Analysis of two binomial proportions in noninferiority confirmatory trials

Hassan Lakkis | Andrew Lakkis

Abstract

In this article, we propose considering an approximate exact score (AES) test for noninferiority comparisons and we derive its test-based confidence interval for the difference between two independent binomial proportions. This test was published in the literature, but not its associated confidence interval. The $p$-value for this test is obtained by using exact binomial probabilities with the nuisance parameter being replaced by its restricted maximum likelihood estimate. Calculated type I errors revealed that the AES method has important advantages for noninferiority comparisons over popular asymptotic methods for adequately powered confirmatory clinical trials, at 80% or 90% statistical power. For unbalanced sample sizes of the compared groups, type I errors for the asymptotic score method were shown to be higher than the nominal level in a systematic pattern over a range of true proportions, but the AES method did not suffer from such a problem. On average, the true type I error of the AES method was closer to the nominal level than all considered methods in the empirical comparisons. In rare cases, type I errors of the AES test exceeded the nominal level, but only by a small amount.Presented examples showed that the AES method can be more attractive in practice than practical exact methods. In addition, $p$-value and confidence interval of the AES method can be obtained in $<30$ s of computer time for most confirmatory trials. Theoretical arguments, combined with empirical evidence and fast computation time should make the AES method attractive in statistical practice.

Keywords

binomial distribution, confidence interval, confirmatory trials, difference in proportions, noninferiority

1 | INTRODUCTION

It is of interest in confirmatory clinical trials to show whether a new treatment is noninferior or equivalent to a standard existing treatment. This is very common in therapeutic areas with well-established efficacious treatments. Clinical trials in the anti-infective therapeutic area are frequently designed to show that a new possibly safer or more tolerable drug is noninferior in treating a specific class of bacterial infections to another active comparator. Common infections that are targeted in these noninferiority (NI) confirmatory clinical trials include, but are not limited to, community-acquired pneumonia, complicated skin and skin structure infections, complicated urinary tract infections, and hospital or ventilator-acquired pneumonia.
In each of these four diseases, it is not ethical to perform placebo-controlled trials due to the seriousness of the infections and the availability of effective treatments. NI designs are also common in vaccine clinical trials for which it is of interest to administer multiple vaccines during a single visit to the clinic. It is of interest to assess whether the conversion in the viral antibody level for each vaccine component in the concomitant administration as noninferior to that produced from the corresponding vaccine given alone. Travelers susceptible to viral infections may need to get their multiple vaccinations concomitantly just before making their trip. Also, pediatric vaccines may need to be administered concomitantly to bring the children up to date in their vaccination schedule.

When designing trials to show that the proportion of responders to the investigational drug is noninferior to that of the comparator, the clinical trialist must choose a test statistic or a confidence interval method for the difference of proportions that will be used for the NI test. Several popular methods are available based on asymptotic normality. Researchers must be careful in their consideration of the appropriate method to avoid selecting a method that is excessively conservative or anticonservative. An anticonservative test does not preserve the type I error at the nominal level and the validity of its statistical significance can be questioned.

Most asymptotic methods for confidence interval of difference in two binomial proportions are known to be anticonservative (liberal), especially when the proportions are higher than 80%. To avoid using anticonservative confidence intervals, methods based on continuity corrections were proposed by Hauck and Anderson\(^1\) and Newcombe.\(^2\) These methods preserve the type I error below the nominal error rate for most cases, but with frequently reduced power.

Various methods based on exact binomial probabilities were proposed to preserve type I error below the nominal rate. Santner and Snell\(^3\) proposed the unconditional exact method by inverting two 1-sided tests based on statistics for the difference between the observed proportions. Chan and Zhang (CZ)\(^4\) showed that the confidence interval method of Santner and Snell\(^3\) leads to a very conservative confidence interval due to severe discreteness of the distribution of the difference in observed proportions being used, especially for small samples. An unconditional exact NI test based on the score statistics was proposed by Chan\(^5\) for the \(p\)-value, this is known as the exact score (ES) method. Further, CZ\(^6\) proposed inverting the two 1-sided unconditional ES tests based on the score statistics to obtain the corresponding confidence interval for NI. The purpose of this exact confidence interval method is that it retains the size of the type I error on each side of the test below the nominal value regardless of the true value of the nuisance parameter, it is much more efficient than that of Santner and Snell\(^3\) as it is not as conservative, and it is available in SAS 9.3 release.\(^6\) However, in the calculation of the \(p\)-value of this ES method, the worst-case scenario for the nuisance parameter is assumed, which could be far from the nuisance parameter's likely true value for the study being considered. This could lead to a more conservative test than what would be necessary, and a method that utilizes the maximum likelihood estimate (MLE) of the true nuisance parameter in calculating the binomial probabilities can lead to a more appropriate test for confirmatory trials. In addition, the ES method requires excess computation time for studies with sample sizes such as those used for Phase 3 confirmatory clinical trials.

Another exact test for NI comparison was proposed by Wang.\(^7,8\) It was reported by Shan and Wang\(^5\) that a practical issue with this method is when the total sample size for a binomial study includes more than 100 subjects. This could cause an excessively long computation time for obtaining its two-sided confidence interval especially when compared with the ES method. For this reason, along with that it does not necessarily have higher statistical power than the ES method (based on examples presented by Garner\(^9\)), this exact method will not be included in the remainder of this article.

Asymptotic methods used for NI hypothesis tests are usually based on the confidence interval for the difference of binomial response rates which include the standard asymptotic normal method known as Wald's method, the method proposed by Miettinen and Nurminen,\(^11\) the uncorrected for continuity method proposed by Newcombe,\(^2\) and the method proposed by Agresti and Caffo (AC).\(^12\) The method of Miettinen and Nurminen\(^11\) is in fact the same as the asymptotic score approach as described by Farrington and Manning\(^13\) for NI hypothesis testing except for bias correction in the variance estimate which has almost no impact on the results when the sample size is large, such as >100 subjects per treatment group which is of interest in this article.

Ripamonti et al.\(^14\) provided an interesting review of tests that involve nuisance parameters for assessing treatment effects based on comparison of two binomial proportions. They covered likelihood and testing theory as well as computational advances for both conditional and unconditional tests. We cover their review of the unconditional tests as these are more relevant in this article. Within an unconditional framework and using binomial probabilities, they presented four methods of correcting approximate tests to properly control size by accounting for the unknown value of the nuisance parameter: maximization (M) by calculating the \(p\)-values that account for the worst possible scenario of the nuisance parameter (such as the method by Chan\(^5\)), partial maximization (B) by using the method of Berger and Boos\(^15\) to maximize over an exact confidence interval of the nuisance parameter under the null, estimation (E) by replacing the nuisance parameter by an estimate (such as the restricted MLE in NI test which is the method of interest in this article),
and estimation followed by maximization (E + M). The E + M approach was initially proposed by Lloyd. From the
least to the most computationally intensive these methods are E, B, M, and E + M. The last three control the type I
error at the nominal significance level, while the first is very close to the nominal level. They noted that the E p-value
is easiest to compute. It is very close to the exact and it is recommended in practice. They also explained how the E
test can be extended to the E + M test which combines the use of the p-value from the E step with the M step. The authors
noted that the E + M p-values are typically slightly more powerful than B p-values. Also, the E + M method is more
powerful than the M method. The only weakness of the E + M method is its computational burden.

Fay and Hansberger evaluated the analysis of comparing two binomial proportions mostly for exact methods.
They provided an extensive review of hypothesis testing with a focus on validity and power. They also discussed confi-
dence interval estimation with emphasis on their coverage probability and expected length. They considered whether a
p-value and its matching confidence interval are compatible, meaning that the p-value rejects at level α if and only if
the (1 − α)100% confidence interval excludes all null parameter values. In Section 7 of their paper, the melding method
for confidence interval by Fay et al was discussed. This confidence interval is compatible with the central Fisher’s
exact test, and its mid-p version. The melding method is fast to calculate, but it is expected to be conservative as is the
case with Fisher’s exact test in comparison to more efficient unconditional exact tests. The authors suggested methods
that allow for one-sided inferences and provided approaches that can reduce computational burden.

Fagerland et al. described and compared asymptotic and exact methods for confidence interval estimation between
two independent proportions based on difference, ratio, and odds ratio. They examined the performances of the two-
sided confidence intervals using coverage probabilities and expected length for each method and made recommenda-
tions for studies with sample sizes up to 40 subjects per treatment group. Their comparisons focused on the accuracy of
the estimation methods for the two-sided confidence intervals and did not address their properties for NI comparison.

Dann and Koch compared the statistical properties of the four asymptotic methods presented in the 7th paragraph in
this section along with the Newcombe with continuity correction method for studies powered at 85%. Based on their results,
they recommended the use of each method for various scenarios. Rothmann et al. provided an extensive summary of the
literature for various methods with applications for NI comparisons. Kang and Chen proposed an approximate exact score
(AES) test statistic using the restricted MLE of the nuisance parameter in the exact binomial probabilities and assessed its
properties only up to a total of 100 subjects per treatment group. However, the properties of this test statistic were not
assessed for common sample sizes considered in adequately powered clinical trials when the NI margin is 10% and 15%. In
this article, we consider the same AES test statistic of Kang and Chen for NI comparisons in difference between investiga-
tional and comparator proportions. We also study the properties of this NI test versus various popular methods for studies
with sample sizes selected at statistical powers of 80% or 90% and with NI margins of 10% or 15% and to a lesser extent with
an NI margin of 5%. In addition, for the AES test by Kang and Chen, we developed a two-sided test-based confidence inter-
val, based on modification of CZ approach, that is consistent with the p-value of the test. The term consistent is utilized to
suggest that the one-sided p-value < α/2 if and only if the lower side of the two-sided 100(1 − α)% confidence interval is
higher than the NI margin, −δ0.

The NI margins considered for assessment of statistical properties in this article are 10% and 15%, such margins are
of interest because NI clinical trials are frequently conducted based on NI margins ranging from 10% to 15%. NI mar-
gins are justified by various historical data for effective comparators or agreed on with regulatory agencies prior to study
initiation. For instance, each of the US Food and Drug Administration (FDA) guidance documents for the indications
of acute bacterial skin and skin structure infections, complicated urinary tract infections, and complicated intra-
abdominal infections reported that a 10% NI margin is justified by historical evidence for clinical response endpoints.
Also, the FDA guidance document for community-acquired bacterial pneumonia reported that a 12.5% NI margin is
justified by historical evidence for a clinical recovery response endpoint. Each of the FDA guidance documents
encourages sponsors to discuss the selection of a NI margin with the FDA in advance of trial initiation, particularly for
a proposed margin greater than the ones supported by historical data in the respective guidance documents. Also,
assessment of the methods is performed with a NI margin of 5% when the proportion of response in the comparator is
95% because it is logical to select a smaller margin as the response in the comparator is expected to be closer to 100%.

2 | BACKGROUND

Assume a binomial trial that compares two groups of sizes N_T and N_C. Let, X and Y be the independent responses dis-
tributed as binomial random variables with parameters (N_T, P_T) and (N_C, P_C), respectively. The test of interest is
regarding the difference (P_T − P_C), where this difference is in the range of [−1, 1].
Throughout this article, for simplicity in the hypothesis formulation, we assume that a higher response indicates a better efficacy outcome such as a clinical cure rate in an anti-infective trial.

The primary interest is to demonstrate that the new treatment is noninferior to the standard treatment by a prespecified margin, such as 10% and 15%. The null hypothesis can be specified as being that the proportion of responders to the new treatment is lower than that of the standard treatment by at least a prespecified small margin versus the alternative that the proportion of responders to the new treatment is not lower than that of the standard treatment by that margin. Let, \( P_T \) and \( P_C \) denote the true proportions of responders to the test (investigational) and to the control treatments, respectively. Then the test of NI (one-sided) focuses on the hypothesis:

\[
H_0 : P_T - P_C \leq -\delta_0 \quad \text{versus} \quad H_1 : P_T - P_C > -\delta_0, \tag{1}
\]

where \( \delta_0 \) is the prespecified NI margin, a positive quantity. Other articles that discussed the selection of the NI margin include Temple\textsuperscript{27} and the ICH E10 Guidelines.\textsuperscript{28}

### 2.1 Asymptotic score test

To test the NI hypothesis in (1), the following test statistic can be used:

\[
Z(X_T, X_C; \delta_0) = \left( \hat{P}_T - \hat{P}_C + \delta_0 \right) / \sqrt{\hat{V}(\delta_0)}, \tag{2}
\]

where \( \hat{P}_T \) and \( \hat{P}_C \) are the random variables of the observed proportions of responders in the two groups, respectively, and \( \hat{V}(\delta_0) = \hat{p}_T(1 - \hat{p}_T)/N_T + \hat{p}_C(1 - \hat{p}_C)/N_C \) is the variance of the numerator and \( \hat{V}(\delta_0) \) is its estimator with \( \hat{V}(\delta_0) \) is its realized value obtained by replacing \( P_T \) and \( P_C \) by their MLEs under the constrained null hypothesis of \( P_T - P_C = -\delta_0 \). Large values of the test statistic favor the alternative hypothesis. The test statistic provided in (2) will be referred to as the asymptotic score test statistic when \( Z(X_T, X_C; \delta_0) \) is considered to follow the asymptotic standard normal distribution. Its \( p \)-value can be obtained by using,

\[
p_{\alpha/2}(x_T, x_C) = \Pr(Z(X_T, X_C; \delta_0) \geq Z(x_T, x_C; \delta_0)) . \tag{3}
\]

The test-based two-sided confidence interval that corresponds to (2) can be obtained by inverting the following equation for \( \delta \)

\[
Z_{\alpha/2}^2 = \left( \hat{p}_T - \hat{p}_C + \delta \right)^2 / (\hat{p}_T(1 - \hat{p}_T)/N_T + \hat{p}_C(1 - \hat{p}_C)/N_C), \tag{4}
\]

where \( \hat{p}_T \) and \( \hat{p}_C \) are the MLEs under the constraint \( P_T - P_C = -\delta \) and \( Z_{\alpha/2} \) is the upper \( \alpha/2 \) percentile of the standard normal distribution. This is in fact an asymptotic likelihood score method and will be referred as the AS method.

Miettinen and Nurminen (MN)\textsuperscript{11} proposed this approach to produce the confidence interval given in (4) and provided expressions for the MLEs \( \hat{p}_T \) and \( \hat{p}_C \), but multiplied the denominator by \( N/(N - 1) \) with \( N = N_T + N_C \) to correct for bias in the variance estimate. Since the focus of this article is on large sample sizes, correcting the denominator of (4) by \( N/(N - 1) \) would not have meaningful impact and therefore we drop this factor for the remainder of this article.

Farrington and Manning\textsuperscript{13} proposed the score test statistic given in (2) and provided an estimate of the sample size for a given power. This is the associated test statistic with the Miettinen and Nurminen\textsuperscript{11} confidence interval method without the bias correction in the denominator. Also, a Farrington and Manning\textsuperscript{13} interval which can lead to the same statistical significance as it is derived from the same test statistic can be obtained by:

\[
(\hat{p}_T - \hat{p}_C) \pm Z_{\alpha/2} \sqrt{\hat{p}_T(1 - \hat{p}_T)/N_T + \hat{p}_C(1 - \hat{p}_C)/N_C}, \tag{5}
\]

where \( \hat{p}_T \) and \( \hat{p}_C \) are the MLEs of \( P_T \) and \( P_C \), respectively, calculated under the fixed NI null hypothesis constraint of \( P_T - P_C = -\delta_0 \). It should be noted that the MLEs \( \hat{p}_T \) and \( \hat{p}_C \) in (5) are obtained under \( P_T - P_C = -\delta_0 \) and the MLEs \( \hat{p}_T \) and \( \hat{p}_C \) in (4) are obtained under \( P_T - P_C = -\delta \).
Hawila and Berg,\textsuperscript{29} in their Theorem A.2, demonstrated that NI conclusion for the fixed null given in (1) would be the same whether the test statistic is based on the \( p \)-value as given in (3) or based on the lower limit, denoted by \( \delta_{\text{LAS}} \), for the confidence interval of the AS method presented in (4). Their proof showed that for any given fixed null \( \delta_0 \):

\[
p_{\delta_0}(x_T,x_C) < \frac{\alpha}{2} \text{ if and only if } \delta_{\text{LAS}} > -\delta_0.
\]  

(6)

Following the same steps presented in Theorem A.2 by Hawila and Berg,\textsuperscript{29} it can be shown that for the given fixed null \( \delta_0 \):

\[
p_{\delta_0}(x_T,x_C) < \frac{\alpha}{2} \text{ if and only if } \delta_{\text{LFM}} > -\delta_0,
\]  

(7)

where \( \delta_{\text{LFM}} \) is the lower limit of the Farrington and Manning\textsuperscript{13} confidence interval as presented in (5). From (6) to (7), we can conclude that

\[
\delta_{\text{LAS}} > -\delta_0 \text{ if and only if } \delta_{\text{LFM}} > -\delta_0.
\]  

(8)

Expression (8) implies that for a given null \( \delta_0 \), both \( \delta_{\text{LAS}} \) and \( \delta_{\text{LFM}} \) are on the same side of \( -\delta_0 \) and they are expected to have different values with some exceptions. For the case where \( \delta_0 \) is selected such that the \( p \)-value given in (3) is equal to \( \alpha/2 \) which is equivalent to \( (\hat{p}_T - \hat{p}_C + \delta_0)/\sqrt{\hat{v}(\delta_0)} = Z_{\alpha/2} \), the lower limits based on the two confidence interval methods become equal and match the negative value of the NI margin, that is \( \delta_{\text{LAS}} = \delta_{\text{LFM}} = -\delta_0 \). That is because \( \delta_{\text{LAS}} \) is a unique value that satisfies \( (\hat{p}_T - \hat{p}_C - \delta_{\text{LAS}})/\sqrt{\hat{v}(\delta_{\text{LAS}})} = Z_{\alpha/2} \) and \( \delta_{\text{LFM}} \) is a unique value that satisfies \( (\hat{p}_T - \hat{p}_C - \delta_{\text{LFM}})/\sqrt{\hat{v}(\delta_{\text{LFM}})} = Z_{\alpha/2} \), where \( Z_{\alpha/2} \) is the upper \( \alpha/2 \) percentile of the standard normal distribution.

Therefore, statistical inference for the hypothesis given in (1) whether the analysis test is based on the \( p \)-value given in (3) or based on any of the lower bounds of the two confidence interval methods given in (4) and (5) would lead to the same conclusion in terms of statistical significance. As a result, we will focus on the \( p \)-value in (3) and its associated confidence interval given in (4) for the AS method in the remainder of the article.

The confidence interval obtained from Expression (4) for the AS method tends to perform well as presented in Miettinen and Nurminen\textsuperscript{11} and it has better properties than that based on Wald’s method where the variance of the numerator is estimated by the sample variance. Consider the following extreme example where \( x_T = 100, N_T = 100, x_C = 0, N_C = 90; \hat{p}_T = 100\% \) and \( \hat{p}_C = 0\% \). In this example, we have the maximum possible difference between the proportions. The two-sided 95\% confidence interval is given by (95.6\%, 100\%) based on the AS method versus (100\%, 100\%) based on Wald’s normal method. This example revealed the aberration of the width of 0 of Wald’s confidence interval versus a realistic nonzero width of the confidence interval based on the AS method.

### 2.2 ES test

Chan\textsuperscript{5} proposed using the exact unconditional binomial distribution of the test statistics \( Z(X_T,X_C;\delta_0) \) for all possible outcomes of the two binomial responses, where each outcome \( (X_T,X_C) \) corresponds to a \( 2 \times 2 \) table. Following his approach, the probability of observing a particular outcome \( (x_T,x_C) \) is a product of two binomial probabilities. Let \( P = P_T \), the likelihood for the test of NI given in (1) can be written as

\[
\Pr(X_T = x_T,X_C = x_C|H_0) = \binom{N_T}{x_T} \binom{N_C}{x_C} P^{x_T} (1-P)^{N_T-x_T} (P+\delta_0)^x_C (1-P-\delta_0)^{N_C-x_C}.
\]  

(9)

Let \( Z(x_T,x_C;\delta_0) \) be the value of \( Z(X_T,X_C;\delta_0) \) at the observed \( (x_T,x_C) \) table using Equation (2). As described by Chan,\textsuperscript{5} the exact significance level is the sum of the probabilities of tables that are at least as extreme as the observed table. As can be seen, the likelihood in (9) depends on an unknown nuisance parameter \( P \). The exact significance level for the test of the hypothesis in (1) is calculated by maximizing the null likelihood over the domain of the nuisance parameter \( P \). The domain of \( P \) is \( D = [0,1-\delta_0] \). The exact significance level or \( p \)-value for the score test is given by,
\[
p_{\delta_0}(x_T,x_C) = \max_{P \in D} \Pr(Z(X_T,X_C;\delta_0) \geq Z(x_T,x_C;\delta_0)|\delta_0,P).
\] (10)

This is the ES test, which was proposed by Chan\(^5\) with \(p\)-value as given in (10). It is an extension of the work of Suissa and Shuster\(^30\) to include null hypotheses of a prespecified difference that is different from 0.

Chan\(^5\) provided detailed steps describing how to calculate the \(p\)-value given in (10). To obtain adequate accuracy of the \(p\)-value, Chan\(^5\) suggested dividing the domain of \(P\) into 1000 equally spaced subintervals and at each value of \(P\) calculate the probability on the right side of Equation (10). The \(p\)-value is the maximum of these 1000 probabilities. CZ\(^4\) derived the exact test-based confidence interval by inverting the ES test statistic.

Due to the numerical search over the domain of the nuisance parameter to obtain the \(p\)-value given in (10), this computation process requires intensive computer work as the sample size increases. Chan\(^5\) indicated that calculating the corresponding test-based confidence interval is much more complicated than performing the hypothesis test because it requires numerical search in the parameter space for the nuisance parameter as well as for the parameter space for each bound of the confidence interval.

It should be noted that the \(p\)-value of the ES method by Chan\(^5\) as presented in (10) is based on the maximization of the tail probability function in (10) over the range of values of \(P\) in \(D = [0,1-\delta_0]\). The domain \(D = [0,1-\delta_0]\) for the nuisance parameter \(P\) is the boundary of the null space \(H_0: P_T - P_C \leq -\delta_0\).

In deriving the unconditional test for the null hypothesis of no difference between two binomial proportions, Barnard\(^31\) discussed the logical requirements that a critical region \((R)\) should be convex. He defined the convex critical region in the sense that if an observed data point from two binomial distributions \((i,j) \in R\), then both \((i-1,j) \in R\) and \((i,j+1) \in R\). This is referred to as the convexity condition \((C)\) and it means that if the outcome \((i,j)\) is in the rejection region of the NI hypothesis defined in (1), then one less success in the control group or one more success in the test group should also lead to the rejection of the null hypothesis given by \(H_0: P_T - P_C \leq -\delta_0\).

Röhmel and Mansmann\(^32\) have pointed out the theoretical desirability of the convexity condition for one-sided unconditional hypothesis tests, including the NI hypothesis as presented in (1), because it allows for the calculation of the exact \(p\)-value by restricting the search for the maximum to the boundary of the null space as given by \(P_T - P_C = -\delta_0\) (or equivalently over the range of values of \(P\) in \(D = [0,1-\delta_0]\)). Röhmel and Mansmann\(^32\) also proved that if an ordering criterion satisfies the convexity condition, then the maximum tail probability under the null hypothesis is attained at a boundary point of the null space.

Röhmel and Mansmann\(^33\) commented that it is unknown whether the test statistic \(Z(X_T,X_C;\delta_0)\) indeed satisfies \((C)\) when used for ordering the sample space for the ES method as presented by Chan.\(^5\) To address their comment, Chan\(^34\) and Chan\(^35\) demonstrated that the \(Z(X_T,X_C;\delta_0)\) as used for ordering the sample space in the ES test statistic satisfies the convexity condition \((C)\). Therefore, the maximization of the tail probability for computing the ES \(p\)-value as given in (10) occurs on the boundary \(P_T - P_C = -\delta_0\) of the null space defined by \(H_0: P_T - P_C \leq -\delta_0\). This proves that the \(p\)-value of the ES method as defined in (10) is the maximum of the tail probability over the full null space defined by \(P_T - P_C \leq -\delta_0\).

### 2.3 Approximate ES test

The AES test described in this paragraph is the one of primary interest in this article. Based on the same approach of the ES test, we will replace the nuisance parameter \(P\) in (10) by its MLE, \(\tilde{P}\), under the constraint \(P_T - P_C = -\delta_0\) (Farrington and Manning\(^11\) provided an expression for this MLE of \(P\)). Then, the \(p\)-value for this AES test becomes much simpler to obtain and can be written as

\[
p_{\delta_0}(x_T,x_C) = \Pr(Z(X_T,X_C;\delta_0) \geq Z(x_T,x_C;\delta_0)|\delta_0,\tilde{P})
\] (11)

This \(p\)-value given in (11) can be obtained by following the two steps:

1. Calculate the \(Z(X_T,X_C;\delta_0)\) statistic for all possible binomial outcomes and order them.
2. The \(p\)-value in (11) can be obtained by summing up all of the probabilities using the likelihood function in (9), with \(P\) replaced by \(\tilde{P}\), for the outcomes whose \(Z(X_T,X_C;\delta_0)\) statistic is greater than or equal to the observed \(Z(x_T,x_C;\delta_0)\) statistic at the \((x_T,x_C)\) table.
Storer and Kim \(^{36}\) proposed the AES test in (11) by obtaining the \(p\)-value using the MLE of the nuisance parameter for testing the case of zero difference under the null hypothesis between the binomial proportions. Also, they discussed the attractive properties of the AES test relative to the exact method of Suissa and Shuster \(^{30}\) by showing that the type I error of their method is closer to the nominal rate and required much less computation time. Kang and Chen \(^{22}\) extended the AES test to the null hypothesis of nonzero difference, based on the same \(p\)-value given in (11), for NI testing. However, as stated earlier, they limited their assessment of the statistical properties to a maximum of 100 subjects per treatment group. In this article, we assess the statistical properties of this method versus other commonly used methods for sample sizes that are required for adequately powered studies for NI confirmatory trials with NI margins 15\%, 10\%, and to a less extent 5\%. Most confirmatory trials have treatment allocations ratio of \(N_T: N_C = 1:1, 2:1, 1:2\) leading to sample sizes frequently being in the range from 100 to over 600 subjects per treatment group.

In addition to the \(p\)-value given in (11), we discuss obtaining a consistent confidence interval which is necessary for reporting NI hypothesis test results for confirmatory trials.

### 3 PROPOSED CONFIDENCE INTERVAL FOR THE NI TEST

For presentation of results for NI clinical trials, clinical reports customarily include confidence intervals along with the \(p\)-values as the basis for clinical interpretation and for NI statistical inference. However, the FDA has been reporting the confidence intervals as the basis for NI confirmatory results in Package Inserts for approved products without inclusion of the \(p\)-values. Accordingly, to make the AES method more appealing for consideration in clinical trial practice, we propose deriving a test-based confidence interval that is consistent with the \(p\)-value given in (11), meaning that the \(p\)-value rejects the fixed null defined in (1) at level \(a/2\) if and only if the lower bound of the two-sided \(1 - \alpha\) confidence interval is \(> -\delta_0\).

When selecting the bounds of the confidence interval, we would like to preserve the decision rule that is consistent with the \(p\)-value for the hypothesis test. Let, \(\delta_L\) and \(\delta_U\) represent the lower and upper bounds of the confidence interval for \(\delta\). The lower bound is obtained by inverting the one-sided, \(a/2\), test of hypothesis for \(H_0: P_T - P_C = \delta_0\) versus \(H_1: P_T - P_C > \delta_0\) and the upper bound is obtained by inverting the one-sided, \(a/2\), test of hypothesis for \(H_0: P_T - P_C = \delta_0\) versus \(H_1: P_T - P_C < \delta_0\). Conventional expressions of the inversions for obtaining the lower and upper bounds, \(\delta_L\) and \(\delta_U\), for the \(100(1-\alpha)\%\) confidence interval for \(\delta\) are provided in Supplement I of this article. When such expressions are used, Supplement I summarized significant computation challenges, to estimate \((\delta_L, \delta_U)\) correctly leading to extensive delay in computer time. Similar challenges were reported by Chen. \(^{37}\)

Instead of using the expressions for the inversions provided in Supplement I, we consider the approach by CZ \(^4\) used for obtaining the ES confidence interval by building the confidence interval based on inverting the following two probabilities denoted by:

\[
g_L(\delta) = \Pr[Z(x_T, x_C; \delta_{LAS}) \geq Z(x_T, x_C; \delta_{LAS}) | \delta, \bar{p}] \tag{12}
\]

\[
g_U(\delta) = \Pr[Z(x_T, x_C; \delta_{UAS}) \leq Z(x_T, x_C; \delta_{UAS}) | \delta, \bar{p}] \tag{13}
\]

where \((\delta_{LAS}, \delta_{UAS})\) are the lower and upper bounds of the confidence interval based on the asymptotic score method given in (4). Note that defining the tail regions by \([Z(x_T, x_C; \delta_{LAS}) \geq Z(x_T, x_C; \delta_{LAS}) | \bar{p}]\) and \([Z(x_T, x_C; \delta_{UAS}) \leq Z(x_T, x_C; \delta_{UAS}) | \bar{p}]\) for (12) and (13), respectively, to be independent of \(\delta\) causes the probabilities \(g_L(\delta)\) and \(g_U(\delta)\) to be monotone and continuous functions in \(\delta\). The lower and upper bounds of the confidence interval \((\delta_L, \delta_U)\) for the AES method can be found by solving \(g_L(\delta) = \alpha/2\) and \(g_U(\delta) = \alpha/2\) for \(\delta\), respectively, where the range of \(\delta\) is \((-1,1)\). The lower bound \(\delta_L\) is a unique solution for \(\delta\) (since \(g_L(\delta)\) is monotone) and can be obtained by solving \(g_L(\delta) = \alpha/2\) using the bisection method with the lower bound of the asymptotic score method as one of the initial values. Similarly, the upper bound \(\delta_U\) is a unique solution for \(\delta\) (since \(g_U(\delta)\) is monotone) and can be obtained by solving \(g_U(\delta) = \alpha/2\) using the bisection method with the upper bound of the asymptotic score method as one of the initial values. The resulting \((\delta_L, \delta_U)\) confidence interval is consistent with the \(p\)-value for most trial outcomes, but it is not guaranteed to be consistent for all possible outcomes. This nonconsistency can occur in some trial outcomes when the \(p\)-value calculated at the \(Z(x_T, x_C; \delta_0)\) for the fixed NI margin \(\delta_0\) is very close to \(\alpha/2\). The reason for the nonconsistency is that the lower bound of the two-sided \(100(1-\alpha)\%\) confidence interval is obtained by inverting the fixed test \(Z(x_T, x_C; \delta_{LAS})\) which depends on different number of points in tail regions \((\{Z(x_T, x_C; \delta_{LAS}) \geq Z(x_T, x_C; \delta_{LAS}) | \bar{p}\})\) than that for calculating the \(p\)-value at \(Z(x_T, x_C; \delta_0)\), for the NI margin \(\delta_0\).
To find a consistent lower bound of the confidence interval when for a trial outcome the lower bound of the confidence interval is not consistent with the p-value based on the solution for $g_L(\delta) = \alpha/2$ with $g_L(\delta)$ defined as in Equation (12), we update the lower bound of the confidence interval for the AES method to be obtained by solving $g_L(\delta) = \alpha/2$ for $\delta$ as follows:

$$g_L(\delta) = \Pr[Z(X_T, X_C; \delta_0) \geq Z(x_T, x_C; \delta_0) | \delta, \bar{p}].$$ (14)

Note that (14) is a monotone function in $\delta$. Also, it has the same form as (12), but instead of using $\delta_{LAS}$ as in (12) to define the fixed tail region we used the tail region that is defined by the fixed NI margin $\delta_0$ as shown in (14). The confidence interval generated using this approach (replacing $\delta_{LAS}$ by the NI margin $\delta_0$) guarantees its consistency with the p-value.

Specifically, if the p-value at the NI margin $\delta_0$ is $> \alpha/2$ and the points in the tail region being the same for calculations of both the p-value and the $g_L(\delta)$ as in (14), then there will be a unique $\delta_L$ such that it is a solution for $g_L(\delta) = \alpha/2$ where $\delta_L < -\delta_0$. On the other hand, if the p-value at the NI margin $\delta_0$ is $\leq \alpha/2$ and the points in the tail region being the same for both the p-value and the $g_L(\delta)$ as in (18), then there will be a unique solution $\delta_L$ such that it is $g_L(\delta) = \alpha/2$ where $-\delta_0 < \delta_L$. These arguments proved that using (14) can lead to a $\delta_L$ that is consistent with the p-value for the AES method.

To illustrate the calculations of the p-value and the confidence interval using the approach presented above we considered the following example, if $\hat{p}_T = 287/328$ and $\hat{p}_C = 98/108$ with NI margin in percent $\delta_0 = 10\%$, then based on the AES method, the NI p-value is 0.025029 and the 95% two-sided confidence interval for the proportion difference is ($-9.9987, 3.1137$). This confidence interval was obtained based on the fixed tail regions of the p-value functions (12) and (13) with $\delta_{LAS} = -9.8328$ and $\delta_{UAS} = 3.36095$ and by inverting $g_L(\delta) = 0.025$ and $g_U(\delta) = 0.025$. Since the lower bound of the 95% two-sided confidence interval, $-9.9997$, is $> -10$ (the NI margin), the confidence interval approach demonstrated NI of the test treatment to the control treatment and that is not consistent with the conclusion of failing the NI test based on the p-value test ($p$-value $>0.025$).

Given that using $\delta_{LAS}$ in defining the fixed tail region for the probability as in (12) led to nonconsistency in this example, we recalculate the lower bound of the confidence interval using $\delta_0 = -10\%$ in defining the fixed tail region as in (14). However, the upper bound is still being estimated using (13). As a result, the two-sided 95% resulting confidence interval of the AES method becomes ($-10.0014, 3.1137$). Since the lower bound of the confidence interval is $-10.0014$, which is lower than the NI margin of $-10$, the statistical test based on the lower bound of the confidence interval leads to a conclusion (failing to conclude NI) that is consistent with the test based on the p-value ($p$-value $>0.025$).

In summary, the proposed AES method involves generating the lower bound and upper bound of the two-sided confidence interval by solving $g_L(\delta) = \alpha/2$ and $g_U(\delta) = \alpha/2$ for $\delta$, respectively, based on using the fixed tail regions as defined in (12) and (13). Only if in the very rare situation where the lower bound of the confidence interval leads to a nonconsistent conclusion with the p-value, then the lower bound should be recalculated based on the tail region that uses the fixed NI margin $\delta_0$ as shown in (14). Also, if it is of interest to use this AES method to obtain the confidence interval without a NI hypothesis test being considered, then it would be sufficient to invert the fixed test statistic $Z(X_T, X_C; \delta_{LAS})$.

With the improvement of the speed of PC computers up to present time, the suggested method for obtaining the confidence interval and the p-value as described in the previous paragraph can be calculated in $<30$ s of real computer time even for studies with total sample size in the two compared groups of 1000 subjects. As examples, we provide the real computer time for obtaining the proposed confidence interval and the p-value in Section 7.

To assess usefulness of the AES method for confirmatory trials, we consider comparing its statistical properties for various reported methods in the literature based on studies with adequate sample sizes for hypothesis tests with statistical powers at 80% and 90%. We also present analyses of four examples based on the AES method and other methods. The next two sections describe the methods that will be included in the assessment along with the statistical properties being used.

### 4 TESTS COMPARED

Based on statistical properties, we compare the AES method to six asymptotic methods for testing NI hypotheses based on two independent proportions.
The methods considered in this comparison of the statistical properties include the following:

1. The AES method with its p-value as presented earlier in (11) and its confidence interval as described in Section 3 in this article.
2. The commonly known Wald method whose confidence interval is given by:

\[
(\hat{p}_T - \hat{p}_C) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_T(1 - \hat{p}_T)}{N_T} + \hat{p}_C(1 - \hat{p}_C)/N_C},
\]

where \(Z_{\alpha/2}\) is the upper \(\alpha/2\) percentile from the standard normal distribution.

3. The AC\(^2\) method whose confidence interval is given by

\[
(\hat{p}_T - \hat{p}_C) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_T(1 - \hat{p}_T)}{N_T + 2} + \hat{p}_C(1 - \hat{p}_C)/(N_C + 2)},
\]

where \(\hat{p}_T = (x_T + 1)/(N_T + 2)\) and \(\hat{p}_C = (x_C + 1)/(N_C + 2)\). In essence, the AC method simply adds one success and one failure to each treatment group and then uses the same Wald's formula for the confidence interval. This simple adjustment gives this method significant advantages over Wald's method, as was shown by AC.\(^{12}\)

4. The Newcombe Hybrid Score interval (NC)\(^2\) whose bounds are given by

\[
\text{Lower Bound : } (\hat{p}_T - \hat{p}_C) - \sqrt{(\hat{p}_T - l_T)^2 + (u_C - \hat{p}_C)^2}
\]

\[
\text{Upper Bound : } (\hat{p}_T - \hat{p}_C) + \sqrt{(u_T - \hat{p}_T)^2 + (\hat{p}_C - l_C)^2},
\]

where \(l_T\) and \(u_T\) are the lower and upper limits of the Wilson Score interval for \(P_T\) and are obtained by solving the following equation for \(P_T\):

\[
|P_T - \hat{p}_T| = Z_{\alpha/2} \sqrt{P_T(1 - P_T)/N_T}.
\]

Similarly, \(l_C\) and \(u_C\) are the lower and upper limits of the Wilson Score interval for \(P_C\) and can be obtained by solving a similar score equation to (19) for \(P_C\).

5. The asymptotic score confidence interval method (AS) as given in (4) and its test statistic as given in (2).
6. The HA continuity\(^1\) corrected interval. Its bounds are given by:

\[
(\hat{p}_T - \hat{p}_C) \pm \left\{ Z_{\alpha/2} \sqrt{\hat{p}_T(1 - \hat{p}_T)/N_T + \hat{p}_C(1 - \hat{p}_C)/N_C + 1/[2 \min(N_T, N_C)]} \right\}
\]

7. The Newcombe Hybrid Score continuity corrected interval (NCC). Newcombe\(^2\) provided this continuity corrected method by solving \(|P_T - \hat{p}_T| - 1/2N_T = Z_{\alpha/2} \sqrt{P_T(1 - P_T)/N_T}\) and \(|P_C - \hat{p}_C| - 1/2N_C = Z_{\alpha/2} \sqrt{P_C(1 - P_C)/N_C}\) for the confidence limits of \(P_T\) and \(P_C\). Newcombe then substituted the appropriate confidence limits of \(P_T\) and \(P_C\) in the interval given in (17) and (18).
5 | EXACT TYPE I ERRORS AND STATISTICAL POWERS

The statistical methods presented in Section 4 for testing NI hypotheses will be compared based on the statistical criteria of the true type I error and the true power.

For the test-based confidence interval methods (AES and AS methods), we define the rejection region of an arbitrary level $\alpha/2$ one-sided test with NI margin $\delta_0$ as follows:

$$R_0 = \{(i,j) : (i,j) \in \Omega \text{ and } p_{\delta_0}(i,j) \leq \alpha/2\}, \quad (21)$$

where $\Omega = \{0,1,\ldots,N_T\} \times \{0,1,\ldots,N_C\}$ represents the entire sample space for the multivariate random variable $(X_T,X_C)$, and $p_{\delta_0}(i,j)$ is the one-sided $p$-value for the NI test, calculated as described in (3) for the AS method and in (11) for the AES method. This $p$-value is calculated for each possible observation $(i,j)$ in the entire sample space $\Omega$ to form the critical region.

The other tests in the comparisons included the Wald, AC, NC, NCC, and HA. These were based on comparing the lower bounds of their confidence intervals versus $-\delta_0$ and these bounds were not inverted from test statistics. For each of these methods, a lower bound of the confidence interval $> -\delta_0$ implies rejection of the null hypothesis and NI is concluded. For these methods, the rejection region of a level $\alpha/2$ one-sided test can be defined as follows:

$$R_0 = \{(i,j) : (i,j) \in \Omega \text{ and } \delta_1(i,j) > -\delta_0\}, \quad (22)$$

where $\delta_1(i,j)$ is the lower bound of the two-sided $100(1-\alpha)\%$ confidence interval for a given method at the $(i,j)$ realization of $(X_T,X_C)$ in the overall domain $\Omega$. Each side outside the two-sided confidence interval has $\alpha/2$ probability.

For given sample sizes $(N_T,N_C)$ and $P_C$, the exact type I error for each of the methods for testing $H_0 : P_T - P_C \leq -\delta_0$ versus $H_1 : P_T - P_C > -\delta_0$ is given by $\alpha_{\delta_0}(P) = P_{H_0}\{(X_T,X_C) \in R_0|P\}$ and can be obtained using the following equation:

$$\alpha_{\delta_0}(P) = \sum_{(i,j) \in R_0} \binom{N_T}{i} \binom{N_C}{j} P^i(1-P)^{N_T-i}(P+\delta_0)^j(1-P-\delta_0)^{N_C-j}, \quad (23)$$

where $P = P_T = P_C - \delta_0$.

The main reason in using the one-sided $p$-value as in (21) for defining the critical region of the exact type 1 errors for the AES and the AS methods and not the lower bound of the confidence intervals as defined in (22) for the other methods is that computing the type 1 errors of these two methods (AES and AS) can take much longer time (especially for the AES method) if based on the confidence interval approach. That is because obtaining the confidence intervals for the two methods involved numerical iterative solutions. In addition, for any given situation (sample sizes, NI margin, and true control rate), the AES method would have the same type 1 error regardless of whether the computation is based on the $p$-value approach or based on the lower bound of the confidence interval approach; the same applies to the AS method. That is because for the hypothesized $\delta_0$, both the AES and the AS methods satisfy that the one-sided $p$-value <0.025 if and only if LB > $-\delta_0$ for any data point in the domain $\Omega$.

The exact statistical power at an $\alpha/2$ one-sided test under the alternative $(P_T,P_C)$ is the probability of rejecting the null when the alternative is true and is given by,

$$\text{Power}_{\delta_0}(P_T,P_C) = \sum_{(i,j) \in R_0} \binom{N_T}{i} \binom{N_C}{j} P_T^i(1-P_T)^{N_T-i}(P_C+\delta_0)^j(1-P_C)^{N_C-j}. \quad (24)$$

For the NI tests considered in this article, the power is calculated under both investigational and control groups having identical responses $(P_T = P_C)$. 
6 | RESULTS

Summary comparisons of the exact type I error rates calculated based on the formula given in (23) for the methods summarized in Section 4 for the test of NI versus the one-sided 0.025 nominal value level are presented in Tables 1 and 2 to assess the performance of the AES method versus the other asymptotic methods. The sample sizes used in Tables 1 and 2 were generated using a SAS program based on 80% and 90% statistical power, respectively, using the Farrington and Manning\textsuperscript{13} method to show NI based on various scenarios of the parameters (NI margin, sample size ratio of test: control, and proportion of response in the control group) for the hypothesis test. The true type I error rates were also computed using SAS IML codes for the following combinations: (1) a NI margin of 10%, a randomization ratio Test:Control of 1:2, 1:1, and 2:1, and a proportion of response in the control group of 25%, 40%, 60%, 75%, 90%, and 95%; (2) for the case of the NI margin of 15%, the same combinations as for NI margin of 10% are used without the case of proportion of response in the control group of 95%; (3) for the case of the NI margin of 5%, the true type I error rate is computed in Tables 1 and 2 for the proportion of response in the control group of 95% for the three subject allocation ratios to Test:Control (1:2, 1:1, and 2:1).

6.1 | Comparison of type I error

From both Tables 1 and 2, each method’s type I error was evaluated based on 72 (36 per table) different combinations of statistical power, randomization ratio, NI margin, and control proportion. For easy identification, type I errors in Tables 1 and 2 higher than the 2.5% nominal value are in bold fonts and those below 2.0% are also in bold fonts. Table 3 summarizes the type I errors across the various scenarios presented in Tables 1 and 2.

6.1.1 | Comparisons of the asymptotic methods

The empirical evidence based on the type I errors for the 72 scenarios summarized in Table 3 and presented in Tables 1 and 2 confirmed that the AS method has the best properties in terms of having its type I error closest to the 2.5% nominal level on average versus all other asymptotic methods. In comparison to the AS method: (1) the NC and AC methods had type I errors that were very frequently higher than the 2.5% nominal level in 69% and 74%, respectively, of the 72 cases; (2) Wald’s method had wide range of the type I errors (1.98%–4.03%) and 49% of them were higher than the 2.5% nominal level versus a narrower range (2.05%–2.84%) of those of the AS method and with 42% of them being higher than the 2.5% nominal level; (3) the methods with continuity correction of HA and NCC were frequently very conservative with many of their type I errors being much lower than the 2.5% nominal level, making them frequently having less statistical power than the AS method.

6.1.2 | Comparisons of the AS and the AES methods

The empirical comparison based on the summary of the type I errors displayed in Table 3 supports that the AES method outperforms the AS method based on all displayed measures. Specifically, the type I errors were higher than the 2.5% nominal level in 8% for the AES method versus 42% for the AS method, their average distance from the 2.5% nominal level was 0.046 for the AES method versus 0.120 for the AS method, they ranged from 2.31% to 2.58% (range = 0.27) for the AES method versus from 2.05% to 2.84% (range = 0.79) for the AS method, the average of those ≤2.5% was 2.45% for the AES method versus 2.38% for the AS method, and the average of those >2.5% was 2.53% for the AES method versus 2.62% for the AS method.

In addition, when the scenario for randomization ratio is 1:2 in test: control groups, the type I errors for the AS method decrease from values systematically higher than the 2.5% nominal level to values lower than the 2.5% level (Tables 1 and 2). Moreover, when the scenario for randomization ratio is 2:1 the type I errors increase from values lower than 2.5% to values systematically higher than 2.5% as the proportion of the control group increases. To illustrate this problem, we will focus on the empirical data from Table 1 for trials with sample sizes selected at 80% statistical power, for the cases of 2:1 randomization ratio, NI margin 10%, proportions of the control group: 0.25, 0.40, 0.60, 0.75,
The type I errors for the AS method (systemically increased beyond 2.5% level) were: 2.39%, 2.47%, 2.49%, 2.59%, 2.69%, 2.82%, respectively, versus the type I errors for the AES method: 2.48%, 2.50%, 2.48%, 2.48%, 2.48%, and 2.38%, respectively.

### Table 1: True type 1 errors for various tests for studies with 80% power at nominal $\alpha = 2.5\%$

| $\delta_0$ | $N_C:N_T$ | $P_C$ | $N_T$ | Wald  | AC     | HA     | NCC    | NC     | AS     | AES    |
|----------|------------|-------|-------|-------|--------|--------|--------|--------|--------|--------|
| 0.10     | 1:2        | 0.25  | 207   | 2.21  | 2.45   | 1.83   | 2.31   | 2.77   | 2.65   | 2.49   |
|          |            | 0.40  | 275   | 2.34  | 2.50   | 2.06   | 2.30   | **2.62**| **2.56**| 2.48   |
|          |            | 0.60  | 285   | 2.46  | 2.47   | 2.19   | 2.18   | 2.46   | 2.44   | 2.44   |
|          |            | 0.75  | 233   | **2.70**| **2.60**| 2.34   | 2.13   | 2.48   | 2.44   | 2.47   |
|          |            | 0.90  | 132   | **3.38**| **2.80**| **2.60**| **1.72**| 2.33   | 2.29   | **2.51**|
|          |            | 0.95  | 90    | **3.94**| **3.08**| **3.05**| **1.41**| 2.05   | 2.05   | 2.39   |
| 1:1      |            | 0.25  | 295   | **2.57**| **2.57**| 2.25   | 2.16   | 2.57   | 2.50   | 2.50   |
|          |            | 0.40  | 374   | 2.46  | **2.54**| 2.27   | 2.28   | 2.57   | 2.51   | 2.46   |
|          |            | 0.60  | 374   | 2.37  | 2.43   | 2.36   | 2.36   | 2.50   | 2.40   | 2.38   |
|          |            | 0.75  | 295   | 2.50  | **2.54**| 2.24   | 2.24   | **2.58**| 2.50   | 2.47   |
|          |            | 0.90  | 154   | **2.78**| **2.76**| 2.19   | **1.97**| 2.55   | 2.36   | 2.47   |
|          |            | 0.95  | 99    | **3.27**| **2.91**| 2.47   | **1.67**| 2.42   | 2.32   | 2.32   |
| 2:1      |            | 0.25  | 466   | **2.88**| **2.67**| 2.48   | 2.02   | 2.42   | 2.39   | 2.48   |
|          |            | 0.40  | 570   | **2.62**| **2.57**| 2.31   | 2.20   | 2.52   | 2.47   | 2.50   |
|          |            | 0.60  | 550   | 2.47  | 2.48   | 2.20   | 2.23   | **2.54**| 2.49   | 2.48   |
|          |            | 0.75  | 414   | 2.29  | 2.49   | **1.94**| 2.27   | **2.65**| **2.59**| 2.48   |
|          |            | 0.90  | 194   | 2.13  | 2.52   | **1.52**| 2.16   | **2.97**| **2.69**| 2.48   |
|          |            | 0.95  | 116   | 2.00  | **2.67**| **1.37**| **1.97**| 2.88   | 2.82   | 2.38   |
| 0.15     | 1:2        | 0.25  | 90    | 2.14  | 2.56   | 1.49   | 2.25   | 2.94   | 2.57   | 2.46   |
|          |            | 0.40  | 120   | 2.22  | 2.48   | **1.81**| 2.26   | 2.76   | 2.63   | 2.48   |
|          |            | 0.60  | 127   | **2.61**| **2.61**| 2.12   | 2.20   | 2.61   | 2.61   | 2.47   |
|          |            | 0.75  | 106   | **2.76**| **2.67**| 2.22   | 2.03   | **2.53**| 2.46   | 2.50   |
|          |            | 0.90  | 65    | **3.55**| **2.87**| **2.56**| **1.65**| 2.22   | 2.19   | **2.51**|
| 1:1      |            | 0.25  | 131   | **2.75**| **2.77**| 2.14   | **1.96**| **2.56**| 2.37   | 2.50   |
|          |            | 0.40  | 165   | 2.47  | **2.63**| 2.11   | 2.16   | **2.66**| **2.58**| **2.52**|
|          |            | 0.60  | 165   | 2.38  | **2.76**| 2.13   | 2.13   | **2.76**| **2.81**| 2.50   |
|          |            | 0.75  | 131   | 2.50  | 2.59   | 2.12   | 2.18   | **2.67**| 2.50   | 2.42   |
|          |            | 0.90  | 73    | **2.78**| **2.74**| 2.08   | **1.83**| 2.73   | 2.37   | 2.38   |
| 2:1      |            | 0.25  | 212   | **3.20**| **2.91**| **2.60**| **1.81**| 2.36   | 2.27   | 2.45   |
|          |            | 0.40  | 254   | **2.73**| **2.65**| 2.23   | 2.07   | **2.52**| 2.47   | 2.47   |
|          |            | 0.60  | 240   | 2.31  | **2.69**| **1.92**| 2.30   | **2.74**| **2.62**| 2.41   |
|          |            | 0.75  | 180   | 2.21  | 2.50   | **1.74**| 2.18   | **2.78**| **2.61**| 2.50   |
|          |            | 0.90  | 88    | 2.02  | 2.41   | **1.21**| 2.27   | **2.92**| **2.83**| 2.41   |
| 0.05     | 1:2        | 0.95  | 289   | **3.41**| **2.81**| **2.82**| **1.76**| 2.17   | 2.17   | 2.43   |
|          |            | 0.95  | 334   | **3.85**| **2.62**| **2.31**| **1.93**| 2.41   | 2.41   | 2.41   |
|          |            | 0.95  | 414   | 2.14  | 2.49   | **1.56**| 2.14   | **2.90**| **2.55**| 2.40   |

Note: Bolded data indicate the type I error value is <2.0% or >2.5%.

Abbreviations: AC, Agresti and Caffo; AES, approximate exact score; AS, asymptotic score; HA, Hauck and Anderson; NC, Newcombe Hybrid Score interval; NCC, Newcombe Hybrid Score continuity corrected interval.
For the scenarios of 1:1 randomization ratio, the AS method controlled the type I error at the 2.5% nominal level in 75% of the cases (18 out of 24 cases) and there was no pattern of systematic lack of control of type I error with various ranges of the proportions in the control group. Based on the scenarios with randomization ratio 1:1, the AES method controlled the type I errors in 92% of the cases at the 2.5% nominal level, in 22 out of the 24 cases.

| $\delta_0$ | $N_T:N_C$ | $P_C$ | $N_T$ | Wald | AC   | HA   | NCC  | NC   | AS   | AES  |
|----------|-----------|-------|-------|-------|------|------|------|------|------|------|
| 0.10     | 1:2       | 0.25  | 280   | 2.24  | 2.48 | 1.89 | 2.32 | 2.71 | 2.59 | 2.49 |
|          |           | 0.40  | 369   | 2.37  | 2.50 | 2.11 | 2.32 | 2.61 | 2.55 | 2.50 |
|          |           | 0.60  | 382   | 2.46  | 2.48 | 2.22 | 2.21 | 2.48 | 2.44 | 2.44 |
|          |           | 0.75  | 310   | 2.67  | 2.60 | 2.35 | 2.17 | 2.49 | 2.46 | 2.49 |
|          |           | 0.90  | 172   | 3.20  | 2.79 | 2.60 | 1.89 | 2.33 | 2.28 | 2.43 |
|          |           | 0.95  | 113   | 4.03  | 2.93 | 2.93 | 1.52 | 2.12 | 2.12 | 2.31 |
| 1:1      |           | 0.25  | 395   | 2.55  | 2.60 | 2.28 | 2.20 | 2.57 | 2.49 | 2.49 |
|          |           | 0.40  | 502   | 2.46  | 2.51 | 2.29 | 2.31 | 2.54 | 2.49 | 2.48 |
|          |           | 0.60  | 502   | 2.53  | 2.53 | 2.20 | 2.21 | 2.53 | 2.53 | 2.53 |
|          |           | 0.75  | 395   | 2.55  | 2.60 | 2.28 | 2.20 | 2.57 | 2.49 | 2.50 |
|          |           | 0.90  | 204   | 2.71  | 2.70 | 2.30 | 2.01 | 2.55 | 2.40 | 2.46 |
| 2:1      |           | 0.25  | 620   | 2.83  | 2.65 | 2.47 | 2.11 | 2.46 | 2.40 | 2.47 |
|          |           | 0.40  | 764   | 2.60  | 2.56 | 2.33 | 2.24 | 2.51 | 2.49 | 2.49 |
|          |           | 0.60  | 738   | 2.40  | 2.57 | 2.14 | 2.33 | 2.58 | 2.57 | 2.54 |
|          |           | 0.75  | 560   | 2.31  | 2.48 | 2.02 | 2.32 | 2.64 | 2.59 | 2.47 |
|          |           | 0.90  | 262   | 2.14  | 2.51 | 1.61 | 2.24 | 2.80 | 2.64 | 2.50 |
|          |           | 0.95  | 154   | 2.16  | 2.67 | 1.29 | 1.95 | 2.83 | 2.59 | 2.34 |
| 0.15     | 1:2       | 0.25  | 121   | 2.15  | 2.56 | 1.63 | 2.30 | 2.82 | 2.66 | 2.42 |
|          |           | 0.40  | 161   | 2.25  | 2.53 | 1.89 | 2.30 | 2.72 | 2.59 | 2.47 |
|          |           | 0.60  | 169   | 2.42  | 2.44 | 2.08 | 2.10 | 2.55 | 2.42 | 2.42 |
|          |           | 0.75  | 140   | 3.35  | 2.98 | 2.67 | 1.70 | 2.36 | 2.20 | 2.36 |
|          |           | 0.90  | 83    | 2.69  | 3.50 | 2.62 | 2.27 | 2.07 | 2.55 | 2.45 | 2.46 |
| 1:1      |           | 0.25  | 176   | 2.67  | 2.80 | 2.21 | 2.08 | 2.64 | 2.49 | 2.49 |
|          |           | 0.40  | 221   | 2.47  | 2.62 | 2.17 | 2.22 | 2.65 | 2.51 | 2.46 |
|          |           | 0.60  | 221   | 2.37  | 2.38 | 2.37 | 2.37 | 2.47 | 2.43 | 2.37 |
|          |           | 0.75  | 176   | 2.50  | 2.62 | 2.11 | 2.16 | 2.64 | 2.56 | 2.50 |
|          |           | 0.90  | 96    | 2.76  | 2.76 | 2.28 | 1.88 | 2.73 | 2.47 | 2.47 |
| 2:1      |           | 0.25  | 280   | 3.16  | 2.84 | 2.54 | 1.87 | 2.42 | 2.28 | 2.45 |
|          |           | 0.40  | 338   | 2.67  | 2.63 | 2.25 | 2.14 | 2.54 | 2.49 | 2.49 |
|          |           | 0.60  | 322   | 2.31  | 2.63 | 1.95 | 2.26 | 2.63 | 2.63 | 2.58 |
|          |           | 0.75  | 242   | 2.22  | 2.50 | 1.81 | 2.26 | 2.76 | 2.62 | 2.50 |
|          |           | 0.90  | 120   | 1.98  | 2.40 | 1.30 | 2.34 | 2.94 | 2.84 | 2.40 |
| 0.05     | 1:2       | 0.95  | 374   | 3.30  | 2.77 | 2.77 | 1.82 | 2.24 | 2.23 | 2.47 |
|          | 1:1       |         |       |       |     |     |     |     |     |     |
|          |           | 0.95  | 440   | 2.82  | 2.65 | 2.35 | 2.02 | 2.48 | 2.37 | 2.48 |
|          | 2:1       |         |       |       |     |     |     |     |     |     |
|          |           | 0.95  | 560   | 2.12  | 2.43 | 1.62 | 2.18 | 2.78 | 2.62 | 2.44 |

Note: Bolded data indicate the type I error value is <2.0% or >2.5%.
Abbreviations: AC, Agresti and Caffo; AES, approximate exact score; AS, asymptotic score; HA, Hauck and Anderson; NC, Newcombe Hybrid Score interval; NCC, Newcombe Hybrid Score continuity corrected interval.

For the scenarios of 1:1 randomization ratio, the AS method controlled the type I error at the 2.5% nominal level in 75% of the cases (18 out of 24 cases) and there was no pattern of systematic lack of control of type I error with various ranges of the proportions in the control group. Based on the scenarios with randomization ratio 1:1, the AES method controlled the type I errors in 92% of the cases at the 2.5% nominal level, in 22 out of the 24 cases.
Comparison of exact statistical power

The exact statistical powers were calculated using Equation (24) and presented in Table 4 for the cases that were included in Table 1 for type 1 errors where the sample sizes were obtained based on nominal 80% power using the Farrington and Manning method. For easy identification, a power value is in bold in Table 4 if its corresponding type 1 error in Table 1 is higher than 2.5%. The bolded power values should not be part of the comparison since their corresponding type 1 errors were not preserved below 2.5%.

Results from Table 4 showed that among the nonbolded power values the statistical power for the AES method is frequently higher than the power values for all other methods or very close to the maximum value of the other methods. In fact, the AES method has statistical power that is almost always higher than those of the two methods with continuity corrections.

7 | EXAMPLES

7.1 | A case study example

Based on Week 48 results from an HIV clinical trial presented by Reynes et al., the proportion of patients who were responders on a new treatment was 84/101 (83.2%) versus the proportion of those who were responders on the standard of care was 89/105 (84.8%). This HIV clinical trial was planned to test for NI based on the ES method, effectively the method by CZ, and it successfully met its prespecified 20% NI margin. Garner summarized these proportions and displayed the confidence interval for their difference based on several methods considering an NI margin of 12%. The confidence intervals from the various methods presented in Section 4 including those based on the AES and the CZ methods are displayed in Table 5 of this article. The one-sided p-values for the test-based methods for NI comparison using NI margins of 20% and 12% are reported as well in Table 5.

As given in Table 5, the study would succeed based on the preplanned NI margin regardless of the considered method since the lower bounds of the CI for all methods are >−20%.

If the NI margin of the study was at 12% with the nominal significant level at 2.5% as presented in Garner, then the CZ method along with NCC and HA would have failed to show NI of the new treatment to the standard one while the AES, AS, Wald, NC, and AC methods would have succeeded in showing NI. The AES method is more appropriate to be considered for the analysis of this trial than the AS, Wald, NC, and AC methods since it has better control of type 1 error. The p-value based on the AES method is 0.0238 and it was computed using Expression (11) by replacing the nuisance parameter P by its restricted MLE of 76.4%.

If the AES method was selected for the primary NI analysis with NI margin of 12% for such data, then a sensitivity analysis would be needed to confirm that the observed significant result would stay significant based on analysis with adjustment to the true type I error at the 2.5% level. Exact methods are useful in such assessment. The review of the significance range based on the profile of the p-value calculated as the tail probability in Expression (10) over the nuisance
parameter in its domain [0, 88\%] can be very helpful. First, the exact p-value based on the ES method is 0.0275. This exact method failed to support the significance based on the AES method. However, the p-value of the ES method was calculated at the nuisance parameter value of 36.9\%, while the restricted MLE of the nuisance parameter (P) was 76.4\%.

### Table 4: True powers for various tests for studies with 80% power at nominal \( \alpha = 2.5\% \).

| \( \delta_0 \) | \( N_T:N_C \) | 1:2 | 2:1 | Pc | N_T | Wald | AC | HA | NCC | NC | AS | AES |
|----------------|----------------|-----|-----|----|-----|------|----|----|-----|----|----|-----|
| 0.10           |                |     |     |    | 0.25| 207  | 78.25| 79.46| 76.11| 78.45| 80.46| 80.26| 79.51|
|                |                |     |     | 0.40| 275 | 79.25| 79.81| 77.62| 78.65| 80.24| 80.01| 79.74|
|                |                |     |     | 0.60| 285 | 80.25| 80.33| 78.78| 78.55| 80.24| 79.92| 79.80|
|                |                |     |     | 0.75| 233 | 81.48| 80.91| 79.75| 78.21| 80.13| 80.11| 80.28|
|                |                |     |     | 0.90| 132 | 85.89| 83.56| 83.23| 76.84| 81.05| 81.05| 81.77|
|                |                |     |     | 0.95| 90  | 90.06| 87.56| 87.56| 77.17| 81.52| 81.52| 84.63|
| 1:1            |                |     |     |    | 0.25| 295  | 80.24| 80.46| 78.66| 78.58| 80.47| 80.21| 79.99|
|                |                |     |     | 0.40| 374 | 80.02| 80.37| 78.43| 78.46| 80.45| 80.34| 80.18|
|                |                |     |     | 0.60| 374 | 80.02| 80.37| 78.43| 78.46| 80.45| 80.34| 80.18|
|                |                |     |     | 0.75| 295 | 80.24| 80.46| 78.66| 78.58| 80.47| 80.21| 79.99|
|                |                |     |     | 0.90| 154 | 83.54| 82.78| 80.90| 78.02| 81.34| 80.85| 80.85|
|                |                |     |     | 0.95| 99  | 88.76| 87.63| 86.02| 76.87| 83.47| 83.46| 83.35|
| 2:1            |                |     |     |    | 0.25| 466  | 81.48| 80.91| 79.75| 78.21| 80.13| 80.11| 80.28|
|                |                |     |     | 0.40| 570 | 80.25| 80.33| 78.78| 78.55| 80.24| 79.92| 79.80|
|                |                |     |     | 0.60| 550 | 79.25| 79.81| 77.62| 78.65| 80.24| 80.01| 79.74|
|                |                |     |     | 0.75| 414 | 78.25| 79.46| 76.11| 78.45| 80.46| 80.26| 79.51|
|                |                |     |     | 0.90| 194 | 79.77| 81.67| 74.60| 79.07| 83.41| 81.29| 80.80|
|                |                |     |     | 0.95| 116 | 85.23| 89.50| 82.06| 81.41| 84.28| 84.28| 82.85|
| 0.15           |                |     |     |    | 0.25| 90   | 77.72| 79.92| 74.14| 78.48| 81.35| 80.38| 79.40|
|                |                |     |     | 0.40| 120 | 78.73| 79.61| 76.26| 78.24| 80.71| 80.21| 79.60|
|                |                |     |     | 0.60| 127 | 80.35| 80.35| 78.20| 78.15| 80.30| 80.26| 80.15|
|                |                |     |     | 0.75| 106 | 82.00| 81.47| 79.55| 77.61| 80.52| 80.18| 80.51|
|                |                |     |     | 0.90| 65  | 87.51| 84.75| 84.22| 76.92| 81.51| 81.51| 82.73|
| 1:1            |                |     |     |    | 0.25| 131  | 80.32| 80.87| 78.08| 78.07| 80.80| 80.27| 79.79|
|                |                |     |     | 0.40| 165 | 80.02| 80.11| 77.17| 77.82| 80.17| 80.08| 80.05|
|                |                |     |     | 0.60| 165 | 80.02| 80.11| 77.17| 77.82| 80.17| 80.08| 80.05|
|                |                |     |     | 0.75| 131 | 80.32| 80.87| 78.08| 78.07| 80.80| 80.27| 79.79|
|                |                |     |     | 0.90| 73  | 85.73| 84.40| 81.49| 76.68| 83.02| 81.68| 81.68|
| 2:1            |                |     |     |    | 0.25| 212  | 82.00| 81.47| 79.55| 77.61| 80.52| 80.18| 80.51|
|                |                |     |     | 0.40| 254 | 80.35| 80.35| 78.20| 78.15| 80.30| 80.26| 80.15|
|                |                |     |     | 0.60| 240 | 78.73| 79.61| 76.26| 78.24| 80.71| 80.21| 79.60|
|                |                |     |     | 0.75| 180 | 77.72| 79.92| 74.14| 78.48| 81.35| 80.38| 79.40|
|                |                |     |     | 0.90| 88  | 82.29| 83.09| 74.27| 79.94| 84.40| 82.36| 82.30|
| 0.05           |                |     |     |    | 0.95| 289  | 86.78| 83.96| 84.28| 77.78| 80.68| 80.68| 82.12|
|                |                |     |     | 1:1 | 0.95| 334  | 84.20| 82.90| 81.94| 78.79| 81.55| 81.55| 81.58|
|                |                |     |     | 2:1 | 0.95| 414  | 79.46| 81.46| 74.99| 79.33| 82.70| 81.19| 81.11|

**Note:** Sample sizes were obtained based on 80% power using the method by Farrington and Manning.\(^{13}\) Bolded power value indicates that its corresponding type 1 error in Table 1 is higher than 2.5\%. Abbreviations: AC, Agresti and Caffo; AES, approximate exact score; AS, asymptotic score; HA, Hauck and Anderson; NC, Newcombe Hybrid Score interval; NCC, Newcombe Hybrid Score continuity corrected interval.
Given that the exact \( p \)-value of the ES method was calculated at nuisance parameter value that is much lower than the MLE of \( P \), the sensitivity analysis should include identification of the range of the nuisance parameter where the profile \( p \)-value is >0.025. The profile \( p \)-value was >0.025 level when the nuisance parameter was in the range from 27.3% to 47.1% and <0.025 in the other regions of the full domain \([0, 88\%]\). If the true nuisance parameter is 60%, then the exact probability of the restricted MLE of the nuisance parameter \( P \) as high as the observed MLE of \( P \) of 76.4% or higher is \( 5.1 \times 10^{-8} \). This very small probability \( 5.1 \times 10^{-8} \) strongly suggests that it is extremely unlikely that the true nuisance parameter \( P \) to be in the range from 0% to 60%. This probability statement along with the \( p \)-value of 0.0244 calculated at the maximum value from the profile of the \( p \)-value of the ES method over the range \((60\%, 88\%)\) of the nuisance parameter support the NI significant finding of the test treatment at the 2.5% level.

Such sensitivity analysis supports the robustness that the study can be considered successful in demonstrating NI based on 12% NI margin if the AES is the primary planned method for NI assessment. However, such a conclusion could not be reached if the ES method is the planned one for the primary analysis in this example.

### Hypothetical examples

Three hypothetical examples of NI analysis are presented in Supplement II. The methods used for the analysis include the AES and the six asymptotic methods. SAS IML programs were prepared and used on a MacBook Pro computer to generate the confidence intervals for these methods and the \( p \)-values for the AES and AS methods. These three examples provide further understanding of the advantage of the AES method over the other methods in the comparisons. Also, the computation time for the confidence interval and the \( p \)-value for the AES method in the example with the largest sample size, 876 subjects, was only 23 s.

### DISCUSSION

Nurminen and Newcombe concluded, in an article published in 2009, that the MN score interval estimates are highly recommended on account of both theoretical arguments and empirical evidence for use in epidemiology and medical research practice. The Miettinen and Nurminen score interval method is the same as the AS method in this article except the denominator of the MN method included a bias correction factor for the variance estimate, \( N/(N - 1) \) with
\[ N = N_T + N_C, \] which makes no practical difference in confirmatory studies with sample sizes frequently above 150 subjects per treatment group. In fact, empirical type I errors and statistical powers summarized for the various scenarios in Tables 1–4 confirm their conclusion in terms of the advantage of the AS method over the other five asymptotic methods. The AS method has average type I errors much closer to the 2.5% nominal level than the two methods with continuity corrections as their type I errors are frequently much lower than the nominal 2.5% level. Also, the AS method is preferred over the Wald, AC, and the NC methods by having a smaller percentage of type I errors being >2.5% nominal level. In addition, for the cases of the AS method with type I errors >2.5% nominal level, on average those type I errors stay closer to the nominal 2.5% level.

On a theoretical basis, both the AES and AS methods depend on the constrained MLEs of the proportions of the compared groups in estimating the variance of the binomial probabilities. The only important difference between the two methods is that the AES method utilizes the exact binomial probabilities with the nuisance parameter being replaced by its constrained MLE of the binomial proportions to obtain the \( p \)-values and the confidence intervals while the AS method approximates the exact binomial through the asymptotic normal distribution to facilitate its computations without the need to deal with the nuisance parameter. Because the AES method does not use asymptotic approximation of the binomial distribution, it is expected to have better statistical properties than the AS method and that was confirmed empirically by the type I errors summarized and displayed in Tables 1–3. Based on the results presented in Tables 1–3 for trials with sample sizes commonly considered in NI confirmatory studies, the type I errors for the AES method were frequently (92%) below or equal to the nominal value of 2.5% than those of the AS method (58%), had much closer distance to the 2.5% nominal level than those of the AS method, had narrower range than those of the AS method, and for the cases with type I errors >2.5% nominal level, the type I errors for the AES method stayed closer to the nominal value than those of the AS method.

The AS method had a systematic problem of not controlling the type I errors at the 2.5% nominal level with specific scenarios based on the empirical type I errors displayed in Tables 1 and 2. Specifically, the AS method had type I errors that are systematically higher than the 2.5% nominal level when the number of subjects in the test: control groups had a ratio of 2:1 and the proportion of response in the control group is 60% or higher and this problem increased as the proportion of the control group increased to 95%. In addition, this problem of systematically higher type I errors than the nominal level appeared for the AS method for the case of randomization ratio of 1:2 combined with low proportions (25% and 40%) in the control group. Based on other empirical data not presented in this article, we observed that a general problem with the AS method is that it is systematically not controlling the type I errors at the nominal level when the randomization ratio is imbalanced between the compared treatment groups for a specific range of the proportion of response in the control group. Unlike the AS method, the type I errors for the AES method are frequently controlled at the nominal level regardless of randomization ratios and/or of range of the proportions of responses of the treatment groups.

In practice of analysis of confirmatory clinical trials, so far as significance testing is concerned, NI effect will confidently be considered significant (rejecting inferiority and concluding NI) only if it remains significant based on sensitivity analysis that corrects for inflated type I error over the nominal level of the primary comparison. As shown earlier, this can affect the AS method when the \( p \)-value is significant and is close to the nominal value, but it is less of a concern for the AES method.

Unlike other exact methods, the AES method was not developed to control the type I error at the nominal level at the full domain of the nuisance parameter and therefore it is not as conservative while it preserves the type I error frequently at the nominal level (as shown in 92% of the considered cases in Tables 1–3). Availability of the confidence interval for analyzing NI clinical trial data is very important as it is the primary statistical method for analyzing and reporting NI tests in clinical reports and product labels.

It is important to note that for any trial outcome for the considered NI hypothesis with fixed NI margin, by definition the \( p \)-value for the AES test is less than or equal to the \( p \)-value for the ES test, leading to the AES test to have higher or similar statistical power to the ES test. Based on the type I error formula provided in (21), we can also conclude that the type I error for the AES test is greater than or equal to that of the ES test. When planning a Phase 3 NI binomial trial, pharmaceutical statisticians are entitled to select a test for the primary NI comparison that uses the full 2.5% type I error value, but the actual size of any exact test that preserves the type I error at 2.5% level is below 2.5% (cannot be exactly 2.5%) leading to an unnecessarily conservative analysis strategy. As shown in the Cases Study Example of Section 7.1, it would be a more appropriate analysis strategy to consider the AES method over the ES method for the primary analysis; only if the \( p \)-value of the AES method is significant and close to 2.5%, then sensitivity analyses with respect to type I error control could be considered for such a situation to confirm the success of the study. The sensitivity analysis strategy can include exact methods such as the ES and the \( E + M \) (as recommended by Ripamonti
et al.\textsuperscript{14}) methods and assessment of potential loss of significance over the domain of the nuisance parameter as shown in the Case Study Example should be carefully investigated.

The actual computation time to generate the confidence interval and the \( p \)-value for the AES method for the difference between two proportions is \(<30\) s for most studies with \(80\%\text{–}90\%\) statistical power. To assess real computer time for the analysis of larger trials we ran the SAS IML program to generate the confidence interval and \( p \)-value based on examples of studies with total sample sizes in the range of 2000–2100 subjects divided between the two compared groups with randomization ratios of 2:1, 1:1, and 1:2. The real computer time for these examples ranged from 1 to 2 min, suggesting the AES method is very feasible for consideration in terms of computation time for trials with sizes in excess of 2000 subjects.

The odds ratio can also be a useful criterion to assess NI in clinical trials. An advantage of the odds ratio is that the same margin can be useful when there is uncertainty about the expected rate in the control treatment group. For example, an odds ratio NI margin of 4/9, or 0.44, can be useful in addressing a rate difference near 0.05 when the control rate is near 0.95, a rate difference near 0.10 when the control rate is near 0.90, and a rate difference near 0.15 when the control rate is near 0.80. Methods that can be used for the analysis of NI based on the odds ratio criterion include the asymptotic score confidence interval for the odds ratio by Miettinen and Nurminen,\textsuperscript{11} the exact mid-\( p \) confidence interval for the odds ratio by inverting two 1-sided (equal tail) noncentral hypergeometric tests, and the exact confidence interval for the odds ratio by inverting two 1-sided noncentral hypergeometric tests. All three confidence interval methods are conveniently available in SAS, PROC FREQ.

Finally, to test the NI hypothesis based on rate difference of the investigational group to the control group, we recommend using the AES method over the AS method for the analysis of confirmatory clinical trials and various similar medical research, given its theoretical advantage and the empirical evidence of being the method with better statistical properties. In addition, given the superior performance of the AES method, we encourage statistical software developers to make this method available in their future updates.

The SAS IML codes used to generate the \( p \)-value and the confidence interval for the AES method are made available in Supplement III to this article.

ACKNOWLEDGMENTS

The authors thank Haiyuan Zhu and anonymous reviewers for comments that improved the article.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

ORCID

Andrew Lakkis \textsuperscript{a} https://orcid.org/0000-0002-7344-736X

REFERENCES

1. Hauck WW, Anderson S. A comparison of large sample confidence interval methods for the differences of two binomial probabilities. \textit{Am Stat}. 1996;40:318-322.
2. Newcombe RG. Interval estimate for the difference between independent proportions: comparison of eleven methods. \textit{Stat Med}. 1998;17:873-890.
3. Santner TJ, Snell MK. Small-sample confidence interval for \( p_1-p_2 \) and \( p_1/p_2 \) in \( 2 \times 2 \) contingency tables. \textit{J Am Stat Assoc}. 1980;75:386-394.
4. Chan ISF, Zhang Z. Test-based exact confidence intervals for the difference of two binomial proportions. \textit{Biometrics}. 1999;55:1201-1209.
5. Chan ISF. Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. \textit{Stat Med}. 1998;17:1403-1413.
6. SAS Institute Inc. \textit{The SAS System, Version 9.3}. SAS Institute Inc.; 2011. \url{http://www.sas.com/}
7. Wang W. On construction of the smallest one-sided confidence interval for the difference of two proportions. \textit{Ann Stat}. 2010;38(2):1227-1243.
8. Wang W. An inductive order construction for the difference of two dependent proportions. \textit{Stat Probab Lett}. 2012;82(8):1623-1628.
9. Shan G, Wang W. ExactCIdiff: Inductive Confidence Intervals for the difference between two proportions 2013. \url{http://CRAN.R-project.org/package=ExactCIdiff}. R package version 1.3. [p. 62].
10. Garner W. \textit{Constructing Confidence Intervals for the Differences of Binomial Proportions in SAS}. Gilead Sciences, Inc; 2022. Accessed March 2022; [p. 1-13]. \url{https://lexjansen.com/wuss/2016/127_Final_Paper_PDF.pdf}
11. Miettinen O, Nurminen M. Comparative analysis of two rates. Stat Med. 1985;4:213-226.
12. Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Am Stat. 2000;54:280-288.
13. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med. 1990;9:1447-1454.
14. Ripamonti E, Lloyd C, Quatto P. Contemporary frequentist views of the $2 \times 2$ binomial trial. Stat Sci. 2017;32:600-615.
15. Berger RL, Boos DD. P values maximized over a confidence set for the nuisance parameter. J Am Stat Assoc. 1994;89:1012-1016.
16. Lloyd CI. Exact p-values for discrete models obtained by estimation and maximization. Aust N Z J Stat. 2008;50:329-345.
17. Fay MP, Hunsberger SA. Practical valid inferences for the two-sample binomial problem. Stat Sur. 2021;15:72-110.
18. Fay MP, Proschan MA, Brittain E. Combining one sample confidence procedures for inference in the two-sample case. Biometrics. 2015;71:146-156.
19. Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. Stat Methods Med Res. 2015;24(2):224-254.
20. Dann RS, Koch GG. Methods for one-sided testing of the difference between proportions and sample size considerations related to non-inferiority clinical trials. Pharm Stat. 2008;7:130-141.
21. Rothmann MD, Wiens BL, Chan ISF. Design and Analysis of Non-inferiority Trials. Chapman & Hall/CRC; 2012; [p. 1-438].
22. Kang SH, Chen JJ. An approximate unconditional test of no-inferiority between two proportions. Stat Med. 2000;19:2089-2100.
23. U.S. Department of Health and Human Services, Food and Drug Administration, CDER. Acute bacterial skin and skin structure infections. Draft 2013; [p. 1-15]. https://www.fda.gov/regulatoryinformation/guidances/default.htm.
24. U.S. Department of Health and Human Services, Food and Drug Administration, CDER. Complicated urinary tract infections. Draft, February 2018; [p. 1-16]. https://www.fda.gov/regulatoryinformation/guidances/default.htm.
25. U.S. Department of Health and Human Services, Food and Drug Administration, CDER. Complicated intra-abdominal infections. Revision 1, May 2018; [p. 1-17]. https://www.fda.gov/regulatoryinformation/guidances/default.htm.
26. U.S. Department of Health and Human Services, Food and Drug Administration, CDER. Community-acquired bacterial pneumonia. Draft, June 2020; [p. 1-17]. https://www.fda.gov/regulatoryinformation/guidances/default.htm.
27. Temple R. Problems in interpreting active control equivalence trials. Account Res. 1996;4:267-275.
28. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) E10: Choice of control group and related issues in clinical trials. July 2000; [p. 1-29]. database.ich.org
29. Hawila N, Berg A. Exact-corrected confidence interval for risk difference in noninferiority binomial trials. Biometrics. 2022;79:1-12. doi: 10.1111/biom.13688
30. Sussa S, Shuster JJ. Exact unconditional sample sizes for the $2 \times 2$ binomial trial. J R Stat Soc A. 1985;148:317-327.
31. Barnad G. Significance tests for $2 \times 2$ tables. Biometrika. 1947;34(1/2):123-138.
32. Röhmel J, Mansmann U. Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. Biometrical J. 1999;41(2):149-170.
33. Röhmel J, Mansmann U. 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). Stat Med. 1999;18(13):1734-1737.
34. Chan IS. Author’s reply on ‘exact tests of equivalence and efficacy with a non-zero lower bound for comparative trials’. Stat Med. 1999; 18:1735-1737.
35. Chan IS. Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. Stat Meth Med Res. 2003;12(1):37-58.
36. Storer BE, Kim C. Exact properties of some exact test statistics for comparing two binomial proportions. J Am Stat Assoc. 1990;85:146-155.
37. Chen X. A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases. Stat Med. 2002;21:943-956.
38. Reynes J, Lawal A, Pulido F, et al. Examination of noninferiority, safety, and tolerability of Lopinavir/ritonavir and raltegravir compared with lopinavir/ritonavir and tenofovir/emtricitabine in antiretroviral-naive subjects. The PROGRESS study, 48-week results. HIV Clinical Trials. 2011;12(5):255-267. doi:10.1310/htc1205-255
39. Nurminen MM, Newcombe RG. Methodological Notes of the Accuracy of the MN Score Intervals for the Difference of Two Independent Binomial Proportions. MarkStat Consultancy; 2009, Commentary. Accessed March 2022; [p. 1-12]. http://www.markstat.net/en/images/stories/methodological_notes_on_score_intervals.pdf

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lakkis H, Lakkis A. Analysis of two binomial proportions in noninferiority confirmatory trials. Pharmaceutical Statistics. 2024;23(2):257-275. doi:10.1002/pst.2351