Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic?

Nick Freemantle

Department of Primary Care and General Practice, University of Birmingham, Birmingham B15 2TT

Nick Freemantle *professor of clinical epidemiology and biostatistics*

N.Freemantle@bham.ac.uk

BMJ 2001;322:989–91

Summary points

- Impressive results in subgroup analyses and secondary outcomes can be hard to interpret
- For individual patients, subgroup analyses and secondary end points can provide the best guide for clinical intervention
- Health policy decisions such as those taken by NICE aim to guide the treatment of future patients and will be difficult to change
- Health policy should be protected from undue inference by considering the results of predetermined primary outcomes

Impressive results for secondary outcomes or subgroup analyses pose problems for those trying to value the benefits observed in clinical trials. In the prospective randomised amlodipine survival evaluation study, comparing amlodipine with placebo in patients with severe heart failure, a prospectively defined subgroup of patients with non-ischaemic heart failure showed a 46% reduction in the risk of death (95% confidence interval 21% to 63%).[1] This was achieved alongside a non-significant reduction in death from any cause or admission to hospital for major cardiovascular events (P = 0.31), the prospectively defined primary outcome measure, and no observed benefits in the ischaemic group. The authors of the report commented: "Although this benefit was seen only in a subgroup of patients, it is likely that it reflects a true effect of amlodipine, since the randomisation procedure was stratified according to the cause of heart failure and a significant difference between the ischaemic and non-ischaemic strata was noted for both the primary and secondary end points of the study."[1]

This article examines the interpretation that may be placed on the results of secondary end points and subgroup analyses in the context of clinical practice and health policy. With regard to health policy, it emphasises the need for discipline in interpreting clinical trials.

Prospectively declared primary outcomes

Randomised trials commonly include a range of patients with a particular disorder and estimate the average effect of the intervention being studied. Clinicians usually want to know the likely benefits and risks for an individual patient. However, attributing benefits to secondary outcomes or specific subgroups in a trial is problematic.

Registration trials require the development of prospective protocols and statistical analysis plans.[2] These describe the inclusion and exclusion criteria for patients, treatment and its delivery, outcome assessment, and the statistical analyses. A key feature is the prospective identification of a primary outcome measure.

Secondary outcomes

Clinical trials are major undertakings for sponsors and investigators. It would be odd for a single outcome to encompass all that interests investigators. Frequently, clinical trials include several outcome measures, raising the problem that the likelihood of finding a statistically significant result by chance alone increases with the number of tests undertaken. This is the "penalty for peeking."[3] One approach is to use a Bonferroni adjustment, modifying the P value to account for the multiple tests performed and the increased probability of chance findings achieving significance. This is too high a price to pay, however, since we are not equally interested in all the statistical tests, and the statistical adjustment increases the probability of failing to detect a true effect of treatment.[4]

Identifying prospectively a primary outcome measure simplifies the situation. Suppose a trial examines the effect of a clinical treatment through a single outcome measure, and the difference between the outcome in the treatment and control groups achieves a two sided P value of 0.05. This means that the observed difference between the groups (or greater) would occur by chance alone only five times in 100. If two outcomes are examined, the situation is complicated. Indeed, if the outcomes are unrelated, the probability of one of the P values being 0.05 is approximately halved (to slightly less than 10 times in 100). Declaring at the outset that an outcome is of principal importance protects the trial from the need to deal with this problem, but it relegates secondary outcomes and prospectively defined subgroup analyses to the status of descriptors.

When licensing pharmaceuticals, the US Food and Drug Administration nearly always requires two well designed randomised trials to achieve a one sided P value of 0.025 (an overall P value of 0.001) for the prospectively identified primary outcome measures against an appropriate comparator.[5] Estimation (using confidence intervals) rather than hypothesis testing (using P values) is likely to be more helpful in interpreting the results of trials. Standard statistical procedures for estimation provide the most likely value (the point estimate of treatment effect) and a plausible range of values (95% confidence intervals) which are taken to describe the probable range of the true population effect.[6]

NICK KILLEM

Purist discipline

In 1980, the US Food and Drug Administration published its critique of the anturane reinfarction trial: "We are aware that it is unusual for an FDA critique of a clinical trial to be published in the medical literature. We believe that it is important in this instance, however, because … it illustrates so clearly the problems that may arise from subgroup analyses and exclusion of patients from analysis after they have completed a study … Our review … indicates that the cause-of-death classification and all conclusions based on it are unreliable, and that the favorable effect of sulfinpyrazone on overall mortality, especially during the first six months, depends heavily on the after-the-fact exclusion of certain deaths from the analysis."[7]

There are good grounds to suggest that a prospectively determined primary outcome based on data from all randomised patients should be used to make policy decisions.[8] This strategy will protect the decision maker from the substantive risk of undue inference.

Significant secondary end point or subgroup result

If the primary outcome measure is not statistically significant, what is the correct interpretation of the results of significant secondary outcomes or subgroup analyses? These analyses are analogous, although people often place greater confidence in secondary outcomes. Moyé comments: "The primary end point, chosen from many possible end points and afforded particular and unique attention during the trial, becomes unceremoniously unseated when it is discovered to be negative at the trial's conclusion. Like the 'crazy aunt in the attic,' the negative primary end point receives little attention in the end, is referred to only obliquely or in passing, and is left to languish in scientific backwaters."[8]

Dr Milton Packer, representing the sponsor, made the following comments to the US Food and Drug Administration representatives (Drs Wood and Shepherd) during the licensing process for carvedilol:[5]

Dr Packer: Almost all of these P values are 0.00 something, so you can do this in a variety of ways, checking for robustness of the data by adding and subtracting endpoints, obviously post hoc, after the fact, and it all comes out the same way.
Dr Wood: Except for the primary endpoints.
Dr Packer: The primary endpoints don't make it, no matter how creative you are.

Without the benefits of hindsight, the decision to license carvedilol may not have served the public interest because of the prospective uncertainty about the result. β blockers have subsequently been shown to be effective in the treatment of mild to moderate heart failure.[9 10] However, when the decision was taken, all available statistical power had been "spent" on the primary outcome, and the play of chance could have considerable influence even though the secondary outcomes seemed to be statistically significant.

Assmann and colleagues argue that statistical inspection of subgroups should not simply rely on P values for the subgroup comparison but on tests for statistical interaction between groups.[11] That is, tests that determine that a group of patients are significantly different from other patients in the trial. They suggest that "only if a statistical interaction test supports a subgroup effect should the results be influenced." The suggestion is not new and echoes that of Peto et al.[12] Although sensible, it is not failsafe, as is shown by the results of the second prospective randomised amlodipine survival evaluation study, which were reported recently.[1] This study, which included only patients with non-ischaemic disease, identified no benefits for amlodipine in the treatment group. Pooled results from both trials indicate no benefits from amlodipine for the patient population as a whole or for the patients with non-ischaemic heart failure. This is despite findings of $P<0.001$ for all cause mortality in the subgroup of patients with non-ischaemic disease and $P=0.004$ for the interaction term between cause of heart failure and treatment in the first study.[1]

Implications

Oxman and Guyatt developed a series of questions to help clinicians decide whether apparent differences in subgroup responses are real.[13] These are given in the box.

An individual patient faced with a serious condition may have only one opportunity to benefit from a potentially helpful treatment. Whatever the statistical results, the subgroup or secondary outcome results could provide the best available estimate of treatment effects for individual patients. Health policy decisions relate not just to the individual patient but to all patients in the future. These decisions require greater rigour because an incorrect decision will be hard to rectify. It may consign future patients to unnecessary treatment with associated risks (but no benefit) and use scarce healthcare resources futilely rather than

Are apparent differences in subgroup response real?

1 Is the magnitude of the difference clinically important?
2 Was the difference statistically significant?
3 Did the hypothesis precede rather than follow the analysis?
4 Was the subgroup analysis one of a small number of hypotheses tested?
5 Was the difference suggested by comparisons within rather than between studies?
6 Was the difference consistent across studies?
7 Is there indirect evidence that supports the hypothesised difference?

allocating them to interventions likely to achieve worthwhile improvements in health status.

Thus, for health policy purposes the list in the box should be prefixed by a question asking whether the primary outcome measure was statistically significant. If the answer to that question is yes, then it may be appropriate to consider the remaining questions. A purist view suggests that when the primary end point is not significant the results should be used only for generating hypotheses. Even when the primary outcome is statistically significant, attention should be directed at the way statistical power is spent in the trial, and consideration should be given to the likelihood that findings in subgroups or secondary end points represent chance rather than reliable findings.

This suggestion has substantial implications. Interim guidance from the National Institute for Clinical Excellence (NICE) to sponsors (box) includes various references to the identification and description of subgroups of patients who will benefit from treatments in a manner that may be considered cost effective.

Methodological arguments counsel against the use of subgroups of patients—particularly those not prospectively defined—and, worse, against using subgroups derived on the basis of observed results. The National Institute for Clinical Excellence's recommendations for considering the cost effectiveness of drugs and devices deviate substantially from purist rigour and may be regarded as ill conceived or even irresponsible.

A review of the experience of the analogous Australian Pharmaceutical Benefits Economics Subcommittee described problems in the economic analysis in two thirds of submissions to the scheme, and it found that two thirds of these problems concerned the interpretation of clinical data.[15] Purist rigour in licensing of pharmaceuticals is challenged by current practice in cost effectiveness analysis.[16] As health systems increasingly consider cost effectiveness analyses as part of the decision making process for the reimbursement of drugs and devices, it is important that research evidence is properly interpreted, otherwise inappropriate pharmaceuticals will be incorporated in clinical practice.

Competing interests: NF has received funding for research from various pharmaceutical and device companies, the Department of Health, the Medical Research Council and other medical charities.

1 Packer M, O'Connor CM, Ghali JK, Pressler ML, Carson PE, Belkin RN, et al. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. *N Engl J Med* 1996;335:1107-14.
2 Department of Health and Human Services. Food and Drug Administration. International conference on harmonisation; guidance on statistical principles for clinical trials. *Federal Register* 1998;63:49583-98. (Docket no 97D-0174.)
3 Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;1:43-6.
4 Perneger TV. What's wrong with Bonferroni adjustments? *BMJ* 1998;316:1236-8.
5 Fisher LD. Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing. *Control Clin Trials* 1999;20:16-39.
6 Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med* 1990;9:657-71.
7 Temple R, Pledger GW. The FDA's critique of the anturane reinfarction trial. *N Engl J Med* 1980;303:1488-92.
8 Moyé LA. End-point interpretation in clinical trials: the case for discipline. *Control Clin Trials* 1999;20:40-9.
9 Cleland JGF, McGowan J, Clark A, Freemantle N. The evidence for β blockers in heart failure. *BMJ* 1999;318:824-5.
10 Hjalmarson Å, Goldstein S, Fagerberg B, Wedel H, Waagstein F, Kjekshus J, et al. Effects of controlled-release metoprolol on total mortality, hospitalizations, and well-being in patients with heart failure: The metoprolol CR/XL randomised intervention trial in congestive heart failure (MERIT-HF). *JAMA* 2000;283:1295-302.

Guidance from NICE on identifying subgroups who may benefit[14]

Information should be provided in order that the clinical effectiveness of the technology can be evaluated—both qualitatively and quantitatively—in relation to those conditions for which it is indicated (both in general and for relevant subgroups)

The manufacturer or sponsor should include data supporting specific claims (for example, improved efficacy, safety, or diagnostic reliability). Data supporting claims in specific target groups of patients, in whom there may be particular advantages, should also be presented even if these are not specifically identified in the product literature

Manufacturers and sponsors should, as appropriate, provide an overall assessment of the health gain that has resulted, or will result, from the routine adoption of the new technology and in special patient subgroups

11 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064-9.
12 Peto R, Pike C, Armitage P, Breslow NE, Cox DR, Howard SV, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 1977;35:1-39.
13 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;116:78-84.
14 National Institute for Clinical Excellence. Interim guidance for manufacturers and sponsors. www.nice.org.uk/ (accessed 24 February 2001).
15 Hill SR, Mitchell AS, Henry DA. Problems with the interpretation of pharmacoeconomic analyses. *JAMA* 2000;283:2116-21.
16 Freemantle N, Maynard A. Something rotten in the state of clinical and economic analyses. *Health Econ* 1994;3:63-7.

(Accepted 7 December 2000)

Corrections and clarifications

Estimating cardiovascular risk for primary prevention: outstanding questions for primary care

In this article last year by John Robson and colleagues (2000;320:702-4), a couple of errors have just come to light. In the fourth paragraph the third sentence should read: "They [the Framingham equations—for predicting cardiovascular risk] are most accurate when using the ratio of concentrations of total cholesterol to high density lipoprotein cholesterol, and they correctly identify 70% [not 85%] of people who develop coronary heart disease, with a false positive rate of 18% [not 30%]."

This Week in the BMJ

It was bound to happen—a grammatical error in the title of one of our TWIBs (the summary paragraphs that appear after the contents page and which give readers a taste of that week's *BMJ*). And it didn't escape the notice of at least one reader. In our quest for a brief title (to fit our style mandate of titles being two lines (ideally), but maximum of three; this reserves space for more text), we wrote (24 March): "Women recover quicker [rather than more quickly] from anaesthesia than men but experience more side effects." Perhaps we could have worked harder at achieving our aim and still remained within grammatical bounds. We have now thought of: "Women are quicker to recover from anaesthesia than men but have more side effects."

Personal view

No one has written to us yet, however, to point out a grammatical error that appeared in "Doctors on the ropes" by Richard Hayward (31 March, p 805). We inadvertently dropped an unattached participle (also known as a dangling modifier) on to the page, and to make more of an impact we did this twice—in large blue type as well as in the final sentence. "By getting angry on behalf of our patients, our status as a profession is assured" should have been changed to: "By getting angry on behalf of our patients, we can assure the status of our profession."