Methodology

Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives

ST Brookes[1]
E Whitley[1*]
TJ Peters[1]
PA Mulheran[2]
M Egger[1]
G Davey Smith[1]

[1] Department of Social Medicine, University of Bristol, UK
[2] Department of Physics, University of Reading, UK

* Corresponding author

The validity of subgroup analyses

Executive summary

Health Technology Assessment 2001; Vol. 5: No. 33

Health Technology Assessment
NHS R&D HTA Programme

Executive summary

Background

Subgroup analyses are common in randomised controlled trials (RCTs). There are many easily accessible guidelines on the selection and analysis of subgroups but the key messages do not seem to be universally accepted and inappropriate analyses continue to appear in the literature. This has potentially serious implications because erroneous identification of differential subgroup effects may lead to inappropriate provision or withholding of treatment.

Objectives

- To quantify the extent to which subgroup analyses may be misleading.
- 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.
- To provide recommendations on the analysis and interpretation of subgroup analyses.

Methods

The performances of subgroup-specific and formal interaction tests were assessed by simulating data with no differential subgroup effects and determining the extent to which the two approaches (incorrectly) identified such an effect, and simulating data with a differential subgroup effect and determining the extent to which the two approaches were able to (correctly) identify it.

Initially, data were simulated to represent the 'simplest case' of two equal-sized treatment groups and two equal-sized subgroups. Data were first simulated with no differential subgroup effect and then with a range of types and magnitudes of subgroup effect with the sample size determined by the nominal power (50–95%) for the overall treatment effect. Additional simulations were conducted to explore the individual impact of the sample size, the magnitude of the overall treatment effect, the size and number of treatment groups and subgroups and, in the case of continuous data, the variability of the data.

The simulated data covered the types of outcomes most commonly used in RCTs, namely continuous (Gaussian) variables, binary outcomes and survival times. All analyses were carried out using appropriate regression models, and subgroup effects were identified on the basis of statistical significance at the 5% level.

Results

While there was some variation for smaller sample sizes, the results for the three types of outcome were very similar for simulations with a total sample size of ≥ 200.

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 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 the treatment effect was seen to be significant in one subgroup only. When a simulated differential subgroup effect was included, the results were dependent on the nominal power of the simulated data and the type and magnitude of the subgroup effect. However, the performance of the formal interaction test was generally superior to that of the subgroup-specific analyses, with more differential effects correctly identified. In addition, the subgroup-specific analyses often suggested the wrong type of differential effect.

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. However, power was considerably reduced for smaller interactions, which are much more likely in practice. The inflation factor required to increase the sample size to enable detection of 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, the inflation factor was 4, and this increased dramatically to ≥ 100 for more subtle interactions of < 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, with 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.

Conclusions

While it is generally recognised that subgroup analyses can produce spurious results, the extent of the problem is almost certainly under-estimated. This is particularly true when subgroup-specific analyses are used. In addition, the increase in sample size required to identify differential subgroup effects may be substantial and the commonly used 'rule of four' may not always be sufficient, especially when interactions are relatively subtle, as is often the case.

Recommendations for subgroup analyses and their interpretation

- Subgroup analyses should, as far as possible, be restricted to those proposed before data collection. Any subgroups chosen after this time should be clearly identified.
- Trials should ideally be powered with subgroup analyses in mind. However, for modest interactions, this may not be feasible.
- Subgroup-specific analyses are particularly unreliable and are affected by many factors. Subgroup analyses should always be based on formal tests of interaction although even these should be interpreted with caution.
- The results from any subgroup analyses should not be over-interpreted. Unless there is strong supporting evidence, they are best viewed as a hypothesis-generation exercise. In particular, one should be wary of evidence suggesting that treatment is effective in one subgroup only.
- Any apparent lack of differential effect should be regarded with caution unless the study was specifically powered with interactions in mind.

Recommendations for research

- The implications of considering confidence intervals rather than *p*-values could be considered.
- The same approach as in this study could be applied to contexts other than RCTs, such as observational studies and meta-analyses.
- The scenarios used in this study could be examined more comprehensively using other statistical methods, incorporating clustering effects, considering other types of outcome variable and using other approaches, such as Bootstrapping or Bayesian methods.

Publication

Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess* 2001;**5**(33).

NHS R&D HTA Programme

The NHS R&D Health Technology Assessment (HTA) Programme was set up in 1993 to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS.

Initially, six HTA panels (pharmaceuticals, acute sector, primary and community care, diagnostics and imaging, population screening, methodology) helped to set the research priorities for the HTA Programme. However, during the past few years there have been a number of changes in and around NHS R&D, such as the establishment of the National Institute for Clinical Excellence (NICE) and the creation of three new research programmes: Service Delivery and Organisation (SDO); New and Emerging Applications of Technology (NEAT); and the Methodology Programme.

Although the National Coordinating Centre for Health Technology Assessment (NCCHTA) commissions research on behalf of the Methodology Programme, it is the Methodology Group that now considers and advises the Methodology Programme Director on the best research projects to pursue.

The research reported in this monograph was funded as project number 97/40/03.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The editors and publisher have tried to ensure the accuracy of this report but do not accept liability for damages or losses arising from material published in this report. They would like to thank the referees for their constructive comments on the draft document.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278