Journal of Biopharmaceutical Statistics, 15: 85–107, 2005

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1081/BIP-200040844

Taylor & Francis
Taylor & Francis Group

REVIEW AND EVALUATION OF METHODS FOR COMPUTING CONFIDENCE INTERVALS FOR THE RATIO OF TWO PROPORTIONS AND CONSIDERATIONS FOR NON-INFERIORITY CLINICAL TRIALS

Rebekkah S. Dann and Gary G. Koch
Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, USA

This article reviews several methods for forming confidence intervals for a risk ratio of two independent binomial proportions (which are both less than 0.50) and evaluates their statistical performance. These methods include use of a Taylor Series expansion to estimate variance, solutions to a quadratic equation, and maximum likelihood methods. In addition, for improvement of the properties of the methods based on large sample approximations, situations where either binomial count was less than or equal to 3 were managed conservatively by having an exact confidence interval for the odds ratio become the confidence interval for its risk ratio counterpart. Methods were initially evaluated by computing confidence limits for certain cases. Second, simulations were used to identify the better methods for controlling the Type I error rate while maintaining power. Last, relationships between methods were evaluated by calculating the percent of disagreement in the decision made regarding noninferiority. Methods in the group using a Taylor Series expansion in variance estimation perform similarly to the Pearson method preferred in the literature. In addition, the group of methods using a Taylor Series expansion are most easily computed. Applications of these findings are discussed for ratios that arise in randomized clinical trials that are conducted to show noninferiority of a new medical product to a reference control. Consideration is given as well to sample size calculations for noninferiority clinical trials.

Key Words: Risk ratio; Noninferiority; Approximate confidence interval; Sample size; Type I error; Power simulations.

Mathematics Subject Classification: Primary 62P10; Secondary 62F12.

INTRODUCTION

Ratios of proportions are often called risk ratios in a clinical trial setting. These ratios are used to compare two independent groups, usually on two different treatments. A noninferiority clinical trial can compare an active-control group to a group taking a new treatment for an efficacy outcome (or a placebo group to a group taking a new treatment for a safety outcome). The goal is to show that the new treatment is not unacceptably worse than the active control (or placebo) treatment

Received December 2003; Accepted August 2004
Address correspondence to Rebekkah S. Dann, CB # 7420, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA; Fax: (919) 966-7285; E-mail: rdann@bios.unc.edu

(Hung et al., 2001). The new treatment may have other beneficial aspects such as a reduction in severity of side effects, easier use, or lower cost.

Assessing noninferiority is often done through a confidence interval for the risk ratio of the two groups (Hung et al., 2001), particularly if control failure rates are small (e.g., ≤ 0.20) or control success rates are large (e.g., ≥ 0.80). If failure rates are very small (e.g., <0.05), then the odds ratio can be conservatively used to approximate the risk ratio (when defined to have the larger expected rate in the numerator and the smaller expected rate in the denominator). For situations where failure rates are larger (e.g., >0.20), then the difference in rates is typically emphasized (Blackwelder, 1982; Farrington and Manning, 1990; nQuery Advisor Version 5.0 User's Guide, 2002). In some cases, if the new treatment group has a risk that is not more than twice that of the control group for a failure outcome through an upper confidence limit of 2 or less, then the treatments are judged noninferior (e.g., when the control failure rate is 0.10, a risk ratio of 2 or less indicates a test failure rate less than 0.20 and an extent of inferiority less than 0.10). Accordingly, a corresponding test of noninferiority has the null hypothesis as H_O: $\theta = \pi_T/\pi_C \geq 2$, and the alternative hypothesis as H_A: $\theta = \pi_T/\pi_C < 2$ where $\theta = \pi_T/\pi_C$ is the population risk ratio for the treatment group versus the control group.

There are many methods in existence for computing a confidence interval for a risk ratio. Several of the methods for forming confidence intervals for ratios of two independent binomial proportions are reviewed and evaluated for their statistical performance. These methods include use of a Taylor Series expansion to estimate variance, solutions to a quadratic equation, and maximum likelihood methods. Methods were initially compared by computing confidence limits for certain cases. Second, simulations were used to identify the better methods for controlling the Type I error rate while maintaining power. Finaly, relationships between methods were evaluated by calculating the percent of disagreement in the decision made regarding noninferiority. Applications of these findings include sample size calculations, which arise in randomized clinical trials conducted to show noninferiority.

METHODS

Taylor Series Expansion Methods

The literature contains many methods for forming confidence intervals for risk ratios. The first group of these uses a variance formed through a Taylor Series expansion. The following method, hereafter called the Taylor Series method, is the simplest in this group discussed by Katz et al. (1978) and used by SAS in the FREQ procedure (1990) and by EquivTest (2000) to form a $100(1-2\alpha)\%$ confidence interval for a risk ratio:

$$\exp\left\{\log_e\left(\frac{y_T}{y_C}\right) \pm z_\alpha\left[\frac{1}{y_T}+\frac{1}{y_C}-\frac{2}{n}\right]^{1/2}\right\}$$

where y_T is the number of events in the treatment group, y_C is the number of events in the control group, and z_α is the $100(1-\alpha)$ percentile for a standard normal

distribution. For simplicity, it will be assumed that the sample sizes in both the treatment and control groups are equal and denoted by n.

In 1988, Gart and Nam (1988) revised this original method so that the confidence interval would be defined if y_T or y_C were equal to zero. The following is this modified confidence interval used by StatXact (1990) for risk ratios:

$$\exp\left\{\log_e\left(\frac{y_T+0.5}{y_C+0.5}\right)+z_\alpha\left[\frac{1}{y_T+0.5}+\frac{1}{y_C+0.5}-\frac{2}{n+0.5}\right]^{1/2}\right\}.$$

This Modified Taylor Series method adds a half to the event count for each group as well as the total sample size for each group.

The last method in this group of Taylor Series expansion methods is adapted from a confidence interval for a single binomial proportion proposed by Agresti and Coull (1998). For a confidence interval for a single binomial proportion, Agresti and Coull suggested adding half the squared z-value to each outcome for each group to produce a more conservative interval. This strategy was adapted for a test of non-inferiority where the null hypothesis is not one of equality. The additional counts must be distributed to each group according to the null hypothesis. The treatment group would receive twice as many of the additional counts as the control group according to a null hypothesis of $\theta=2$. This Adapted Agresti confidence interval for a ratio of two binomial proportions is as follows:

$$\exp\left\{\log_e\left(\frac{y_T+2.67}{y_C+1.33}\right)+z_\alpha\left[\frac{1}{y_T+2.67}+\frac{1}{y_C+1.33}-\frac{2}{n+4}\right]^{1/2}\right\}$$

for a one-sided alpha level of 0.025 and

$$\exp\left\{\log_e\left(\frac{y_T+2}{y_C+1}\right)+z_\alpha\left[\frac{1}{y_T+2}+\frac{1}{y_C+1}-\frac{2}{n+3}\right]^{1/2}\right\}$$

for a one-sided alpha level of 0.05. For $\alpha=0.025$, the four additional events result from the square of $(z_{0.025})^2=(1.96)^2\approx 4$. These four events are added with 2.67 additional events in the treatment group and 1.33 additional events in the control group. Similarly, for an alpha level of 0.05, a total of $(z_{0.05})^2=(1.645)^2\approx 3$ events are added, with two events added to the treatment group and one event added to the control group.

Solution to Quadratic Equation Methods

The next group of methods is slightly more complicated because the confidence limits are the solutions to a quadratic equation. After algebraic manipulations, a quadratic form of the equations provided below are then solved for θ. The upper and lower confidence limits are the smaller and larger of the two solutions, respectively. However, these methods may produce complex-valued results (when square roots of negative numbers are involved).

Fieller (1944) first presented the most basic of these methods in 1944 as seen below, hereafter called the Quadratic method:

$$\frac{(y_T - \theta y_C)^2 (n-1)}{\{y_T(n-y_T) + \theta^2 y_C(n-y_C)\}} = z_\alpha^2.$$

The second of this group of methods was proposed by Bailey (1987) in 1987 which is a modification of the Quadratic method to produce limits with more desirable properties as will be discussed in more detail in the literature review section:

$$\frac{(y_T{}^{1/3} - \theta^{1/3} y_C{}^{1/3})^2 n}{(1/9)\left\{y_T{}^{-1/3}(n-y_T) + \theta^{2/3} y_C{}^{-1/3}(n-y_C)\right\}} = z_\alpha^2.$$

The last of this group of methods was proposed by Farrington and Manning (1990) with three possible variations on the following equation:

$$\frac{(\hat{\pi}_T - \theta\hat{\pi}_C)^2 n}{\{\tilde{\pi}_T(n-\tilde{\pi}_T) + \theta^2\tilde{\pi}_C(n-\tilde{\pi}_C)\}} = z_\alpha^2$$

where $\hat{\pi}_C = y_c/n$ and $\hat{\pi}_T = y_T/n$. Each variation suggests computing $\tilde{\pi}_C$ and $\tilde{\pi}_T$ in a different manner. The first of these, F–M 1, uses the observed values and sets $\tilde{\pi}_C = \hat{\pi}_C$ and $\tilde{\pi}_T = \hat{\pi}_T$. The second variation, F–M 2, uses fixed marginal totals to compute $\tilde{\pi}_C = (\theta/(\theta+1))(\hat{\pi}_C + \hat{\pi}_T)$ and $\tilde{\pi}_T = (1/(\theta+1))(\hat{\pi}_C + \hat{\pi}_T)$. The third variation, F–M 3, uses maximum likelihood estimation under the null hypothesis to obtain $\tilde{\pi}_C$ and $\tilde{\pi}_T$ with details found in Farrington and Manning (1990).

Maximum Likelihood Methods

The third group of confidence interval methods includes those that use maximum likelihood estimators for the proportion of events in the treatment and control groups based on the joint distribution of the events as the product of two independent binomial distributions for the treatment and control groups. The first of these methods calculates a deviance statistic as follows

$$\text{Deviance} = 2^*[\log \mathrm{L}(\hat{\pi}_T, \hat{\pi}_C) - \log \mathrm{L}(2\hat{\pi}^*, \hat{\pi}^*)]$$

where $\hat{\pi}_T$ and $\hat{\pi}_C$ are the maximum likelihood estimators of π_T and π_C under the alternative hypothesis, and $2\hat{\pi}^*$ and $\hat{\pi}^*$ are the corresponding maximum likelihood estimators under the null hypothesis $\theta = 2$.

The second of these maximum likelihood methods is based on a Pearson statistic in the form of [(observed – expected)^2/expected] as follows:

$$\frac{\{y_T - 2n\hat{\pi}^*\}^2}{2n\hat{\pi}^*} + \frac{\{(n-y_T) - n(1-2\hat{\pi}^*)\}^2}{n(1-2\hat{\pi}^*)} + \frac{\{y_C - n\hat{\pi}^*\}^2}{n\hat{\pi}^*} + \frac{\{(n-y_C) - n(1-\hat{\pi}^*)\}^2}{n(1-\hat{\pi}^*)}$$

using $2\hat{\pi}^*$ and $\hat{\pi}^*$, the maximum likelihood estimators of π_T and π_C under the null hypothesis $\theta = 2$. Koopman (1984) proposed this method in 1984, and StatXact is a software package that provides these confidence intervals.

The Deviance and Pearson methods produce test statistics for which p-values can be obtained by using the chi-square distribution under one degree of freedom. The appropriate confidence limits can be found through an iterative process. The hypothesized ratio θ of π_T to π_C is modified until the desired p-value (e.g., 0.05 or 0.025) is obtained. This process identifies the largest θ_0 that would not be rejected as H_O: $\theta \geq \theta_0$. The ratio that produces the desired p-value is then the upper confidence limit. This iterative process requires changing the maximum likelihood estimator pertaining to the null hypothesis as θ_0 changes. This group of methods is more complicated than the others due to the iterative nature of finding the confidence intervals as all hypotheses not rejected, thus requiring intensive computer resources.

A summary of available software resources for the computation of the methods described can be found in Table 1.

REVIEW OF LITERATURE

Different combinations of the methods described above have been compared in the literature. In 1978, Katz et al. (1978) compared the Taylor Series method and the Quadratic method by using simulations and calculating coverage probabilities. Katz et al. (1978) suggested that the Quadratic method could be erratic and may not produce confidence limits at all; the Taylor Series method was recommended for use instead of the Quadratic method.

Again in 1984, Koopman (1984) used simulations and coverage probabilities to compare the Taylor Series method and the Pearson maximum likelihood method. Findings suggested that the Pearson method maintains a coverage probability closer to the $(1 - 2\alpha)$ level, and in addition, the one-sided probabilities of exceeding the upper limit or being lower than the lower limit are much closer to α. Therefore, Koopman recommended use of the Pearson method.

Table 1 Software resources for methods

Method	Software resources
Taylor Series	SAS using PROC FREQ (1990) EquivTest (2000)
Taylor Series Adjusted Alpha	SAS using PROC FREQ (1990)[a]
Modified Taylor Series	SAS using PROC FREQ (1990)[b] StatXact (1990)
Adapted Agresti	SAS using PROC FREQ (1990)[b]
Quadratic	No resources available
Bailey	No resources available
Farrington–Manning 1	No resources available
Farrington–Manning 2	No resources available
Farrington–Manning 3	No resources available
Deviance	SAS using PROC GENMOD (1990)[c]
Pearson	SAS using PROC GENMOD (1990)[c] StatXact (1990)

[a]The alpha level can be modified to produce this interval.
[b]Event counts can be modified to produce this interval.
[c]Additional programming is required to use this computer resource.

In 1987, Bailey (1987) extended the Quadratic method to produce Bailey's method, which should reduce the skewness of the confidence interval as well as maintain the nominal coverage probability better than the Quadratic method. This new method is also compared with the Taylor Series method and the Pearson method. Bailey concluded that his method results in confidence limits that are closer to the nominal level than the Taylor Series method. In addition, Bailey's method more often maintains the nominal coverage probability better than the Pearson method.

Gart and Nam (1988) produced a comprehensive comparison of the methods presented previous to 1988. They indicated that the Quadratic method and Bailey's method tend to produce confidence limits that are either above or below the nominal coverage probability, whereas the Modified Taylor Series method and the Pearson method achieve coverage probabilities close to the nominal level, with the Pearson method slightly better.

In 1990, Farrington and Manning (1990) presented results on the three variations of quadratic methods for producing confidence limits for risk ratios. Their recommendation was the third of these methods based on maximum likelihood estimation for the proportions, F–M 3.

In summary, the Pearson method is more often preferred in the literature. Use of the Taylor Series method and the Modified Taylor Series method was mentioned as potentially beneficial. Conflicting opinions over the group of quadratic methods exist with some authors suggesting they are best and others avoiding their use.

CONFIDENCE LIMIT COMPARISONS

An initial comparison of the methods includes computing the upper confidence limits for selected cases. At one-sided alpha levels of 0.025 and 0.05 seen in Table 2, the upper confidence limits are presented for each of the 11 methods. The methods are grouped by the three method types: the Taylor Series variance expansion methods, the quadratic methods, and the maximum likelihood methods.

Within the Taylor Series variance expansion methods, the Adapted Agresti method tends to produce slightly higher upper confidence limits than the other three methods. As mentioned above, the Quadratic method and Farrington–Manning method 1 produce very similar upper confidence limits due to their similiarity in computation. Farrington–Manning methods 2 and 3 also produce similar upper confidence limits for the selected cases presented. The Deviance and Pearson methods produce similar confidence limits.

SIMULATIONS

Data were generated from known distributions to compare the behavior of the confidence intervals for the methods. The values of four main parameters were varied. The number of trials, n, was set at 100, 140, and 200 per group. The probability of occurrence for the control group, π_C, took on the following range of values: 0.10, 0.15, 0.20, and 0.25. The actual ratio of the two proportions, θ, was set at 0.667, 0.800, 1.000, 1.250, 1.500, and 2.000. The value of $\pi_T = \theta\pi_C$ also varied but was completely specified after knowing π_C and θ. Two one-sided alpha levels were used: 0.025 and 0.050.

Table 2 Upper confidence limits for selected cases

				Taylor Series variance expansion methods				Quadratic methods					Maximum likelihood methods	
n	α	y_C	y_T	Taylor Series	Taylor Series Adjusted Alpha	Modified Taylor Series	Adapted Agresti	Quadratic	Bailey	F–M 1	F–M 2	F–M 3	Deviance	Pearson
100	0.025	15	10	1.41	1.44	1.41	1.53	1.45	1.41	1.45	1.36	1.36	1.40	1.39
		15	15	1.93	1.96	1.91	2.00	2.08	1.96	2.07	1.91	1.90	1.95	1.91
		15	18	2.25	2.28	2.21	2.28	2.46	2.29	2.45	2.28	2.26	2.29	2.23
		20	15	1.38	1.40	1.38	1.46	1.41	1.38	1.40	1.34	1.34	1.38	1.37
		20	20	1.74	1.76	1.73	1.80	1.82	1.75	1.81	1.72	1.71	1.76	1.74
		20	23	1.96	1.98	1.94	2.00	2.06	1.98	2.05	1.95	1.94	1.98	1.95
	0.05	15	10	1.25	1.26	1.25	1.35	1.26	1.25	1.26	1.22	1.22	1.23	1.23
		15	15	1.74	1.75	1.72	1.80	1.82	1.75	1.81	1.71	1.71	1.75	1.73
		15	18	2.03	2.05	2.00	2.06	2.14	2.05	2.14	2.03	2.02	2.05	2.02
		20	15	1.25	1.26	1.25	1.32	1.26	1.25	1.26	1.22	1.22	1.25	1.25
		20	20	1.59	1.60	1.58	1.64	1.63	1.60	1.63	1.57	1.56	1.60	1.59
		20	23	1.80	1.81	1.78	1.83	1.86	1.81	1.85	1.78	1.77	1.81	1.80

For each combination of n, π_C, and θ, 100,000 simulations were generated by using a random sample from the two binomial distributions of $y_T \sim \text{bin}(n, \pi_T = \theta\pi_C)$ and $y_C \sim \text{bin}(n, \pi_C)$. For each combination of y_T and y_C, upper confidence limits for all methods were found. If y_T or y_C were less than or equal to 3, then the exact confidence limit for the odds ratio was the default for the following methods because of the inability of these methods to yield a valid result in this situation: Taylor Series, Taylor Series Adjusted Alpha, Quadratic, Farrington–Manning 1, Farrington–Manning 2, Farrington–Manning 3, Deviance, and Pearson methods. This modification using the odds ratio is conservative because it uses exact methodology and because the odds ratio exceeds the risk ratio when both exceed 1. As a note, if $y_C = 0$, then the upper confidence limit for the odds ratio is essentially infinite, and so it was set to 100, and the null hypothesis of inferiority was not rejected. This modification, where the upper confidence limit was set to 100, is also necessary in cases where the group of quadratic methods lead to square roots of negative numbers (i.e., have complex solutions) or where the Deviance or Pearson methods fail to produce interpretable results because of computational singularities, although in the 100,000 simulations none of these latter difficulties occurred for y_T or y_C greater than 3. No modifications were necessary for the Modified Taylor Series or the Adapted Agresti methods.

For each method, an indicator variable was created for each simulation that takes the value of 1 if the upper confidence limit produced was less than 2 and 0 otherwise. This indicator was then averaged across all 100,000 simulations to produce a probability. For $\theta < 2$, this probability is the power for the test of noninferiority and can be written in the following manner: power = pr(reject H_O: $\theta = \pi_T/\pi_C \geq 2 \mid H_A$: $\theta < 2$ true). For $\theta = 2$, this probability is the Type I error rate for the test of noninferiority and can be written in the following manner: α = Type I error = pr(reject H_O: $\theta = \pi_T/\pi_C \geq 2 \mid H_O$: $\theta \geq 2$ true).

A summary of the power of the methods generated from the 100,000 simulations can be seen in Table 3 ($n = 100$), ($n = 140$), ($n = 200$) for $\alpha = 0.025$. Farrington–Manning method 1 is dropped from summaries due to its similarities to the Quadratic method. In addition, Farrington–Manning method 3 is also removed due to its similarity to Farrington–Manning method 2 and its inflation of the Type I error rate over that of Farrington–Manning method 2.

The Taylor Series Adjusted Alpha method was added to correct inflation of Type I error by the Taylor Series method seen in initial simulations. This method is the Taylor Series method with an alpha level that is 0.0025 less than the alpha level specified. For example, this method would use an alpha level of $0.025 - 0.0025 = 0.0225$ when $\alpha = 0.025$ was specified. The choice of 0.0025 was motivated by findings from the simulations for the specific scenarios presented where this modification was needed to offset the small inflation in Type I error of the Taylor Series method. The same adjustments were used for this method as with the Taylor Series method when y_T or y_C was less than or equal to 3 and the alpha level was maintained at 0.025. This 0.0025 adjustment of the alpha level is dependent on the application at hand, and simulations can be used to determine the appropriate adjustment for any scenario. This adjustment to the alpha level is a way to address studies with finite samples rather than infinite (or very large) samples by increasing the z-criterion for significance slightly (i.e., for $\alpha = 0.025$ the z-criterion would increase from 1.96 to 2.00). Sample SAS code is attached in the Appendix to produce these simulations.

Table 3 Summary of power from simulations $\alpha = 0.025$

			Taylor Series variance expansion methods				Quadratic methods			Maximum likelihood methods	
n	π_C	θ	Taylor Series	Taylor Series Adjusted Alpha	Modified Taylor Series	Adapted Agresti	Quadratic	Bailey	F–M 2	Deviance	Pearson
100	0.1	0.667	0.628	0.624	0.645	0.563	0.543	0.624	0.650	0.627	0.666
		0.8	0.517	0.510	0.536	0.455	0.430	0.510	0.541	0.517	0.550
		1	0.367	0.353	0.379	0.308	0.284	0.353	0.389	0.364	0.386
		1.25	0.223	0.206	0.227	0.177	0.158	0.206	0.246	0.217	0.229
		1.5	0.120	0.107	0.121	0.091	0.078	0.107	0.149	0.113	0.121
	0.15	0.667	0.825	0.808	0.827	0.780	0.757	0.808	0.827	0.818	0.828
		0.8	0.718	0.693	0.721	0.664	0.634	0.693	0.719	0.703	0.720
		1	0.543	0.516	0.549	0.490	0.457	0.516	0.544	0.522	0.544
		1.25	0.323	0.306	0.337	0.288	0.258	0.306	0.324	0.307	0.327
		1.5	0.163	0.155	0.175	0.146	0.125	0.155	0.164	0.155	0.168
	0.2	0.667	0.925	0.917	0.929	0.907	0.892	0.917	0.925	0.918	0.926
		0.8	0.849	0.838	0.858	0.826	0.803	0.838	0.849	0.839	0.852
		1	0.676	0.665	0.694	0.654	0.617	0.666	0.676	0.665	0.684
		1.25	0.425	0.411	0.443	0.406	0.362	0.414	0.424	0.411	0.432
		1.5	0.211	0.197	0.222	0.197	0.165	0.203	0.208	0.197	0.213
	0.25	0.667	0.973	0.971	0.976	0.969	0.961	0.971	0.973	0.971	0.975
		0.8	0.928	0.924	0.934	0.921	0.903	0.924	0.928	0.924	0.931
		1	0.793	0.781	0.804	0.780	0.745	0.786	0.791	0.781	0.796
		1.25	0.536	0.514	0.548	0.514	0.470	0.524	0.527	0.514	0.536
		1.5	0.276	0.257	0.285	0.257	0.223	0.265	0.265	0.256	0.273

(continued)

Table 3. Continued.

			Taylor Series variance expansion methods				Quadratic methods			Maximum likelihood methods	
n	π_C	θ	Taylor Series	Taylor Series Adjusted Alpha	Modified Taylor Series	Adapted Agresti	Quadratic	Bailey	F–M 2	Deviance	Pearson
140	0.1	0.667	0.772	0.763	0.792	0.734	0.704	0.772	0.792	0.772	0.792
		0.8	0.658	0.653	0.685	0.621	0.583	0.658	0.683	0.658	0.685
		1	0.480	0.478	0.510	0.446	0.405	0.480	0.504	0.480	0.510
		1.25	0.284	0.282	0.309	0.259	0.227	0.283	0.298	0.283	0.309
		1.5	0.146	0.143	0.160	0.129	0.111	0.143	0.152	0.143	0.160
	0.15	0.667	0.923	0.923	0.933	0.912	0.895	0.923	0.927	0.923	0.933
		0.8	0.847	0.843	0.860	0.828	0.805	0.843	0.849	0.843	0.860
		1	0.676	0.664	0.689	0.649	0.614	0.664	0.677	0.664	0.689
		1.25	0.418	0.401	0.429	0.395	0.355	0.401	0.418	0.401	0.428
		1.5	0.208	0.196	0.218	0.195	0.166	0.196	0.208	0.196	0.214
	0.2	0.667	0.981	0.979	0.982	0.977	0.971	0.979	0.981	0.979	0.982
		0.8	0.940	0.935	0.944	0.933	0.918	0.935	0.940	0.935	0.943
		1	0.812	0.801	0.821	0.800	0.768	0.801	0.811	0.801	0.817
		1.25	0.551	0.533	0.563	0.533	0.489	0.533	0.545	0.533	0.553
		1.5	0.280	0.266	0.292	0.266	0.230	0.266	0.275	0.266	0.280
	0.25	0.667	0.995	0.995	0.996	0.995	0.993	0.995	0.995	0.995	0.996
		0.8	0.981	0.979	0.982	0.979	0.972	0.979	0.980	0.979	0.981
		1	0.906	0.898	0.911	0.898	0.877	0.898	0.903	0.898	0.906
		1.25	0.670	0.658	0.684	0.658	0.618	0.658	0.666	0.658	0.670
		1.5	0.351	0.337	0.366	0.338	0.305	0.338	0.345	0.337	0.350

200	0.1	0.667	0.905	0.892	0.905	0.879	0.863	0.892	0.906	0.895	0.906
		0.8	0.818	0.800	0.819	0.783	0.759	0.800	0.819	0.802	0.819
		1	0.641	0.618	0.644	0.603	0.565	0.618	0.642	0.618	0.642
		1.25	0.387	0.370	0.396	0.361	0.319	0.370	0.390	0.370	0.390
		1.5	0.190	0.182	0.201	0.176	0.148	0.182	0.195	0.182	0.195
	0.15	0.667	0.982	0.980	0.983	0.979	0.973	0.980	0.982	0.980	0.982
		0.8	0.944	0.941	0.948	0.939	0.924	0.941	0.946	0.941	0.946
		1	0.817	0.810	0.826	0.807	0.778	0.810	0.820	0.810	0.820
		1.25	0.552	0.539	0.566	0.539	0.499	0.539	0.552	0.539	0.553
		1.5	0.273	0.261	0.286	0.261	0.230	0.261	0.271	0.261	0.273
	0.2	0.667	0.998	0.997	0.998	0.997	0.996	0.997	0.998	0.997	0.998
		0.8	0.987	0.986	0.988	0.986	0.982	0.986	0.987	0.986	0.987
		1	0.923	0.918	0.928	0.918	0.903	0.918	0.921	0.918	0.923
		1.25	0.699	0.684	0.710	0.684	0.652	0.684	0.692	0.684	0.699
		1.5	0.368	0.353	0.382	0.357	0.323	0.353	0.361	0.353	0.368
	0.25	0.667	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.8	0.998	0.997	0.998	0.997	0.997	0.997	0.997	0.997	0.998
		1	0.974	0.971	0.976	0.971	0.965	0.971	0.972	0.971	0.974
		1.25	0.815	0.805	0.823	0.809	0.782	0.805	0.809	0.805	0.814
		1.5	0.471	0.454	0.481	0.462	0.424	0.456	0.462	0.454	0.466

Among the group of Taylor Series variance expansion methods, the Modified Taylor Series method consistently produces the highest power and the Adapted Agresti method has the lowest power. As is expected, the Taylor Series Adjusted Alpha method produces a slightly lower power than the Taylor Series method. In the group of Quadratic methods, Farrington–Manning method 2 produces the highest power. The Pearson method also has higher power than the Deviance method in the maximum likelihood group of methods.

Table 4 summarizes the Type I error from the 100,000 simulations for $\alpha = 0.025$. Although the Modified Taylor Series method produces the highest power in the Taylor Series variance expansion group, it also has the most inflated Type I error rate. The Taylor Series Adjusted Alpha method and the Adapted Agresti method seem to control the Type I error the best in this group. Farrington–Manning method 2 appears to have noteworthy Type I error inflation in some cases, whereas Bailey's method controls this error the best in the group of quadratic methods. The Pearson method also inflates the Type I error somewhat, whereas the Deviance method produces Type I errors close to the nominal level.

AGREEMENT

It is apparent that the methods do not produce exactly the same confidence limits and, therefore, may disagree on the decision to reject the null hypothesis of inferiority. For the simulations, the number of times one method rejected the null and another method did not reject the null was counted. A percent disagreement was calculated out of the 100,000 simulations performed. A summary of these percents can be found in Table 5 for selected cases where $\alpha = 0.025$, $n = 100$, and $\pi_C = 0.1$.

Of the Taylor Series Variance Expansion methods, the Adapted Agresti method appears to produce consistently higher, more conservative upper confidence limits because of its relatively high percent of time the null is not rejected for $\theta = 1$ when the Taylor Series, Taylor Series Adjusted Alpha, and Modified Taylor Series methods do reject the null hypothesis. Bailey's method tends to reject the null hypothesis more often than the quadratic method when they disagree on the decision made. Overall, it appears that the Adapted Agresti and Quadratic methods are most often not rejecting the null hypothesis when the other methods do conclude noninferiority by rejecting the null.

SAMPLE SIZE CALCULATIONS

An immediate application of these results arises in the design of noninferiority clinical trials. The Taylor Series method provides a fairly straightforward form from which to obtain sample size calculations. A conservative form of the variance is the following:

$$\operatorname{var}\left\{ \log_e \frac{y_T}{y_C} \right\} \approx \left\{ \frac{1}{n\pi_C}\left(\frac{1}{\theta}+1\right) - \frac{2}{n} \right\} < \left\{ \frac{1}{n\pi_C}\left(\frac{1}{\theta}+1\right) \right\} = v^*. \qquad (1)$$

Table 4 Summary of Type I error from simulations ($\theta = 2$) $\alpha = 0.025$

		Taylor Series variance expansion methods				Quadratic methods			Maximum likelihood methods	
n	π_C	Taylor Series	Taylor Series Adjusted Alpha	Modified Taylor Series	Adapted Agresti	Quadratic	Bailey	F–M 2	Deviance	Pearson
100	0.1	0.027	0.023	0.028	0.020	0.016	0.023	0.063	0.024	0.027
	0.15	0.027	0.025	0.030	0.025	0.018	0.026	0.027	0.025	0.028
	0.2	0.029	0.025	0.031	0.025	0.019	0.027	0.027	0.025	0.029
	0.25	0.029	0.025	0.031	0.026	0.020	0.026	0.026	0.024	0.027
140	0.1	0.028	0.025	0.030	0.023	0.019	0.025	0.030	0.025	0.030
	0.15	0.027	0.025	0.029	0.025	0.019	0.025	0.027	0.025	0.028
	0.2	0.028	0.025	0.030	0.025	0.020	0.025	0.027	0.025	0.028
	0.25	0.028	0.025	0.031	0.026	0.021	0.026	0.026	0.025	0.027
200	0.1	0.026	0.025	0.029	0.024	0.019	0.025	0.027	0.025	0.027
	0.15	0.028	0.025	0.030	0.025	0.021	0.025	0.026	0.025	0.028
	0.2	0.027	0.024	0.028	0.025	0.020	0.024	0.025	0.024	0.026
	0.25	0.028	0.025	0.030	0.027	0.022	0.026	0.026	0.025	0.027

Table 5 Summary of percent disagreement between methods $\alpha = 0.025, n = 100$ and $\pi_C = 0.1$

		Null not rejected								
θ	Null rejected	Taylor Series	Taylor Series Adjusted Alpha	Modified Taylor Series	Adapted Agresti	Quadratic	Bailey	F–M 2	Deviance	Pearson
1	Taylor Series	—	1.4	<0.1	5.9	8.4	1.4	0	0.3	0
	Taylor Series Adj. Alpha	0	—	<0.1	4.4	6.9	0	0	0	0
	Modified Taylor Series	1.2	2.6	—	7.1	9.5	2.6	<0.1	1.4	<0.1
	Adapted Agresti	0	0	0	—	2.5	0	0	0	0
	Quadratic	0	0	<0.1	<0.1	—	0	0	0	0
	Bailey	0	0	<0.1	4.4	6.9	—	0	0	0
	F–M 2	2.2	3.7	1.0	8.1	10.6	3.7	—	2.5	1.0
	Deviance	0	1.2	<0.1	5.6	8.1	1.2	0	—	0
	Pearson	1.9	3.3	0.7	7.8	10.2	3.3	0.7	2.1	—
2	Taylor Series	—	0.4	0	0.7	1.1	0.4	0	0.3	0
	Taylor Series Adj. Alpha	0	—	0	0.3	0.7	0	0	0	0
	Modified Taylor Series	0.1	0.5	—	0.8	1.2	0.5	0.1	0.5	0.1
	Adapted Agresti	0	0	0	—	0.4	0	0	0	0
	Quadratic	0	0	0	0	—	0	0	0	0
	Bailey	0	0	0	0.3	0.7	—	0	0	0
	F–M 2	3.6	4.0	3.6	4.3	4.7	4.0	—	3.9	3.6
	Deviance	0	<0.1	0	0.3	0.8	<0.1	0	—	0
	Pearson	<0.1	0.4	0	0.7	1.1	0.4	<0.1	0.4	—

Motivation for obtaining a sample size formula begins with formulation of a z-statistic:

$$z = \frac{\log_e \theta - \log_e \theta_0}{\sqrt{v^*}} \tag{2}$$

where θ_0 is the value of θ under the null hypothesis, and in the examples presented thus far, θ_0 has taken the value 2.

The equation below results from squaring Eq. (2) and writing z in terms of the Type I and Type II errors:

$$(z_\alpha + z_\beta)^2 = \frac{\left(\log_e \frac{\theta}{\theta_0}\right)^2}{v^*} \tag{3}$$

which produces the following equation after algebraic manipulations

$$\frac{(z_\alpha + z_\beta)^2 \left\{ \frac{1}{n\pi_C} \left(\frac{1}{\theta} + 1 \right) \right\}}{\left\{ \log_e \left(\frac{\theta}{\theta_0} \right) \right\}^2} = 1. \tag{4}$$

This form is then solved for the sample size, n, and can be written as

$$n = \frac{(z_\alpha + z_\beta)^2 \left\{ \frac{1}{\pi_C} \left(\frac{1}{\theta} + 1 \right) \right\}}{\left\{ \log_e \left(\frac{\theta}{\theta_0} \right) \right\}^2} \tag{5}$$

which depends only on a prespecified one-sided type I error (α), power ($1 - \beta$), event rate in the control group (π_C), and a hypothesized ratio of events in the treatment vs. the control group (θ) with θ_0, the null hypothesis, specified. This formula is useful in practice due to ease of computation.

To evaluate whether Eq. (5) produces sample sizes that maintain the prespecified power, results were compared with those obtained from simulations. These results were based on 100,000 simulations. The sample size Eq. (5) was written in terms of power as seen in the following equation:

$$z_\beta = \sqrt{\frac{n\pi_C \theta \left\{ \log_e \left(\frac{\theta}{\theta_0} \right) \right\}^2}{1 + \theta}} - z_\alpha \tag{6}$$

where power is $\Phi(z_\beta)$ and $\Phi(\cdot)$ is the standard normal probability. In addition, this sample size formula and power calculation in Eqs. (5) and (6) can be modified for the Taylor Series Adjusted Alpha method, which controls Type I error better than the Taylor Series method. This adjustment uses an alpha level of 0.0025 lower than that specified. For example, at a specified $\alpha = 0.025$, the critical value would be calculated at $0.025 - 0.0025 = 0.0225$.

For combinations of α, π_C, θ, and n generated in the 100,000 simulations, a power based on the sample size formula was calculated by using Eq. (6) and the

Table 6 Power comparisons for sample size formulas $\alpha = 0.025$

			Taylor Series method		Taylor Series Adjusted Alpha Method		Farrington–Manning method 1		Farrington–Manning method 2	
n	π_C	θ	Calculated	Simulated	Calculated	Simulated	Calculated	Simulated	Calculated	Simulated
100	0.1	0.667	0.709	0.746	0.701	0.728	0.658	0.687	0.624	0.755
		0.8	0.613	0.645	0.603	0.628	0.570	0.587	0.529	0.661
		1	0.462	0.485	0.452	0.474	0.439	0.437	0.391	0.512
		1.25	0.296	0.316	0.287	0.308	0.291	0.278	0.243	0.343
		1.5	0.174	0.186	0.167	0.180	0.177	0.158	0.137	0.208
	0.2	0.667	0.928	0.958	0.925	0.955	0.923	0.947	0.913	0.960
		0.8	0.861	0.907	0.856	0.903	0.860	0.888	0.844	0.908
		1	0.708	0.777	0.699	0.772	0.723	0.744	0.695	0.774
		1.25	0.469	0.550	0.459	0.541	0.502	0.505	0.464	0.541
		1.5	0.258	0.314	0.250	0.307	0.288	0.273	0.252	0.307
140	0.1	0.667	0.830	0.853	0.824	0.853	0.783	0.813	0.755	0.856
		0.8	0.739	0.765	0.731	0.765	0.695	0.713	0.658	0.767
		1	0.575	0.602	0.565	0.602	0.547	0.546	0.498	0.606
		1.25	0.369	0.395	0.360	0.395	0.363	0.350	0.310	0.404
		1.5	0.209	0.225	0.202	0.225	0.213	0.193	0.168	0.233

	0.2	0.667	0.979	0.991	0.978	0.990	0.976	0.988	0.973	0.991
		0.8	0.944	0.968	0.941	0.966	0.943	0.958	0.935	0.968
		1	0.829	0.880	0.822	0.878	0.842	0.857	0.821	0.880
		1.25	0.583	0.661	0.573	0.656	0.620	0.622	0.584	0.661
		1.5	0.321	0.385	0.312	0.378	0.359	0.348	0.320	0.383
200	0.1	0.667	0.928	0.946	0.925	0.943	0.895	0.928	0.878	0.946
		0.8	0.861	0.884	0.856	0.881	0.824	0.855	0.796	0.884
		1	0.708	0.740	0.699	0.734	0.678	0.698	0.633	0.740
		1.25	0.469	0.505	0.459	0.493	0.461	0.457	0.404	0.505
		1.5	0.258	0.288	0.250	0.276	0.264	0.250	0.212	0.288
	0.2	0.667	0.997	0.999	0.997	0.999	0.996	0.999	0.996	0.999
		0.8	0.987	0.994	0.986	0.994	0.987	0.992	0.984	0.994
		1	0.927	0.957	0.924	0.956	0.935	0.948	0.925	0.957
		1.25	0.716	0.791	0.707	0.785	0.754	0.766	0.724	0.788
		1.5	0.407	0.485	0.397	0.479	0.456	0.452	0.414	0.479

Taylor Series Adjusted Alpha formula. Power was also calculated by using the sample size formulas presented by Farrington and Manning (1990) for F–M method 1 and F–M method 2. This calculated power was then compared with the power obtained from the simulations for each method. Results can be seen in Table 6 for $\alpha = 0.025$.

For all methods presented, the calculated power is less than the simulated power; therefore, the sample size formulas are somewhat conservative, which is beneficial when determining sample size for clinical trials. Farrington–Manning method 2 has the largest discrepancy between calculated and simulated power, with the simulated power being larger although possibly because of its somewhat inflated Type I error.

CONCLUSIONS

The properties and behavior of many different methods for computing confidence intervals for risk ratios have been reviewed. The Pearson method is favored in the literature and produces fairly high power in the simulations, but this method has inflated Type I error rates and requires computer-intensive resources. The Taylor Series Adjusted Alpha method is a fairly simple method in practice while producing good power and maintaining nominal Type I error rates. Bailey's method also has fairly good power and is better at maintaining Type I error rates than the other quadratic methods but can be problematic in certain cases that produce undefined values. The Pearson method, Taylor Series Adjusted Alpha method, and Bailey's method also have fairly low percentages of disagreement in the noninferiority decision made.

Problems due to small event rates were avoided because of use of methods for exact odds ratios for counts less than or equal to 3. Because of this modification, the performance of the methods may be slightly different from that presented previously in the literature.

The fairly straightforward sample size formulas for the Taylor Series and Taylor Series Adjusted Alpha methods make them attractive for use in designing sample size in noninferiority clinical trials. The sample size formulas presented by Farrington and Manning are also useful, although the power and Type I error of these methods are not quite as attractive as other methods.

APPENDIX

The following is SAS code to produce simulations for the Taylor Series method and resulting summaries of power and Type I error. These results can be used to make appropriate adjustments to the alpha level when implementing the Taylor Series Adjusted Alpha method.

```
/*****************************************************/
/** Generates dataset of random numbers used for simulations **/
/*****************************************************/
%macro sims;

data sims;
keep n pi_c theta sims alpha pi_t y_c y_t ts_ul ts;
```

```
  do n = 100,140,200; /** CHANGE these parameters **/
    do pi_c = 0.10,0.15,0.20,0.25; /** CHANGE these parameters **/
      do theta = 0.667,0.800,1.000,1.250,1.500,2.000; /** CHANGE these
      parameters **/
        do sims = 1 to &numsims;
            pi_t = theta*pi_c;
            /** Sample from Binomial Distributions **/
            y_c = ranbin(&seed1,n,pi_c);
            y_t = ranbin(&seed2,n,pi_t);
          do alpha = 0.025, 0.05;
           z_alpha = probit(1-alpha);

           /***********************/
           /** Taylor Series Method **/
           /***********************/
           if (y_t ne 0 & y_c ne 0) then do; /** Check for undefined values
           **/
            ts_1 = log(y_t/y_c);
            ts_2 = (1/y_t) + (1/y_c) - (2/n);
            ts_ul = exp(ts_1 + z_alpha*sqrt(ts_2)); /** Upper confidence
            limit **/
            if (. < ts_ul < &null) then ts = 1; else ts = 0; /** Decision **/
           end;

            output;
          end; /** alpha **/
        end; /** sims **/
      end; /** theta **/
    end; /** pi_c **/
  end; /** n **/

run;

%mend sims;

/************************************/
/** Macro to produce Exact Odds Ratio CL **/
/************************************/
%macro xor(data = );

data temp;
  set &data(keep = n pi_c theta sims pi_t y_c y_t);
  outcome = 1; count = y_t; type = 'T'; output;
  outcome = 0; count = n-y_t; type = 'T'; output;
  outcome = 1; count = y_c; type = 'C'; output;
  outcome = 0; count = n-y_c; type = 'C'; output;
run;
```

```
/** Exact OR using Proc Freq when y_t or y_c <= 3 **/
%if &data = adj %then %do;
proc freq data = temp order = data noprint;
  by n pi_c theta sims pi_t y_c y_t;
  weight count;
  tables type*outcome;
  exact or;
  output out = or(keep = n pi_c theta sims pi_t y_c y_t xu_rror)
        or;
run;
%end;

/** Exact Logistic Regression when y_t = 0 **/
%else %if &data = adj_0 %then %do;
ods output ExactOddsRatio = or_0
        (keep = n pi_c theta sims pi_t y_c y_t uppercl
          rename = (uppercl = xu_rror));
proc logistic data = temp order = data;
  by n pi_c theta sims pi_t y_c y_t;
  class type(ref = 'C') / param = ref;
  freq count;
  model outcome = type;
  exact 'Treatment' type / estimate = odds;
run;
%end;

%mend xor;

/****************/
/** Body of Code **/
/****************/

%macro main;

ods listing close;

%sims; /** Produce simulations & calculate Taylor
Series confidence limit **/

/** Adjustments for small number of events **/
%let or_flag = 0;
%let or_flag_0 = 0;

data adj adj_0;
  set sims;
  if (y_c = 0) then do; /** NOTHING **/ end;
  else if (y_t = 0) then do;
```

```
    call symput('or_flag_0', 1);
    output adj_0;
  end;
  else if (0 < y_t le 3 | 0 < y_c le 3) then do;
    call symput('or_flag', 1);
    output adj;
  end;
run;

/** Find Exact Upper CL for OR **/
%if &or_flag = 1 %then %do;
  %xor(data = adj);
%end;
%if &or_flag_0 = 1 %then %do;
  %xor(data = adj_0);
%end;

data sims;
%if (&or_flag = 1 & &or_flag_0 = 1) %then %do;
  merge sims
        or
        or_0;
%end;
%else %if &or_flag = 1 %then %do;
  merge sims
        or;
%end;
%else %if &or_flag_0 = 1 %then %do;
  merge sims
        or_0;
%end;
%else %do;
  set sims;
%end;
  by n pi_c theta sims pi_t y_c y_t;

  %if (&or_flag = 1 | &or_flag_0 = 1) %then %do;
  if (xu_rror < &null) then xor = 1; else xor = 0;
  %end;

  if (y_c = 0) then do;
    ts_ul = 100; ts = 0;
  end;
  else if (y_t <= 3 | y_c <= 3) then do;
      ts_ul = xu_rror; ts = xor;
  end;
run;
```

```
/*******************/
/** Find mean of data **/
/*******************/
proc means data=sims noprint;
  class n pi_c theta pi_t alpha;
  var   ts_ul ts;
  output out=mn_sims(drop=_type_ _freq_
              where=(n ne . & pi_c ne . & theta ne . & pi_t ne . & alpha ne .))
      mean=ts_ul ts
      std=  sd_ts_ul sd_ts;
run;


/*************************************/
/** Produce report of Power and Type I Error **/
/*************************************/
ods listing;
title ''Summary of Power from Simulations for Taylor Series Method";
proc report data=mn_sims headline headskip nowd spacing=5 missing;
  where theta ne &null;
  column alpha n pi_c theta ts;
  define alpha / ''Alpha'' order order=internal width=5 center;
  define n / ''N'' order order=internal width=3;
  define pi_c / ''Pi C'' width=4 order order=internal;
  define theta / ''Theta'' width=5;
  define ts / ''Taylor Series/Power'' width=20 center format=5.3;

  break after alpha / page;
  break after n / skip;
  break after pi_c / skip;
run;


title ''Summary of Type I Error from Simulations for Taylor Series Method";
proc report data=mn_sims headline headskip nowd spacing=5 missing;
  where theta = &null;
  column alpha n pi_c theta ts;
  define alpha / ''Alpha'' order order=internal width=5 center;
  define n / ''N'' order order=internal width=3;
  define pi_c / ''Pi C'' width=4 order order=internal;
  define theta / ''Theta'' width=5;
  define ts / ''Taylor Series/Type I Error'' width=20 center
  format=5.3;

  break after alpha / page;
  break after n / skip;
run;
%mend main;
```

```
/*******************/
/** Run Main macro **/
/*******************/

%let numsims = 1000; /** PROVIDE number of simulations **/
%let null = 2;  /** PROVIDE null hypothesis **/
%let seed1 = 839; /** PROVIDE first seed for simulations **/
%let seed2 = 3921; /** PROVIDE second seed for simulations **/

%main;
```

REFERENCES

Agresti, A., Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *Am. Statist.* 52:119–126.

Bailey, B. J. R. (1987). Confidence limits to the risk ratio. *Biometrics* 43:201–205.

Blackwelder, W. C. (1982). Proving the null hypothesis in clinical trials. *Control. Clin. Trials* 3:345–353.

EquivTest 1.0. (2000). *Software for the Statistical Analysis of Equivalence and Bioavailability Studies*. Cork, Ireland: Statistical Solutions Ltd.

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

Fieller, E. C. (1944). A fundamental formula in the statistics of biological assay and some applications. *Q. J. Pharm. Pharmacol.* 17:117–123.

Gart, J. J., Nam, J. (1988). Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. *Biometrics* 44:323–338.

Hung, H. M. J., Wang, S. J., Tsong, Y., Lawrence, J., O'Neill, R. T. (2001). Some fundamental issues with non-inferiority testing in active controlled trials. *Stat. Med.* 22:213–225.

Katz, D., Baptista, J., Azen, S. P., Pike, M. C. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* 34:469–474.

Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics* 40:513–517.

(2002). *nQuery Advisor Version 5.0 User's Guide*. Cork, Ireland: Statistical Solutions Ltd.

(1990). *SAS/STAT User's Guide*. Vol. 1. Version 6. Cary, North Carolina: SAS Institute Inc., pp. 871–872.

(1990). *StatXact4 for Windows User Manual*. Cambridge, Massachusetts: CYTEL Software Corporation, pp. 435–452.