Received: 25 October 2021 | Accepted: 25 April 2022

DOI: 10.1111/biom.13688

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Exact-corrected confidence interval for risk difference in noninferiority binomial trials

Nour Hawila | Arthur Berg

Division of Biostatistics and Bioinformatics, College of Medicine, Penn State University, Pennsylvania, USA

Correspondence
Arthur Berg, Department of Public Health Sciences, A210 Penn State College of Medicine, 90 Hope Drive, Suite 2200, Hershey, PA 17033-0855.
Email: berg@psu.edu

Abstract

A novel confidence interval estimator is proposed for the risk difference in noninferiority binomial trials. The proposed confidence interval, which is dependent on the prespecified noninferiority margin, is consistent with an exact unconditional test that preserves the type-I error and has improved power, particularly for smaller sample sizes, compared to the confidence interval by Chan and Zhang. The improved performance of the proposed confidence interval is theoretically justified and demonstrated with simulations and examples. An R package is also distributed that implements the proposed methods along with other confidence interval estimators.

KEYWORDS
confidence interval estimation, exact test, noninferiority clinical trial

1 | INTRODUCTION

We consider a noninferiority trial with a binary outcome and risk difference as the treatment effect. The noninferiority trial design incorporates a noninferiority margin, δ_0, and generalizes the standard comparative binomial trial corresponding to $\delta_0 = 0$. In Chan (1998), a class of exact-based tests is described for noninferiority binomial trials with type-I error rates that are guaranteed to be bounded by the level of the test. These exact-based procedures do not leverage the conditional distribution of a sufficient statistic, like that of the Fisher's exact test, but rather produce an exact unconditional test using a maximization method (Boschloo, 1970; McDonald *et al.*, 1977; Lehmann and Romano, 2006; Basu, 2011). For standard comparative trials ($\delta_0 = 0$), such exact unconditional tests were shown to be more powerful than Fisher's conditional exact test (Suissa and Shuster, 1985; Haber, 1986).

As described in Wasserstein *et al.* (2019), it is often not appropriate to just report *p*-value results; interval estimates of the effect size should also be reported. The unconditional exact method of Chan (1998) does not immediately yield a corresponding confidence interval estimator, but Chan and Zhang (1999) produce a confidence interval estimator that does leverage the exact unconditional test. However, we show that this confidence interval corresponds to a statistical test that is more conservative than the exact unconditional test of Chan (1998). Less conservative confidence interval estimators have been proposed in Miettinen and Nurminen (1985) and Farrington and Manning (1990), but these confidence interval estimators are based on asymptotic distributions and correspond to statistical tests that do not necessarily preserve type-I error rates.

We introduce a novel confidence interval estimator—called the exact-corrected (EC) estimator—that is less conservative and more powerful than the Chan & Zhang interval, but that also corresponds to a statistical test with preserved type-I error. The approach modifies the pivotal quantity used to produce the asymptotic confidence interval in Miettinen and Nurminen (1985), referred to as the δ-projected Z-score, and tacks on a correction factor to produce a confidence interval that is consistent with the exact test of Chan (1998). The proposed EC interval estimator is particularly unique in that it explicitly incorporates the noninferiority margin in the estimator, so different

TABLE 1 Description of the noninferiority hypotheses

Hypothesis	Positive outcome	Negative outcome	Interpretation
H_0	$\delta \leq -\delta_0$	$\delta \geq \delta_0$	"inferior trial," T is inferior to C
H_1	$\delta > -\delta_0$	$\delta < \delta_0$	"noninferior trial;" T is not inferior to C

predetermined noninferiority margins will result in different confidence intervals.

Next, we precisely define the statistical hypotheses being considered for a noninferiority binomial trial and formalize the statistical modeling framework. Then we discuss the implementation of Chan's exact p-value followed by introducing the δ-projected Z-score as the choice of the test statistic for building an asymptotic confidence interval, which also serves as the basis of the Chan & Zhang confidence interval. We then introduce the proposed EC δ-projected confidence interval method along with its favorable properties in size and power. We finally illustrate those properties through carefully conducted simulations and real data examples.

2 | METHODS

We consider a noninferiority trial with treatment group (T) and control/standard-of-care group (C) having a binary endpoint representing whether or not an outcome is observed. Let P_T and P_C be the probabilities the outcome is observed, and let $\delta = P_T - P_C$ represent the risk difference. Depending on whether we are considering a positive outcome (e.g., resolution of a disease) or a negative outcome (e.g., cancer recurrence), we will use the following hypotheses for a noninferiority trial with prespecified noninferiority margin $\delta_0 \geq 0$ (Table 1).

We will consider a positive outcome for the rest of this paper. We model the binary outcomes of the treatment and control groups with the following binomial distributions:

$$X_T \sim \text{Binomial}(N_T, P_T)$$

$$X_C \sim \text{Binomial}(N_C, P_C).$$

Under this binomial model, the joint probability for $X_T = x_T$ and $X_C = x_C$ is

$$\Pr(X_T = x_T, X_C = x_c \mid P_T, P_C) = \binom{N_T}{x_T}\binom{N_C}{x_C}(P_T)^{x_T}(1-P_T)^{N_T-x_T}P_C^{x_C}(1-P_C)^{N_C-x_C},$$

where $0 \leq x_T \leq N_T$ and $0 \leq x_C \leq N_C$. The likelihood can be reformulated in terms of P_T and δ with the substitution $P_C = P_T - \delta$.

$$\Pr(X_T = x_T, X_C = x_c \mid P_T, \delta) = \binom{N_T}{x_T}\binom{N_C}{x_C}(P_T)^{x_T}(1-P_T)^{N_T-x_T} \times (P_T-\delta)^{x_C}(1+\delta-P_T)^{N_C-x_C}, \quad (1)$$

where P_T and δ must satisfy the condition

$$\max(0, \delta) \leq P_T \leq \min(1, 1+\delta).$$

If $\delta_0 = 0$, then $X_T + X_C$ is a sufficient statistic for P_T under the null hypothesis, which forms the basis of Fisher's exact test procedure (Fisher, 1935). However, for the more general setup of $\delta_0 \neq 0$, a different approach is needed. One such approach is described in the next section.

2.1 | Chan's exact test

Chan (1998), and subsequently Röhmel and Mansmann (1999b), proposed an unconditional exact p-value approach based on the maximization/minimax principle. This approach starts with specifying a preorder on the sample space

$$\Omega = \{(x_T, x_C) : 0 \leq x_T \leq N_T \quad \text{and} \quad 0 \leq x_C \leq N_C\}.$$

More specifically, given the preorder, we can index and arrange the n elements of Ω in the following manner:

$$\omega_1 \leq \omega_2 \leq \cdots \leq \omega_n.$$

A natural approach to specifying a preorder is to use a test statistic, or more generally any function, that maps the elements of Ω to $\mathbb{R}$. Let S be a statistic that induces a preorder on Ω. Without loss of generality, suppose that greater values of S favor the alternative hypothesis (otherwise replace S with $-S$). This statistic will likely depend on N_T and N_C and may also depend on δ_0. The so-called exact unconditional p-value, as described in Chan (1998), Röhmel and Mansmann (1999b), and Chan (2003), can be expressed as

$$p_S^{\text{exact}}(x_T, x_C) = \sup_{P_T, \delta \leq -\delta_0} \Pr\left[S(X_T, X_C) \geq S(x_T, x_C) \mid P_T, \delta\right]. \quad (2)$$

This approach uses the maximization/minimax principle (Lehmann and Romano, 2006; Basu, 2011) to eliminate the nuisance parameters P_T and δ. Taking the supremum over both P_T and δ can be simplified when the statistic S satisfies the so-called Barnard criteria, stemming from Barnard (1947), which is given by the following two conditions:

$$\begin{aligned} S(x_T, x_C) &\geq S(x_T, x_C+1) \text{forall} (x_T, x_C), (x_T, x_C+1) \in \Omega \\ S(x_T, x_C) &\geq S(x_T-1, x_C) \text{forall} (x_T, x_C), (x_T-1, x_C) \in \Omega. \end{aligned} \tag{3}$$

This condition is intuitively clear: for any observed outcome, having one more success in the control or one less success in the treatment should lead to a smaller value of the test statistic. Röhmel and Mansmann (1999b) proved that when the inequalities in (3) are satisfied, the supremum in (2) occurs on the boundary of H_0; that is, the supremum is the maximum under the restriction $P_T - P_C = -\delta_0$. Frick (2000) generalized this and proved that if either inequality in (3) is satisfied, then again the supremum in (2) occurs at the boundary of H_0. We shall assume that the statistic S satisfies (3), thus allowing us to rewrite Equation (2) as follows:

$$\begin{aligned} p_S^{\text{exact}}(x_T, x_C) = \max_{P_T \in [0, 1-\delta_0]} &\Pr[S(X_T, X_C) \\ &\geq S(x_T, x_C) \mid P_T, \delta = -\delta_0]. \end{aligned} \tag{4}$$

For a general statistic p and specified level α, we define the critical region, Ω_α, to be the set of elements of Ω that reject the null hypothesis based on a level α test; that is,

$$\Omega_\alpha = \{(x_T, x_C) \text{ such that } p(x_T, x_C) \leq \alpha\}. \tag{5}$$

Conditioned on P_T and δ, we define the conditional size of p under the null to be

$$\alpha(p \mid P_T, \delta) = \sum_{(x_T, x_C) \in \Omega_\alpha} \Pr[X_T = x_T, X_C = x_C \mid P_T, \delta], \tag{6}$$

and the maximal size is defined as

$$\alpha^*(p) = \sup_{P_T, \delta \leq -\delta_0} \alpha(p \mid P_T, \delta). \tag{7}$$

Following terminology of Berger and Boos (1994) and Röhmel and Mansmann (1999a), we will call p a *valid p-value* if

$$\alpha^*(p) \leq \alpha \text{ for all } \alpha \in [0, 1]. \tag{8}$$

In Theorem A.1, proved in Appendix A, we show that $p_S^{\text{exact}}(x_T, x_C)$ is a valid p-value. Ultimately, we will propose a confidence interval for δ that corresponds to p_S^{exact} for a particular choice of S—the so-called δ-projected Z-score—which is described in the next section.

2.2 | δ-Projected Z-score

There are several choices of statistics to define the preorder in Chan's method including Dunnett and Gent (1977), Santner and Snell (1980), Blackwelder (1982), Miettinen and Nurminen (1985), Farrington and Manning (1990), Chan and Zhang (1999), and Röhmel and Mansmann (1999b). Chan (2003, 1998) is particularly favorable to what he calls the δ-projected Z-score, originally described in Miettinen and Nurminen (1985), given by

$$Z_\delta(X_T, X_C) = \frac{\widehat{P}_T - \widehat{P}_C + \delta}{\widehat{\sigma}_\delta}, \tag{9}$$

where

$$\begin{aligned} \widehat{P}_T &= \frac{X_T}{N_T}, \quad \widehat{P}_C = \frac{X_C}{N_C}, \\ \widehat{\sigma}_\delta &= \sqrt{\frac{\widetilde{P}_T(1-\widetilde{P}_T)}{N_T} + \frac{\widetilde{P}_C(1-\widetilde{P}_C)}{N_C}}, \end{aligned} \tag{10}$$

and $\widetilde{P}_T$ and $\widetilde{P}_C$ represent the maximum likelihood estimators of P_T and P_C, respectively, under the null hypothesis constraint $\widetilde{P}_T - \widetilde{P}_C = -\delta$. In particular, in calculating the exact p-value with Equation (4), Chan (2003, 1998) advocates the use of the statistic $S(X_T, X_C) = Z_{\delta_0}(X_T, X_C)$. We will simply write $p^{\text{exact}}(x_T, x_C)$ to refer to Chan's exact p-value with this statistic; that is,,

$$\begin{aligned} p^{\text{exact}}(x_T, x_C) = \max_{P_T \in [0, 1-\delta_0]} &\Pr[Z_{\delta_0}(X_T, X_C) \\ &\geq Z_{\delta_0}(x_T, x_C) \mid P_T, \delta = -\delta_0]. \end{aligned} \tag{11}$$

Chan (1999, 2003) has provided justification that $Z_{\delta_0}(X_T, X_C)$ satisfies the Barnard criteria (conditions in Equation (3)), so as in Equation (4), the maximization in Equation (11) occurs on the boundary of H_0. Closed formulas for calculating the restricted maximum likelihood estimators $\widetilde{P}_T$ and $\widetilde{P}_C$ are given in Miettinen and Nurminen (1985) and Farrington and Manning (1990).

Asymptotically, $Z_{-\delta}(X_T, X_C)$ has a standard normal distribution, so it can be used as an asymptotic pivotal quantity to form a confidence set for δ (Miettinen and Nurminen, 1985). In particular, if Z_δ is monotonic in δ, this confidence set would be a continuous interval. We do not prove monotonicity due to the complexity of the statistic, but monotonicity is easy to confirm for any given set of parameters. We numerically confirmed that $Z_\delta(x_T, x_C)$ is monotonically increasing for all tables up to $N_T = 100$ and $N_C = 100$ with a grid size on δ of 0.01. Therefore, in the discussion below, we simply assume that $Z_\delta(x_T, x_C)$ is monotonically increasing in δ.

The asymptotic $(1-\alpha)$ confidence interval of Miettinen and Nurminen (1985) is $(\delta^{\text{asy}}_{L,\alpha}, \delta^{\text{asy}}_{U,\alpha})$, where $\delta^{\text{asy}}_{L,\alpha}$ and $\delta^{\text{asy}}_{U,\alpha}$ satisfy

$$\mathfrak{z}_{1-\alpha/2} = Z_{-\delta^{\text{asy}}_{L,\alpha}}(x_T, x_C) \quad \text{and} \quad \mathfrak{z}_{\alpha/2} = Z_{-\delta^{\text{asy}}_{U,\alpha}}(x_T, x_C); \tag{12}$$

the notation $\mathfrak{z}_\alpha$ represents the α quantile of the standard normal distribution. The following probability coverage calculation validates this asymptotic confidence interval formulation:

$$\Pr\left[\delta^{\text{asy}}_{L,\alpha} \le \delta \le \delta^{\text{asy}}_{U,\alpha}\right] = \Pr\left[-\delta^{\text{asy}}_{U,\alpha} \le -\delta \le -\delta^{\text{asy}}_{L,\alpha}\right] \tag{13}$$

$$= \Pr\left[Z_{-\delta^{\text{asy}}_{U,\alpha}}(X_T, X_C) \le Z_{-\delta}(X_T, X_C) \le Z_{-\delta^{\text{asy}}_{L,\alpha}}(X_T, X_C)\right] \tag{14}$$

$$= \Pr\left[\mathfrak{z}_{\alpha/2} \le Z_{-\delta}(X_T, X_C) \le \mathfrak{z}_{1-\alpha/2}\right] \tag{15}$$

$$\approx 1 - \alpha. \tag{16}$$

We define the asymptotic p-value to be

$$\begin{aligned} p^{\text{asy}}(x_T, x_C) &= \Pr\left[Z \ge Z_{\delta_0}(x_T, x_C)\right] \\ &= 1 - \Phi\big(Z_{\delta_0}(x_T, x_C)\big), \end{aligned} \tag{17}$$

where Z represents the standard normal distribution and $\Phi(x)$ is the cumulative distribution function for the standard normal distribution. Theorem A.2, proved in Appendix A, establishes a connection between $p^{\text{asy}}(x_T, x_C)$ and $\delta^{\text{asy}}_{L,\alpha}$. Namely, assuming that $Z_\delta(x_T, x_C)$ is monotonically increasing in δ, then $p^{\text{asy}}(x_T, x_C)$ is less than $\alpha/2$ if and only if $\delta^{\text{asy}}_{L,\alpha}$ is larger than $-\delta_0$.

We note that the confidence interval $(\delta^{\text{asy}}_{L,\alpha}, \delta^{\text{asy}}_{U,\alpha})$, which is based on Z_δ, does not depend on the prespecified value of δ_0. It is also noted that $\alpha^*(p^{\text{asy}})$ is not necessarily bounded by $\alpha/2$; furthermore, there are many examples in which the type-I error exceeds the level $\alpha/2$ (see Section 3.4 below). Hence, p^{asy} is not a valid p-value per the definition stated above in Section 2.1. Next, we describe a confidence interval that utilizes Chan's exact p-value with guaranteed probability coverage.

2.3 | Chan and Zhang confidence interval

Chan and Zhang (1999) proposed an "exact" two-sided $(1-\alpha)\%$ confidence interval for the risk difference δ. The method is based on inverting two one-sided hypotheses using the δ-projected Z-score Z_δ. We first define the following quantities:

$$\begin{aligned} P_{L,\delta}(x_T, x_C) = \max_{P_T \in [0, 1-\delta]} \Pr[Z_{-\delta}(X_T, X_C) \\ \ge Z_{-\delta}(x_T, x_C)], \\ P_{U,\delta}(x_T, x_C) = \max_{P_T \in [0, 1-\delta]} \Pr[Z_{-\delta}(X_T, X_C) \\ \le Z_{-\delta}(x_T, x_C)]. \end{aligned}$$

In particular, we note that the exact p-value, given in Equation (11), can be written as

$$p^{\text{exact}}(x_T, x_C) = P_{L,-\delta_0}(x_T, x_C).$$

The Chan & Zhang confidence interval, denoted as $(\delta^{\text{CZ}}_{L,\alpha}, \delta^{\text{CZ}}_{U,\alpha})$, is defined by the following expressions:

$$\begin{aligned} \delta^{\text{CZ}}_{L,\alpha}(x_T, x_C) &= \inf_{\delta}\left\{\delta : P_{L,\delta}(x_T, x_C) > \alpha/2\right\}, \\ \delta^{\text{CZ}}_{U,\alpha}(x_T, x_C) &= \sup_{\delta}\left\{\delta : P_{U,\delta}(x_T, x_C) > \alpha/2\right\}. \end{aligned} \tag{18}$$

It is noted that this confidence interval, like the asymptotic confidence interval in the previous section, does not depend on the noninferiority margin δ_0. We correspond the Chan & Zhang confidence interval with a Chan & Zhang p-value defined as

$$p^{\text{CZ}}(x_T, x_C) = \max_{\delta \in [-1, -\delta_0]} P_{L,\delta}(x_T, x_C), \tag{19}$$

where the correspondence is established in Theorem A.3 proved in Appendix A. Theorem A.3 also shows that $p^{\text{CZ}}(x_T, x_C)$ is bounded below by $p^{\text{exact}}(x_T, x_C)$, which is equivalent to the following statement:

$$p^{\text{CZ}}(x_T, x_C) \text{ rejects } H_0 \Rightarrow p^{\text{exact}}(x_T, x_C) \text{ rejects } H_0.$$

So, anytime the Chan & Zhang confidence interval rejects H_0, the exact test will also necessarily reject H_0, but the converse is not always true. This indicates that the exact test will have at least as much statistical power as the test induced by the Chan & Zhang confidence interval.

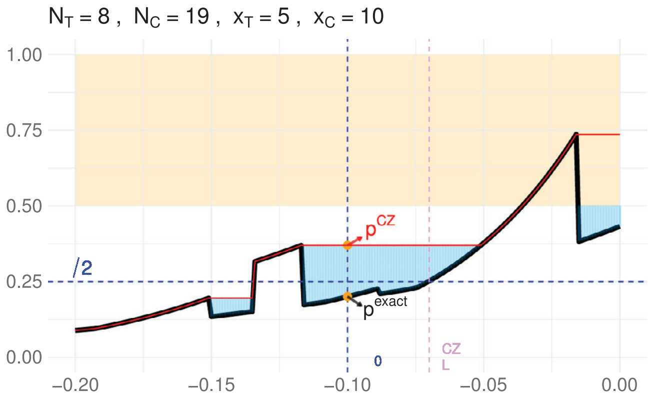

FIGURE 1 This graphic shows the relationship between $p^{\mathrm{CZ}}(x_T, x_C)$ (red line) and $p^{\mathrm{exact}}(x_T, x_C)$ (black line). The regions shaded in blue indicate the values of $\alpha/2$ and δ_0 what would cause $p^{\mathrm{exact}}(x_T, x_C)$ to reject H_0 and $p^{\mathrm{CZ}}(x_T, x_C)$ not to reject H_0 for a level α test

Figure 1 illustrates the relationship between $p^{\mathrm{CZ}}(x_T, x_C)$ and $p^{\mathrm{exact}}(x_T, x_C)$ with a concrete example taking $N_T = 9$, $N_C = 19$, $x_T = 5$, and $x_C = 10$. In the figure, the black line is $P_{L,\delta}(x_T, x_C)$ and the red line is $\max_{\delta\in[-1,-\delta]} P_{L,\delta}(x_T, x_C)$. The values for which these lines intersect at $\delta = -\delta_0$ correspond to $p^{\mathrm{exact}}(x_T, x_C)$ and $p^{\mathrm{CZ}}(x_T, x_C)$, respectively, as indicated on the figure with $\delta_0 = 0.1$. The regions shaded in blue indicate values of $\alpha/2$ and δ_0 that would cause $p^{\mathrm{exact}}(x_T, x_C)$ to reject H_0 and $p^{\mathrm{CZ}}(x_T, x_C)$ not to reject H_0 for a level α test. Also illustrated on the graphic is the lower bound of the Chan & Zhang confidence interval for $\alpha = 0.5$; although not a standard significance level, the chosen parameters were selected to clearly highlight key features of p^{exact}. The region shaded in orange is not of interest as the corresponding α would be greater than 1 in this region.

As illustrated in Figure 1, there are many situations in which the strict inequality $p^{\mathrm{CZ}}(x_T, x_C) < p^{\mathrm{exact}}(x_T, x_C)$ holds. Using terminology described in Röhmel and Mansmann (1999b), we say that p^{exact} *strictly dominates* p^{CZ}. A p-value that is not strictly dominated is called *acceptable*. It is much easier establishing a p-value is not acceptable, like p^{CZ}, than to prove a given p-value, say p_S^{exact}, is acceptable. Frick (2000) provides various necessary and sufficient conditions for acceptable p-values.

Next, we propose a novel "EC" confidence interval, $(\delta_{L,\alpha}^{\mathrm{EC}}, \delta_{U,\alpha}^{\mathrm{EC}})$, which corresponds to p^{exact}; that is, $p^{\mathrm{exact}}(x_T, x_C) \le \alpha/2$ if and only if $\delta_{L,\alpha}^{\mathrm{EC}} > -\delta_0$.

2.4 | Exact-corrected δ-projected Z-score

We consider a modification of the δ-projected Z-score, which we call the EC δ-projected Z-score. This EC δ-projected Z-score, labeled $Z_\delta^{\mathrm{EC}}(X_T, X_C)$, is given in (20). The motivation and derivation of $Z_\delta^{\mathrm{EC}}(X_T, X_C)$ is presented in Appendix B.

$$
\begin{aligned}
Z_\delta^{\mathrm{EC}}(X_T, X_C) &= \frac{\hat{P}_T - \hat{P}_C + \delta}{\hat{\sigma}_\delta} - \frac{\hat{\sigma}_{\delta_0}}{\hat{\sigma}_\delta}\big(Z_{\delta_0}(X_T, X_C) \\
&\quad - \Phi^{-1}\big(1 - p^{\mathrm{exact}}(X_T, X_C)\big)\big) \\
&= \frac{\hat{P}_T - \hat{P}_C + \delta}{\hat{\sigma}_\delta} - \frac{\hat{\sigma}_{\delta_0}}{\hat{\sigma}_\delta}\big(\Phi^{-1}(1 - p^{\mathrm{asy}}(X_T, X_C)) \\
&\quad - \Phi^{-1}(1 - p^{\mathrm{exact}}(X_T, X_C))\big) \\
&= Z_\delta(X_T, X_C) - \mathrm{EC}_\delta(X_T, X_C),
\end{aligned} \tag{20}
$$

where Φ^{-1} denotes the quantile function of the standard normal distribution (also called the probit function), $\hat{\sigma}_\delta$ is as defined in (10), $\hat{\sigma}_{\delta_0}$ is $\hat{\sigma}_\delta$ evaluated at $\delta = \delta_0$, and $\mathrm{EC}_\delta(X_T, X_C)$ denotes the exact correction term given by

$$
\begin{aligned}
\mathrm{EC}_\delta(X_T, X_C) = \frac{\hat{\sigma}_{\delta_0}}{\hat{\sigma}_\delta}\big(&\Phi^{-1}(1 - p^{\mathrm{asy}}(X_T, X_C)) \\
&- \Phi^{-1}(1 - p^{\mathrm{exact}}(X_T, X_C))\big).
\end{aligned}
$$

In particular, when evaluating $Z_\delta^{\mathrm{EC}}(X_T, X_C)$ at $\delta = \delta_0$, $X_T = x_T$, and $X_C = x_C$, we have

$$
Z_{\delta_0}^{\mathrm{EC}}(x_T, x_C) = \Phi^{-1}\big(1 - p^{\mathrm{exact}}(x_T, x_C)\big). \tag{21}
$$

In the simulation section, we numerically show that the expectation of $\mathrm{EC}_\delta(X_T, X_C)$ tends to zero with increasing sample size over selected values of P_T, P_C, and δ_0. We do not prove monotonicity due to the complexity of the statistic, but monotonicity is easy to confirm for any given set of parameters. The software that we developed to implement the EC methods (detailed in Section 3.1) confirms monotonicity of $Z_\delta^{\mathrm{EC}}(x_T, x_C)$ in δ when applied to user-specified data. In the subsequent discussion, we assume that $Z_\delta^{\mathrm{EC}}(x_T, x_C)$ is monotonic in δ, so inverting $Z_\delta^{\mathrm{EC}}(x_T, x_C)$ will produce $(1-\alpha)\%$ "EC" confidence interval, $(\delta_{L,\alpha}^{\mathrm{EC}}, \delta_{U,\alpha}^{\mathrm{EC}})$, defined by the following equations

$$
\mathfrak{z}_{1-\alpha/2} = Z_{-\delta_{L,\alpha}^{\mathrm{EC}}}(X_T, X_C) \quad \text{and} \quad \mathfrak{z}_{\alpha/2} = Z_{-\delta_{U,\alpha}^{\mathrm{EC}}}(X_T, X_C). \tag{22}
$$

A connection between $\delta_{L,\alpha}^{\mathrm{EC}}$, derived in (22), and $p^{\mathrm{exact}}(x_T, x_C)$, is established in Theorem A.4 (proved in Appendix A). Namely, assuming that $Z_\delta^{\mathrm{EC}}(x_T, x_C)$ is monotonically increasing in δ, then $p^{\mathrm{exact}}(x_T, x_C)$ is less than $\alpha/2$ if and only if $\delta_{L,\alpha}^{\mathrm{EC}}(x_T, x_C)$ is larger than $-\delta_0$.

3 | SIMULATIONS AND EXAMPLES

3.1 | Software and data sharing

The R package `EC` (see Supporting Information) allows the user to easily compute the confidence intervals, p-values, and maximal sizes discussed in this paper. Examples are provided in the package for each of these functions. The function to calculate the Chan & Zhang confidence intervals was verified against SAS PROC FREQ by using the EXACT statement within the TABLES statement (using the RISKDIFF statistic option coupled with the SCORE method). Automated checking of the functions was performed with the `devtools::check` function. All data that support the findings of this research are provided within the paper.

3.2 | Asymptotic assessments

Here, we present simulations to show the asymptotic behavior of the expected value of $\text{EC}_\delta(X_T, X_C)$. Seven different sample sizes of $N = N_T = N_C$, doubling each time from 10 to 640, were considered along with three different values of δ_0 (0, 0.1, 0.2), three different values of P_T (0.3, 0.5, 0.7), and three different values of P_C (0.3, 0.5, 0.7). Each expected value is computed over 10,000 realizations of the data, thus yielding very precise estimates. Figure 2 shows $\boldsymbol{E}[\text{EC}_\delta(X_T, X_C)] \approx 0$ for large values of N. This, in turn, suggests that $Z_\delta^{\text{EC}}(X_T, X_C)$ is close to $Z_\delta(X_T, X_C)$ for large N.

3.3 | Power and size

The performance of the confidence interval estimators is compared in terms of power and size.

- The method "MN" (Miettinen and Nurminen, 1985) corresponds to $(\delta_{L,\alpha}^{\text{asy}}, \delta_{U,\alpha}^{\text{asy}})$.
- The method "CZ" (Chan and Zhang, 1999) corresponds to $(\delta_{L,\alpha}^{\text{CZ}}, \delta_{U,\alpha}^{\text{CZ}})$.
- The method "EC" corresponds to our proposed "exact-corrected" confidence interval estimator $(\delta_{L,\alpha}^{\text{EC}}, \delta_{U,\alpha}^{\text{EC}})$.

The two examples displayed in Figure 3 showcase the potential differences in power and size across the three methods. Once the values of P_T, N_T, N_C, δ_0, and α are determined, the probability of rejecting H_0 for different values of δ is calculated from the likelihoods of the $(N_T + 1)(N_C + 1)$ tables using Equation (1). Values of δ that are smaller than $-\delta_0$ correspond to H_0 being true, whereas values of δ that are larger than $-\delta_0$ correspond to H_1 being true. The values `n.AA`, `n.AR`, and `n.RR` displayed on the graphic represent the number of the $(N_T + 1)(N_C + 1)$ tables for which CZ and EC both accept H_0 (`n.AA`), CZ accepts H_0 and EC rejects H_0 (`n.AR`), and CZ and EC both reject H_0 (`n.RR`). Note that the term "accept" here is used synonymously with "failed to reject."

Figure 3a sets $P_T = 0.95$, $N_T = 5$, $N_C = 11$, $\delta_0 = 0.03$, and $\alpha = 0.7$. In this example, there are four tables for which EC rejects H_0 but CZ fails to reject H_0. This causes EC to have greater power compared to CZ, though both methods have controlled size under H_0. We also see that the MN method has better power than both CZ and EC, but also rejects H_0 with probability greater than α when H_0 is true.

Figure 3b sets $P_T = 0.1$, $N_T = 12$, $N_C = 5$, $\delta_0 = 0.33$, and $\alpha = 0.1$. In this example, there is only one table for which EC rejects H_0 but CZ fails to reject H_0, yet this one table occurs with a high enough probability to produce a measurable difference in power between the EC and CZ methods. The MN and EC methods reject/accept H_0 for all tables (even though they produce different confidence intervals) causing them to have identical power curves. In this example, all methods have a type-I error that is bounded by α.

3.4 | Data examples

In addition to the three previously discussed methods—EC, CZ, and MN—we also consider the commonly used Wald's method (Altman *et al.*, 2013; Fagerland *et al.*, 2015). The Wald Z-statistic is given by

$$Z^{\text{Wald}} = \frac{\hat{P}_T - \hat{P}_C + \delta_0}{\sqrt{\frac{\hat{P}_T(1-\hat{P}_T)}{N_T} + \frac{\hat{P}_C(1-\hat{P}_C)}{N_C}}},$$

and the corresponding confidence interval and p-value are given by

$$\hat{P}_T - \hat{P}_C \pm z_{1-\alpha/2}\sqrt{\frac{\hat{P}_T(1-\hat{P}_T)}{N_T} + \frac{\hat{P}_C(1-\hat{P}_C)}{N_C}}$$

$$p^{\text{Wald}} = 1 - \Phi(Z^{\text{Wald}}).$$

We first present three examples in which the EC and CZ confidence intervals produce different hypothesis test decisions. As shown in Theorem A.3, $p^{\text{exact}} \leq p^{CZ}$, so if the hypothesis test decisions differ between EC and CZ, it must be that EC rejects the null and CZ fails to reject the null. The parameters for the first example are the same as the parameters presented in Figure 1. The second and third examples also show advantages of the EC method over the CZ method but with the more standard $\alpha = 0.05$. Confidence intervals for all four methods are presented in

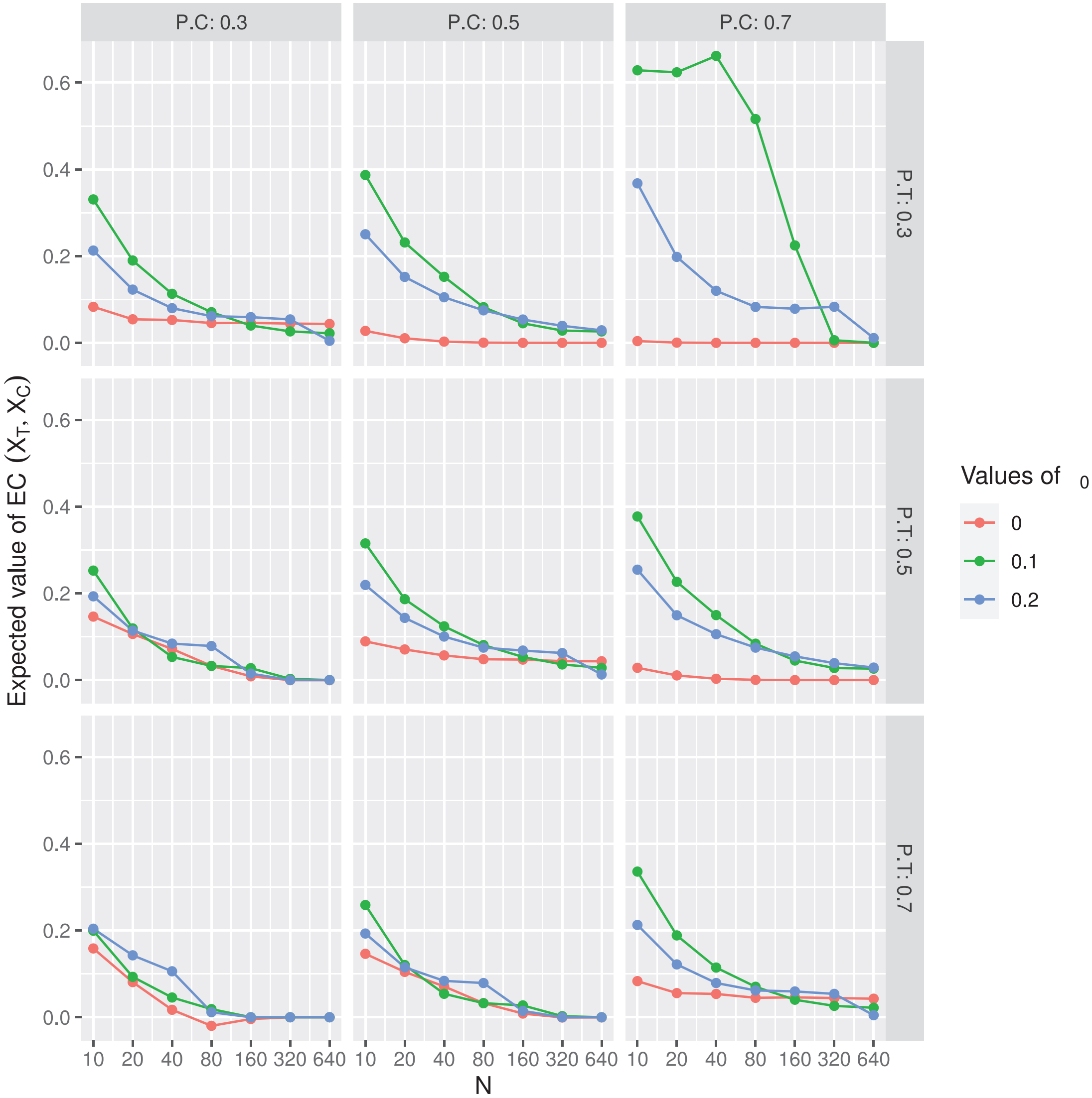

FIGURE 2 Expected value of $\mathrm{EC}_\delta(X_T, X_C)$ is approximated for different combinations of P_T, P_C, δ_0, and $N = N_T = N_C$

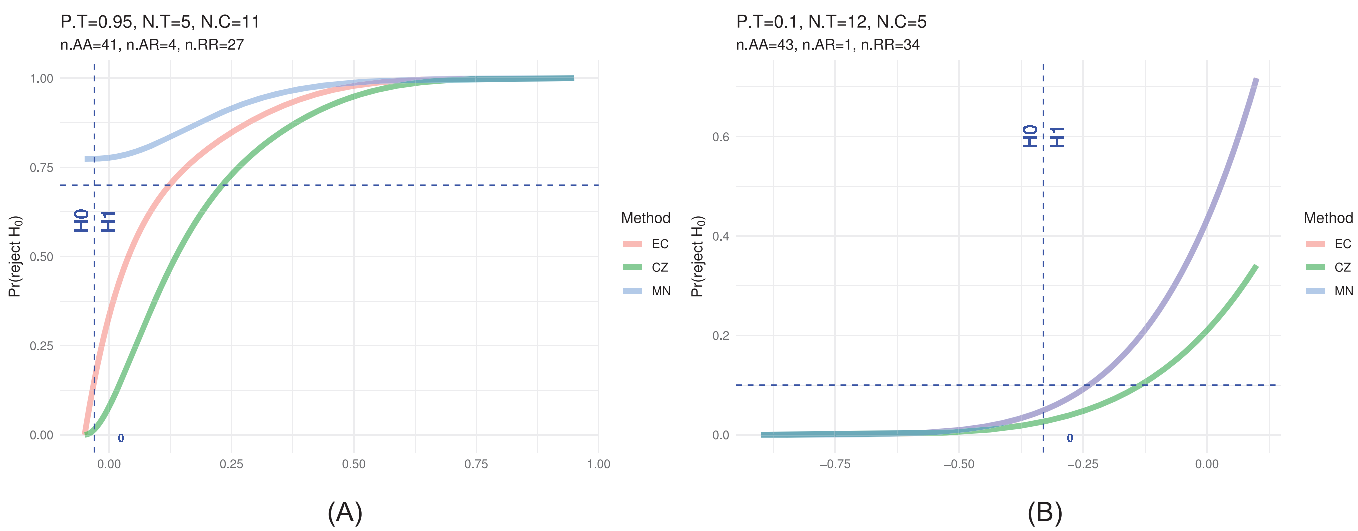

FIGURE 3 The probability of rejecting H_0 over different values of δ for the three confidence interval methods EC, CZ, and MN

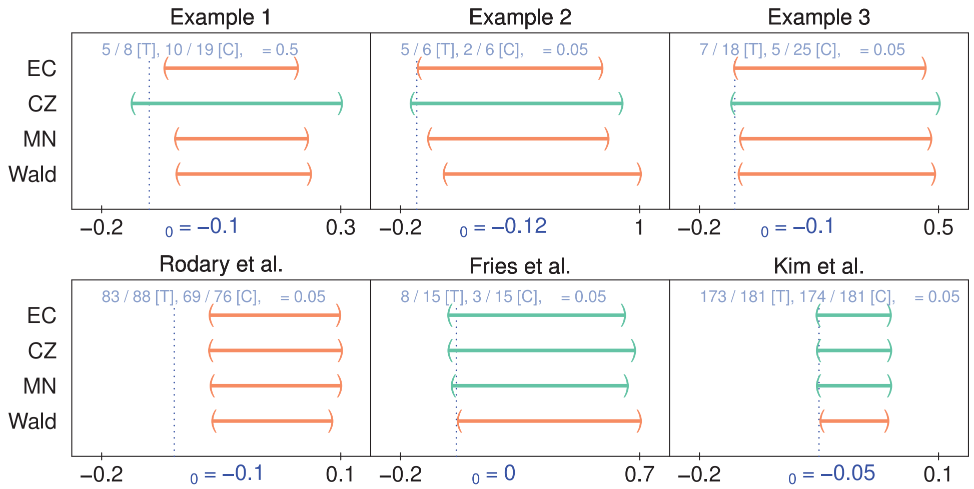

FIGURE 4 Comparison of the four confidence interval methods over the six data examples presented in Section 3.4

TABLE 2 Three examples show that the EC method is less conservative than the CZ method yet still corresponding to an exact test that controls type-I error

	Data parameters				*p*-Values				maximal size			
	x_T/N_T	x_C/N_C	δ_0	$\alpha/2$	EC	CZ	MN	Wald	EC	CZ	MN	Wald
Example 1	5/8	10/19	0.10	0.25	0.200	0.370	0.172	0.167	0.197	0.197	0.430	0.430
Example 2	5/6	2/6	0.12	0.025	0.023	0.030	0.014	0.006	0.022	0.012	0.030	0.464
Example 3	7/18	5/25	0.10	0.025	0.024	0.027	0.018	0.020	0.024	0.021	0.028	0.150

Figure 4. Table 2 shows the associated *p*-values and maximal sizes for the four methods. We remind the reader here that the *p*-values and maximal sizes associated with the EC method are equivalent to p^{exact} and $\alpha^*(p^{\text{exact}})$, respectively. Consistent with Theorems A.2, A.3, and A.4, we see that the associated *p*-values are shown to declare noninferiority (reject the null with the corresponding *p*-value being less than $\alpha/2$) if and only if the lower bound of the respective confidence intervals is bigger than $-\delta_0$.

Theorem A.3 also establishes the following inequalities on the maximal sizes of the CZ and EC methods:

$$\alpha^*(p^{\text{CZ}}) \le \alpha^*(p^{\text{exact}}) \le \alpha/2.$$

Note that maximal size does not depend on specific values of x_T and x_C. The maximal size calculations shown in Table 2 demonstrate this conservativeness of the CZ method over the EC method. Additionally, in each of these three examples, the maximal sizes for both the MN and Wald methods exceed the $\alpha/2$ threshold, which shows that the type-I error rates associated with confidence intervals produced from the MN and Wald methods can be inflated.

In the next three examples, we compare the different methods from published studies. The first example, originally published in Lemerle *et al.* (1983) and later reanalyzed in Rodary *et al.* (1989) and Chan (1998), considers a randomized trial in childhood nephroblastoma comparing neoadjuvant chemotherapy (treatment) to radiation therapy (control) with the outcome of preventing tumor rupture during surgery. A noninferiority margin is taken to be $\delta_0 = 0.1$, and the chemotherapy treatment would be considered noninferior to radiation if $\delta = P_T - P_C > -0.1$. Eighty three of the 88 chemotherapy subjects had a positive outcome ($\hat{P}_T = 0.943$), and 69 of the 76 radiation subjects had a positive outcome ($\hat{P}_C = 0.908$). There is pretty strong evidence that neoadjuvant chemotherapy is not inferior to radiation in preventing surgical tumor rupture in this study.

The second example, originally published in Fries *et al.* (1993) and later reanalyzed in Chan (1998), considers the protective efficacy against illness of a recombinant protein flu vaccine in response to exposure to the H1N1 virus. The noninferiority margin is set to $\delta_0 = 0$, so the vaccine would be considered meaningful if $\delta = P_T - P_C > 0$. Eight out of 15 subjects who received the vaccine (treatment group) avoided any kind of clinical illness ($\hat{P}_T = 0.533$), whereas only 3 out of 15 subjects who received the placebo (control group) avoided illness ($\hat{P}_C = 0.200$). Even with the small sample size, this study gives pretty strong evidence that the recombinant protein vaccine is more effective than a placebo in preventing illness from exposure to H1N1.

The third example, published in Kim *et al.* (2013), considers whether the success rate of subclavian venous catheterization using a neutral shoulder position (treatment group) is not inferior to the often recommended retracted shoulder position (control group). The noninferiority margin is set to $\delta_0 = 0.05$, so the neutral shoulder position would be noninferior if $\delta = P_T - P_C > -0.05$. One hundred seventy three out of 181 subjects in the neutral position had a successful catheterization ($\widehat{P}_T = 0.956$), and 174 out of 181 subjects in the retracted position had a successful catheterization ($\widehat{P}_C = 0.961$). The success rates in this study are quite comparable for the two groups. We note that this study reports a confidence interval for δ based on Wald's method, which we show produces a decision that is inconsistent with the other three methods.

The four confidence interval methods for each of these three examples with $\alpha = 0.05$ are presented in Figure 4. In the Rodary et al. example, all confidence intervals have a lower bound around -0.05 declaring noninferiority for the chemotherapy treatment. In the Fries et al. example, EC, CZ, and MN intervals report a different decision compared to the Wald interval. The Fisher's exact and Chan's exact p-values for this example are 0.128 and 0.008, respectively, thus leading to a different statistical decision based on a 5% level test. The lower bound of the EC and CZ intervals match, but the upper bound of the EC interval is somewhat smaller. The Kim et al. example has relatively large sample sizes, so the EC, CZ, and MN methods produce confidence intervals that are all fairly similar to each other. However, these three methods produce a different conclusion about noninferiority compared to the Wald interval that was reported in the paper. That is, the EC, CZ, and MN intervals fail to conclude noninferiority at the 5% margin, whereas the Wald interval establishes noninferiority.

4 | DISCUSSION

A novel confidence interval estimator is proposed that bridges the divide between the generally more powerful asymptotic confidence interval of Miettinen and Nurminen (1985) and the less powerful but correctly sized exact confidence interval of Chan and Zhang (1999) to yield a correctly sized exact confidence interval that is more powerful than the Chan & Zhang interval under the setting of a noninferiority design. The proposed confidence interval fully leverages the noninferiority trial design by incorporating the noninferiority margin, whereas the other methods do not involve the prespecified noninferiority margin.

It is natural to expect confidence intervals for the effect size δ not to depend on noninferiority design δ_0. However, the Chan exact p-value requires the specification of δ_0 in addition to usual data values that go into confidence interval constructions (x_T, x_C, N_T, N_C). Therefore, any confidence interval method that is consistent with the Chan exact p-value will naturally be dependent on δ_0. We note that with small sample sizes, Chan's exact p-value estimator can behave somewhat erratic as demonstrated in Figure 1. This leads to the lower bound of the proposed estimator potentially being sensitive to δ_0 in the presence of small sample sizes.

For larger sample sizes, the methods all produce similar confidence intervals, but the Chan & Zhang confidence interval method is substantially more computationally demanding. Moderate sample sizes, such as the examples explored in Section 3.4, can take from several minutes to several hours, depending on the level of precision required, whereas the other methods, including the proposed method, will compute the confidence interval within a few seconds.

The differences in results are usually not very dramatic, but with smaller sample sizes and certain values of the parameters P_T, α, and δ_0, the proposed method can provide a pretty substantial improvement in power as demonstrated in Section 3.3. Theorems A.3 and A.4 also theoretically establish that the proposed EC confidence interval estimator is at least as powerful as the Chan & Zhang confidence interval estimator and that they both correspond to valid p-value estimators with controlled size. Therefore, the proposed EC risk difference confidence interval estimator is recommended for noninferiority binomial trials as it is computationally efficient, preserves the type-I error, and has improved power over the Chan & Zhang interval. Finally, we briefly mention that the EC confidence interval has a one-to-one correspondence with Chan's exact test, but a much different and modern approach that decouples this link and uses randomized-based interval estimation is worth considering; see Wang and Rosenberger (2020) for more discussion on this topic.

ACKNOWLEDGMENTS

We appreciate the comments and revision suggestions of the editor, associate editor, and three anonymous reviewers that led to a much-improved manuscript.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at https://github.com/NourHawila/EC.

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are openly available in the R package `EC` accessible from GitHub at https://github.com/NourHawila/EC.

ORCID

Arthur Berg https://orcid.org/0000-0002-4097-7348

REFERENCES

Altman, D., Machin, D., Bryant, T. & Gardner, M. (2013) *Statistics with confidence: confidence intervals and statistical guidelines*. New York: John Wiley & Sons.

Barnard, G. (1947) Significance tests for 2 x 2 tables. *Biometrika*, 34(1/2), 123–138.

Basu, D. (2011) On the elimination of nuisance parameters. In: Basu, D. & DasGupta, A. (Eds.) *Selected works of Debabrata Basu*. New York: Springer, pp. 279–290.

Berger, R.L. & Boos, D.D. (1994) P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427), 1012–1016.

Blackwelder, W.C. (1982) "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials*, 3(4), 345–353.

Boschloo, R. (1970) Raised conditional level of significance for the 2 x 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24(1), 1–9.

Chan, I.S. (1998) Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine*, 17(12), 1403–1413.

Chan, I.S. (1999) Author's reply on 'exact tests of equivalence and efficacy with a non-zero lower bound for comparative trials'. *Statistics in Medicine*, 18, 1735–1737.

Chan, I.S. (2003) Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. *Statistical Methods in Medical Research*, 12(1), 37–58.

Chan, I.S. & Zhang, Z. (1999) Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics*, 55(4), 1202–1209.

Dunnett, C. & Gent, M. (1977) Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 x 2 tables. *Biometrics*, 33, 593–602.

Fagerland, M.W., Lydersen, S. & Laake, P. (2015) Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, 24(2), 224–254.

Farrington, C.P. & Manning, G. (1990) Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9(12), 1447–1454.

Fisher, R.A. (1935) *The design of experiments*. Edinburgh, UK: Oliver & Boyd.

Frick, H. (2000) Undominated p-values and property c for unconditional one-sided two-sample binomial tests. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 42(6), 715–728.

Fries, L.F., Dillon, S.B., Hildreth, J.E., Karron, R.A., Funkhouser, A.W., Friedman, C.J., Jones, C.S., Culleton, V.G. & Clements, M.L. (1993) Safety and immunogenicity of a recombinant protein influenza a vaccine in adult human volunteers and protective efficacy against wild-type H1N1 virus challenge. *Journal of Infectious Diseases*, 167(3), 593–601.

Haber, M. (1986) An exact unconditional test for the 2 x 2 comparative trial. *Psychological Bulletin*, 99(1), 129.

Kim, H., Jung, S., Min, J., Hong, D., Jeon, Y. & Bahk, J.-H. (2013) Comparison of the neutral and retracted shoulder positions for infraclavicular subclavian venous catheterization: a randomized, non-inferiority trial. *British Journal of Anaesthesia*, 111(2), 191–196.

Lehmann, E.L. & Romano, J.P. (2006) *Testing statistical hypotheses*. New York: Springer Science & Business Media.

Lemerle, J., Voute, P., Tournade, M., Rodary, C., Delemarre, J., Sarrazin, D., Burgers, J., Sandstedt, B., Mildenberger, H. & Carli, M. (1983) Effectiveness of preoperative chemotherapy in Wilms' tumor: results of an international society of paediatric oncology (siop) clinical trial. *Journal of Clinical Oncology*, 1(10), 604–609.

McDonald, L.L., Davis, B.M. & Milliken, G.A. (1977) A nonrandomized unconditional test for comparing two proportions in 2 x 2 contingency tables. *Technometrics*, 19(2), 145–157.

Miettinen, O. & Nurminen, M. (1985) Comparative analysis of two rates. *Statistics in Medicine*, 4(2), 213–226.

Rodary, C., Com-Nougue, C. & Tournade, M.-F. (1989) How to establish equivalence between treatments: a one-sided clinical trial in paediatric oncology. *Statistics in Medicine*, 8(5), 593–598.

Röhmel, J. & Mansmann, U. (1999a) Re: exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies by ISF Chan, Statistics in Medicine, 17, 1403–1413 (1998). *Statistics in Medicine*, 18(13), 1734–1737.

Röhmel, J. & Mansmann, U. (1999b) Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 41(2), 149–170.

Santner, T.J. & Snell, M.K. (1980) Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2 x 2 contingency tables. *Journal of the American Statistical Association*, 75(370), 386–394.

Suissa, S. & Shuster, J.J. (1985) Exact unconditional sample sizes for the 2 x 2 binomial trial. *Journal of the Royal Statistical Society: Series A (General)*, 148(4), 317–327.

Wang, Y. & Rosenberger, W.F. (2020) Randomization-based interval estimation in randomized clinical trials. *Statistics in Medicine*, 39(21), 2843–2854.

Wasserstein, R.L., Schirm, A.L. & Lazar, N.A. (2019) Moving to a world beyond "p< 0.05". *The American Statistician*, 73(1), 1–19.

SUPPORTING INFORMATION

A ready to run Rmd README file is available with this paper at the Biometrics website on Wiley Online Library. The file describes the R package EC that is developed and used to implement the proposed methods in this paper, demonstrates the installation process and runs a few examples.

Supporting Information

How to cite this article: Hawila, N., Berg, A. (2022) Exact-corrected confidence interval for risk difference in noninferiority binomial trials. *Biometrics*1–12. https://doi.org/10.1111/biom.13688

APPENDIX A: THEOREMS AND PROOFS

Theorem A.1. *Let p_S^{exact} be the exact unconditional* p-value *given in Equation (2). Then p_S^{exact} is a valid* p-value*; that is,*

$$\alpha^*\left(p^{\text{exact}}\right) \leq \alpha \text{for all } \alpha \in [0,1].$$

Proof. Let $\alpha \in [0,1]$, and let Ω_α be the critical region for p^{exact} as defined in Equation (5). Let (x_T^α, x_C^α) be a minimal element of Ω_α; that is,

$$S(x_T, x_C) \geq S(x_T^\alpha, x_C^\alpha) \text{for all } (x_T, x_C) \in \Omega_\alpha.$$

It is noted that this minimal element may not be unique. Define $\widetilde{\Omega}_\alpha$ as follows:

$$\widetilde{\Omega}_\alpha = \left\{(x_T, x_C) \in \Omega \text{such that } S(x_T, x_C) \geq S(x_T^\alpha, x_C^\alpha)\right\}.$$

If $(x_T, x_C) \in \Omega_\alpha$, then $S(x_T, x_C) \geq S(x_T^\alpha, x_C^\alpha)$, as (x_T^α, x_C^α) is a minimal element, which implies $\Omega_\alpha \subset \widetilde{\Omega}_\alpha$. Therefore,

$$\begin{aligned}
&\sup_{P_T, \delta \leq -\delta_0} \left[\sum_{(x_T, x_C) \in \Omega_\alpha} \Pr\left[X_T = x_T, X_C = x_C \mid P_T, \delta\right]\right] \\
&\leq \sup_{P_T, \delta \leq -\delta_0} \left[\sum_{(x_T, x_C) \in \widetilde{\Omega}_\alpha} \Pr\left[X_T = x_T, X_C = x_C \mid P_T, \delta\right]\right] \\
&= p_S^{\text{exact}}(x_T^\alpha, x_C^\alpha) \\
&\leq \alpha.
\end{aligned}$$

$\square$

Theorem A.2. *Let $\delta_{L,\alpha}^{\text{asy}}$ be the asymptotic lower bound given in Equation (12), and let $p^{\text{asy}}(x_T, x_C)$ be the* p-value *defined in Equation (17). Assuming that $Z_\delta(x_T, x_C)$ is monotonically increasing in δ, then*

$$p^{\text{asy}}(x_T, x_C) \leq \alpha/2 \quad \text{ifandonlyif} \quad \delta_{L,\alpha}^{\text{asy}}(x_T, x_C) > -\delta_0.$$

Proof. Suppose $p^{\text{asy}}(x_T, x_C) \leq \alpha/2$. From Equation (17), we have

$$1 - \Phi\left(Z_{\delta_0}(x_T, x_C)\right) \leq \alpha/2,$$

which implies

$$Z_{\delta_0}(x_T, x_C) \geq \Phi^{-1}(1 - \alpha/2) = z_{1-\alpha/2} = Z_{-\delta_{L,\alpha}^{\text{asy}}}(x_T, x_C),$$

using Equation (12). From the assumption that $Z_\delta(x_T, x_C)$ is monotonically increasing in δ, we have that $\delta_{L,\alpha}^{\text{asy}} \geq -\delta_0$. The same steps can be used backward to show that $\delta_{L,\alpha}^{\text{asy}}(x_T, x_C) > -\delta_0$ implies $p^{\text{asy}}(x_T, x_C) \leq \alpha/2$. $\square$

Theorem A.3. *Let $\delta_{L,\alpha}^{\text{CZ}}(x_T, x_C)$ be the Chan & Zhang lower bound given in Equation (18), and let $p^{\text{CZ}}(x_T, x_C)$ be the Chan & Zhang* p-value *given in Equation (19).*

(i) $p^{\text{CZ}}(x_T, x_C) \leq \alpha/2$ if and only if $\delta_{L,\alpha}^{\text{CZ}}(x_T, x_C) > -\delta_0$.
(ii) $p^{\text{CZ}}(x_T, x_C)$ is bounded below by $p^{\text{exact}}(x_T, x_C)$. In particular, the Chan & Zhang p-value *is valid and satisfies the following inequalities:*

$$p^{\text{CZ}}(x_T, x_C) \geq p^{\text{exact}}(x_T, x_C) \tag{A.1}$$

$$\alpha^*(p^{\text{CZ}}) \leq \alpha^*(p^{\text{exact}}) \leq \alpha/2. \tag{A.2}$$

Proof. Suppose

$$p^{\text{CZ}}(x_T, x_C) \triangleq \max_{\delta \in [-1, -\delta_0]} P_{L,\delta}(x_T, x_C) \leq \alpha/2.$$

Then

$$\delta_{L,\alpha}^{\text{CZ}}(x_T, x_C) = \inf_\delta \left\{\delta : P_{L,\delta}(x_T, x_C) > \alpha/2\right\} > -\delta_0,$$

which establishes $\delta_{L,\alpha}^{\text{CZ}} > -\delta_0$.

Now suppose

$$\delta_L^{\text{CZ}}(x_T, x_C) \triangleq \inf_\delta \left\{\delta : P_{L,\delta}(x_T, x_C) > \alpha/2\right\} > -\delta_0.$$

This implies

$$\max_{\delta \in [-1, -\delta_0]} P_{L,\delta}(x_T, x_C) = p^{\text{CZ}}(x_T, x_C) < \alpha/2.$$

This establishes part (i). Part (ii) immediately follows from

$$\begin{aligned}
p^{\text{CZ}}(x_T, x_C) &= \max_{\delta \in [-1, -\delta_0]} P_{L,\delta}(x_T, x_C) \\
&\geq P_{L,-\delta_0}(x_T, x_C) = p^{\text{exact}}(x_T, x_C)
\end{aligned}$$

and the definition of $\alpha^*(\cdot)$ provided in Equation (7). $\square$

Theorem A.4. *Let $\delta_{L,\alpha}^{\text{EC}}$ be the EC lower bound given in Equation (22), and let $p^{\text{exact}}(x_T, x_C)$ be Chan's exact* p-value *based on the δ-projected Z-score as given in Equation (11). Assuming that $Z_\delta^{\text{EC}}(x_T, x_C)$ is monotonically increasing in δ, then*

$$p^{\text{exact}}(x_T, x_C) \leq \alpha/2 \quad \textit{if and only if} \quad \delta_{L,\alpha}^{\text{EC}}(x_T, x_C) > -\delta_0.$$

Proof. Note that

$$\begin{aligned}\Pr\left[Z \geq Z_{\delta_0}^{\mathrm{EC}}(x_T, x_C)\right] &= \Pr\left[Z \geq \Phi^{-1}\left(1 - p^{\mathrm{exact}}(x_T, x_C)\right)\right] \\ &= 1 - \Phi\left(\Phi^{-1}\left(1 - p^{\mathrm{exact}}(x_T, x_C)\right)\right) \\ &= p^{\mathrm{exact}}(x_T, x_C). \qquad \text{(A.3)}\end{aligned}$$

Suppose $p^{\mathrm{exact}}(x_T, x_C) \leq \alpha/2$. From Equation (A.3), we have

$$\Pr\left[Z \geq Z_{\delta_0}^{\mathrm{EC}}(x_T, x_C)\right] = 1 - \Phi\left(Z_{\delta_0}^{\mathrm{EC}}(x_T, x_C)\right) \leq \alpha/2.$$

This implies

$$Z_{\delta_0}^{\mathrm{EC}}(x_T, x_C) \geq \Phi^{-1}(1 - \alpha/2) = z_{1-\alpha/2} = Z_{-\delta_{L,\alpha}^{\mathrm{EC}}}(x_T, x_C),$$

using Equation (22). From the assumption $Z_\delta^{\mathrm{EC}}(x_T, x_C)$ is monotonically increasing in δ, we have that $\delta_{L,\alpha}^{\mathrm{EC}} \geq -\delta_0$. The same steps can be used backward to show that $\delta_{L,\alpha}^{\mathrm{EC}}(x_T, x_C) > -\delta_0$ implies that $p^{\mathrm{exact}}(x_T, x_C) \leq \alpha/2$. □

APPENDIX B: MOTIVATION AND DERIVATION OF Z_δ^{EC}

Given x_T, x_C, and δ_0, we wish to derive a function of δ, written $f(\delta)$, that has properties like the Miettinen & Nurminen function Z_δ in (9), but whose associated confidence interval lower bound corresponds with Chan's exact unconditional p-value in (2) with the Miettinen & Nurminen statistic.

We first note that in order for $f(\delta)$ to invert to a confidence interval, we require $f(\delta)$ to be monotonic in δ and span the entire range $(-\infty, \infty)$. We associate a p-value to the function $f(\delta)$ in a manner similar to (17) by defining

$$p^f = 1 - \Phi(f(\delta_0)). \qquad \text{(B.1)}$$

And for any $\alpha \in (0,1)$, we define $\delta_{L,\alpha}^f$ to be the $(1-\alpha) \times 100\%$ confidence interval lower bound associated with $f(\delta)$ in a manner similar to (12) as follows:

$$f\left(-\delta_L^f\right) = \Phi^{-1}(1 - \alpha/2). \qquad \text{(B.2)}$$

Following a similar approach to the proof of Theorem A.2, we can establish the following relationship between p^f and $\delta_{L,\alpha}^f$:

$$p^f \leq \alpha/2 \quad \text{if and only if} \quad \delta_{L,\alpha}^f > -\delta_0.$$

Therefore, in order to make $f(\delta)$ consistent with Chan's exact p-value, p^{exact}, we need $p^f = p^{\mathrm{exact}}$, which, when combined with (B.1), yields the condition

$$1 - \Phi(f(\delta_0)) = p^{\mathrm{exact}}. \qquad \text{(B.3)}$$

For $f(\delta) = Z_\delta^{\mathrm{EC}}$, we show that the condition (B.3) is indeed established in (21). Although there are infinitely many monotonic functions $f(\delta)$ that span the entire range $(-\infty, \infty)$ whose value at δ_0 satisfies (B.3), we show how the function Z_δ^{EC} is a very natural choice with characteristics similar to the Miettinen & Nurminen function given by

$$Z_\delta = \frac{\widehat{P}_T - \widehat{P}_C + \delta}{\widehat{\sigma}_\delta}.$$

In particular, we can write Z_δ^{EC} as

$$\begin{aligned}Z_\delta^{\mathrm{EC}} &= \frac{\widehat{P}_T - \widehat{P}_C + \delta + \overbrace{\widehat{\sigma}_{\delta_0}\Phi^{-1}(1 - p^{\mathrm{exact}}) - \widehat{\sigma}_{\delta_0} Z_{\delta_0}}^{\text{constant}}}{\widehat{\sigma}_\delta} \\ &= \frac{\widehat{P}_T - \widehat{P}_C + \delta + \text{constant}}{\widehat{\sigma}_\delta}. \qquad \text{(B.4)}\end{aligned}$$

So, with the representation in (B.4), we see the resemblance of Z_δ^{EC} to Z_δ, and in Section 3.2, we numerically show similar asymptotic properties of these two functions as estimators.