Reports of the Institute of Biostatistics

No 02 / 2008

Leibniz University of Hannover

Natural Sciences Faculty

Title: *Properties of confidence intervals for the comparison of small binomial proportions when sample sizes are large*

Authors: *Frank Schaarschmidt*

1 Introduction

Consider two independent binomial random variables Y_i, $i = 1, 2$, with $Y_i \sim Bin(n_i, \pi_i)$, with $i = 1$ specifying an untreated control group and $i = 2$ specifying a group exposed to a treatment of interest. Assume further that π specifies the proportion of a rare detrimental event, and at least $\pi_1 << 0.5$, and large sample sizes are available, i.e., $n_1 \geq 1000$. In safety assessment, aim is to measure the possible dissimilarity of π_2 compared to π_1. There are three commonly used measures for dissimilarity of proportions, the risk difference $\delta = \pi_2 - \pi_1$, the risk ratio $\rho = \pi_2/\pi_1$ and the odds ratio $\psi = \frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)}$ Here, methods for constructing confidence intervals for ψ, ρ, and δ are considered.

1.1 Proof of Safety and Proof of Hazard

A **Proof of Safety** for a novel treatment 2 compared to a standard treatment 1 is expressed by the following pairs of hypotheses with respect to $\psi = \frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}}$, $\delta = \pi_2 - \pi_1$, and $\rho = \pi_2/\pi_1$:

$$H_0 : \psi \geq \psi_0 \text{ vs. } H_A : \psi < \psi_0 \text{ , with } \psi_0 > 1 \tag{1}$$

$$H_0 : \delta \geq \delta_0 \text{ vs. } H_A : \delta < \delta_0 \text{ , with } \delta_0 > 0 \tag{2}$$

$$H_0 : \rho \geq \rho_0 \text{ vs. } H_A : \rho < \rho_0 \text{ , with } \rho_0 > 1 \tag{3}$$

All alternative hypotheses H_A define a state that π_2 is not relevantly increased over π_1, where relevance is defined by ψ_0, δ_0, and ρ_0. Hence, for general problems without a-priori definition of the null-parameters, the estimation of upper confidence limits is of interest.

A **Proof of Hazard** for a novel treatment 2 compared to a standard treatment 1 is expressed by the following pairs of hypotheses with respect to $\psi = \frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}}$, $\delta = \pi_2 - \pi_1$, and $\rho = \pi_2/\pi_1$:

$$H_0 : \psi \leq \psi_0 \text{ vs. } H_A : \psi > \psi_0 \text{ , with } \psi_0 \geq 1 \tag{4}$$

$$H_0 : \delta \leq \delta_0 \text{ vs. } H_A : \delta > \delta_0 \text{ , with } \delta_0 \geq 0 \tag{5}$$

$$H_0 : \rho \leq \rho_0 \text{ vs. } H_A : \rho > \rho_0 \text{ , with } \rho_0 \geq 1 \tag{6}$$

Here, the alternative hypotheses H_A define a state of hazardousness of the novel treatment, i.e., $\pi_2 > \pi_1$, where either marginal hazardousness ($\psi_0 = 1$, $\delta_0 = 0$, $\rho_0 = 1$) or relevant hazardousness ($\psi_0 > 1$, $\delta_0 > 0$, $\rho_0 > 1$) could be of interest. Then, the estimation of upper confidence limits is of interest. Often, tests for the hypotheses above are inappropriately interpreted in contradiction to Neyman-Pearson concept of statistical test, i.e. concluding for safety when the null hypothesis can not be rejected. If this is common practice, conservative procedures should be avoided in favour of liberal procedure, with the aim to increase power and improve the 'confidence in negative results'.

1.2 Properties of δ, ρ, and ψ with respect to safety assessment

In statistical safety assessment, a safety margin ρ_0, δ_0, ψ_0 as to be defined which describes acceptable deviations from the state of exact equality of risks, $\pi_2 = \pi_1$. As outlined by Wellek (2005) for the general problem of proving bioequivalence, the definition of safety margins based on δ is problematic, since it depends on the proportion in the control group, π_1. Simply, a difference of $\delta = 0.01$ might be acceptable when $\pi_1 = 0.1$, but might be not when $\pi_1 = 0.01$ or $\pi_1 = 0.001$. The risk ratio and odds ratio do not have this property. However, in general settings with $\pi \in [0, 1]$, the risk ratio is not invariant with respect to the definition of success and falure, while the odds ratio and the risk difference are. An alternative hypothesis $\frac{(1-\pi_2)/(\pi_2)}{(1-\pi_1)/(\pi_1)} < \psi_0$ has the same meaning as $\frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} < \frac{1}{\psi_0}$. According to Wellek (2005): if a treatment is non-inferior to the control with regard to the event "success", it is also non-inferior to the control with regard to failure, if non-inferiority is defined in terms of the odds ratio. For the difference, $\pi_1 - \pi_2 < \delta_0$ has the same meaning as $(1 - \pi_1) - (1 - \pi_2) > -\delta_0$. However, for the risk ratio, π_2/π_2, and $(1 - \pi_2) / (1 - \pi_1)$ do not have the same meaning.

For these reasons, Wellek (2005) strongly recommends the odds ratio as natural measure of dissimilarity in the general problem of proving non-inferiority. For rare detrimental events, where the proportion of interest is clearly defined and $\pi << 0.5$ also the risk ratio has acceptable properties.

1.3 Important requirements on the properties of confidence limits for δ, ρ, and ψ

Except an acceptable coverage probability, the confidence intervals should exhibit the following properties:

1. Decisions should be invariant with regard to the exchange of control and treatment, i.e., if the upper $(1-\alpha)$ limit for π_2/π_1 excludes ρ_0, the lower $(1-\alpha)$ limit for π_1/π_2 should also exclude $1/\rho_0$.

2. Confidence limits should be computable for each reasonable outcome of the experiment $\{y_1, y_2\}$, which provides information on the parameter of interest. Confidence intervals should not be empty.

3. Confidence limits should change monotonically for monotone change in dissimilarity of the outcome, i.e., considering confidence limits for $\rho = \pi_2/\pi_1$, denoting the bound resulting from the event $\{y_2 = 3, y_1 = 2\}$ $\hat{\rho}_{32}$ and the bound resulting from the event $\{y_2 = 3, y_1 = 1\}$ $\hat{\rho}_{31}$, then $\hat{\rho}_{31} \geq \hat{\rho}_{32}$, since the evidence for $\pi_2 > \pi_1$ is larger in $\{y_2 = 3, y_1 = 1\}$ than in $\{y_2 = 3, y_1 = 2\}$.

4. Confidence limits should always contain the point estimates, and span only in the range, where the parameter of interest is defined.

The first requirement is violated by only very few confidence interval methods proposed in the literature, see e.g. the discussion in Lecoutre and Faure (2007) and Agresti and Min (2005) who discourage the use of the method proposed by Zhou, Tsao and Qin (2004) for δ and the Bayesian Highest posterior density credible intervals for ρ and ψ Lecoutre and Faure (2007).

The second requirement is discussed by Gart and Nam (1988) for the risk ratio ρ. The problem, that certain confidence interval methods are not computable for certain outcomes $\{y_2, y_1\}$ occurs for extreme events involving $y_i = n_i$ or $y_i = 0$. Here, only the latter case is of importance. Note, that the event $\{y_2 = 0, y_1 = 0\}$ is non-informative for the risk ratio ρ and odds ratio ψ, but is informative for the risk difference δ.

2 Methods

2.1 Confidence limits for the odds ratio ψ

The point estimator for ψ is $\hat{\psi} = \frac{y_2/(n_2-y_2)}{y_1/(n_1-y_1)}$. A simple large sample interval with nominal confidence coefficient $(1-\alpha)$ can be computed according to Equation (7).

$$\exp\left(\log\left(\tilde{\psi}\right) \pm z_{1-\alpha/2}\tilde{\sigma}\right), \tag{7}$$

with $\tilde{\psi} = \frac{(y_2+0.5)/(n_2-y_2+0.5)}{(y_1+0.5)/(n_1-y_1+0.5)}$, and $\tilde{\sigma} = \sqrt{\frac{1}{y_1+0.5} + \frac{1}{n_1-y_1+0.5} + \frac{1}{y_2+0.5} + \frac{1}{n_2-y_2+0.5}}$. This interval is computable even in the case $\{y_2 = 0, y_1 = 0\}$ and as resonable small sample performance in two-sided application. This method is referred to as **adjusted Woolf** in the simulation study of Lawson (2004).

Alternatively, estimates from the generalized linear model with binomial family and logit link could be used to derive estimates for $\log(\psi)$ and its standard error and replacing $\tilde{\psi}$ and $\tilde{\sigma}$ in Equation (7). See Gerhard (2007) for detailed information. Here, this method is referred to as **GLM**.

In `R` an exact confidence interval based on the inversion of Fishers exact test for shifted values of ψ, based on the central and non-central hypergeometric distribution. It is available also for large sample sizes. The algorithm is described in Clarkson et al. (1993), and we will refer to it as **Exact**.

Figures 1 and 2 display calculated upper and lower bounds of two-sided 95% adjusted Woolf, GLM and Exact confidence intervals with for $n_1 = n_2 = 1000$, and certain events $\{y_1, y_2\}$. An important disadvantage of the GLM method is obvious, i.e., the violation of requirement 3 in Section 1.3, whenever $y_1 = 0$ or $y_2 = 0$ occurs. The Exact and the adjusted Woolf interval do not show this property. However, the adjusted Woolf interval does not contain the point estimates $\infty, 0$ in the cases $\{y_1 = 1, y_2 = 0\}$, $\{y_1 = 0, y_2 = 5\}$, respectively.

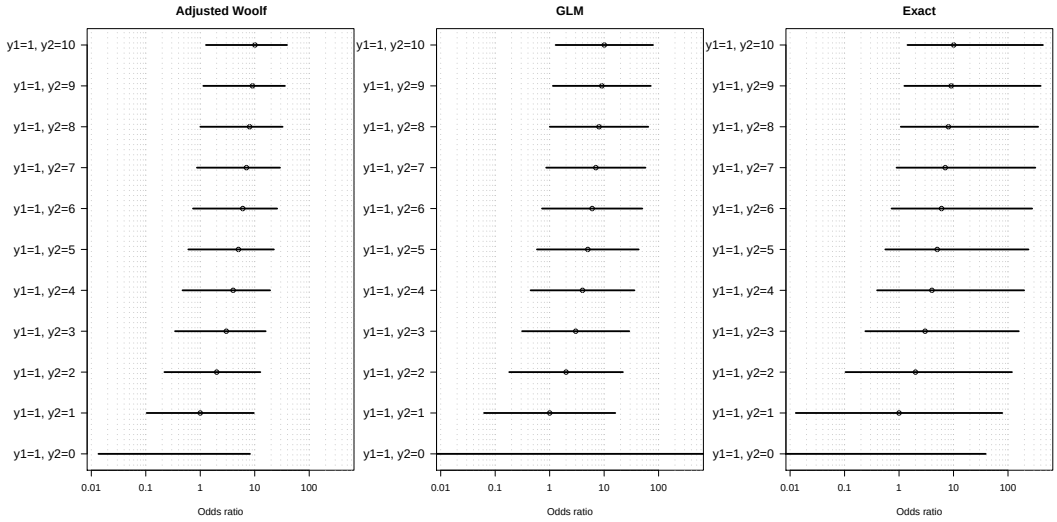

Figure 1: Two-sided 0.95 confidence intervals according to adjusted Woolf, GLM and Exact method for the events $y_1 = 1$, $y_2 = 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0$ with $n_1 = n_2 = 1000$. The intervals for the odds ratio are displayed on the x-axis in logarithmic scale.

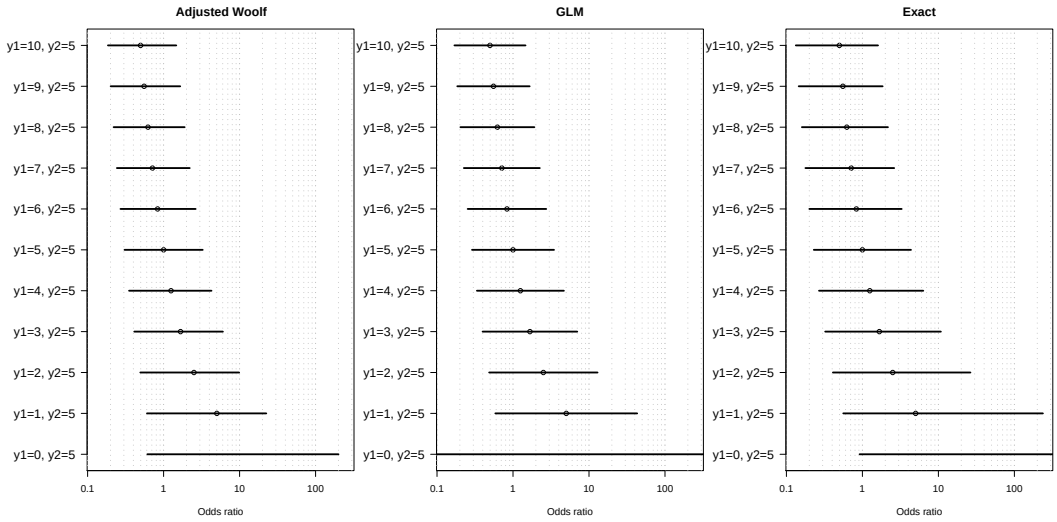

Figure 2: Two-sided 0.95 confidence intervals according to adjusted Woolf, GLM and Exact method for the events $y_1 = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, $y_2 = 5$ with $n_1 = n_2 = 1000$. The intervals for the odds ratio are displayed on the x-axis in logarithmic scale.

2.2 Confidence limits for the risk ratio ρ

The point estimate for ρ is $\hat{\rho} = p_2/p_1$, with $p_i = y_i/n_i$. Gart and Nam (1988) discuss a simple large sample interval presented in equation (8), called **Crude** in the following.

$$\exp\left(\log\left(\tilde{\rho}\right) \pm z_{1-\alpha/2}\sqrt{\hat{u}}\right) \tag{8}$$

where

$$\tilde{\rho} = \frac{\left(y_2 + 0.5\right) / \left(n_2 + 0.5\right)}{\left(y_1 + 0.5\right) / \left(n_1 + 0.5\right)},$$

and

$$\hat{u} = \hat{V}\left(\log\left(\tilde{\rho}\right)\right) = \frac{1}{y_2 + 0.5} + \frac{1}{y_1 + 0.5} - \frac{1}{n_2 + 0.5} - \frac{1}{n_1 + 0.5},$$

and $z_{1-\alpha/2}$ is the quantile of the standard normal distribution.

This method yields degenerate intervals $[1, 1]$ in case $y_2 = n_2$ and $y_1 = n_1$, which is of no importance in the problem of considering rare event rates. It is one of the few intervals considered in Gart and Nam (1988) which can be computed in case of the event $y_2 = 0$ and $y_1 = 0$. It has the advantage that calculated bounds for ρ are the same as the reciprocal of the bounds calculated for $1/\rho$ Gart and Nam (1988), i.e. they are invariant with respect to exchanging numerator and denominator. Dann and Koch (2005) call this method "Modified Taylor Series".

The second method considered here is the "Score" method discussed in Gart and Nam (1988), section 3.3 (Methods based on Likelihood Methods). Confidence bounds for ρ are found by iterative process, involving the solution of quadratic equations. In the limited simulation study presented by Gart and Nam (1988), this method shows best coverage probabilities among the considered methods. For computational details, refer to Gart and Nam (1988). Gart and Nam (1988) state, that the method is not computable for the case $y_1 = y_2 = 0$. However, in my implementation, problems in the iterative process occured also for a number of other events, like $y_1 = 0$, $y_2 = 0$, $y_1 = n_1$, $y_2 = n_2$. In the simulation study, I replace $y_1 = 0.5$, $y_1 = 0.5$, $y_1 = n_1 - 0.5$, $y_1 = n_2 - 0.5$, in the iterative process if these events occured, respectively. Therefore, the method referred to as **Score** in this report is not exactly the same as that described by Gart and Nam (1988).

2.3 Confidence limits for δ

In the recent years, various papers have been published considering the construction of confidence intervals for the difference of binomial proportions between two independent samples. The recommendations in these publications are based on simulations of the coverage probability of two-sided confidence intervals. The most recent simulation studies by Newcombe (1998) and Brown and Li (2005), comparing several of these methods, recommend certain approximate methods. However, in toxicological applications, interest is usually in directional decisions, i.e., only the upper or only the lower limits are of interest. Such comparisons have not been considered in the literature so far. Cai (2005) showed for the problem of one binomial proportion, that the coverage probabilities of approximate upper or lower limits can be seriously asymmetric. Therefore, additional simulation studies are needed.

Let $\hat{\pi}_i = y_i/n_i$ denote the estimated proportions the two independent samples $i = 1, 2$. The Wald (**Wald**) interval for the difference of proportions is:

$$\hat{\pi_2} - \hat{\pi}_1 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \tag{9}$$

Agresti and Caffo (2000) add four pseudo-observations, one to each cell in the 2×2-table, and use the formula to construct confidence intervals. Denoting $\tilde{\pi}_i = (y_i + 1) / (n_i + 2)$, the Add-4 (**Add4**) interval is:

$$\tilde{\pi}_2 - \tilde{\pi}_1 \pm z_{1-\alpha/2}\sqrt{\frac{\tilde{\pi}_1(1-\tilde{\pi}_1)}{n_1+2} + \frac{\tilde{\pi}_2(1-\tilde{\pi}_2)}{n_2+2}} \tag{10}$$

A slightly less conservative interval can be derived by adding only 0.5 pseudo-observations to each cell in the 2×2-table. Denoting $\tilde{\pi}_i = (y_i + 0.5) / (n_i + 1)$, the Add-2 (**Add2**) interval is:

$$\tilde{\pi}_2 - \tilde{\pi}_1 \pm z_{1-\alpha/2}\sqrt{\frac{\tilde{\pi}_1(1-\tilde{\pi}_1)}{n_1+1} + \frac{\tilde{\pi}_2(1-\tilde{\pi}_2)}{n_2+1}} \tag{11}$$

Based on the Score interval proposed by Wilson (1927), Newcombe (1998) introduced an interval for the difference of proportions (referred to as Newcombes Hybrid Score interval, **NHS**). Its variance term is constructed based on Wilson Score confidence limits for the single proportions. The lower $(1-\alpha/2)$-bound is:

$$\hat{\pi}_2 - \hat{\pi}_1 - z_{1-\alpha/2}\sqrt{\frac{l_2(1-l_2)}{n_2} + \frac{u_1(1-u_1)}{n_1}}, \tag{12}$$

and the upper $(1-\alpha/2)$-bound is:

$$\hat{\pi}_2 - \hat{\pi}_1 + z_{1-\alpha/2}\sqrt{\frac{u_2(1-u_2)}{n_2} + \frac{l_1(1-l_1)}{n_1}} \tag{13}$$

where l_i, u_i are the lower and upper bounds of the $(1-\alpha)$ Wilson Score interval for π_i, which can be calculated according to equation 14:

$$[l_i, u_i] = \left[\frac{y_i + \frac{z^2_{1-\alpha/2}}{2}}{n_i + z^2_{1-\alpha/2}} \pm \frac{z_{1-\alpha/2}\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i) + \frac{z^2_{1-\alpha/2}}{4}}}{n_i + z^2_{1-\alpha/2}}\right] \tag{14}$$

The method used in the R function `prop.test` is the continuity corrected (**CC**) confidence interval as described in Newcombe (1998).

3 Simulation study

3.1 Coverage probability

- Sample size $n_1 = n_2 = 1000, 2000$, $n_1 = 2000$, $n_2 = 1000$;

- all combinations of $\pi_1 = 0.001, 0.002, ..., 0.04$, and $\pi_2 = 0.001, 0.002, ..., 0.04$;
- lower (0.95)-confidence limits to investigate the α-control when used in a proof of hazard for rare detrimental events;
- upper (0.95)-confidence limits to investigate the α-control when used in a proof of safety for rare detrimental events;
- coverage probability estimated based on 10000 simulation runs.

3.1.1 Lower 0.95 confidence limits for the risk difference (proof of hazard)

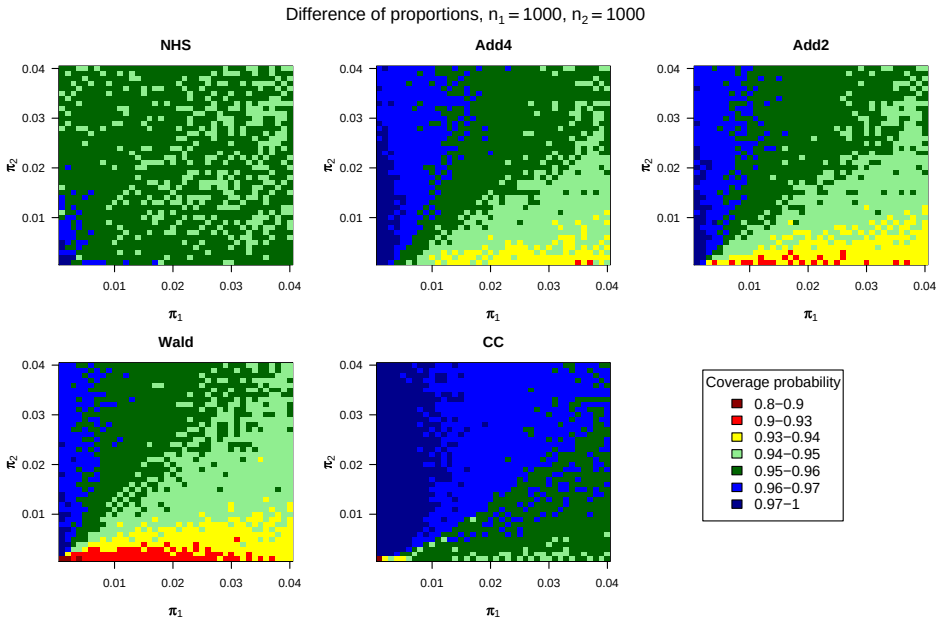

Figure 3: Coverage probability of lower, nominal 0.95 confidence limits for $n_1, n_2 = 1000$

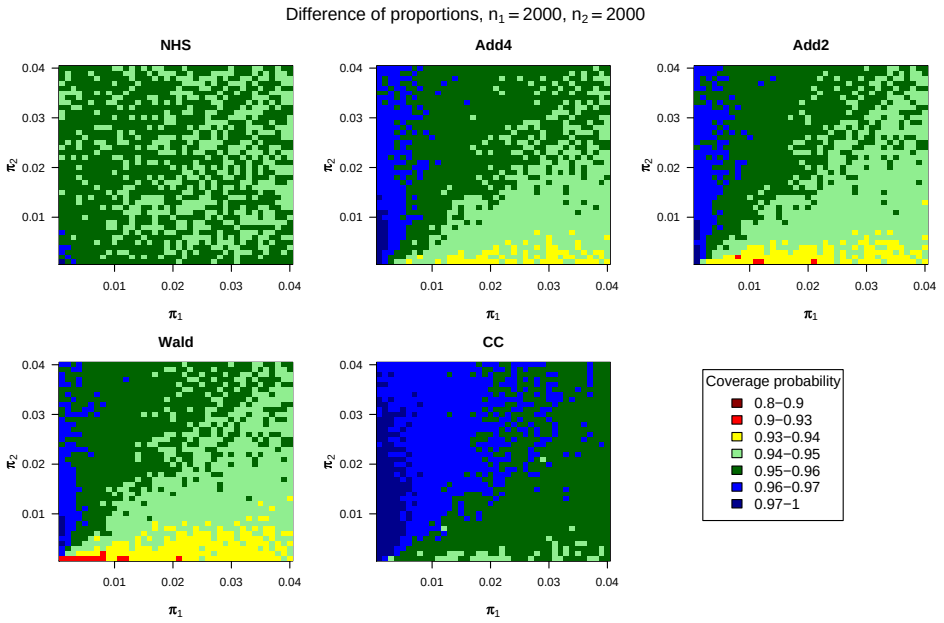

Figure 4: Coverage probability of lower, nominal 0.95 confidence limits for $n_1, n_2 = 2000$

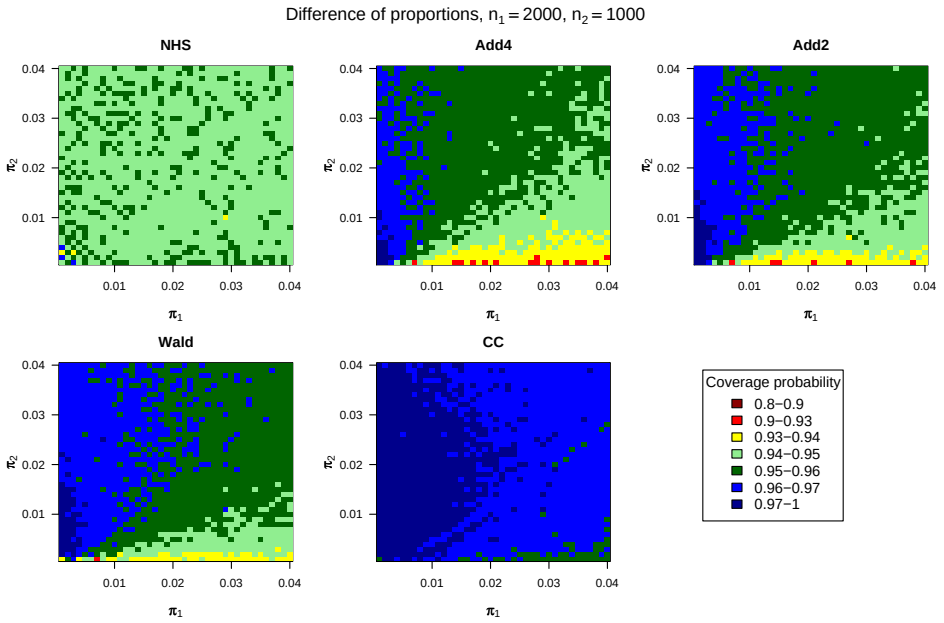

Figure 5: Coverage probability of lower, nominal 0.95 confidence limits for $n_1 = 2000, n_2 = 1000$

3.1.2 Upper 0.95 confidence limits for the risk difference (proof of safety)

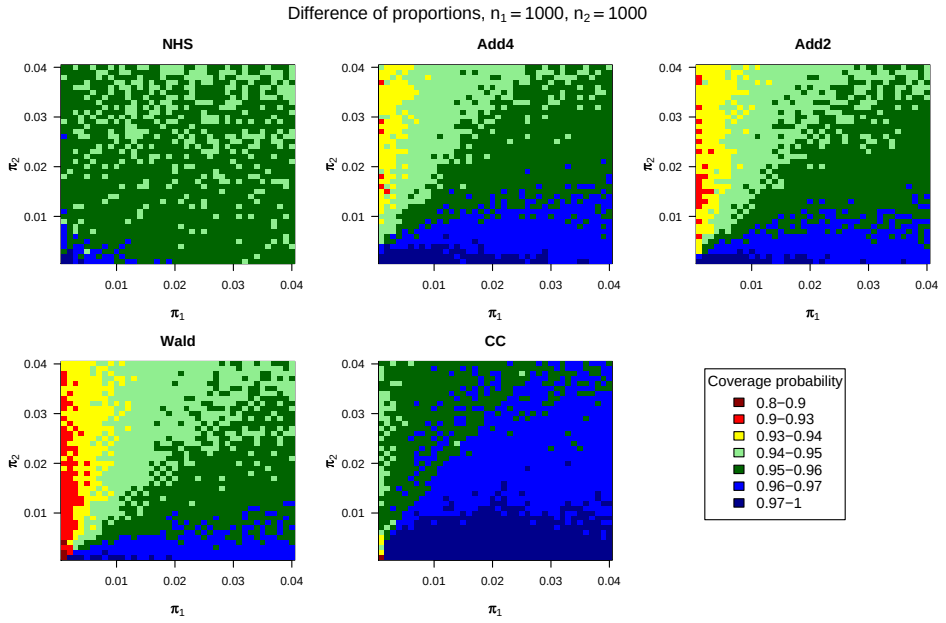

Figure 6: Coverage probability of upper, nominal 0.95 confidence limits for $n_1, n_2 = 1000$

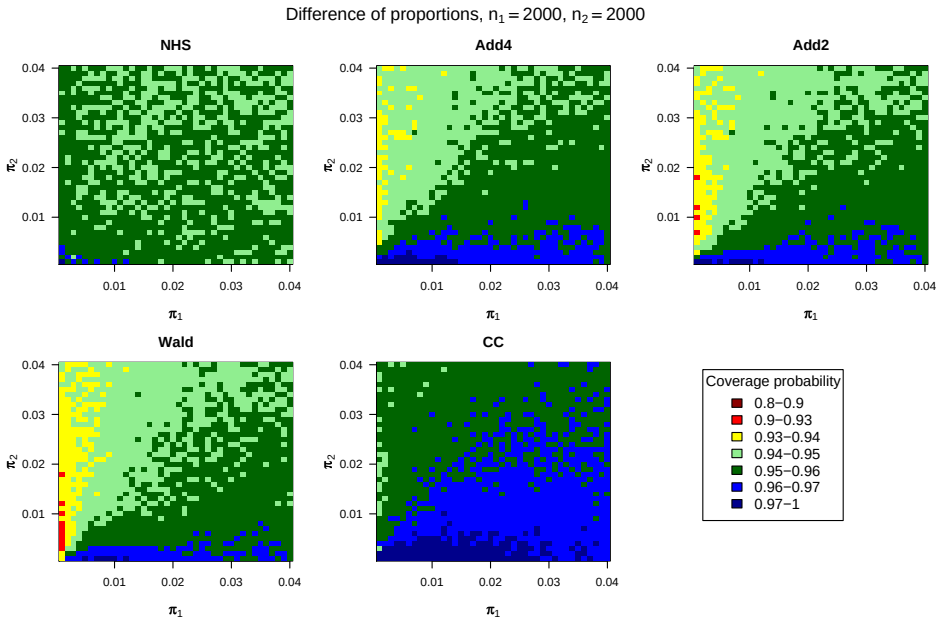

Figure 7: Coverage probability of upper, nominal 0.95 confidence limits for $n_1, n_2 = 2000$

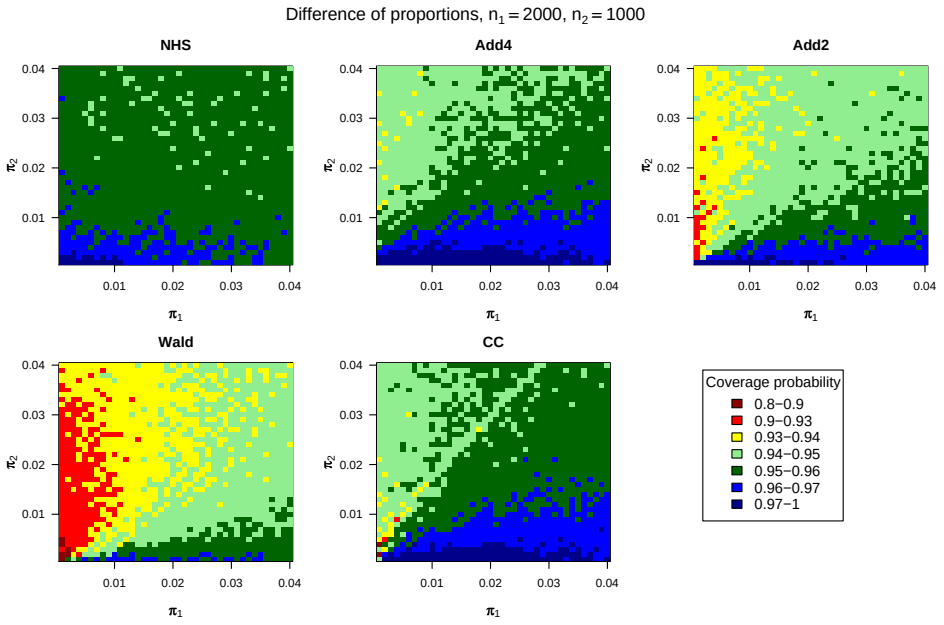

Figure 8: Coverage probability of upper, nominal 0.95 confidence limits for $n_1 = 2000, n_2 = 1000$

3.1.3 Lower 0.95 confidence limits for the risk ratio (proof of hazard)

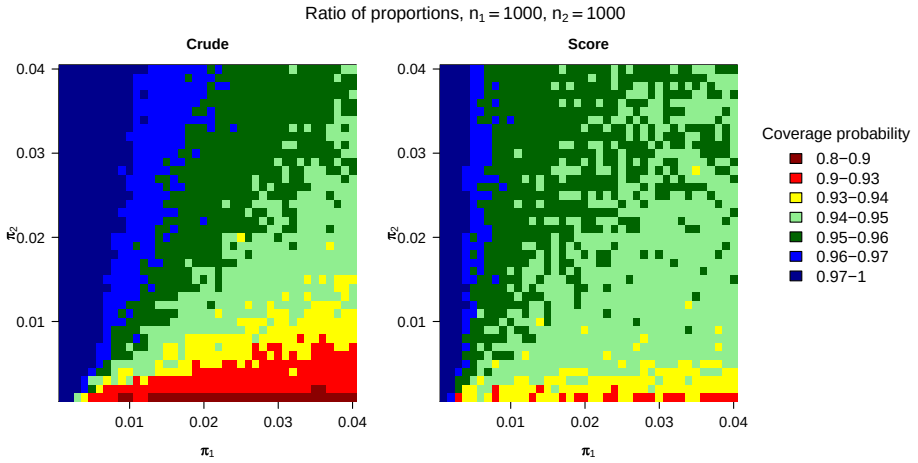

Figure 9: Coverage probability of lower, nominal 0.95 confidence limits for $n_1, n_2 = 1000$

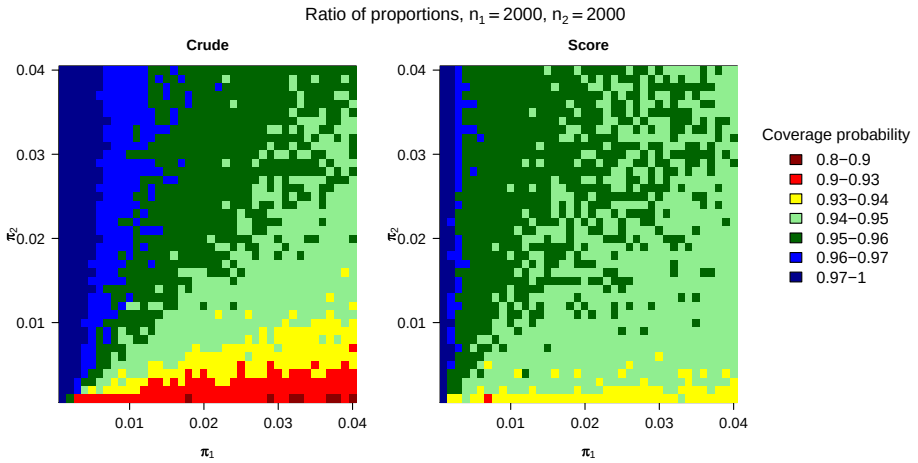

Figure 10: Coverage probability of lower, nominal 0.95 confidence limits for $n_1, n_2 = 2000$

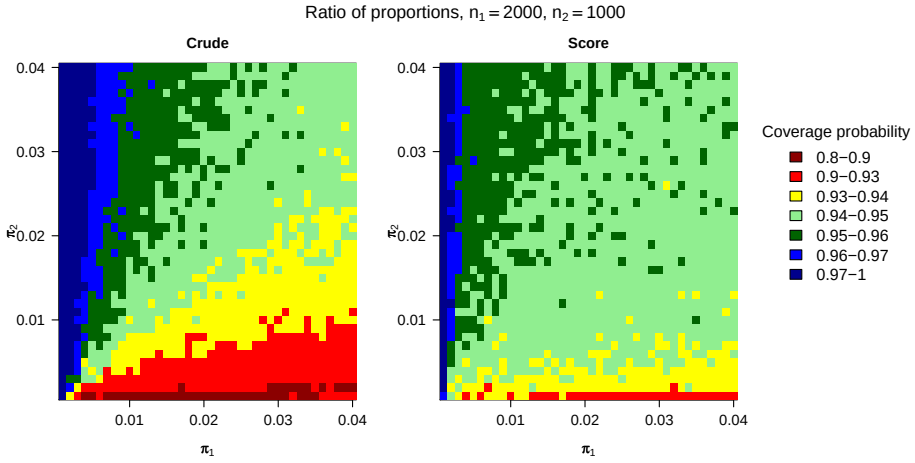

Figure 11: Coverage probability of lower, nominal 0.95 confidence limits for $n_1 = 2000, n_2 = 1000$

3.1.4 Upper 0.95 confidence limits for the risk difference (proof of safety)

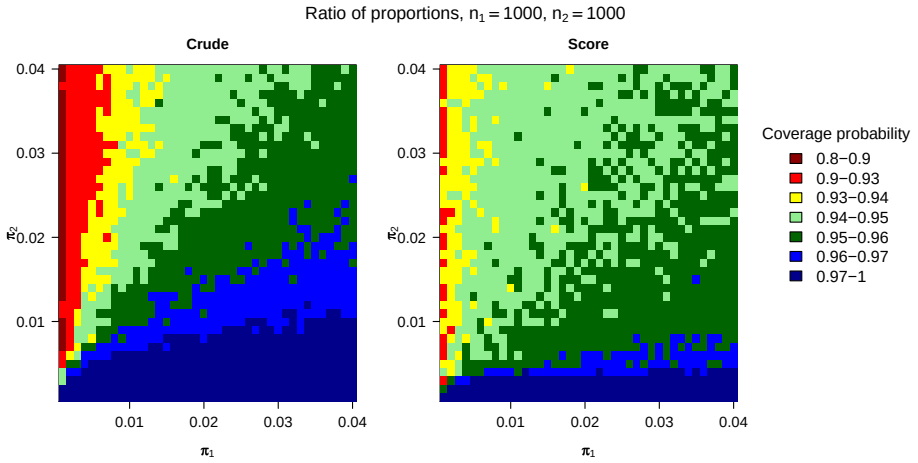

Figure 12: Coverage probability of upper, nominal 0.95 confidence limits for $n_1, n_2 = 1000$

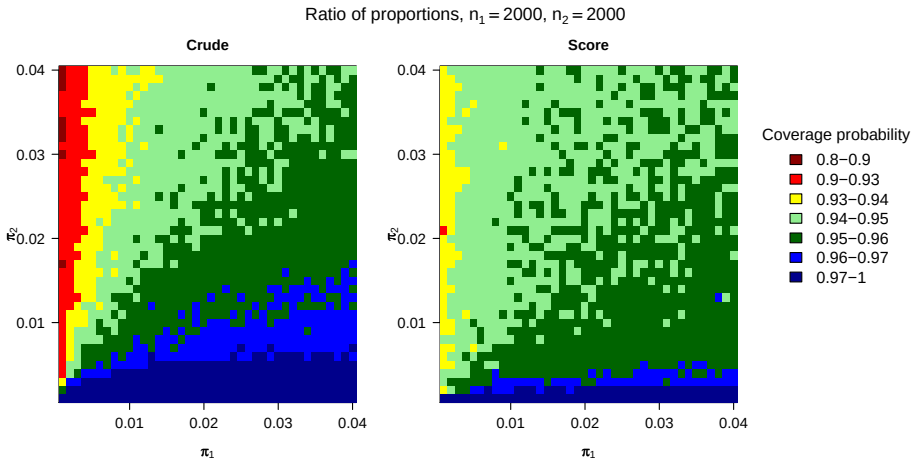

Figure 13: Coverage probability of upper, nominal 0.95 confidence limits for $n_1, n_2 = 2000$

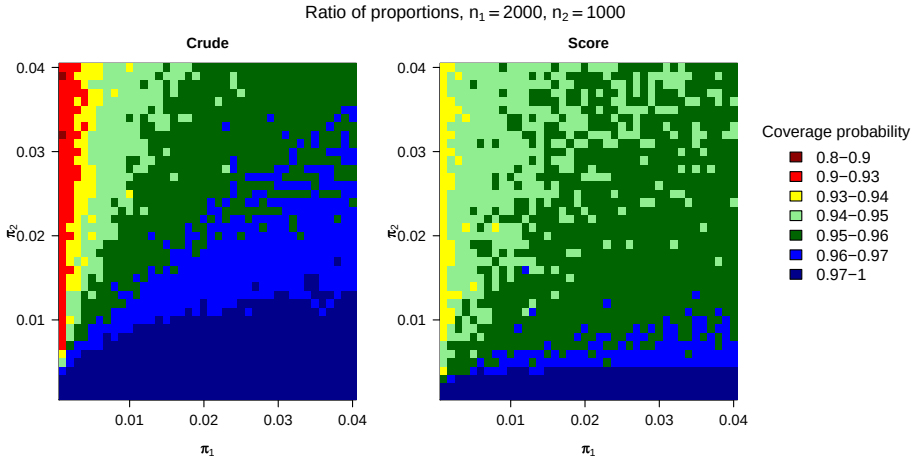

Figure 14: Coverage probability of upper, nominal 0.95 confidence limits for $n_1 = 2000, n_2 = 1000$

3.2 Lower confidence limits for the odds ratio (proof of hazard)

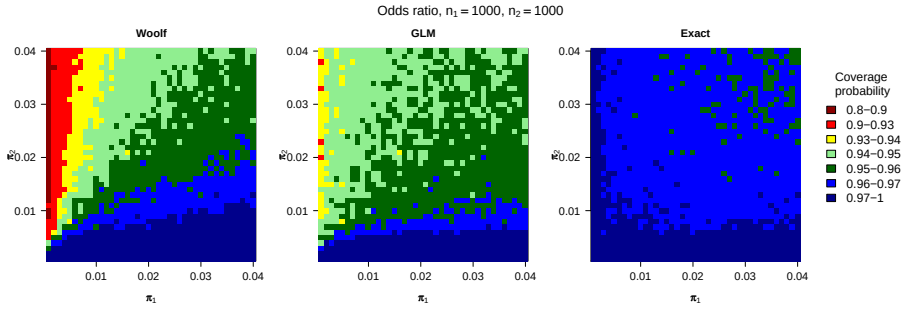

Figure 15: Coverage probability of lower, nominal 0.95 confidence limits for $n_1, n_2 = 1000$

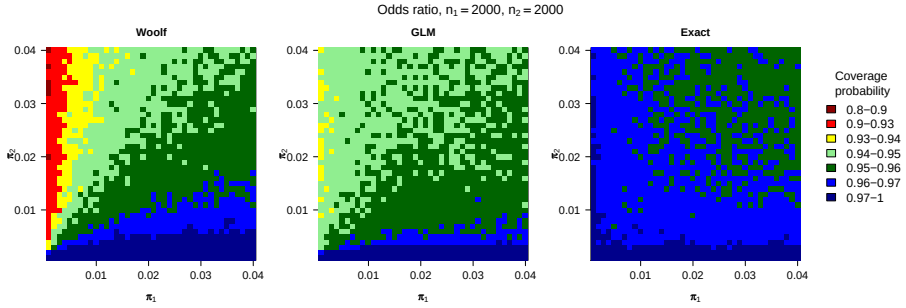

Figure 16: Coverage probability of lower, nominal 0.95 confidence limits for $n_1, n_2 = 2000$

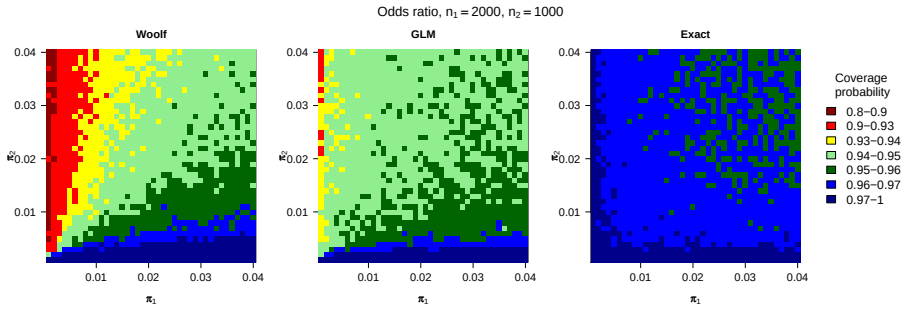

Figure 17: Coverage probability of lower, nominal 0.95 confidence limits for $n_1 = 2000, n_2 = 1000$

3.3 Upper confidence limits for the odds ratio (proof of safety)

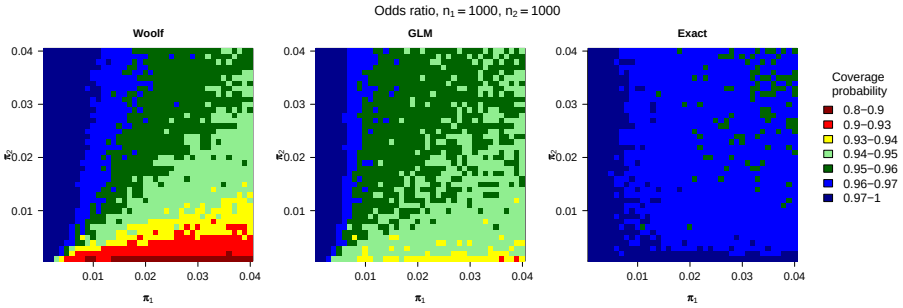

Figure 18: Coverage probability of upper, nominal 0.95 confidence limits for $n_1, n_2 = 1000$

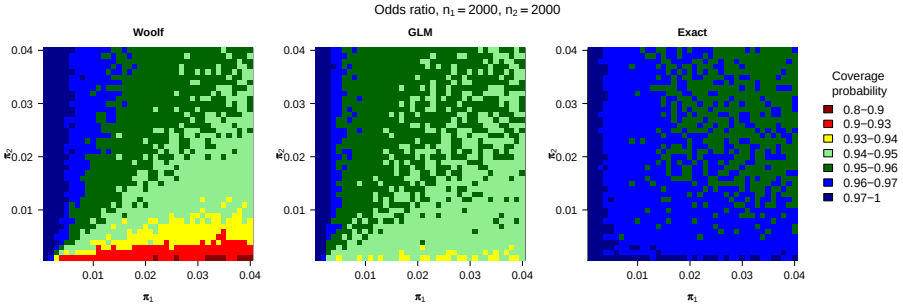

Figure 19: Coverage probability of upper, nominal 0.95 confidence limits for $n_1, n_2 = 2000$

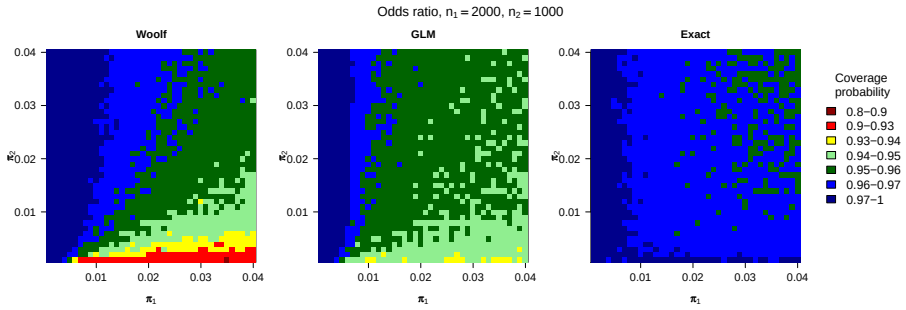

Figure 20: Coverage probability of upper, nominal 0.95 confidence limits for $n_1 = 2000, n_2 = 1000$

3.4 Power for the proof of hazard

Testing the null hypothesis $H_0 : \pi_2 - \pi_1 \leq 0$, vs. the alternative $H_1 : \pi_2 - \pi_1 > 0$ is a proof of hazard when π_i specifies the probability of a detrimental event. The five confidence intervals above can be used to decied on such hypotheses. The alternative can be rejected with approximately 0.05 error probability, if the value 0 is excluded by the lower (0.95)-limit for δ. Here, the probability to reject H_0 under different settings is estimated. As a standard, the probability to reject H_0 using Fishers exact test is additionally estimated (bold, solid line). Note that, due to the discreteness of the binomial distribution, there is a limited number of events $\{y_1, y_2\}$ that may lead to the rejection of $H_0 : \delta = \delta_0$ for given n_1, n_2. Therefore, power of two methods can be exactly equal for one choice of n_1, n_2, δ_0 and different for another choice of n_1, n_2, δ_0.
In all considered cases, the NHS and Wald confidence intervals have higher power than Fishers exact test. The Add4 and Add2 intervals have equal power as Fishers exact test in one case and higher power in all others. The CC confidence interval has usually the same power as the Fisher test, but occasionally has even lower power.

The difference of power between the NHS confidence interval and the Fisher test was 0.12 in the most pronounced case.

- Sample size $n_1 = n_2 = 1000, 2000$, $n_1 = 2000$, $n_2 = 1000$;
- all combinations of $\pi_1 = 0.001, 0.005, 0.01$, and $\pi_2 = \rho\pi_1$, $\rho = 1, 1.1, ..., 4$;
- lower (0.95) limits, and one-sided 0.05 Fishers test, respectively;
- 10000 simulation runs.

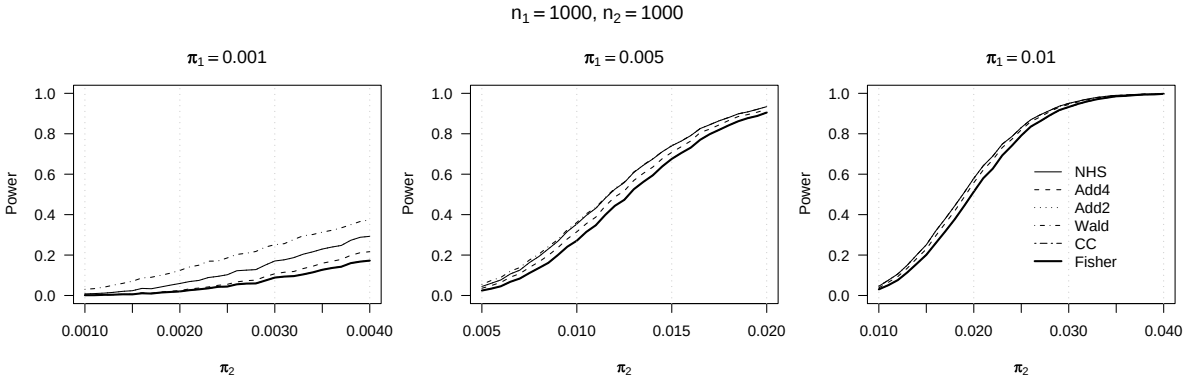

Figure 21: Power to reject $H_0 : \pi_2 - \pi_1 \leq 0$ with type-I-error = 0.05, $n_1 = 2000, n_2 = 1000$

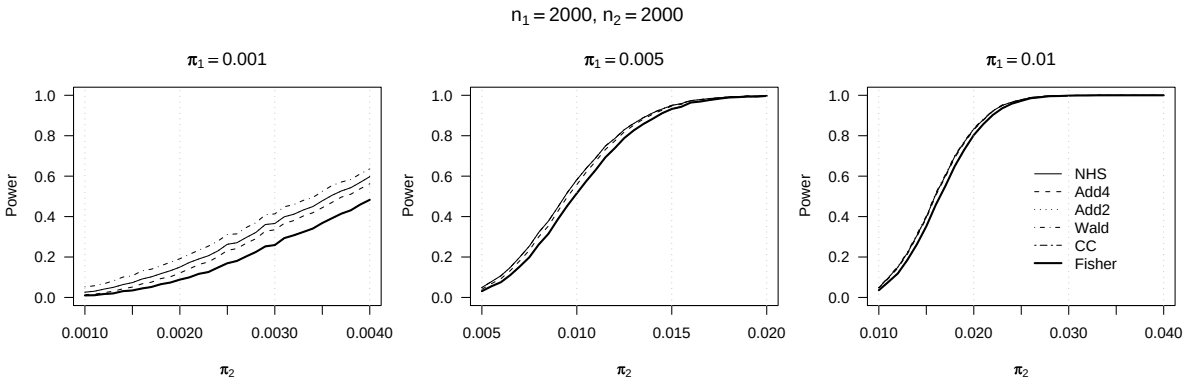

Figure 22: Power to reject $H_0 : \pi_2 - \pi_1 \leq 0$ with type-I-error = 0.05, $n_1 = 2000, n_2 = 1000$

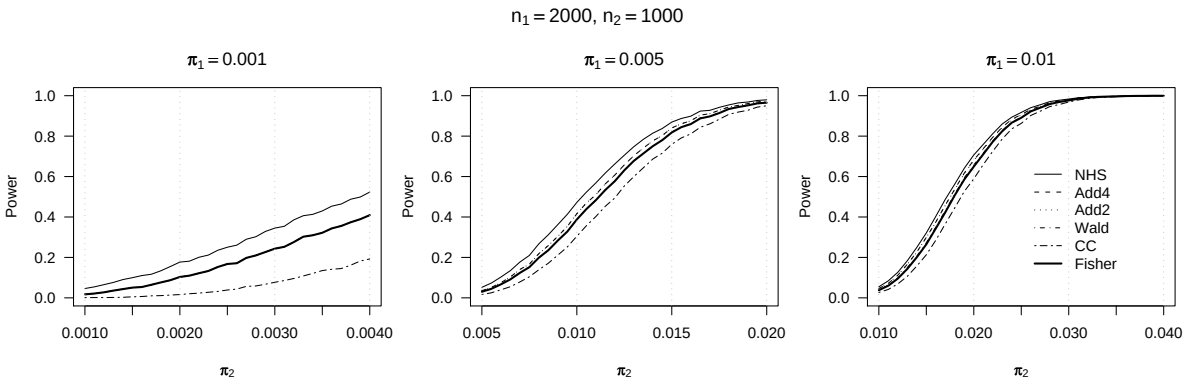

Figure 23: Power to reject $H_0 : \pi_2 - \pi_1 \leq 0$ with type-I-error = 0.05, $n_1 = 2000, n_2 = 1000$

4 Recommendation

4.1 Odds ratio

When applied to perform a proof of safety and considering the problems of non-informative bounds of the GLM method, the Exact confidence is the recommended method, since conservative performance is acceptable in the situation. However, when confidence intervals are applied in a proof of hazard with inverse interpretation (i.e., concluding for safety, when the null hypothesis of a Proof of Hazard can not be rejected), the Crude method is recommended, since in this situation, conservative performance is not acceptable (here, aim is 'being confident in negative results').

4.2 Risk difference

For the considered parameter combinations of very small proportions of success ($\pi_i = 0.001, ..., 0.04$) and large sample sizes ($n_i > 1000$), the NHS method achieves coverage probabilities closest to the nominal level, and avoids severe violations of the nominal level in all considered cases. All other methods under comparison do show more conservative performance for $\pi_2 - \pi_1 > 0$. The Add4, Add2, and Wald show liberal performance for $\pi_2 - \pi_1 < 0$. Hence, the NHS method is the best among the considered confidence intervals. This recommendation is restricted to lower (0.95)-bounds $\pi_2 - \pi_1$. Note further, that the NHS method might lead to liberal confidence limits in the case $\pi_2 >> 0.5 \cap \pi_1 << 0.5$ which are not of interest here (Sill (2007); Schaarschmidt et al. (submitted)).
For the rejection of the null hypothesis $H_0 : \pi_2 - \pi_1 \leq 0$ the NHS interval has higher power than the Fisher exact test in all considered cases.

4.3 Risk ratio

The Score method in its current implementation clearly outperforms the Crude method, with confidence limits closer to the nominal level, showing conservative as well as liberal performance in a smaller proportion of cases. In a proof of hazard for increasing rates $\pi_2/\pi_1 > 1$, both methods are conservative, when the proportion in the control group, π_1, is very small. When $n_1\pi_1 > 4$, the Score methods avoids severely conservative performance. In a proof of safety using upper bounds, the Score method is conservative when π_2 is very small. As a rule of thumb, when $n_2\pi_2 > 4$, severely conservative performance is avoided. The Score method is liberal, when $\pi_2 >> \pi_1$ AND $n_1\pi_1 < 4$. This situation is not of high relevance in a proof of safety, with commonly used safety margins ($\rho_0 = 1.11, 1.25, 2$).

References

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician* 54:280-288.

Agresti, A. and Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency tables. *Biometrics* 61: 515-523.

Brown, L. and Li, X. (2005). Confidence intervals for two sample binomial distributions. *Journal of Statistical Planning and Inference* 130:359-375.

Cai, T.T. (2005). One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference* 131: 63-88.

Clarkson, D. B., Fan, Y. and Joe, H. (1993). A Remark on Algorithm 643: FEXACT: An Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables. *ACM Transactions on Mathematical Software* 19: 484–488.

Dann, R.S., and Koch, G.G. (2005): Review and Evaluation of Methods for Computing Confidence Intervals for the Ratio of Two Proportions and Considerations for Non-inferirority Clinical Trials. Journal of Biopharmaceutical Statistics, 15:85-107.

Gart, J.J. and Nam, J. (1988): Approximate Interval Estimation of the Ratio of Binomial Parameters: A Review and Corrections for Skewness. Biometrics 44, 323-338.

Gerhard, D. (2007): CI for Odds Ratios accounting for Extra Variation between Replicated Experiments. Report of the Institute of Biostatistics No 10 / 2007, Natural Sciences Faculty, Leibniz University of Hannover.

Lawson, R. (2004): Small sample confidence intervals for the odds ratio. *Communications in Statistics Simulation and Computation* 33: 1095-1113.

Lecoutre, B. and Faure, S. (2007). A note on new confidence intervals for the difference between two proportions based on the Edgeworth expansion. *Journal of Statistical Planning and Inference* 137: 355-356.

Newcombe, R.G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine* 17:873-890.

Sill, M. (2007). Approximate one-sided two-sample confidence limits for the comparison of a treatment versus a near-zero spontaneous rate in control. Reports of the Institute of Biostatistics No 09 / 2007, Leibniz University of Hannover, Natural Sciences Faculty.

Schaarschmidt, F., Biesheuvel, E. and Hothorn, L.A. (submitted). Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials.

Wellek, S. (2005): Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes. *Biometrical Journal* 47: 48-61.

Wilson, E.B. (1927). Probable inference, the law of succession and statistical inference. *Journal of the American Statistical Association* 22:209-212.

Zhou,X.-H., Tsao, M. and Qin, G. (2004). New intervals for the difference between two independent binomial proportions. *Journal of Statistical Planning and Inference* 123: 97-115.