Calculating Confidence Intervals for the Number Needed to Treat

Ralf Bender, PhD

Department of Epidemiology and Medical Statistics, School of Public Health, University of Bielefeld, Bielefeld, Germany

ABSTRACT: The number needed to treat (NNT) has gained much attention in the past years as a useful way of reporting the results of randomized controlled trials with a binary outcome. Defined as the reciprocal of the absolute risk reduction (ARR), NNT is the estimated average number of patients needed 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 NNT can be obtained by inverting and exchanging the confidence limits for the ARR provided that the NNT scale ranging from 1 through ∞ to -1 is taken into account. Unfortunately, the only method used in practice to calculate confidence intervals for ARR seems to be the simple Wald method, which yields too short confidence intervals 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. *Control Clin Trials* 2001;22:102–110 © Elsevier Science Inc. 2001

KEY WORDS: *Absolute risk reduction, confidence interval, evidence-based medicine, equivalence, number needed to treat*

INTRODUCTION

The number needed to treat (NNT) has gained much attention in the past years as a useful way of reporting the results of randomized controlled trials with a binary outcome [1–3]. Defined as the reciprocal of the absolute risk reduction (ARR), the number needed to treat is the estimated average number of patients needed to be treated to prevent an adverse outcome in one additional patient. A negative NNT is the estimated average number of patients needed to be treated with the new rather than the standard treatment for one additional patient to be harmed. While this measure is often 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

Address reprint requests to: Ralf Bender, PhD, Department of Epidemiology and Medical Statistics, School of Public Health, University of Bielefeld, P.O. Box 100131, D-33501 Bielefeld, Germany (ralf.bender@uni-bielefeld.de)

Received June 27, 2000; accepted November 28, 2000.

Controlled Clinical Trials 22:102–110 (2001)
© Elsevier Science Inc. 2001
655 Avenue of the Americas, New York, NY 10010

0197-2456/01/$–see front matter
PII S0197-2456(00)00134-3

was recently given by Altman [4]. The mathematical and statistical properties of the NNT statistic are described in more detail by Lesaffre and Pledger [5].

The key to understanding the confidence interval for NNT is that principally the domain of NNT is the union of 1 to ∞ and $-\infty$ to -1. The best value of NNT indicating the largest possible beneficial treatment effect is 1, 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, the result NNT = 10 with confidence limits 4 and -20 means that the two regions 4 to ∞ and -20 to $-\infty$ form the confidence interval. Altman proposed to use two new abbreviations, namely number needed to treat for one patient to benefit (NNTB) or be harmed (NNTH) [4]. This concept avoids the awkward term "number needed to harm" (NNH), which is used, for example, in the journal *Evidence-Based Medicine.* The result of an estimated NNT with confidence interval can then be presented as NNTB = 10 (NNTB 4 to ∞ to NNTH 20) [4].

Altman recommended that a confidence interval should always be given when an NNT is reported as a study result [4]. However, the usual Wald method for calculating such confidence intervals is frequently inappropriate. By using examples from the literature and artificial examples, it is shown that the application of the Wilson score method [6] improves the calculation and presentation of confidence intervals for the number needed to treat.

METHODS TO CALCULATE CONFIDENCE INTERVALS FOR NNT

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 ARR is the difference of the two risks $\pi_1 - \pi_2$. The true NNT is the reciprocal $1/(\pi_1 - \pi_2)$ of the true ARR. To estimate these measures a randomized clinical trial can be performed. Let n_1 and n_2 be the number of patients randomized 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 the proportions $p_1 = e_1/n_1$ and $p_2 = e_2/n_2$. The true effect measures can be estimated by ARR = $p_1 - p_2$ and NNT = $1/(p_1 - p_2)$.

Under regularity conditions (continuity, one-to-one transformation) a confidence interval for NNT can be obtained by inverting and exchanging the confidence limits for ARR [7]. Let LL(ARR) and UL(ARR) be the lower and upper confidence limits for ARR, then the confidence interval for NNT can be expressed as [1/UL(ARR),1/LL(ARR)]. However, it should be recognized that the continuity condition is violated for the reciprocal transformation if the confidence interval for ARR encloses 0. In this case, the confidence interval for NNT is the union $(-\infty,1/\text{LL(ARR)}]\cup[1/\text{UL(ARR)},\infty)$ of two half intervals [4, 5]. One possibility to take the violation of the continuity condition into account is Altman's suggestion to write the confidence interval for an estimated positive NNT value as "NNTB 1/UL(ARR) to ∞ to NNTH 1/LL(ARR)" [4]. Thus, confidence limits for NNT can be calculated from confidence limits for ARR in all cases. Hence, we concentrate on the interval estimation of ARR.

The standard method of calculating confidence intervals for ARR makes use of the asymptotic normality and the usual formula for the standard error (SE) of the estimated ARR. Using the notations above the estimated standard error of $p_1 - p_2$ is given by:

$$\mathrm{SE}(p_1 - p_2) = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \tag{1}$$

Let $z_{1-\alpha/2}$ be the $1 - \alpha/2$ quantile of the standard normal distribution. The simple Wald-type $100 \times (1 - \alpha)\%$ confidence interval for ARR is then given by:

$$p_1 - p_2 \pm z_{1-\alpha/2} \times \mathrm{SE}\ (p_1 - p_2) \tag{2}$$

While Wald confidence intervals are adequate for large sample sizes and probabilities not close to 0 or 1, they have poor coverage characteristics and a propensity to aberrations in many practical situations. Especially in small samples, unbalanced designs, and probabilities close to 0 or 1, the Wald method leads to unreliable or even theoretically impossible results. This is well known in the statistical literature [6, 8–10] and was also noted in the medical literature several years ago [11]. However, up to now, confidence intervals for NNT—if at all—are calculated by applying the simple Wald method [4, 7, 12]. There are a number of better methods that can be used instead of the simple Wald method [6, 8–10]. However, some of these methods require complex computations. Buchan [11] proposed to use exact confidence intervals, which are now provided by StatXact [13]. However, exact methods for interval estimation of proportions are conservative, i.e., they yield confidence intervals that are unnecessarily wide [14].

It has been shown that confidence intervals based upon Wilson scores have coverage probabilities close to the nominal confidence level [6, 14–16]. Moreover, they are easier to calculate than exact confidence intervals. Hence, after investigating 11 methods for interval estimation of ARR, Newcombe proposed to use the Wilson score method [6]. The $100 \times (1 - \alpha)\%$ confidence interval for ARR based upon Wilson scores is given by:

$$\mathrm{LL(ARR)} = p_1 - p_2 - \delta \text{ and } \mathrm{UL(ARR)} = p_1 - p_2 + \epsilon \tag{3}$$

where:

$$\delta = \sqrt{(p_1 - l_1)^2 + (u_2 - p_2)^2}, \quad \epsilon = \sqrt{(u_1 - p_1)^2 + (p_2 - l_2)^2}$$

$$l_i = \varphi_i - \sqrt{\varphi_1^2 - \psi_i}, \quad u_i = \varphi_i + \sqrt{\varphi_i^2 - \psi_i}, \quad i = 1, 2$$

$$\varphi_i = \frac{2e_i + z_{1-\alpha/2}^2}{2(n_i + z_{1-\alpha/2}^2)}, \quad \psi_i = \frac{e_i^2}{n_i^2 + n_i z_{1-\alpha/2}^2}, \quad i = 1, 2 \tag{4}$$

The corresponding approximate confidence limits for NNT can then be calculated by LL(NNT) = 1/UL(ARR) and UL(NNT) = 1/LL(ARR) in consideration of the NNT scale ranging from 1 through ∞ to −1 (see above). For calculations, a SAS/IML [17] program can be used that is available via the internet <<http://www.uni-bielefeld.de/~rbender/SOFTWARE/nnt_ci.sas>> or from the author on request.

SHORTCOMINGS OF THE SIMPLE WALD METHOD

Principally, the shortcomings of the Wald confidence intervals transmit from ARR to NNT. However, for interpretation the NNT scale has to be taken into account. In the following the confidence intervals for NNT based on Wilson scores are compared with the Wald confidence intervals by means of published and artificial examples. The published examples are estimated NNT values found in the journal *Evidence-Based Medicine* [18–21]. Here, we concentrate on the comparison of the confidence intervals and do not discuss the clinical background of the studies. The adequacy of the Wald confidence intervals is mainly dependent on the sample size and the distance of the risks from the extreme points 0 and 1. Nevertheless, in the following the properties of the Wald confidence intervals for NNT are described with reference to the sample size and the size of the NNT value, because this information is mostly given in articles whereas the risks themselves are frequently missing.

In Table 1, example 1 shows that the Wald method could be used if NNT is low (say, NNTB < 10) and the sample size is moderate ($n > 100$). Note that NNT values below 10 correspond to ARR values above 0.1, which are only possible if at least one of the estimated risks is larger than 10%. It can be expected that the Wald confidence intervals are inadequate if both risk estimates are close to 0, which results in higher NNT values. For very high NNT values (say, NNTH > 100), the sample size has to be extremely large ($n > 10{,}000$) to get reliable confidence limits by means of the Wald method (example 2). For high NNT estimates (say, NNTB > 10), the Wald method is insufficient in the case of moderate sample size ($n > 100$, example 3) but improves markedly for large sample sizes ($n > 1000$, example 4), although the Wald confidence interval is still too short in this case. These examples are based on published results demonstrating that the application of the Wald method may lead to inappropriate confidence intervals in situations occurring in practice.

The deficiencies of the simple Wald method are pointed out more clearly by means of artificial examples. For high NNT estimates especially the upper Wald confidence limit is unreliable, even for moderate sample sizes (artificial example 1). In most cases, the upper Wald confidence limit will be too low. However, in the case of quite different sample sizes between the two groups, the opposite may be true. In artificial example 2 the Wald UL of 486 is much larger than the UL of 64 calculated by means of the Wilson score method. At first sight, the Wald confidence interval seems to be wider than the Wilson confidence interval. However, this is true only in the NNT scale. In the ARR scale the Wald confidence interval is shorter than the Wilson confidence interval and therefore inadequate. The magnitude of the difference between the Wald and Wilson lower confidence limits ($|11.4 - 9.4| = 2$ in NNT scale, but $|0.088 - 0.107| = 0.019$ in ARR scale) is larger in the ARR scale than between the upper limits ($|486 - 64| = 422$ in NNT scale, but $|0.002 - 0.016| = 0.014$ in ARR scale). This makes it difficult to interpret small and large NNT values. On one hand, there is no substantial difference between large NNT values, say between NNT = 1000 and NNT = 5000. In terms of probabilities this is only a difference of $0.001 - 0.0002 = 0.0008$. On the other hand, for public health decisions it may be important to treat 1000 or 5000 patients to prevent one death. However, whether relying on the NNT or the ARR scale, both Wald confidence limits are unreliable in the case of small risks and a highly unbalanced design.

The Wald method leads to several aberrations. NNT estimates close to 1 and low sample size can lead to a theoretically impossible lower Wald confidence limit (artificial example 3). If the ARR estimate is exactly 1, no meaningful

Table 1 Confidence Intervals Calculated by the Wald and the Wilson Score Method for NNT Values of Published and Artificial Examples

			Treatments		Control Group			Treatment Group					95% Confidence Intervals	
Example	Reference	Outcome	Control	New	e_1	n_1	p_1	e_2	n_2	p_2	ARR	NNT	Wald	Wilson
1	Stobart [18]	stroke in children with sickle–cell anemia	standard care	blood transfusions	11[a]	67	0.164	1[a]	63	0.016	0.148	NNTB 6.7	NNTB 4.1 to 18.4	NNTB 3.9 to 19.9
2	Walma and Thomas [19]	stroke in patients with hypertension	standard care	captopril	148[a]	5493	0.027	192[a]	5492	0.035	−0.008	NNTH 125	NNTH 649 to 69	NNTH 651 to 69
3	Faris [20]	stroke in patients with carotid endarterectomy	vein–patch closure	primary closure	7[a]	135	0.052	1[a]	130	0.008	0.044	NNTB 22.6	NNTB 12 to 260	NNTB 10 to 1735
4	Woods [21]	stroke in patients with myocardial infarction	placebo	pravastatin	47[a]	643	0.073	29[a]	640	0.045	0.028	NNTB 36.0	NNTB 19 to 499	NNTB 18 to 549
5	Artificial example 1: Wald UL unreliable if both risks are low				10	200	0.05	1	100	0.01	0.04	NNTB 25.0	NNTB 13 to 247	NNTB 12 to ∞ to NNTH 101
6	Artificial example 2: both Wald confidence limits unreliable if design is unbalanced				5	100	0.05	5	1000	0.005	0.045	NNTB 22.2	NNTB 11.4 to 486	NNTB 9.4 to 63
7	Artificial example 3: Wald LL theoretically impossible				6	7	0.857	1	7	0.143	0.714	NNTB 1.4	NNTB 0.9 to 2.9	NNTB 1.1 to 5.2
8	Artificial example 4: no meaningful Wald CI if ARR = 1				5	5	1	0	5	0	1	NNTB 1.0	NNTB 1.0 to 1.0	NNTB 1.0 to 2.6
9	Artificial example 5: no meaningful Wald CI if $p_1 = p_2 = 0$				0	100	0	0	100	0	0	∞	∞ to ∞	NNTB 27 to ∞ to NNTH 27
10	Artifical example 6: Wald CI inadequate to prove equivalence				5	100	0.05	1	100	0.01	0.04	NNTB 25.0	NNTB 11 to ∞ to NNTH 144	NNTB 10 to ∞ to NNTH 78

ARR = absolute risk reduction; NNT = number needed to treat; NNTB = number needed to treat to benefit; NNTH = number needed to treat to be harmed; UL = upper limit; LL = lower limit; CI = confidence interval.

[a] The numbers of events are estimated from percent data given in the articles.

Wald confidence interval can be calculated because the standard error of ARR is erroneously 0 (artificial example 4). The same holds when both risk estimates are exactly 0 (artificial example 5).

USING NNT FOR EQUIVALENCE TRIALS

The possible aberrations of the simple Wald method to calculate confidence intervals for ARR and NNT are meaningful especially for equivalence trials [22]. To demonstrate equivalence in therapeutic clinical trials the use of confidence intervals with coverage probability of 95% or more is recommended [23]. Frequently, the objective of a study is to show that the new treatment is not inferior to the standard treatment. In such trials, one possibility to demonstrate equivalence between treatments at one-sided significance level α is to show that the value of the $100 \times (1 - 2\alpha)\%$ confidence limit corresponding to the deterioration of the effect is better than a predefined acceptable difference.

As clinicians argue more and more in terms of NNT, it seems logical to use NNT also as an effect measure in equivalence trials. If NNT is better understood than, for example, the odds ratio, it should be easier to define an appropriate acceptable difference for NNT than for the odds ratio. For example, if the risk of the standard treatment group is expected to be 5%, a possible acceptable difference for NNT could be the value NNTH = 100. Thus, it is defined that the new treatment is not inferior to the standard treatment if 100 or more patients are needed to be treated for one additional patient to be harmed. To demonstrate one-sided equivalence between the new and the standard treatment the upper confidence limit for NNT must be larger than NNTH = 100 or must lie within the range of NNTB 1 to ∞. The latter would mean that the new treatment is even superior to the standard treatment.

In artificial example 6 the upper 95% Wald confidence limit of NNTH = 144 would lead to the decision of equivalence. This decision, however, is questionable because the Wald confidence interval is probably too short, as is shown by the Wilson score confidence interval of NNTB 10 to ∞ to NNTH 78. This means that there may be up to 1 of 78 treated patients who is harmed instead of 1 of 100 treated patients. Thus, the upper confidence limit is beyond the acceptable difference of NNTH = 100. If NNT is used as an effect measure in equivalence trials, the usual Wald confidence intervals for ARR and NNT should not be applied even in the case of moderate sample sizes. The decision that two treatments are equivalent with regard to NNT should be based upon appropriate confidence limits to ensure adequate decisions.

DISCUSSION AND CONCLUSION

NNT has become a popular summary statistic to describe the absolute effect of a given treatment in comparison to a standard treatment or control. It was first introduced for use in randomized placebo-controlled clinical trials [24], then adopted as the primary outcome measure for systematic reviews such as meta-analyses [25], extended to the statistic “number needed to screen” to compare strategies for disease screening [26], and is now applied also in epidemiology to express the magnitude of adverse effects in case-control studies [27]. NNTs are popular among clinicians because at first sight they are easier to understand than odds ratios or even ARRs. However, there are different opinions about what is easy to understand. Some authors still prefer to use ARR rather than NNT [28–31]. In my opinion, ARR and NNT contain equivalent information. In trials with a beneficial effect of the treatment, ARR expresses this effect in terms of numbers of additional adverse events prevented per 100 people treated (if ARR is presented in percentages), while NNT is the number of people needed to be treated to prevent one additional adverse event. Both measures can be applied; however, to use and interpret them adequately, the underlying scale has to be understood. As NNTs are used more and more in biomedical research, it is apparent that appropriate methods to calculate confidence intervals for NNTs are required.

In the current medical literature the calculation and reporting of confidence intervals for NNT is quite unsatisfactory. A systematic search through all issues of the journal *Evidence-Based Medicine* (1995–1999) revealed that confidence intervals for estimated NNTs are given only for significant results. If confidence intervals are reported, the method used for calculation is frequently unclear. One reason for this is that a definition of NNT is given only for the simple situation of a randomized clinical trial comparing a new with a standard treatment concerning a binary outcome over a fixed follow-up time. However, in practice NNT values are also calculated for trials with variable follow-up times. For example, in the UK Prospective Diabetes Study (UKPDS), NNTs have been calculated for the comparison of less tight blood pressure control (control group) and tight blood pressure control (treatment group) [32]. For the outcome of diabetes-related death, the result NNTB = 15 (95% confidence interval: 12.1 to 17.9) was obtained [32]. However, 1 year later, for the same data, NNTB = 20 (95% confidence interval: 10 to 100) was calculated [33]. Such different results concerning NNT estimation and confidence intervals are probably due to the application of unclear and questionable ad hoc methods in studies with varying follow-up times. To estimate NNT with appropriate confidence intervals in trials in which the outcome is time to an event rather than a simple binary variable, more complex methods are required [34].

In trials with fixed follow-up time and a binary outcome, the only method routinely used in practice seems to be the inverting and exchanging of the simple Wald confidence limits for ARR. This procedure, however, leads to unreliable confidence intervals for NNT in many cases, especially in studies with low sample sizes, risks close to 0 or 1, unbalanced designs, and equivalence trials. In situations in which it is particularly important to quantify the uncertainty of estimations, the usual Wald method fails. The application of the Wilson score method leads to confidence intervals for NNT that have much better coverage properties, are free of aberrations, and are much easier to calculate than exact confidence intervals. Any estimated NNT should be complemented by an adequate confidence interval and the calculation method should be stated. For interval estimation of NNTs in trials with fixed follow-up times and binary outcomes I recommend the replacement of the usual Wald method with the Wilson score method [6].

I thank Robert G. Newcombe for his valuable and helpful comments, which improved the paper considerably.

REFERENCES

1. Cook RJ, Sackett DL. The number needed to treat: A clinically useful measure of treatment effect. *BMJ* 1995;310:452–454.
2. Sackett DL. On some clinically useful measures of the effects of treatment. *Evidence-Based Med* 1996;1:37–38.
3. Chatellier G, Zapletal E, Lemaitre D, Ménard J, Degoulet P. The number needed to treat: A clinically useful nomogram in its proper context. *BMJ* 1996;312:426–429.
4. Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998;317:1309–1312.
5. Lesaffre E, Pledger G. A note on the number needed to treat. *Control Clin Trials* 1999;20:439–447.
6. Newcombe RG. Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Stat Med* 1998;17:873–890.
7. Daly LE. Confidence limits made easy: Interval estimation using a substitution method. *Am J Epidemiol* 1998;147:783–790.
8. Miettinen OS, Nurminen M. Comparative analysis of two rates. *Stat Med* 1985;4:213–226.
9. Beal SL. Asymptotic confidence intervals for the difference between binomial parameters for the use with small samples. *Biometrics* 1987;43:941–950.
10. Wallenstein S. A non-iterative accurate asymptotic confidence interval for the difference between two proportions. *Stat Med* 1997;16:1329–1336.
11. Buchan IE. Computer software that can calculate confidence intervals is now available (letter). *BMJ* 1995;310:1269–1270.
12. Gardner MJ, Altman DG. Confidence intervals rather than P values: Estimating rather than hypothesis testing. *BMJ* 1986;292:746–750.
13. Mehta CR, Patel NR. *StatXact 4 for Windows. Statistical Software for Exact Nonparametric Inference.* Cambridge, MA: CYTEL Software Corporation; 1999.
14. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. *Am Statistn* 1998;52:119–126.
15. Vollset SE. Confidence intervals for a binomial proportion. *Stat Med* 1993;12:809–824.
16. Newcombe RG. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat Med* 1998;17:857–872.
17. SAS. *SAS/IML User’s Guide, Version 5 Edition.* Cary, NC: SAS Institute Inc.; 1985.
18. Stobart K. Periodic blood transfusions reduced the risk for stroke in children with sickle-cell anaemia. *Evidence-Based Med* 1999;4:54.
19. Walma E, Thomas S. Captopril was not more effective than conventional treatment in hypertension and led to an increase in stroke. *Evidence-Based Med* 1999;4:110.
20. Faris I. Vein-patch closure was better than primary closure in decreasing early strokes and arterial occlusion in carotid endarterectomy. *Evidence-Based Med* 1997;2:117.
21. Woods KL. Pravastatin reduced cardiovascular events in older patients with myocardial infarction and average cholesterol levels. *Evidence-Based Med* 1999;4:52.
22. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: The importance of rigorous methods. *BMJ* 1996;313:36–39.
23. The CPMP Working Party on Efficacy of Medical Products. Biostatistical methodology in clinical trials in applications for marketing authorizations for medical products. *Stat Med* 1995;14:1659–1682.
24. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;318:1728–1733.
25. McQuay HJ, Moore A. Using numerical results from systematic reviews in clinical practice. *Ann Intern Med* 1997;126:712–720.

26. Rembold CM. Number needed to screen: Development of a statistic for disease screening. *BMJ* 1998;317:307–312.
27. Bjerre LM, LeLorier L. Expressing the magnitude of adverse effects in case-control studies: "The number of patients needed to be treated for one additional patient to be harmed." *BMJ* 2000;320:503–506.
28. North D. Number needed to treat: Absolute risk reduction may be easier for patients to understand (letter). *BMJ* 1995;310:1269.
29. Pickin M, Nicholl J. Number who benefit per unit of treatment may be a more appropriate measure (letter). *BMJ* 1995;310:1270.
30. Newcombe RG. Confidence intervals for the number needed to treat: Absolute risk reduction is less likely to be misunderstood (letter). *BMJ* 1999;318:1765.
31. Hutton JL. Number needed to treat: Properties and problems. *J R Stat Soc A* 2000; 163: 403–415.
32. The UK Prospective Diabetes Study (UKPDS) Group. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. *BMJ* 1998;317:703–713.
33. Almbrand B, Malmberg K, Ryden L. Tight blood pressure control reduced diabetes mellitus-related deaths and complications and was cost-effective in type 2 diabetes. *Evidence-Based Med* 1999;4:12–13.
34. Altman DG, Andersen PK. Calculating the number needed to treat where the outcome is time to an event. *BMJ* 1999;319:1492–1495.