# 1 Absolute Risk Reduction

*Robert Newcomb*



## 1.1 Introduction

Many response variables in clinical trials are binary: the treatment was successful or unsuccessful; the adverse effect did or did not occur. Binary variables are summarized by proportions, which may be compared between different arms of a study by calculating either an absolute difference of proportions or a relative measure, the relative risk or the odds ratio. In this article we consider several point and interval estimates for the absolute difference between two proportions, for both unpaired and paired study designs. The simplest methods encounter problems when numerators or denominators are small; accordingly, better methods are introduced. Because confidence interval methods for differences of proportions are derived from related methods for the simpler case of the single proportion, which itself can also be of interest in a clinical trial, this case is also considered in some depth. Illustrative examples relating to data from two clinical trials are shown.

## 1.2 Preliminary Issues

In most clinical trials, the *unit of data* is the *individual*, and statistical analyses for efficacy and safety outcomes compare responses between the two (or more) treatment groups. When subjects are randomized between these groups, responses of subjects in one group are *independent* of those in the other group. This leads to *unpaired* analyses. *Crossover* and *split-unit* designs require *paired* analyses. These have many features in common with the unpaired analyses and will be described in the final section.

Thus, we study $n_1$ individuals in group 1 and $n_2$ individuals in group 2. Usually, all analyses are *conditional* on $n_1$ and $n_2$. Analyses conditional on $n_1$ and $n_2$ would also be appropriate in other types of prospective studies or in cross-sectional designs. (Some hypothesis testing procedures such as the Fisher exact test are conditional also on the total number of "successes" in the two groups combined. This alternative conditioning is inappropriate for confidence intervals for a difference of proportions; in particular in the event that no successes are observed in either group, this approach fails to produce an interval.) The outcome variable is *binary:* 1 if the event of interest occurs, 0 if it does not. (We do not consider here the case of an integer-valued outcome variable; typically, this involves the number of episodes of relapse or hospitalization, number of accidents, or similar events occurring

within a defined follow-up period. Such an outcome would instead be modeled by the Poisson distribution.) We observe that $r_1$ subjects in group 1 and $r_2$ subjects in group 2 experience the event of interest. Then the proportions having the event in the two groups are given by $p_1 = r_1/n_1$ and $p_2 = r_2/n_2$. If responses in different individuals in each group are independent, then the distribution of the number of events in each group is *binomial*.

Several *effect size measures* are widely used for comparison of two independent proportions:

Difference of proportions $p_1 - p_2$

Ratio of proportions (risk ratio or relative risk) $p_1/p_2$

Odds ratio $(p_1/(1-p_1))/(p_2/(1-p_2))$

In this article we consider in particular the *difference between two proportions*, $p_1 - p_2$, as a measure of effect size. This is variously referred to as the absolute risk reduction, risk difference, or success rate difference. Other articles in this work describe the risk ratio or relative risk and the odds ratio. We consider both point and interval estimates, in recognition that "confidence intervals convey information about magnitude and precision of effect simultaneously, keeping these two aspects of measurement closely linked" [1]. In the clinical trial context, a difference between two proportions is often referred to as an absolute risk reduction. However, it should be borne in mind that any term that includes the word "reduction" really presupposes that the direction of the difference will be a reduction in risk—such terminology becomes awkward when the anticipated benefit does not materialize, including the nonsignificant case when the confidence interval for the difference extends beyond the null hypothesis value of zero. The same applies to the relative risk reduction, $1 - p_1/p_2$. Whenever results are presented, it is vitally important that the direction of the observed difference should be made unequivocally clear. Moreover, sometimes confusing labels are used, which might be interpreted to mean something other than $p_1 - p_2$; for example, Hashemi et al. [2] refer to $p_1 - p_2$ as attributable risk. It is also vital to distinguish between relative and absolute risk reduction.

In clinical trials, as in other prospective and cross-sectional designs already described, each of the three quantities we have discussed may validly be used as a measure of effect size. The risk difference and risk ratio compare two proportions *from different perspectives*. A halving of risk will have much greater population impact for a common outcome than for an infrequent one. Schechtman [3] recommends that both a relative and an absolute measure should always be reported, with appropriate confidence intervals.

The odds ratio is discussed at length by Agresti [4]. It is widely regarded as having a special preferred status on account of its role in retrospective case-control studies and in logistic regression and meta-analysis. Nevertheless, it should not be regarded as having gold standard status as a measure of effect size for the $2 \times 2$ table [3,5].

## 1.3 Point and Interval Estimates for a Single Proportion

Before considering the difference between two independent proportions in detail, we first consider some of the issues that arise in relation to the fundamental task of estimating a *single proportion*. These issues have repercussions for the comparison of proportions because confidence interval methods for $p_1 - p_2$ are generally based closely on those for proportions. The single proportion is also relevant to clinical trials in its own right. For example, in a

clinical trial comparing surgical versus conservative management, we would be concerned with estimating the incidence of a particular complication of surgery such as postoperative bleeding, even though there is no question of obtaining a contrasting value in the conservative group or of formally comparing these.

The most commonly used estimator for the population proportion $\pi$ is the familiar *empirical estimate*, namely, the observed proportion $p = r/n$. Given $n$, the random variable $R$ denoting the number of subjects in which the response occurs has the binomial $B(n,p)$ distribution, with $Pr[R = r] = \{n!/r!(n - r!)\}\ p^r\ q^{n-r}$ where $q = 1 - p$. The simple empirical estimator is also the *maximum likelihood estimate* for the binomial distribution, and it is *unbiased*—in the usual statistical sense that the expectation of $R$ given $n$,

$$E[R|n] = \pi.$$

However, when $r = 0$, many users of statistical methods are uneasy with the idea that $p = 0$ is an unbiased estimate. The range of possible values for $\pi$ is the interval from 0 to 1. Generally, this means the open interval $0 < \pi < 1$, not the closed interval $0 \leq \pi \leq 1$, as usually it would already be known that the event sometimes occurs and sometimes does not. As the true value of $\pi$ cannot then be negative or zero but must be greater than zero, the notion that $p = 0$ should be regarded as an unbiased estimate of $\pi$ seems highly counterintuitive.

Largely with this issue in mind, alternative estimators known as *shrinkage estimators* are available. These generally take the form $p_\psi = (r + \psi)/(n + 2\psi)$ for some $\psi > 0$. The quantity $\psi$ is known as a *pseudo-frequency*. Essentially, $\psi$ observations are added to the number of "successes" and also to the number of "failures." The resulting estimate $p_\psi$ is intermediate between the empirical estimate $p = r/n$ and $\frac{1}{2}$, which is the midpoint and center of symmetry of the support scale from 0 to 1. The degree of shrinkage toward $\frac{1}{2}$ is great when $n$ is small and minor for large $n$. Bayesian analyses of proportions lead naturally to shrinkage estimators, with $\psi = 1$ and $\frac{1}{2}$ corresponding to the most widely used uninformative conjugate priors, the uniform prior $B(1,1)$ and the Jeffreys prior $B(\frac{1}{2}, \frac{1}{2})$.

It also is important to report *confidence intervals*, to express the uncertainty due to sampling variation and finite sample size. The simplest interval $p \pm z\times$ SE$(p)$, where SE$(p) = \sqrt{(pq/n)}$, remains the most commonly used. This is usually called the *Wald interval*. Here, $z$ denotes the relevant quantile of the standard Gaussian distribution. Standard practice is to use intervals that aim to have 95% coverage, with 2.5% noncoverage in each tail, leading to $z = 1.9600$.

Unfortunately, confidence intervals for proportions and their differences do not achieve their nominal coverage properties. This is because the sample space is *discrete* and *bounded*. The Wald method for the single proportion has three unfavorable properties [6–9]. These can all be traced to the interval's simple symmetry about the empirical estimate.

The achieved *coverage* is much lower than the nominal value. For some values of $\pi$, the achieved coverage probability is close to zero.

The noncoverage probabilities in the two tails are very different. The *location* of the interval is *too distal*—too far out from the center of symmetry of the scale, $\frac{1}{2}$. The noncoverage of the interval is predominantly *mesial*.

The calculated limits often *violate the boundaries* at 0 and 1. In particular, when $r = 0$, a degenerate, zero-width interval results. For small non-zero values of $r$ (1, 2, and sometimes 3 for a 95% interval), the calculated lower limit is below zero. The resulting interval is usually truncated at

zero, but this is unsatisfactory as the data tells us that 0 is an impossible value for $\pi$. Corresponding anomalous behavior at the upper boundary occurs when $n - r$ is 0 or small.

Many *improved methods* for confidence intervals for proportions have been developed. The properties of these methods are evaluated by choosing suitable *parameter space points* (here, combinations of $n$ and $\pi$), using these to generate large numbers of simulated random samples, and recording how often the resulting confidence interval includes the true value $\pi$. The resulting *coverage probabilities* are then summarized by calculating the *mean coverage* and *minimum coverage* across the simulated datasets.

Generally, the improved methods obviate the boundary violation problem, and improve coverage and location. The most widely researched options are as follows.

A *continuity correction* may be incorporated: $p \pm \{z\sqrt{(pq/n)} + 1/(2n)\}$. This certainly improves coverage and obviates zero-width intervals but increases the incidence of boundary overflow.

The Wilson *score method* [10] uses the theoretical value $\pi$, not the empirical estimate $p$, in the formula for the standard error of $p$. Lower and upper limits are obtained as the two solutions of the equation $p = \pi \pm z \times \mathrm{SE}(\pi) = \pi \pm z \times \sqrt{(\pi(1-\pi)/n)}$, which reduces to a quadratic in $\pi$. The two roots are given in closed form as

$$\{2p + z^2 \pm z\sqrt{(z^2 + 4rq)}\}/\{2(n + z^2)\}.$$

It is easily demonstrated [7] that the resulting interval is symmetrical on the logit scale—the other natural scale for proportions—by considering the product of the two roots for $\pi$, and likewise for $1 - \pi$. The resulting interval is boundary respecting and has appropriate mean coverage. In contrast to the Wald interval, location is rather too mesial.

The midpoint of the score interval, on the ordinary additive scale, is a shrinkage estimator with $\psi = (\frac{1}{2})z^2$, which is 1.92 for the default 95% interval. With this (and also Bayesian intervals) in mind, Agresti and Coull [8] proposed a *pseudo-frequency method*, which adds $\psi = 2$ to the numbers of successes ($r$) and failures ($n - r$) before using the ordinary Wald formula. This is also a great improvement over the Wald method, and is computationally and conceptually very simple. It reduces but does not eliminate the boundary violation problem. A variety of alternatives can be formulated, with different choices for $\psi$, and also using something other than $n + 2\psi$ as the denominator of the variance.

Alternatively, the *Bayesian* approach described elsewhere in this work may be used. The resulting intervals are best referred to as *credible intervals*, in recognition that the interpretation is slightly different from that of *frequentist* confidence intervals such as those previously described.

Bayesian inference starts with a *prior distribution* for the parameter of interest, in this instance the proportion $\pi$. This is then combined with the *likelihood function* comprising the evidence from the sample to form a *posterior distribution* that represents beliefs about the parameter after the data have been obtained. When a *conjugate prior* is chosen from the *beta distribution* family, the posterior distribution takes a relatively simple form: it is also a beta distribution. If substantial information about $\pi$ exists, an *informative prior* may be chosen to encapsulate this information.

More often, an *uninformative prior* is used. The simplest is the *uniform* prior $B(1,1)$, which assumes that all possible values of $\pi$ between 0 and 1 start off equally likely. An alternative uninformative prior with some advantages is the Jeffreys prior $B(\frac{1}{2}, \frac{1}{2})$. Both are *diffuse* priors, which spread the probability thinly across the

whole range of possible values from 0 to 1.

The resulting posterior distribution may be displayed graphically, or may be summarized by salient summary statistics such as the posterior mean and median and selected centiles. The $2\frac{1}{2}$ and $97\frac{1}{2}$ centiles of the posterior distribution delimit the *tail-based* 95% credible interval. Alternatively, a *highest posterior density* interval may be reported. The tail-based interval is considered preferable because it produces equivalent results when a transformed scale (e.g., logit) is used [11].

These Bayesian intervals perform well in a frequentist sense [12]. Hence, it is now appropriate to regard them as confidence interval methods in their own right, with theoretical justification in the Bayesian paradigm but empirical validation from a frequentist standpoint. They may thus be termed *beta intervals*. They are readily calculated using software for the incomplete beta function, which is included in statistical packages and also spreadsheet software such as Microsoft Excel. As such, they should now be regarded as computationally of "closed form," though less transparent than Wald methods.

Many statisticians consider that a coverage level should represent minimum, not average, coverage. The Clopper-Pearson "exact" or *tail-based method* [13] achieves this, at the cost of being excessively *conservative;* intervals are unnecessarily wide. There is a trade-off between coverage and width; it is always possible to increase coverage by widening intervals, and the aim is to attain good coverage without excessive width. A variant on the "exact" method involving a *mid-P* accumulation of tail probabilities [14,15] aligns mean coverage closely with the nominal $1 - \alpha$. Both methods have appropriate location. The Clopper-Pearson interval, but not the mid-$P$ one, is readily programmed as a beta interval, of similar form to Bayes intervals. A variety of *shortened* intervals have also been developed that maintain minimum coverage but substantially shrink interval length [16,17]. Shortened intervals are much more complex, both computationally and conceptually. They also have the disadvantage that what is optimized is the interval, not the lower and upper limits separately; consequently, they are unsuitable when interest centers on one of the limits rather than the other.

Numerical examples illustrating these calculations are based on some results from a very small randomized phase II clinical trial performed by the Eastern Cooperative Oncology Group [18]. Table 1 shows the results for two outcomes, treatment success defined as shrinkage of the tumor by 50% or more, and life-threatening treatment toxicity, for the two treatment groups A and B.

Table 2 shows 95% confidence intervals for both outcomes for treatment A. These examples show how Wald and derived intervals often produce inappropriate limits (see asterisks) in boundary and near-boundary cases.

## 1.4 An Unpaired Difference of Proportions

We return to the unpaired difference case. As described elsewhere in this work, hypothesis testing for the comparison of two proportions takes a quite different form according to whether the objective of the trial is to ascertain *difference* or *equivalence*. When we report the contrast between two proportions with an appropriately constructed confidence interval, this issue is taken into account only when we come to interpret the calculated point and interval estimates. In this respect, in comparison with hypothesis testing, the confidence interval approach leads to much simpler, more flexible patterns of inference.

The quantity of interest is the difference

Table 1: Some Results from a Very Small Randomized Phase II Clinical Trial Performed by the Eastern Cooperative Oncology Group

| | Treatment A | Treatment B |
|---|---|---|
| Number of patients | 14 | 11 |
| Number with successful outcome: tumor shrinkage by $\geq$50% | 0 | 0 |
| Number with life-threatening treatment toxicity | 2 | 1 |

*Source:* Parzen et al. *J Comput Graph Stat.* 2002: **11**; 420–436.

Table 2: 95% Confidence Intervals for Proportions of Patients with Successful Outcome and with Life-Threatening Toxicity on Treatment A in the Eastern Cooperative Oncology Group Trial

| Outcome | Successful Tumor Shrinkage | Life-Threatening Toxicity |
|---|---|---|
| Empirical estimate | 0 | 0.1429 |
| Wald interval | 0 to 0* | <0* to 0.3262 |
| Wald interval with continuity correction | <0* to 0.0357 | <0* to 0.3619 |
| Wilson score interval | 0 to 0.2153 | 0.0401 to 0.3994 |
| Agresti-Coull shrinkage estimate | 0.1111 | 0.2222 |
| Agresti-Coull interval | <0* to 0.2563 | 0.0302 to 0.4143 |
| Bayes interval, $B(1,1)$ prior | 0 to 0.2180 | 0.0433 to 0.4046 |
| Bayes interval, $B(\frac{1}{2},\frac{1}{2})$ prior | 0 to 0.1616 | 0.0309 to 0.3849 |
| Clopper-Pearson "exact" interval | 0 to 0.2316 | 0.0178 to 0.4281 |
| Mid-$P$ interval | 0 to 0.1926 | 0.0247 to 0.3974 |

*Note:* Asterisks denote boundary violations.
*Source:* Parzen et al. *J Comput Graph Stat.* 2002: **11**; 420–436.

between two binomial proportions, $\pi_1$ and $\pi_2$. The *empirical estimate* is $p_1 - p_2 = r_1/n_1 - r_2/n_2$. It is well known that, when comparing means, there is a direct correspondence between hypothesis tests and confidence intervals. Specifically, the null hypothesis is rejected at the conventional two-tailed $\alpha = 0.05$ level if and only if the $100(1 - \alpha) = 95\%$ confidence interval for the difference excludes the null hypothesis value of zero. A similar property applies also to the comparison of proportions—usually, but not invariably. This is because there are several options for constructing a confidence interval for the difference of proportions, which have different characteristics and do not all correspond directly to purpose-built hypothesis tests.

The *Wald interval* is calculated as $p_1 - p_2 \pm z\sqrt{(p_1q_1/n_1 + p_2q_2/n_2)}$. It has poor mean and minimum coverage and fails to produce an interval when both $p_1$ and $p_2$ are 0 or 1. Overshoot can occur when one proportion is close to 1 and the other is close to 0, but this situation is expected to occur infrequently in practice. Use of a continuity correction improves mean coverage, but minimum coverage remains low.

Several better methods have been developed, some of which are based on specific mathematical models. Any model for the comparison of two proportions necessarily involves both the *parameter of interest*, $\delta = \pi_1 - \pi_2$, and an additional *nuisance parameter* $\gamma$. The model may be parametrized in terms of $\delta$ and $\pi_1+\pi_2$, or $\delta$ and $(\pi_1+\pi_2)/2$, or $\delta$ and $\pi_1$. We will define the nuisance parameter as $\gamma = (\pi_1+\pi_2)/2$.

Some of the better methods substitute the *profile estimate* $\gamma_\delta$, which is the maximum likelihood estimate of $\gamma$ conditional on a hypothesized value of $\delta$. These include *score-type asymptotic intervals* developed by Mee [19] and Miettinen and Nurminen [20]. Newcombe [21] developed *tail-based exact and mid-P intervals* involving substitution of the profile estimate.

All these intervals are boundary respecting. The "exact" method aligns the minimum coverage quite well with the nominal $1 - \alpha$; the others align mean coverage well with $1-\alpha$, at the expense of fairly complex iterative calculation.

*Bayesian* intervals for $p_1 - p_2$ and other comparative measures may be constructed [2,11], but they are computationally much more complex than in the single proportion case, requiring use of numerical integration or computer-intensive methodology such as Markov chain Monte Carlo (MCMC) methods. It may be more appropriate to incorporate a prior for $p_1 - p_2$ itself rather than independent priors for $p_1$ and $p_2$ [22]. The Bayesian formulation is readily adapted to incorporate functional constraints such as $\delta \geq 0$ [22]. Walters [23] and Agresti and Min [11] have shown that Bayes intervals for $p_1 - p_2$ with uninformative beta priors have favorable frequentist properties.

Two computationally simpler, effective approaches have been developed. Newcombe [21] also formulated *square-and-add* intervals for differences of proportions. The concept is a very simple one. Assuming independence, the variance of a difference between two quantities is the sum of their variances. In other words, standard errors "square and add"—they combine in the same way that differences in $x$ and in $y$ coordinates combine to give the Euclidean distance along the diagonal, as in Pythagoras' theorem. This is precisely how the Wald interval for $p_1 - p_2$ is constructed. The same principle may be applied starting with other, better intervals for $p_1$ and $p_2$ separately. The Wilson score interval is a natural choice as it already involves square roots, though squaring and adding would work equally effectively starting with, for instance, tail-based [24] or Bayes intervals. It is easily demonstrated that the square-and-add process preserves the property of respecting boundaries.

Thus, the square-and-add interval is obtained as follows. Let $(l_i, u_i)$ denote the score interval for $p_i$, for $i = 1,2$. Then the square-and-add limits are

$$p_1 - p_2 - \sqrt{\{(p_1 - l_1)^2 + (u_2 - p_2)^2\}},$$
$$p_1 - p_2 + \sqrt{\{(u_1 - p_1)^2 + (p_2 - l_2)^2\}}.$$

This easily computed interval aligns mean coverage closely with the nominal $1 - \alpha$. A continuity correction is readily incorporated, resulting in more conservative coverage. Both intervals tend to be more mesially positioned than the $\gamma_\delta$-based intervals discussed previously.

The square-and-add approach may be applied a second time to obtain a confidence interval for a difference between differences of proportions [25]; this is the linear scale analogue of assessing an *interaction* effect in logistic regression.

Another simple approach that is a great improvement over the Wald method is the *pseudo-frequency* method [26,27]. A pseudo-frequency $\psi$ is added to each of the four cells of the $2 \times 2$ table, resulting in the *shrinkage estimator* $(r_1 + \psi)/(n_1 + 2\psi) - (r_2 + \psi)/(n_2 + 2\psi)$.

The Wald formula then produces the limits

$$p_{\psi 1} - p_{\psi 2} \pm z\sqrt{\{p_{\psi 1}(1 - p_{\psi 1})/(n_1 + 2\psi) + p_{\psi 2}(1 - p_{\psi 2})/(n_2 + 2\psi)\}},$$

where

$$p_{\psi i} = (r_i + \psi)/(n_i + 2\psi)$$
$$i = 1, 2.$$

Agresti and Caffo [27] evaluated the effect of choosing different values of $\psi$, and they reported that adding 1 to each cell is optimal here. So here, just as for the single proportion case, in total four pseudo-observations are added. This approach also aligns mean coverage effectively with $1 - \alpha$. Interval location is rather too mesial, very similar to that of the square-and-add method. Zero-width intervals cannot occur. Boundary violation is not ruled out but is expected to be infrequent.

Table 3 shows 95% confidence intervals calculated by these methods, comparing treatments A and B in the ECOG trial [18].

## 1.5 Number Needed to Treat

In the clinical trial setting, it has become common practice to report the *number needed to treat (NNT)*, defined as the reciprocal of the absolute risk difference: NNT $= 1/(p_1 - p_2)$ [28,29]. This measure has considerable intuitive appeal, simply because we are used to assimilating proportions expressed in the form of "1 in $n$," such as a 1 in 7 risk of life-threatening toxicity for treatment A in Table 1.

The same principle applies to differences of proportions. These tend to be small decimal numbers, often with a leading zero after the decimal point, which risk being misinterpreted by the less numerate. Thus if $p_1 = 0.35$ and $p_2 = 0.24$, we could equivalently report $p_1 - p_2 = 0.11$, or as an absolute difference of 11% or an NNT of 9. The latter may well be an effective way to summarize the information when a clinician discusses a possible treatment with a patient. As always, we need to pay careful attention to the direction of the difference. By default, NNT is read as "number needed to treat for (one person to) benefit," or NNTB. If the intervention of interest proves to be worse than the control regime, we report the number needed to harm (NNTH).

A confidence interval for the NNT may be derived from any good confidence interval method for $p_1 - p_2$ by *inverting* the two limits. For example, Bender [30] suggests an interval obtained by inverting square-and-add limits [21]. But it is when we turn attention to confidence intervals that the drawback of the NNT approach becomes apparent. Consider first the case

Table 3: 95% Confidence Intervals for Differences in Proportions of Patients with Successful Outcome and with Life-Threatening Toxicity between Treatments A and B in the Eastern Cooperative Oncology Group Trial

| Outcome | Successful Tumor Shrinkage | Life-Threatening Toxicity |
|---|---|---|
| Empirical estimate | 0 | 0.0519 |
| Wald interval | 0* to 0* | −0.1980 to 0.3019 |
| Mee interval | −0.2588 to 0.2153 | −0.2619 to 0.3312 |
| Miettinen-Nurminen interval | −0.2667 to 0.2223 | 0.2693 to 0.3374 |
| Tail-based "exact" interval | −0.2849 to 0.2316 | −0.2721 to 0.3514 |
| Tail-based mid-$P$ interval | −0.2384 to 0.1926 | −0.2539 to 0.3352 |
| Bayes interval, B(1,1) priors for $p_1$ and $p_2$ | −0.2198 to 0.1685 | −0.2432 to 0.2986 |
| Bayes interval, B($\frac{1}{2}$,$\frac{1}{2}$) priors for $p_1$ and $p_2$ | −0.1768 to 0.1361 | −0.2288 to 0.3008 |
| Square-and-add Wilson interval | −0.2588 to 0.2153 | −0.2524 to 0.3192 |
| Agresti-Caffo shrinkage estimate | −0.0144 | 0.0337 |
| Agresti-Caffo interval | −0.2016 to 0.1728 | −0.2403 to 0.3076 |

*Note:* Asterisks denote boundary violations.
*Source:* Parzen et al. *J Comput Graph Stat.* 2002; **11**; 420–436.

of a statistically significant difference, with $p_1 - p_2 = +0.25$, and 95% confidence interval from +0.10 to +0.40. Then an NNT of 4 is reported, with 95% confidence interval from 2.5 to 10. This has two notable features. The lower limit for $p_1 - p_2$ gives rise to the upper limit for the NNT and vice versa. Furthermore, the interval is very skewed, and the point estimate is far from the midpoint. Neither of these is a serious contraindication to use of the NNT.

But often the difference is not statistically significant—and, arguably, reporting confidence intervals is even more important in this case than when the difference is significant. Consider, for example, $p_1 - p_2 = +0.10$, with 95% confidence interval from −0.05 to +0.25. Here, the estimated NNT is $1/0.10 = +10$. Inverting the lower and upper confidence limits for $p_1 - p_2$ gives −20 and +4. This time, the two limits do not change places apparently. But there are two problems. The point estimate, +10, is not intermediate between −20 and +4. Moreover, the interval from −20 to +4 does not comprise the values of the NNT that are compatible with the data, but rather the ones that are not compatible with it. In fact, the confidence region for the NNT in this case consists of two intervals that extend to infinity, one from +4 to $+\infty$ in the direction of benefit, the other from −20 to $-\infty$ in the direction of harm. It could be a challenge to clinicians and researchers at large to comprehend this *singularity* that arises when a confidence interval spanning 0 is inverted [31].

Accordingly, it seems preferable to report absolute risk reductions in percentage rather than reciprocal form. The most appropriate uses of the NNT are in giving simple bottomline figures to patients (in which situation, usually only the point estimate would be given), and in labeling a secondary axis on a graph.

## 1.6 A Paired Difference of Proportions

*Crossover* and *split-unit* trial designs lead to paired analyses. Regimes that aim to produce a cure are generally not suitable for evaluation in these designs, because in the event that a treatment is effective, there would be a carryover effect into the next treatment period. For this reason, these designs tend to be used for evaluation of regimes that seek to control symptomatology, and thus most often give rise to continuous outcome measures. Examples of paired analyses of binary data in clinical trials include comparisons of different antinauseant regimes administered in randomized order during different cycles of chemotherapy, comparisons of treatments for headache pain, and split-unit studies in ophthalmology and dermatology. Results can be reported in either risk difference or NNT form, though the latter appears not to be frequently used in this context. Other examples in settings other than clinical trials include longitudinal comparison of oral carriage of an organism before and after third molar extraction, and twin studies.

Let $a$, $b$, $c$, and $d$ denote the four cells of the paired contingency table. Here, $b$ and $c$ are the *discordant cells*, and interest centers on the *difference of marginals:*

$$\begin{aligned} p_1 - p_2 &= (a+b)/n - (a+c)/n \\ &= (b-c)/n. \end{aligned}$$

Hypothesis testing is most commonly performed using the *McNemar* approach [32], using either an asymptotic test statistic expressed as $z$ or chi-square, or an aggregated tail probability. In both situations, inference is conditional on the total number of discordant pairs, $b + c$.

Newcombe [33] reviewed confidence interval methods for the paired difference case. Many of these are closely analogous to unpaired methods. The Wald interval performs poorly. So does a conditional approach, based on an interval for the simple proportion $b/(b+c)$. Exact and tail-based profile methods perform well; although, as before, these are computationally complex. A closed-form square-and-add approach, modified to take account of the nonindependence, also aligns mean coverage with $1-\alpha$, provided that a novel form of continuity correction is incorporated.

Tango [34] developed a score interval, which is boundary respecting and was subsequently shown to perform excellently [35]. Several further modifications were suggested by Tang, Tang, and Chan [36]. Agresti and Min [11] proposed pseudo-frequency methods involving adding $\psi = 0.5$ to each cell and demonstrated good agreement of mean coverage with $1-\alpha$. However, overshoot can occasionally occur.

The above methods are appropriate for a paired difference of proportions. But for crossover and simultaneous split-unit studies, a slightly different approach is preferable. Thus, in a crossover study, if the numbers of subjects who get the two treatment sequences AB and BA are not identical, the simple difference of marginals contains a contribution from period differences. A more appropriate analysis is based on the analysis of differences of paired differences described by Newcombe [25]. The example in Table 4 relates to a crossover trial of home versus hospital physiotherapy for chronic multiple sclerosis [37]. Twenty-one patients were randomized to receive home physiotherapy followed by hospital physiotherapy, and 19 to receive these treatments in the reverse order. Following Hills and Armitage [38] and Koch [39], the treatment effect is estimated as half the difference between the within-subjects period differences in the two treatment order groups. The resulting estimate, +0.1454 and 95% confidence interval, −0.0486 to +0.3238, are very similar but not identi-

Table 4: Crossover Trial of Home Versus Hospital Physiotherapy: Treating Physiotherapist's Assessment of Whether the Patient Benefited from Either Type of Treatment

| | Number of Patients Benefiting From: | | | | Difference (First Minus sec377-ond Treatment) | |
|---|---|---|---|---|---|---|
| Treatment Sequence | Both | First | sec377-ond | Neither | Estimate | 95% Confidence Interval |
| Home versus hospital | 11 | 6 | 1 | 3 | +0.2381 | −0.0127 to +0.4534 |
| Hospital versus home | 9 | 4 | 5 | 1 | −0.0526 | −0.3372 to +0.2434 |
| Difference | | | | | +0.2907 | −0.0973 to +0.6475 |
| Half | | | | | +0.1454 | −0.0486 to +0.3238 |

*Source:* R. G. Newcombe, Estimating the difference between differences: measurement of additive scale interaction for proportions. *Stat Med.* 2001; **20**: 2885–2893. Reproduced with permission.

cal to those obtained by direct application of the modified square-and-add approach [33], +0.1500 and −0.0488 to +0.3339.

## References


[1] K. Rothman, *Modern Epidemiology.* Boston: Little, Brown, 1986.

[2] L. Hashemi, B. Nandram, and R. Goldberg, Bayesian analysis for a single 2 × 2 table. *Stat Med.* 1997; **16**: 1311–1328.

[3] E. Schechtman, Odds ratio, relative risk, absolute risk reduction, and the number needed to treat—which of these should we use? *Value Health.* 2002; **5**: 431–436.

[4] A. Agresti, *Categorical Data Analysis*, 2nd ed. Hoboken, NJ: Wiley, 2002.

[5] R. G. Newcombe, A deficiency of the odds ratio as a measure of effect size. *Stat Med.* 2006; **25**: 4235–4240.

[6] S. E. Vollset, Confidence intervals for a binomial proportion. *Stat Med.* 1993; **12**: 809–824.

[7] R. G. Newcombe, Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med.* 1998; **17**: 857–872.

[8] A. Agresti and B. A. Coull, Approximate is better than "exact" for interval estimation of binomial proportions. *Am Stat.* 1998; **52**: 119–126.

[9] L. D. Brown, T. T. Cai, and A. DasGupta, Interval estimation for a proportion. *Stat Sci.* 2001; **16**: 101–133.

[10] E. B. Wilson, Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc.* 1927; **22**: 209–212.

[11] A. Agresti and Y. Min, Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 × 2 contingency tables. *Biometrics.* 2005; **61**: 515–523.

[12] B. P. Carlin and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman & Hall, 1996.

[13] C. J. Clopper and E. S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika.* 1934; **26**: 404–413.

[14] H. O. Lancaster, The combination of probabilities arising from data in discrete distributions. *Biometrika.* 1949; **36**: 370–382.

[15] G. Berry and P. Armitage, Mid-P confidence intervals: a brief review. *Statistician.* 1995; **44**: 417–423.

[16] H. Blaker, Confidence curves and improved exact confidence intervals for discrete distributions. *Can J Stat.* 2000; **28**: 783–798.

[17] J. Reiczigel, Confidence intervals for the binomial parameter: some new considerations. *Stat Med.* 2003: **22**; 611–621.

[18] M. Parzen, S. Lipsitz, J. Ibrahim, and N. Klar, An estimate of the odds ratio that always exists. *J Comput Graph Stat.* 2002: **11**; 420–436.

[19] R. W. Mee, Confidence bounds for the difference between two probabilities. *Biometrics.* 1984; **40**: 1175–1176.

[20] O. S. Miettinen and M. Nurminen, Comparative analysis of two rates. *Stat Med.* 1985; **4**: 213–226.

[21] R. G. Newcombe, Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med.* 1998; **17**: 873–890.

[22] R. G. Newcombe, Bayesian estimation of false negative rate in a clinical trial of sentinel node biopsy. *Stat Med.* 2007; **26**: 3429–3442.

[23] D. E. Walters, On the reliability of Bayesian confidence limits for a difference of two proportions. *Biom. J.* 1986; **28**: 337–346.

[24] T. Fagan, Exact 95% confidence intervals for differences in binomial proportions. *Comput Biol Med.* 1999; **29**: 83–87.

[25] R. G. Newcombe, Estimating the difference between differences: measurement of additive scale interaction for proportions. *Stat Med.* 2001; **20**: 2885–2893.

[26] W. W. Hauck and S. Anderson, A comparison of large-sample confidence interval methods for the difference of two binomial probabilities. *Am Stat.* 1986; **40**: 318 322.

[27] A. Agresti and B. Caffo, Simple and effective confidence intervals for proportions and differences of proportions result from adding 2 successes and 2 failures. *Am Stat.* 2000; **54**: 280–288.

[28] A. Laupacis, D. L. Sackett, and R. S. Roberts, An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med.* 1988; **318**: 1728–1733.

[29] D. G. Altman, Confidence intervals for the number needed to treat. *BMJ.* 1998; **317**: 1309–1312.

[30] R. Bender, Calculating confidence intervals for the number needed to treat. *Control Clin Trials.* 2001; **22**: 102–110.

[31] R. G. Newcombe, Confidence intervals for the number needed to treat—absolute risk reduction is less likely to be misunderstood. *BMJ.* 1999; **318**: 1765.

[32] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* 1947; **12**: 153–157.

[33] R. G. Newcombe, Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat Med.* 1998; **17**: 2635–2650.

[34] T. Tango, Equivalence test and CI for the difference in proportions for the paired-sample design. *Stat Med.* 1998; **17**: 891–908.

[35] R. G. Newcombe, Confidence intervals for the mean of a variable taking the values 0, 1 and 2. *Stat Med.* 2003; **22**: 2737–2750.

[36] M. L. Tang, N. S. Tang, and I. S. F. Chan, Confidence interval construction for proportion difference in small sample paired studies. *Stat Med.* 2005; **24**: 3565–3579.

[37] C. M. Wiles, R. C. Newcombe, K. J. Fuller, S. Shaw, J. Furnival-Doran, et al., Controlled randomised crossover trial of physiotherapy on mobility in chronic multiple sclerosis. *J Neurol Neurosurg Psych.* 2001; **70**: 174 179.

[38] M. Hills and P. Armitage, The two-period cross-over clinical trial. *Br J Clin Pharmacol.* 1979; **8**: 7–20.

[39] G. G. Koch, The use of non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics.* 1972; **28**: 577–584.

## Further Reading

[1] Microsoft Excel spreadsheets that implement chosen methods for the single

proportion, unpaired and paired difference, and interaction cases can be downloaded from the author's website: http://www.cardiff.ac.uk/medicine/epidemiology_statistics/research/statistics/newcombe

[2] The availability of procedures to calculate confidence intervals for differences of proportions is quite patchy in commercial software. StatXact (Cytel Statistical Software) includes confidence intervals for differences and ratios of proportions and odds ratios. These are "exact" intervals, designed to guarantee minimum coverage $1 - \alpha$. The resulting intervals are likely to be relatively wide compared with methods that seek to align the mean coverage approximately with $1 - \alpha$.