INAHTA Briefs

Issue 2002/96

Title **Subgroup Analyses in Randomized Controlled Trials: Quantifying the Risks of False-positives and False-negatives**

Agency **NCCHTA, National Coordinating Centre for Health Technology Assessment**
Mailpoint 728, Boldrewood, University of Southampton, Southampton SO16 7PX, United Kingdom
tel: +44 2380 595586, fax: +44 2380 595639

Reference Health Technol Assess 2001;5(33) Oct 2001 www.ncchta.org/execsumm/summ533.htm

Aim

To quantify the extent to which subgroup analyses may be misleading and to compare the relative merits and weaknesses of the two most common approaches to subgroup analysis: separate (subgroup-specific) analyses of treatment effect and formal statistical tests of interaction. To establish what factors affect the performance of the two approaches. To provide estimates of the increase in sample size required to detect differential subgroup effects and to provide recommendations on the analysis and interpretation of subgroup analyses.

Conclusions and results

With simulated simplest case data with no differential subgroup effects, the formal tests of interaction were significant in 5% of cases as expected, while subgroup-specific tests were less reliable and (incorrectly) identified effects in 7%–66% of cases depending on whether there was an overall treatment effect. The most common type of subgroup effect identified in this way was where treatment effect was significant in one subgroup only. When a differential subgroup effect was included, the performance of the formal interaction test was generally superior to that of the subgroup-specific analyses, with more differential effects correctly identified using interaction tests. The ability of formal interaction tests to (correctly) identify subgroup effects improved as the size of the interaction increased relative to the overall treatment effect. When the size of the interaction was twice the overall effect or greater, the interaction tests had at least the same power as the overall treatment effect, but was considerably reduced for smaller interactions more likely to occur in practice. The inflation factor required to increase the sample size to detect the interaction with the same power as the overall effect varied with the size of the interaction. For an interaction of the same magnitude as the overall effect, this inflation factor was four. This increased dramatically for more subtle interactions to 100 or greater for interactions smaller than 20% of the overall effect. Formal interaction tests were generally robust to alterations in the number and size of the treatment and subgroups and, for continuous data, the variance in the treatment groups; the only exception being a change in the variance in one of the subgroups. In contrast, the performance of the subgroup-specific tests was affected by almost all of these factors with only a change in the number of treatment groups having no impact at all. While it is generally recognized that subgroup analyses can produce spurious results, the extent of the problem is almost certainly underestimated. This is particularly true for subgroup-specific analyses. In addition, the increase in sample size required to identify differential subgroup effects may be substantial, and the previously used 'rule of four' may not always be sufficient, especially when interactions are relatively subtle.

Recommendations

Trials should ideally be powered with subgroup analyses in mind. Subgroup analyses should be restricted to those proposed before data collection, Any subgroups chosen after this time should be clearly identified. Subgroup-specific analyses are particularly unreliable and are affected by many factors. Subgroup analyses should be based on formal tests of interaction, but even these should be interpreted with caution. Results from subgroup analyses should not be over-interpreted, and unless there is strong supporting evidence, they are best viewed as a hypothesis-generation exercise. Any apparent lack of differential effect should be regarded with caution unless the study was specifically powered with interactions in mind.

Title **Subgroup Analyses in Randomized Controlled Trials: Quantifying the Risks of False-positives and False-negatives**

Agency **NCCHTA, National Coordinating Centre for Health Technology Assessment**
Mailpoint 728, Boldrewood, University of Southampton, Southampton SO16 7PX, United Kingdom
tel: +44 2380 595586, fax: +44 2380 595639

Reference Health Technol Assess 2001;5(33) Oct 2001 www.ncchta.org/execsumm/summ533.htm

Methods

Subgroup-specific and formal interaction tests were assessed, firstly, by simulating data with no differential subgroup effects and determining the extent to which the two approaches (incorrectly) identified such an effect. Secondly, data were simulated with a range of types and magnitudes of subgroup effect (sample size determined by the nominal power (50%–95%) for the overall treatment effect) and the extent to which the two approaches were able to (correctly) identify the subgroup effect determined. Initially, data were simulated to represent the 'simplest case' of two equal-sized treatment groups and two equal-sized subgroups. Additional simulations were used to explore the impact of various trial specifications.

Further research/reviews required

Implications of considering confidence intervals rather than p-values could be considered. The same approach could be applied to contexts other than randomized controlled trials, eg, observational studies and meta-analyses.

Written by S Brookes, Lecturer in Medical Statistics, Dept Social Medicine, Univ Bristol, UK