BMC Medical Research Methodology

Debate

Meta-analysis, Simpson's paradox, and the number needed to treat

Douglas G Altman* and Jonathan J Deeks

Address: Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Headington, Oxford, UK

E-mail: Douglas G Altman* - doug.altman@cancer.org.uk; Jonathan J Deeks - jon.deeks@cancer.org.uk
*Corresponding author

Published: 25 January 2002

BMC Medical Research Methodology 2002, **2**:3

Received: 27 December 2001
Accepted: 25 January 2002

This article is available from: http://www.biomedcentral.com/1471-2288/2/3

Abstract

Background: There is debate concerning methods for calculating numbers needed to treat (NNT) from results of systematic reviews.

Methods: We investigate the susceptibility to bias for alternative methods for calculating NNTs through illustrative examples and mathematical theory.

Results: Two competing methods have been recommended: one method involves calculating the NNT from meta-analytical estimates, the other by treating the data as if it all arose from a single trial. The 'treat-as-one-trial' method was found to be susceptible to bias when there were imbalances between groups within one or more trials in the meta-analysis (Simpson's paradox). Calculation of NNTs from meta-analytical estimates is not prone to the same bias. The method of calculating the NNT from a meta-analysis depends on the treatment effect used. When relative measures of treatment effect are used the estimates of NNTs can be tailored to the level of baseline risk.

Conclusions: The treat-as-one-trial method of calculating numbers needed to treat should not be used as it is prone to bias. Analysts should always report the method they use to compute estimates to enable readers to judge whether it is appropriate.

Introduction

Cates [1] considers "how should we pool data?", focusing in particular on the calculation of the number needed to treat (NNT). He explains how Simpson's paradox may lead to the wrong answer when the NNT is calculated in a particular way. Moore and colleagues [2], responding to Cates, focus on the specific example used by Cates and address the question "which data should be pooled?". Unfortunately, they largely ignore Cates' methodological point. We suspect that many readers of these two articles will come away rather confused, so our aim is to try to clarify some of these issues.

First, we note that incorrect methods may give the right answer some or even most of the time, but that does not mean that they should be used or recommended. They might be copied by others and used in situations where they do not work. For example, in the fraction 16/64 we can cancel the two sixes to give the right answer of $^1/_4$. This method will almost always give the wrong answer, however, such as in the fraction 26/61.

Methods of meta-analysis

Combining results of several studies

Meta-analysis is a statistical method for combining the results from two or more independent studies. It is used when it is felt that the studies are similar enough to make combining the results a sensible thing to do. Judging combinability requires what Moore et al refer to as clinical common sense, although there are additional considerations such as the quality of the separate studies.

Meta-analysis is a two-stage process. First the data for each study are summarised, and then those summaries are statistically combined. All widely used methods of meta-analysis are of this type [3]. The method used (and recommended) by Moore and colleagues for calculating the number needed to treat uses the data as if they came from a single study, and is not a valid approach to meta-analysis; indeed it is not really a meta-analysis. They argue that it generally gives similar answers to those obtained using the correct approach. This is true, but sometimes it gives the wrong answer, as Cates shows and we will also show.

One complication is that both of these articles refer to 'pooling', a term that does not have a unique meaning. Cates [1] uses it to mean combining the treatment effects from separate studies while Moore et al [2] use the term to mean adding up the numbers as if the data were all from one large study. We will refer to two stage meta-analysis (as recommended by Cates) as standard meta-analysis, and adding up the numbers (as recommended by Moore et al) as the 'treat-as-one-trial' method.

Calculating the number needed to treat (NNT)

For a single trial with a binary outcome the number needed to treat (NNT) is estimated as the reciprocal of the absolute risk difference. In meta-analysis the situation is complicated by the fact that there are several possible ways of summarising effect – risk ratio, odds ratio, and risk difference (and there are in fact two risk ratios) [4,5]. The principles of meta-analysis and the arguments below apply regardless of the choice of outcome summary.

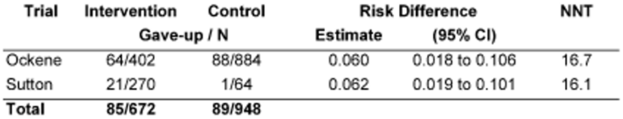
The NNT can be obtained directly from a meta-analysis that pools risk differences from several trials. However, it is common for systematic reviews to include trials with different baseline risks (estimated by the event rate in the control group) due to differences in trial characteristics such as case-mix and duration of follow-up. Inclusion of trials with different baseline risks is not always a problem. This is because the treatment effect may be quite constant across trials, if expressed as relative effects (such as the odds ratio or relative risk), rather than as absolute effects (such as the risk difference)[4–6]. As Cates notes, the NNT can be estimated from an analysis based on relative effect measures, if one specifies the anticipated (baseline) risk in untreated patients. Further, this allows the NNT to be tailored to patients with differing baseline risks, rather than quoting one overall NNT [7,8]. Moore et al calculate a single NNT from several trials directly from the proportions, with events summed over all studies.

Impact of unbalanced numbers

The vast majority of RCTs have approximately equal numbers in each arm. Unequal numbers may occasionally arise from deliberate unbalanced randomisation but more often because two or more groups receiving similar treatments are combined for a meta-analysis.

Imbalance in one or more trials has no effect on standard meta-analysis. Separate estimates are obtained from each trial that take correct account of the numbers per arm. No trial influences the impact of the other trials. By contrast, the treat-as-one-trial approach, which adds all the data for like arms across trials, certainly can be influenced by imbalance. As Cates notes, this is a form of Simpson's paradox (which is not a true paradox but rather a bias).

Consider, for example, two RCTs from the systematic review of nicotine gum [9] referred to by Cates. These trials, both with unbalanced allocation, are summarised as follows:



Trial	Intervention	Control	Risk Difference		NNT
	Gave-up / N		Estimate	(95% CI)	
Ockene	64/402	88/884	0.060	0.018 to 0.106	16.7
Sutton	21/270	1/64	0.062	0.019 to 0.101	16.1
Total	**85/672**	**89/948**			

Both trials showed a risk difference of 0.06 (or 6%). Common sense (statistical rather than clinical) tells us that when we pool the results of these two trials we should get an answer that lies between those of the two trials being combined. Results of meta-analyses of these two trials are shown in Table 1(a). While the standard approach indeed gives a pooled risk difference of 6%, the treat-as-one-trial method pools two trials each with a risk difference of 6% and gets an estimated effect of 3%. Consequently the NNT from this approach is 31 compared to the correct value of 17. This simple example shows why Moore et al are wrong to dismiss Simpson's paradox (and Cates' argument).

The full review [10] includes 51 trials of which several are unbalanced. Overall, summing across the 51 trials the quit rates in the intervention and control groups were 1508/7674 (19.7%) and 1110/9613 (11.5%) respectively. Meta-analyses of the 51 trials using the two methods just discussed are shown in Table 1(b). Even here, where most trials are balanced, the treat-as-one-trial approach gives an incorrect answer. In fact, the 95% confidence intervals for the two estimates barely overlap. Thus the statement of Moore et al that "the problem of trial imbalance

Table 1: Results of meta-analyses of trials of nicotine gum to reduce smoking using standard and 'treat-as-one-trial' methods, with risk difference as effect measure

	Pooled risk difference (95% CI)	P	NNT
(a) Two selected trials (see text)			
Standard meta-analysis (Mantel-Haenszel method)	0.060 (0.025 to 0.095)	0.001	16.7
Treat-as-one-trial	0.033 (0.001 to 0.064)	0.04	30.7
(b) All 51 trials			
Standard meta-analysis (Mantel-Haenszel method)	0.061 (0.050 to 0.071)	<0.001	16.4
Treat-as-one-trial	0.081 (0.070 to 0.092)	<0.001	12.3

comes from combining different interventions as if they were a single intervention" is incorrect. We also note that, compared to standard meta-analysis, the treat-as-one-trial method gives greater weight to large trials and will tend to give narrower confidence intervals.

The risk of such imbalance having an effect increases with (a) increasing discrepancy in size of treatment groups, (b) increasing variation in control group event rates, (c) increasing heterogeneity in treatment effects between the studies. Standard (stratified) meta-analytical methods are not affected by imbalance and so are not affected by Simpson's paradox.

Discussion

Cates' illustrative example

Cates used for illustration a subset of the trials included in a systematic review published on the Cochrane Library [10]. He used the ten trials of high intensity nursing interventions to encourage smoking cessation – his paper shows the actual results of each trial. He demonstrates that the treat-as-one-trial approach gives an answer in the opposite direction to that from standard meta-analysis, and attributes this to the fact that the analysis is not immune to the effect of Simpson's paradox. Moore and colleagues believe that these trials should not all be grouped together and that the paradoxical answer arose from the inappropriate pooling. They thus split the trials by setting – seven trials done in hospital settings and three in primary care. We agree that it may be wise to investigate whether setting affects effectiveness before combining all of these trials, but we do not agree with the method by which Moore et al undertakes this investigation, nor their implication that the methodological problem could only occur if one pools trials inappropriately.

Moore et al show that the results are somewhat different in the two subgroups (their table 1). They argue that among hospital patients the relative treatment benefit was statistically significant (RR = 1.3, 95% CI 1.1 to 1.6) and that the NNT of 14 (95% CI 9 to 26) is a useful result. In unselected primary care patients (their definition) there was a not a statistically significant result and the NNT was 222. They conclude that nursing interventions are "probably ineffective" in these patients. Our comments are:

1 The evidence of an effect in hospital patients is fairly weak, being only marginally statistically significant.

2 The estimated effect in primary care has a confidence interval that goes way above the whole CI for secondary care, so that it is quite inappropriate to dismiss the intervention on such slight evidence. (However, it seems as if the result for this group is based on a random effects analysis.)

3 The comparison of the subgroups should not be based on comparison of P values (one significant and one not), whether explicit or, as here, implicit [11]. By a formal test of interaction the pooled results from the two groups of trials are not significantly different.

4 No account is taken of the quality of these trials. For example, two trials (including the largest) were not properly randomised and another was a cluster randomised trial that was analysed wrongly [10].

5 Interested readers should consult the Cochrane review [10] to get the 'real' results. The review includes data from additional trials and analyses stratified by patient type and type of intervention. The authors conclude that the intervention is beneficial in both hospitalised and non-hospitalised patients (RR = 1.28 (1.03 to 1.61) (random effects model)), there being no-significant difference in RR between primary care and secondary care patients (P = 0.42). Applying the overall relative increase in cessation rates of 28% gives: (a) an NNT of 89 for primary care trials, based on the median placebo quit rate of 4%; (b) an NNT of 12 for secondary care trials, based on the median placebo quit rate of 30%.

Table 2: Results of meta-analyses of trials of high intensity nursing to reduce smoking using standard§ and 'treat-as-one-trial' methods, with relative risk as effect measure

Setting (no. of trials)	**Number quitting/total**		**Pooled risk ratio (95% CI)**	**NNT (95% CI)**
	Intervention	Control		
Hospital (7)	435/1367	318/1295		
Meta-analysis			1.30 (1.16 to 1.47)	13.6 (8.7 to 25.5)¶¶
Treat-as-one-trial			1.30 (1.15 to 1.47)	13.8 (9.4 to 25.9)
Primary Care (3)*	111/2453	41/1006		
Meta-analysis			1.01 (0.71 to 1.42)	2454 (58.4 to H84.6¶)¶¶
Treat-as-one-trial			1.11 (0.78 to 1.58)	222.5 (52.0 to H97.7¶)
Primary Care (2)**	87/2246	25/958		
Meta-analysis			1.54 (0.97 to 2.44)	71.0 (26.6 to H1277¶)¶¶
Treat-as-one-trial			1.48 (0.96 to 2.30)	79.1(39.2 to H4369¶)

§ Mantel-Haenszel method (fixed effect) * as defined by Moore et al ** excluding reference [12]¶ NNT for harm [13]¶¶NNT calculated using the event rate among controls and the relative risk reduction [4]

In addition, one of the trials that Moore et al included as a trial of "unselected primary care patients", was in fact done in patients with cardiovascular problems [12]. Our common sense tells us to exclude that trial. We summarise the results of meta-analyses in Table 32, here using the risk ratio (relative risk) as Moore et al did. (We cannot exactly reproduce the results given by Moore et al as we are not sure which method they used to obtain the relative risks.)

We agree with Moore et al that it helps to split the trials by setting to gauge the differential impact of the intervention – the NNTs in the two settings are clearly different. But notably, once the trial Moore et al inappropriately include as a primary care trial is excluded, the results expressed as risk ratios are surprisingly similar for both settings, both in the subset of data presented by Cates [1], and the full results of the review [10]. There is also little difference between the results using standard meta-analysis and the treat-as-one-trial method, but as we noted above, although use of the treat as one trial method increases the risk of bias, bias will not always be seen.

Other points

Moore's method of grouping trials with similar control-group event rates does appear to reduce the problem, as would be predicted, but it cannot eradicate it. It is important to note that this is a 'results based' categorisation that is not based on a priori clinical criteria. Also, grouping by control group event rate only reduces the bias for analysis of risk differences, and not for relative effect measures. But grouping by control group event rate leads to worse problems as the treatment effect is correlated with the control group event rate [14].

Moore et al also say that an analysis based on pooling risk differences assumes that the control group event rate is the same in all trials. This statement is incorrect – what is assumed (in a fixed effect analysis) is that the true treatment effect expressed as a risk difference was the same in all trials. There is no statistical assumption that the event rates per arm are similar across trials. There may be other reasons to worry about this issue, as discussed above.

Cates comments on the choice of effect measure for binary data. It is true that there is empirical evidence that relative effect measures are more likely to be homogenous across trials this does not mean that absolute measures should never be used. More seriously, it does not help us unravel the choice between odds ratio and risk ratio, where the empirical evidence shows no such dominance of one measure [5,6].

Systematic reviews involve subjectivity, for example in deciding which studies to analyse. It is essential that reviews include the summary data from each study so that readers can examine the implications of some of these judgements [15]. The methods of analysis should also be specified, including the method to derive an estimated NNT. For example, it would be misleading not to report the use of the treat-as-one-trial method.

Conclusions

We agree with earlier comments from Moore and McQuay: "NNT is a tool. Like any tool, when used appropriately it will be helpful and effective. What we have to do is to ensure that in any given situation we know what the rules are for using the tools correctly." [16]

Given the choice between a method that always gives a right answer and a method that sometimes or even usually gives the right answer, it is common sense to use the one that always gives the right answer. Adding numbers may have some value for simple descriptive purposes, but the treat-as-one-trial method should not be used for substantive analysis.

Competing interests

None declared

Acknowledgements

DGA is funded by Cancer Research UK and JJD by the NHS Research and Development programme.

References

1. Cates C: **Simpson's paradox and calculation of number needed to treat from meta-analysis.** *BMC Medical Research Methodology* 2002, **2**:1
2. Moore RA, Gavaghan DJ, Edwards JE, Wiffen P, McQuay HJ: **Pooling data for number needed to treat: no problems for apples.** *BMC Medical Research Methodology 2002, 2:2* 2002, **2**:2
3. Deeks JJ, Altman DG, Bradburn MJ: **Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis.** *In: Systematic Reviews in Health Care. Meta-analysis in Context.* 2001285-312
4. Deeks JJ, Altman DG: **Effect measures for meta-analysis of trials with binary outcomes.** *In: Systematic Reviews in Health Care. Meta-analysis in Context.* 2001313-335
5. Deeks JJ: **Issues in selection of a summary statistic for meta-analysis of clinical trials with binary data.** *Stat Med*
6. 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
7. Cook RJ, Sackett DL: **The number needed to treat: a clinically useful measure of treatment effect.** *BMJ* 1995, **310**:452-454
8. Ebrahim S: **Numbers needed to treat derived from meta-analyses: pitfalls and cautions.** *In: Systematic Reviews in Health Care. Meta-analysis in Context.* 2001386-399
9. Silagy C, Lancaster T, Stead L, Mant D, Fowler G: **Nicotine replacement therapy for smoking cessation (Cochrane Review).** *In: The Cochrane Library,* 2001
10. Rice VH, Stead LF: **Nursing interventions for smoking cessation (Cochrane Review).** *In: The Cochrane Library,* 2001
11. Matthews JNS, Altman DG: **Interaction 2: Compare effect sizes not P values.** *BMJ* 1996, **313**:808
12. Rice VH, Fox DH, Lepczyk M, Sieggreen M, Mullin M, Jarosz P, Templin T: **A comparison of nursing interventions for smoking cessation in adults with cardiovascular health problems.** *Heart Lung* 1994, **23**:473-86
13. Altman DG: **Confidence intervals for the number needed to treat.** *BMJ* 1998, **317**:1309-1312
14. Sharp S: **Analysing the relationship between treatment benefit and underlying risk: precautions and recommendations.** *In: Systematic Reviews in Health Care. Meta-analysis in Context.* 2001176-188
15. Altman DG, Cates C: **Authors should make their data available.** *BMJ* 2001, **323**:1069-1070
16. Moore A, McQuay H: **Numbers needed to treat derived from meta analysis. NNT is a tool, to be used appropriately.** *BMJ* 1999, **319**:1200