Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–31

DOI: 10.1177/0962280211415469
smm.sagepub.com
SAGE

# Recommended confidence intervals for two independent binomial proportions

**Morten W Fagerland,[1] Stian Lydersen[2] and Petter Laake[3]**

**Abstract**

The relationship between two independent binomial proportions is commonly estimated and presented using the difference between proportions, the number needed to treat, the ratio of proportions or the odds ratio. Several different confidence intervals are available, but they can produce markedly different results. Some of the traditional approaches, such as the Wald interval for the difference between proportions and the Katz log interval for the ratio of proportions, do not perform well unless the sample size is large. Better intervals are available. This article describes and compares approximate and exact confidence intervals that are – with one exception – easy to calculate or available in common software packages. We illustrate the performances of the intervals and make recommendations for both small and moderate-to-large sample sizes.

**Keywords**

$2 \times 2$ table, NNT, odds ratio, relative risk, risk difference

## 1 Introduction and notation

A frequent task in medical statistics is to compare two independent binomial proportions. It occurs both in experimental trials and in observational studies when a dichotomous variable is compared in two independent samples. The dichotomous variable is often the occurrence of an event, for example in randomized controlled trials (RCTs) and cohort studies, where the event may be the primary outcome of interest. In observational studies, some patient characteristics, such as the presence of certain diseases, are dichotomous, and the proportions of diseased patients are compared between exposed and unexposed groups. Another example is unmatched case–control studies where the proportions of exposed subjects are compared between cases and controls.

[1]Unit of Biostatistics and Epidemiology, Oslo University Hospital, Norway.
[2]Department of Neuroscience, Norwegian University of Science and Technology, Trondheim, Norway.
[3]Department of Biostatistics, University of Oslo, Norway.

**Corresponding author:**
Morten W Fagerland, Unit of Biostatistics and Epidemiology, Oslo University Hospital, Norway.
Email: morten.fagerland@medisin.uio.no

The two possible outcomes of a dichotomous variable are often referred to as success and failure. The successes do not necessarily indicate a favourable outcome, but rather the outcome of interest, which is the outcome we count. We may summarize the outcomes of two independent groups in a $2 \times 2$ table (Table 1). The number of subjects in each group ($n_{1+}$ and $n_{2+}$) is assumed to be fixed by the design. We further assume that the subjects in group 1 have probability of success equal to $p_1$, and that the subjects in group 2 have probability of success equal to $p_2$. This design is usually referred to as the one margin fixed design, which is the basic design for RCTs, cohort studies and case–control studies. Other designs may also be summarized in a $2 \times 2$ table, most notably the both margin fixed design – hardly ever used in practice – and the total number fixed design, which is used in cross-sectional studies. Neither of these two designs leads to two binomial samples. However, $2 \times 2$ tables from cross-sectional studies are often analysed conditionally, as if they were sampled under the one margin fixed design.

The number of successes in group 1 is binomially distributed with parameters $n_{1+}$ and $p_1$. In a similar manner, the number of successes in group 2 is binomially distributed with parameters $n_{2+}$ and $p_2$. We shall estimate the parameters $p_1$ and $p_2$ by the sample proportions

$$\hat{p}_1 = \frac{n_{11}}{n_{1+}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{21}}{n_{2+}}, \tag{1}$$

which are the maximum likelihood estimates. Let $\boldsymbol{n} = \{n_{11}, n_{12}, n_{21}, n_{22}\}$ denote the observed table, and $z_{\alpha/2}$ the upper $\alpha/2$ percentile of the standard normal distribution. For example, if $\alpha = 0.05$, $z_{\alpha/2} = 1.96$.

The purpose of this article is to describe and recommend two-sided $100(1 - \alpha)\%$ confidence intervals for the difference between proportions (Section 4), the number needed to treat (NNT; Section 5), the ratio of proportions (Section 6) and the odds ratio (OR; Section 7). We do not aim to present a systematic or complete review of all available intervals, but we shall consider the most commonly used and recommended ones. Special attention will be paid to simple intervals, i.e. intervals that can be easily explained and computed or intervals that are readily available in common software packages. We do not consider intervals that have not been thoroughly evaluated – even though they may show initial promise – but some are mentioned and referenced. There are many more measures of relationship between binomial proportions than the four presented here. We refer to Hamilton[1] for an overview.

In Section 2, we present the data from a RCT of epinephrine in children with cardiac arrest, which are subsequently used to illustrate the difference between the confidence intervals we consider in Sections 4–7. Section 3 outlines our criteria for how we compare different confidence intervals. Recommendations are given in Section 8.

Our focus here is estimation of confidence interval. Hypothesis tests for association in $2 \times 2$ tables have been described and recommended by Lydersen et al.[2]

**Table 1.** The observed counts of a $2 \times 2$ table

| | Success | Failure | Sum |
|---|---|---|---|
| Group 1 | $n_{11}$ | $n_{12}$ | $n_{1+}$[a] |
| Group 2 | $n_{21}$ | $n_{22}$ | $n_{2+}$[a] |
| Sum | $n_{+1}$ | $n_{+2}$ | $N$[a] |

[a]Fixed by design.

## 2 Example: epinephrine in children with cardiac arrest

Children who remain in cardiac arrest after cardiopulmonary resuscitation are administered with an initial standard dose of epinephrine. If resuscitation is unsuccessful, should the next dose be the same dose or a higher dose? Perondi *et al.*[3] randomized 34 patients to receive a high dose and 34 patients to receive the standard dose of epinephrine. The primary outcome measure was survival 24 h after cardiac arrest. The results are displayed in Table 2. The authors estimated the OR for death with the high-dose therapy to be 8.6 with 95% confidence interval from 1.0 to 397.0 and $p = 0.05$.

Based on these results (and similar results from logistic regression), Perondi *et al.* suggest that 'high-dose therapy may be worse than standard-dose therapy'. The $p$-value given above was calculated using Fisher's exact test ($p = 0.054$).

It has previously been shown that better tests than Fisher's exact test are available, such as exact unconditional tests ($p = 0.028$) and conditional mid-$p$ tests ($p = 0.030$).[2] As is the case with test statistics, different methods of calculating confidence intervals can also give markedly different results, particularly when we consider the lengths of the intervals.

## 3 Criteria for comparing confidence intervals

The main property for evaluating the performance of a confidence interval is coverage probability. The coverage probability is the probability that the confidence interval contains the true value. We prefer this probability to be close to $1 - \alpha$, the nominal coverage probability, usually set to 95%. Exact confidence intervals are required to have coverage probability at least the nominal size, whereas approximate confidence intervals satisfy no such criterion. An interval that has too large coverage probability is denoted as conservative. A liberal interval has coverage probability below $1 - \alpha$.

The coverage probability can be calculated exactly. Under the one margin fixed design, the coverage probability can be expressed as

$$\mathrm{CP} = \sum_{x_{11}=0}^{n_{1+}} \sum_{x_{21}=0}^{n_{2+}} f(\boldsymbol{x}|\, p_1, p_2) \cdot I(\boldsymbol{x}|\, p_1, p_2), \tag{2}$$

where $\boldsymbol{x}$ denotes the table $\{x_{11}, n_{1+} - x_{11}, x_{21}, n_{2+} - x_{21}\}$, $f$ the probability of observing $\boldsymbol{x}$ given $p_1$ and $p_2$ and $I$ an indicator function that equals 1 if the confidence interval for table $\boldsymbol{x}$ includes the true value – as defined by $p_1$ and $p_2$ – else it is 0. In short, the coverage probability for a given interval and point in the parameter space is the sum of the probability of all possible tables having confidence limits that enclose the true value.

**Table 2.** The results of a RCT of epinephrine in children with cardiac arrest[3]

| | Survival at 24 h | | |
|---|---|---|---|
| Treatment | Yes | No | Sum |
| Standard dose | 7 | 27 | 34[a] |
| High dose | 1 | 33 | 34[a] |
| Sum | 8 | 60 | 68[a] |

[a]Fixed by design.

If two or more intervals have similar coverage probabilities, we can compare their lengths, and we prefer the shorter one. Note that an interval with low coverage probability is usually shorter than an interval with large coverage probability.

In our presentation of confidence intervals (Sections 4.2, 6.2 and 7.2), we note whether the intervals produce non-sensical or uninformative results. Overshoot happens when one or both confidence limits lie outside the permissible range of the measure. The difference between proportions, for example, is limited to the range $[-1, 1]$. Uninformative intervals can be either of zero-width, for example (1, 1), or with incomputable limits.

In cases where an interval has incomputable limits, we set the limits to the entire range of the measure – e.g. $(0, \infty)$ for the ratio of proportions – such that the probabilities of those tables are included in the calculations of the coverage probability (Equation (2)).

Other properties we consider beneficial for a confidence interval are simplicity, availability in software packages, consistency with tests and symmetry of coverage. Detailed criteria for evaluating the confidence intervals that go beyond the ones presented here have been set out by Newcombe.[4,5]

# 4 The difference between proportions

## 4.1 Introduction and estimate

The difference between proportions, or success probabilities,

$$\Delta = p_1 - p_2,$$

is an important effect measure for RCTs and cohort studies. In addition to its inherent value as an effect measure, its estimate and confidence interval are used to derive the estimate and confidence interval of the NNT (Section 5). The difference between proportions is also called the probability difference, or when the event in question is a harmful one, the risk difference. In epidemiology, it is often called the absolute risk reduction or the attributable risk (reduction). We estimate the difference between proportions using the sample proportions:

$$\hat{\Delta} = \hat{p}_1 - \hat{p}_2 = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}. \tag{3}$$

## 4.2 Confidence intervals

### 4.2.1 Wald

The traditional Wald confidence interval for $\Delta$ is based on the asymptotic normal distribution of $\hat{\Delta}$:[6]

$$\hat{\Delta} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_{1+}} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_{2+}}}. \tag{4}$$

The Wald interval has zero-width when (1) $n_{11} = n_{21} = 0$ or $n_{12} = n_{22} = 0$, which gives the interval (0, 0); (2) $n_{11} = n_{22} = 0$, which gives the interval $(-1, -1)$; and (3) $n_{12} = n_{21} = 0$, which gives the interval (1, 1).

A continuity corrected version due to Yates[7] can be expressed, as shown in Equation (5):[8]

$$\hat{\Delta} \pm \left[ z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_{1+}} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_{2+}}} + \frac{1}{2}\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)\right]. \tag{5}$$

The Wald interval with continuity correction avoids zero-width but has a higher overshoot rate.[5] Overshoot (intervals outside [−1, 1]) is also possible with the Agresti–Caffo interval (Section 4.2.2). The problem of overshoot can be easily eliminated by truncation for Wald, Wald with continuity correction and Agresti–Caffo, although the resulting interval may not be entirely satisfactorily.

### 4.2.2 Agresti–Caffo

Agresti and Caffo[9] proposed a simple, yet effective procedure for computing a confidence interval: add one success and one failure in each sample and calculate the Wald confidence interval on the resulting data:

$$\breve{p}_1 - \breve{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\breve{p}_1(1-\breve{p}_1)}{\breve{n}_{1+}} + \frac{\breve{p}_2(1-\breve{p}_2)}{\breve{n}_{2+}}}, \tag{6}$$

where

$$\breve{n}_{1+} = n_{1+} + 2, \quad \breve{n}_{2+} = n_{2+} + 2, \quad \breve{p}_1 = (n_{11} + 1)/\breve{n}_{1+}, \quad \breve{p}_2 = (n_{21} + 1)/\breve{n}_{2+}.$$

Note that our estimate of the difference between proportions is still given by the difference in sample proportions (Equation (3)), the calculations of $\breve{p}_1$ and $\breve{p}_2$ go only into the calculations of the confidence interval.

The Agresti–Caffo interval is usually consistent with the results of the much-used Pearson's chi-squared test,[9] but it can produce overshoot (Section 4.2.1). We note that adjustments, such as adding values to the observed counts – as it is done in the Agresti–Caffo interval – are discouraged on general principles by some authors, for instance Hirji[10] (p. 78).

### 4.2.3 Newcombe hybrid score

Newcombe[5] proposed a confidence interval for $\Delta$ based on the Wilson[11] score confidence interval for a single proportion. We calculate the intervals for $p_1$ and $p_2$ by

$$\hat{p}_i\left(\frac{n_{i+}}{n_{i+} + z^2_{\alpha/2}}\right) + \frac{1}{2}\left(\frac{z^2_{\alpha/2}}{n_{i+} + z^2_{\alpha/2}}\right) \pm z^2_{\alpha/2}\sqrt{\frac{1}{n_{i+} + z^2_{\alpha/2}}\left[\hat{p}_i(1-\hat{p}_i)\left(\frac{n_{i+}}{n_{i+} + z^2_{\alpha/2}}\right) + \frac{1}{4}\left(\frac{z^2_{\alpha/2}}{n_{i+} + z^2_{\alpha/2}}\right)\right]} \quad \text{for} \quad i = 1, 2.$$

Denote the interval for $p_1$ by $(l_1, u_1)$ and the one for $p_2$ by $(l_2, u_2)$. The Newcombe hybrid score confidence interval for $\Delta$ is given by

$$\hat{\Delta} - \sqrt{(\hat{p}_1 - l_1)^2 + (u_2 - \hat{p}_2)^2} \quad \text{to} \quad \hat{\Delta} + \sqrt{(\hat{p}_2 - l_2)^2 + (u_1 - \hat{p}_1)^2}. \tag{7}$$

### 4.2.4 Miettinen–Nurminen asymptotic score

We obtain a score interval by inverting two one-sided $\alpha/2$-level score tests (the tail method), or one two-sided $\alpha$-level score test. For a specified value $\Delta_0 \in (-1, 1)$, the score test statistic is

$$T(\boldsymbol{n}|\Delta_0) = \frac{\hat{p}_1 - \hat{p}_2 - \Delta_0}{\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1+}} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2+}}}}, \tag{8}$$

where $\boldsymbol{n}$ denotes the observed table $\{n_{11}, n_{12}, n_{21}, n_{22}\}$ and $\tilde{p}_1$ and $\tilde{p}_2$ the maximum likelihood estimates of $p_1$ and $p_2$ subject to $p_1 - p_2 = \Delta_0$. An asymptotic confidence interval based on inverting two one-sided score tests (Equation (8)) was first proposed by Mee.[12] Miettinen and Nurminen[13] suggested a similar interval based on the test statistic

$$T(\boldsymbol{n}|\Delta_0)_{\mathrm{MN}} = \frac{\hat{p}_1 - \hat{p}_2 - \Delta_0}{\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1+}} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2+}}}} \cdot \sqrt{1 - \frac{1}{n_{1+} + n_{2+}}}. \tag{9}$$

The correction term in Equation (9) makes a difference only for small sample sizes. The Miettinen–Nurminen asymptotic score confidence interval $(L, U)$ for $\Delta$ is based on Equation (9) and obtained by solving

$$T(\boldsymbol{n}|L)_{\mathrm{MN}} = -z_{\alpha/2} \tag{10}$$

and

$$T(\boldsymbol{n}|U)_{\mathrm{MN}} = z_{\alpha/2}. \tag{11}$$

Miettinen and Nurminen showed that the restricted maximum likelihood estimates ($\tilde{p}_1$ and $\tilde{p}_2$) can be obtained by solving a cubic equation and gave unique closed-form expressions for them. We refer to Ref. 13 and to the almost equal expressions in Farrington and Manning[14] for details.

### 4.2.5 Exact unconditional intervals

In the previous section, we inverted two asymptotic tests to obtain an asymptotic score interval. Equation (8) can also be used to construct exact tests, which can be inverted to obtain exact unconditional score intervals. Under the restriction $p_1 - p_2 = \Delta_0$, the domain of $p_1$ given $\Delta_0$ is

$$I(\Delta_0) = \{p_1 : \max(0, \Delta_0) \leq p_1 \leq \min(1, 1 + \Delta_0)\}. \tag{12}$$

Let $\boldsymbol{x} = \{x_{11}, x_{12}, x_{21}, x_{22}\}$ denote any $2 \times 2$ table that might be observed given the fixed row sums. The probability of observing $\boldsymbol{x}$ is the product of the likelihoods for the number of successes in the two samples:

$$f(\boldsymbol{x}|\, p_1, \Delta_0) = \binom{x_{1+}}{x_{11}} p_1^{x_{11}}(1-p_1)^{x_{12}} \times \binom{x_{2+}}{x_{21}} (p_1 - \Delta_0)^{x_{21}}(1 - p_1 + \Delta_0)^{x_{22}}.$$

*Inverting two one-sided score tests (Chan–Zhang).* The interval by Chan and Zhang[15] is based on inverting two one-sided exact score tests of size at most $\alpha/2$ (the tail method). Define

$$P\big(T(\boldsymbol{n})|\, p_1, \Delta_0\big) = \sum_{T(\boldsymbol{x}) \leq T(\boldsymbol{n})} f(\boldsymbol{x}|\, p_1, \Delta_0)$$

and

$$Q\big(T(\boldsymbol{n})|\, p_1, \Delta_0\big) = \sum_{T(\boldsymbol{x}) \geq T(\boldsymbol{n})} f(\boldsymbol{x}|\, p_1, \Delta_0).$$

where $T(\boldsymbol{n})$ refers to the score test statistic in Equation (8). We eliminate the nuisance parameter $p_1$ by taking the supremum over the range $I(\Delta_0)$ given in Equation (12):

$$P\big(T(\boldsymbol{n})|\Delta_0\big) = \sup_{p_1 \in I(\Delta_0)} P\big(T(\boldsymbol{n})|\, p_1, \Delta_0\big) \tag{13}$$

and

$$Q\big(T(\boldsymbol{n})|\Delta_0\big) = \sup_{p_1 \in I(\Delta_0)} Q\big(T(\boldsymbol{n})|\, p_1, \Delta_0\big). \tag{14}$$

The Chan–Zhang confidence interval $(L, U)$ for $\Delta$ is the solution of

$$Q\big(T(\boldsymbol{n})|L\big) = \alpha/2 \tag{15}$$

and

$$P\big(T(\boldsymbol{n})|U\big) = \alpha/2. \tag{16}$$

*Inverting one two-sided score test (Agresti–Min).* Instead of inverting two one-sided tests, Agresti and Min[16] proposed to invert one two-sided test of size at most $\alpha$. Using the score test statistic in Equation (8), we define

$$R\big(T(\boldsymbol{n})|\, p_1, \Delta_0\big) = \sum_{|T(\boldsymbol{x})| \geq |T(\boldsymbol{n})|} f(\boldsymbol{x}|\, p_1, \Delta_0),$$

and eliminate the nuisance parameter $p_1$ by maximizing over all possible values, i.e.

$$R\big(T(\boldsymbol{n})|\Delta_0\big) = \sup_{p_1 \in I(\Delta_0)} R\big(T(\boldsymbol{n})|\, p_1, \Delta_0\big). \tag{17}$$

The Agresti–Min confidence interval $(L, U)$ for $\Delta$ is the solution of

$$R\big(T(\boldsymbol{n})|L\big) = \alpha \tag{18}$$

and

$$R\big(T(\boldsymbol{n})|U\big) = \alpha, \tag{19}$$

such that $R(T(\boldsymbol{n})|\Delta_0) < \alpha$ when $\Delta_0 < L$ and $R(T(\boldsymbol{n})|\Delta_0) < \alpha$ when $\Delta_0 > U$.

*Inverting two one-sided unstandardized tests (Santner–Snell).* Santner and Snell[17] used the tail method (see the description of the Chan–Zhang interval) with the unstandardized difference between proportions,

$$T(\boldsymbol{n})_{SS} = \hat{p}_1 - \hat{p}_2, \tag{20}$$

as test statistic. We include the Santner–Snell interval in our selection of intervals for the difference between proportions because it is available in the much-used software package SAS (SAS Institute Inc.).

*The Berger and Boos procedure.* The procedure by Berger and Boos[18] is a general approach to reduce the conservatism of exact unconditional methods. Instead of maximizing over the entire range of the nuisance parameter, as it is done, for example, in Equations (13), (14) and (17), the

maximization is done over a restricted range of values. This range is taken to be a $100(1-\gamma)\%$ confidence interval for the nuisance parameter, where $\gamma$ is very small, say 0.0001.

## 4.3 Comparisons of intervals

The comparisons in this section are based on exact calculations of coverage probabilities (Section 3). Several sample sizes and combinations of $p_1$ and $p_2$ were used. For the figures shown – which are used as a starting point for the discussion on interval performance – we selected sample sizes and fixed values of $\Delta$ in a manner that serves to illustrate typical differences and similarities between the intervals. In addition, we sometimes refer to calculated coverage probabilities that are not shown in the figures. The results from our calculations are consistent with those published in previous papers.

### 4.3.1 Approximate intervals

Figure 1 shows the coverage probabilities of four of the five approximate intervals presented in Section 4.2 for one small sample size and one medium sample size. We did not include Wald with continuity correction in these plots as this interval is generally too conservative, even for large sample sizes, to consider it further.[5,9] For small sample sizes, such as $n_{1+}=n_{2+}=10$, the Newcombe hybrid score and the Miettinen–Nurminen asymptotic score intervals perform generally well; both these intervals have coverage probabilities close to the nominal level of 95%, but they can be liberal. The Agresti–Caffo interval is slightly conservative with a mean coverage probability of 97% for the sample size and parameter values shown in the left panel of Figure 1. The Wald interval is seriously liberal with coverage probabilities in the range 92–94% for most situations, and even lower coverage probabilities for proportions below 5% or above 95%.

With increasing sample size, the coverage probabilities of the Agresti–Caffo, Newcombe hybrid score, and Miettinen–Nurminen asymptotic score intervals become more and more similar (right panel of Figure 1). When $n_{1+}=n_{2+}=40$, the three intervals perform almost equally, except when one or both proportions are close to 0 or 1, where the asymptotic score interval can be somewhat liberal. In contrast, the Wald interval is still quite liberal for most parameter values at

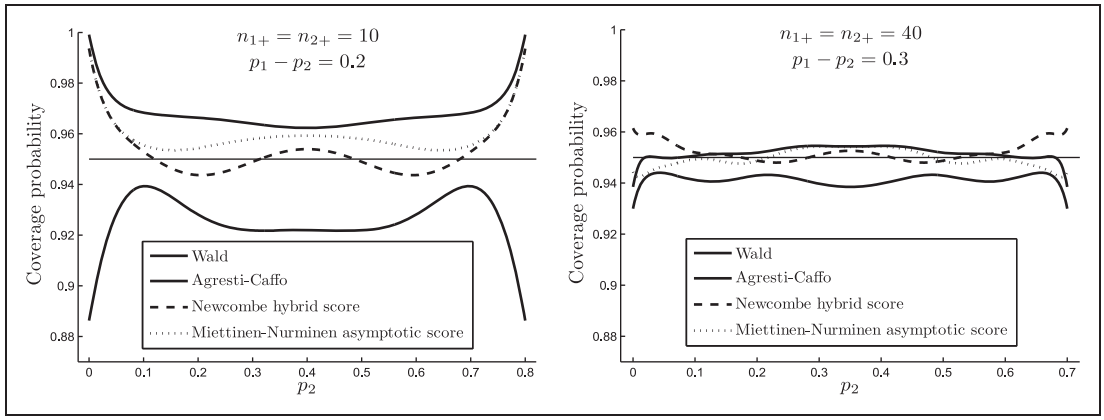


**Figure 1.** Coverage probabilities of the Wald, Agresti–Caffo, Newcombe hybrid score and Miettinen–Nurminen asymptotic score intervals for sample sizes $n_{1+}=n_{2+}=10$ (left) and $n_{1+}=n_{2+}=40$ (right).
Note: The difference between proportions is fixed at 0.2 (left) and 0.3 (right).

$n_{1+} = n_{2+} = 40$, and it does not, in general, achieve approximately nominal coverage probabilities until $n_{1+} = n_{2+} = 100$.

The Agresti–Caffo, Newcombe hybrid score and Miettinen–Nurminen asymptotic score intervals all cope well with unequal sample sizes, for instance, $n_{1+} = 20$, $n_{2+} = 10$ (results not shown). The Wald interval has coverage probabilities further lowered by unbalanced samples.

For small sample sizes, none of the intervals perform particularly well when both proportions are close to either 0 or 1. This improves with increasing sample size. For 30 or more in each sample, the Agresti–Caffo interval performs acceptably. For the Newcombe hybrid score interval, 40 or more in each sample ensures acceptable coverage probabilities. A recent bootstrap interval by Lin *et al.*,[19] based on the median unbiased estimate,[20] is reported to have favourable properties for small sample sizes and when the success probability is close to 0 or 1.

The Newcombe hybrid score and the Miettinen–Nurminen asymptotic score intervals have approximately equal lengths,[5] and both tend to be shorter than the Agresti–Caffo interval.[9]

### 4.3.2 Exact intervals

Figure 2 illustrates the difference in coverage probabilities of the Santner–Snell, Chan–Zhang and Agresti–Min exact unconditional intervals. The unstandardized test statistic used in the Santner–Snell interval can be seriously discrete with small sample sizes, which makes the interval overly conservative. The lower bound on the coverage probability of tail method intervals such as Santner–Snell is sometimes $1 - \alpha/2$ instead of $1 - \alpha$.[21]

The two intervals based on inverting exact score tests sometimes perform similarly, for instance, when $n_{1+} = n_{2+} = 15$ and $p_1 - p_2 = 0.2$ (left panel of Figure 2). However, the Chan–Zhang interval can be seriously conservative, as shown in the right panel of Figure 2. Agresti and Min[16] and Agresti[21] demonstrate that inverting one two-sided test (Agresti–Min interval) is less conservative and gives shorter intervals than inverting two one-sided tests (Chan–Zhang interval). This difference in coverage probabilities is often exaggerated when the sample sizes are unequal,[16] as shown in the right panel of Figure 2.

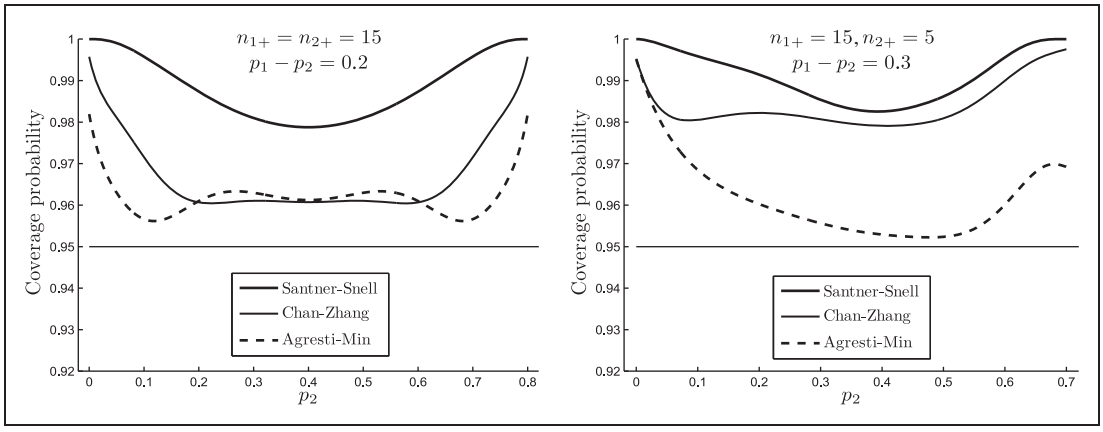


**Figure 2.** Coverage probabilities of the Santner–Snell, Chan–Zhang and Agresti–Min exact unconditional intervals for sample sizes $n_{1+} = n_{2+} = 15$ (left) and $n_{1+} = 15$, $n_{2+} = 5$ (right).
Note: The difference between proportions is fixed at 0.2 (left) and 0.3 (right).

Intervals based on inverting two one-sided exact tests are defined so that the non-coverage probability in each tail is no more than $\alpha/2$, whereas intervals that invert one two-sided exact test are constructed so that the sum of the non-coverage probabilities in the two tails is no more than $\alpha$. The only practical disadvantage of using an interval based on a two-sided test is that the endpoints are not consistent with the results of one-sided tests, which are used in studies that aim to show that a new treatment improves upon a standard one. In such studies, Agresti[21] suggests that a one-sided confidence bound is used instead of a confidence interval.

The Berger and Boos procedure has been recommended for the Chan–Zhang and Agresti–Min confidence intervals,[22] however, both Chan and Zhang[15] and Agresti and Min[16] claim that it does not improve the performance of their intervals. Nevertheless, by limiting the range over which the maximization is performed, the procedure may reduce the computation time. Another feature of using the Berger and Boos procedure is that we exclude values of the nuisance parameter that are highly unlikely given the observed data, thereby accommodating one element of the criticism against exact unconditional inference (Agresti,[23] p. 95). In our calculations, we have used the Berger and Boos procedure with $\gamma = 0.000001$ (default in StatXact 9, Cytel Inc.). We note that the Santner–Snell interval is quite sensitive to the value of $\gamma$. For larger values of $\gamma$ than the above, such as 0.001 and 0.0001, and for no Berger and Boos correction ($\gamma = 0$), the Santner–Snell interval can be even more conservative than what is shown in Figure 2. No optimal choice for $\gamma$ has been suggested for exact unconditional intervals, but Lydersen et al.[24] recommend $\gamma = 0.0001$ for the exact unconditional $z$-pooled test.

A detailed comparison of asymptotic and exact score intervals is given in Santner et al.[25] The authors recommend the Coe–Tamhane interval[26] followed by the Agresti–Min and the Chan–Zhang intervals. They do not recommend the asymptotic score interval because the coverage probabilities of that interval was below the nominal level (90%) for nearly half of the studied cases. However, that conclusion has been disputed by Newcombe and Nurminen,[27] who point out that Santner *et al*. used a criterion of strict conservatism – thus favouring exact intervals – and that the asymptotic score interval in that study is the Mee interval, not the Miettinen–Nurminen interval.

### 4.3.3 The best performing approximate and exact intervals

In Figure 3, we compare the three best performing confidence intervals for the difference between proportions using two small sample size combinations. The Agresti–Min exact unconditional interval is generally superior to the other two intervals, particularly when proportions are close to 0 or 1. Its coverage probability is consistently close to the nominal level, whereas the Agresti–Caffo and the Newcombe hybrid score intervals can be moderately conservative and liberal, respectively.

### 4.3.4 Example: data from Table 2

In Section 2, we presented the results from a RCT of high-dose *versus* standard-dose epinephrine in children with cardiac arrest. The observed proportion of survival in the standard-dose group was $\hat{p}_1 = 7/34 = 0.21$, and in the high-dose group, it was $\hat{p}_2 = 1/34 = 0.029$. We estimate the difference between proportions as $\hat{\Delta} = 0.18$. Seven different confidence intervals are shown in Table 3. The Wald interval is shorter than the other intervals, which is not surprising as the coverage probability of the Wald interval is generally below the nominal level. All the intervals give similar results, except for the Santner–Snell interval, which is about 50% wider than the other intervals and

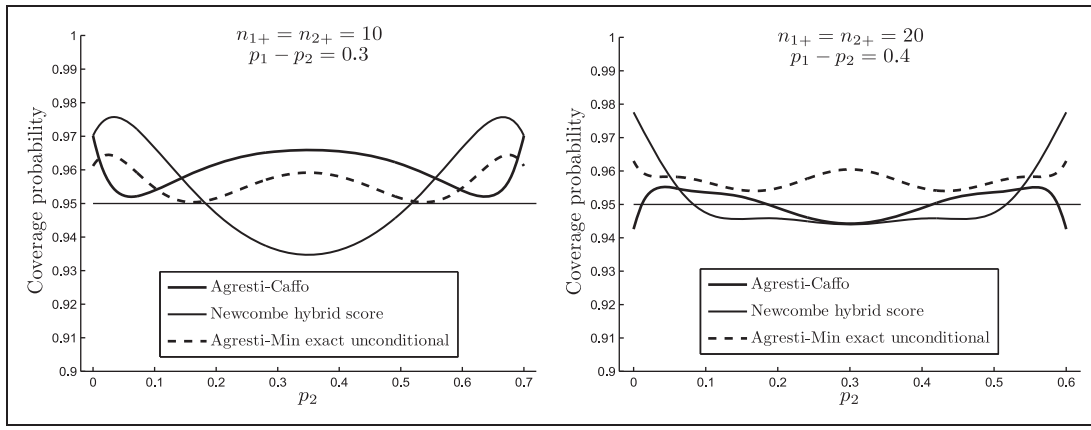


**Figure 3.** Coverage probabilities of the three best performing confidence intervals for the difference between proportions.

**Table 3.** Confidence intervals for the difference between proportions using data from Table 2

| | Confidence interval | | |
|---|---|---|---|
| | Lower | Upper | Length |
| Wald | 0.029 | 0.32 | 0.29 |
| Agresti–Caffo | 0.012 | 0.32 | 0.31 |
| Newcombe hybrid score | 0.019 | 0.34 | 0.32 |
| Miettinen–Nurminen asymptotic score | 0.028 | 0.34 | 0.31 |
| Santner–Snell exact unconditional | −0.069 | 0.41 | 0.48 |
| Chan–Zhang exact unconditional | 0.019 | 0.36 | 0.34 |
| Agresti–Min exact unconditional | 0.024 | 0.35 | 0.33 |

Note: The estimate is $\hat{\Delta} = 0.18$.

the only one to include zero. The other two exact intervals (Chan–Zhang and Agresti–Min) are slightly wider than the approximate intervals. The results indicate a reduced survival with high-dose therapy.

## 5 The NNT

The NNT is a useful way of summarizing the results from studies of two treatments or exposures.[28] When presented with a proper confidence interval, the NNT is particularly suitable for clinical decision making, as it incorporates both statistical and clinical significance by dealing with numbers of patients rather than probabilities.[29] This view is, however, not universally acknowledged, and some authors argue that the NNT can be confusing, particularly for clinicians.[30]

As the name indicates, it is most often used in clinical trials of two treatments, but it may also be used in observational studies, where the NNT is sometimes called the number needed to be exposed.[31] In the following, we shall explain and illustrate the NNT in the context of two competing treatments: one new (group 1) *versus* one standard (group 2) treatment.

The NNT is the number of patients that would have to be treated with a new treatment instead of a standard treatment for one additional patient to benefit. It can be estimated by the reciprocal of the difference between proportions:

$$\text{NNT} = \frac{1}{\hat{p}_1 - \hat{p}_2}.$$

We have now assumed that a positive value of $\hat{p}_1 - \hat{p}_2$, and thereby a positive value of NNT, indicates that the new treatment is superior to the standard treatment (i.e. $p_1$ and $p_2$ are the probabilities of a beneficial event). To calculate a confidence interval for the NNT, we first compute a confidence interval for the difference between proportions using one of the intervals in Section 4. Denote the lower and upper limits of that interval by $L$ and $U$.

We need to distinguish between positive and negative values of NNT. As suggested by Altman,[32] it may be informative to denote positive values of NNT by NNTB: the number of patients needed to be treated for one additional patient to benefit. In a similar manner, negative values of NNT can be made positive and denoted by NNTH: the number of patients needed to be treated for one additional patient to be harmed.

If the confidence interval for the difference between proportions does not include zero, the confidence interval for NNTB and NNTH can be obtained by taking the reciprocals of the absolute values of $L$ and $U$ and reversing their order:

$$1/|U| \text{ to } 1/|L|. \tag{21}$$

If, on the other hand, the interval $(L, U)$ contains zero, the confidence interval for the NNT should[32] be denoted by

$$\text{NNTH } 1/|L| \text{ to } \infty \text{ to NNTB } 1/U. \tag{22}$$

Example: data from Table 2. The estimate of the number needed to be treated with the standard compared with the high dose of epinephrine for one additional patient to survive is

$$\text{NNT} = \frac{1}{\hat{p}_1 - \hat{p}_2} = \frac{1}{0.18} = 5.6.$$

Based on the computed confidence intervals for the difference between proportions in Table 3, we calculate seven different confidence intervals for the NNT (Table 4). We note that calculating confidence intervals for the NNT using the Newcombe hybrid score interval has been recommended by Bender.[33]

## 6 The ratio of proportions

### 6.1 Introduction and estimate

In Section 4, we considered the difference between proportions. Another measure of interest is the ratio of proportions:

$$\phi = \frac{p_1}{p_2}.$$

**Table 4.** Confidence intervals for the NNT with standard *versus* high-dose epinephrine using the confidence intervals for the difference between proportions in Table 3

| | |
|---|---|
| Wald | NNTB 3.1 to 34 |
| Agresti–Caffo | NNTB 3.1 to 83 |
| Newcombe hybrid score | NNTB 2.9 to 53 |
| Miettinen–Nurminen asymptotic score | NNTB 2.9 to 36 |
| Santner–Snell exact unconditional | NNTB 2.4 to $\infty$ to NNTH 14 |
| Chan–Zhang exact unconditional | NNTB 2.8 to 53 |
| Agresti–Min exact unconditional | NNTB 2.9 to 42 |

Note: The estimate is NNT = 5.6.

It is mostly used in cohort studies where the proportions are defined as the probabilities of a harmful event. The ratio of proportions is then called the risk ratio or the relative risk and measures the elevated risk in one group compared with the other.

We estimate the ratio of proportions using the sample proportions:

$$\hat{\phi} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}. \tag{23}$$

## 6.2 Confidence intervals

### 6.2.1 Katz log

Katz *et al.*[34] showed that the logarithm of the ratio of proportions is approximately normal distributed. Using the delta method, we obtain a confidence interval for $\phi$ by exponentiating the endpoints of

$$\log(\hat{\phi}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} - \frac{1}{n_{1+}} - \frac{1}{n_{2+}}}. \tag{24}$$

This interval cannot be computed when $n_{11} = 0$ or $n_{21} = 0$. When $n_{11} = n_{1+}$ and $n_{21} = n_{2+}$, the estimate of standard error is zero, resulting in the interval (1, 1).

### 6.2.2 Adjusted log

An adjusted log interval for $\phi$ can be obtained by adding 1/2 success to each group (Walter[35]):

$$\hat{\phi}_{1/2} = \frac{(n_{11} + 0.5)/(n_{1+} + 0.5)}{(n_{21} + 0.5)/(n_{2+} + 0.5)}.$$

Using the variance estimate of Pettigrew, *et al.*[36] a confidence interval for $\phi$ is given by exponentiating the endpoints of

$$\log(\hat{\phi}_{1/2}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11} + 0.5} + \frac{1}{n_{21} + 0.5} - \frac{1}{n_{1+} + 0.5} - \frac{1}{n_{2+} + 0.5}}. \tag{25}$$

This method allows for zero events in either group but produces the zero-width interval (1, 1) when $n_{11} = n_{1+}$ and $n_{21} = n_{2+}$. The adjusted log interval excludes the estimate of the ratio of proportions

given by Equation (23) in two cases: (1) when $n_{11}=0$ and $n_{21}\neq 0$, where $\hat{\phi}=0$ and the lower endpoint is $L>0$; and (2) when $n_{11}\neq 0$ and $n_{21}=0$, where $\hat{\phi}=\infty$ and the upper endpoint is finite.

### 6.2.3 Inverse hyperbolic sine

Using the inverse hyperbolic sine transformation, Newcombe[37] forms a confidence interval for $\phi$ by exponentiating the endpoints of

$$\log(\hat{\phi}) \pm 2\sinh^{-1}\left(\frac{z_{\alpha/2}}{2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{21}}-\frac{1}{n_{1+}}-\frac{1}{n_{2+}}}\right). \tag{26}$$

If $n_{11}=0$ or $n_{21}=0$, substitute the zero cell entry with $z^2_{\alpha/2}$ before calculating the interval. As was the case for the Katz log and the adjusted log intervals, the inverse sinh interval is equal to (1, 1) when $n_{11}=n_{1+}$ and $n_{21}=n_{2+}$. The inverse sinh interval excludes $\hat{\phi}$ for the same situations as does the adjusted log interval.

The inverse hyperbolic sine transformation was originally presented as an approach to confidence interval estimation for a single binomial proportion and the OR. The extension to the ratio of proportions was mentioned only briefly in Newcombe,[37] but was later included in the evaluations of Price and Bonett.[38] The inverse hyperbolic sine interval for the OR has not been properly evaluated yet and we do not consider it in Section 7.

### 6.2.4 Koopman asymptotic score

Koopman[39] proposed an asymptotic score confidence interval for the ratio of proportions that is always consistent with Pearson's chi-squared test.[39,40] The interval can be computed iteratively. Define the function

$$U(\phi)=\frac{(n_{11}-n_{1+}\tilde{p}_1)^2}{n_{1+}\tilde{p}_1(1-\tilde{p}_1)}\left\{1+\frac{n_{11}(\phi-\tilde{p}_1)}{n_{2+}(1-\tilde{p}_1)}\right\},$$

where

$$\tilde{p}_1=\frac{\phi\cdot(n_{1+}+n_{21})+n_{11}+n_{2+}-\sqrt{\left[\phi\cdot(n_{1+}+n_{21})+n_{11}+n_{2+}\right]^2-4\phi\cdot N(n_{11}+n_{21})}}{2N}.$$

$U$ is a convex function of $\phi$. $(\phi_L, \phi_U)$ is a confidence interval for $\phi$ if $\phi_L$ and $\phi_U$ are the two solutions to the equation

$$U(\phi)=\chi^2_{1,1-\alpha}, \tag{27}$$

where $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ percentile of the chi-squared distribution with one degree of freedom and $\phi_L<\phi_U$. If $n_{11}=0$, let $\phi_L=0$. If $n_{21}=0$, let $\phi_U=\infty$.

Using a series of substitutions, Nam[41] solved Equation (27) analytically. The resulting closed-form expression for the score interval is rather elaborate, and as such, we do not give it here.

The Koopman interval can distribute tail probabilities unevenly. If one-sided confidence intervals are of interest, the skewness-corrected score interval by Gart and Nam[40] can be used. In that same paper, Gart and Nam show that the Koopman interval is almost identical to the asymptotic score interval by Miettinen and Nurminen[13]–an interval for $\phi$ analogous to that for $\Delta$ (Section 4.2.4). Both intervals can be derived from the general theory of score methods by Gart.[42]

### 6.2.5 Exact unconditional intervals

Exact unconditional intervals can be obtained by inverting two one-sided $\alpha/2$-level exact tests (the tail method), or one two-sided $\alpha$-level exact test. For a specified value $\phi_0$, the score test statistic for the ratio of proportions is

$$T(\boldsymbol{n}|\phi_0) = \frac{\hat{p}_1 - \phi_0\hat{p}_2}{\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1+}} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2+}}}}, \tag{28}$$

where $\boldsymbol{n}$ denotes the observed table, and $\tilde{p}_1$ and $\tilde{p}_2$ the maximum likelihood estimates of $p_1$ and $p_2$ subject to $p_1/p_2 = \phi_0$. Closed-form expressions for $\tilde{p}_1$ and $\tilde{p}_2$ can be found in Miettinen and Nurminen[13] and Farrington and Manning.[14]

Under the restriction $p_1/p_2 = \phi_0$, the domain of $p_1$ given $\phi_0$ is

$$I(\phi_0) = \{p_1 : 0 \le p_1 \le \min(\phi_0, 1)\}. \tag{29}$$

As before, let $\boldsymbol{x} = \{x_{11}, x_{12}, x_{21}, x_{22}\}$ denote any $2 \times 2$ table that might be observed given the fixed row sums. The probability of observing $\boldsymbol{x}$ is the product of the likelihoods for the number of successes in the two samples:

$$f(\boldsymbol{x}|\, p_1, \phi_0) = \binom{x_{1+}}{x_{11}} p_1^{x_{11}}(1-p_1)^{x_{12}} \times \binom{x_{2+}}{x_{21}} (p_1/\phi_0)^{x_{21}}(1-p_1/\phi_0)^{x_{22}}.$$

*Inverting two one-sided score tests (Chan–Zhang).* Using the method by Chan and Zhang,[15] we invert two one-sided exact score tests (Equation (28)) of size at most $\alpha/2$ (the tail method). Define

$$P\big(T(\boldsymbol{n})|\, p_1, \phi_0\big) = \sum_{T(\boldsymbol{x}) \le T(\boldsymbol{n})} f(\boldsymbol{x}|\, p_1, \phi_0)$$

and

$$Q\big(T(\boldsymbol{n})|\, p_1, \phi_0\big) = \sum_{T(\boldsymbol{x}) \ge T(\boldsymbol{n})} f(\boldsymbol{x}|\, p_1, \phi_0).$$

The nuisance parameter $p_1$ is eliminated by taking the supremum over the range $I(\phi_0)$ given in (29):

$$P\big(T(\boldsymbol{n})|\phi_0\big) = \sup_{p_1 \in I(\phi_0)} P\big(T(\boldsymbol{n})|\, p_1, \phi_0\big)$$

and

$$Q\big(T(\boldsymbol{n})|\phi_0\big) = \sup_{p_1 \in I(\phi_0)} Q\big(T(\boldsymbol{n})|\, p_1, \phi_0\big).$$

The Chan–Zhang confidence interval $(L, U)$ for $\phi$ is the solution of

$$Q\big(T(\boldsymbol{n})|L\big) = \alpha/2 \tag{30}$$

and

$$P\big(T(\boldsymbol{n})|U\big) = \alpha/2. \tag{31}$$

*Inverting one two-sided score test (Agresti–Min).* Agresti and Min[16] suggest a method based on inverting one two-sided exact score test of size at most $\alpha$. Define

$$R\big(T(\boldsymbol{n})|\, p_1, \phi_0\big) = \sum_{|T(\boldsymbol{x})|\geq|T(\boldsymbol{n})|} f(\boldsymbol{x}|\, p_1, \phi_0).$$

We eliminate the nuisance parameter $p_1$ by maximizing over all possible values, i.e.

$$R\big(T(\boldsymbol{n})|\phi_0\big) = \sup_{p_1\in I(\phi_0)} R\big(T(\boldsymbol{n})|\, p_1, \phi_0\big).$$

The Agresti–Min confidence interval $(L, U)$ for $\phi$ is the solution of

$$R\big(T(\boldsymbol{n})|L\big) = \alpha \tag{32}$$

and

$$R\big(T(\boldsymbol{n})|U\big) = \alpha, \tag{33}$$

such that $R(T(\boldsymbol{n})|\phi_0) < \alpha$ when $\phi_0 < L$ and $R(T(\boldsymbol{n})|\phi_0) < \alpha$ when $\phi_0 > U$.

*The Berger and Boos procedure.* The discussions in Sections 4.2.5 and 4.3.2 on the benefits of using the Berger and Boos procedure also apply to exact unconditional intervals for the ratio of proportions.

## 6.3 Comparisons of intervals

The discussions in Sections 6.3.1 and 6.3.2 are based on exact calculations of coverage probabilities (Section 3) for several combinations of $n_{1+}$, $n_{2+}$, $p_1$ and $p_2$. The figures show some typical cases that were selected for the purpose of illustration. Some of the statements on interval performance refer to calculations that are not shown in the figures. Our findings are consistent with previous literature, except for the noted discrepancy at the end of Section 6.3.1.

### 6.3.1 Approximate intervals

We illustrate the coverage probabilities of the Katz log, adjusted log, inverse hyperbolic sine and Koopman asymptotic score intervals in Figure 4. The Katz log interval is usually quite conservative. It improves with increasing sample size, but its coverage probabilities are almost always further from the nominal level than those of the other three intervals. The adjusted log interval is slightly less conservative, but it can have coverage probabilities markedly below the nominal level for unproblematic parameter values – values for which other intervals perform quite well. The inverse hyperbolic sine interval has coverage probabilities fairly close to the nominal level when the minimal of the two proportions is above a certain value. This value seems to be about 0.15–0.25 for small sample sizes, such as $n_{1+} = n_{2+} = 10$, and decreases rapidly for increasing sample size. For proportions below this value, the coverage probability of the inverse sinh interval can be very low.

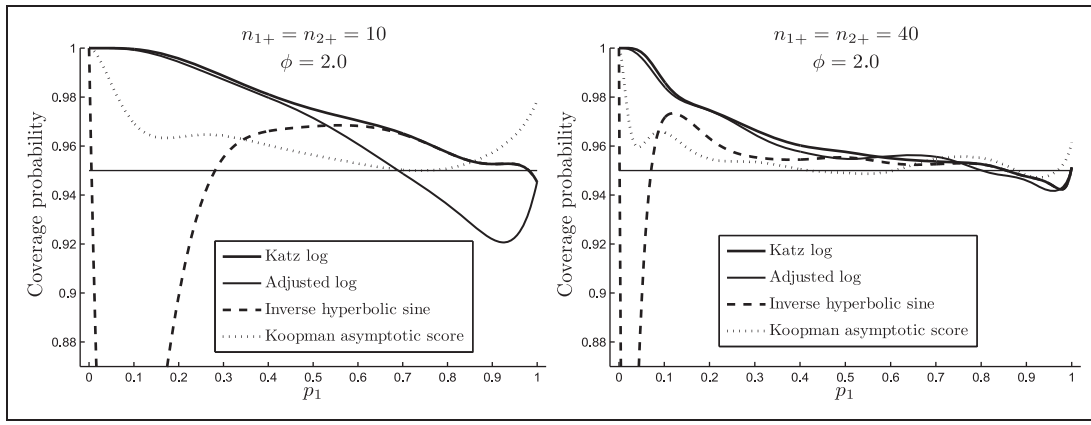


**Figure 4.** Coverage probabilities of the Katz log, adjusted log, inverse hyperbolic sine and Koopman asymptotic score intervals for sample sizes $n_{1+} = n_{2+} = 10$ (left) and $n_{1+} = n_{2+} = 40$ (right).
Note: The ratio of proportions is fixed at 2.0.

The Koopman asymptotic score interval performs almost always better than the other intervals, and it works quite well even for small sample sizes.

All four intervals in Figure 4 are negatively affected by unequal sample size; particularly, the inverse sinh interval, which has coverage probabilities close to the nominal level only for a narrow range of parameter values. The Katz log and Koopman asymptotic score intervals are affected the least, but the Koopman interval can have coverage probabilities below the nominal level for small proportions.

Price and Bonett[38] consider several large sample intervals for the ratio of proportions, including the Katz log, adjusted log, inverse hyperbolic sine and Koopman asymptotic score intervals. The authors find that the Koopman interval is clearly superior to the other intervals. They also find that the Katz log interval has coverage probabilities far below the nominal level, even for moderately large sample sizes. That result is not replicated in our calculations, where the Katz log interval is usually quite conservative. This difference may be due to different ways of dealing with incomputable limits. As noted in Section 3, we explicitly set the confidence limits to $(0, \infty)$ when an interval has incomputable limits. If, instead, we ignore those cases from the calculations of coverage probability, we get results consistent with the findings in Price and Bonett.[38]

### 6.3.2 *Exact intervals*

We compare the coverage probabilities of the two exact unconditional intervals, Chan–Zhang and Agresti–Min, in Figure 5. For equal sample sizes, the Agresti–Min interval is superior to the Chan–Zhang interval (left panel). This holds for most values of $\phi$ and for both small and moderate sample sizes, however, the difference in coverage probabilities between the two intervals decreases with increasing sample size. For small sample sizes, the Chan–Zhang interval is usually quite conservative.

When sample sizes are unequal, both intervals can be negatively affected. The coverage probabilities depend heavily on the particular combination of sample sizes and the value of $\phi$. Sometimes, as illustrated in the right panel of Figure 5, the Chan–Zhang and Agresti–Min

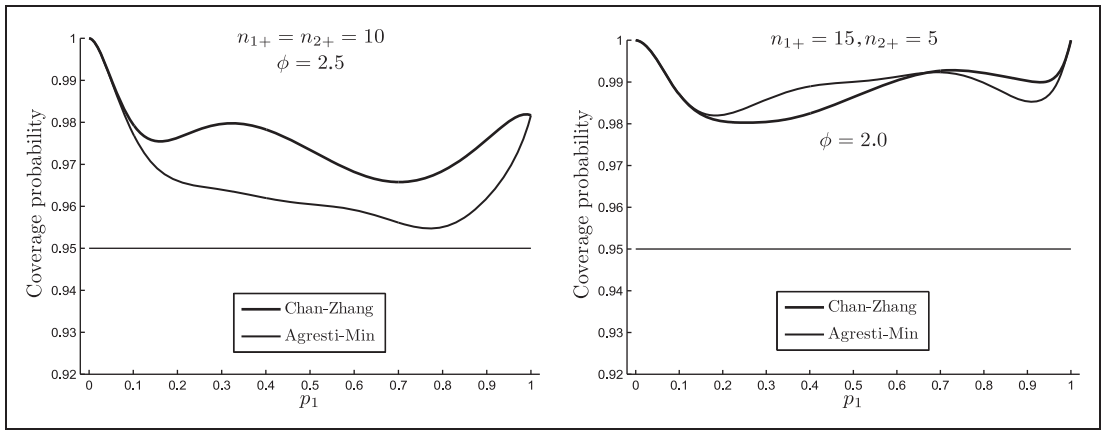


**Figure 5.** Coverage probabilities of the Chan–Zhang and Agresti–Min exact unconditional intervals for sample sizes $n_{1+} = n_{2+} = 10$ (left) and $n_{1+} = 15$, $n_{2+} = 5$ (right).
Note: The ratio of proportions is fixed at 2.5 (left) and 2.0 (right).

intervals perform almost equally, here with mean coverage probabilities equal to 98.8% and 98.9%, respectively. For the same sample size but with $\phi = 3.0$ instead of $\phi = 2.0$, the Agresti–Min interval performs much better with a mean coverage probability of 96.2%, whereas the Chan–Zhang interval is even more conservative with a mean coverage probability of 99.4%. If we change the sample sizes, for instance to $n_{1+} = 13$, $n_{2+} = 7$ or to $n_{1+} = 12$, $n_{2+} = 8$, different patterns of coverage probabilities emerge. As a rule, however, the Agresti–Min interval has coverage probabilities closer to the nominal level than does the Chan–Zhang interval.

We refer to Section 4.3.2 for a discussion on the non-consistency of intervals based on inverting one two-sided test – such as the Agresti–Min interval – and the results of one-sided tests.

### 6.3.3 The best performing approximate and exact intervals

The two best performing intervals for the ratio of proportions are the Koopman asymptotic score and the Agresti–Min exact unconditional intervals. A direct comparison of the two intervals for two situations with small and unequal sample sizes is shown in Figure 6. The coverage probability of the Koopman interval is generally closer to the nominal level than that of the Agresti–Min interval; however, the Koopman interval can be quite liberal for combinations of unequal sample sizes and small proportions.

### 6.3.4 Example: data from Table 2

For the data in Table 2, the estimate of the ratio of proportions is $\hat{\phi} = (7/34) / (1/34) = 7.0$. Table 5 shows the results of calculating the six confidence intervals considered in Section 6.2 using the observations in Table 2. In contrast to most of the intervals for the difference between proportions, these intervals differ markedly. The adjusted log interval is not to be trusted as this interval can have coverage probabilities well below the nominal level. The Koopman asymptotic score interval always perform well and has shorter length than the Katz log interval. The inverse hyperbolic sine interval is similar to the Koopman interval. Both exact unconditional intervals are clearly quite conservative, particularly the Chan–Zhang interval, which is considerably wider than the Koopman interval.

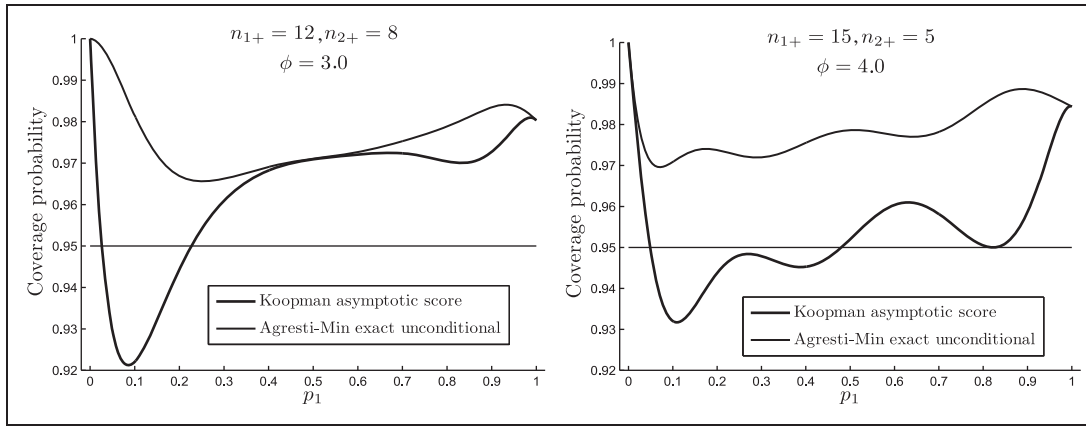


**Figure 6.** Coverage probabilities of the two best performing confidence intervals for the ratio of proportions.

**Table 5.** Confidence intervals for the ratio of proportions using data from Table 2.

| | Confidence interval | | |
|---|---|---|---|
| | Lower | Upper | Length[a] |
| Katz log | 0.91 | 54 | 4.08 |
| Adjusted log | 0.92 | 27 | 3.38 |
| Inverse sinh | 1.17 | 42 | 3.58 |
| Koopman asymptotic score | 1.21 | 43 | 3.57 |
| Chan–Zhang | 1.22 | 181 | 5.00 |
| Agresti–Min | 1.15 | 89 | 4.35 |

[a]Length = log(upper) − log(lower).
Note: The estimate is $\hat{\phi} = 7.0$.

# 7 The OR

## 7.1 Introduction and estimate

The odds of an event is the probability that the event occurs divided by the probability that it does not occur. The odds of success in group 1 is $p_1/(1-p_1)$, and in group 2, it is $p_2/(1-p_2)$. The ratio of the two odds is called the OR and is denoted by

$$\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

The OR is the natural measure of effect in case–control studies. It also plays an important role in logistic regression, where the relationship between the OR and a regression coefficient (β) is $\theta = \exp(\beta)$. Due to its mathematical properties, the OR is commonly used as a summary measure in meta-analysis.[43]

As for the difference between proportions and the ratio of proportions, we estimate the OR using the sample proportions

$$\hat{\theta} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{34}$$

## 7.2 Confidence intervals

### 7.2.1 Woolf logit

A confidence interval based on the approximate normal distribution of $\log(\hat{\theta})$ was first proposed by Woolf[44] and is often referred to as the logit interval. By the delta method, we obtain a confidence interval for θ by exponentiating the endpoints of

$$\log\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{35}$$

If one of the cell counts is zero, the standard error will be infinite and the confidence interval uninformative.

### 7.2.2 Gart adjusted logit

Gart[45] suggested an adjustment to the logit interval by adding 0.5 to all cell counts. The resulting confidence interval for θ is obtained by exponentiating the endpoints of

$$\log\tilde{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{\tilde{n}_{11}} + \frac{1}{\tilde{n}_{12}} + \frac{1}{\tilde{n}_{21}} + \frac{1}{\tilde{n}_{22}}}, \tag{36}$$

where

$$\tilde{\theta} = \frac{\tilde{n}_{11}\tilde{n}_{22}}{\tilde{n}_{12}\tilde{n}_{21}}, \quad \tilde{n}_{11} = n_{11} + 0.5, \quad \tilde{n}_{12} = n_{12} + 0.5, \quad \tilde{n}_{21} = n_{21} + 0.5, \quad \tilde{n}_{22} = n_{22} + 0.5.$$

This interval can be computed also when one or more cell counts are zero, however, it will, in some cases, exclude the estimate of the OR given by Equation (34). If $n_{11} = 0$ and $n_{21} \neq 0$, the estimate is $\hat{\theta} = 0$, but the lower endpoint is $L > 0$. If $n_{11} = n_{1+}$ and $n_{21} \neq n_{2+}$, or if $n_{11} \neq 0$ and $n_{21} = 0$, the estimate is $\hat{\theta} = \infty$, but the upper endpoint is finite.

The adjusted logit interval is sometimes referred to as the Haldane–Anscombe correction.

### 7.2.3 Independence-smoothed logit

The above adjustment, Equation (36), is a special case of a general class of adjustments. Suppose that we add a non–negative quantity $c$ to each cell count and proceed as above. For $c = 0$, we obtain the Woolf logit interval, and for $c = 0.5$, the Gart adjusted logit interval. Agresti[46] suggests individual values for each cell based on the observed data:

$$c_{ij} = 2n_{i+}n_{+j}/N^2, \quad i,j = 1,2.$$

The resulting confidence interval is called the independence-smoothed logit interval. It excludes the estimate of the OR for the same situations, as described in Section 7.2.2.

### 7.2.4 *Cornfield exact conditional*

In Sections 4.2.5 and 6.2.5, we developed exact unconditional confidence intervals for the difference between proportions and the ratio of proportions. These methods coped with the nuisance parameter $p_1$ by maximizing the $p$-value over the range of $p_1$. Another approach to eliminate the nuisance parameter is to condition on a sufficient statistic for it. Methods based on the conditional approach has the benefit that they are much simpler and computationally easier than unconditional methods. Like their unconditional counterparts, exact conditional intervals are guaranteed to have at least nominal coverage probabilities. Conditional confidence intervals can be constructed for the OR but not for the difference between proportions nor the ratio of proportions (Agresti[23], p.101).

If we condition on the number of successes ($n_{+1}$) and the number of failures ($n_{+2}$), such that all marginal totals in Table 1 are fixed, the probability of observing a table with $x_{11}$ successes follows the non-central hypergeometric distribution

$$f(x_{11}|\theta) = \frac{\binom{n_{1+}}{x_{11}}\binom{n_{2+}}{n_{+1}-x_{11}}\theta^{x_{11}}}{\sum_{i=n_{\min}}^{n_{\max}}\binom{n_{1+}}{i}\binom{n_{2+}}{n_{+1}-i}\theta^{i}},$$

where $n_{\max} = \max(0,\ n_{+1} - n_{2+})$ and $n_{\min} = \min(n_{1+},\ n_{+1})$.

Cornfield[47] constructs an exact conditional confidence interval ($L$, $U$) for θ by inverting two one-sided exact conditional tests (the tail method). $L$ and $U$ can be obtained by solving the following equations iteratively

$$\sum_{x_{11}=n_{11}}^{n_{\min}} f(x_{11}|L) = \alpha/2 \tag{37}$$

and

$$\sum_{x_{11}=n_{\max}}^{n_{11}} f(x_{11}|U) = \alpha/2. \tag{38}$$

The Cornfield exact interval is always consistent with the Fisher–Irwin exact test.[47]

Cornfield also proposed an asymptotic approximation to the exact conditional interval. It was originally presented with a continuity correction and performs quite similarly to the exact interval.[46,48] Miettinen and Nurminen[13] suggested an interval that is equal to the approximate Cornfield interval without the continuity correction. Cornfield's unadjusted interval is supported in Stata, whereas the adjusted interval is unsupported in the most common statistical software packages.

### 7.2.5 *Baptista-Pike exact conditional*

Baptista and Pike[49] use the method by Sterne[50] and invert a two-sided test with acceptance region formed by ordered null probabilities. We solve the following equations instead of (37) and (38):

$$\sum_{x_{11}=n_{\max}}^{n_{\min}} f(x_{11}|L) \cdot I\Big(f(x_{11}|L) \le f(n_{11}|L)\Big) = \alpha \tag{39}$$

and

$$\sum_{x_{11}=n_{\max}}^{n_{\min}} f(x_{11}|U) \cdot I\Big(f(x_{11}|U) \le f(n_{11}|U)\Big) = \alpha, \tag{40}$$

such that $L < U$. $I$ is an indicator function, and $n_{\max}$, $n_{\min}$ and $f$ are as defined in Section 7.2.4. The interval given by $(L, U)$ is an exact conditional confidence interval for the OR.

### 7.2.6 Quasi-exact intervals (mid-p)

The mid-$p$ approach, first proposed by Lancaster,[51] is a general approach for statistical inference that has been applied in a wide range of settings, particularly for small sample and sparse data (Hirji[10], pp.50–51 and 218–219). The mid-$p$ concept has mostly been used in relation to significance testing – the Fisher's – exact mid-$p$ test is one of the recommended tests for association in $2 \times 2$ tables in Lydersen et al.[2] - however, it can also be used for confidence interval estimation.[52]

To calculate a mid-$p$ value, we subtract half the point probability of the observed table from the ordinary $p$-value. Contrary to an exact test, a mid-$p$ test is not guaranteed to have type I error probabilities below the nominal level. In a similar manner, a mid-$p$ interval is not guaranteed to have at least nominal coverage probabilities. For both tests and intervals, however, the nominal level is seldom violated, and when that happens, the degree of infringement is usually low.[53,54]

Because mid-$p$ inference is based on exact distributions but without the guarantee to maintain the nominal level, mid-$p$ tests and intervals are often called quasi-exact.[55]

A mid-$p$ confidence interval for $\theta$ based on the Cornfield exact conditional interval is obtained by making the following adjustments to Equations (37) and (38):

$$\sum_{x_{11}=n_{11}}^{n_{\min}} f(x_{11}|L) - \frac{1}{2} f(n_{11}|L) = \alpha/2 \tag{41}$$

and

$$\sum_{x_{11}=n_{\max}}^{n_{11}} f(x_{11}|U) - \frac{1}{2} f(n_{11}|U) = \alpha/2. \tag{42}$$

To obtain the Baptista–Pike mid-$p$ confidence interval, we substitute Equations (39) and (40) with

$$\sum_{x_{11}=n_{\max}}^{n_{\min}} f(x_{11}|L) \cdot I\Big(f(x_{11}|L) \le f(n_{11}|L)\Big) - \frac{1}{2} f(n_{11}|L) = \alpha \tag{43}$$

and

$$\sum_{x_{11}=n_{\max}}^{n_{\min}} f(x_{11}|U) \cdot I\Big(f(x_{11}|U) \le f(n_{11}|U)\Big) - \frac{1}{2} f(n_{11}|U) = \alpha. \tag{44}$$

### 7.2.7 Agresti–Min exact unconditional

Agresti and Min[56] consider exact unconditional intervals for the OR, and in particular, an interval based on inverting one two-sided exact unconditional score test. The approach is similar to the Agresti–Min exact unconditional intervals for the difference of proportions (Section 4.2.5) and the ratio of proportions (Section 6.2.5). For a given value $\theta_0$, we have the constraint $\theta_0 = p_1(1-p_2)/p_2(1-p_1)$. The score test statistic for the OR is[56]

$$T(\boldsymbol{n}|\theta_0) = \left[n_{1+}(\hat{p}_1 - \tilde{p}_1)\right]^2 \left[\frac{1}{n_{1+}\tilde{p}_1(1-\tilde{p}_1)} + \frac{1}{n_{2+}\tilde{p}_2(1-\tilde{p}_2)}\right], \tag{45}$$

where $\boldsymbol{n}$ denotes the observed table, and $\tilde{p}_1$ and $\tilde{p}_2$ the maximum likelihood estimates of $p_1$ and $p_2$. We refer to Miettinen and Nurminen[13] for closed-form expressions of $\tilde{p}_1$ and $\tilde{p}_2$. Let $\boldsymbol{x} = \{x_{11}, x_{12}, x_{21}, x_{22}\}$ denote any $2 \times 2$ table that might be observed given the fixed row sums. The probability of observing $\boldsymbol{x}$ is the product of the likelihoods for the number of successes in the two samples:

$$f(\boldsymbol{x}|\, p_1, \theta_0) = \binom{x_{1+}}{x_{11}} p_1^{x_{11}}(1-p_1)^{x_{12}} \times \binom{x_{2+}}{x_{21}} \left[p_1/(p_1+\theta_0-p_1\theta_0)\right]^{x_{21}} \left[1-p_1/(p_1+\theta_0-p_1\theta_0)\right]^{x_{22}}.$$

Define

$$R\big(T(\boldsymbol{n})|\, p_1, \theta_0\big) = \sum_{|T(\boldsymbol{x})| \geq |T(\boldsymbol{n})|} f(\boldsymbol{x}|\, p_1, \theta_0),$$

and eliminate the nuisance parameter $p_1$ by maximizing over all possible values:

$$R\big(T(\boldsymbol{n})|\theta_0\big) = \sup_{p_1} R\big(T(\boldsymbol{n})|\, p_1, \theta_0\big).$$

The range of $p_1$ may be reduced with the Berger and Boos procedure (Section 4.2.5). The Agresti–Min confidence interval $(L, U)$ for $\theta$ is the solution of

$$R\big(T(\boldsymbol{n})|L\big) = \alpha \tag{46}$$

and

$$R\big(T(\boldsymbol{n})|U\big) = \alpha, \tag{47}$$

such that $R(T(\boldsymbol{n})|\theta_0) < \alpha$ when $\theta_0 < L$ and $R(T(\boldsymbol{n})|\theta_0) < \alpha$ when $\theta_0 > U$.

## 7.3 Comparisons of intervals

The comparisons in this section are based on several exact calculations of coverage probabilities (Section 3). The figures present illustrative cases and are used to initiate the discussion on differences and similarities between the intervals. Unreferenced statements on interval performance are based on our calculations, which are consistent with previous published literature.

### 7.3.1 Approximate intervals

Figure 7 presents typical coverage probabilities of the Woolf logit, Gart adjusted logit and independence-smoothed logit intervals for one small sample size and one medium sample size. The three logit intervals perform similarly. For small sample sizes, such as 10 subjects in each sample (left panel of Figure 7), the Gart adjusted logit interval is slightly less conservative compared with the other two intervals. All three intervals are conservative for small and large proportions.

None of the intervals are particularly affected by unbalanced sample size. For instance, when $n_{1+} = 40$ and $n_{2+} = 20$ (results not shown), all three intervals have similar coverage probabilities to the ones seen in the right panel of Figure 7.

With increasing OR, the performances of the three logit intervals diverge (Figure 8). The Gart adjusted logit interval can have coverage probabilities lower than the nominal level, starting at about $\theta = 8$. This problem increases with increasing OR, and when $\log(\theta) > 4$, the coverage probability of the Gart interval can be very low.[46]

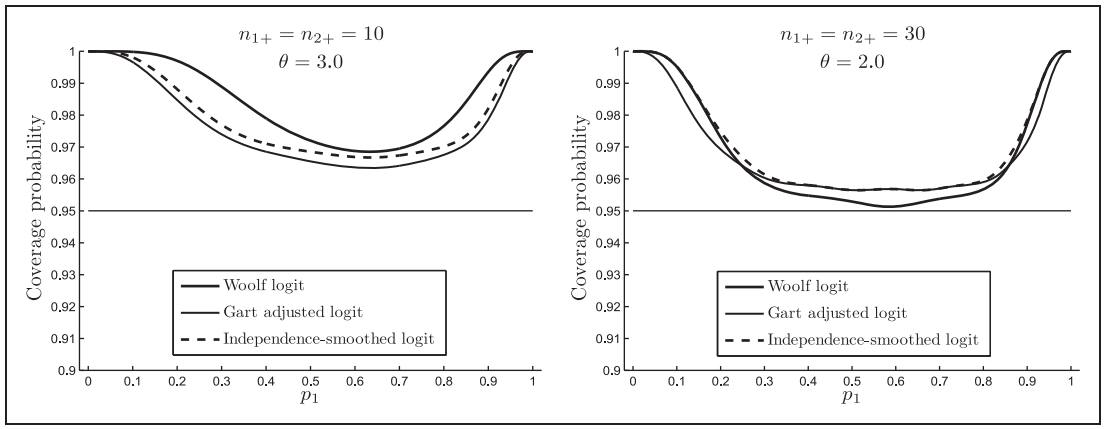


**Figure 7.** Coverage probabilities of the Woolf logit, Gart adjusted logit and independence-smoothed logit intervals for sample sizes $n_{1+} = n_{2+} = 10$ (left) and $n_{1+} = n_{2+} = 30$ (right).
Note: The OR is fixed at 3.0 (left) and 2.0 (right).

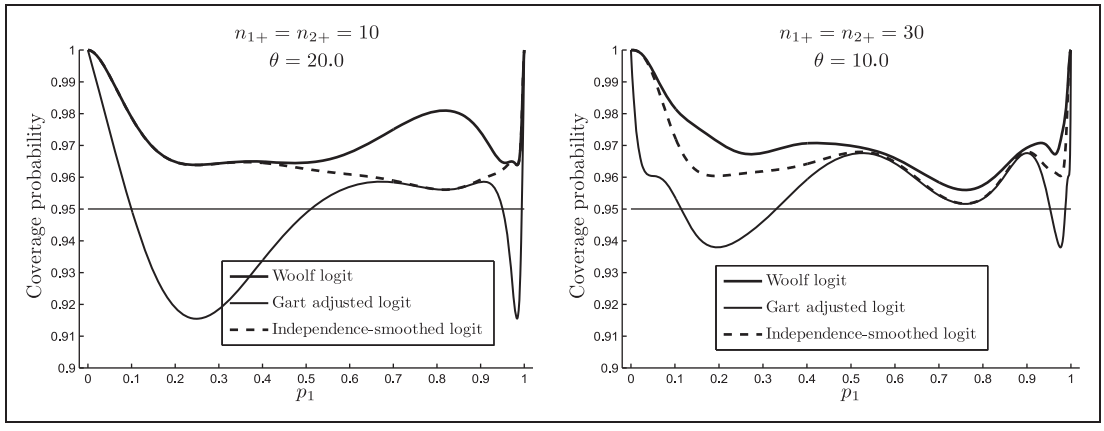


**Figure 8.** Coverage probabilities of the Woolf logit, Gart adjusted logit and independence-smoothed logit intervals for sample sizes $n_{1+} = n_{2+} = 10$ (left) and $n_{1+} = n_{2+} = 30$ (right).
Note: The OR is fixed at 20.0 (left) and 10.0 (right).

The independence-smoothed logit interval performs well in nearly all cases, while the Woolf logit interval is slightly more conservative. Even though Figures 7 and 8 do not show it, both the Woolf logit and the independence-smoothed logit intervals can have coverage probabilities below the nominal level. This happens rarely, particularly for the independence-smoothed logit interval.

The Gart logit interval is slightly shorter than the independence-smoothed logit interval.[46]

### 7.3.2 *Exact and quasi-exact intervals*

We present typical coverage probabilities for exact and quasi-exact intervals for the OR in Figure 9. For small sample sizes, the Cornfield exact conditional interval is very conservative, and it does not improve much with increasing sample size; it is still quite conservative with 75 subjects in each

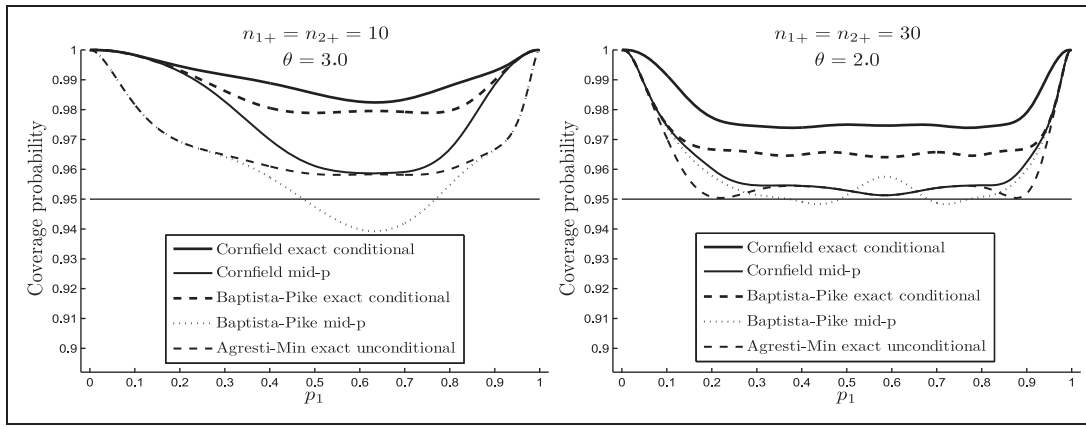


**Figure 9.** Coverage probabilities of the Cornfield exact conditional, Baptista–Pike exact conditional, Cornfield mid-*p*, Baptista–Pike mid-*p* and Agresti–Min exact unconditional intervals for sample sizes $n_{1+} = n_{2+} = 10$ (left) and $n_{1+} = n_{2+} = 30$ (right).
Note: The OR is fixed at 3.0 (left) and 2.0 (right).

sample (results not shown). The Baptista–Pike exact conditional interval improves upon the Cornfield exact interval, but it is also quite conservative, particularly for small sample sizes. The mid-*p* intervals have coverage probabilities closer to the nominal level than their exact counterparts. The Baptista–Pike mid-*p* interval can have coverage probabilities below the nominal level but usually not by much. All four intervals cope well with unequal sample sizes.

For increasing ORs, the Baptista–Pike mid-*p* interval performs very well, especially when $\theta \geq 10$ and there are 30 or more in each sample.

The Agresti–Min exact unconditional interval is clearly the superior exact interval. It is far less conservative than the Cornfield and Baptista–Pike exact conditional intervals. In Figure 9, the Agresti-Min interval performs quite similarly to the Baptista–Pike mid-*p* interval. In general, when $\theta < 5$ and $n_{1+} = n_{2+}$, the Baptista–Pike mid-*p* interval has coverage probabilities slightly closer to the nominal level than the Agresti–Min interval, but not by much. If, however, the sample size is unbalanced or $\theta \geq 5$, the Baptista–Pike mid-*p* interval will outperform the Agresti–Min interval (results not shown here).

Except for the Cornfield exact conditional interval, none of the intervals mentioned in this section are widely available in software packages or particularly easy to calculate.

### 7.3.3 The best performing approximate and exact intervals

Figure 10 is a head-to-head comparison of the four best performing intervals for the OR. If we accept coverage probabilities slightly below the nominal level, the Baptista–Pike mid-*p* interval is clearly superior to the logit intervals. This result persists for other combinations of sample sizes and θ-values. As discussed in Section 7.3.2, the Baptista–Pike mid-*p* interval usually performs somewhat better than the Agresti–Min exact unconditional interval. This is not obvious in the left panel of Figure 10, but can be seen in the right panel of Figure 10.

### 7.3.4 Example: data from Table 2

The OR was the measure of effect used in Perondi *et al.*[3] As mentioned in Section 2, the estimated OR for death with the high-dose therapy is 8.6. The authors used the Cornfield exact conditional

interval and presented a 95% confidence interval from 1.0 to 397.0. This interval is very wide, as are the Cornfield mid-*p* and Baptista–Pike exact conditional intervals (Table 6). The Baptista–Pike mid-*p* interval has equal length to the Woolf logit interval, but the endpoints are not equal; the Woolf interval contains the null value ($\theta = 1$), whereas the Baptista–Pike mid-*p* interval does not. A similar situation is seen with the independence-smoothed logit and the Agresti–Min exact unconditional intervals. The shortest interval is the Gart adjusted logit, which also includes the null value. The logit intervals thus have short lengths but are conservative, particularly for small and large proportions (Figure 7). For this example, the Baptista–Pike mid-*p* and the Agresti–Min exact unconditional intervals are the only intervals for the OR that are consistent with the Koopman asymptotic score interval for the ratio of proportions and the Newcombe hybrid score interval for the difference between proportions.

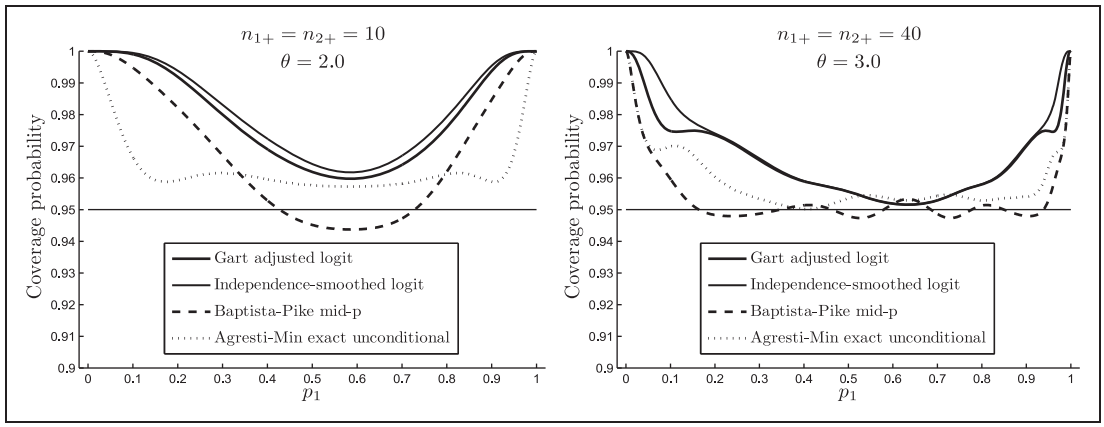


**Figure 10.** Coverage probabilities of the four best performing confidence intervals for the OR.

**Table 6.** Confidence intervals for the OR using data from Table 2.

| | Confidence interval | | |
|---|---|---|---|
| | Lower | Upper | Length[a] |
| Woolf logit | 0.99 | 74 | 4.31 |
| Gart adjusted logit | 0.98 | 38 | 3.65 |
| Independence-smoothed logit | 0.99 | 60 | 4.11 |
| Cornfield exact conditional | 0.97 | 397 | 6.01 |
| Cornfield mid-*p* | 1.19 | 200 | 5.12 |
| Baptista–Pike exact conditional | 1.00 | 195 | 5.28 |
| Baptista–Pike mid-*p* | 1.33 | 99 | 4.31 |
| Agresti–Min exact unconditional | 1.19 | 72 | 4.10 |

[a]Length = log(upper) − log(lower).
Note: The estimate is $\hat{\theta} = 8.6$.

## 8 Recommendations

We present a summary of recommended confidence intervals in Table 7. These are our prime choices, but they are not the only intervals we consider appropriate. A more detailed picture is given in Sections 8.1–8.4. Our recommended confidence intervals comply with the following criteria. The coverage probability is close to the nominal level for a wide range of parameter values. It is allowed, occasionally, to dip below the nominal level, as long as the infringement is small. For intervals with similar coverage probabilities, we prefer the interval with shortest length. With one exception, all intervals we recommend are either easy to calculate – usually meaning that it has a relatively simple closed-form expression – or readily available in one or more standard software packages (Table 8).

### 8.1 The difference between proportions

For small sample sizes (less than 30 in each sample), the Newcombe hybrid score interval performs well but can be somewhat liberal. It is easy to calculate and also available in several software packages (Table 8). The Agresti–Caffo interval is slightly more conservative but even easier to calculate than the Newcombe hybrid score interval. The Miettinen–Nurminen asymptotic score interval performs well in several situations, but it requires iterative calculations and is not widely available in software packages. The Agresti–Min exact unconditional interval is usually more conservative than the Newcombe hybrid score interval; however, it has better coverage probability when proportions are close to 0 or 1 and avoids coverage probabilities below the nominal level. The Agresti–Min interval is therefore our prime recommendation for small sample sizes; however, both the Newcombe and Agresti–Caffo intervals are good alternatives. The widely available Wald interval, with or without continuity correction, is not recommended.

For moderate and large sample sizes (more than 30 in each sample), the Agresti–Caffo, Newcombe hybrid score and Miettinen–Nurminen asymptotic score intervals perform similarly and all have coverage probabilities close to the nominal level. We prefer the Newcombe hybrid score interval, as it copes slightly better with proportions close to 0 or 1 than do the other intervals. Nevertheless, all three intervals are quite safe to use. The Wald interval has coverage probabilities close to the nominal level for 100 or more in each sample.

If the coverage probability must be at least the nominal size, we recommend the Agresti–Min exact unconditional interval.

**Table 7.** Summary of the recommended confidence intervals

| Measure | Small samples | Moderate and large samples |
|---|---|---|
| Difference between proportions | Agresti–Min exact unconditional | Newcombe hybrid score |
| NNT | The reciprocal of the above confidence interval limits for difference between proportions + Altman's notation | |
| Ratio of proportions | Koopman asymptotic score | Koopman asymptotic score |
| OR | Baptista–Pike mid-$p$[a] | Baptista–Pike mid-$p$[a] |

[a]Not available in standard software packages. We refer to Section 8.4 for available alternatives.

Note: Other appropriate intervals are considered in Sections 8.1–8.4, which also details what constitutes small and large sample sizes.

**Table 8.** Availability of confidence intervals in standard software packages

| Confidence interval | Closed form[a] | CIA 2.1 | R 2.12 | SAS 9.2 | SPSS 18 | Stata 11 | StatXact 9 |
|---|---|---|---|---|---|---|---|
| Difference between proportions | | | | | | | |
| *Approximate intervals* | | | | | | | |
| Wald | √ | √ | √ | √ | √[b] | √ | – |
| Wald with continuity correction | √ | – | √ | √ | – | – | – |
| **Agresti–Caffo (3)** | √ | – | √[c] | – | – | √[d] | – |
| **Newcombe hybrid score (2)** | √ | √ | √[c] | √ | – | √[d] | – |
| Miettinen–Nurminen asymptotic score | – | – | √[e] | – | – | √[d] | √ |
| *Exact intervals* | | | | | | | |
| Santner–Snell exact unconditional | – | – | – | √ | – | – | √ |
| Chan–Zhang exact unconditional | – | – | – | – | – | – | √ |
| **Agresti-Min exact unconditional (1)** | – | – | – | – | – | – | √ |
| Ratio of proportions | | | | | | | |
| *Approximate intervals* | | | | | | | |
| Katz log | √ | √ | – | √ | √ | √ | – |
| Adjusted log | √ | – | √[c] | – | – | – | – |
| Inverse hyperbolic sine | √ | – | – | – | – | – | – |
| **Koopman asymptotic score (1)** | – | – | √[c,e] | – | – | √[f] | √ |
| *Exact intervals* | | | | | | | |
| Chan–Zhang exact unconditional | – | – | – | – | – | – | √ |
| **Agresti–Min exact unconditional (2)** | – | – | – | – | – | – | √ |
| OR | | | | | | | |
| *Approximate intervals* | | | | | | | |
| Woolf logit | √ | √ | – | √ | √ | √ | √ |
| **Gart adjusted logit (3)** | √ | – | √[c] | – | – | √[g] | – |
| **Independent-smoothed logit (4)** | √ | – | – | – | – | √[g] | – |
| Cornfield mid-*p* | – | – | – | – | – | – | – |
| **Baptista–Pike mid-*p* (1)** | – | – | – | – | – | – | – |
| *Exact intervals* | | | | | | | |
| Cornfield exact conditional | – | – | √ | √ | – | √ | √ |
| Baptista–Pike exact conditional | – | – | – | – | – | – | – |
| **Agresti–Min exact unconditional (2)** | – | – | – | – | – | – | – |

[a] Intervals with closed-form expressions can be calculated manually;
[b] Available with the Complex Samples module;
[c] Available with the pairwiseCI package;
[d] Available with the package rdci;
[e] Available with the PropCIs package;
[f] Available with the package sg154;
[g] Available with the package sbe30.

Note: Recommended intervals are typed in bold style with their preferred order in parentheses. Some of the packages also calculate other intervals than the ones listed here. Exact intervals are guaranteed to have coverage probability at least the nominal size.

## 8.2 The NNT

The computation of a confidence interval for the NNT is based on a confidence interval for the associated difference between proportions. One of the recommended intervals in Section 8.1 should be used for this purpose. We further recommend that the notation by Altman[32] is used, particularly when the interval for the difference between proportions contains zero.

## 8.3 The ratio of proportions

For both small and large sample sizes, the Koopman asymptotic score interval performs generally well. The Agresti–Min exact unconditional interval also performs well, except for some cases with unequal sample sizes, where it can be quite conservative. The Koopman interval has almost always coverage probabilities closer to the nominal level than does the Agresti–Min interval; however, it can be quite liberal for unequal sample sizes and small proportions. Based on the many situations where the Koopman interval has coverage probability close to the nominal level, we recommend the Koopman interval ahead of the Agresti–Min interval. The Agresti–Min interval is a safe but conservative choice, and we recommended it when the coverage probability needs to be at least the nominal level.

For 40 subjects or more in each sample, the inverse hyperbolic sine interval performs well when proportions are greater than 0.1. It is easy to calculate and can be a good alternative to the Koopman interval, which needs to be computed iteratively. We do not recommend the widely available Katz log interval nor the adjusted log interval, unless the sample size is at least 75 in each sample.

## 8.4 The OR

For small sample sizes, the three logit intervals – due to Gart, Woolf and Agresti – can be rather conservative with Gart as the least and Woolf as the most conservative. When the sample size increases, the performances of the three logit intervals become similar and their coverage probabilities get closer to the nominal level. Still, all three intervals are conservative when proportions are close to 0 or 1. The Gart interval is slightly better than the other two intervals.

The quasi-exact Baptista–Pike mid-$p$ interval performs well for most sample sizes and proportions. It can be slightly liberal but has coverage probabilities closer to the nominal level than do the logit intervals. The Agresti–Min exact unconditional interval performs almost as well as the Baptista–Pike mid-$p$ interval, but it is even more complex to calculate. We thus recommend the Baptista–Pike mid-$p$ interval ahead of the Agresti–Min exact unconditional interval for all sample sizes. If neither of these intervals are available – they are not supported by any standard software package – or the sample size is too large to compute them, we recommend the Gart adjusted logit interval, as long as the OR is not too far from one ($\theta = 10$ or $\theta = 0.1$ should be no problem). The independence-smoothed logit interval is also a good alternative.

If it is necessary to use a confidence interval that is guaranteed to have at least the nominal coverage probability, the only widely available exact interval is the Cornfield exact conditional interval. It can, however, be extremely conservative, so we recommend it only if no other options are available. The Baptista–Pike exact conditional and the Agresti–Min exact unconditional intervals both outperform the Cornfield interval, but neither are available in any standard software package.

## Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## Conflict of interest statement

The authors declare that there is no conflict of interest.

## References


1. Hamilton MA. Choosing the parameter for a $2 \times 2$ table or a $2 \times 2 \times 2$ table analysis. *Am J Epidemiol* 1979; **109**: 362–375.
2. Lydersen S, Fagerland MW and Laake P. Tutorial in biostatistics: recommended tests for association in $2 \times 2$ tables. *Stat Med* 2009; **28**: 1159–1175.
3. Perondi MBM, Reis AG, Paiva EF, Nadkarni VM and Berg RA. A comparison of high-dose and standard-dose epinephrine in children with cardiac arrest. *N Engl J Med* 2004; **350**: 1722–1730.
4. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998; **17**: 857–872.
5. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med* 1998; **17**: 873–890.
6. Altman DG, Machin D, Bryant TN and Gardner MJ (eds). *Statistics with confidence* (2nd edn). London: BMJ Books, 2000, pp. 48–49.
7. Yates F. Contingency tables involving small numbers and the $\chi^2$test. *J R Statist Assoc* 1934; (Suppl. 1): 217–235.
8. Fleiss JL, Levin B and Paik MC. *Statistical methods for rates and proportions* (3rd edn). Hoboken, NJ: John Wiley & Sons, Inc., 2003, p.60.
9. Agresti A and Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *Am Stat* 2000; **54**(4): 280–288.
10. Hirji KF. *Exact analysis of discrete data*. Boca Raton, FL: Chapman & Hall/CRC, 2006.
11. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 1927; **22**: 209–212.
12. Mee RW. Confidence bounds for the difference between two probabilities. *Biometrics* 1984; **40**: 1175–1176.
13. Miettinen O and Nurminen M. Comparative analysis of two rates. *Stat Med* 1985; **4**: 213–226.
14. Farrington CP and Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat Med* 1990; **9**: 1447–1454.
15. Chan ISF and Zhang Z. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* 1999; **55**: 1202–1209.
16. Agresti A and Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 2001; **57**: 963–971.
17. Santner TJ and Snell MK. Small-sample confidence intervals for $p_1 - p_2$ and $p_1/p_2$ in $2 \times 2$ contingency tables. *J Am Stat Assoc* 1980; **75**: 386–394.
18. Berger RL and Boos DD. P values maximized over a confidence set for the nuisance parameter. *J Am Stat Assoc* 1994; **89**: 1012–1016.
19. Lin Y, Newcombe RG, Lipsitz S and Carter RE. Fully specified bootstrap confidence intervals for the difference of two independent binomial proportions based on the median unbiased estimator. *Stat Med* 2009; **28**: 2876–2890.
20. Hirji KF, Tsiatis AA and Mehta CR. Median unbiased estimation for binary data. *Am Stat* 1989; **43**: 7–11.
21. Agresti A. Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Stat Meth Med Res* 2003; **12**: 3–21.
22. Cytel. *StatXact 9 user manual*. Cambridge, MA: Cytel Inc, 2010 (www.cytel.com), p. 505.
23. Agresti A. *Categorical data analysis* (2nd edn). Hoboken, NJ: John Wiley & Sons, Inc., 2002.
24. Lydersen S, Langaas M and Bakke Ø. The exact unconditional $z$-pooled test for equality of two binomial probabilities: optimal choice of the Berger and Boos confidence coefficient. *J Stat Comput Simul*. doi: 10.1080/00949655.2011.579969.
25. Santner TJ, Pradhan V, Senchaudhuri P, Mehta CR and Tamhane A. Small-sample comparisons of confidence intervals for the difference of two independent binomial proportions. *Computat Stat Data Anal* 2007; **51**: 5791–5799.
26. Coe PR and Tamhane AC. Small sample confidence intervals for the difference, ratio, and odds ratio of two success probabilities. *Commun Stat Simul Comput* 1993; **22**: 925–938.
27. Newcombe RG and Nurminen MM. In defence of score intervals for proportions and their differences. *Commun Stat Theory Meth* 2011; **40**: 1271–1282.
28. Laupacis A, Sackett DL and Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988; **318**(26): 1728–1733.
29. Cook RJ and Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995; **310**: 452–454.
30. Newcombe RG. Confidence intervals for the number needed to treat—absolute risk reduction is less likely to be misunderstood. *BMJ* 1999; **318**: 1765.
31. Bender R and Blettner M. Calculating the ''number needed to be exposed'' with adjustment for confounding variables in epidemiologic studies. *J Clin Epidemiol* 2002; **55**: 525–530.
32. Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998; **317**: 1309–1312.
33. Bender R. Calculating confidence intervals for the number needed to treat. *Control Clin Trials* 2001; **22**: 102–110.
34. Katz D, Baptista J, Azen SP and Pike MC. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* 1978; **34**: 469–474.
35. Walter SD. The distribution of Levin's measure of attributable risk. *Biometrika* 1975; **62**(2): 371–374.
36. Pettigrew HM, Gart JJ and Thomas DG. The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika* 1986; **73**(2): 425–435.
37. Newcombe RG. Logit confidence intervals and the inverse sinh transformation. *Am Stat* 2001; **55**: 200–202.
38. Price RM and Bonett DG. Confidence intervals for a ratio of two independent binomial proportions. *Stat Med* 2008; **27**: 5497–5508.
39. Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 1984; **40**: 513–517.
40. Gart JJ and Nam J. Approximate interval estimation of the ratio of binomial parameters: a review and correction for skewness. *Biometrics* 1988; **44**: 323–338.
41. Nam J. Confidence limits for the ratio of two binomial proportions based on likelihood scores: non-iterative method. *Biometrical J* 1995; **3**: 375–379.
42. Gart JJ. Approximate tests and interval estimation of the common relative risk in the combination of $2 \times 2$ tables. *Biometrika* 1985; **72**(3): 673–677.

43. Deeks JJ and Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G and Altman DG (eds) *Systematic reviews in health care: meta-analysis in context* (2nd edn). London: BMJ Publishing Group, 2001, pp. 313–335.
44. Woolf B. On estimating the relation between blood group and disease. *Ann Human Gene* 1955; **19**: 251–253.
45. Gart JJ. Alternative analyses of contingency tables. *J R Stat Soc B Stat Meth* 1966; **28**: 164–179.
46. Agresti A. On logit confidence intervals for the odds ratio with small samples. *Biometrics* 1999; **55**: 597–602.
47. Cornfield J. A statistical problem arising from retrospective studies. *Pro Third Berkeley Sym Math Stat Probab* 1956; **4**: 135–148.
48. Lawson R. Small sample confidence intervals for the odds ratio. *Commun Stat Simul Comput* 2004; **33**(4): 1095–1113.
49. Baptista J and Pike MC. Exact two-sided confidence limits for the odds ratio in a $2 \times 2$ table. *J R Stat Soc C Appl Stat* 1977; **26**: 214–220.
50. Sterne TE. Some remarks on confidence or fiducial limits. *Biometrika* 1954; **41**: 275–278.
51. Lancaster HO. Significance tests in discrete distributions. *J Am Stat Assoc* 1961; **56**: 223–234.
52. Berry G and Armitage P. Mid-*p* confidence intervals: a brief review. *Stat* 1995; **44**: 417–423.
53. Mehta CR and Walsh SJ. Comparison of exact, mid-*p* and Mantel-Haenszel confidence intervals for the common odds ratio across several $2 \times 2$ contingency tables. *Am Stat* 1992; **46**: 146–150.
54. Lydersen S and Laake P. Power comparison of two-sided exact tests for association in $2 \times 2$ contingency tables using standard, mid-*p* and randomized test versions. *Stat Med* 2003; **22**: 3859–3871.
55. Hirji KF, Tan S-J and Elashoff RM. A quasi-exact test for comparing two binomial proportions. *Stat Med* 1991; **10**: 1137–1153.
56. Agresti A and Min Y. Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics* 2002; **3**: 379–386.