PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.287

MAIN PAPER

Methods for one-sided testing of the difference between proportions and sample size considerations related to non-inferiority clinical trials

Rebekkah S. Dann[1,*,†] and Gary G. Koch[2]
[1] *GlaxoSmithKline, Research Triangle Park, NC, USA*
[2] *Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*

Assessment of non-inferiority is often performed using a one-sided statistical test through an analogous one-sided confidence limit. When the focus of attention is the difference in success rates between test and active control proportions, the lower confidence limit is computed, and many methods exist in the literature to address this objective. This paper considers methods which have been shown to be popular in the literature and have surfaced in this research as having good performance with respect to controlling type I error at the specified level. Performance of these methods is assessed with respect to power and type I error through simulations. Sample size considerations are also included to aid in the planning stages of non-inferiority trials focusing on the difference in proportions. Results suggest that the appropriate method to use depends on the sample size allocation of subjects in the test and active control groups. Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: ***difference in proportions; sample size; one-sided testing; non-inferiority trials; sample size allocation***

1. INTRODUCTION

Proportions are used in many clinical trials to describe the distributions of dichotomous response variables with independent binomial distributions for treatments under study. In a non-inferiority setting, the goal is to show that the investigational treatment (test) group is no worse than an active control group by a predetermined non-inferiority margin. The ICH-E9 guidelines [1] and more recently the European Medicines Agency guidelines [2] suggest that in this non-inferiority setting comparisons between treatment groups be made

*Correspondence to: Rebekkah S. Dann, GlaxoSmithKline, 5 Moore Drive, Research Triangle Park, NC 27709, USA.
†E-mail: rebekkah.s.dann@gsk.com

with a one-sided test through an analogous confidence interval for the difference between the two treatment group proportions. For proportions pertaining to favorable response, this approach to testing requires the lower confidence bound on the difference between the test and control groups to be larger than this margin in order to demonstrate that the test treatment is not inferior to the active control with respect to efficacy.

There are many methods in the statistical literature for computing the confidence interval for the difference between two independent binomial proportions, and their lower bounds are potentially applicable to the testing of the one-sided non-inferiority hypothesis. This discussion is not intended to include all possible methods, but will consider methods which have been shown to be popular in the literature and have surfaced in this research as having good performance with respect to controlling type I error at the specified level. The results in this paper will extend findings by Roebruck and Kühn [3] to include two additional methods which appeared later in the literature and the findings of Li and Chuang-Stein [4] to include broader discussions related to power and sample size calculation for various sample size allocations.

The lower limit of the confidence interval for the difference in proportions exceeding the non-inferiority margin (Δ_0) is used as the counterpart to a test statistic for H_0: $(\pi_T - \pi_C) \leqslant \Delta_0$ versus H_A: $(\pi_T - \pi_C) > \Delta_0$ as the alternative hypothesis for non-inferiority, for values of $\Delta_0 < 0$. Discussion and results will focus on the lower confidence limit through its provision of a one-sided test of the null hypothesis H_0.

2. METHODS

Historically, the most well-known method for computing the confidence interval for a difference

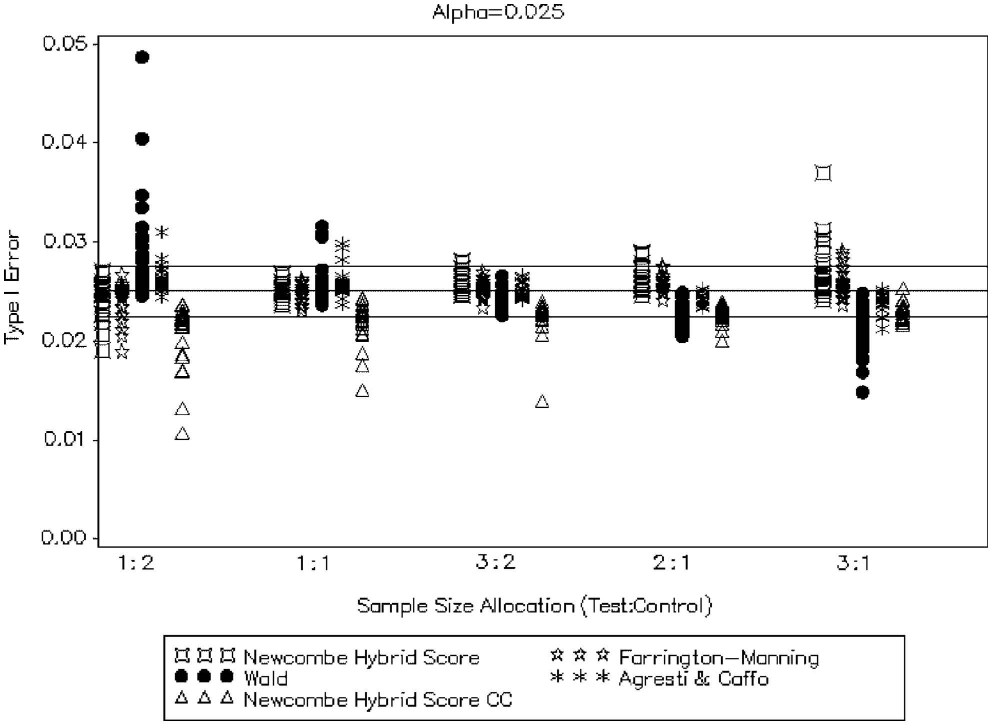
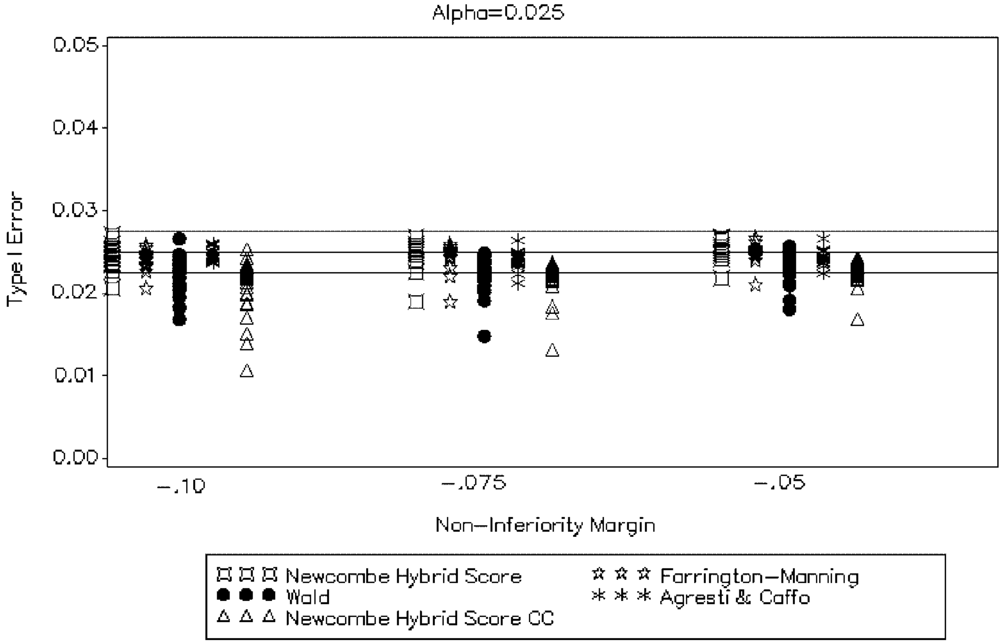


Figure 1. Summary of simulated type I error by sample size allocation.

DOI: 10.1002/pst



Note: Newcombe hybrid score and Farrington–Manning methods summarized for 1:2 & 1:1 allocations only
Wald and Agresti & Caffo methods summarized for 3:2, 2:1, and 3:1 allocations only
Newcombe hybrid score CC summarized for all allocations

Figure 2. Summary of simulated type I error by non-inferiority margin.

between two independent binomial proportions is the Wald method as based on a normal approximation. As shown in (1), this method is straightforward for computation and understanding, and so it is presented in most basic statistics textbooks and implemented in standard statistical software packages

$$\{\hat{p}_T - \hat{p}_C\} - z_{1-\alpha}\sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}} \qquad (1)$$

In (1), $\hat{p}_T = y_T/n_T$ is the observed proportion for favorable response in the test group with y_T representing the total number of such outcomes in a total sample size of n_T, and $\hat{p}_C = y_C/n_C$ is the observed proportion for favorable response in the control group with y_C representing the total number of such outcomes in a total sample size of n_C. Also, $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution, where α is the significance level set for the one-sided test. This method has traditionally been shown to have poor performance for even moderate sample sizes with respect to excessive inflation of the type I error rate when consideration includes the entire confidence interval using both the upper and lower bounds [5,6]. However, Roebruck and Kühn [3] found this method to perform adequately for sample sizes large enough to yield power of at least 0.70 where the sample size allocation for test:control is 3:2 for the one-sided limit as a one-sided test. Li and Chuang-Stein also found this method to perform well in an equal allocation setting when event rates were moderate enough so as to provide expected cell frequencies of at least 15 for all cells [4].

Farrington and Manning [7] proposed a method for the difference in proportions with the form shown in (2) which is the same as the Wald method in (1), but they determined the proportions $\tilde{\pi}_T$ and $\tilde{\pi}_C$ in the expression for variance using maximum

DOI: 10.1002/pst

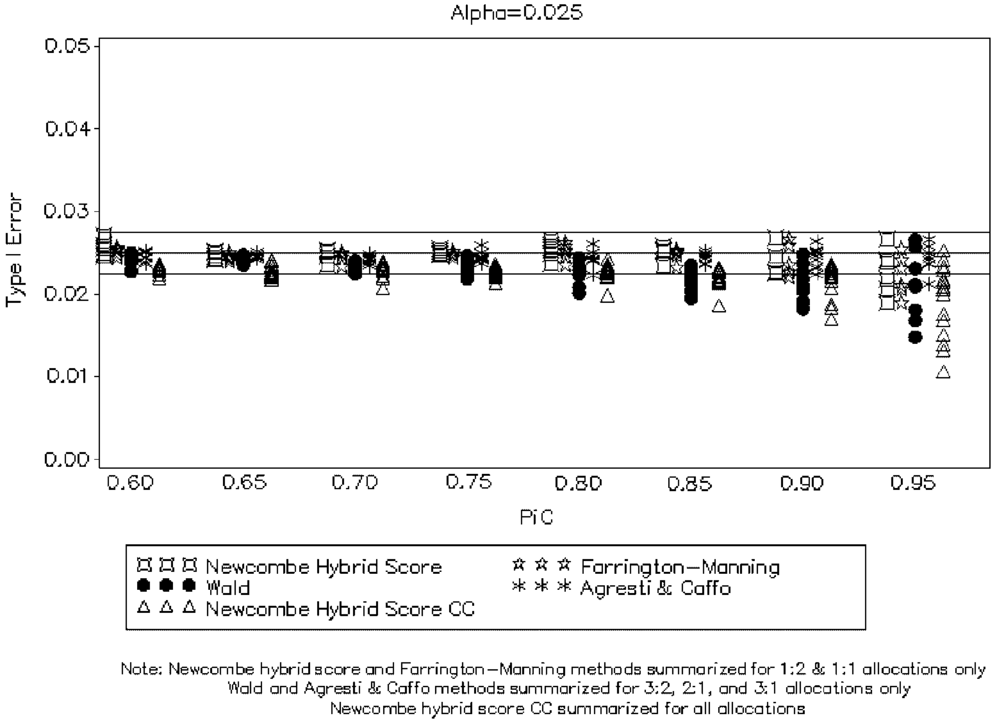


Figure 3. Summary of simulated type I error by Pi C.

likelihood estimates under the null hypothesis H_0 of inferiority at Δ_0

$$\{\hat{p}_{\mathrm{T}} - \hat{p}_{\mathrm{C}}\} - z_{1-\alpha}\sqrt{\frac{\tilde{\pi}_{\mathrm{T}}(1-\tilde{\pi}_{\mathrm{T}})}{n_{\mathrm{T}}} + \frac{\tilde{\pi}_{\mathrm{C}}(1-\tilde{\pi}_{\mathrm{C}})}{n_{\mathrm{C}}}} \tag{2}$$

Farrington and Manning [7] discussed their computation as closed-form solutions for $\tilde{\pi}_{\mathrm{T}}$ and $\tilde{\pi}_{\mathrm{C}}$, which are somewhat complicated to implement. Software, such as SAS [8], can be used to compute these maximum likelihood estimates, and these values can then be placed in (2) to produce Farrington–Manning confidence limits. In this regard for implementation in SAS through PROC GENMOD, a procedure used to fit general linear models, there would be specification of a binomial distribution with an identity link. The model statement fits only the intercept and includes an offset term where the offset for the control group is zero and for the test group is set equal to the specified non-inferiority margin Δ_0. Roebruck and Kühn [3] indicate that the Farrington–Manning method is generally an appropriate method and suggest that its behavior with respect to controlling type I error at the nominal level is similar across sample size allocations (2:3, 1:1, 3:2) of test:control.

Agresti and Caffo [5] developed an adjustment to the Wald confidence interval to produce results that maintain the nominal type I error better than the Wald method, while still being straightforward to calculate. As shown in (3), this method uses adjusted proportions when computing the confidence limits

$$\{\tilde{p}_{\mathrm{T}} - \tilde{p}_{\mathrm{C}}\} - z_{1-\alpha}\sqrt{\frac{\tilde{p}_{\mathrm{T}}(1-\tilde{p}_{\mathrm{T}})}{n_{\mathrm{T}}+2} + \frac{\tilde{p}_{\mathrm{C}}(1-\tilde{p}_{\mathrm{C}})}{n_{\mathrm{C}}+2}} \tag{3}$$

In (3), $\tilde{p}_{\mathrm{T}} = (y_{\mathrm{T}}+1)/(n_{\mathrm{T}}+2)$ and $\tilde{p}_{\mathrm{C}} = (y_{\mathrm{C}}+1)/(n_{\mathrm{C}}+2)$ where these proportions are calculated by adding one success and one failure to each group and thereby two successes and two failures in total.

DOI: 10.1002/pst

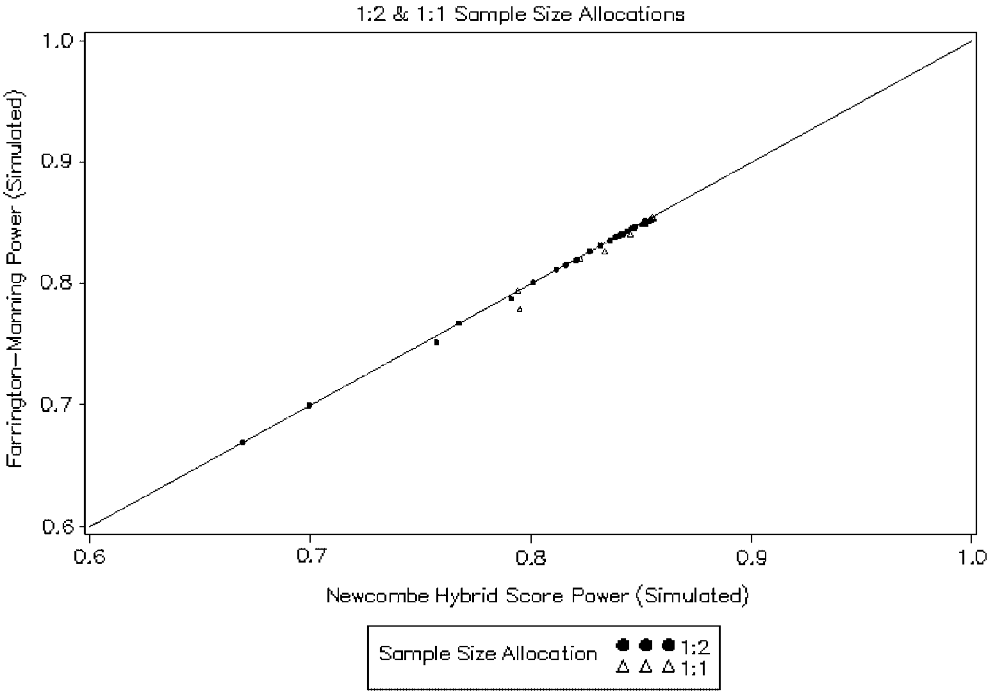


Figure 4. Summary of simulated power.

Newcombe [6] provided a method based on the Wilson score method for a single proportion. For this Newcombe hybrid score interval, one solves $|p_{\mathrm{T}} - \hat{p}_{\mathrm{T}}| = z_{1-\alpha}\sqrt{p_{\mathrm{T}}(1-p_{\mathrm{T}})/n_{\mathrm{T}}}$ for p_{T} resulting in two solutions, l_{T} and u_{T}. Similarly, the equation $|p_{\mathrm{C}} - \hat{p}_{\mathrm{C}}| = z_{1-\alpha}\sqrt{p_{\mathrm{C}}(1-p_{\mathrm{C}})/n_{\mathrm{C}}}$ is solved for p_{C} yielding solutions l_{C} and u_{C}. The lower and upper bounds of the interval are then computed from the solutions previously obtained, although the discussion in this paper only uses the lower bound in (4)

$$\begin{aligned}&\{\hat{p}_{\mathrm{T}} - \hat{p}_{\mathrm{C}}\} - \sqrt{(\hat{p}_{\mathrm{T}} - l_{\mathrm{T}})^2 + (u_{\mathrm{C}} - \hat{p}_{\mathrm{C}})^2}\\ &\{\hat{p}_{\mathrm{T}} - \hat{p}_{\mathrm{C}}\} + \sqrt{(u_{\mathrm{T}} - \hat{p}_{\mathrm{T}})^2 + (\hat{p}_{\mathrm{C}} - l_{\mathrm{C}})^2} \qquad (4)\end{aligned}$$

For the Newcombe hybrid score interval, the lower bound can also be written as $\{\hat{p}_{\mathrm{T}} - \hat{p}_{\mathrm{C}}\} - z_{1-\alpha} \times \sqrt{l_{\mathrm{T}}(1-l_{\mathrm{T}})/n_{\mathrm{T}} + u_{\mathrm{C}}(1-u_{\mathrm{C}})/n_{\mathrm{C}}}$. Newcombe [6] recommends this method over the Wald method because of its performance with respect to coverage in the setting which involves both upper and lower limits. In addition, Agresti and Caffo [5] suggest that the method in Newcombe [6] is an appropriate method with better coverage properties than their method except when proportions are close to 0 or 1, although it has the limitation of being more complicated to implement.

In addition, Newcombe [6] provides a continuity corrected version of this interval where one solves $|p_{\mathrm{T}} - \hat{p}_{\mathrm{T}}| - 1/2n_{\mathrm{T}} = z_{1-\alpha}\sqrt{p_{\mathrm{T}}(1-p_{\mathrm{T}})/n_{\mathrm{T}}}$ and $|p_{\mathrm{C}} - \hat{p}_{\mathrm{C}}| - 1/2n_{\mathrm{C}} = z_{1-\alpha}\sqrt{p_{\mathrm{C}}(1-p_{\mathrm{C}})/n_{\mathrm{C}}}$ as in the Newcombe method and substitutes the appropriate solutions in the interval in (4).

3. SIMULATIONS

Simulations were used to study the properties of the methods for the statistical test based on the lower confidence bound to demonstrate non-

DOI: 10.1002/pst

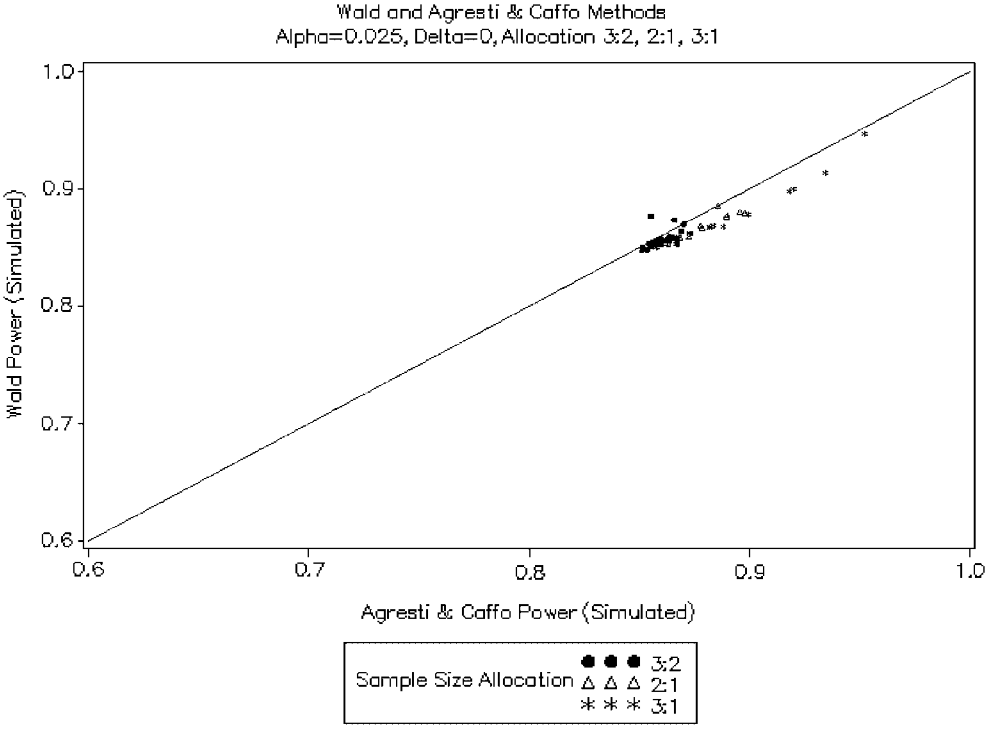


Figure 5. Comparison of simulated power.

inferiority. The assessment included proportions in the control group of 0.60–0.95 with non-inferiority margins of −0.05, −0.075, and −0.10. The sample sizes used were based on providing approximately 0.85 power to contradict the hypothesis H_0 for the non-inferiority margin given the specified active control proportion and no difference between treatments. In addition, the simulation study also focused on various sample size allocations for test versus active control groups including 1:2, 1:1, 3:2, 2:1, and 3:1. Due to the one sided nature of the hypothesis pertaining to non-inferiority, the alpha level of interest was set at one sided 0.025, although similar results were seen for $\alpha = 0.005$ and 0.05.

For each of the 100 000 replications performed, a sample from the specified binomial distribution was drawn separately for the test group and for the control group. All of the lower confidence limits were computed for the same replication, and a conclusion of non-inferiority or not was determined for the applicable one-sided test according to whether or not the one-sided lower confidence limit exceeded the specified non-inferiority margin. The average of these indicator variables for demonstration of non-inferiority produced a simulated power for the methods when the true difference in proportions exceeded the non-inferiority margin as under the alternative hypothesis and a type I error rate when the difference in true proportions was equal to (or poorer than) the non-inferiority margin as under the null hypothesis.

When the number of events in either group was zero or if a method failed to produce a logical lower confidence limit (because of intermediate computations pertaining to estimates of π_T or π_C being outside of the (0,1) range), then the Agresti and Caffo method was used in place of the Wald, Farrington–Manning, and Newcombe hybrid score methods. In practice, an alternative (exact)

DOI: 10.1002/pst

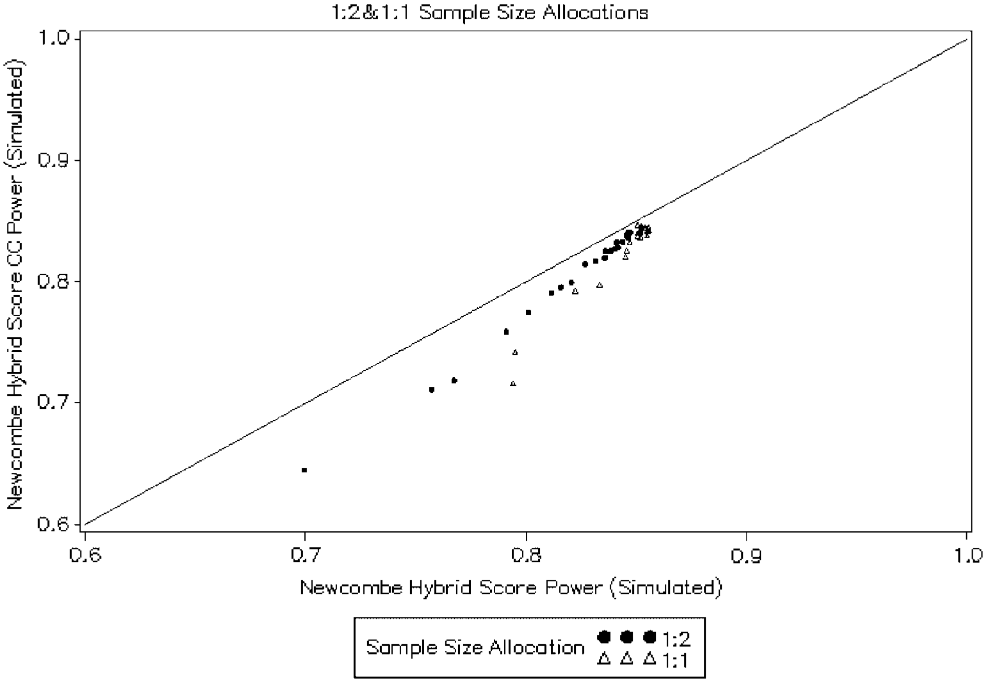


Figure 6. Summary of simulated power.

method might be used for small event rates, but the performance of the methods in this paper with this minimal modification is adequate for this discussion.

The performance of the methods will first be discussed with respect to maintaining a nominal type I error rate as summarized in Figures 1–3. Appropriate type I error will be those values included within $[0.9\alpha, 1.1\alpha] = [0.0225, 0.0275]$ as suggested by Roebruck and Kühn [3] and denoted with horizontal lines on the figures. The type I error for the methods is summarized in Figure 1 by sample size allocation, in Figure 2 by non-inferiority margins, and in Figure 3 by the event rate in the control group. The performance of the methods depends strongly on the sample size allocation and the event rate in the control group. Results appear to be similar across choices of non-inferiority margin as seen in Figure 2, with summaries limited to allocations in which methods have close to nominal type I error rates. In addition, results are similar for choice of alpha level although these are not provided.

The Wald and Agresti and Caffo methods tend to produce type I errors higher than the nominal level for the 1:2 allocation, with the Wald method becoming increasingly more conservative (in the sense of type I error being less than or equal to the nominal level) as relatively more sample size is placed in the test group. The Agresti and Caffo method tends to produce nearly nominal type I error rates for the 3:2, 2:1, and 3:1 allocations.

The opposite phenomenon occurs for the Farrington–Manning and Newcombe hybrid score methods with the type I error becoming progressively higher so as to exceed the nominal level as relatively more sample size is placed in the test group. These methods produce nominal type I errors for the 1:2 and 1:1 allocations and slightly higher than nominal type I errors for the 3:2, 2:1, and 3:1 allocations. The Newcombe hybrid score

DOI: 10.1002/pst

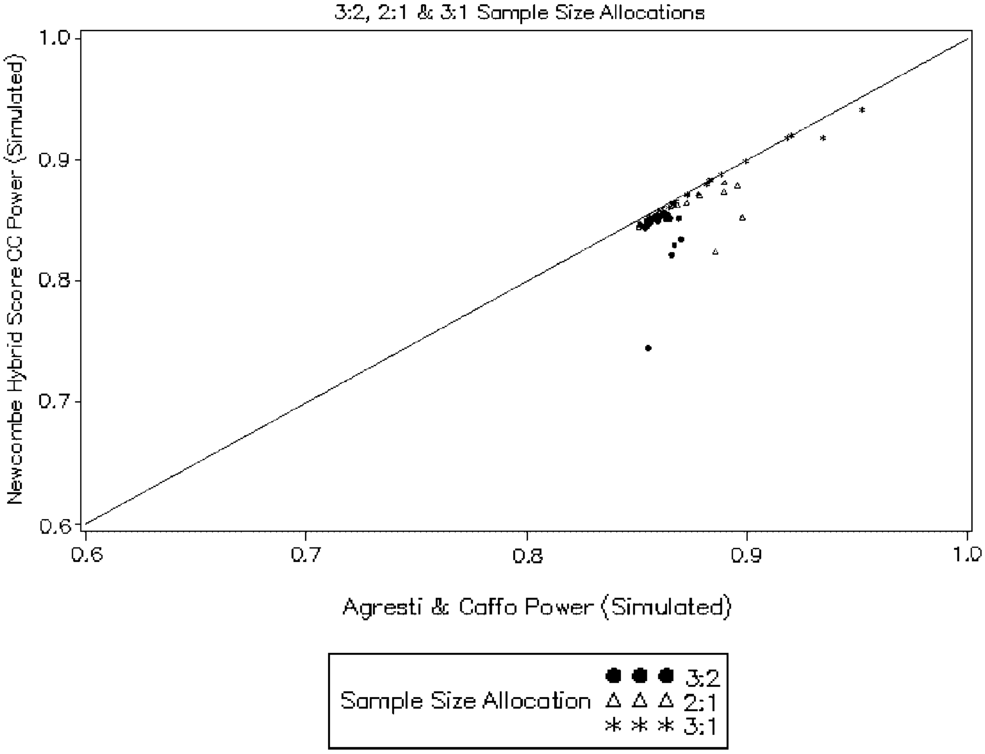


Figure 7. Summary of simulated power.

continuity corrected method produces lower than nominal type I errors for all allocation scenarios.

The type I errors of all of the methods seem to become more variable as the event rate in the control group increases toward 1 as seen in Figure 3, with summaries again limited to allocations in which methods have close to nominal type I error rates. The most plausible reason for this is that observed by Li and Chuang-Stein [4] in that the expected cell frequencies become smaller as π_C increases. Therefore, the asymptotic assumption may not be appropriate for larger π_C depending on the sample sizes in each treatment group.

The power of these methods is summarized in Figures 4–7 only for sample size allocations in which the methods appropriately control the type I error at the nominal level. As seen in Figure 4, the Farrington–Manning and Newcombe hybrid score methods yield fairly similar power levels for the 1:2 and 1:1 allocation settings, with the Newcombe hybrid score method yielding slightly higher power in the 1:1 setting. In the allocations with more sample size in the test group (3:2, 2:1, 3:1) seen in Figure 5, the Agresti and Caffo method generally yields a higher power than the Wald method especially as the sample size imbalance becomes greater.

The Newcombe hybrid score continuity corrected method has lower power than the Newcombe hybrid score method for the 1:2 and 1:1 allocations, and this power is much lower for the 1:1 setting as seen in Figure 6. The Newcombe hybrid score continuity corrected method also has lower power than the Agresti and Caffo method for the 3:2, 2:1, and 3:1 allocation settings seen in Figure 7, where the difference in power is more distinct for the 3:2 and 2:1 settings.

DOI: 10.1002/pst

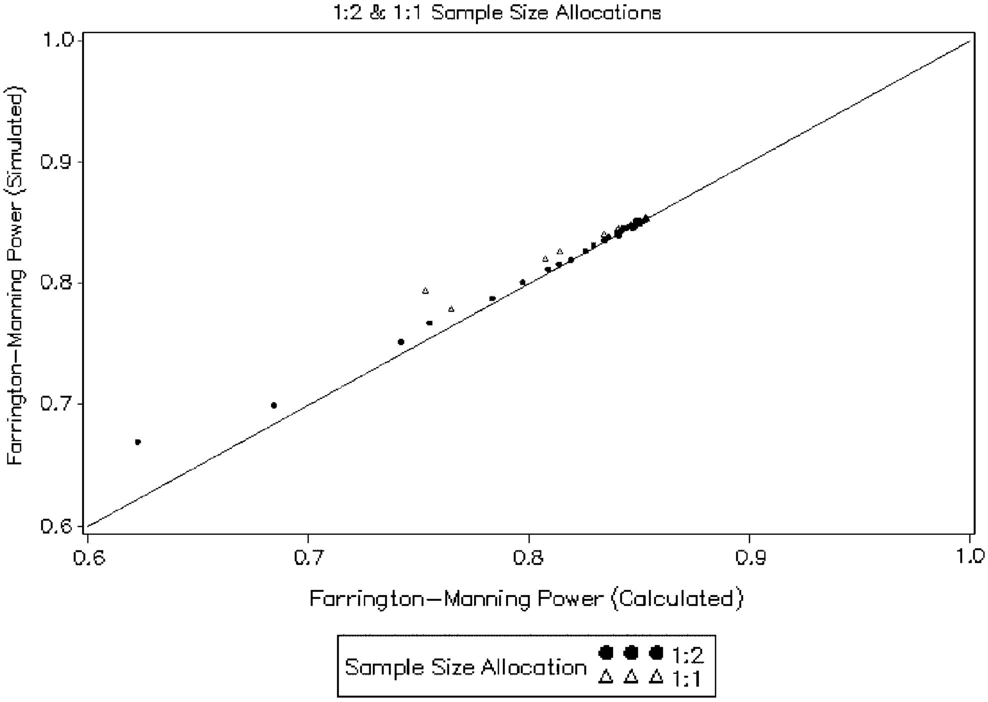


Figure 8. Summary of simulated versus calculated power.

4. SAMPLE SIZE CONSIDERATIONS

In addition to appropriate methods for analyses when the difference between proportions is the measure of interest for treatment comparisons in a one-sided non-inferiority setting, it is important to have corresponding sample size formulas in the planning stages of the trial. Sample size calculation based on the Wald method is a popular and straightforward way to plan for patient recruitment in non-inferiority trials for the difference in proportions. This sample size calculation (with respect to $\Delta_0 < 0$) is shown in (5) for the test group, with the sample size in the control group defined as $n_C = n_T/R$ where $R = n_T/n_C$.

$$n_T = \frac{\{z_{1-\alpha} + z_{1-\beta}\}^2\{\pi_T(1-\pi_T) + R\pi_C(1-\pi_C)\}}{\{\pi_T - \pi_C - \Delta_0\}^2} \tag{5}$$

This sample size formula, through algebraic manipulation, can be written to produce power for specified sample sizes in the test and control groups as in (6)

$$z_{1-\beta} = \frac{\sqrt{n_T}\{\pi_T - \pi_C - \Delta_0\}}{\sqrt{\pi_T(1-\pi_T) + R\pi_C(1-\pi_C)}} - z_{1-\alpha} \tag{6}$$

where power is obtained as the probability ($1-\beta$) from $z_{1-\beta}$ as the ($1-\beta$) quantile of the standard normal distribution.

Additionally, Farrington and Manning [7] provide the sample size formula in (7) that is analogous to their methods, with the appropriate proportions substituted for $\tilde{\pi}_T$ and $\tilde{\pi}_C$ obtained from solving the maximum likelihood equations under the null hypothesis. As a note, (7) reduces to (5) if $\tilde{\pi}_T = \pi_T$ and $\tilde{\pi}_C = \pi_C$

$$n_T = \frac{\left\{z_{1-\alpha}\sqrt{\tilde{\pi}_T(1-\tilde{\pi}_T) + R\tilde{\pi}_C(1-\tilde{\pi}_C)} + z_{1-\beta}\sqrt{\pi_T(1-\pi_T) + R\pi_C(1-\pi_C)}\right\}^2}{\{\pi_T - \pi_C - \Delta_0\}^2} \tag{7}$$

DOI: 10.1002/pst

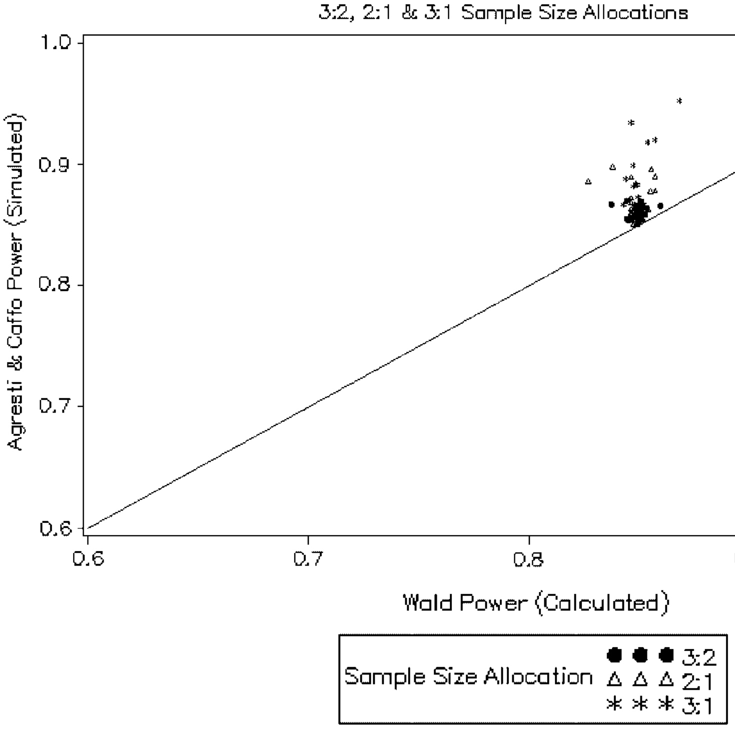


Figure 9. Summary of simulated versus calculated power.

The corresponding power calculation is seen in (8)

$$z_{1-\beta} = \frac{\left\{\sqrt{n_{\mathrm{T}}}(\pi_{\mathrm{T}} - \pi_{\mathrm{C}} - \Delta_0) - z_{1-\alpha}\sqrt{\tilde{\pi}_{\mathrm{T}}(1-\tilde{\pi}_{\mathrm{T}}) + R\tilde{\pi}_{\mathrm{C}}(1-\tilde{\pi}_{\mathrm{C}})}\right\}}{\sqrt{\pi_{\mathrm{T}}(1-\pi_{\mathrm{T}}) + R\pi_{\mathrm{C}}(1-\pi_{\mathrm{C}})}} \qquad (8)$$

The sample size formula from (7) analogous to the Farrington–Manning method is appropriate in trial design for the 1:2 and 1:1 allocation settings. As seen in Figure 8, the calculated power defined in (8) is consistently lower than the simulated power using the Farrington–Manning method.

For settings with more sample size in the test group as in the 3:2, 2:1, and 3:1 settings, the Wald sample size formula from (5) may be used in designing non-inferiority trials. Figure 9 also shows that the calculated power from (6) is consistently slightly smaller than the simulated power from the Agresti and Caffo method. Higher simulated power is beneficial in trial design as the calculated sample sizes will be slightly conservative.

The Farrington–Manning calculated power is compared to the Newcombe hybrid score continuity corrected simulated power in Figure 10, where the simulated power is consistently lower than the calculated power. The Farrington–Manning sample size formula should be used cautiously when implementing the Newcombe hybrid score continuity corrected method for analysis in these trials.

5. DISCUSSION

There are many choices of methods for determining a one-sided confidence interval for the difference in proportions in the setting of hypothesis

DOI: 10.1002/pst

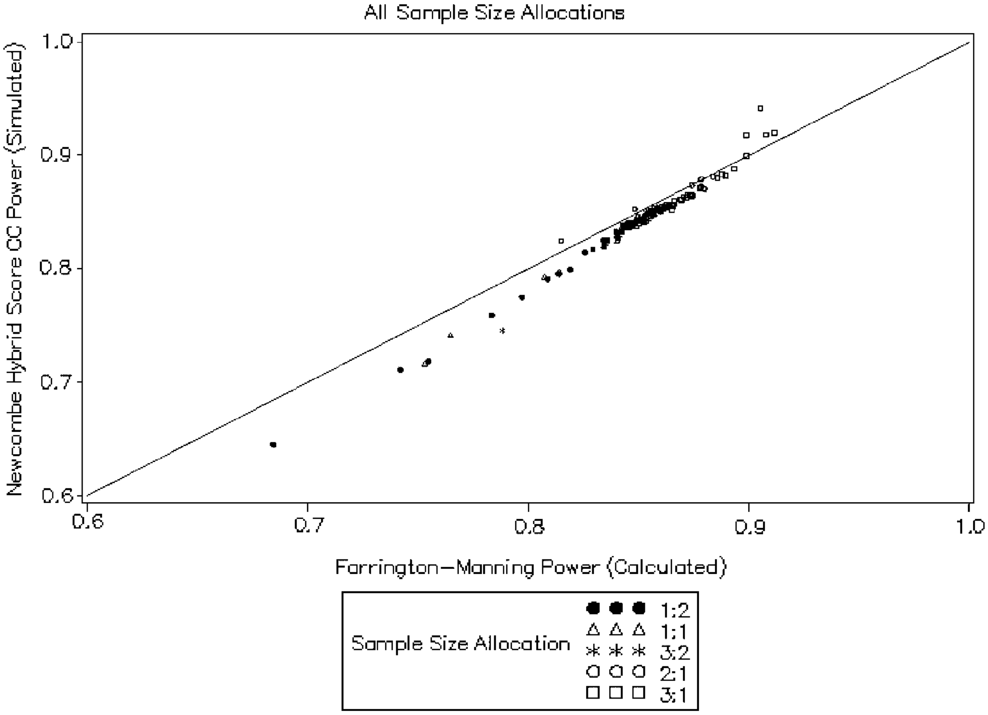


Figure 10. Summary of simulated versus calculated power.

testing for non-inferiority. The Wald method is most easily applied and it performs appropriately in terms of control of type I error at the nominal level when relatively more sample size is placed in the test group. However, the Agresti and Caffo method yields higher power in these situations and is also easily applied. The Wald sample size formula may be used in these settings to determine sample size in the design stages of a non-inferiority clinical trial. The Farrington–Manning method seems to perform most consistently in the 1:2 and 1:1 allocations and has a corresponding sample size formula which is always conservative. These results suggest that choice of an appropriate method for assessing non-inferiority of a risk difference is dependent on the sample size allocation.

While it may be desirable to have a general method for use in all scenarios, it is important to ensure both appropriate control of type I error while providing enough power to assess the null hypothesis of interest. The Newcombe hybrid score continuity corrected method provides appropriate type I error control in all allocation settings, but this overall control results in a loss of power.

Similar results are expected for values of the parameters within the range studied, but simulations can be performed to ensure this consistency. Additionally, results may be different when assessment could additionally include the upper confidence limit for hypothesis testing pertaining to equivalence through a two-sided confidence interval. This topic for which the null hypothesis is H_0: $|\pi_T - \pi_C| \geqslant \Delta_0$ is beyond the scope of this paper, although one can recognize that equivalence is comparable to each treatment being non-inferior to the other and so can be addressed by the two corresponding one-sided tests (and confidence intervals). Also, the scope of this paper is restricted to the type I error and power for the use of the

DOI: 10.1002/pst

one-sided confidence interval as a test for the one-sided hypothesis pertaining to non-inferiority. The coverage properties of such confidence intervals are beyond the scope of this paper.

REFERENCES

1. ICH Expert Working Group. ICH harmonized tripartite guideline, statistical principles for clinical trials. *Statistics in Medicine* 1999; **18**:1905–1942.
2. Efficacy Working Party. European Medicines Agency Committee for Medicinal Products for Human Use (CHMP) guideline on the choice of the non-inferiority margin. *Statistics in Medicine* 2006; **25**:1628–1638.
3. Roebruck P, Kühn A. Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. *Statistics in Medicine* 1995; **14**: 1583–1594.
4. Li Z, Chuang-Stein C. A note on comparing two binomial proportions in confirmatory noninferiority trials. *Drug Information Journal* 2006; **40**:203–208.
5. Agresti A, Caffo B. Simple and effective confidence intervals for proportions and difference of proportions results from adding two successes and two failures. *The American Statistician* 2000; **54**(4): 280–288.
6. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 1998; **17**: 873–890.
7. 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. *Statistics in Medicine* 1990; **9**:1447–1454.
8. SAS®, Version 8.02. SAS Institute Inc.: Cary, NC, 1999.

DOI: 10.1002/pst