z_{1-\frac{\alpha}{2}}^2n(1 + \theta)\tilde{p}_{21} &= x_{1\cdot}^2\theta^2 - \left[ 2x_{1\cdot}x_{1\cdot} + z_{1-\frac{\alpha}{2}}^2(x_{11} + x_{12} + x_{21}) \right]\theta + x_{1\cdot}^2 + z_{1-\frac{\alpha}{2}}^2(x_{11} + x_{12} + x_{21}). \end{aligned}$$

Input  $\tilde{p}_{21}$  as an expression of  $\theta$ , we have

$$\frac{1}{2}z_{1-\frac{\alpha}{2}}^2\sqrt{T} = \left( x_{1\cdot}^2 + \frac{1}{2}z_{1-\frac{\alpha}{2}}^2x_{1\cdot} \right)\theta^2 - \left[ 2x_{1\cdot}x_{1\cdot} + z_{1-\frac{\alpha}{2}}^2(x_{11} + x_{12} + x_{21}) \right]\theta + x_{1\cdot}^2 + \frac{1}{2}z_{1-\frac{\alpha}{2}}^2x_{1\cdot}. \tag{10}$$



where  $T \geq 0$  and  $T = x_1^2 \theta^4 - 2(x_{11}^2 + x_{11}x_{12} + x_{11}x_{21} - x_{21}x_{12})\theta^2 + x_1^2$ . To calculate  $\theta$ , we need to square both sides of Eq. (10). However, the right side of Eq. (10), denoted as  $V$ , is not necessarily nonnegative and redundant solutions of  $\theta$  are added in such a case. Therefore, if the solutions of  $\theta$  satisfy  $V < 0$ , the solutions should be discarded. Squaring both sides of Eq. (10), we can obtain the quartic equation as

$$a\theta^4 + b\theta^3 + c\theta^2 + d\theta + e = 0, \tag{11}$$

where

$$\begin{aligned} a &= x_1^4 + z_{1-\frac{\alpha}{2}}^2 x_1^3 \\ b &= -(2x_1^2 + z_{1-\frac{\alpha}{2}}^2 x_1)[2x_1 x_1 + z_{1-\frac{\alpha}{2}}^2 (x_{11} + x_{12} + x_{21})] \\ c &= 6x_1^2 x_1^2 + z_{1-\frac{\alpha}{2}}^4 (x_1 + x_1)(x_{11} + x_{12} + x_{21}) + z_{1-\frac{\alpha}{2}}^2 x_1 x_1 (6x_{11} + 5x_{12} + 5x_{21}) \\ d &= -(2x_1^2 + z_{1-\frac{\alpha}{2}}^2 x_1)[2x_1 x_1 + z_{1-\frac{\alpha}{2}}^2 (x_{11} + x_{12} + x_{21})] \\ e &= x_1^4 + z_{1-\frac{\alpha}{2}}^2 x_1^3. \end{aligned}$$

Using the Ferrari’s method, the formulas of the roots are

$$\begin{aligned} \theta_1 &= -\frac{b}{4a} - S - \frac{1}{2} \sqrt{-4S^2 - 2k + \frac{h}{S}} \\ \theta_2 &= -\frac{b}{4a} - S + \frac{1}{2} \sqrt{-4S^2 - 2k + \frac{h}{S}} \\ \theta_3 &= -\frac{b}{4a} + S - \frac{1}{2} \sqrt{-4S^2 - 2k - \frac{h}{S}} \\ \theta_4 &= -\frac{b}{4a} + S + \frac{1}{2} \sqrt{-4S^2 - 2k - \frac{h}{S}}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} k &= \frac{8ac - 3b^2}{8a^2} \\ h &= \frac{b^3 - 4abc + 8a^2 d}{8a^3}. \end{aligned}$$

The value of  $S$  is calculated using different formulas under various conditions. We denote that  $\Delta_0 = c^2 - 3bd + 12ae$ ,  $\Delta_1 = 2c^3 - 9bcd + 27b^2e + 27ad^2 - 72ace$ , and  $\Delta = (4\Delta_0^3 - \Delta_1^2)/27$ . If  $\Delta < 0$  and  $\Delta_0 \neq 0$ ,

$$Q = \sqrt[3]{\frac{\Delta_1 + \sqrt{(\Delta_1^2 - 4\Delta_0^3)}}{2}}$$

$$S = \frac{1}{2} \sqrt{-\frac{2}{3}k + \frac{1}{3a}(Q + \frac{\Delta_0}{Q})}$$

If  $\Delta < 0$  and  $\Delta_0 = 0$ ,

$$Q = \sqrt[3]{\Delta_1}$$

$$S = \frac{1}{2} \sqrt{-\frac{2}{3}k + \frac{1}{3a}(Q + \frac{\Delta_0}{Q})}$$

If  $\Delta > 0$ ,

$$\varphi = \arccos\left(\frac{\Delta_1}{2\sqrt{\Delta_0^3}}\right)$$

$$S = \frac{1}{2} \sqrt{-\frac{2}{3}k + \frac{2}{3a}\sqrt{\Delta_0} \cos \frac{\varphi}{3}}$$

If  $\Delta = 0$  and  $\Delta_0 \neq 0$ , the formula in case  $\Delta > 0$  can be used to calculate  $S$ . If  $\Delta = 0$  and  $\Delta_0 = 0$ , thus  $\Delta_1 = 0$ , at least three roots of Eq. (11) are equal. In this special case, four roots of Eq. (11) are denoted as the triple root  $\theta_{tri}$  and the unique root  $\theta_{uni}$ . The roots are calculated using the following procedure. Solving the equation,  $6a\theta^2 + 3b\theta + c = 0$ , to obtain two values. Plugging in the two values to the left side of Eq. (11) to obtain the value satisfied the quartic equation. This common root of the two equations is  $\theta_{tri}$ , then the unique root is calculated by  $\theta_{uni} = -3\theta_{tri} - b/a$ . If  $S = 0$ , the associated depressed quartic equation of Eq. (11) is a biquadratic equation,

$$\theta^4 + \left(\frac{8ac - 3b^2}{8a^2}\right)\theta^2 + \frac{256a^3e + 16ab^2c - 3b^4 - 64a^2bd}{256a^4} = 0,$$

the four roots of Eq. (11) can be obtained by solving the upper equation.

Because of symmetry, the lower confidence limit also satisfies Eq. (11). After the four roots are obtained, calculating the corresponding values of  $V$ . The roots with  $V < 0$  are discarded, then plugging in the left roots to Eq. (7) to select the upper and lower confidence limits  $\theta_U$  and  $\theta_L$  satisfying  $S(\theta_U) = -z_{1-\frac{\alpha}{2}}$  and  $S(\theta_L) = z_{1-\frac{\alpha}{2}}$ .

To calculate confidence limits of the  $100(1 - \alpha)\%$  Continuity-corrected Score Asymptotic CI, we need to solve the equation,

$$\frac{|x_1 - x_1\theta| - \frac{x_1}{\delta n}}{\sqrt{n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)}} = z_{1-\frac{\alpha}{2}},$$

where  $\delta$  is a constant (e.g.,  $\delta = 2$ ). Thus, solving the equation,

$$\frac{x_{.1}\theta - (x_{.1} + \frac{x_{.1}}{\delta n})}{\sqrt{n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)}} = z_{1-\frac{\alpha}{2}}, \tag{13}$$

to obtain  $\theta_U$ . And solving the equation,

$$\frac{(x_{.1} - \frac{x_{.1}}{\delta n}) - x_{.1}\theta}{\sqrt{n(1 + \theta)\tilde{p}_{21} + (x_{11} + x_{12} + x_{21})(\theta - 1)}} = z_{1-\frac{\alpha}{2}}, \tag{14}$$

to obtain  $\theta_L$ . From Eq. (13), we can obtain the quartic equation like Eq. (11) where

$$\begin{aligned} a &= x_{.1}^4 + z_{1-\frac{\alpha}{2}}^2 x_{.1}^3 \\ b &= -(2x_{.1}^2 + z_{1-\frac{\alpha}{2}}^2 x_{.1}) \left[ 2x_{.1} \left( x_{.1} + \frac{x_{.1}}{\delta n} \right) + z_{1-\frac{\alpha}{2}}^2 (x_{11} + x_{12} + x_{21}) \right] \\ c &= 6x_{.1}^2 \left( x_{.1} + \frac{x_{.1}}{\delta n} \right)^2 + z_{1-\frac{\alpha}{2}}^4 (x_{.1} + x_{.1})(x_{11} + x_{12} + x_{21}) + z_{1-\frac{\alpha}{2}}^2 x_{.1} \left[ \left( x_{.1} + \frac{x_{.1}}{\delta n} \right)^2 \right. \\ &\quad \left. + 4(x_{11} + x_{12} + x_{21}) \left( x_{.1} + \frac{x_{.1}}{\delta n} \right) + x_{.1}x_{.1} \right] \\ d &= - \left[ 2 \left( x_{.1} + \frac{x_{.1}}{\delta n} \right)^2 + z_{1-\frac{\alpha}{2}}^2 x_{.1} \right] \left[ 2x_{.1} \left( x_{.1} + \frac{x_{.1}}{\delta n} \right) + z_{1-\frac{\alpha}{2}}^2 (x_{11} + x_{12} + x_{21}) \right] \\ e &= \left( x_{.1} + \frac{x_{.1}}{\delta n} \right)^4 + z_{1-\frac{\alpha}{2}}^2 x_{.1} \left( x_{.1} + \frac{x_{.1}}{\delta n} \right)^2. \end{aligned}$$

Then the four roots are calculated by Eq. (12) obtained from the Ferrari’s method. Because  $\theta_U$  satisfies Eq. (13), only one of the four roots is selected as  $\theta_U$ . From Eq. (14), the quartic equation like Eq. (11) is obtained where

$$\begin{aligned} a &= x_{.1}^4 + z_{1-\frac{\alpha}{2}}^2 x_{.1}^3 \\ b &= -(2x_{.1}^2 + z_{1-\frac{\alpha}{2}}^2 x_{.1}) \left[ 2x_{.1} \left( x_{.1} - \frac{x_{.1}}{\delta n} \right) + z_{1-\frac{\alpha}{2}}^2 (x_{11} + x_{12} + x_{21}) \right] \\ c &= 6x_{.1}^2 \left( x_{.1} - \frac{x_{.1}}{\delta n} \right)^2 + z_{1-\frac{\alpha}{2}}^4 (x_{.1} + x_{.1})(x_{11} + x_{12} + x_{21}) + z_{1-\frac{\alpha}{2}}^2 x_{.1} \left[ \left( x_{.1} - \frac{x_{.1}}{\delta n} \right)^2 \right. \\ &\quad \left. + 4(x_{11} + x_{12} + x_{21}) \left( x_{.1} - \frac{x_{.1}}{\delta n} \right) + x_{.1}x_{.1} \right] \\ d &= - \left[ 2 \left( x_{.1} - \frac{x_{.1}}{\delta n} \right)^2 + z_{1-\frac{\alpha}{2}}^2 x_{.1} \right] \left[ 2x_{.1} \left( x_{.1} - \frac{x_{.1}}{\delta n} \right) + z_{1-\frac{\alpha}{2}}^2 (x_{11} + x_{12} + x_{21}) \right] \\ e &= \left( x_{.1} - \frac{x_{.1}}{\delta n} \right)^4 + z_{1-\frac{\alpha}{2}}^2 x_{.1} \left( x_{.1} - \frac{x_{.1}}{\delta n} \right)^2. \end{aligned}$$

The four roots are calculated by Eq. (12) and  $\theta_L$  is selected from the four roots using the condition that  $\theta_L$  satisfies Eq. (14).

### Appendix 4: The Asymptotic Based Nonparametric Confidence Interval

For correlated binary data of two tests, Duan et al. [32] estimated the variance–covariance matrix of  $(p_1, p_2)^T$  as

$$\hat{V} = \begin{bmatrix} \frac{(x_{11}+x_{12})(x_{21}+x_{22})}{n^2(n-1)} & \frac{x_{11}x_{22}-x_{12}x_{21}}{n^2(n-1)} \\ \frac{x_{11}x_{22}-x_{12}x_{21}}{n^2(n-1)} & \frac{(x_{11}+x_{21})(x_{12}+x_{22})}{n^2(n-1)} \end{bmatrix}.$$

The proof of this estimated covariance matrix could be found in their paper’s appendix using Langes’s rank-based method.

For the log relative risk  $g(p_1, p_2) = \log(p_1/p_2)$ , using the first-order Taylor expansion, we have

$$g(p_1, p_2) \approx \log\left(\frac{\hat{p}_1}{\hat{p}_2}\right) + \left[\frac{1}{\hat{p}_1} \quad -\frac{1}{\hat{p}_2}\right] \begin{bmatrix} p_1 - \hat{p}_1 \\ p_2 - \hat{p}_2 \end{bmatrix},$$

where  $\hat{p}_1 = (x_{11} + x_{12})/n$ ,  $\hat{p}_2 = (x_{11} + x_{21})/n$ , and  $\hat{p}_1, \hat{p}_2 > 0$ . Therefore, the estimated variance of the log relative risk can be obtained using the Delta method

$$\begin{aligned} \hat{V}_{LRR} &= \begin{bmatrix} \frac{1}{\hat{p}_1} & -\frac{1}{\hat{p}_2} \end{bmatrix} \hat{V} \begin{bmatrix} \frac{1}{\hat{p}_1} \\ -\frac{1}{\hat{p}_2} \end{bmatrix} \\ &= \frac{1}{n^2(n-1)\hat{p}_1^2\hat{p}_2^2} [(x_{11}x_{12} + x_{21}x_{22})\hat{p}_1^2 + (x_{11}x_{21} + x_{12}x_{22})\hat{p}_2^2 \\ &\quad + x_{11}x_{22}(\hat{p}_1 - \hat{p}_2)^2 + x_{12}x_{21}(\hat{p}_1 + \hat{p}_2)^2]. \end{aligned}$$

It is straightforward that  $\hat{V}_{LRR} > 0$  because  $\hat{p}_1$  and  $\hat{p}_2$  are nonzero positive values.

Based on  $\hat{V}_{LRR}$ , we then constructed a approximated  $100(1 - \alpha)\%$  CI for log relative risk. The simulation results in Duan et al. [32] showed that the  $t$  approximation with the degree of freedom  $df_0 = n - 1$  provided preferable performance of the corresponding CI of risk differences. Compared with the normal approximation, the  $t$  approximation with degrees of freedom  $df_0$  leads to slightly conservative test following Brunner et al. [44, 45]. So, a  $100(1 - \alpha)\%$  CI of the log relative risk is constructed as (L, U):

$$\begin{aligned} L &= \log\left(\frac{\hat{p}_1}{\hat{p}_2}\right) - t_{n-1, 1-\alpha/2} \hat{V}_{LRR} \\ U &= \log\left(\frac{\hat{p}_1}{\hat{p}_2}\right) + t_{n-1, 1-\alpha/2} \hat{V}_{LRR}, \end{aligned}$$

where  $t_{n-1, 1-\alpha/2}$  is the  $\alpha/2$  upper quantile of the  $t$  distribution with  $n - 1$  degrees of freedom. Moreover, the  $100(1 - \alpha)\%$  CI of the relative risk is obtained as  $(\exp(L), \exp(U))$ .

**Acknowledgements** The authors are very grateful to the Editor, Associate Editor, and two reviewers for their insightful comments that help improve the manuscript. We would like to thank Dr. Pete Laud from University of Sheffield who gave us the permission to use, modify, and redistribute the functions in his developed R package *ratesci*. The authors thank Dr. Duan's group who shared their R code for their nonparametric interval. Shan's research is partially supported by grants from NIH: R01AG070849 and R03CA248006.

## References

- Piantadosi S (2005) Clinical trials: a methodological perspective, 2nd edn. Wiley, Hoboken
- Fagerland MW, Lydersen S, Laake P (2014) Recommended tests and confidence intervals for paired binomial proportions. *Stat Med* 33(16):2850–2875
- Newcombe RG (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 17(May 1998):857–872
- Shan G (2015) Exact statistical inference for categorical data, 1st edn. Academic Press, San Diego
- Shan G, Ma C, Hutson AD, Wilding GE (2012) An efficient and exact approach for detecting trends with binary endpoints. *Stat Med* 31(2):155–164
- Okely JA, Corley J, Welstead M, Taylor AM, Page D, Skarabela B, Redmond P, Cox SR, Russ TC (2021) Change in physical activity, sleep quality, and psychosocial variables during COVID-19 lockdown: evidence from the lothian birth cohort 1936. *Int J Environ Res Public Health* 18(1):1–16
- Shan G, Wang W (2021) Advanced statistical methods and designs for clinical trials for COVID-19. *Int J Antimicrob Agents* 57(1):106167
- Shan G (2013) Exact unconditional testing procedures for comparing two independent Poisson rates. *J Stat Comput Simul* 85(5):947–955
- Shan G (2014) New nonparametric rank-based tests for paired data. *Open J Stat* 04(07):495–503
- Shan G (2016) Exact confidence intervals for randomized response strategies. *J Appl Stat* 43(7):1279–1290
- Shan G (2014) Exact approaches for testing non-inferiority or superiority of two incidence rates. *Stat Probab Lett* 85:129–134
- Shan G, Gerstenberger S (2017) Fisher's exact approach for post hoc analysis of a chi-squared test. *PLoS ONE* 12(12):e0188709
- Shan G (2015b) Improved confidence intervals for the Youden Index. *PLoS ONE* 10(7), e0127272+
- Shan G (2020) Accurate confidence intervals for proportion in studies with clustered binary outcome. *Stat Methods Med Res* 29(10):3006–3018
- Casella G, Berger RL (2002) Statistical inference, 2nd edn. Wadsworth Cengage Learning, Mason
- Gart JJ, Nam J-M (1988) Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. *Int Biometric Soc* 44(2):323–338
- Lang JB (2008) Score and profile likelihood confidence intervals for contingency table parameters. *Stat Med* 27:5975–5990
- Nam J-M, Blackwelder WC (2002) Analysis of the ratio of marginal probabilities in a matched-pair setting. *Stat Med* 21(5):689–699
- Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken
- Shan G (2021) Optimal two-stage designs based on restricted mean survival time for a single-arm study. *Contemp Clin Trials Commun* 21:100732
- Shan G, Ma C, Hutson AD, Wilding GE (2013) Randomized two-stage phase II clinical trial designs based on Barnard's exact test. *J Biopharm Stat* 23(5):1081–1090
- Shan G, Wang W (2017) Exact one-sided confidence limits for Cohen's kappa as a measurement of agreement. *Stat Methods Med Res* 26(2):615–632
- Shan G, Dodge-Francis C, Wilding GE (2020) Exact unconditional tests for dichotomous data when comparing multiple treatments with a single control. *Ther Innov Regul Sci* 54(2):411–417
- Shan G, Zhang H, Barbour J (2021) Bootstrap confidence intervals for correlation between continuous repeated measures. *Stat Methods Appl*, 1–21
- Wilson EB (1927) Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 22(158):209–212

26. Nam J-M (2009) Efficient interval estimation of a ratio of marginal probabilities in matched-pair data: non-iterative method. *Stat Med* 28(23):2929–2935
27. Bonett DG, Price RM (2006) Confidence intervals for a ratio of binomial proportions based on paired data. *Stat Med* 25(17):3039–3047
28. Tang NS, Tang ML, Chan ISF (2003) On tests of equivalence via non-unity relative risk for matched-pair design. *Stat Med* 22(8):1217–1233
29. Donner A, Zou GY (2012) Closed-form confidence intervals for functions of the normal mean and standard deviation. *Stat Methods Med Res* 21(4):347–359
30. Tang ML, Li HQ, Tang NS (2012) Confidence interval construction for proportion ratio in paired studies based on hybrid method. *Stat Methods Med Res* 21(4):361–378
31. Plackett RL (1964) The continuity correction in 2x2 tables. *Biometrika* 51(3):327–337
32. Duan C, Cao Y, Zhou L, Tan MT, Chen P (2018) A novel nonparametric confidence interval for differences of proportions for correlated binary data. *Stat Methods Med Res* 27(8):2249–2263
33. Leisch F, Weingessel A, Hornik K (1998) On the generation of correlated artificial binary data on the generation of correlated artificial binary data
34. Kowalski J, Xin T (2008) *Modern applied U-statistics*. Wiley, Hoboken
35. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3):837–845
36. R Core Team (2021) *R: a language and environment for statistical computing*
37. Laud P (2018) *ratesci: confidence intervals for comparisons of binomial or Poisson rates*
38. Goldfeld K, Wujciak-Jens J (2020) *simstudy: simulation of study data*
39. Bentur L, Lapidot M, Livnat G, Hakim F, Lidroneta-Katz C, Porat I, Vilozni D, Elhasid R (2009) Airway reactivity in children before and after stem cell transplantation. *Pediatric Pulmonol* 44(9):845–850
40. Postma DS, Kerstjens HAM, Postma S (1998) Characteristics of airway hyperresponsiveness in asthma and chronic obstructive pulmonary disease RELATIONSHIP BETWEEN SEVERITY OF HYPERRESPONSIVENESS AND FEV 1 LEVEL. Technical report
41. Comstock CE, Gatsonis C, Newstead GM, Snyder BS, Gareen IF, Bergin JT, Rahbar H, Sung JS, Jacobs C, Harvey JA, Nicholson MH, Ward RC, Holt J, Prather A, Miller KD, Schnall MD, Kuhl CK (2020) Comparison of abbreviated breast MRI vs digital breast tomosynthesis for breast cancer detection among women with dense breasts undergoing screening. *JAMA* 323(8):746–756
42. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, Paterson AD (2005) Mammographic breast density as an intermediate phenotype for breast cancer. *Lancet Oncol* 6:798–808
43. Wang W, Shan G (2015) Exact confidence intervals for the relative risk and the odds ratio. *Biometrics* 71(4):985–995
44. Brunner E, Dette H, Munk A (1997) Box-type approximations in nonparametric factorial designs. *J Am Stat Assoc* 92(440):1494–1502
45. Brunner E, Munzel U, Puri ML (2002) The multivariate nonparametric Behrens-Fisher problem. *J Stat Plan Inference* 108(1–2):37–53