BMC Medical Research Methodology

Debate

Pooling data for Number Needed to Treat: no problems for apples

R Andrew Moore*[1], David J Gavaghan[2], Jayne E Edwards[1], Phillip Wiffen[1] and Henry J McQuay[1]

Address: [1]Pain Research & Nuffield Department of Anaesthetics, University of Oxford, Oxford Radcliffe Hospital, The Churchill, Headington, Oxford, UK and [2]Oxford University Computing Laboratory, Wolfson Building, Parks Rd, Oxford OX1 3QD, UK

E-mail: R Andrew Moore* - andrew.moore@pru.ox.ac.uk; David J Gavaghan - David.Gavaghan@comlab.oxford.ac.uk; Jayne E Edwards - jayne.edwards@pru.ox.ac.uk; Phillip Wiffen - phil.wiffen@pru.ox.ac.uk; Henry J McQuay - henry.mcquay@pru.ox.ac.uk

*Corresponding author

Published: 25 January 2002

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

Received: 24 October 2001
Accepted: 25 January 2002

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

Abstract

Objective: To consider the problem of the calculation of number needed to treat (NNT) derived from risk difference, odds ratio, and raw pooled events shown to give different results using data from a review of nursing interventions for smoking cessation.

Discussion: A review of nursing interventions for smoking cessation from the Cochrane Library provided different values for NNT depending on how NNTs were calculated. The Cochrane review was evaluated for clinical heterogeneity using L'Abbé plot and subsequent analysis by secondary and primary care settings.

Three studies in primary care had low (4%) baseline quit rates, and nursing interventions were without effect. Seven trials in hospital settings with patients after cardiac surgery, or heart attack, or even with cancer, had high baseline quit rates (25%). Nursing intervention to stop smoking in the hospital setting was effective, with an NNT of 14 (95% confidence interval 9 to 26). The assumptions involved in using risk difference and odds ratio scales for calculating NNTs are discussed.

Summary: Clinical common sense and concentration on raw data helps to detect clinical heterogeneity. Once robust statistical tests have told us that an intervention works, we then need to know how well it works. The number needed to treat or harm is just one way of showing that, and when used sensibly can be a useful tool.

Background

Cates [1] concentrates on Simpson's paradox, which relates to problems that can arise when there is an imbalance between treatment and placebo arms in controlled trials. This "paradox" is hardly new, having first been discussed by E.H. Simpson 50 years ago [2], and is now a staple of any undergraduate statistics course. Cates further contends that NNTs should be calculated from weighted risk differences (or odds ratios) rather than pooled raw events, although this is relevant to Simpson's paradox only if inappropriate statistical methods are being used in inappropriate circumstances.

It all comes down to the old problem of meta-analysis, of whether you are comparing apples with something else, and how you count the apples when you've got them.

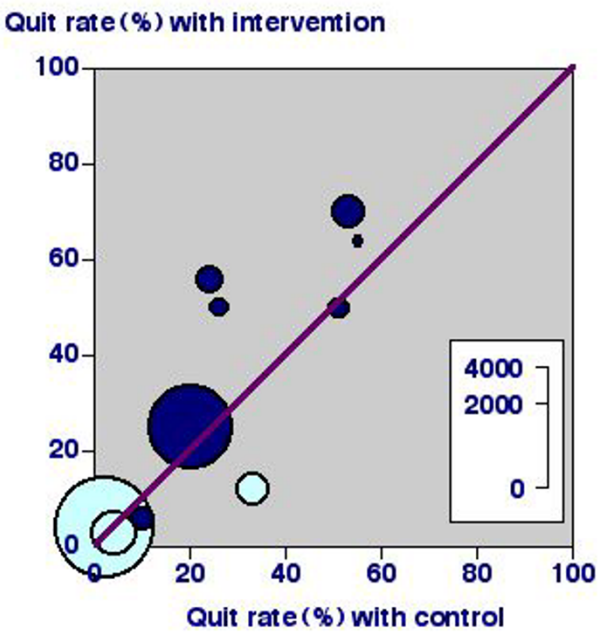

Figure 1
L'Abbé plot of nursing interventions versus control for smoking cessation at longest follow up. Dark blue symbols indicate studies in a hospital setting and light blue symbols those in a primary care setting.

The problem

All of this is based on a numerical analysis of a Cochrane review of nursing interventions for smoking cessation [3]. The pooled raw data show that fewer people (14.3%) stop smoking with a nursing intervention than with control (15.6%): that is, the intervention does not work. Cates wants us to believe that the "real" answer is different, and that 3.7% more patients stop smoking with the intervention than with control.

Clinical heterogeneity

This is, indeed, a paradox. But complicated statistical arguments may not be the best way of dealing with it. When faced with something that looks wrong, the first rule is to look at the raw data. In this case a simple graphical representation [4] of what happened in each trial helps.

Figure 1 shows a plot of the percentage of quitters for intervention (Y-axis) and control (X-axis) for individual trials. There is a huge variation, from about 2% to 60-70% in each case. Since stopping smoking is universally judged to be very difficult for most people, trials showing quit rates of up to 55% without any intervention need a second look. When examining the individual trials we find that three (light blue) were done in primary care populations with no particular desire to stop smoking. We find that seven trials (dark blue) were done in a hospital setting, and included patients who had heart attacks, cardiac surgery, or even had cancer. It is not surprising that their attitude to stopping smoking was somewhat different.

L'Abbé plots using raw data will almost always show up clinical heterogeneity, whereas Forrest plots, in which data have been manipulated to create statistical outputs like odds ratios or risk differences, will not.

Of course there are many other sources of clinical heterogeneity in these ten trials, apart from populations tested. It was unlikely, for instance, that any two interventions were the same, and we know that criteria for cessation were different even *within* studies. Moreover, the problem of trial imbalance comes from combining different interventions as if they were a single intervention [5,6].

The "real" results

If we believe that patients after coronary artery bypass, for instance, are different in their motivation to stop smoking from unselected general practice patients smoking at least one cigarette per day, and analyse them separately, a more sensible picture emerges (Table 1).

In hospital patients there was a significant relative benefit from nursing interventions (using both random and fixed effects models), with 7% more quitting smoking, and generating an NNT of 14. That is, for every 14 patients given a nursing intervention, one more will quit smoking than would have done without the nursing intervention. Many will see this as a useful result, especially as these patients need advice about other aspects of their lifestyle, like diet and exercise.

In unselected primary care patients there was no benefit from nursing interventions (using both random and fixed effects models). Two of these three trials were unbalanced, but even choosing the most effective of three interventions in each to lose the imbalance would not affect the result. Nursing interventions in unselected primary care patients are probably not effective.

These are the real results, and they are quite clear, despite some misgivings about the trials.

Discussion

Different methods of calculating NNTs, using pooled raw event rates, or from odds ratios, relative risk, or risk difference, will generally give much the same answer when pooling information where the same outcome is measured over the same time for the same intervention in similar patients, when the effect is large and where there is a sufficiency of information. Variation in event rates may just be a product of size [7], but when large variations exist

Table 1: Results for nursing interventions versus control analysed by hospital and primary care setting

		Number quitting/total		Percent quitting (95% CI)			
Setting	**Number of studies**	**Intervention**	**Control**	**Intervention**	**Control**	**Relative benefit (95% CI)**	**NNT (95% CI)**
Hospital	7	435/1367	318/1295	32 (30 to 34)	25 (23 to 27)	1.3 (1.1 to 1.6)	14 (9 to 26)
Primary care	3	111/2453	41/1006	4.5 (3.7 to 5.3)	4.1 (2.9 to 5.3)	0.8 (0.3 to 2.6)	222 (52 to -98)
Combined	10	546/3820	359/2301	14 (12 to 16)	16 (14 to 18)	1.2 (0.9 to 1.6)	-76 (184 to -32)

the presence of clinical heterogeneity should first be sought.

Unfortunately much of the discussion on statistical techniques put forward in Cates' article is confused and misleading. Other authors have discussed these issues cogently and coherently and interested readers are referred to these articles [8–12]. We will, however, comment on one particular point which recurs throughout the article, on the validity of pooling data, since this is fundamental to meta-analysis.

Any technique for combining data from a series of studies or trials of a particular treatment or intervention must be based on a set of assumptions about the nature of any positive or negative effect that results. These assumptions are discussed below.

1 In the risk difference scale, the traditional assumption is that the event rates are fixed in each of the control (control event rate or CER) and treatment groups (experimental event rate or EER). Any variation in the observed event rates is then attributed to random chance. If the trials being combined are truly clinically homogeneous and have been designed properly (for example, with balanced arms), which is the situation that will commonly pertain, then in this (and only in this) case it is appropriate to pool raw data to obtain combined measures such as NNTs.

More recently the "random effects" model [10] has been suggested to allow calculation of summary measures when the degree of "statistical" heterogeneity is greater than that occurring by random chance. This technique is based on what Thompson & Pocock [11] have described as "the peculiar premise that the trials done are representative of some hypothetical population of trials, and on the unrealistic assumption that the heterogeneity between studies can be represented by a single variance". We agree with other authors [11,12] who contend that where considerable heterogeneity is observed it is more useful to investigate what may have caused those differences (such as the underlying differences between the inhospital and primary care patients in the nursing intervention study) than to attempt to overcome them by statistical methods of unproven validity.

2 The assumptions underlying the odds ratio scale are very different. Here we assume that the ratio of the odds of observing an effect (e.g. smoking cessation) in the treatment group to the odds of observing that effect in the control group are constant between trials. This scale is appropriate where it can be demonstrated that whilst the underlying event rates in both the control and treatment arms of the trial may vary, the relative odds of those in whom we observe a particular effect remains fixed.

Techniques for combining odds ratios from several studies were developed primarily for case control studies (particularly cancer trials) to overcome problems due to possible confounding factors (such as age) by stratifying the data into internally homogeneous strata, then testing the hypothesis that the odds ratio remains constant across the strata. The odds ratio has been proposed as an appropriate technique for meta-analysis since it allows combination of the results from trials with widely differing control event rates, but it is clearly a matter of some contention whether such trials can be considered to be clinically homogeneous. In particular, it seems to us to be very unwise to use a summary odds ratio to calculate an NNT value (even if the associated CER is quoted) since the NNT is, by definition, dependent on the assumption of a fixed underlying control event rate, whilst the odds ratio, also by definition, is not. Any such NNT would therefore be of very questionable value.

Our practice (as reflected in the two articles published in Bandolier that Cates comments on; [14], [15]) of pooling raw events to calculate an NNT has always been predicated on having clinically homogeneous trials in the first place, and when outcomes, interventions and duration are similar. Only then is an NNT useful, and only then will an NNT calculated in this way be correct.

Conclusions

The lesson is that systematic reviews and meta-analyses have to be done to high quality. Quality comes in different guises, which might include gross imbalance between the size of groups. What is needed is some clinical common sense and concentration on raw data. Yes, we need robust statistical tests to tell us that an intervention works, but we need also to know how well an intervention works. The number needed to treat or harm is just one way of showing how well an intervention works, and when used sensibly can be a useful tool. Among GPs in Essex it was the tool they felt most confident about using [13].

If we have only apples, then counting them should not be a problem.

Competing Interests

None declared.

Editorial note

An additional commentary on the article by Cates [1] is published alongside this [16].

Acknowledgements

The authors wish to acknowledge the incredible hard work and dedication of those who produce any systematic review and especially a Cochrane Review. Finding better ways forward is part of the learning process and the fun; we have learned much from our mistakes, and from those who continue to point them out.

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. Simpson EH: **The Interpretation of Interaction in Contingency Tables.** *Journal of the Royal Statistical Society, Ser. B* 1951, **13**:238-241
3. Rice VH, Stead LF: **Nursing interventions for smoking cessation (Cochrane Review).** *In: The Cochrane Library,* 2001
4. L'Abbé KA, Detsky AS, O'Rourke K: **Meta-analysis in clinical research.** *Ann Intern Med* 1987, **107**:224-33
5. Hollis JF, Lichstenstein E, Vogt TM, Stevens VJ, Biglan A: **Nurse-assisted counseling for smokers in primary care.** *Ann Intern Med* 1993, **118**:521-5
6. Rice VH, Fox DH, Lepczyk M, Siegreen 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
7. Moore RA, Gavaghan D, Tramèr MR, Collins SL, McQuay HJ: **Size is everything – large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects.** *Pain* 1998, **78**:209-16
8. Altman DG, Deeks JJ, Sackett DL: **Odds ratios should be avoided when events are common.** *BMJ* 1998, **317**:1318
9. Senn S: **Rare distinction and common fallacy.** *BMJ electronic letters* 2001 [www.bmj.com/cgi/eletters/317/7168/1318#EL3]
10. DerSimonian R, Laird N: **Meta-analysis of clinical trials.** *Control Clin Trial* 1986, **7**:177-88
11. Thompson SG, Pocock SJ: **Can Meta-analyses be trusted** *Lancet* 1991, **338**:1127-1130
12. Fleiss JL: **Statistical Methods for rates and Proportions.** *Wiley, New York,* 1991
13. McColl A, Smith H, White P, Field J: **General practitioners' perceptions of the route to evidence based medicine: a questionnaire survey.** *BMJ* 1998, **316**:361-365
14. **Number needed to treat (NNT).** *Bandolier* 1999, **59**:1-4 [http://www.jr2.ox.ac.uk/bandolier/band59/NNT1.html]
15. **Nicotine replacment and smoking cessation.** *Bandolier* 2001, **86**:5-6 [http://www.jr2.ox.ac.uk/bandolier/band86/b86-2.html]
16. Altman DG, Deeks JJ: **Meta analysis, Simpson's paradox, and the number needed to treat.** *BMC Medical Research Methodology* 2002