

Debate

## Simpson's paradox and calculation of number needed to treat from meta-analysis

Christopher J Cates

Address: Manor View Surgery, Bushey Health Centre, London Road, Bushey, Watford, UK

E-mail: [chriscates@email.msn.com](mailto:chriscates@email.msn.com)

Published: 25 January 2002

Received: 27 December 2001

*BMC Medical Research Methodology* 2002, **2**:1

Accepted: 25 January 2002

This article is available from: <http://www.biomedcentral.com/1471-2288/2/1>

© 2002 Cates; licensee BioMed Central Ltd. Verbatim copying and redistribution of this article are permitted in any medium for any purpose, provided this notice is preserved along with the article's original URL.

### Abstract

**Background:** Calculation of numbers needed to treat (NNT) is more complex from meta-analysis than from single trials. Treating the data as if it all came from one trial may lead to misleading results when the trial arms are imbalanced.

**Discussion:** An example is shown from a published Cochrane review in which the benefit of nursing intervention for smoking cessation is shown by formal meta-analysis of the individual trial results. However if these patients were added together as if they all came from one trial the direction of the effect appears to be reversed (due to Simpson's paradox).

Whilst NNT from meta-analysis can be calculated from pooled Risk Differences, this is unlikely to be a stable method unless the event rates in the control groups are very similar. Since in practice event rates vary considerably, the use a relative measure, such as Odds Ratio or Relative Risk is advocated. These can be applied to different levels of baseline risk to generate a risk specific NNT for the treatment.

**Summary:** The method used to calculate NNT from meta-analysis should be clearly stated, and adding the patients from separate trials as if they all came from one trial should be avoided.

### Introduction

#### *Calculation of summary statistics from single trials and pooled data*

In a single trial that reports outcomes in a dichotomous (binary) fashion the results can be reported in a variety of ways. The relative effect of treatment may be reported as an Odds Ratio, or as a Risk Ratio. These ratios describe the effect of treatment in reducing or increasing the odds or risks of events, and may be fairly independent of the patients' baseline risk status. In contrast the reduction in risk may be described as a Risk Difference (otherwise described as Absolute Risk Reduction), and this may be turned into a Number Needed to Treat (NNT) by taking the inverse of the Risk Difference.

### Discussion

#### *Example of Simpson's paradox from a Cochrane review of nursing interventions for smoking cessation*

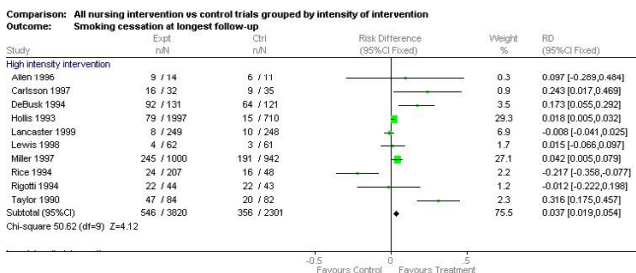
Thus, for example, a single trial of nursing interventions [1] to promote smoking cessation showed that 245 out of 1000 (24.5%) smokers were able to stop with intensive help from nurses whilst 191 out 942 (20.3%) smokers stopped without such help. In this trial 3.2% more smokers gave up with the help of nurses. This means that the number of patients needing high intensity nursing intervention in this trial in order for one extra patient to stop smoking is  $100/3.2 = 31$ .

The results of all such high intensity nursing interventions are summarised in a systematic review on the Cochrane Library [2] and are shown in Figure 1.

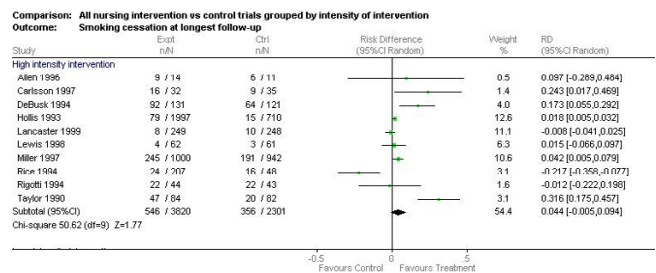
The raw totals from each trial can be added together and these are shown at the bottom of each column, 546/3820 stopped smoking in the nursing group and 356/2301 did so without intervention from nurses. Some authors have advocated the use of these proportions to calculate NNT from the pooled results of systematic reviews. [3] I would argue that this is not a reliable method to derive NNTs from pooled data, because Simpson's paradox [4] can occur when there is imbalance in the number of included patients between the arms of individual trials.

Adding together the total number of patients from each trial who stop smoking following intensive nursing intervention gives a cessation rate 14.3% (546 out of 3820), whilst adding those who stop in the placebo groups yields 15.5% (356 out of 2301). Thus it would appear that overall 1.2% less patients stop smoking when nurses intervene, suggesting a Number Needed to Harm or NNT(H) of  $100/1.2 = 83$ . This is in contrast to the result of calculating the Risk Difference from each trial and combining the weighted trial results, which yield a pooled Risk Difference of 0.037, that is to say that 3.7% more patients stop with help from the nurses ( $NNT\ 100/0.037 = 27$ ).

The direction of the effect has been reversed using the raw totals due to Simpson's paradox, because the arms in the individual trials are not equal in size. For example the Hollis study has roughly three times as many patients who have a nursing intervention as the control group, because there are four active arms in the trial. Three of these arms used nursing intervention and these have been added together for the purposes of the meta-analysis. This avoids the danger of triple-counting the control group, which



**Figure 1**  
Risk Difference meta-analysis of trials of high intensity nursing intervention for smoking cessation (showing proportion of patients in each trial arm who had ceased smoking at the longest follow-up). The trial results are combined using the Mantel-Haenszel fixed effects method.



**Figure 2**  
Risk Difference meta-analysis of trials of high intensity nursing intervention for smoking cessation (showing proportion of patients in each trial arm who had ceased smoking at the longest follow-up). The meta-analysis is performed using a random effects model and the confidence interval of the pooled result is wider than for the fixed effects method.

would occur if each arm were entered separately against the control arm in the meta-analysis.

There are further theoretical reasons why the raw totals should not be used to calculate NNT or any other pooled effect from meta-analysis of controlled trials. In each trial the patients are randomly assigned to either the treatment or control group, and conventional analysis using weighted trial effects preserves the benefit of randomisation by considering the patients in each trial only in comparison with that trial's controls. When raw totals are added together the treated patients in one trial are compared to the controls in all the trials and the benefits of randomisation are lost.

Finally adding the patients as if they all came from a single trial prevents us from considering the differences in treatment effects between the trials, and as a consequence the confidence intervals may be too narrow. In the example shown in Figure 1 the heterogeneity is high and there is an argument for the use of a random effects model that incorporates the possibility of non-random differences between the included studies. The results of such a model are shown in Figure 2.

The point estimate is altered slightly using the random effects model because the weights given to each trial are different and more weight is given to the smaller trials. The confidence interval is considerably wider using a random effects model and now includes a risk difference of zero. The fact that the pooled result is altered when a random model is used should be included, as part of a sensitivity analysis, but this cannot be done if only the sum of the raw totals is considered.

It should be noted that Simpson's paradox will also alter the direction of effect of all the other summary statistics if

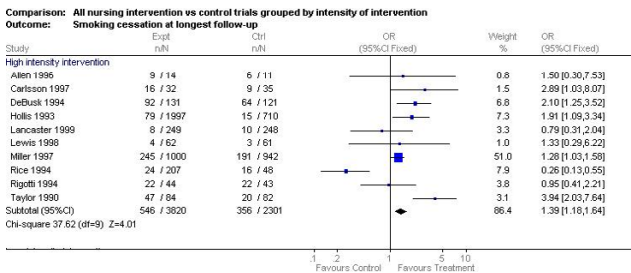
the raw totals are used, and that this is independent of whether statistical significance is present. The shift is caused by imbalance in the size or the treatment and control arms (not the heterogeneity that is present in this case).

**Limitations of Number Needed to Treat as a summary statistic**

Thus far I have considered the disadvantages of using raw totals from meta-analysis to calculate NNT from pooled data instead of a pooled Risk Difference, which can be inverted (with its confidence interval) to form a pooled NNT [5]. There are however inherent limitations in the use of NNT and Risk Difference as summary statistics. The strength and weakness of these absolute measures is that they are very dependent upon the baseline risk of the patients included in the constituent trials and on the duration of follow-up [6].

As can be seen from Figure 1 the average control event rate for the patients who do not receive nursing intervention is 15.5%. In other words 15.5% manage to stop smoking without the help of the nurses. This is a high figure and inspection of the individual trial placebo arms reveals a good deal of variation from over 50% who stop smoking in the Allen and DeBusk trials to 2% in the Hollis trial. This may well reflect variation in the type of patients included in the different trials and in the co-interventions that were used. However the pooled event rate of 15.5% is heavily influenced by the characteristics of the patients and co-interventions in the largest trials. This in turn will influence the pooled risk difference, and its corresponding inverse, the NNT.

If we pool the results of these studies using Odds Ratios (which have better properties for calculation [7,8]) the result is shown in Figure 3.



**Figure 3**  
Odds Ratio meta-analysis of trials of high intensity nursing intervention for smoking cessation (showing proportion of patients in each trial arm who had ceased smoking at the longest follow-up). The trial results are combined using the Mantel-Haenszel fixed effects method.

The pooled Odds Ratio of 1.39 can be applied to the control group odds and the NNT derived from the type of patients in the Hollis study with low rates of smoking cessation (2% in the control group) would be 125, whilst the patients of the type included in the Miller study (20% cessation rate in the control group) would yield an NNT of 17. For the "average" patient (15.5% cessation rates in the pooled control groups) the NNT is 21.

If NNT is used as the main descriptive measure of the result of trials or meta-analyses without reference to the baseline risks of the included patients there is a danger of seriously misleading the reader, who in the above case might wrongly assume that the type of nursing intervention used by Miller was far more effective than that used by Hollis. In trying to make the results easier for clinicians to apply, the use of NNT without reference to baseline risk may inadvertently give the impression of spurious differences between treatments.

An example of this problem can be seen in a recent Bandolier article on Nicotine replacement for smoking cessation [9], in which the authors conclude "the evidence for gum is a bit flakey, because the NNTs increase substantially with larger trials and in those with lower control cessation rates." Inspection of the table in this article shows that the overall placebo cessation rate is 12% for gum and 8% for patches; in the analysis of those trials with control cessation rates of under 10%, the placebo rate is 6% for gum and patches. The NNT would be expected to rise with lower control cessation rates since the relative effect of treatment is fairly constant (Pooled Odds Ratio of 1.63 for all gum trials and 1.64 for those with a control rate under 10%), so the greater rise in NNT in the gum group may just be due to the difference in the control rates.

Moreover there is imbalance in the size of the trial arms in the included studies, so use of the raw totals for calculation is subject to Simpson's paradox as well. Meta-analysis with pooled risk differences for all trials would produce an NNT of 17 for gum (compared to the reported NNT of 12 from the raw totals) and an NNT of 18 for patches (compared to the reported NNT of 17 from raw totals).

**A suggested way forward**

Egger [10] and Engels [11] have argued for the use of relative measures (Odds Ratios or Risk Ratios) as a preferred summary statistic for meta-analysis, and Engels [11] and Deeks [12] have shown that Risk Differences are on average more heterogeneous than Odds Ratios and Risk Ratios when used to pool effects in meta-analyses. I would support Engels and Senn, who suggest using Odds Ratios as the summary statistic since, in contrast to Relative Risks, the weights and effect sizes are less dependent upon whether the data is entered as beneficial or adverse out-

comes. The pooled Odds Ratio can then be converted into an NNT for individual types of patient by assessing their baseline risk from the characteristics of those groups of patients that have been included in the different trials, or by using other information (from observational studies for example) to assess their risk. The confidence intervals of the Odds Ratios can also be converted into a confidence interval of the NNT, and using this method the confidence interval reflects uncertainty around the effect of treatment but not the individual patient risk (in contrast to the confidence interval derived from the pooled risk difference).

It is wise to limit this conversion to patients whose expected risk lies within the control event rates of the included trials (which in the example from the Cochrane review is 2% to 50% cessation of smoking), and to use this method cautiously where, as in this example, there is heterogeneity in the Odds Ratio.

The formula for conversion of the pooled Odds ratio to an individual NNT is rather tedious [12] and could be somewhat time consuming to calculate for clinicians, but there are NNT calculators available on the internet which will do the job quickly and some will also produce a graphical display of results to aid explanation. [13,14]

## Conclusion

More clarity is required from the authors of meta-analyses that calculate NNT from pooled data; the method used to derive the NNT should be specified (as well as the control event rate), and simple addition of the patients from each trial should be avoided. The data from individual trials in a meta-analysis should always be included in published meta-analyses, as electronic publication has relieved the constraints of space in paper journals. One of the principles of systematic reviewing is that the processes should be transparent, so that others can repeat the analysis of the data so as to assess the influence of using different methodologies.

## Competing Interests

CC devised Visual Rx, which is a free calculator available on the Internet for use with Java-enabled browsers to allow clinicians to convert pooled Odds Ratios into a Number Needed to Treat.

## Editorial note

Two commentaries on this article are published alongside: from Andrew Moore and Henry McQuay [15] and from Doug Altman and John Deeks [16]. Andrew Moore and Henry McQuay are the editors of *Bandolier*, which published the two articles commented on above [3],[9].

## Acknowledgements

I am grateful to Jon Deeks and Steven Senn for helpful comments on the assessment of binary outcomes in Meta-analysis.

## References

1. Miller NH, Smith PM, DeBusk RF, Sobel DS, Taylor CB: **Smoking cessation in hospitalized patients – Results of a randomized trial.** *Arch Intern Med* 1997, **157**:409-15
2. Rice VH, Stead LF: **Nursing interventions for smoking cessation (Cochrane Review).** In: *The Cochrane Library*, 2001
3. **Number Needed to Treat (NNT)** *Bandolier* 1999, **59**:i-iv [http://www.jr2.ox.ac.uk/bandolier/band59/NNT1.html]
4. Blyth C: **On Simpson's paradox and the sure-thing principle.** *J Am Statistical Assn* 1972, **67**:364-366
5. Altman DG: **Confidence intervals for the number needed to treat.** *BMJ* 1998, **317**:1309-1312
6. Smeeth L, Haines A, Ebrahim S: **Numbers needed to treat derived from meta-analyses – sometimes informative, usually misleading.** *BMJ* 1999, **318**:1548-1551
7. Senn S: **At Odds with reality.** *BMJ Rapid Response.* [http://www.bmj.com/cgi/eletters/318/7197/1548#EL2]
8. Hutton JL: **Number needed to treat: properties and problems.** *J R Statist Soc A* 2000, **163**:403-419
9. **Nicotine replacement for smoking cessation.** *Bandolier* 2001, **86**:5-6 [http://www.jr2.ox.ac.uk/bandolier/band86/b86-3.html]
10. Egger M, Smith GD, Phillips AN: **Meta-analysis: Principles and procedures.** *BMJ* 1997, **315**:1533-1537
11. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J: **Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses.** *Stat Med* 2000, **19**:1707-1728
12. Deeks JJ, Altman DG: **Effect measures for meta-analysis of trials with binary outcomes.** In: *Systematic reviews in Health Care: Meta-Analysis in Context.* 2001:322-325
13. [http://www.anu.edu.au/nceph/surfstat/surfstat-home/tables/nnt.html]
14. [http://www.nntonline.net]
15. Moore RA, Gavaghan D, Edwards JE, Wiffen P, McQuay HJ: **Pooling data for number needed to treat: no problems for apples.** *BMC Medical Research Methodology* 2002
16. Altman D, Deeks J: **Meta-analysis, Simpson's paradox, and the number needed to treat.** *BMC Medical Research Methodology* 2002:3

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMedCentral will be the most significant development for disseminating the results of biomedical research in our lifetime."

Paul Nurse, Director-General, Imperial Cancer Research Fund

Publish with **BMC** and your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours - you keep the copyright



Submit your manuscript here:

http://www.biomedcentral.com/manuscript/

editorial@biomedcentral.com