Commentary

Debate: Subgroup analyses in clinical trials – fun to look at, but don't believe them!

Peter Sleight
John Radcliffe Hospital, Oxford, UK

Received: 4 July 2000
Accepted: 7 July 2000
Published: 31 July 2000

Curr Control Trials Cardiovasc Med 2000, **1**:25–27

(Print ISSN 1468-6708; Online 1468-6694)

Abstract

Analysis of subgroup results in a clinical trial is surprisingly unreliable, even in a large trial. This is the result of a combination of reduced statistical power, increased variance and the play of chance. Reliance on such analyses is likely to be more erroneous, and hence harmful, than application of the overall proportional (or relative) result in the whole trial to the estimate of absolute risk in that subgroup. Plausible explanations can usually be found for effects that are, in reality, simply due to the play of chance. When clinicians believe such subgroup analyses, there is a real danger of harm to the individual patient.

Keywords: astrology, clinical trials, overview, play of chance, randomization

Introduction

Large, properly randomized clinical trials have transformed the quality of the evidence on which we base modern therapy [1]. The aim of proper randomization is to eliminate systematic bias in the allocation of the patient to active versus placebo (or control) therapy. Systematic bias is probable in trials in which the physician in charge of the allocation has some insight into which treatment his or her patient will receive. It is particularly likely to occur with envelope randomization, in which the physician may open the first envelope, decides that he or she does not like the allocation and then opens the next envelope. This has happened quite recently, for example in the CAPP (Captopril Prevention Project) study [2]. Another similar source of bias is when randomization is by the day of the week, or by hospital A versus hospital B. In such cases the physician can again influence the allocation. Large effects can result from such seemingly small biases, even on the main results of the trial. Such biases are exaggerated in any subgroup analyses. For this reason many studies use computer allocation, with randomization only after all the necessary entry/baseline data have been collected at the randomization centre, for example by telephone (as in the ISIS [International Study of Infarct Survival] trials [3]) or by some form of fax system (as in the Heart Outcomes Prevention Evaluation [HOPE] study, which used computer-scanned fax entry [4]).

Physicians, particularly those in cardiovascular specialities, have embraced the concept of randomized trials with enthusiasm. However, they have not fully appreciated the extent to which the play of chance can produce erroneous results, even if strenuous efforts have been made to reduce all bias. After all, a P value of 0.05 means that there is a 1 in 20 chance that the result may be wrong. These are not extreme odds – many people are happy to back horses at this level.

MI = myocardial infarction; t-PA = tissue plasminogen activator.

The play of chance is even more likely to produce spurious results when we examine subgroups in a trial, because of the diminished power to detect real differences, the increase in the variance around the mean estimate, and the increasing statistical likelihood of a false finding when many subgroups are examined. If we divide a large multicentre international trial with a negative result into 40 subgroups (by country, age, sex, blood pressure, severity of disease, treatment, etc), then (at $P < 0.05$) we would expect a positive result in two subgroups.

Erroneous interpretation of subgroup analyses

The following are examples of erroneous interpretation of subgroup analyses that have caused harm to patients.

Restriction of thrombolytic therapy to anterior infarction

In the early trials of thrombolytic therapy in acute myocardial infarction (MI), treatment with streptokinase was clearly effective in reducing mortality in patients with anterior infarction, who were at higher risk. This at first led to the erroneous conclusion that thrombolytic therapy did not work in inferior infarction.

As more data accumulated it became obvious that lytic therapy was also effective in inferior infarction. Of course, as with any effective treatment, it is necessary to balance the benefit against the risk of treatment. In very low risk small infarctions this risk may outweigh the benefit, although for most MIs some form of reperfusion is best [5].

Restriction in use of β-blockade after myocardial infarction to only anterior infarction

In the same way, the early studies of β-blockade after MI showed benefit in patients with anterior MI, but only a nonsignificant trend toward benefit in patients with inferior MI (who had fewer events overall). It can readily be shown that, when the number of events available for analysis is increased by performing a meta-analysis of all of the studies [6], β-blocker therapy is beneficial in inferior infarction.

Astrology in the International Study of Infarct Survival trials

In retrospect, perhaps one of the most important results in the ISIS trials was the analysis of the results by astrological star sign. All of the patients had their date of birth entered as an important ‘identifier’. We were therefore able to divide our population into 12 subgroups by astrological star sign. Even in a highly positive trial such as ISIS-2 [3], in which the overall statistical benefit for aspirin over placebo was extreme ($P < 0.00001$), division into only 12 subgroups threw up two (Gemini and Libra) for which aspirin had a nonsignificantly adverse effect (9% ± 13%)

Of course most physicians (but not all!) laughed when they were presented with these results. However, when presented with other less ridiculous subgroup analyses they are likely to believe the results, and forget the example from astrology, particularly if the result can be justified by some pet theory.

When one divides a trial by a seemingly more legitimate grouping (eg by the individual countries in a multinational study), then it is highly probable that a negative or neutral result will be seen in one country. Indeed, this was a point of discussion during the 1 May 2000 US Food and Drug Administration hearings (Yusuf S, personal communication) on the results of the recent HOPE study [4], in which ramipril had no significant effect in the US participants. We have seen similar results in the ISIS trials, but did not report these because of the possibility of harm caused by misinterpretation of such statistical ‘flukes’ (and hence a failure to use a useful treatment in that country).

ISIS-2 was carried out in 16 countries. For the streptokinase randomization, two countries had nonsignificantly negative results, and a single (different) country was nonsignificantly negative for aspirin.

There is no plausible explanation for such findings except for the entirely expected operation of the statistical play of chance. Of course, another reasonable explanation for negative or curious results in a subgroup is that the statistical power to detect a result is reduced by either a low event rate (eg in a low-risk subgroup such as young hypertensive persons) or by a low number of subjects in a particular subgroup (eg old age or female sex).

It is very important to realize that lack of a statistically significant effect is not evidence of lack of a real effect. Unfortunately, this error is often made by physicians.

What is the best way to estimate benefit in a subgroup?

One way to estimate the benefit of a particular treatment in a subgroup is to prespecify that an analysis will be done in particular subgroups (eg high versus low risk, high versus moderate versus mild hypertension) and to prespecify what is expected. Such analyses carry more weight than retrospective analyses that, if positive, are used to support some hypothesis. Another way is to apply the overall proportional (relative) risk reduction by treatment X (obtained from the whole trial) and then to apply this to the absolute risk in the subgroup of interest. Although possibly counter intuitive, this is more reliable statistically than examining the actual result obtained on that subgroup in the trial in question [1,7]. The best estimate of risk in such a subgroup might be from large overviews of the condition, rather than from the trial itself, particularly if the subgroup is small.

In general it is unlikely that the results in a particular subgroup are qualitatively different from those of the main trial result, although they might well be quantitatively different in, say, young versus older individuals, or higher versus lower risk groups.

Undue emphasis on a particular subgroup

The GUSTO (Global Use of Strategies to Open Occluded Coronary Arteries) trial [8] compared two tissue plasminogen activator (t-PA)-based regimens with two streptokinase-based regimens for the thrombolytic treatment of MI. Before the trial results were known, the investigators had hoped that the combination arm with streptokinase and t-PA would be the best. In the event, accelerated t-PA was better and hence was emphasized. Indirect comparisons of various trials suggested that there was little difference between the various agents. In a later analysis of all the data from ISIS-3, Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto miocardico (GISSI)-2 and GUSTO, we concluded that there was no significant superiority of t-PA over streptokinase, despite its undoubtedly better lysis rate if 90 min is selected (by 3 h, the patency rates are similar).

I have also been concerned that t-PA was significantly superior to streptokinase only in North American patients, but that this was not so in the rest of the world (approximately 20 000 randomized in each group) [8]. This might have been caused by chance, but also might have been because of a North American tendency to greater withdrawal from streptokinase if hypotension occurred, which is less likely to occur in non-American centres that are more familiar with the hypotension that commonly occurs with streptokinase [9]. I also argued that the lower blood pressure with streptokinase might have protected against cerebral haemorrhage, which was significantly higher with t-PA, especially in the elderly [10]. Recent outcome data (non-randomized) from the Medicare database has suggested little benefit, but possible harm, from thrombolysis in elderly patients (>75 years old) in the USA [11]. The higher rate of cerebral haemorrhage in elderly patients with t-PA (the commonly preferred lytic agent in the US after the GUSTO trial) might be responsible for this lack of benefit. Certainly the Fibrinolytic Therapy Trialists' overview (largely streptokinase trials) [5] suggested a higher absolute benefit of thrombolysis in older patients. We have emphasized the importance of considering all the information available from randomized controlled trials, rather than selective emphasis on one subgroup in one trial [1,12].

Conclusion

Cardiologists have been at the forefront in carrying out large randomized clinical trials. They have based their practice on the evidence that has resulted from overviews of these trials. However, this enthusiasm for trial evidence may be harmful when subgroup analyses are carried out without a proper appreciation of the statistical pitfalls. Undue emphasis on a particular subgroup may result in inappropriate treatment.

References

1. Collins R, Peto R, Gray R, Parish S: **Large-scale randomized evidence: trials and overviews.** In: Oxford Textbook of Medicine, edn 3. Edited by Weatherall DJ, Ledingham JGG, Warrell DA. Oxford, UK: Oxford University Press, 1996:21–32.
2. Hansson L, Lindholm LH, Niskanen L, Lanke J, Hedner T, Niklason A, Luomanmaki K, Dahlof B, de Faire U, Morlin C, Karlberg BE, Wester PO, Bjorck JE: **Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial.** *Lancet* 1999, **353**:611–616.
3. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group: **Randomised trial of intravenous streptokinase, oral aspirin, both or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2.** *Lancet* 1988, **ii**:349–360.
4. The Heart Outcomes Prevention Evaluation Study Investigators: **Effects of an angiotensin-converting enzyme inhibitor, ramipril, on death from cardiovascular causes, myocardial infarction, and stroke in high-risk patients.** *N Engl J Med* 2000, **342**:145–153.
5. Fibrinolytic Therapy Trialists' (FTT) Collaborative Group: **Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of mortality and major morbidity results from all randomised trials of more than 1000 patients.** *Lancet* 1994, **343**:311–322.
6. Yusuf S, Peto R, Lewis J, Collins R, Sleight P: **Beta blockade during and after myocardial infarction: an overview of the randomised trials.** *Prog Cardiovasc Dis* 1985, **27**:335–371.
7. Effron B, Morris C: **Steins paradox in statistics.** *Sci Am* 1977, **236**: 119–127.
8. The GUSTO Investigators: **An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction.** *N Engl J Med* 1993, **329**:673–682.
9. Sleight P: **Streptokinase is still the agent of choice for most patients with myocardial infarction.** *Am J Ther* 1995, **2**:128–135.
10. ISIS-3 (Third International Study of Infarct Survival) Collaborative Group: **ISIS-3: a randomised comparison of streptokinase vs. tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs. aspirin alone among 41,299 cases of suspected acute myocardial infarction.** *Lancet* 1992, **339**:753–770.
11. Thiemann DR, Coresh J, Schulman SP, Gerstenblith G, Oetgen WJ, Powe NR: **Lack of benefit for intravenous thrombolysis in patients with myocardial infarction who are older than 75 years.** *Circulation* 2000, **101**:2239–2246.
12. Collins R, Peto R, Baigent C, Sleight P: **Aspirin, heparin and fibrinolytic therapy in suspected acute myocardial infarction.** *N Engl J Med* 1997, **336**:847–860.

Author's affiliation: University of Oxford, John Radcliffe Hospital, Oxford, UK

Correspondence: Professor Peter Sleight, MD, FRCS, FACC, Department of Cardiovascular Medicine, Level 5, John Radcliffe Hospital, Oxford OX3 9DU, UK. Tel: +44 (0)1865 760 564; Fax: +44 (0)1865 768 844; e-mail: peter.sleight@cardiov.ox.ac.uk