Number Needed to Treat (NNT)

RALF BENDER

Volume 6, pp. 3752–3761

In

Encyclopedia of Biostatistics
Second Edition
(ISBN 0-470-84907-X)

Edited by

Peter Armitage & Theodore Colton

Number Needed to Treat (NNT)

Introduction

The number needed to treat (NNT) was originally proposed as a way of presenting the results of randomized **clinical trials** with **binary** outcome [16, 32, 33, 47, 48, 50]. Defined as the inverse of the **absolute risk** reduction (ARR), the number needed to treat is the average number of patients needed to be treated to prevent an adverse outcome in one additional patient compared to a control or standard treatment group. For example, in the Diabetes Control and Complications Trial (DCCT) the five-year **risk** of neuropathy in type 1 diabetic patients was 16.9% in the standard treatment group compared to 6.7% in the intensive insulin treatment group [18]. The absolute effect of the treatment can be described by $\text{ARR} = 16.9\% - 6.7\% = 10.2\%$. This translates to $\text{NNT} = 1/0.102 = 9.8 \approx 10$, that is, on average 10 patients are needed to be treated with intensive diabetes therapy to prevent one additional case of neuropathy compared to the standard therapy. For an adequate interpretation of NNTs, the characteristics of patients being treated, the outcome being measured, and the type and duration of interventions being compared have to be known.

NNT as well as ARR represent absolute measures of the treatment effect. Relative effect measures such as the **odds ratio** (OR), the **relative risk** (RR), or the relative risk reduction (RRR) frequently result in impressive numbers, even though the absolute effect of the treatment might be low. For example, if the two risks are $\pi_0 = 0.6$ and $\pi_1 = 0.1$, then $\text{RRR} = 83\%$, $\text{ARR} = 0.5$ and $\text{NNT} = 2$; if the two risks are $\pi_0 = 0.006$ and $\pi_1 = 0.001$, then $\text{RRR} = 83\%$ remains the same, but $\text{ARR} = 0.005$ and $\text{NNT} = 200$. Owing to the low baseline risk, the absolute effect of the treatment is also low, which is described by ARR and NNT. The information given by ARR and NNT is mathematically identical. However, the statement "200 patients are needed to be treated in order to avoid one event" is potentially more informative and comprehensible than "the treatment reduces the risk of an event by 0.005". Several studies demonstrated that assessment of health-care intervention effects by consumers is affected by the way in which study results are presented. The inclination of physicians to prescribe drugs and to treat patients is stronger when study results are presented by means of relative effect measures than when the same study is described by using absolute effect measures [12, 23, 40]. Health authority members are more willing to support health programs when results are expressed as RRRs compared with absolute effect measures [22]. Likewise, more patients assent to receive a therapy when potential benefits are reported in terms of RRR rather than ARR or NNT [29].

NNT has become the standard for presenting results of randomized clinical trials in the journal *Evidence-Based Medicine* [47] and the *ACP Journal Club* [1] and use of NNT to express study results is suggested in the **CONSORT** explanation and elaboration document [6]. However, the widespread application and extension of NNT in different settings is not without difficulties and care is required to use and interpret NNT appropriately. Recent developments regarding NNT are given by the development of methods to express benefit as well as harm, the calculation, presentation, and interpretation of confidence intervals, the application in screening studies, public health research, **epidemiology** (case–control and cohort studies), crossover studies, studies measuring continuous and time-to-event data, risk-benefit analyses, and systematic reviews. In the following, the characteristics and application areas of NNT are summarized.

General Characteristics

Relation to Other Effect Measures

A large number of effect measures exist to express the magnitude of difference between two groups concerning the risk of an adverse event. Let π_0 be the risk in the control group and π_1 be the risk in the treatment group. In the case of a beneficial treatment ($\pi_0 > \pi_1$) the most frequently used effect measures derived from a simple **2 x 2 table** are

$$\begin{aligned}
&\text{Absolute risk reduction:} && ARR = \pi_0 - \pi_1 \\
&\text{Relative risk:} && RR = \frac{\pi_1}{\pi_0} \\
&\text{Relative risk reduction:} && RRR = \frac{\pi_0 - \pi_1}{\pi_0} \\
& && \quad = 1 - RR
\end{aligned}$$

Odds ratio: $$OR = \frac{\pi_1 \times (1-\pi_0)}{\pi_0 \times (1-\pi_1)}$$

Number needed to treat: $$NNT = \frac{1}{\pi_0 - \pi_1} = \frac{1}{ARR}$$

The relation between NNT and ARR is obvious. It is helpful in practice to also express NNT as a function of RR, RRR, OR, and the control event rate π_0. The respective formulae are given by

$$NNT = \frac{1}{(1-RR) \times \pi_0} = \frac{1}{RRR \times \pi_0} \quad (1)$$

$$NNT = \frac{1-(1-OR) \times \pi_0}{(1-OR) \times \pi_0 \times (1-\pi_0)} = \frac{1}{(1-OR) \times \pi_0} + \frac{OR}{(1-OR) \times (1-\pi_0)} \quad (2)$$

Similar formulas are published for the case of harmful treatments [7, 9], for considering desirable instead of adverse outcomes [39], and for the inverse definitions of RR and OR [28].

Quantifying Benefit and Harm

NNT represents the inverse of the difference of two risks. On principle, the difference of two risks can be positive, zero, or negative. The concept of NNT was originally developed for the situation of a beneficial treatment, so that the risk of an adverse event in the treatment group is lower than in the control group [33]. Thus, calculating the risk difference as control minus treatment leads to a positive ARR value. Considering only beneficial treatments, the term "number needed to treat" was proposed to describe the inverse of ARR. In the case of a harmful treatment, this calculation leads to a negative risk difference and a negative NNT. To avoid negative numbers, the risk difference is calculated as treatment minus control if the risk of the treatment group is higher than that of the control group leading to a positive value called absolute risk increase (ARI). To describe the inverse of ARI, the unfavorable term "number needed to harm" (NNH) was used [39]. Recognizing that NNT and NNH are not good abbreviations, Altman suggested the terminology "number of patients needed to be treated for one additional patient to benefit" (NNTB) or "be harmed" (NNTH) [2]. This terminology should be used when it is necessary to indicate the direction of the effect. In the case of desirable outcomes, such as healing or improvement of **quality of life**, the order of the two probabilities in the calculation of NNT is reversed. Here, NNTB represents the average number of patients needed to be treated to gain one additional beneficial outcome compared to a control or standard treatment group [55].

Confidence Intervals

As with other estimated effect measures, it is important to document the uncertainty of the estimation by means of an appropriate **confidence interval**. In principle, confidence intervals for NNTs can be obtained by inverting and exchanging the confidence limits of the corresponding risk difference [17]. Nevertheless, calculating, presenting, and interpreting confidence intervals for NNTs is not straightforward. Owing to the reciprocal **transformation**, the NNT has undesirable statistical properties [34]. To obtain meaningful confidence intervals for NNT two issues have to be considered. Firstly, the unusual scale of NNT has to be taken into account, and secondly, an appropriate method to calculate confidence intervals for the risk difference is required.

The key to understand the confidence interval for NNT is that the domain of NNT is the union of 1 to ∞ (in the NNTB region) and $-\infty$ to -1 (in the NNTH region). 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. Values between -1 and 1 are impossible for NNT. Owing to estimation uncertainty, the estimated NNT may be negative even when the true NNT is positive and vice versa. Even when the sign of the estimated and true NNT are identical, the estimation uncertainty can be so large that neither a harmful nor a beneficial effect can be excluded. In this case, the confidence interval covers both the NNTB and the NNTH region. 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. To make this clear, a confidence interval for an NNT estimate that is not statistically significant should be presented as NNTB = 10 (NNTB 4 to ∞ to NNTH 20) [2]. This presentation indicates that a beneficial treatment effect of NNTB = 10 is estimated, but the uncertainty

of this estimation is so large that a more beneficial effect up to NNTB = 4 and a less beneficial effect up to NNTB = ∞ (no effect at all) as well as a harmful effect up to NNTH = 20 is compatible with the observed data.

For large sample sizes and risks not close to 0 or 1, the usual Wald method can be used to calculate confidence intervals for risk differences (*see* **Estimation, Interval**). However, Wald confidence intervals have poor coverage probabilities and a propensity to aberrations in many practical situations. Thus, Newcombe proposed to calculate confidence intervals for risk differences based upon Wilson scores [42]. This method was also recommended for NNT [8].

Let n_0 and n_1 be the number of patients in the control and the treatment group, respectively, and let e_0 and e_1 be the number of patients having an event in the control and the treatment group, respectively. The risks of an event in the two groups can then be estimated by the proportions $p_0 = e_0/n_0$ and $p_1 = e_1/n_1$. The effect measures can be estimated by ARR $= p_0 - p_1$ and NNT $= 1/(p_0 - p_1)$. Using this notation, the $100 \times (1 - \alpha)\%$ confidence interval for ARR based upon Wilson scores is given by:

$$LL(ARR) = p_0 - p_1 - \delta \text{ and}$$
$$UL(ARR) = p_0 - p_1 + \varepsilon, \qquad (3)$$

where

$$\delta = \sqrt{(p_0 - l_0)^2 + (u_1 - p_1)^2},$$
$$\varepsilon = \sqrt{(u_0 - p_0)^2 + (p_1 - l_1)^2},$$
$$\mathrm{l}_i = \varphi_i - \sqrt{\varphi_i^2 - \psi_i}, \mathrm{u}_i = \varphi_i + \sqrt{\varphi_i^2 - \psi_i}, \mathrm{i} = 0, 1,$$
$$\varphi_i = \frac{2e_i + z_{1-\alpha/2}^2}{2(n_i + z_{1-\alpha/2}^2)}, \psi_i = \frac{e_i^2}{n_i^2 + n_i z_{1-\alpha/2}^2}, \mathrm{i} = 0, 1,$$

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

The corresponding 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). An SAS program can be used for calculations [8].

Confidence intervals for NNT based upon Wilson scores seem to be adequate for most practical applications. For very small sample sizes or applications, which require that the true confidence level under no circumstances remains under the nominal level, exact [14] or quasi-exact methods [15] should be used (*see* **Exact Inference for Categorical Data**).

Extensions and Applications

The principle of NNT has been extended and suggested for use in a wide variety of circumstances. The most important ones are summarized below.

Screening

Rembold extended the NNT concept to compare strategies for disease **screening** [44]. The analogous statistic termed "number needed to screen" (NNS) describes the number of people that need to be screened to prevent one death or adverse event. In clinical trials that directly investigate the benefit of a screening strategy, the point and interval estimation of NNS is identical to that of NNT. However, the intervention under study is a screening strategy applied to a population, rather than a treatment applied to patients. If no study exists that evaluates directly the benefit of a screening strategy, NNS estimation can be performed by combining the knowledge of clinical trials investigating the benefit of treating **risk factors** and the **prevalence** of persons with inadequately treated risk factors in the community. Under the assumption that screened individuals with positive results will show full compliance with subsequent treatment, NNS can be calculated by dividing the corresponding NNT by the prevalence of unaware or untreated disease.

Expressing the absolute effect of screening strategies as NNS values has the same advantages as the presentation of treatment effects by means of NNTs. However, the NNS approach has some limitations. Firstly, the division of NNTs by an estimated prevalence of untreated disease is subject to propagation of errors. A method to calculate confidence intervals for NNS taking the uncertainty of both the NNT and the prevalence estimation into account is required. Secondly, NNS values calculated from clinical trials investigating the benefit of a screening strategy directly (*see* **Screening Trials**) may not be comparable to NNS values calculated from NNTs divided by the prevalence of unaware or untreated disease. The former may be more affected by participation and selection effects than the latter. Hence, Richardson suggested to multiply the directly estimated NNS

by the participation rate adjusted for **selection** to obtain an NNS value free of participation and selection effects [45]. However, this method is even more exposed to propagation of errors. Moreover, the benefit of a screening strategy should be described including participation and selection effects. Analogous to the **intention to treat** analysis of clinical trials, the gold standard is the unadjusted NNS estimated from trials directly investigating the benefit of screening strategies.

Public Health Research

The NNT statistic relates to those patients actually treated and gives no information how many people of all patients with the disease or of the total population will benefit from the treatment. Heller & Dobson proposed two new statistics offering a public health perspective [27]. The idea is similar to that of NNS calculated by NNT divided by the prevalence of unaware or untreated disease. The "disease impact number" (DIN) takes into account the number of people in the population with the disease, not just those eligible for treatment according to the entry criteria of the considered clinical trial. DIN is calculated by dividing NNT by the proportion of patients with the disease who are eligible for treatment. The "population impact number" (PIN) takes into account the total size of the population from which the patients with the disease are drawn. PIN is calculated by dividing DIN by the prevalence of disease in the population. DINs and PINs suffer from limitations similar to those of indirectly estimated NNS values. Owing to the division of NNTs by estimated proportions they are subject to greater **random error** than NNT. However, they may play a role as communication tool for treatment effects from a population perspective [52].

Case–control Studies

Bjerre & LeLorier proposed to use the NNTH statistic to express the magnitude of harmful exposures effects in **case–control studies** [11]. As information about the absolute risk is not directly available from case–control studies, they calculated NNTH by using the odds ratio provided by the case–control study and the unexposed event rate obtained from external sources. Although not mentioned by the authors, an additional advantage of this approach is that adjusted NNTs can be calculated by using adjusted ORs to estimate the corresponding NNT values (see next section). Confidence intervals for NNTH are calculated by transforming the confidence limits of OR. Unfortunately, to calculate NNTH as function of OR, formula (1) was used, which actually represents the relation between NNT and RR. Thus, NNTH is systematically underestimated, that is, the exposure effect is overestimated. The magnitude of this error is negligible if OR and RR are approximately equal. Thus, in case-control studies, in which usually rare diseases are investigated, the error is unimportant. However, in situations where OR and RR are quite different, either formula (1) with RR or formula (2) with OR must be applied to obtain correct results. Let $NNTH_{1,\mathrm{OR}}$ be the *NNTH* value calculated by formula (1) with OR and let $NNTH_{\mathrm{true}}$ be the true *NNTH*. It can be shown that $(NNTH_{\mathrm{true}} - NNTH_{1,\mathrm{OR}})/NNTH_{\mathrm{true}} = \pi_1$, that is, the relative error of $NNTH_{1,\mathrm{OR}}$ equals the exposed event rate [7]. Even, if the correct formula is used, a limitation of this approach is that the confidence interval for NNTH takes into account the uncertainty of the OR estimation but not that of the unexposed event rate. A possible solution is given by the methods developed by King & Zeng for point and interval estimation of risk differences in case–control studies based upon **Bayesian methods** or a range of possible values for the unexposed event rate [30].

Cohort Studies

The NNT concept has been applied to compare exposed and unexposed persons in **cohort studies** [9]. For this application, the term "number needed to be exposed" (NNE) was suggested. When it is necessary to distinguish between harmful and beneficial **exposures**, the abbreviations NNEH and NNEB should be used. In the case of a harmful exposure, NNEH represents the average number of persons needed to be exposed for one additional case of disease or death compared to the unexposed persons. NNEs are calculated as a function of the odds ratio and the unexposed event rate by means of formula (2). This approach allows the calculation of adjusted NNEs by using adjusted odds ratios, estimated, for example, by multiple **logistic regression**. Within the framework of logistic regression, the adjusted odds ratio is constant over the distribution of the considered **confounders**. However, the event rates and their differences are dependent on the confounder

values. Thus, NNE also varies with the values of the confounding variables, which has to be taken into account when adjusted NNEs are estimated. Two methods were proposed to calculate adjusted NNEs. In the first approach, the mean risk of the unexposed persons is used and NNE is calculated for the corresponding confounder profile. In the second approach, NNE is calculated for some fixed confounder profiles, which gives an impression about different absolute effects of the exposure in cohorts with varying confounder values. A similar principle is applied to calculate pooled NNTs in meta-analysis (see below).

Confidence intervals for adjusted NNEs can be calculated indirectly via confidence intervals for the corresponding risk difference. Within the framework of logistic regression analysis applied to prospective cohort data, risk differences between the exposed and unexposed persons can be expressed as functions of the logistic regression coefficients. Thus, approximate **standard errors** and confidence intervals for risk differences can be calculated by means of the multivariate **delta method** [9]. In contrast with the calculation of NNTs in case–control studies, this method takes the estimation uncertainties of both the odds ratio and the unexposed event rate into account. The adequacy of the approximate confidence intervals was investigated via **simulations** demonstrating sufficient quality for most epidemiological applications [10].

Continuous Data

NNT represents a summary statistic for the comparison of two groups concerning a binary outcome. Nevertheless, in some applications, investigators want to express their study results in terms of NNT although the outcome variable is measured in a continuous scale (*see* **Random Variable**). One obvious method to calculate NNTs for continuous outcomes is to dichotomize the response in both groups and to apply the usual methods. Alternatively, one can dichotomize the difference of the responses between the two groups. Walter examined the probability that the difference of the responses between the two groups is larger than the minimally important difference (*see* **Sample Size Determination for Clinical Trials**) [55]. Without loss of generality, we assume that higher response values correspond to adverse outcomes (such as hypertension). Let X_0 and X_1 be the control and treatment responses of a given subject and c be the minimally important difference. The probability described above is given by $\theta = P(X_0 - X_1 > c)$. The continuous data version of NNT is then calculated by $NNT = 1/\theta$. Under the assumption of **bivariate normality** of (X_0,X_1), θ is given by

$$\theta = \Phi\left(\frac{\mu_0 - \mu_1 - c}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1}}\right), \qquad (4)$$

where Φ denotes the distribution function of the standard normal distribution, μ_0 and μ_1 and σ_0 and σ_1 are the means and standard deviations of X_0 and X_1, respectively, and ρ is correlation of X_0 and X_1 [55]. Estimation of θ and NNT is performed by substituting the usual estimates of μ_0, μ_1, σ_0, σ_1 and ρ into (4). Formulas for the standard error of the estimated probability θ can be derived by means of the delta method both for paired and unpaired data [55].

It should be noted that formula (4) is first of all only useful in studies, which provide an estimate of ρ (such as **crossover** studies, see below). In all designs considered so far (randomized clinical trials with parallel group design, cohort studies, and case–control studies with two independent groups) the within-subject correlation is not estimable. In this case, Walter proposed to use a variety of different assumed values of ρ and investigate the **sensitivity** of θ to the unknown correlation value [55]. Alternatively, in studies observing independent groups, the first mentioned approach of dichotomizing the response in both groups could be used.

In practice, continuous outcomes are frequently subject to random **measurement error**. Even in the case of **nondifferential** measurement error, dichotomization of continuous variables leads to a **bias** in the estimated proportions and estimated NNTs. Walter & Irwig investigated the effect of measurement error in continuous outcomes on NNT estimation [56], and methods to reduce the bias by adjusting for measurement error are in development [38]. In general, even in the case of no measurement error, one should be aware of the potential loss of information due to **categorizing of continuous variables**. Hence, calculation of NNTs from continuous data can only serve as supplement to the analysis of data in the original continuous scale by using means and differences of means.

Crossover Studies

Originally, the NNT statistic was developed for use in studies investigating two independent groups. Walter systematically examined NNT estimators and their variances for both crossover and parallel group designs [55]. Owing to the undesirable statistical properties of NNT, it is preferable to calculate the standard errors of the corresponding risk differences instead of the NNTs themselves. The NNT estimators are identical in both designs, whereas standard errors are different. Approximate confidence intervals for risk differences can be calculated in both designs by using the Wald method [55]. As described before, confidence intervals for NNTs can be obtained by inverting and exchanging confidence limits of the corresponding risk difference [17]. For the parallel group design, it was shown that the Wald method is unreliable in many practical situations. The same holds for crossover studies, in which it is preferable to calculate confidence intervals for the difference between paired proportions based upon the Wilson score method [41].

As crossover studies provide an estimate of the within-subject correlation, the continuous data version of NNT based upon the minimally important difference (see above) can be estimated directly. Under the assumption of normality of the continuous response, NNT can be estimated by using equation (4). Without making distributional assumptions, NNT is given by the inverse proportion of subjects for which the difference between the responses is larger than the minimally important difference [55].

Survival Data

The concept of NNT was originally developed for binary outcomes measured at a specific fixed time point. Nevertheless, NNTs are also calculated and presented for studies where the outcome is the time to an event (*see* **Survival Analysis, Overview**). Unfortunately, unclear and questionable methods have been used for point and interval estimation of NNT in studies in which follow-up times are not equal for all patients. Owing to the application of questionable *ad hoc* methods, different and confusing results have been published for the same data [8].

First, it should be noticed that in studies with varying follow-up times, NNT would also vary according to the length of follow-up. In such studies, no single NNT value exists. NNT can be calculated at any time point after the start of the treatment. Frequently used methods to analyze survival times are given by **Kaplan–Meier** survival curves providing estimates of the survival probabilities $S_0(t)$ and $S_1(t)$ of the control and treatment group, respectively, and the **Cox regression model**, providing an estimate of the **hazard ratio** (HR), possibly adjusted for other **prognostic** variables. Altman & Andersen proposed to estimate NNT by means of

$$NNT(t) = \frac{1}{S_1(t) - S_0(t)} \tag{5}$$

if the survival probabilities $S_0(t)$ and $S_1(t)$ are given, or by

$$NNT(t) = \frac{1}{(S_0(t))^{\mathrm{HR}} - S_0(t)} \tag{6}$$

if the assumption of **proportional hazards** is fulfilled and $S_0(t)$ and the HR for the comparison of the control and treatment group are given [3]. If one fixed time point is specified, one NNT value is obtained. Otherwise, (5) and (6) will lead to a NNT curve as a function of time.

To get an NNT statistic independent of time, Lubsen et al. proposed to calculate NNTs by the reciprocal of the difference of two **hazards** [36]. However, this approach requires the assumption of constant hazards. Moreover, the difference of hazards is not the same as the difference of risks. Thus, this approach leads to a statistic with a different meaning than that of the usual NNT. It should be noted that in the presence of confounders survival probabilities are dependent on the confounder values even if we can assume a constant HR. Thus, NNT not only depends on time but also on confounders. Altman & Andersen proposed to calculate NNT curves for different subsets of patients with varying **prognosis** [3]. However, more work is required to develop methods for estimation of adjusted NNTs from survival times.

Combining and Pooling

Risk-benefit Analysis

The decision about the use of a treatment should not be based upon its effect on the target event alone. Adverse side effects attributable to treatment as well as costs of therapy and costs avoided by preventing target events should also be considered (*see* **Decision Analysis in Diagnosis and Treatment Choice**). The

"threshold NNT" (NNT_T) was defined as the NNT value at which the therapeutic benefit equals the therapeutic risks [25, 26, 51]. If the estimated NNT is below the threshold NNT, then treatment should be administered. If the estimated NNT is above the threshold NNT, the patients should not be treated because the risks and costs of treatment are larger than the expected benefit. The threshold NNT is given by

$$NNT_T = \frac{TEC + TEV}{DC + AER \times (AEC + AEV)}, \quad (7)$$

where TEC represents the costs of treating one target event, TEV the value of one target event avoided (given in the same economic units as costs), DC the direct costs of therapy, AEC the costs of treating one adverse side effect, AEV the value of the side effect and AER the event rate of the side effect [51]. Similar formulas for considering multiple side effects and omitting costs can be found elsewhere [51].

While the concept of the threshold NNT seems to be appealing, the practical application is challenging. For an adequate decision making, the estimation uncertainties should be taken into account. The specification of the data (costs and values) required for the calculation of the threshold NNT is not easy and the quantification of these data uncertainties is much more difficult. Especially, the values one is willing to pay for one target or one side effect avoided are highly subjective. Thus, it is quite important to disclose all data and assumptions used for calculating a threshold NNT.

Combined NNT Measures for Different Outcomes

Several approaches have been published to combine the NNTB of the target event and the NNTH of a side effect into one measure incorporating benefit as well as harm. Let π_0, π_1 be risks of the target event and ν_0, ν_1 the risks of the side effect in the control and the treatment group, respectively. We consider the case of an adverse target event, an adverse side effect and a treatment that is beneficial concerning the target event ($\pi_0 > \pi_1$) but harmful concerning the side effect ($\nu_1 > \nu_0$). For other situations, appropriate modifications of the following measures are required. Riegelman & Schroth proposed the combined measure

$$NNT_{\text{comb}} = \frac{1}{(\pi_0 - \pi_1) - (\nu_1 - \nu_0)}, \quad (8)$$

that is, the reciprocal of the difference between ARR of the target event and ARI of the side effect [46]. The authors proceeded by adjusting this measure for the qualities and timings of the considered outcomes [46]. This procedure was criticized because a decision analysis has to be carried out before the quality-adjusted NNT can be calculated [19]. Thus, the intuitive meaning, which is one advantage of the NNT statistic, is lost. It is only possible to interpret a quality-adjusted NNT if the underlying decision analysis is understood. The statistical properties of the quality-adjusted NNT statistic have not been investigated and no methods to calculate confidence intervals have been developed.

A second approach of an NNT measure incorporating benefit and harm was proposed by Schulzer & Mancini [49]. They tried to calculate the number of patients needed to treat to produce one "unqualified success" (US), that is, the situation in which one adverse target event is avoided while simultaneously no treatment-induced side effect occurred. The NNT for one unqualified success is given by

$$NNT_{\text{US}} = \frac{1}{(\pi_0 - \pi_1) \times [1 - (\nu_1 - \nu_0)]}. \quad (9)$$

Formula (9) is based upon the assumption that the target event and the adverse side effect are independent in both the untreated and the treated population. This assumption will rarely be true in practical applications. Although a procedure was proposed to handle situations in which an association between the prevention of a target event and the induction of a side effect is expected [37], this approach suffers from the lack of an appropriate method to estimate the association from the data. Moreover, no adequate method to calculate confidence intervals for NNT_{US} has been developed.

Willan et al. proposed the benefit–risk ratio

$$R = \frac{NNTH\,(\text{side effect})}{NNTB\,(\text{target effect})} = \frac{\pi_0 - \pi_1}{\nu_1 - \nu_0}$$
$$= (\pi_0 - \pi_1) \times NNTH\,(\text{side effect}), \quad (10)$$

which can be interpreted as increase in the expected number of prevented target events achieved for each additional adverse side effect induced by treatment [57]. For large sample sizes, Willan et al. developed a statistical procedure to construct confidence intervals for the benefit–risk ratio based upon **Fieller's theorem** [57].

The development of a combined NNT statistic incorporating benefit and harm of multiple events is not straightforward. Before one of the proposed combined NNT measures considering multiple events can be routinely applied in practice, more work is required concerning the practical utility of these measure as well as their statistical properties, especially in small samples.

Meta-analysis

Since NNT has been advocated as a useful effect measure for systematic reviews [39], a number of authors have pointed out that particular caution is needed in deriving pooled NNTs in **meta-analyses** [5, 13, 20, 21, 53]. A single pooled NNT value over all studies in a meta-analysis may be misleading, especially if there is a variation in the baseline risk, different lengths of follow-up, differences in the outcomes considered, or different clinical settings. The naive approach of simply adding the raw totals of all considered trials as if the data came from one trial should be avoided. The calculation of a pooled NNT should be based upon a pooled effect measure, which should be independent of the baseline risk. Using empirical data, Furukawa et al. showed that the relative effect measures OR and RR calculated by means of an appropriate **fixed** or **random effects** regression model often appear to be reasonably constant across different baseline risks [24]. Meaningful NNTs can be obtained by inserting the pooled RR or OR from meta-analyses in formula (1) or (2). If there is variation in the baseline risk, different NNTs relevant to specific patient subgroups should be calculated [20, 24, 53]. If there is evidence that even the relative effect measures vary substantially between subgroups in a meta-analysis, no meaningful pooled NNT can be calculated.

Conclusion

The use of NNT as effect measure for the comparison of risks between two groups has been advocated in medical journals for several years [16, 33, 39, 43, 47, 50] but was recently criticized [53, 58] or even rejected [28]. There seems to be a gap in the assessment of the practical usefulness of NNTs between some statisticians and clinicians [4, 28, 35]. Some mathematical arguments against the use of NNTs, such as undesirable distributional properties, are surely justified. However, strict mathematical arguments lose their importance when NNT is considered as a way of presenting results, not as a tool for statistical computations [4, 35]. A clear distinction should be made between data analysis and subsequent risk communication [54]. In the light of the effects on consumers of the scale in which benefits and risks are reported, it is frequently advisable to choose a statistical model and a corresponding appropriate summary measure for the task of data analysis, but alternative effect measures to report the most important results. For the translation of research findings to consumers, the number needed to treat may represent a useful tool, because it gives an intuitive impression of the absolute effect of a therapy or an intervention. NNTs contain the same information as risk differences, but in the unit of patient numbers instead of probabilities, which is easier to understand.

The attempt to extend and apply the simple NNT concept developed for randomized clinical trials with two independent groups and a binary outcome for a variety of other settings led to the development of more sophisticated approaches and procedures for NNT calculation. Some useful approaches have been developed, but situations remain for which further work is needed to calculate meaningful NNTs, for example, survival time studies or the combination of NNTs for multiple outcomes. These extension and adjustment procedures can alleviate problems with NNTs. However, the extended and adjusted NNTs can no more be considered as "one simple single yardstick" [31]. Particular caution is required to apply and interpret NNTs adequately in practice, especially in meta-analyses and in the presence of confounders. Nevertheless, if handled appropriately, NNTs represent a useful communication tool to express the absolute effects of interventions and exposures.

References

[1] ACP Journal Club (1997). Glossary, *ACP Journal Club* **126**, 28.

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

[3] Altman, D.G. & Andersen, P.K. (1999). Calculating the number needed to treat where the outcome is time to an event, *British Medical Journal* **319**, 1492–1495.

[4] Altman, D.G. & Deeks, J.J. (2000). Comments on the paper by Hutton, *Journal of the Royal Statistical Society – A* **163**, 415–416.

[5] Altman, D.G. & Deeks, J.J. (2002). Meta-analysis, Simpson's paradox, and the number needed to treat, *BMC Medical Research Methodology* **2**, 3.

[6] Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D.R., Gøtzsche, P.C., & Lang, T. for the CONSORT Group. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration, *Annals of Internal Medicine* **134**, 663–694.

[7] Bender, R. (2000). Expressing the number needed to treat as a function of the odds ratio and the unexposed event rate (Rapid Electronic Letter), *British Medical Journal* (eBMJ: `http://www.bmj.com/cgi/eletters/320/7233/503#EL7`).

[8] Bender, R. (2001). Calculating confidence intervals for the number needed to treat, *Controlled Clinical Trials* **22**, 102–110.

[9] Bender, R. & Blettner, M. (2002). Calculating the "number needed to be exposed" with adjustment for confounding variables in epidemiological studies, *Journal of Clinical Epidemiology* **55**, 525–530.

[10] Bender, R. & Kuss, O. (2003). Confidence intervals for adjusted NNEs: a simulation study (Letter), *Journal of Clinical Epidemiology* **56**, 205–206.

[11] Bjerre, L.M. & LeLorier, L. (2000). 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", *British Medical Journal* **320**, 503–506.

[12] Bobbio, M., Demichelis, B. & Giustetto, G. (1994). Completeness of reporting trial results: Effect on physicians' willingness to prescribe, *Lancet* **343**, 1209–1211.

[13] Cates, C. (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis, *BMC Medical Research Methodology* **2**, 1.

[14] Chan, I.S.F. & Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions, *Biometrics* **55**, 1202–1209.

[15] Chen, X. (2002). A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small samples, *Statistics in Medicine* **21**, 943–956.

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

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

[18] DCCT Research Group (1995). The effect of intensive diabetes therapy on the development and progression of neuropathy, *Annals of Internal Medicine* **122**, 561–568.

[19] Dowie, J. (1998). The 'number needed to treat' and the 'adjusted NNT' in health care decision-making, *Journal of Health Services Research and Policy* **3**, 44–49.

[20] Ebrahim, S. (2001). The use of numbers needed to treat derived from systematic reviews and meta-analysis. Caveats and pitfalls, *Evaluation & the Health Professions* **24**, 152–164.

[21] Ebrahim, S. & Smith, G.D. (1999). The 'number need to treat': Does it help clinical decision making? *Journal of Human Hypertension* **13**, 721–724.

[22] Fahey, T., Griffiths, S. & Peters, T.J. (1995). Evidence based purchasing: understanding results of clinical trials and systematic reviews, *British Medical Journal* **311**, 1056–1095.

[23] Forrow, L., Taylor, W.C. & Arnold, R.M. (1992). Absolutely relative: how research results are summarized can affect treatment decisions, *American Journal of Medicine* **92**, 121–124.

[24] Furukawa, T.A., Guyatt, G.H. & Griffith, L.E. (2002). Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses, *International Journal of Epidemiology* **31**, 72–76.

[25] Guyatt, G.H., Sackett, D.L., Sinclair, J.C., Hayward, R., Cook, D.J. & Cook, R.J. for the Evidence-Based Medicine Working Group. (1995). Users' guides to the medical literature. IX. A method for grading health care recommendations, *Journal of the American Medical Association* **274**, 1800–1804.

[26] Guyatt, G.H., Sinclair, J.C., Cook, D.J. & Glasziou, P. for the Evidence-Based Medicine Working Group & for the Cochrane Applicability Methods Working Group. (1999). Users' guides to the medical literature. XVI. How to use a treatment recommendation, *Journal of the American Medical Association* **281**, 1836–1843.

[27] Heller, R.F. & Dobson, A. (2000). Disease impact number and population impact number: population perspectives to measures of risk and benefit, *British Medical Journal* **321**, 950–952.

[28] Hutton, J.L. (2000). Number needed to treat: properties and problems, *Journal of the Royal Statistical Society – A* **163**, 403–415.

[29] Hux, J.E. & Naylor, C.D. (1995). Communicating the benefits of chronic preventive therapy: does the format of efficacy data determine patients' acceptance of treatment? *Medical Decision Making* **15**, 152–157.

[30] King, G. & Zeng, L. (2002). Estimating risk and rate levels, ratios and differences in case-control studies, *Statistics in Medicine* **21**, 1409–1427.

[31] Kristiansen, I.S., Gyrd-Hansen, D., Nexøe, J. & Nielsen, J.B. (2002). Number needed to treat: easily understood and intuitively meaningful? Theoretical considerations and a randomized trial, *Journal of Clinical Epidemiology* **55**, 888–892.

[32] Laupacis, A. & Sackett, D.L. (1998). Number needed to treat (NNT), in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds. Wiley, Chichester, pp. 3081–3088.

[33] Laupacis, A., Sackett, D.L. & Roberts, R.S. (1988). An assessment of clinically useful measures of the consequences of treatment, *New England Journal of Medicine* **318**, 1728–1733.

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

[35] Lesaffre, E. & Pledger, G. (2000). Comments on the paper by Hutton, *Journal of the Royal Statistical Society – A* **163**, 417.

[36] Lubsen, J., Hoes, A. & Grobbee, D. (2000). Implications of trial results: the potentially misleading notations of number needed to treat and average duration life gained, *Lancet* **356**, 1757–1759.

[37] Mancini, G.B. & Schulzer, M. (1999). Reporting risks and benefits of therapy by use of the concepts of unqualified success and unmitigated failure: applications to highly cited trials in cardiovascular medicine, *Circulation* **99**, 377–383.

[38] Marschner, I.C., Emberson, J., Irwig, L. & Walter, S.D. (2003). Adjustment for bias in the number needed to treat (NNT) when outcome is based on a continuous quantity measured with error, *Journal of Clinical Epidemiology* (submitted for publication).

[39] McQuay, H.J. & Moore, A. (1997). Using numerical results from systematic reviews in clinical practice, *Annals of Internal Medicine* **126**, 712–720.

[40] Naylor, C.D., Chen, E. & Strauss, B. (1992). Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Annals of Internal Medicine* **117**, 916–921.

[41] Newcombe, R.G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data, *Statistics in Medicine* **17**, 2635–2650.

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

[43] Nuovo, J., Melnikow, J. & Chang, D. (2002). Reporting number needed to treat and absolute risk reduction in randomized controlled trials, *Journal of the American Medical Association* **287**, 2813–2814.

[44] Rembold, C.M. (1998). Number needed to screen: development of a statistic for disease screening, *British Medical Journal* **317**, 307–312.

[45] Richardson, A. (2001). Screening and the number needed to treat, *Journal of Medical Screening* **8**, 125–127.

[46] Riegelman, R. & Schroth, W.S. (1993). Adjusting the number needed to treat: Incorporating adjustments for the utility and timing of benefits and harms, *Medical Decision Making* **13**, 247–252.

[47] Sackett, D.L. (1996). On some clinically useful measures of the effects of treatment, *Evidence-Based Medicine* **1**, 37–38.

[48] Sackett, D.L. & Cook, R.J. (1994). Understanding clinical trials. What measures of efficacy should journals provide busy clinicians, *British Medical Journal* **309**, 755–756.

[49] Schulzer, M. & Mancini, G.B. (1996). 'Unqualified success' and 'unmitigated failure': number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment-induced adverse events, *International Journal of Epidemiology* **25**, 704–712.

[50] Sinclair, J.C. & Bracken, M.B. (1994). Clinically useful measures of effect in binary analyses of randomized trials, *Journal of Clinical Epidemiology* **47**, 881–889.

[51] Sinclair, J.C., Cook, R.J., Guyatt, G.H., Pauker, S.G. & Cook, D.J. (2001). When should an effective treatment be used? derivation of the threshold number needed to treat and the minimum event rate for treatment, *Journal of Clinical Epidemiology* **54**, 253–262.

[52] Smeeth, L. & Ebrahim, S. (2000). Commentary: DINs, PINs, and things – clinical and population perspectives on treatment effects, *British Medical Journal* **321**, 952–953.

[53] Smeeth, L., Haines, A. & Ebrahim, S. (1999). Numbers needed to treat derived from meta-analyses – sometimes informative, usually misleading, *British Medical Journal* **318**, 1548–1551.

[54] Walter, S.D. (2000). Choice of effect measure for epidemiological data, *Journal of Clinical Epidemiology* **53**, 931–939.

[55] Walter, S.D. (2001). Number needed to treat (NNT): estimation of a measure of clinical benefit, *Statistics in Medicine* **20**, 3947–3962.

[56] Walter, S.D. & Irwig, L. (2001). Estimating the number needed to treat (NNT) index when the data are subject to error, *Statistics in Medicine* **20**, 893–906.

[57] Willan, A.R., O'Brien, B.J. & Cook, D.J. (1997). Benefit-risk ratios in the assessment of the clinical evidence of a new therapy, *Controlled Clinical Trials* **18**, 121–130.

[58] Wu, L.A. & Kottke, T.E. (2001). Number needed to treat: Caveat emptor, *Journal of Clinical Epidemiology* **54**, 111–116.

(*See also* **Categorical Data Analysis**; **Evidence-based Medicine**; **Risk Assessment in Clinical Decision Making**)

RALF BENDER