Medical Infobahn for Europe
A. Hasman et al. (Eds.)
IOS Press, 2000

Improving the calculation of confidence intervals for the number needed to treat

Ralf Bender
Institute of Epidemiology and Medical Statistics
School of Public Health, University of Bielefeld,
P.O. Box 100131, D-33501 Bielefeld, Germany

Abstract. The number needed to treat has gained much attention in the past years as a useful way of reporting the results of randomised controlled trials with a binary outcome. Defined as reciprocal of the absolute risk reduction the number needed to treat is the estimated number of patients who need to be treated to prevent an adverse outcome in one additional patient. As with other estimated effect measures, it is important to document the uncertainty of the estimation by means of an appropriate confidence interval. Confidence intervals for the number needed to treat can be obtained by inverting and exchanging the confidence limits for the absolute risk reduction. Unfortunately, the only method used in practice for calculating a confidence interval for the absolute risk reduction seems to be the usual asymptotic method, which yields confidence intervals which are too short in many cases. In this paper it is shown that the application of the Wilson score method improves the calculation and presentation of confidence intervals for the number needed to treat.

1. Background

The number needed to treat (NNT) has gained much attention in the past years as a useful way of reporting the results of randomised controlled trials with a binary outcome [1]. Defined as reciprocal of the absolute risk reduction (ARR) the number needed to treat is the estimated number of patients who need to be treated to prevent an adverse outcome in one additional patient. A negative NNT is the estimated number of patients who need to be treated with the new rather than the standard treatment for one additional patient to be harmed. While this measure is better understood than risk ratios or risk reductions by clinicians and patients, the NNT has undesirable mathematical and statistical properties. The understanding of the confidence interval for NNT is not straightforward. However, an excellent explanation was recently given by Altman [2]. The mathematical and statistical properties of the NNT are described in more detail by Lesaffre and Pledger [3].

The key to understand the confidence interval for NNT is that principally the domain of NNT is the union of 1 to ∞ and $-\infty$ to -1. Values between -1 and 1 are impossible for NNT. The best value of NNT indicating the largest possible beneficial effect of a new treatment is 1, which means that one patient has to be treated to prevent one adverse outcome. The NNT value indicating no treatment effect (ARR=0) is $\pm\infty$, and the worst NNT value indicating the largest possible harmful effect is -1. Thus, a NNT of 10 with confidence interval of 4 to -20 means that the two regions of 4 to ∞ and of -20 to $-\infty$ form the confidence interval.

Altman recommended that a confidence interval should always be given when a NNT is reported as study result [2]. However, the usual method of calculating such confidence intervals can be improved markedly. In the medical literature concerning NNT it is stated that a confidence interval for NNT can be obtained by inverting and exchanging the confidence limits for the ARR [1]. Unfortunately, the only method used in practice for calculating a confidence interval for ARR seems to be the usual asymptotic method [2,4]. In the statistical literature it is well known that the usual asymptotic method yields confidence intervals for the ARR which are too short in many cases [5-7]. Recently, Newcombe proposed a method based upon Wilson score intervals [7]. This method was strongly recommended over the usual asymptotic method and is applied in this paper. By using artificial examples it is shown that the application of the Wilson score method improves the calculation and presentation of confidence intervals for the number needed to treat.

2. Methodology

Let π_1 and π_2 be the true probabilities (risks) of an adverse event in the control group (group 1) and the treatment group (group 2), respectively. The true absolute risk reduction is simply the difference of the two risks $ARR=\pi_1-\pi_2$. The true number needed to treat is the inverse of ARR, i.e. $NNT=1/ARR=1/(\pi_1-\pi_2)$. To estimate these measures a randomised clinical trial can be performed. Let n_1 and n_2 be the number of patients randomised in the control group and the treatment group, respectively, and let e_1 and e_2 be the number of patients having an event in the control group and the treatment group, respectively. The two risks can then be estimated by means of $p_1=e_1/n_1$ and $p_2=e_2/n_2$. An estimate of the absolute risk reduction is given by $ARR=p_1-p_2$ and NNT can be estimated by $NNT=1/(p_1-p_2)$. As noted before, an approximate confidence interval for NNT can be obtained by inverting and exchanging the confidence limits for the ARR. Hence, we concentrate on the confidence interval calculation for ARR. The standard method of calculating confidence intervals for ARR makes use of the asymptotic normality and the usual formula for the standard error of the estimated ARR [4].

While the usual asymptotic method is adequate for large sample size and large ARR values it yields too short confidence intervals in many cases [5-7]. However, in practice, confidence intervals for NNT – if at all – are calculated by applying the usual asymptotic method. Exact confidence intervals for ARR are now provided by StatXact. However, exact methods for interval estimation of binomial proportions are conservative, i.e. yield confidence intervals which are unnecessarily too wide [8].

It has been shown that confidence intervals based upon Wilson scores have coverage probabilities close to the nominal confidence level [7-9]. Moreover, they are easier to calculate than exact confidence intervals because it is only required to solve a quadratic equation. Hence, after investigating eleven methods to calculate confidence intervals for ARR, Newcombe proposed to use the Wilson score method for interval estimation of ARR [7]. Explicit formulas for the lower (LL) and upper limits (UL) of the confidence interval for ARR based upon Wilson scores can be found elsewhere [10]. For calculations a SAS/IML [11] program can be used which is available from the author on request.

3. Comparison of the Asymptotic and the Wilson Sore Method

The usual asymptotic method to calculate confidence intervals for the absolute risk reduction has poor coverage characteristics and a propensity to aberrations [7]. Especially in small samples and ARR values close to 0 or 1 the asymptotic method leads to unreliable results. For estimated ARR values of 0 or 1 the asymptotic method gives no meaningful confidence interval. These shortcomings have considerable importance in trials with low treatment effect and equivalence trials. In these situations where it is particularly important to quantify the uncertainty of estimations the usual asymptotic method fails. In the following the confidence intervals based on Wilson scores are compared with the usual asymptotic confidence intervals by means of artificial examples.

Table 1: Confidence intervals calculated by the usual asymptotic method and by the Wilson score method for NNT values of artificial examples

	Description of examples	Control group			Treatment group			ARR	NNT	95% CIs	
		e_1	n_1	p_1	e_2	n_2	p_2			As.	Wilson
1	as. UL unreliable (too low)	10	200	0.050	3	200	0.015	0.035	28.6	14 to 2411	13 to –890
2	as. LL theoretically impossible (<1)	6	7	0.857	1	7	0.143	0.714	1.4	0.9 to 2.9	1.1 to 5.2
3	no meaningful as. CI if ARR=1	5	5	1.000	0	5	0.000	1.000	1.0	1.0 to 1.0	1.0 to 2.6
4	no meaningful as. CI if p_1=p_2=0	0	100	0.000	0	100	0.000	0.000	∞	∞ to ∞	27 to -27
5	as. CI inadequate to prove equivalence	1	500	0.002	2	500	0.004	-0.002	-500	209 to –114	130 to –79

abbreviations: as. = asymptotic, ARR = absolute risk reduction, CI = confidence interval, NNT = number needed to treat, UL = upper limit, LL = lower limit,

The deficiencies of the usual asymptotic method are pointed out by the examples presented in Table 1. For high NNT estimates and moderate sample size especially the upper asymptotic confidence limit is unreliable (example 1). The usual asymptotic method leads to several aberrations. Low NNT estimates and low sample size can lead to a theoretically impossible lower confidence limit (example 2). If the ARR estimate is exactly 1 no meaningful confidence interval can be calculated by using the asymptotic method because the standard error of ARR is erroneously 0 (example 3). The same holds when both risk estimates are exactly zero (example 4). The Wilson score method overcomes all these deficiencies.

The possible aberrations of the usual asymptotic method to calculate confidence intervals for ARR and NNT are meaningful especially for equivalence trials [12]. In such trials one possibility to demonstrate equivalence between treatments is to show that the 95% confidence interval of the effect measure lies entirely in a predefined range of equivalence. For example, a possible equivalence region for NNT could be the interval of 100 to –100. This means that the two treatments are equivalent if 100 or more patients are needed to be treated for one patient to benefit as well as to be harmed. If both treatments are highly effective to prevent patients from adverse events the number of observed events in the study will be low. If the number of observed events is zero in both groups the standard method gives no meaningful confidence interval at all (example 4). If the number of observed events is low in both groups, the usual asymptotic confidence interval is unreliable even for large sample size. In example 5 the confidence interval of 209 to –114 calculated by the asymptotic method would lead to the decision of equivalence. This decision, however, is questionable because the asymptotic confidence interval is probably too short shown by the Wilson score confidence interval of 130 to –79. This means that there may be up to 1 of 79 treated patients who is harmed instead of 1 of 100 treated patients. Thus, the upper confidence limit exceeds the equivalence limit of NNT = –100. Hence, if NNT is used as effect measure in equivalence trials the usual asymptotic method of calculating confidence intervals for ARR and NNT should not be applied even in the case of large sample sizes.

4. Conclusion

In the current medical literature the calculation and reporting of confidence intervals for the number needed to treat is quite unsatisfactory. A systematic search through all issues of the journal *Evidence-Based Medicine* (1995-1999) revealed that NNT estimates with confidence intervals are given only for significant results [10]. The only method routinely used in practice seems to be the inverting and exchanging of the usual asymptotic confidence limits for the absolute risk reduction. This procedure, however, leads to unreliable confidence intervals for the number needed to treat in a many cases, especially in studies with low sample size, low absolute risk reduction, and equivalence trials. The application of the Wilson score method leads to confidence intervals for the number needed to treat which have much better coverage properties, are free of aberrations, and are quite easier to calculate than exact confidence intervals. Any estimated number needed to treat should be complemented by an adequate confidence interval and the calculation method should be stated. It is recommended to replace the usual asymptotic method to calculate confidence intervals for the number needed to treat by the Wilson score method or another method with adequate coverage properties [7].

References

[1] R.J. Cook and D.L. Sackett, The number needed to treat: A clinically useful measure of treatment effect, *British Medical Journal* **310** (1995) 452-454.

[2] D.G. Altman, Confidence intervals for the number needed to treat. *British Medical Journal* **317** (1998) 1309-1312.

[3] E. Lesaffre and G. Pledger, A note on the number needed to treat, *Controlled Clinical Trials* **20** (1999) 439-447.

[4] Daly, L.E. (1998): Confidence limits made easy: Interval estimation using a substitution method. *American Journal of Epidemiology* **147**, 783-790.

[5] S.L. Beal, Asymptotic confidence intervals for the difference between binomial parameters for the use with small samples, *Biometrics* **43** (1987) 941-950.

[6] S. Wallenstein, A non-iterative accurate asymptotic confidence interval for the difference between two proportions, *Statistics in Medicine* **16** (1997) 1329-1336.

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

[8] A. Agresti and B.A. Coull, Approximate is better than "exact" for interval estimation of binomial proportions, *American Statistician* **52** (1998) 119-126.

[9] S.E. Vollset, Confidence intervals for a binomial proportion, *Statistics in Medicine* **12** (1993) 809-824.

[10] R. Bender, Calculating confidence intervals for the number needed to treat. *Controlled Clinical Trials* (submitted for publication).

[11] SAS, SAS/IML User's Guide, Version 5 Edition. SAS Institute Inc., Cary, NC, 1985.

[12] B. Jones, P. Jarvis, J.A. Lewis and A.F. Ebbutt, Trials to assess equivalence: The importance of rigorous methods, *British Medical Journal* **313** (1996) 36-39.