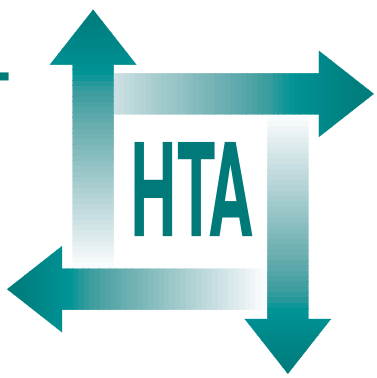
Health Technology Assessment 2001; Vol. 5: No. 33

Methodology

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

ST Brookes
E Whitley
TJ Peters
PA Mulheran
M Egger
G Davey Smith

Health Technology Assessment
NHS R&D HTA Programme



How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

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

Competing interests: none declared

Published September 2001

This report should be referenced as follows:

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).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE. Copies of the Executive Summaries are available from the NCCHTA website (see opposite).

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.

The views expressed in this publication are those of the authors and not necessarily those of the Methodology Programme, HTA Programme or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for any recommendations made by the authors.

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.

ISSN 1366-5278

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

M

Contents

List of abbreviations

CI	confidence interval
CONSORT	Consolidated Standards for Reporting of Trials
df	degrees of freedom*
MS	mean square*
MSE	mean square error*
RCT	randomised controlled trial
S1	subgroup 1
S2	subgroup 2
SS	sum of squares*
T1	treatment group 1
T2	treatment group 2

* Used only in tables

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.

Chapter 1
Introduction

The presentation of subgroup analyses in reports of randomised controlled trials (RCTs) is common[1] and it is important that researchers and clinicians are able to assess their validity and to interpret their results. Subgroups arise in many different settings, for example, centres in a multicentre trial or groups of patients defined by age, sex and baseline risk. Inappropriate analyses of such subgroups are widespread and many reports put too much emphasis on subgroup analyses that frequently lack statistical power.[1–5] Guidelines are available on the analysis and interpretation of subgroups, but the debate as to when and how subgroup analyses can be legitimately carried out continues.[6–9] One recent paper examined 50 consecutive RCT reports in four major medical journals (*British Medical Journal*, *Journal of the American Medical Association*, *Lancet* and *New England Journal of Medicine*) during a 3-month period in 1997. Over two-thirds of the reports presented subgroup findings. While only 43% of these used appropriate statistical tests for interaction, 60% claimed differential subgroup effects with the majority reporting it in the summary and/or conclusions.[2] Even this year, misleading reports of subgroup analyses can be found in leading journals.[10]

The analysis of an RCT typically begins with the investigation of differences in health outcomes between patients in two (or more) treatment groups, that is the overall treatment effect. It is then common for further analyses to be carried out to determine if the effects of treatment are different across particular groups of patients. There may be good biological reasons why such differences occur and it is certainly important that such subgroup analyses be considered if treatment is to be selected appropriately for future patients. For example, a moderately beneficial treatment effect overall may be masking a beneficial treatment effect in one subgroup of patients and a detrimental effect in another.[11] However, an inappropriate subgroup analysis may lead to an incorrect conclusion, for example, that treatment is not beneficial in a particular group, with the result that treatment is withheld from those who would benefit from it.[12]

The choice of which subgroups should be considered in this type of analysis is not straightforward and there are two extreme approaches that might be adopted.[13] The first approach is to work through all possible subgroup analyses on the basis that one or more might reveal some differential treatment effect. This 'data dredging' approach has serious implications in terms of multiple testing. Specifically, every statistical test carries the risk of a false-positive result (a statistically significant finding that is actually due to chance rather than to any inherent difference in the comparison groups) and as more tests are performed the probability of a false-positive finding increases. The second approach is to specify what subgroups might be of interest, generally based on findings from similar trials, at the design stage of the RCT and then to look only at these in the analysis. This approach reduces the problem of multiple testing and is often recommended as an ideal. However, it may be rather conservative and may rigidly test only pre-specified hypotheses that may lead to unexpected, yet clinically important, differences in treatment effects being missed.

A reasonable compromise is, perhaps, to specify a small number of key subgroup analyses in advance and to be much more cautious about conclusions drawn from any other subgroup analyses that are carried out, including being more conservative in terms of correcting for multiple testing. This approach is analogous to that commonly adopted for primary and secondary outcomes in RCTs.[14] Moreover, with the emphasis on the importance of pre-specifying subgroup analyses in the trial protocol, it is consistent with Consolidated Standards for Reporting of Trials (CONSORT) guidelines.[15] The issue of how to choose subgroups will not be addressed in this report, although, in line with the above philosophy, they should, in general, be based on biological factors and not derived data.[5] The issue of corrections for multiple testing will also not be covered and thus the results presented here relate to a single subgroup analysis without reference to any others being performed. While this may be a reasonable representation of reality for (a limited number of) pre-specified subgroups, it should be borne in mind that, particularly for other situations, the false-positive rates presented here may be considered optimistic. However, the patterns of findings would not be expected to change.

Having identified which subgroups are to be compared, the next question is how this comparison is to be made. One frequent approach is to conduct separate (stratified) analyses of the treatment effect in each subgroup. However, it is generally accepted that a more appropriate analysis is one in which the interaction between treatment and subgroup is formally tested[1,2,15,16] in a suitable regression analysis. Not only does the formal test of interaction involve just one additional test (regardless of the number of subgroups) while subgroup-specific analyses involve two or more, but also the interaction is the only approach that tests (and estimates) the differential effect directly.

Another related issue is that of false-negative results, that is failing to detect a true difference in effect. Power calculations for RCTs are generally based on detecting the overall treatment effect. Subgroup-specific testing requires the data to be split and these smaller datasets will have reduced power to detect a similar treatment effect. This reduction in power might lead to a number of erroneous conclusions. For example, an apparently significant treatment effect overall may vanish, or be apparent in only one subgroup in the secondary analysis.[17] Moreover, as confidence intervals (CIs) around estimates of effect will be wider in the reduced datasets, there is a reduced chance of concluding that there is a differential treatment effect across subgroups based on CIs. As discussed above, formal tests of interaction are superior to subgroup-specific tests, but the problem of false-negatives may remain since interaction tests may be relatively underpowered (although, due to changes in estimates of random variation this is not always straightforward). Again, it is unlikely that a trial will be powered with an interaction test in mind.

Ideally, the decision of whether and how to look for subgroup effects should be made on the basis of scientific advancement. However, in reality, the pressures on researchers to publish may also be a factor and the problems of publication bias with respect to 'negative' findings are well known.[18] The decision to conduct subgroup analyses may well be influenced by the result of the analysis of the overall treatment effect, with subgroup analyses being more common in RCTs showing no overall difference between treatments.[19,20]

There are existing guidelines available on the analysis and interpretation of subgroups,[2,21,22] which suggest that statistical tests of interaction should be used rather than inspection of subgroup-specific *p*-values. In addition, subgroup analyses should be confined to the primary outcome, and to a few predefined subgroups on the basis of biologically plausible hypotheses. Subgroup findings should essentially be considered exploratory in nature and should only affect the conclusions drawn from the trial in exceptional circumstances. However, in spite of the existence of these recommendations, inappropriate analysis, presentation and interpretation of differential subgroup effects continue to appear in the literature.[2,10]

Subgroup analyses are problematic both in terms of false-positive and false-negative results, and although these problems can be reduced by the use of formal tests of interaction rather than subgroup-specific tests they are by no means eliminated. It is, therefore, important to understand the extent of the problem.

Objectives of the study

The aim of this report was to provide more quantitative guidelines on the analysis and interpretation of subgroups based on formal statistical methods. Simulated data were analysed using standard methods in order to explore the impact of different strategies of analysis on false-positive and false-negative rates. There are many factors that might affect the outcome of subgroup analyses, including the magnitude of the overall treatment effect, the magnitude and type of subgroup effect, the variability in the data and the number and size of treatment groups and subgroups. Each of these was varied in a controlled manner and their relative impact assessed. The results of these simulations then formed the basis of recommendations for researchers carrying out or interpreting subgroup analyses.

The context of this study is the RCT. However, subgroup analyses are also common in observational studies and, while issues of confounding have been ignored, the results presented here are also pertinent in this context in general.

Chapter 2
Methods

The rationale for using simulations

To investigate the issues surrounding the reliability of subgroup analyses, real or simulated data could be considered or a theoretical approach could be used. While real data have a number of obvious advantages in terms of the extent to which the variations in scenarios are realistic, the disadvantage is that there is little control over the underlying distributions. Inevitably, more than one parameter varies at once in an unknown way in real data making it difficult, if not impossible, to draw general conclusions about their separate influences. Simulated data, on the other hand, afford the investigator complete control over both the underlying distributions and the nature of the alterations to the parameters. It is important to note that since (as described in detail below) the simulations involve repeated random sampling from the relevant distribution, there remains a stochastic element. In general, about 5% of simulated test statistics would thus be expected to be significant (by chance) at the 5% level, and, hence, interest was focused on any substantial variation from this expectation.

At the same time, simulations were not strictly necessary for all of the calculations presented here. However, using theory throughout would have been far from straightforward and, in making the procedure less transparent, such an approach would severely hamper interpretation for a general audience. A consistent approach was, therefore, used throughout the study in the form of simulations developed from the simplest scenario. Notwithstanding this, it is pointed out where certain findings are theoretically impossible.

Simulation strategy

Separate simulations were carried out for each of the three main types of outcome variable commonly encountered in RCTs: continuous (such as blood pressure or cholesterol levels), binary (for example, survival versus death at a given time of follow-up or relief of symptoms versus none) and survival times (for instance, time to death or remission). The exact nature of the simulated data and the analyses depended on the outcome in question, however, the basic strategy in each case was the same (see *Figure 1*).

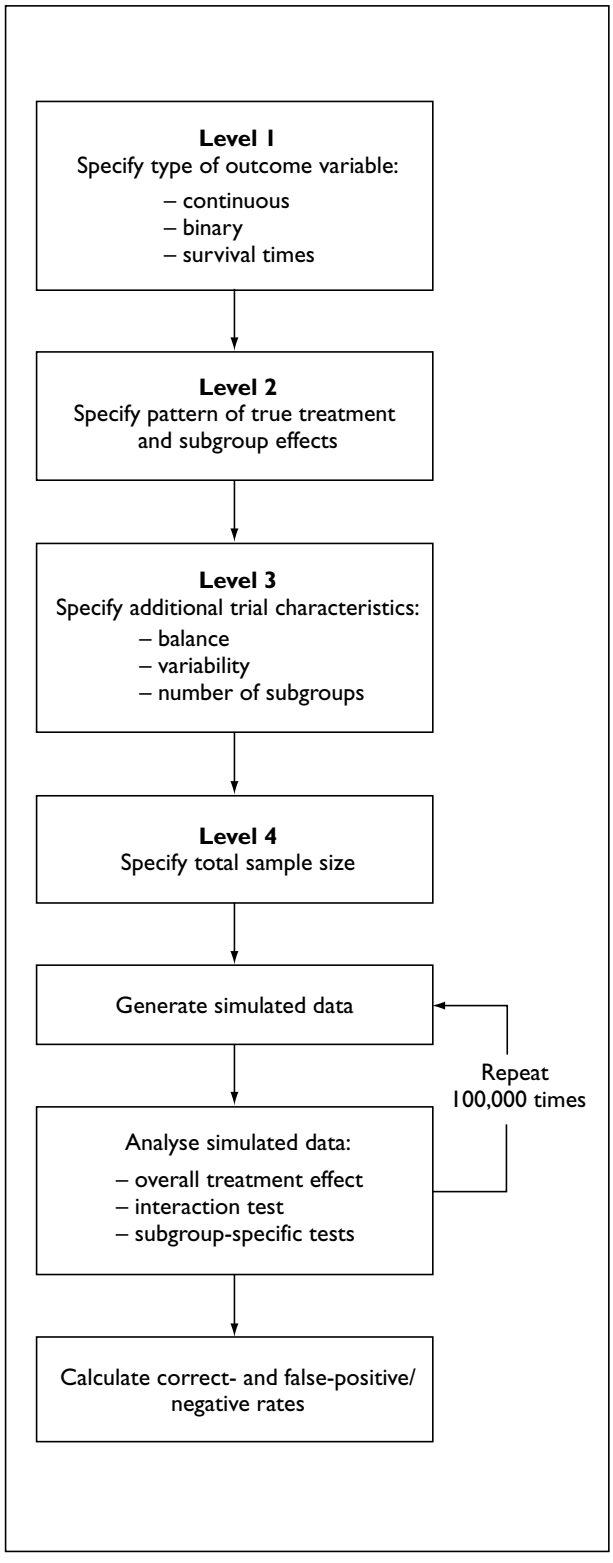


FIGURE 1 *Simulation strategy*

Firstly, streams of pseudo-random numbers between 0 and 1 were generated and then simulated data values were obtained by applying transformations to obtain the relevant distributional form for that type of outcome variable (level 1 in *Figure 1*). Analyses were then conducted using standard statistical methods. Separate details for each type of outcome variable are given later in this chapter.

Each round of simulations was based on a fixed set of parameters defining the treatment and subgroup effects in the (hypothetical) populations from which the simulated data were considered to be drawn. Repeated simulations were then carried out to give estimates of the correct- and false-positive/negative rates associated with different analytical approaches. In this way, it was possible to explore (a) the probability of wrongly concluding that a differential subgroup effect exists and (b) the reduced power associated with subgroup-specific and interaction tests.

Initial simulations concentrated on the situation where there were just two treatments groups and two subgroups. For these analyses, data were generated in four distinct categories: (A) treatment group 1 (T1), subgroup 1 (S1); (B) treatment group 2 (T2), S1; (C) T1, subgroup 2 (S2) and (D) T2, S2 (see *Table 1*). The full definition of the 'simplest' case was an arrangement as in *Table 1* but where, in addition, each category was of equal size and (for continuous data) had equal variance.

TABLE 1 *Data categories required for the 'simplest' case*

	T1	**T2**
S1	A	B
S2	C	D

There are several factors that may impact on the results of a subgroup analysis and each of these were explored in turn by varying specific parameters in successive simulations. Data in each of the four categories were generated to represent situations where an overall treatment or subgroup effect did or did not exist (level 2 in *Figure 1*). Further simulations explored the additional effects of imbalance in the categories in *Table 1*, the variability in the data and the number of subgroups (level 3 in *Figure 1*). Finally, all simulations were repeated for a number of different overall sample sizes (level 4 in *Figure 1*), that is, the total number of data values across all four categories in *Table 1*.

Each simulated dataset was analysed to examine the overall (main) treatment effect and to look for any evidence of subgroup effects. Subgroup analyses were then carried out in two ways using standard statistical techniques. The first of these analyses concentrated on formal tests of interaction between treatment and subgroup using regression techniques. The second followed the naïve approach often seen in the literature in which separate (stratified) analyses were carried out for each subgroup. All tests used a cut-off for statistical significance of 5%, and repeated simulations gave estimates of the associated correct- and false-positive/negative rates (see below for details).

Note that, for simplicity, the current study used a decision-based approach based on statistical significance to look for differential treatment effects across subgroups. In reality, it would also be appropriate to consider subgroup-specific parameter estimates and CIs. The reasons for focusing on statistical significance were that (a) this approach remains very common in the literature, especially in the context of subgroup and interaction tests, and (b) interpreting CIs is highly subjective and it would, therefore, be extremely difficult to derive a systematic method for determining what conclusions a researcher might reach using these quantities. Although the most common (5%) threshold for 'statistical significance' was used, there is no reason to expect that the general patterns would be different for other thresholds, and, hence, the results can be interpreted as portraying the influences on the *p*-value itself generally, rather than being restricted to the (arbitrary) 5% threshold. The findings from the simulations would, therefore, be relevant within the paradigm of presenting and interpreting actual *p*-values, rather than relying on the inappropriate use of arbitrary thresholds.[23]

A total of 100,000 simulations were conducted for each set of parameters to allow the correct- and false-positive/negative rates to be estimated with sufficient precision. For instance, this number of repeated simulations gives a margin of error of 0.14% around a false-positive rate of 5%. In view of the large number of simulations, it was important to use a programming language offering speed and efficiency. Hence, all simulations (including statistical analyses) were programmed in FORTRAN, rather than using a standard statistical package. However, a representative sample of all analyses was repeated using the Stata statistical package to confirm that the programming was correct.

Definition of treatment and subgroup effects

Each set of simulations was based on one of four possible scenarios:

- No overall treatment effect and no subgroup-specific effect
- Overall treatment effect but no differential subgroup effect
- No overall treatment effect but a differential subgroup effect
- Overall treatment effect and a differential subgroup effect.

In the first case, data in all treatment group/subgroup categories were based on the same underlying distribution. In the remaining cases, known treatment and subgroup effects were introduced. Details for each type of outcome are given later in this chapter.

The overall treatment effect was controlled by fixing the difference between (the average of) categories A and C versus (the average of) categories B and D in *Table 1*. In (a) and (c) above this difference was set to 0. In (b), differences that would be detectable at the different powers typically considered in RCTs, namely 80, 90 and 95%, were calculated and implemented in the data. In addition, consideration was also given to the scenario where a study is substantially underpowered, for example, at 50% power. If patterns in error rates were similar for different powers in the simplest case, only 80% power was considered for further scenarios.

In terms of the nature and magnitudes of the differential effects specified for (c) and (d), the simulations covered a wide range of feasible sizes of interaction effects relative to the overall effect and these are described in detail in chapter 3. In order not to confound different influences, the specification of differential subgroup effects in the simulations were considered only for the simplest case depicted in *Table 1* in which the four groups were of equal size and variability. Four types of differential effects were considered as follows:

(1) Treatment differences in both subgroups in the same direction but of different magnitudes (a 'quantitative' interaction), for example, the treatment difference for A versus B is greater than that for C versus D
(2) Treatment difference only occurs in one subgroup (viewed here as a special case of a quantitative interaction), for example, there is a treatment difference for A versus B but not for C versus D
(3) Treatment differences in different directions and of different magnitudes (a 'qualitative' interaction), for example, the treatment difference for A versus B is greater than that for D versus C
(4) The effect of treatment is exactly reversed in the two subgroups (viewed here as a special case of a qualitative interaction), for example, one treatment is beneficial compared with another treatment for A versus B, but equally harmful for C versus D.

Type 4 would be concomitant with no overall treatment effect, whereas in the other three types of differential effect there would also be an overall treatment effect.

Finally, the inflation factor for the sample size required to yield the same power for the interaction as provided for the overall effect by the original sample size was obtained for a suitable range of interaction effects.

Additional variations

Initially, the simulations concentrated on the simplest case described in *Table 1*, that is, two treatment groups and two subgroups of equal size with the same variance. In subsequent simulations, factors were varied **one at a time** in a controlled manner to examine their impact on the correct- and false-positive/negative rates arising from different analytical approaches.

Firstly, in general, it is unlikely that a subgroup analysis would lead to the same number of observations in each of the four categories in *Table 1* (the exception to this being when the randomisation is stratified with respect to the subgroups and randomisation is kept at a 1:1 ratio within each subgroup) and any imbalance in the categories would be likely to affect the significance level and power of subgroup analyses. Separate simulations were, therefore, used to examine the impact of various types and levels of imbalance by considering separate treatment group and subgroup ratios of 1:2, 1:3, 1:4 and 1:5. Secondly, the variance of the simulated data is also an issue, and its impact in the continuous case, where the scale parameter is specified independently of that representing location, was assessed by altering the relevant variances (for more details see later in this chapter). Thirdly, the initial set of simulations considered the situation where there were only two subgroups, which was extended to three, four or five subgroups.

In addition, all simulations were performed for a range of total sample sizes (specifically, the total number of simulated data points across all treatment groups and subgroups). The sample sizes were chosen to cover the range typically seen in RCTs. For the simplest case of two treatment groups, two subgroups, complete balance and equal variability amongst categories, the following 24 total sample sizes were considered: 20, 40, 60, 80, 100, 150, 200, 250, 300, 350, 400, 800, 1200, 1600, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10,000 and 50,000.

A reduced number of sample sizes covering the same range was used for additional simulations where changes were made to the number and balance of subgroups and the variability in the data. This was due, in part, to the similarity of results across the sample sizes, particularly the larger ones, and, in part, to the extensive running time of each set of simulations. The following 16 total sample sizes were considered for the case of continuous outcomes: 20, 40, 60, 80, 100, 150, 200, 250, 300, 350, 400, 1200, 2000, 5000, 10,000 and 50,000.

The initial simulations for binary data and survival times were conducted for the same 24 total sample sizes as in the continuous case. However, the running times of the FORTRAN programmes for these outcomes were even longer and any subsequent simulations were carried out for just the following nine total sample sizes: 60, 100, 200, 300, 400, 1200, 2000, 5000 and 10,000.

Data analyses

The statistical tests used for the analyses were determined by the type of outcome in question, and details are given later in this chapter. However, the basic philosophy in each case was the same (see *Figure 2*). The simulated data were analysed in the first instance to look for evidence of an overall significant treatment effect – that is, ignoring the subgroups. Following this, two analyses were performed to investigate subgroup effects: a formal statistical test of interaction between treatment and subgroup, and a naïve approach in which separate analyses were performed for each subgroup. The results of the tests for overall treatment effect and interaction between treatment and subgroup were considered simply in terms of statistical significance – more specifically, the percentages of simulations yielding correct- or false-negative/positive results. However, the subgroup-specific tests for the simplest case of two treatment groups and two subgroups resulted in four possible scenarios:

- Significant treatment effect apparent in both subgroups with treatment effect in the same direction in both
- Significant treatment effect apparent in both subgroups with treatment effects in opposite directions
- Significant treatment effect apparent in one subgroup only
- No significant treatment effect apparent in either subgroup.

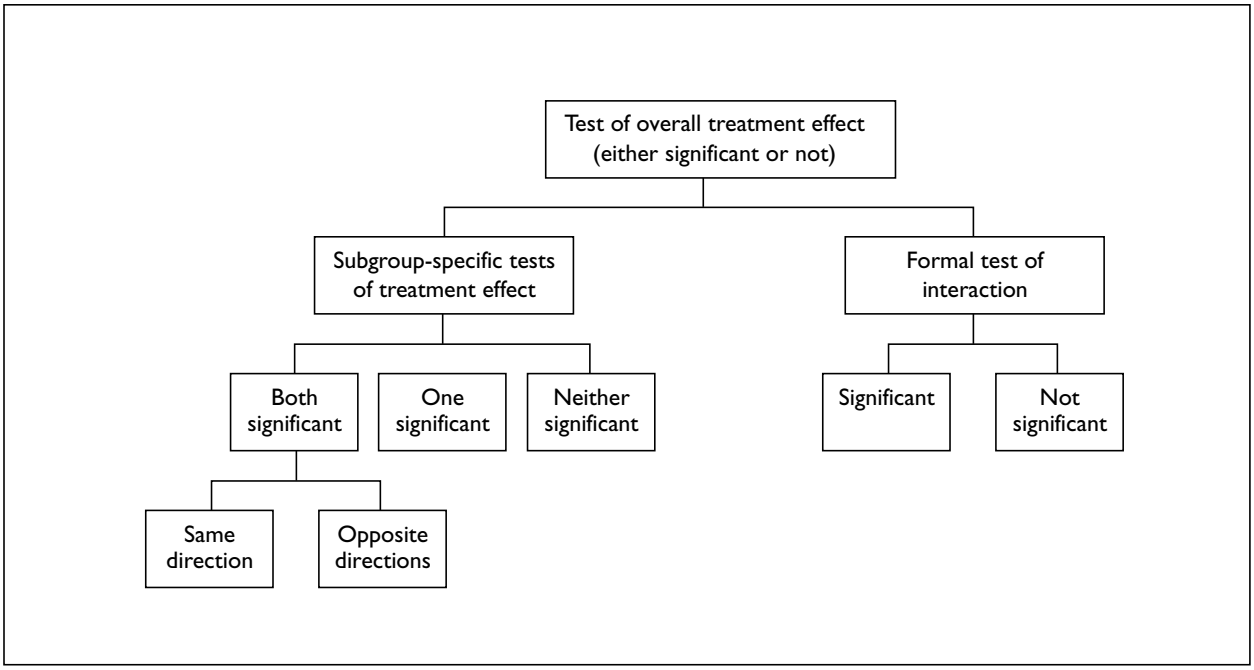


FIGURE 2 *Framework for the analyses*

These four scenarios were considered separately to allow the identification of common situations in which researchers might conclude that there are differential effects across subgroups. It is impossible to systematically identify every case in which this conclusion might be drawn without looking directly at estimates of effect and CIs. However, it was hypothesised that scenarios (b) and (c), in particular, might be interpreted in this way.

A similar approach was adopted in simulations involving more than two subgroups, resulting in the following scenarios of actual results:

(a) Significant treatment effect apparent in all subgroups with treatment effect in the same direction in all
(b) Significant treatment effect apparent in all subgroups with treatment effects not in the same direction
(c) Significant treatment effect apparent (in the same direction) in one or more subgroups but not all subgroups
(d) No significant treatment effect apparent in any subgroup.

Estimates of correct- and false-positive/negative rates for the tests of overall treatment effect and interaction were obtained by calculating the proportion of the 100,000 simulations with a significant result. Similar rates for the subgroup-specific tests were calculated for each of the four scenarios described above.

In order to mimic what might happen in reality as closely as possible, results of the subgroup analyses are presented separately according to whether or not the test for the overall treatment effect was significant or not (irrespective of whether or not the data were generated to have a treatment effect). The justification for this was that the results of the overall treatment comparison might influence whether and to what extent a researcher would conduct additional subgroup analyses, for example, a researcher finding no overall treatment effect might search more extensively for subgroup-specific effects in an attempt to find a 'positive' result for publication.

Details of analytical methods by type of outcome variable

Continuous data

Simulated data for the category defined by treatment group i and subgroup j followed a Gaussian distribution defined by the mean (μ_{ij}) and variance (σ_{ij}^2). Standardised Gaussian data ($\mu = 0$, $\sigma^2 = 1$) were obtained for each treatment group/subgroup category from pseudo-random numbers (between 0 and 1) using the Box-Muller transformation.[24] Treatment and subgroup effects were obtained simply by altering the means in each category. The effect of variability was explored by altering the variances for treatment groups and subgroups (see chapter 3 for details).

The continuous data were analysed using ordinary least squares regression techniques for the comparison of means. Significance tests for the overall treatment effect and subgroup-specific treatment effects were based on univariable regression (that is, one-way analysis of variance (ANOVA)) models. The formal tests of interaction between treatment and subgroup were obtained from multivariable regression (two-way ANOVA) models.

Binary data

Simulated data in category i, j followed a binomial distribution defined by the size of the category (n_{ij}) and probability of an event (π_{ij}). In each category n_{ij}, pseudo-random numbers between 0 and 1 were converted to the appropriate (0 or 1) data values according to whether or not they exceeded the specified π_{ij} for that category. Treatment and subgroup effects were specified by appropriately altering n_{ij} and π_{ij}. The variability of binary data is a function of n_{ij} and π_{ij} and thus it was not possible to look separately at the effect of variability when n_{ij} and π_{ij} were fixed.

Analysis of the binary data was based on logistic regression models using maximum likelihood methods. Significance tests of overall and subgroup-specific treatment effects were based on simple univariable models. Formal tests of interaction between treatment and subgroup were obtained by including an additional interaction term in a multivariable logistic regression model.

Survival data

Survival times were generated using the exponential distribution. This model is commonly used in practice and represents the situation where the hazard is constant over time. Simulated survival times (x) in treatment group/subgroup category i, j were generated from pseudo-random numbers, q, in the range 0 to 1 according to the equality:

$$x = \frac{-\ln(1 - q)}{\lambda_{ij}}$$

where λ_{ij} is a (positive) parameter representing the hazard in category i, j. Treatment and subgroup effects were then controlled by altering the mean survival time,

$$\frac{1}{\lambda_{ij}}$$

in each category. As in the binary case, there is no independent measure of variability in survival data.

Analyses of survival data were based on Cox proportional hazards models using maximum likelihood methods. Overall, interaction and subgroup-specific tests were obtained from univariable or multivariable regression models as appropriate.

Alternative analytical methods

The methods described above are, at least for the binary and survival cases, only one possible analytical approach that could be employed. Other methods that might have been considered include chi-squared (for differences in proportions), Mantel–Haenzel and log-rank tests. The rationale for the current choice was simply to cover the methods most commonly used by researchers.

Chapter 3

Results – continuous outcome data

Simplest case – two subgroups, complete balance and equal variability

Data simulated with no overall treatment or subgroup effects

Overall treatment effect found to be non-significant (correct-negative result)

The percentage of the 100,000 simulated datasets in which the overall treatment effect was correctly found to be non-significant ($p > 0.05$) fluctuated about 95% as expected. Within the range considered, the total sample size had no consistent effect on this percentage, which ranged between 94.9 and 95.3% for different sample sizes. The reason for the lack of sensitivity to sample size is that, while the standard errors for each test are greater for the smaller samples, sampling variation inherent in the simulations are correspondingly larger (that is, there is more fluctuation in the sample means derived from them), but still, by definition, 5% of tests will be significant. As discussed in the methods section, additional imprecision through repeated sampling with 100,000 simulations would be expected to be trivial.

Within this (approximately) 95% of datasets correctly finding no evidence of an overall treatment effect, the percentage of interaction tests with a statistically significant finding also fluctuated about 5% (4.8–5.2%) for different overall sample sizes (*Figure 3*). That is, 5% of the 95% with a non-significant overall treatment effect had a significant interaction.

Table 2 and *Figure 4* show the results of subgroup-specific treatment effect tests performed on the 95% of datasets correctly finding no overall treatment effect. Unsurprisingly, given the non-significant overall effect, none of the tests found significant treatment effects in the same direction within both of the subgroups. As expected, the majority of subgroup-specific tests found the treatment effect to be non-significant in both subgroups.

However, a small percentage of subgroup-specific tests found significant treatment effects within both subgroups but in opposite directions. The arrows in *Table 2* (and in the rest of this report) indicate the percentage tended to as sample size increased. The left arrow indicates that the tendency was towards the lower end of the range, the right arrow signifying the upper end of the range and where no arrow is presented, there was no consistent pattern across sample sizes. About 7% of subgroup-specific tests found a significant treatment effect in only one subgroup with a tendency towards the lower value (6.6%) as

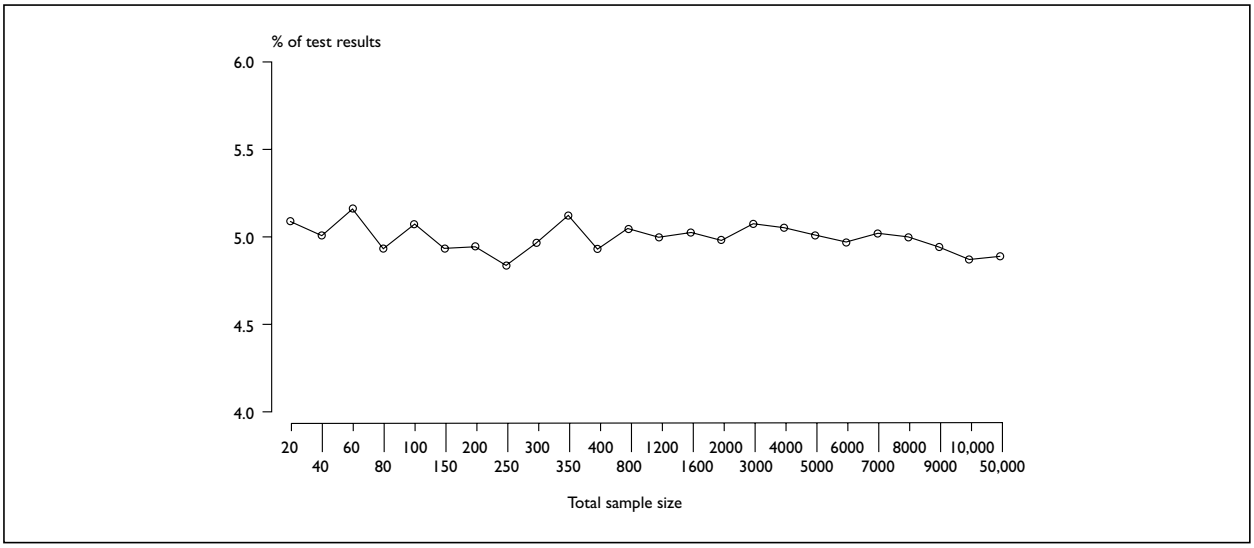


FIGURE 3 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of interaction test in datasets with a correct-negative overall result*

TABLE 2 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of subgroup-specific tests of treatment effect in datasets with a correct-negative overall result*

Subgroup treatment effects found	% with finding (range across sample sizes)
One subgroup significant	← 6.60–7.10
Both subgroups significant in opposite directions	0.12–0.15
Both subgroups significant in the same direction	Theoretically impossible
Neither subgroup significant	92.70–93.30 →

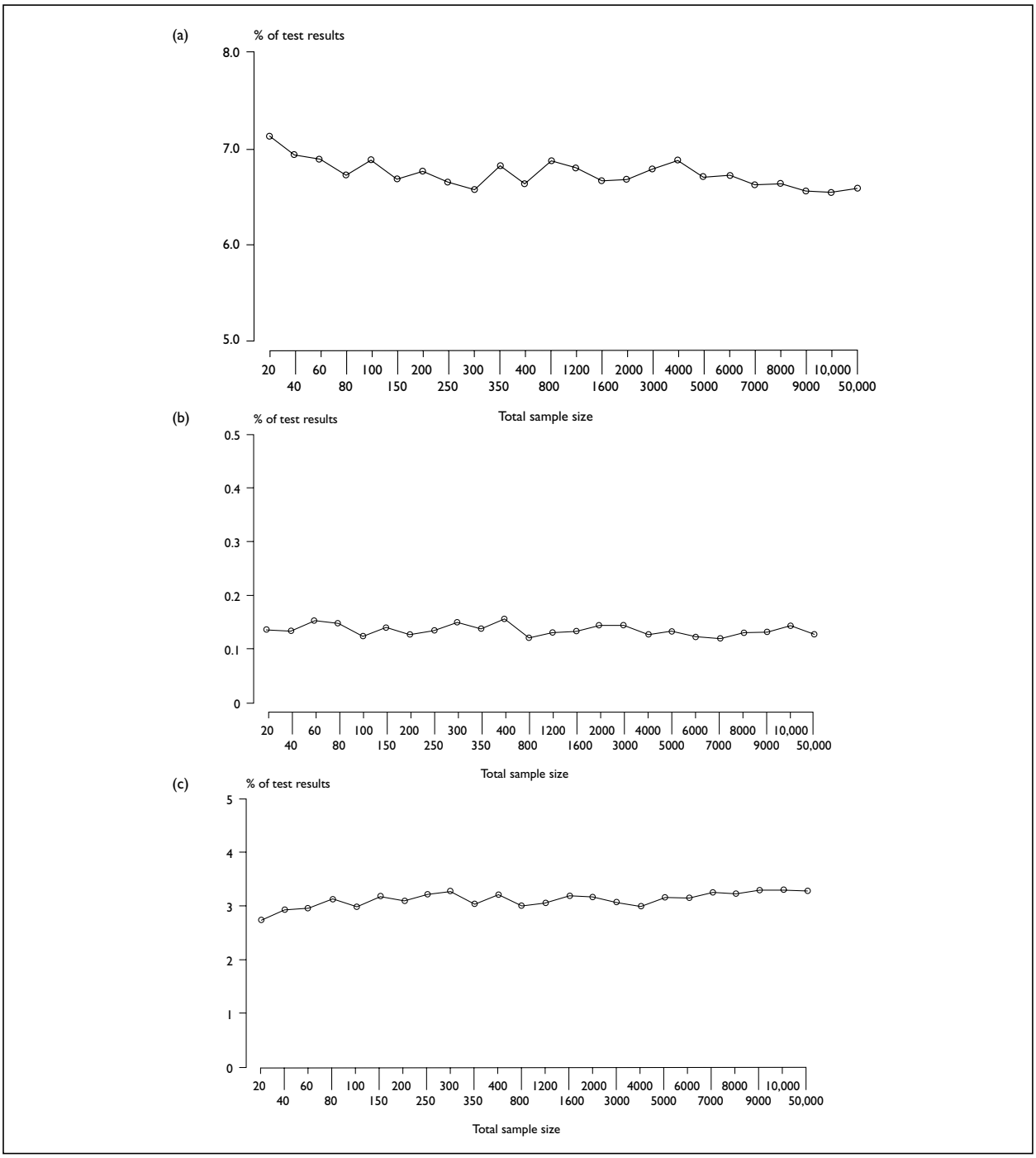


FIGURE 4 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of subgroup-specific tests of treatment effect in datasets with a correct-negative overall finding with (a) one subgroup significant, (b) both subgroups significant in opposite directions, (c) neither subgroup significant*

sample size increased, as indicated by the arrow in *Table 2*. Although this percentage is only marginally higher than the nominal 5% significance level, the observed differences in treatment effects across the subgroups are likely to be high given the smaller sample sizes in this situation. In this case, there may be a much higher potential for researchers to reach an incorrect conclusion of differential treatment effects across the subgroups. Moreover, such apparently large differential effects could be especially harmful when translated into practice.

Overall treatment effect found to be significant (false-positive result, type I error)

As expected, the percentage of the 100,000 simulated datasets which (incorrectly) found the overall treatment effect to be significant at the 5% level (that is, a type I error) fluctuated about 5% (4.7–5.1%) for different sample sizes as shown in *Figure 5*.

Among this 5% of datasets with an incorrectly significant overall treatment effect, the percentage of significant interaction tests also fluctuated about 5% (4.4–5.8%) for different sample sizes.

Table 3 and *Figure 6* show the results of the subgroup-specific tests performed on the 5% of datasets with an incorrect significant overall finding. As before, a large proportion of subgroup-specific tests found the treatment effect to be non-significant within both subgroups. However, as expected, this percentage was lower than in the previous section because a significant overall treatment effect had been observed. Correspondingly, there was also a slight increase in the percentage of subgroup-specific tests that found both subgroups to have significant treatment effects in the same direction. There were no instances of a significant treatment effect in both subgroups but in opposite directions. In contrast to the situation where the overall treatment effect was non-significant, a large percentage of the analyses (between 55 and 64%) found a significant treatment effect in one subgroup only. This is potentially a very misleading finding because it might be interpreted as showing treatment to be beneficial in one group and not in the other. Also in contrast to the previous situation, there was some evidence of a pattern across the total sample size with the percentage significant within one subgroup decreasing for sample sizes below about 80. This is the point

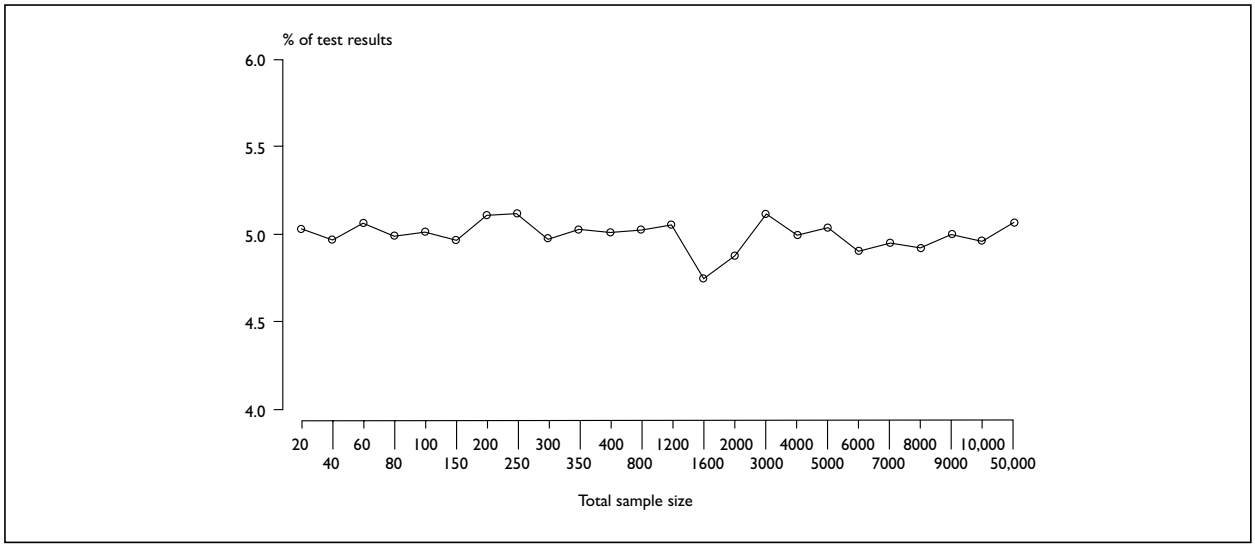


FIGURE 5 *Simplest case: data simulated with no overall treatment or subgroup effects. Percentage of primary analyses finding a significant overall treatment effect (false-positive result, type I error)*

TABLE 3 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of subgroup-specific tests of treatment effect in datasets with a false-positive overall result*

Subgroup treatment effects found	% with finding (range across sample sizes)
One subgroup significant	54.60–64.10 →
Both subgroups significant in opposite directions	0.00 for all
Both subgroups significant in the same direction	2.10–2.90
Neither subgroup significant	← 33.30–42.90

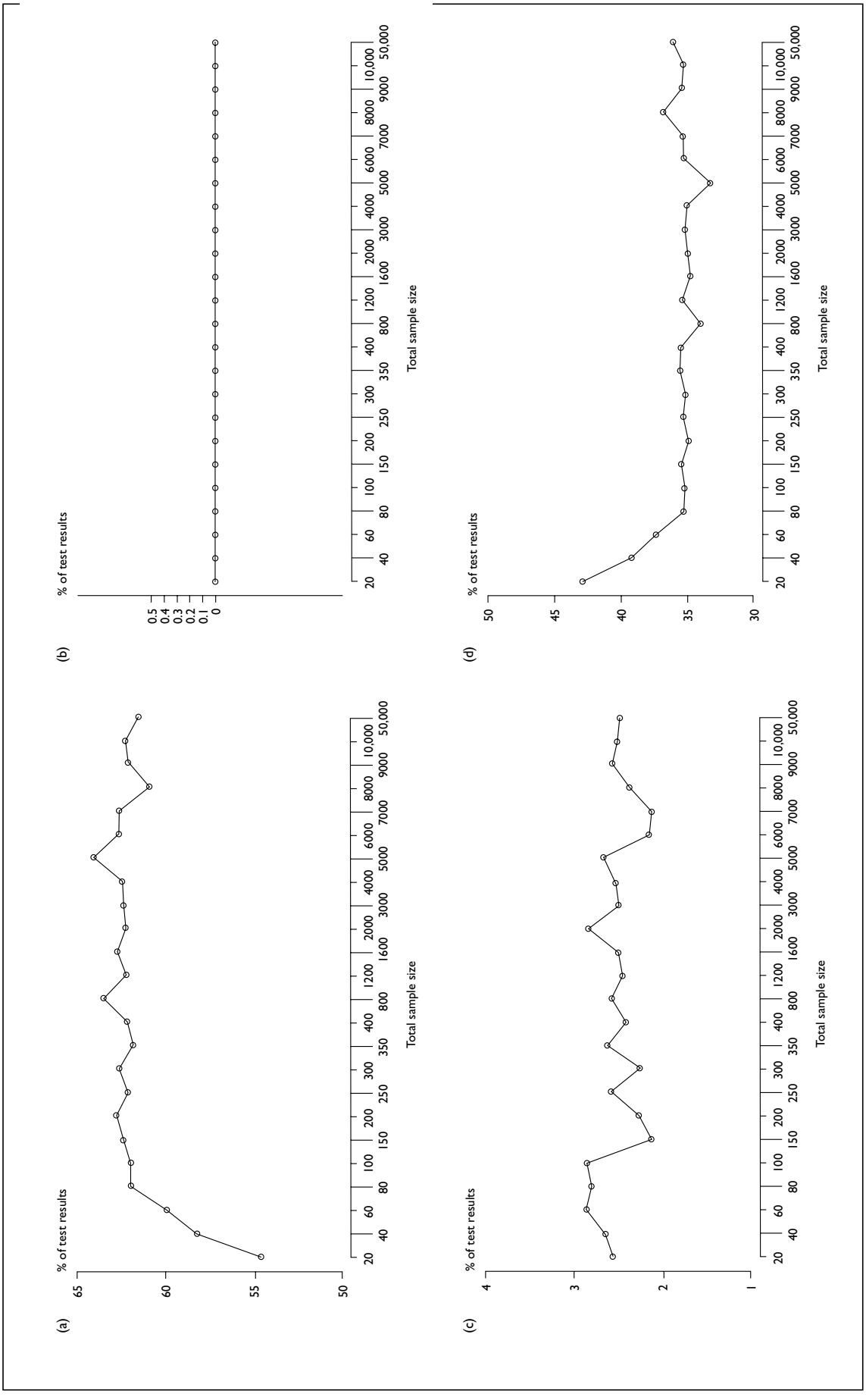


FIGURE 6 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of subgroup-specific tests of treatment effect in datasets with a false-positive overall finding (type I error) with (a) one subgroup significant, (b) both subgroups significant in opposite directions, (c) both subgroups significant in the same direction, (d) neither subgroup significant*

below which the degrees of freedom for the (within-subgroup) treatment comparison would be expected to result in more conservative tests.

Data simulated with an overall treatment effect but no differential subgroup effects

The overall treatment effect differences detectable with 50, 80, 90 and 95% power for a two-sided 5% significance level were calculated for each sample size (see appendix 1).

Overall treatment effect found to be significant (correct-positive result)

The percentage of primary analyses that correctly found the overall treatment effect to be significant fluctuated about the nominal power for all sample sizes. For each of the respective (approximately) 50, 80, 90 or 95% of datasets that found a significant overall treatment effect, the percentage of formal tests of interaction that were significant fluctuated about 5% (4.8–5.2%; *Table 4*).

Table 4 and *Figure 7* show the results of the subgroup-specific tests of the treatment effect within the datasets with a correct-positive overall finding. No simulations for any sample size found both subgroups to have a significant treatment effect but in opposite directions. The percentages of the three remaining combinations of subgroup-specific test results varied for smaller sample sizes, but all became reasonably stable at about 200 observations in total. Specifically, the percentage where a significant result in the same direction was observed in both subgroups increased with increasing power (16% for 50% power, 33% for 80% power, 44% for 90% power and 55% for 95% power). Correspondingly, the percentage for which only one or neither subgroup-specific test was significant reduced as power increased. One subgroup was significant in about 66% of cases for 50% power and this decreased to 57% for 80% power, 49% for 90% power and 41% for 95% power.

Overall treatment effect found to be non-significant (false-negative result, type II error)

The percentage of primary analyses that failed to detect a significant overall treatment effect were consistent with the nominal powers, and, in each case, the false-negative rate did not vary systematically with sample size. For each nominal power, within the datasets that (incorrectly) found a non-significant overall treatment effect, the percentage of formal tests of interaction that were significant fluctuated about 5% (4.4–5.7%; *Table 5*).

Table 5 and *Figure 8* show the results of the subgroup-specific tests of treatment effect within the datasets with a false-negative overall finding. A negligible number of simulations found both subgroups to have a significant treatment effect in opposite directions. As would be expected, the percentage where only one subgroup was significant increased with nominal power. However, surprisingly, given the non-significant overall effect, this percentage was rather high (20–26%). The percentage where neither subgroup was significant decreased correspondingly with power, although not appreciably so (*Table 5*).

Given the similarity of the results across the different nominal powers, only those for 80% are presented in the rest of this report for clarity.

Differential subgroup effects

Specification of differential effects

As described briefly in the methods, the general strategy for generating differential effects was to stipulate the interaction to be detected (θ) relative to the overall (target) effect (δ), in a ratio denoted by ψ ($= \theta/\delta$). For example, this ratio would take the value 1 for an interaction that is the same magnitude as the overall effect and 0.67 where the interaction is two-thirds as large as the overall effect. One advantage of specifying the differential effects in this way

TABLE 4 *Simplest case: data simulated with an overall treatment effect but no differential subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a (correct) significant overall treatment effect*

Nominal power	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
50%	4.80–5.20	60.10–65.50 →	0.00 for all	12.60–16.20 →	← 18.70–27.30
80%	4.80–5.20	56.70–57.50	0.00 for all	25.50–32.80 →	← 10.20–17.50
90%	4.90–5.20	← 49.00–52.10	0.00 for all	35.60–44.20 →	← 6.50–12.30
95%	4.80–5.10	← 40.80–46.50	0.00 for all	44.90–55.10 →	← 4.10–8.50

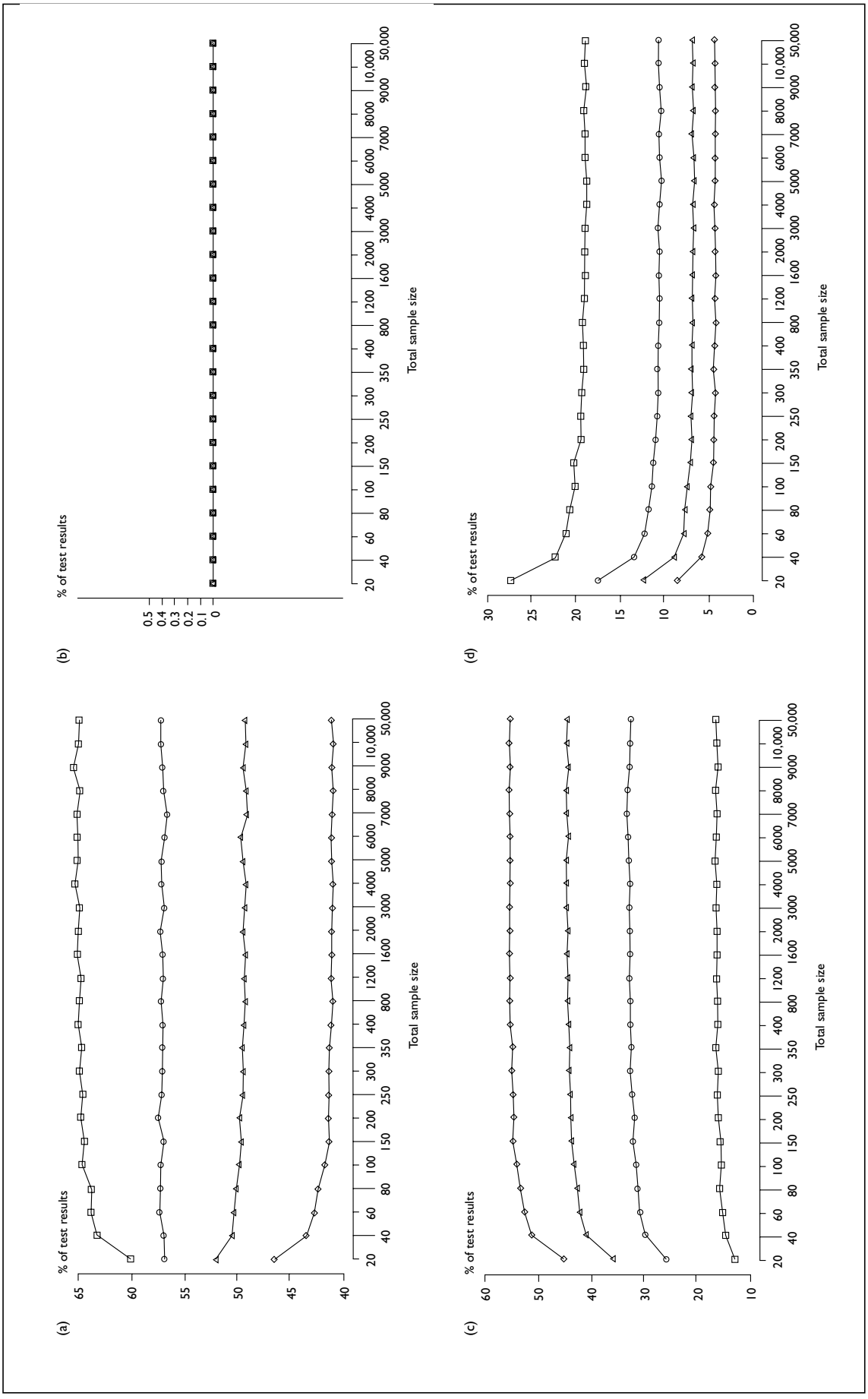


FIGURE 7 *Simplest case: data simulated with an overall treatment effect but no differential subgroup effects. Results of subgroup-specific tests of treatment effect within datasets with a correct-positive overall finding with (a) one subgroup significant, (b) both subgroups significant in opposite directions, (c) both subgroups significant in the same direction, (d) neither subgroup significant.* □, *50% power;* ○, *80% power;* △, *90% power;* ◇, *95% power*

TABLE 5 *Simplest case: data simulated with an overall treatment effect but no differential subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a (incorrect) non-significant treatment effect (type II error)*

Nominal power	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
50%	4.70–5.20	15.70–16.30	0.03–0.06	Theoretically impossible	83.70–84.20
80%	4.60–5.20	20.20–21.40	0.005–0.05	Theoretically impossible	78.50–79.80
90%	4.70–5.40	22.10–24.10	0.00–0.20	Theoretically impossible	75.90–77.80
95%	4.40–5.70	23.30–26.00	0.00–0.06	Theoretically impossible	74.00–76.70

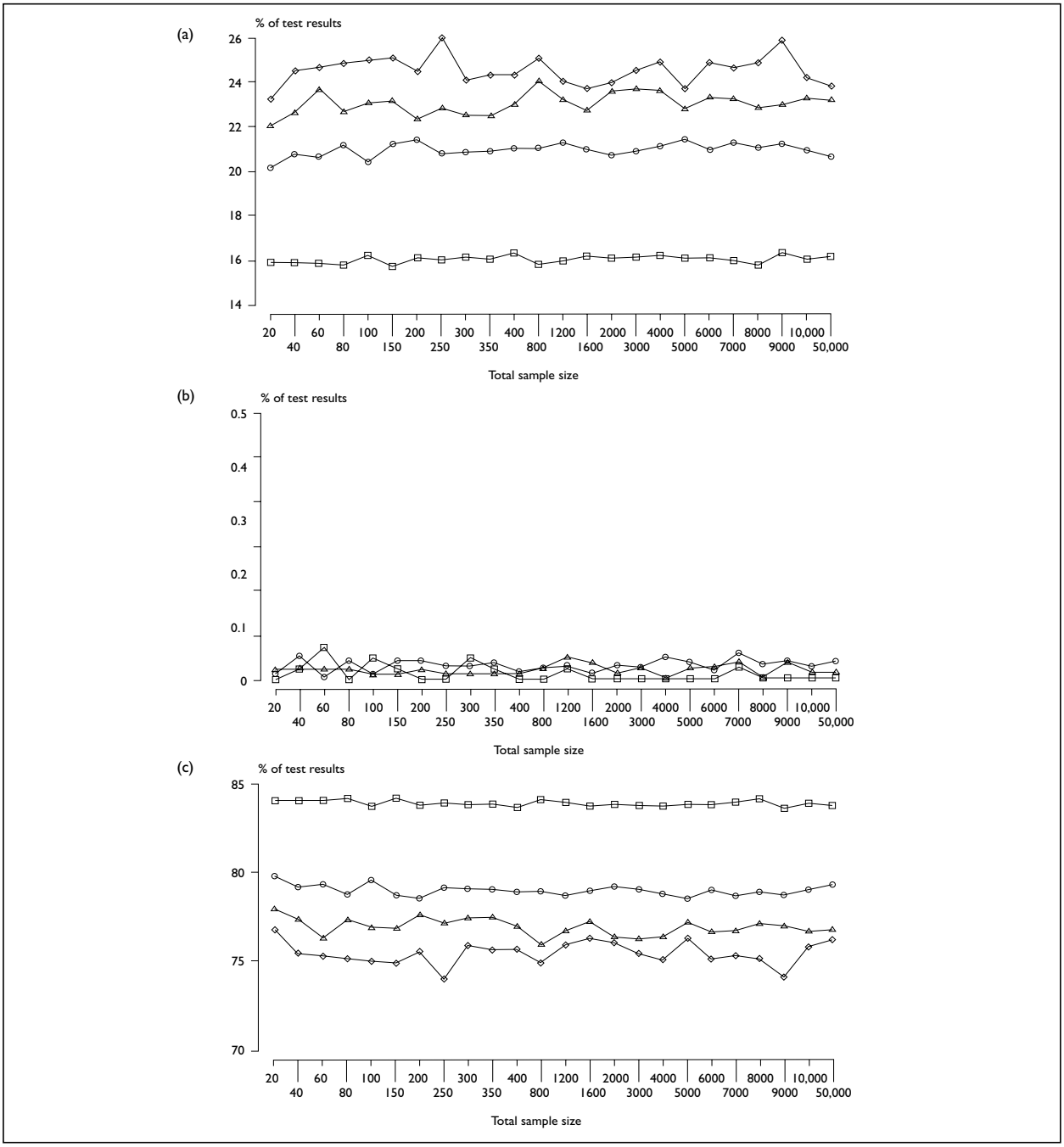


FIGURE 8 *Simplest case: data simulated with an overall treatment effect but no differential subgroup effects. Results of subgroup-specific tests of treatment effect within datasets with a false-negative overall result (type II error) with (a) one subgroup significant, (b) both subgroups significant in opposite directions, (c) neither subgroup significant. □, 50% power; ○, 80% power; △, 90% power; ◇, 95% power*

is that the ratio ψ remains constant over different sample sizes for the same nominal power. This follows from the manner in which the simulations were constructed in which the true treatment effect difference was fixed just by the nominal power and hence reduced as the sample size increased. The range of ratios used for the simulations covered all the possible types of differential effects defined in chapter 2:

(1) Treatment effect differences in the same direction but of different magnitudes
(2) Treatment effect difference in one subgroup only
(3) Treatment effect differences in different directions and of different magnitudes
(4) Treatment effect exactly reversed in the two subgroups.

Examples of the calculation of the ratio ψ for general quantitative and qualitative differential effects are as follows. For a quantitative interaction (type 1) with (standardised) treatment effects in the two subgroups of 0.1 and 0.2 then the overall effect (δ) would be 0.15 and the interaction (θ) would be 0.1 given equal-sized groups, hence $\psi = \theta/\delta = 0.1/0.15 = 0.67$. For a qualitative interaction (type 3) with subgroup-specific effects of –0.05 and 0.1, $\psi = 0.15/0.025 = 6$.

In almost all situations, the specification of ψ was made in the following way. Firstly, the target overall treatment effect was fixed by the nominal power specified and the simulations were then performed with this difference incorporated into the means. In other words, the true difference was effectively fixed to be the same as the target difference δ. The interaction (θ) was then specified relative to δ in terms of ψ. The only exception to this was for interactions of type 4, for which the overall (true) treatment effect difference incorporated into the simulations was, by definition, equal to zero. In this case, setting the true overall effect to be δ would mean that ψ is not defined. In order to retain this special case in the general presentation of results (however unlikely it is to occur in practice), δ was set here to be equal to the (overall) target difference that would be detectable at the nominal power with the sample size involved. By doing this, ψ was, again, constant over different sample sizes for the same nominal (overall) power, and was, therefore, consistent with the other three types of interaction.

The possible values of ψ for the four types of interaction defined above are as follows. For quantitative interactions, ψ is constrained to be 2 or below, being less than 2 for the general case of type 1 and always equalling 2 for the special case of type 2. General qualitative interactions (type 3) have values of ψ greater than 2. Once again, the special case of type 4 interactions (equal and opposite subgroup-specific effects) is an exception: given the method of specifying these that were adopted, ψ for this type of interaction can take any value.

It is reiterated that the simulations underlying the results that follow were all performed for the situation where the treatment groups and subgroups were all of equal size.

Performance of formal interaction tests

Figure 9 shows the power of the interaction test for various values of ψ for trials powered for the

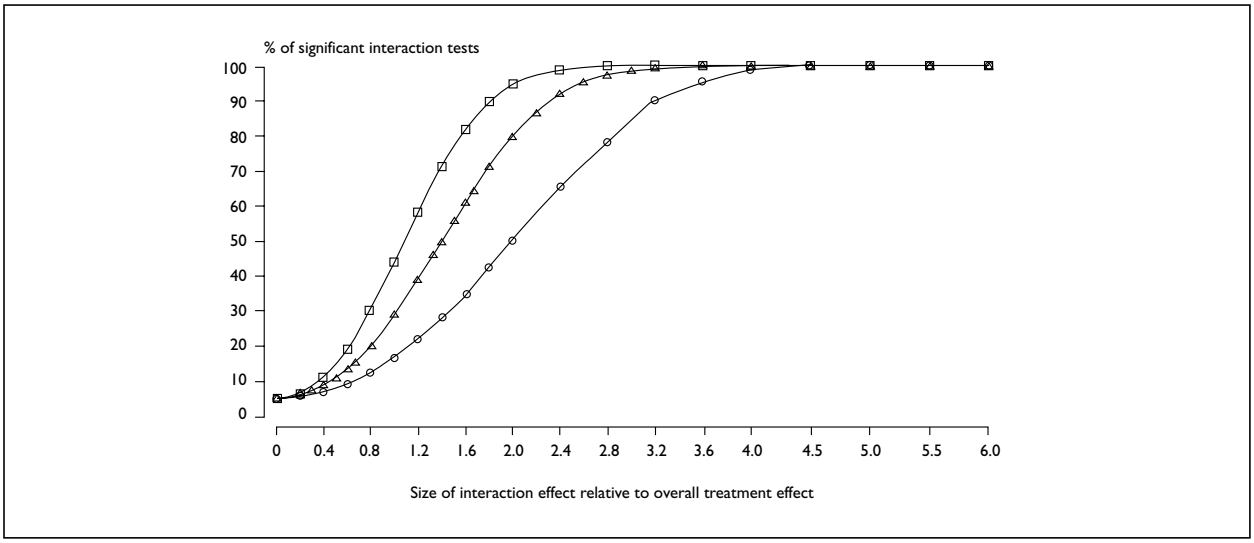


FIGURE 9 *Simplest case: data simulated with differential subgroup effects. Results of interaction tests for trials powered on the overall treatment effect. ○, 50% power; △, 80% power; □, 95% power*

overall treatment effect at 50, 80 and 95% levels. For example, if the interaction effect is the same magnitude as the overall effect ($\psi = 1$) then, in a trial with 80% power for the overall treatment effect, the test for interaction will only have about 29% power. On the other hand, if the interaction effect is four times as large as the overall effect then the interaction test will have power approaching 100%. As stated above, it is emphasised that this method of specifying the overall and differential effects (δ and θ, respectively) means that the results in *Figure 9* are independent of sample size. Moreover, considering the results underlying this figure separately according to whether or not the (simulated) overall test result was significant had little effect either on the percentages of significant interaction tests or the lack of dependence on sample size.

Performance of subgroup-specific tests

There was some variation with sample size in the performance of the subgroup-specific tests. For an example total sample size of 1200, *Figures 10* and *11* present the results from subgroup-specific tests across the range of values for ψ where the overall treatment effect is not equal to zero (differential effect types 1, 2 and 3). In contrast to *Figure 9*, the limitations of this analytical approach mean that

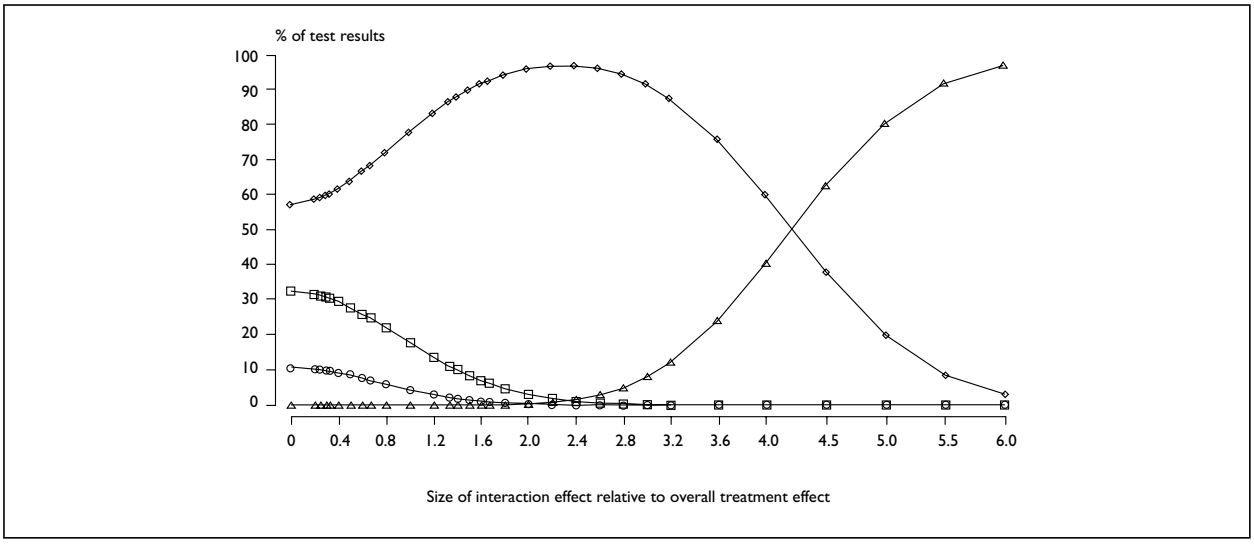


FIGURE 10 *Simplest case: data simulated with differential subgroup effects. Results of subgroup-specific tests of treatment effect in datasets with a (correct) significant overall treatment effect. □, Both significant in the same direction; △, both significant in opposite directions; ◇, only one significant; ○, neither significant*

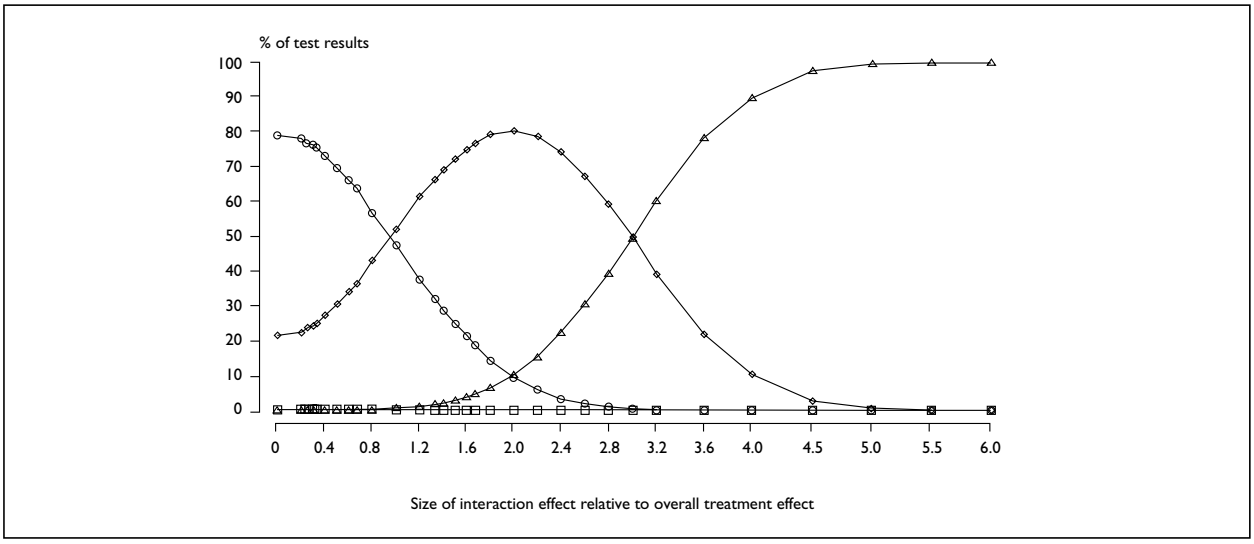


FIGURE 11 *Simplest case: data simulated with differential subgroup effects. Results of subgroup-specific tests of treatment effect in datasets with an (incorrect) non-significant overall treatment effect. □, Both significant in the same direction; △, both significant in opposite directions; ◇, only one significant; ○, neither significant*

Figures 10 and *11* are for illustrative purposes rather than practical use. Each figure depicts the percentage of test results in the four possible combinations of subgroup-specific results, that is, both significant in the same or opposite directions, only one significant or neither significant. *Figure 10* relates to the case where the observed overall treatment effect was (correctly) identified as statistically significant and *Figure 11* relates to false-negative overall results. As would be expected, the relative likelihoods of the various results vary considerably across the range of ψ; more importantly, apart from exceptional circumstances, such as when ψ = 2 for quantitative interactions (that is, not type 4 interactions) or is extremely large, the subgroup-specific tests, again, have a high chance of leading to inappropriate conclusions.

Generally, the results for subgroup-specific tests were unaffected by sample size. For qualitative interactions where the interaction effect was more than twice the overall effect, however, both tests being significant in opposite directions was observed more frequently as sample size increased, and, correspondingly, just one significant subgroup-specific test was a less frequent observation. As would be expected, the other two eventualities (neither significant or both significant in the same direction) occur very rarely in these circumstances.

Equivalent data were simulated for type 4 differential effects and these produced very similar patterns. As would be expected, an exception was the case when the overall effect was (correctly) non-significant and the chance of neither subgroup-specific test being significant was increased compared with *Figure 11*. In this case, the chance of only one subgroup-specific test being significant was correspondingly lower (below 50%).

Inflation factors for interaction tests

Figure 12 presents the factor by which the sample size would have to be inflated in order for the interaction test to have the same power as that provided by the original sample size for the overall treatment effect. These inflation factors are given for a range of magnitudes of the interaction relative to the overall effect (ψ) up to the value 2, by which point the inflation factor has effectively reached unity. Given the specification for deriving *Figure 12* (in particular, that the inflation factor applies multiplicatively to the original sample size in order to yield the nominal power), the inflation factors presented are independent of the original sample size, the nominal power it provided and the (absolute) magnitudes of the interaction and overall treatment effects.

One limitation of the presentation in *Figure 12* is that to accommodate the very large inflation factors required for interactions that are more subtle than, say, half the size of the overall effect, the scale is difficult to read for values of ψ between 0.5 and 2.0. A clearer version for practical use employing the log scale is, therefore, given in chapter 6. Examples of the potential use of this figure are that for interactions of the same magnitude as the overall effect, sample size should be inflated approximately four-fold, and for inter-

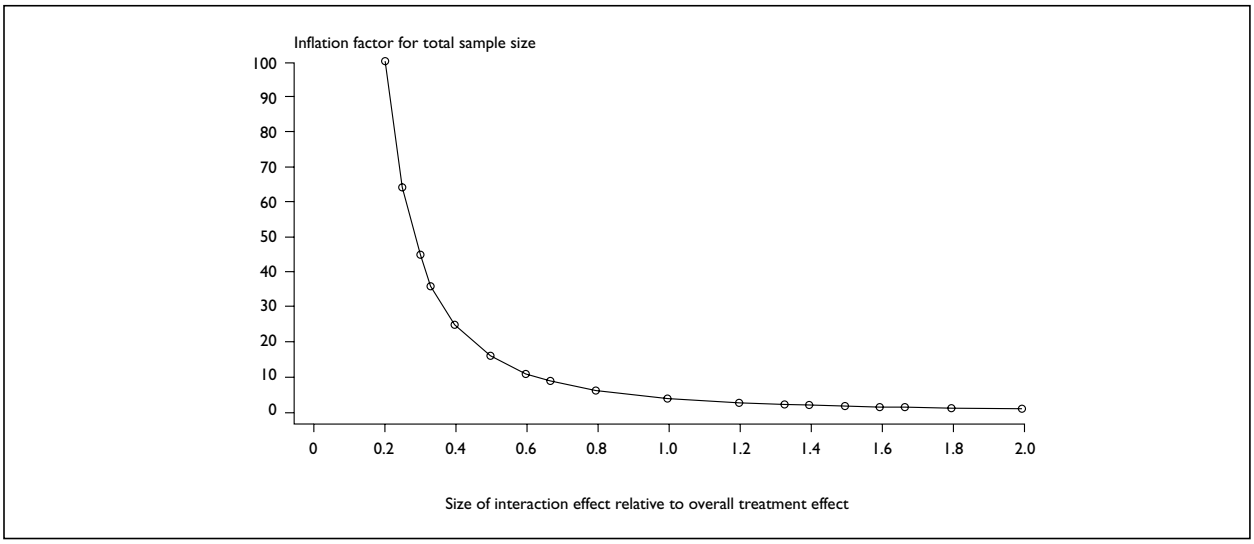


FIGURE 12 *Inflation factors required to increase sample sizes so that interaction tests have the same power as that for the overall treatment effect*

actions half the size of the overall effect, the inflation factor is about 16. As would be expected, the inflation factor has no upper bound as ψ approaches zero; more importantly, for interactions more subtle than, say, one-third to one-fifth of the overall effect, the inflation factor approaches levels that are most unlikely to be achievable in practice (about 50 or higher).

Comparison with a theoretical model

All the above figures were obtained by the use of simulations, as described in the methods. In principle, at least some of them could be obtained from a theoretical perspective, as, for example, used previously for a binary outcome in the special case of an interaction the same size as the overall effect (that is, ψ = 1 in our terminology).[25] The reasons for using a simulation approach throughout this report were given in chapter 2. Nevertheless, it would seem of value to investigate such an alternative approach for a selected example here. Partly to compare with this previous method and partly due to ease of specification, in addition to the simulations for a continuous outcome, the power for the interaction test and hence the inflation factor for ψ = 1 was obtained for a binary outcome using the computer program Power developed by the National Cancer Institute and based on a model for binary outcomes.[26,27]

The levels of power and the inflation factors were consistent (at about 29% and 4, respectively, for this example) across all these analyses and with previous work.[25] However, it is stressed again that a value of 1 for ψ is very large and, as shown in *Figure 12*, the commonly used 'rule of four' is a gross underestimate of the inflation factor required for more subtle (and realistic) interactions. The particular theoretical approach used previously[25] needs to be re-considered for situations where the interactions are much smaller than the overall effect.

The effect of modifying the treatment group ratio

As the ratio of the size of the two treatment groups was varied, the subgroup ratio remained at 1:1 throughout and the overall sample sizes remained as before.

Data simulated with no overall treatment or subgroup effects

Overall treatment effect found to be non-significant (correct-negative result)

The percentage of the 100,000 simulated datasets in which a non-significant overall treatment effect was correctly observed ranged (across sample sizes) between 94.9 and 95.1% for a treatment group ratio of 1:2, between 94.9 and 95.2% for a ratio of 1:3, between 94.9 and 95.2% for 1:4 and 94.9 and 95.1% for 1:5. *Table 6* shows the results of the formal tests of interaction and subgroup-specific tests of treatment effect. All the results were very similar to those of the simplest case with equal-sized treatment groups, that is, no patterns with increasing treatment group ratio.

Overall treatment effect found to be significant (false-positive result, type I error)

Across the sample sizes, about 5% of the 100,000 simulated datasets resulted in an incorrect significant overall treatment effect. The results of the interaction and subgroup-specific tests of treatment effect performed on these (false-positive) datasets are shown in *Table 7*. The results are again very similar to those for the simplest case with equal-sized treatment groups.

Data simulated with an overall treatment effect but no differential subgroup effects

Overall treatment effect found to be significant (correct-positive result)

The percentage of the 100,000 simulated datasets for each different sample size that (correctly)

TABLE 6 *The effect of varying the treatment group ratio: data simulated with no treatment or subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a non-significant overall treatment effect*

Treatment group ratio	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
1:1	4.80–5.20	← 6.60–7.10	0.12–0.15	Theoretically impossible	92.70–93.30 →
1:2	4.80–5.10	← 6.60–7.10	0.12–0.17	Theoretically impossible	92.80–93.30 →
1:3	4.90–5.10	← 6.60–6.90	0.12–0.15	Theoretically impossible	93.00–93.30 →
1:4	4.90–5.20	← 6.60–7.10	0.10–0.15	Theoretically impossible	92.80–93.30 →
1:5	4.90–5.10	← 6.70–7.00	0.12–0.15	Theoretically impossible	92.90–93.20 →

TABLE 7 *The effect of varying the treatment group ratio: data simulated with no treatment or subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a significant overall treatment effect*

Treatment group ratio	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
1:1	4.40–5.80	54.60–64.10 →	0.00 for all	2.10–2.90	← 33.30–42.90
1:2	4.40–5.60	55.00–63.50 →	0.00–0.02	2.00–2.90	← 34.10–42.40
1:3	4.50–5.50	59.60–62.60 →	0.00 for all	2.10–2.90	← 35.00–38.00
1:4	4.90–5.10	56.80–63.40 →	0.00 for all	2.20–3.20	← 34.30–41.00
1:5	4.60–5.60	59.30–63.30 →	0.00 for all	2.20–2.10	← 34.10–38.10

TABLE 8 *The effect of varying the treatment group ratio: data simulated with an overall treatment effect but no subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a significant overall treatment effect at a nominal power of 80%*

Treatment group ratio	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
1:1	4.80–5.20	56.70–57.50	0.00 for all	25.50–32.80 →	← 10.20–17.50
1:2	4.80–5.10	56.80–57.40	0.00 for all	25.80–32.60 →	← 10.40–17.10
1:3	4.90–5.10	56.90–57.60	0.00 for all	29.90–32.50 →	← 10.50–12.90
1:4	4.90–5.10	56.10–57.60	0.00 for all	27.50–32.50 →	← 10.40–16.40
1:5	4.90–5.10	56.20–57.40	0.00–0.001	31.30–32.40 →	← 10.50–12.50

TABLE 9 *The effect of varying the treatment group ratio: data simulated with an overall treatment effect but no subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a non-significant overall treatment effect at a nominal power of 80%*

Treatment group ratio	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
1:1	4.60–5.20	20.20–21.40	0.005–0.05	Theoretically impossible	78.50–79.80
1:2	4.60–5.20	19.40–21.10	0.01–0.05	Theoretically impossible	78.80–80.60
1:3	4.70–5.20	20.30–21.50	0.01–0.03	Theoretically impossible	78.50–79.70
1:4	4.70–5.40	19.90–21.50	0.00–0.05	Theoretically impossible	78.50–80.10
1:5	4.70–5.20	20.50–21.30	0.01–0.05	Theoretically impossible	78.60–79.50

demonstrated a significant overall treatment effect was consistent with the nominal power for the overall treatment effect difference. *Table 8* shows the results of the interaction and subgroup-specific tests in those datasets with a significant overall treatment effect, and, as before, the results were very similar to those of the corresponding simplest case (with a 1:1 treatment group ratio).

Overall treatment effect found to be non-significant (false-negative result, type II error)

The percentage of datasets that failed to demonstrate a significant overall treatment effect was about 20% for datasets with 80% nominal power for each of the different treatment group ratios. *Table 9* again shows that varying treatment group ratio within equal-sized subgroups had

no consistent effect on the results of either the formal tests of interaction or the subgroup-specific tests.

The effect of modifying the subgroup ratio

The ratio of the size of the two subgroups was varied while the treatment group ratio was fixed at 1:1 and the overall sample sizes remained as before.

Data simulated with no overall treatment or subgroup effects

Overall treatment effect found to be non-significant (correct-negative result)

The percentage of the 100,000 simulated datasets in which a non-significant overall treatment effect was (correctly) observed was 94.8–95.0% for a subgroup ratio of 1:2, 94.9–95.1% for 1:3, 94.9–95.1% for 1:4 and 94.9–95.2% for a ratio of 1:5.

The results of the interaction and subgroup-specific tests are shown in *Table 10*. The percentage of significant interaction tests ranged about 5% across sample sizes for each subgroup ratio as in the simplest case. The percentage finding the treatment effect to be significant within both subgroups but in opposite directions was very small for each subgroup ratio. The percentage that found only one subgroup to have a significant treatment effect decreased slightly with an increase in subgroup ratio (*Figure 13*) from the (approximately) 7% found for a 1:1 ratio. This is, presumably, the net effect of two competing influences. The first is that for, say, a 1:5 ratio, the single (chance) large treatment effect could occur in either the larger or the smaller subgroup but is much less likely to be significant if it occurs in the smaller. Secondly, for a 1:1 ratio, the observed difference would need to be even larger to be significant, but it would not matter in which subgroup it was observed. However, the reduction in this percentage was only marginal and thus,

TABLE 10 *The effect of varying the subgroup ratio: data simulated with no overall treatment or subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a non-significant overall treatment effect*

Subgroup ratio	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
1:1	4.80–5.20	← 6.60–7.10	0.12–0.15	Theoretically impossible	92.70–93.30 →
1:2	4.90–5.10	← 6.50–7.10	0.11–0.14	Theoretically impossible	92.80–93.40 →
1:3	4.80–5.10	← 6.40–6.70	0.11–0.15	Theoretically impossible	93.20–93.50 →
1:4	4.90–5.20	← 6.30–6.70	0.10–0.16	Theoretically impossible	93.20–93.60 →
1:5	4.90–5.20	← 6.10–6.40	0.09–0.15	Theoretically impossible	93.40–93.80 →

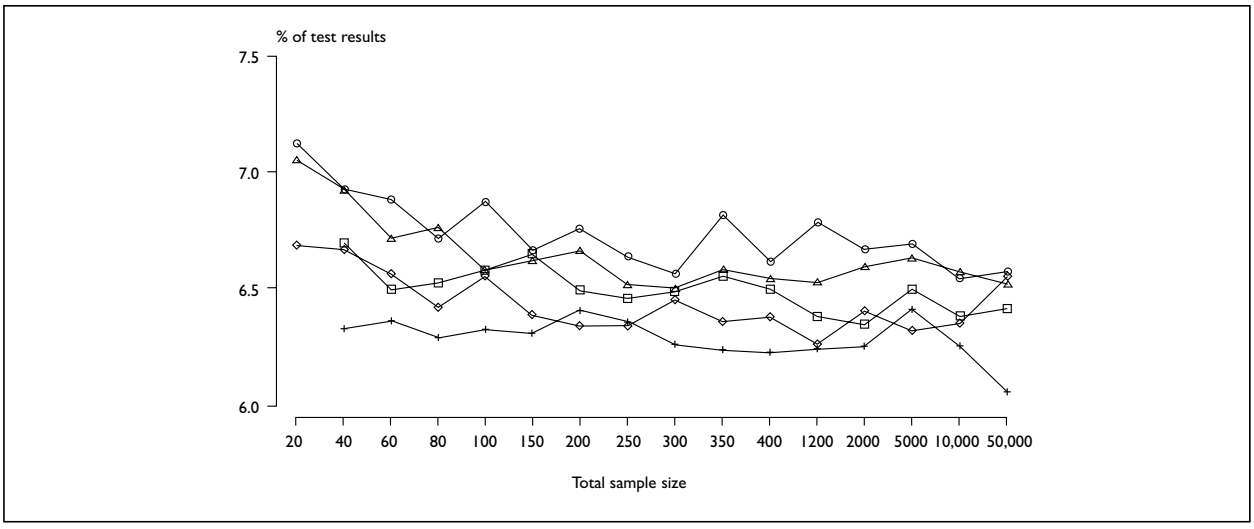


FIGURE 13 *The effect of varying the subgroup ratio: data simulated with no overall treatment or subgroup effects. Percentage of simulated datasets resulting in a non-significant overall treatment effect with a significant treatment effect in only one subgroup. ○, 1:1; △, 1:2; □, 1:3; ◇, 1:4; +, 1:5*

in this case, it appeared that the two competing influences more or less cancelled out.

Overall treatment effect found to be significant (false-positive result, type I error)

The percentage of the 100,000 simulated datasets that found a significant overall treatment effect was about 5% for all subgroup ratios and there was a similar percentage of significant interaction tests.

The results of the subgroup-specific tests within these (false-positive) overall results are shown in *Table 11*. Not surprisingly, the occasions where both tests were significant but in opposite directions were negligible. As shown in *Figure 14*, compared with the simplest (1:1) case, the proportion of analyses in which the treatment effect was significant in only one subgroup increased with more extreme subgroup ratios (to about 71% for a ratio of 1:5). Hence, the balance of influences described previously appeared not to cancel out when there was a (observed) significant overall treatment effect, with the existence of a larger subgroup having a greater effect. The proportion where both subgroup-specific tests were significant in the same direction was about 2.5% for all subgroup ratios.

Data simulated with an overall treatment effect but no differential subgroup effects

Overall treatment effect found to be significant (correct-positive result)

As before, for each subgroup ratio, the percentage of analyses detecting a significant treatment effect was consistent with the nominal power assigned for the treatment effect difference.

For data simulated to have 80% nominal power and with a (correct) significant overall test result, the interaction tests performed had approximately 5% of results significant as anticipated (*Table 12*). As might be expected, the percentage of subgroup-specific tests finding a significant treatment effect in the same direction for both subgroups

TABLE 11 *The effect of varying the subgroup ratio: data simulated with no overall treatment or subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a significant overall treatment effect*

Subgroup ratio	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
1:1	4.40–5.80	54.60–64.10 →	0.00 for all	2.10–2.90	← 33.30–42.90
1:2	4.60–5.60	57.50–65.70 →	0.00–0.02	2.00–3.00	← 31.80–40.10
1:3	4.50–5.90	63.80–68.20 →	0.00–0.06	2.00–2.80	← 29.60–33.70
1:4	4.50–5.30	60.00–69.90 →	0.00–0.10	2.10–3.00	← 27.80–37.30
1:5	4.70–5.60	67.70–71.40 →	0.02–0.21	2.00–2.70	← 25.90–29.80

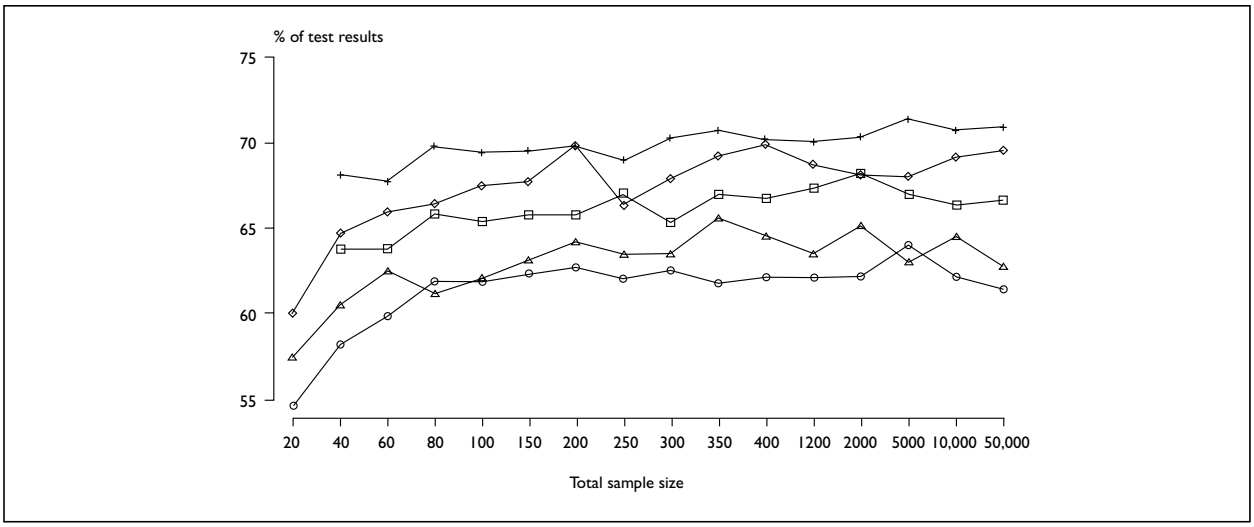


FIGURE 14 *The effect of varying the subgroup ratio: data simulated with no overall treatment or subgroup effects. Percentage of simulated datasets resulting in a significant overall treatment effect with a significant treatment effect in only one subgroup. ○, 1:1; △, 1:2; □, 1:3; ◇, 1:4; +, 1:5*

TABLE 12 *The effect of varying the subgroup ratio: data simulated with an overall treatment effect but no subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a significant overall treatment effect at a nominal power of 80%*

Subgroup ratio	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
1:1	4.80–5.20	56.70–57.50	0.00 for all	25.50–32.80 →	← 10.20–17.50
1:2	4.80–5.40	← 61.30–65.60	0.00–0.01	18.80–28.90 →	← 9.40–15.60
1:3	4.80–5.10	← 66.80–68.70	0.003–0.02	20.40–24.40 →	← 8.30–10.90
1:4	4.90–5.20	← 71.20–76.10	0.01–0.09	10.80–21.50 →	← 7.30–13.00
1:5	4.90–5.10	← 74.30–78.80	0.02–0.11	12.70–19.10 →	← 6.60–8.40

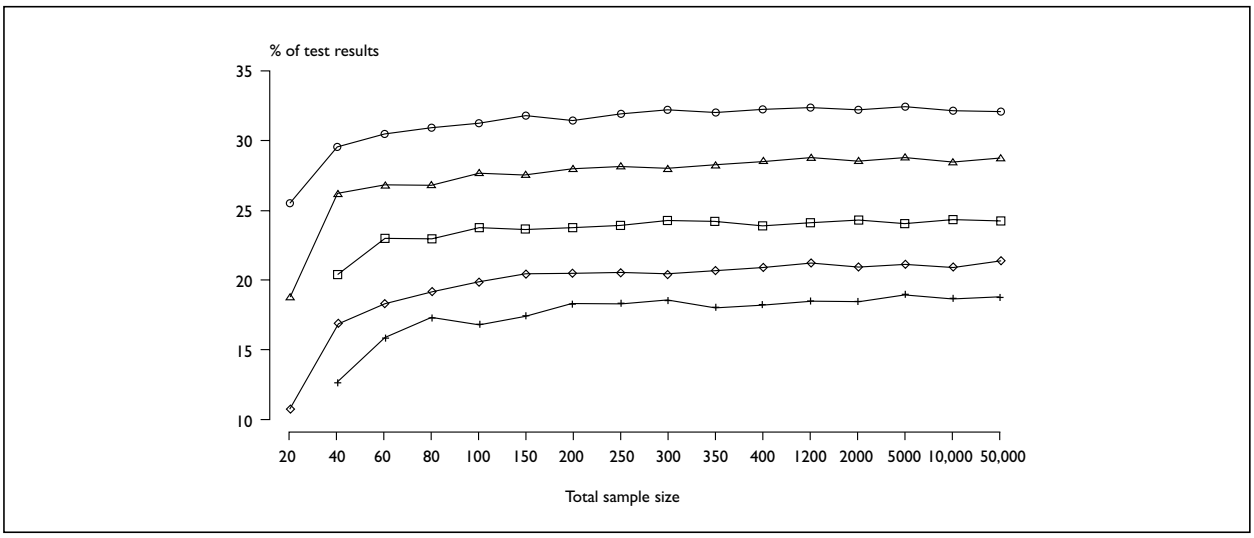


FIGURE 15 *The effect of varying the subgroup ratio: data simulated with an overall treatment effect but no subgroup effects. Percentage of simulated datasets resulting in a significant overall treatment effect with a significant treatment effect in both subgroups in the same direction.* ○, *1:1;* △, *1:2;* □, *1:3;* ◇, *1:4;* +, *1:5*

decreased as the subgroup ratio was increased (*Figure 15*). Conversely, the percentage where just one subgroup-specific test was significant increased quite dramatically as the subgroup ratio departed from 1:1 (*Figure 16*). This is very similar to the situation in the previous section (*Table 11*) where the overall treatment effect was significant (albeit then as a false-positive). The percentage of subgroup-specific tests in which a significant treatment effect was observed for both subgroups but in opposite directions was again close to zero irrespective of subgroup ratio.

Overall treatment effect found to be non-significant (false-negative result, type II error)

The percentage of datasets giving a false-negative overall result reflected the specified nominal power, and the percentage of interaction tests finding a differential treatment effect across subgroups fluctuated about 5% regardless of the subgroup ratio (*Table 13*). Among these datasets, the percentage where both subgroup-specific analyses were significant was negligible (either in the same or opposite directions). The percentage where just one subgroup-specific test was significant declined as the subgroup ratio increased (*Table 13* and *Figure 17*). Although the percentages were higher overall in this case, the pattern was similar to that observed when there was no overall treatment effect (*Table 10*).

The effect of modifying the variance of the data

The impact of departing from the assumption of equal variances across treatment groups and subgroups was explored. In all cases, the number of treatment groups and subgroups was fixed at two and the sizes of the treatment groups and subgroups remained equal.

The variances were modified in a number of different ways. Firstly, the variance for one of the

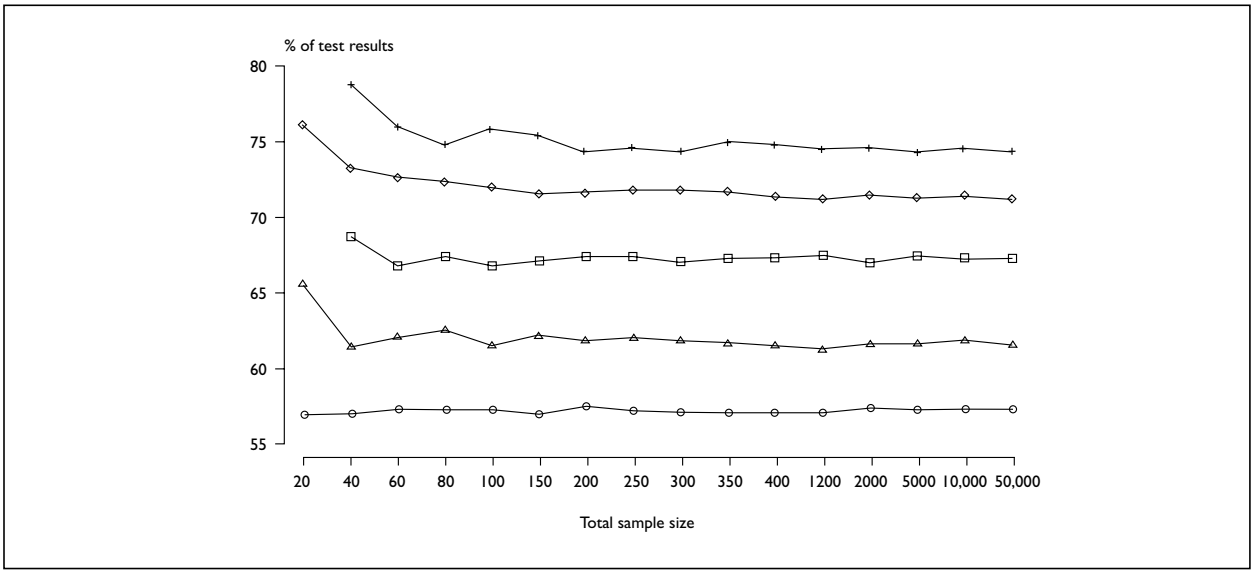


FIGURE 16 *The effect of varying the subgroup ratio: data simulated with an overall treatment effect but no subgroup effects. Percentage of simulated datasets resulting in a significant overall treatment effect with a significant treatment effect in only one subgroup. ○, 1:1; △, 1:2; □, 1:3; ◇, 1:4; +, 1:5*

TABLE 13 *The effect of varying the subgroup ratio: data simulated with an overall treatment effect but no subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a non-significant overall treatment effect at a nominal power of 80%*

Subgroup ratio	% of significant interaction tests	% of significant subgroup-specific tests			
		One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
1:1	4.60–5.20	20.20–21.40	0.005–0.05	Theoretically impossible	78.50–79.80
1:2	4.90–5.30	19.10–20.90	0.03–0.10	Theoretically impossible	79.10–80.80
1:3	4.60–5.20	18.90–20.00	0.07–0.16	Theoretically impossible	79.90–81.00
1:4	4.70–5.20	16.90–19.10	0.10–0.28	Theoretically impossible	80.70–82.80
1:5	4.60–5.20	16.10–18.50	0.16–0.36	Theoretically impossible	81.40–83.50

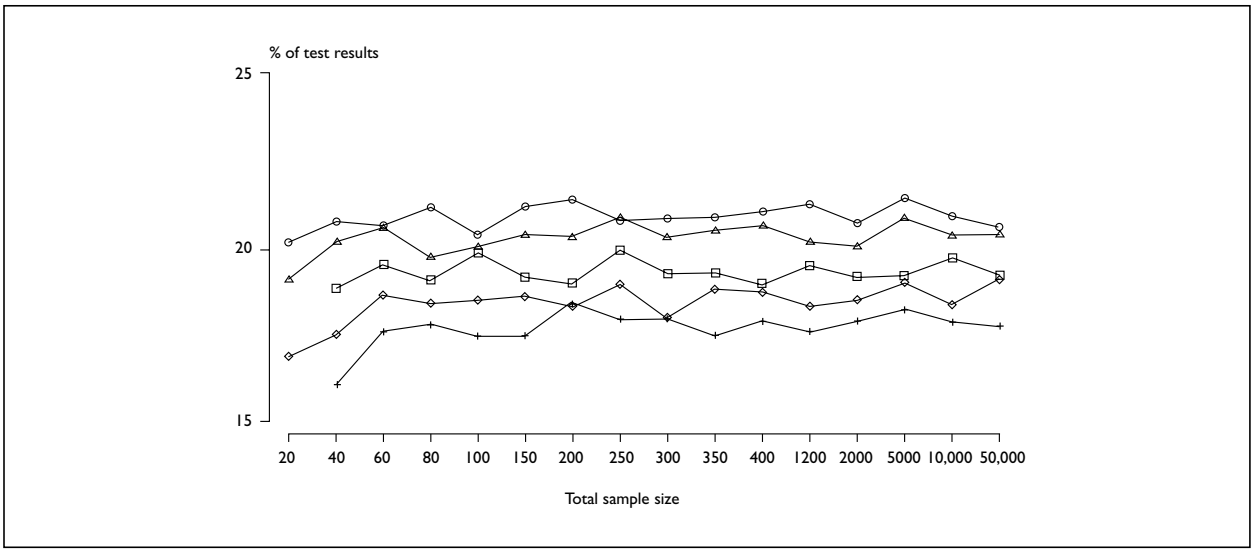


FIGURE 17 *The effect of varying the subgroup ratio: data simulated with an overall treatment effect but no subgroup effects. Percentage of simulated datasets resulting in a non-significant overall treatment effect with a significant treatment effect in only one subgroup. ○, 1:1; △, 1:2; □, 1:3; ◇, 1:4; +, 1:5*

treatment groups or one of the subgroups was changed from 1 to 2 and then to 5. The alterations to the treatment group- and subgroup-specific variances were then made jointly (for values of both 2 and 5), assuming that when the two variances changed together they did so multiplicatively. For example, if the variance in T2, S1 took the value 2 and that for T1, S2 was 5, then the variance in T2, S2 was taken as 10.

It is acknowledged that some of these specifications are somewhat extreme (particularly given the assumption of equal variances in the regression models). The advantage, however, is that all realistic situations will fall between them and, in the event of researchers transforming their data to improve homogeneity, their situation, if anything, would tend towards the simplest case model already covered.

Data simulated with no overall treatment or subgroup effects

Overall treatment effect found to be non-significant (correct-negative result)

Within these datasets (95% of the total), altering the variance in one of the treatment groups had no effect on the formal interaction test, regardless of whether or not the subgroup variances were altered (*Table 14*). Conversely, regardless of the pattern of variances across treatment groups, increasing the variance of one of the subgroups generally reduced the percentage of interaction tests that were significant (in this correct-negative situation), although not markedly so.

The subgroup-specific tests of treatment effect produced very similar findings to those of the simplest case of equal variability (with, as might be anticipated, a slight widening of the ranges across sample sizes when at least one of the variances was relatively large).

Overall treatment effect found to be significant (false-positive result, type I error)

Within these (5%) datasets, altering the variance in one of the treatment groups had almost no effect on the interaction test or the subgroup-specific tests (*Table 15*). However, increasing the variance in one of the subgroups was rather more influential. In contrast to the previous case, the impact on the interaction test was a dramatic increase in the number of significant tests.

Regarding the subgroup-specific tests, the percentages where both were significant in opposite directions (virtually zero) and both were significant in the same direction (about 2.5%) appeared to be unaffected by the pattern of variances. The percentage where just one was significant had a similar pattern to the interaction test in terms of the impact of altering the variances (*Table 15*). This general pattern (i.e. an impact of subgroup but not treatment group variances) can be explained in the

TABLE 14 *The effect of varying the variance of the data: data simulated with no overall treatment or subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a non-significant overall treatment effect*

	T1, S1	T1, S2	T2, S1	T2, S2	% of significant interaction tests	% of significant subgroup-specific tests			
						One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
Simplest case	*N*(0,1)	*N*(0,1)	*N*(0,1)	*N*(0,1)	4.80–5.20	← 6.60–7.10	0.12–0.15	Theoretically impossible	92.70–93.30 →
One treatment group variance increased	*N*(0,1)	*N*(0,1)	*N*(0,2)	*N*(0,2)	4.90–5.20	← 6.60–7.40	0.12–0.16	Theoretically impossible	92.40–93.20 →
	N(0,1)	*N*(0,1)	*N*(0,5)	*N*(0,5)	← 4.90–5.60	← 6.60–8.60	0.12–0.18	Theoretically impossible	91.20–93.20 →
One subgroup variance increased	*N*(0,1)	*N*(0,2)	*N*(0,1)	*N*(0,2)	← 4.60–5.00	← 6.40–7.00	0.12–0.15	Theoretically impossible	92.90–93.40 →
	N(0,1)	*N*(0,5)	*N*(0,1)	*N*(0,5)	← 3.60–4.70	← 6.10–7.10	0.12–0.14	Theoretically impossible	92.80–93.70 →
One treatment group and one subgroup variance increased	*N*(0,1)	*N*(0,2)	*N*(0,2)	*N*(0,4)	← 4.60–5.10	← 6.50–7.50	0.11–0.15	Theoretically impossible	92.30–93.30 →
	N(0,1)	*N*(0,5)	*N*(0,2)	*N*(0,10)	← 3.60–4.70	← 6.20–7.10	0.11–0.15	Theoretically impossible	92.70–93.70 →
	N(0,1)	*N*(0,2)	*N*(0,5)	*N*(0,10)	← 4.60–5.70	← 6.60–8.70	0.13–0.20	Theoretically impossible	91.10–93.30 →
	N(0,1)	*N*(0,5)	*N*(0,5)	*N*(0,25)	← 3.60–5.30	← 6.20–8.60	0.12–0.18	Theoretically impossible	91.20–93.70 →

TABLE 15 *The effect of varying the variance of the data: data simulated with no overall treatment or subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a significant overall treatment effect*

	T1, S1	T1, S2	T2, S1	T2, S2	% of significant interaction tests	% of significant subgroup-specific tests			
						One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
Simplest case	*N*(0,1)	*N*(0,1)	*N*(0,1)	*N*(0,1)	4.60–5.80	54.60–64.10 →	0.00 for all	2.10–2.90	← 33.30–42.90
One treatment group variance increased	*N*(0,1)	*N*(0,1)	*N*(0,2)	*N*(0,2)	4.60–5.60	55.00–64.30 →	0.00 for all	2.20–2.90	← 33.40–42.20
	N(0,1)	*N*(0,1)	*N*(0,5)	*N*(0,5)	4.60–5.70	56.10–63.00 →	0.00 for all	2.10–3.20	← 34.20–40.60
One subgroup variance increased	*N*(0,1)	*N*(0,2)	*N*(0,1)	*N*(0,2)	9.00–11.20 →	56.50–65.60 →	0.00–0.04	2.20–3.40	← 32.10–41.20
	N(0,1)	*N*(0,5)	*N*(0,1)	*N*(0,5)	26.50–30.90 →	62.10–71.30 →	0.00–0.12	1.80–3.00	← 26.50–34.90
One treatment group and one subgroup variance increased	*N*(0,1)	*N*(0,2)	*N*(0,2)	*N*(0,4)	9.80–11.50 →	58.10–65.20 →	0.00–0.02	2.20–2.80	← 32.40–39.50
	N(0,1)	*N*(0,5)	*N*(0,2)	*N*(0,10)	26.60–30.60 →	61.50–71.70 →	0.00–0.12	2.20–2.90	← 25.60–35.60
	N(0,1)	*N*(0,2)	*N*(0,5)	*N*(0,10)	9.80–11.10 →	58.90–66.00 →	0.00–0.02	2.00–3.20	← 31.80–38.00
	N(0,1)	*N*(0,5)	*N*(0,5)	*N*(0,25)	25.20–30.60 →	64.10–71.20 →	0.00–0.12	2.30–3.60	← 26.20–32.30

situation of a false-positive overall treatment effect by the following argument. If the subgroup variances are different but the treatment group variances are equal, this might simply reflect a (chance) differential effect as a result of one of the larger variances if there is an (chance) observed overall treatment effect. However, in the case of different treatment group but equal subgroup variances, an (chance) overall treatment effect does not mean that a differential effect is any more likely than expected by chance (5%). It is emphasised, however, that even though some of the observations in *Table 15* are explained in largely artefactual terms, the error rates remain a true reflection of the dangers of researchers being influenced by the overall treatment effect result in the decisions about whether or not to perform subgroup analyses (even interaction tests).

In addition, the effect of small sample sizes (a total of less than 100) in terms of reducing the percentage where one subgroup test was significant was slightly greater for the more extreme specifications of variances (*Table 15*).

Data simulated with an overall treatment effect but no differential subgroup effects

The overall treatment effects incorporated into these simulations took into account not only the power, as before, but also the (changing) variances (see appendix 1 for details).

Overall treatment effect found to be significant (correct-positive result)

Within these datasets (approximately 80% of the total), the pattern of results from the interaction tests in *Table 16* was very similar to that seen in *Table 14* (the correct-negative case), where the subgroup variances had a small effect on the interaction tests and the treatment group variances were not influential at all.

However, in this case, while differences in the treatment group variances still had no impact on the results of the subgroup-specific results, altering the subgroup variances did. Specifically, the percentage of subgroup-specific tests with one or both subgroup-specific treatment effects that were significant increased marginally with increasing subgroup variances while the percentage where neither was significant decreased correspondingly (*Table 16*).

Overall treatment effect found to be non-significant (false-negative result, type II error)

In these (20%) datasets, an influence of subgroup but not treatment group variance was again observed for the interaction tests (*Table 17*). Not surprisingly, in the case of an overall false-negative result, situations where both subgroup-specific tests were significant (either in the same or opposite directions) were very rare. The percentage where only one subgroup-specific treatment effect was significant dramatically increased with increasing variance in one subgroup with a corresponding decrease in the percentage where neither subgroup-specific test was significant. The results for this situation were intermediate between the correct-negative (where the trend by subgroup variance was very weakly in the opposite direction) and the false-positive cases

TABLE 16 *The effect of varying the variance of the data: data simulated with an overall treatment effect but no differential subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a significant overall treatment effect at a nominal power of 80%*

	T1, S1	T1, S2	T2, S1	T2, S2	% of significant interaction tests	% of significant subgroup-specific tests			
						One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
Simplest case	$N(0,1)$	$N(0,1)$	$N(\delta,1)$	$N(\delta,1)$	4.90–5.20	57.00–57.50	0.00 for all	25.50–32.50 →	← 10.20–17.50
One treatment group variance increased	$N(0,1)$	$N(0,1)$	$N(\delta,2)$	$N(\delta,2)$	4.90–5.20	57.40–58.10	0.00 for all	24.80–31.40 →	← 10.80–17.50
	$N(0,1)$	$N(0,1)$	$N(\delta,5)$	$N(\delta,5)$	← 4.90–5.60	57.90–59.90	0.00–0.001	23.70–28.20 →	← 12.20–18.50
One subgroup variance increased	$N(0,1)$	$N(0,2)$	$N(\delta,1)$	$N(\delta,2)$	← 4.60–5.00	58.10–59.10	0.00 for all	26.40–34.40 →	← 8.40–15.00
	$N(0,1)$	$N(0,5)$	$N(\delta,1)$	$N(\delta,5)$	← 3.70–4.80	61.20–64.00	0.00–0.001	29.80–36.40 →	← 2.20–6.20
One treatment group and one subgroup variance increased	$N(0,1)$	$N(0,2)$	$N(\delta,2)$	$N(\delta,4)$	← 4.50–5.10	58.60–59.20	0.00 for all	26.00–32.50 →	← 8.70–15.20
	$N(0,1)$	$N(0,5)$	$N(\delta,2)$	$N(\delta,10)$	← 3.70–5.00	61.40–63.90	0.00–0.001	29.50–35.80 →	← 2.80–6.70
	$N(0,1)$	$N(0,2)$	$N(\delta,5)$	$N(\delta,10)$	← 4.60–5.60	59.60–60.90	0.00–0.001	24.40–29.40 →	← 10.20–16.00
	$N(0,1)$	$N(0,5)$	$N(\delta,5)$	$N(\delta,25)$	← 3.90–5.80	62.80–63.30	0.00–0.002	28.90–33.30 →	← 3.70–8.00

TABLE 17 *The effect of varying the variance of the data: data simulated with an overall treatment effect but no subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a non-significant overall treatment effect at a nominal power of 80%*

	T1, S1	T1, S2	T2, S1	T2, S2	% of significant interaction tests	% of significant subgroup-specific tests			
						One subgroup only	Both subgroups in opposite directions	Both subgroups in the same direction	Neither subgroup
Simplest case	$N(0,1)$	$N(0,1)$	$N(\delta,1)$	$N(\delta,1)$	4.60–5.20	20.20–21.40	0.005–0.04	Theoretically impossible	78.50–79.80
One treatment group variance increased	$N(0,1)$	$N(0,1)$	$N(\delta,2)$	$N(\delta,2)$	4.70–5.30	20.00–21.20	0.01–0.05	Theoretically impossible	78.80–80.00
	$N(0,1)$	$N(0,1)$	$N(\delta,5)$	$N(\delta,5)$	4.80–5.80	19.80–22.10	0.008–0.05	Theoretically impossible	77.80–80.20
One subgroup variance increased	$N(0,1)$	$N(0,2)$	$N(\delta,1)$	$N(\delta,2)$	6.10–6.60	34.30–36.40	0.01–0.08	Theoretically impossible	63.50–67.60
	$N(0,1)$	$N(0,5)$	$N(\delta,1)$	$N(\delta,5)$	8.20–9.10	71.50–78.50	0.07–0.17	Theoretically impossible	21.40–28.30
One treatment group and one subgroup variance increased	$N(0,1)$	$N(0,2)$	$N(\delta,2)$	$N(\delta,4)$	6.00–6.50	32.90–35.40	0.02–0.08	Theoretically impossible	64.60–67.00
	$N(0,1)$	$N(0,5)$	$N(\delta,2)$	$N(\delta,10)$	8.00–8.60	70.60–77.60	0.07–0.20	Theoretically impossible	22.30–29.30
	$N(0,1)$	$N(0,2)$	$N(\delta,5)$	$N(\delta,10)$	5.70–6.40	32.00–33.00	0.02–0.09	Theoretically impossible	66.90–67.90
	$N(0,1)$	$N(0,5)$	$N(\delta,5)$	$N(\delta,25)$	7.10–8.00	66.00–73.40	0.10–0.20	Theoretically impossible	26.50–33.80

seen in the previous section (where the impact was dramatic if essentially artefactual). This is likely to reflect the fact that in the current (false-negative) situation, there is a proportion of datasets where the observed difference is non-zero but just not large enough to be significant.

It should be remembered here that, for all the simulations, only one specification was varied at a time. Hence, the above results for unequal variances relate just to the case of equal treatment group and subgroup sizes. It would be anticipated that the impact of unequal variances would increase for unequal group sizes.

The effect of modifying the number of subgroups

All simulations up to this point were restricted to the simplest case of two equal-sized treatment groups and two equal-sized subgroups. In this section, the effect of having more than two subgroups was explored. An additional complication was then that the actual results from subgroup-specific tests could take a number of different forms. As a simplification, just the four types described in the methods section were distinguished for the following tables. This complication does not arise for the interaction test, which is a global test of differences between the subgroup-specific effects across all subgroups.

Data simulated with no overall treatment or subgroup effects

Overall treatment effect found to be non-significant (correct-negative result)

The percentage of the 100,000 simulated datasets with a correct non-significant overall treatment effect fluctuated about 95% as expected (94.8–95.2% for three, four and five subgroups), and the formal interaction test resulted in a significant finding in about 5% of cases, again as anticipated (*Table 18*).

In terms of subgroup-specific tests of the treatment effect, the percentage where all subgroup-specific analyses resulted in a significant treatment effect, either in the same or different directions, was unsurprisingly negligible (*Table 18*). In addition, the percentage finding no subgroups to have a significant treatment effect decreased with increasing numbers of subgroups. Correspondingly, the percentage with one or more but not all subgroups being significant increased quite considerably with increasing numbers of subgroups, from about 7% for two subgroups to about 21% for five subgroups (*Figure 18*).

Overall treatment effect found to be significant (false-positive result, type I error)

Within these datasets (approximately 5% of the total), the interaction test was significant approximately 5% of the time. Other than for the case of two subgroups already discussed, the percentages where all subgroup-specific tests were significant (in the same or opposite directions) were negligible. In contrast to the preceding section, the number of subgroups did not appreciably affect the percentages of either no subgroup-specific tests being significant (tending to 35% for total sample sizes over 200) or of one or more but not all results being significant (tending to 65% over 200; *Table 19*). The patterns across sample sizes were consistent with those seen in the simplest case.

Data simulated with an overall treatment effect but no differential subgroup effects

Overall treatment effect found to be significant (correct-positive result)

Within these datasets (approximately 80% of the total), the interaction test was again significant approximately 5% of the time. As was the case for two subgroups, significant results in all subgroups but in inconsistent directions were virtually never observed (*Table 20*). As would be expected, the percentage where all subgroup-specific treatment effects were significant in the same direction reduced dramatically with the number of subgroups from about 33% for two subgroups to < 1% for four or more subgroups (with these and the other limits in

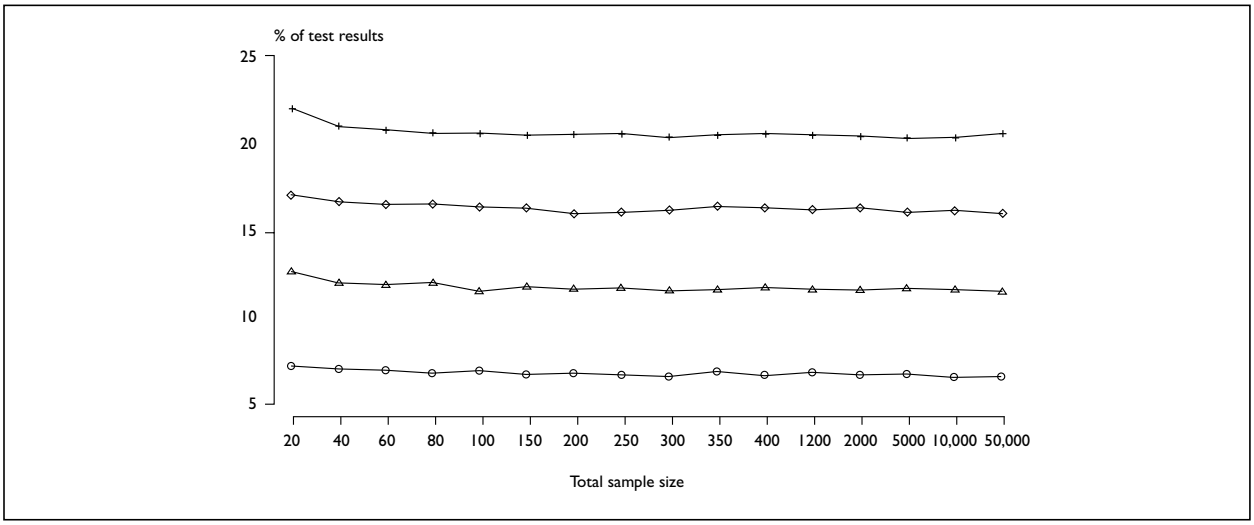


FIGURE 18 *The effect of varying the number of subgroups: data simulated with no overall treatment or subgroup effects. Percentage of simulated datasets resulting in a non-significant overall treatment effect with a significant treatment effect in one or more but not all subgroups. ○, two subgroups; △, three subgroups; ◇, four subgroups; +, five subgroups*

TABLE 18 *The effect of varying the number of subgroups: data simulated with no overall treatment or subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a non-significant overall treatment effect*

Number of subgroups	% of significant interaction tests	% of significant subgroup-specific tests			
		One or more but not all subgroups	All subgroups in opposite directions	All subgroups in the same direction	No subgroups
2	4.80–5.20	← 6.60–7.10	0.10–0.20	Theoretically impossible	92.70–93.30 →
3	4.70–5.00	← 11.50–12.60	0.002–0.02	Theoretically impossible	87.40–88.50 →
4	4.90–5.20	← 16.00–17.00	0.00–0.002	Theoretically impossible	83.00–84.00 →
5	4.90–5.20	← 20.30–21.90	0.00–0.001	Theoretically impossible	78.10–79.70 →

TABLE 19 *The effect of varying the number of subgroups: data simulated with no overall treatment or subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a significant overall treatment effect*

Number of subgroups	% of significant interaction tests	% of significant subgroup-specific tests			
		One or more but not all subgroups	All subgroups in opposite directions	All subgroups in the same direction	No subgroups
2	4.40–5.80	54.60–64.10 →	0.00 for all	2.10–2.90	← 33.30–42.90
3	4.60–5.50	47.30–64.40 →	0.00–0.06	0.00–0.80	← 35.60–52.60
4	4.50–5.60	47.80–64.80 →	0.00–0.02	0.00–0.02	← 35.20–52.20
5	4.40–5.60	40.50–65.70 →	0.00 for all	0.00 for all	← 34.30–59.50

TABLE 20 *The effect of varying the number of subgroups: data simulated with an overall treatment but no subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a significant overall treatment effect at a nominal power of 80%*

Number of subgroups	% of significant interaction tests	% of significant subgroup-specific tests			
		One or more but not all subgroups	All subgroups in opposite directions	All subgroups in the same direction	No subgroups
2	4.80–5.20	56.70–57.50	0.00 for all	25.50–32.80 →	← 10.20–17.50
3	4.90–5.20	67.00–80.70 →	0.00–0.01	2.30–6.30 →	← 13.30–30.80
4	4.80–5.20	68.30–84.70 →	0.00–0.01	0.20–0.90	← 14.50–31.50
5	4.90–5.10	54.80–84.80 →	0.00–0.003	0.01–0.10	← 15.10–45.20

TABLE 21 *The effect of varying the number of subgroups: data simulated with an overall treatment but no subgroup effects. Percentage of significant results (range across sample sizes) in datasets with a non-significant overall treatment effect at a nominal power of 80%*

Number of subgroups	% of significant interaction tests	% of significant subgroup-specific tests			
		One or more but not all subgroups	All subgroups in opposite directions	All subgroups in the same direction	No subgroups
2	4.60–5.20	20.20–21.40	0.005–0.05	Theoretically impossible	78.50–79.80
3	4.70–5.30	23.00–27.50 →	0.01–0.05	Theoretically impossible	← 72.50–77.00
4	4.70–5.20	27.30–31.50 →	0.00–0.005	Theoretically impossible	← 68.50–72.70
5	4.70–5.20	28.50–35.70 →	0.00 for all	Theoretically impossible	← 64.30–71.50

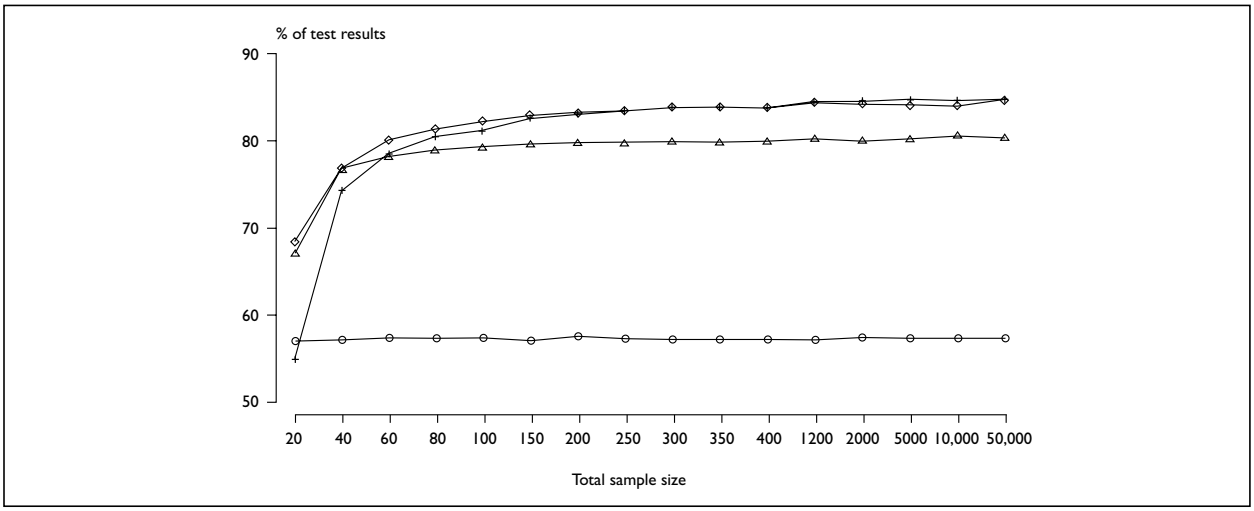


FIGURE 19 *The effect of varying the number of subgroups: data simulated with an overall treatment but no subgroup effects. Percentage of simulated datasets resulting in a significant overall treatment effect with a significant treatment effect in one or more but not all subgroups. ○, two subgroups; △, three subgroups; ◇, four subgroups; +, five subgroups*

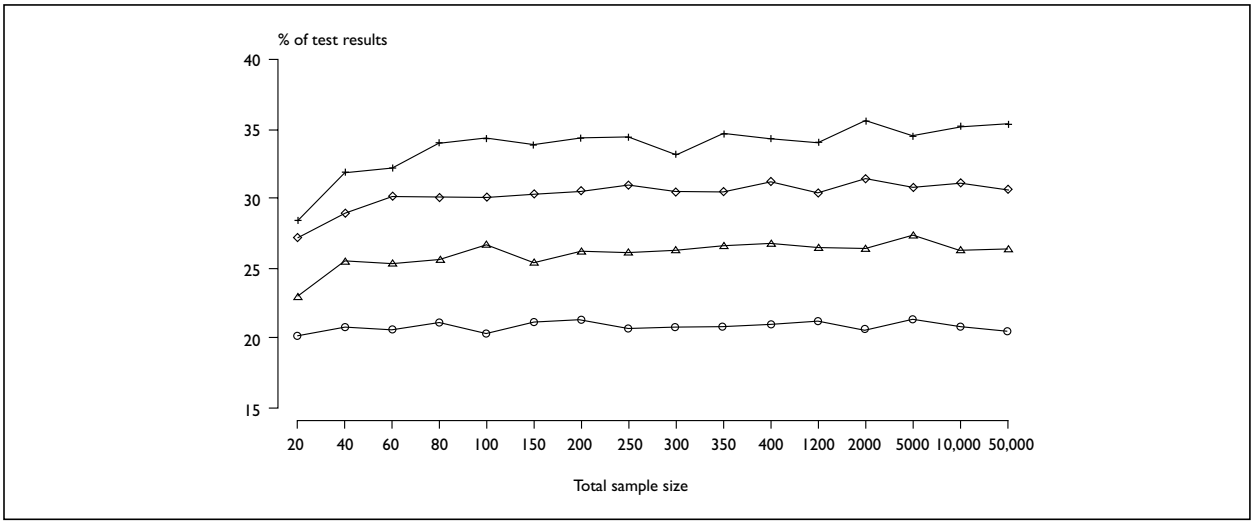


FIGURE 20 *The effect of varying the number of subgroups: data simulated with an overall treatment but no subgroup effects. Percentage of simulated datasets resulting in a non-significant overall treatment effect with a significant treatment effect in one or more but not all subgroups. ○, two subgroups; △, three subgroups; ◇, four subgroups; +, five subgroups*

Table 20 reached for total sample sizes over about 80). The percentage of no subgroup-specific test results being significant slightly increased in correspondence with the number of subgroups (from about 10% with two subgroups to about 15% for five subgroups). In addition, the percentage with one or more but not all subgroup-specific results being significant increased with the number of subgroups. However, this increase was not linear and for more than two subgroups the percentage seemed to converge rapidly to about 85% (*Table 20* and *Figure 19*).

Overall treatment effect found to be non-significant (false-negative result, type II error)

Within these datasets (approximately 20% of the total), the interaction test was again significant approximately 5% of the time (*Table 21*). In terms of subgroup-specific tests, not surprisingly, the percentages where all subgroup-specific analyses resulted in a significant treatment effect, either in the same or different directions, were negligible. In addition, as might be expected, the percentage finding no subgroups to have a significant treatment effect decreased with an increasing number of subgroups. Correspondingly, the percentage with one or more but not all subgroups significant increased with increasing numbers of subgroups, from about 21% for two subgroups to about 36% for five subgroups (*Figure 20*).

The main findings and conclusions of chapter 3 are summarised and discussed in chapter 6.

Chapter 4

Results – binary outcome data

The performance of subgroup analyses conducted on binary data was very similar to that of the continuous case. For brevity, only the results for the simplest case (two equal-sized treatment groups and two equal-sized subgroups) are summarised in this chapter. The variation observed in the percentages for sample sizes less than 200 was greater than that for the continuous case and this increased fluctuation amongst small datasets means that researchers should be very cautious of the findings of subgroup-specific tests for binary outcomes with total sample sizes below 200, and especially those below 100.

The results presented in *Tables 22* to *25* are the approximate values tended to for total sample sizes over 200, with the proportion positive on the binary outcome set (essentially arbitrarily) at 0.5 for one or both treatment groups. 'Reference' values other than 0.5, such as 0.1 and 0.2, were also specified for the simulations, but this had no appreciable effect on the findings and, in particular, on the percentages for larger sample sizes.

Simplest case – two subgroups and complete balance

Data simulated with no overall treatment or subgroup effects

The percentage of the 100,000 simulated datasets within which the overall treatment effect was correctly found to be non-significant fluctuated about 95%, and 5% (incorrectly) found the overall treatment effect to be significant (type I error) as expected. *Tables 22* and *23* present the results of the interaction and subgroup-specific tests within the corresponding datasets. The findings are very similar to those seen for the continuous case (see *Tables 2* and *3*).

Data simulated with an overall treatment effect but no differential subgroup effects

Treatment effect differences that would be detectable with 80% power were calculated for each sample size considered (see appendix 1).

The percentage of analyses that correctly found the overall treatment effect to be significant

TABLE 22 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of interaction and subgroup-specific tests of treatment effect in datasets with a correct-negative overall result*

Test results	Approximate % tended to
Interaction test significant	5.0
Subgroup-specific tests	
One subgroup significant	7.0
Both subgroups significant in opposite directions	< 1.0
Both subgroups significant in the same direction	Theoretically impossible
Neither subgroup significant	93.0

TABLE 23 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of interaction and subgroup-specific tests of treatment effect in datasets with a false-positive overall result*

Test results	Approximate % tended to
Interaction test significant	5.0
Subgroup-specific tests	
One subgroup significant	63.0
Both subgroups significant in opposite directions	0.0
Both subgroups significant in the same direction	2.5
Neither subgroup significant	34.0

fluctuated about 80% (or lower for a sample size of 40) and about 20% (incorrectly) found a non-significant overall treatment effect (type II error). *Tables 24* and *25* present the corresponding results of the interaction and subgroup-specific tests in datasets with each of these overall findings. The findings are very similar to those seen for the 80% powered continuous case (see *Tables 4* and *5*). As a result of this similarity, 90 and 95% powers were not considered for the binary outcome case.

TABLE 24 *Simplest case: data simulated with an overall treatment effect but no differential subgroup effects. Results of interaction and subgroup-specific tests of treatment effect in datasets with a correct-positive overall result*

Test results	Approximate % tended to
Interaction test significant	5.0
Subgroup-specific tests	
One subgroup significant	57.0
Both subgroups significant in opposite directions	0.0
Both subgroups significant in the same direction	33.0
Neither subgroup significant	10.0

TABLE 25 *Simplest case: data simulated with an overall treatment effect but no differential subgroup effects. Results of interaction and subgroup-specific tests of treatment effect in datasets with a false-negative overall result*

Test results	Approximate % tended to
Interaction test significant	5.0
Subgroup-specific tests	
One subgroup significant	21.0
Both subgroups significant in opposite directions	< 0.1
Both subgroups significant in the same direction	Theoretically impossible
Neither subgroup significant	79.0

Chapter 5

Results – survival outcome data

The performance of subgroup analyses carried out on survival data was very similar to that of the continuous case. The results of the interaction tests and subgroup-specific tests for the survival case are summarised in this chapter. The percentages presented in *Tables 26* to *29* are the approximate values tended to for total sample sizes over 200 due to the extent of the fluctuation for sample sizes smaller than this, although this fluctuation was less for survival data than for binary data. However, there remained a degree of fluctuation for larger sample sizes.

The results presented here are based on a total follow-up period of 60 months and a median survival time of 36 months. Simulations based on much shorter median survival times (for example, 6 months) were also performed. Although small systematic differences were seen in terms of absolute values (as the median survival time declined, the ability of the Cox proportional hazards model to reject the null hypothesis declined slightly), the general conclusions remained the same.

Only the results of the simplest case (two equal-sized treatment groups and two equal-sized subgroups) are presented here.

Simplest case – two subgroups and complete balance

Data simulated with no overall treatment or subgroup effects

The percentage of the 100,000 simulated datasets within which the overall treatment effect was correctly found to be non-significant fluctuated about 95%, and 5% (incorrectly) found the overall treatment effect to be significant (type I error). *Tables 26* and *27* present the results of the interaction and subgroup-specific tests within datasets of each type of overall test result. The findings were very similar to those seen for the continuous case (see *Tables 2* and *3*).

TABLE 26 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of interaction and subgroup-specific tests of treatment effect in datasets with a correct-negative overall result*

Test results	Approximate % tended to
Interaction test significant	5.0
Subgroup-specific tests	
One subgroup significant	6.5
Both subgroups significant in opposite directions	< 1.0
Both subgroups significant in the same direction	Theoretically impossible
Neither subgroup significant	93.0

TABLE 27 *Simplest case: data simulated with no overall treatment or subgroup effects. Results of interaction and subgroup-specific tests of treatment effect in datasets with a false-positive overall result*

Test results	Approximate % tended to
Interaction test significant	5.0
Subgroup-specific tests	
One subgroup significant	61.0
Both subgroups significant in opposite directions	0.0
Both subgroups significant in the same direction	2.5
Neither subgroup significant	36.0

Data simulated with an overall treatment effect but no differential subgroup effects

Treatment effect differences detectable with 80% power were calculated for each sample size considered (see appendix 1). The percentage of analyses that correctly found the overall treatment effect to be significant fluctuated about 80%, and about 20% (incorrectly) found a non-significant overall treatment effect (type II error). *Tables 28* and *29* present the results of the interaction and subgroup-specific tests in datasets with each of these overall test results. The findings were very similar to those seen for the continuous case (see *Tables 4* and *5*).

TABLE 28 *Simplest case: data simulated with an overall treatment effect but no differential subgroup effects. Results of interaction and subgroup-specific tests of treatment effect in datasets with a correct-positive overall result*

Test results	Approximate % tended to
Interaction test significant	5.0
Subgroup-specific tests	
One subgroup significant	57.0
Both subgroups significant in opposite directions	0.0
Both subgroups significant in the same direction	32.0
Neither subgroup significant	11.0

TABLE 29 *Simplest case: data simulated with an overall treatment effect but no differential subgroup effects. Results of interaction and subgroup-specific tests of treatment effect in datasets with a false-negative overall result*

Test results	Approximate % tended to
Interaction test significant	5.0
Subgroup-specific tests	
One subgroup significant	21.0
Both subgroups significant in opposite directions	< 0.1
Both subgroups significant in the same direction	Theoretically impossible
Neither subgroup significant	79.0

Chapter 6
Discussion

Influence of the type of outcome

The findings of the continuous, binary and survival outcome data were very similar with the exception of greater instability in the percentages for smaller sample size with binary and survival data. The reason for the high degree of stability across sample sizes in the results for the continuous outcome data was that the data for these simulations were drawn from a (theoretical) Gaussian distribution and thus the behaviour of summary statistics from the (simulated) samples did not rely upon the central-limit theorem. In other words, the performance of the test statistics did not depend on sample size. The factor that was influential with respect to stability was the number of simulations performed for each test statistic. A total of 100,000 simulations were chosen (and used for all sample sizes) to provide adequate precision of the estimates of the percentages that were statistically significant, irrespective of total sample size.

For the binary and survival cases, the (theoretical) distributions from which data were sampled are clearly not Gaussian. Even using maximum likelihood methods (that is, using measures of deviance for tests), the approximation of the distribution of test statistics to chi-squared distributions is an asymptotic result, and, hence, some dependency on sample size would be anticipated.

Other than this, due to the overall similarities between the findings for the three different types of outcome data, the results summarised in this concluding chapter are drawn from the continuous case. The rationale for this is that, as explained in chapters 4 and 5, chapter 3 covered a more comprehensive set of scenarios for changing specifications. Moreover, there was greater consistency in the specifications for the simulations in the continuous case, for example, in terms of true differential effects. However, it remains that the general patterns and, in many cases, the absolute findings were very similar for the other types of outcome data.

Practical considerations

The aim of this study was to quantify the extent to which subgroup analyses may be misleading. However, in addition to the percentages presented in this report, it is also important to consider how the results from subgroup analyses might be interpreted and translated into practice. Of particular concern is the scenario in which subgroup-specific tests indicate that treatment is effective in one subgroup only. If this situation arises then treatment may be erroneously withheld in one group and, moreover, the effectiveness of the treatment in the other group is likely to be over-estimated. As a result of this potential impact, even a small increase in the number of misleading results has serious implications for the provision of suitable treatment.

In chapters 3, 4 and 5, the results were presented according to whether or not the simulated data were generated with the relevant (overall or differential) treatment effect and, within each of these two underlying scenarios, whether or not the overall test was significant. The reason for this approach was two-fold. Firstly, when simulating data, the underlying distributions for the treatment group/subgroup categories, the overall treatment effects and any differential effects must be specified. Secondly, at least initially, it was valuable to keep the various eventualities (for example, correct-negative, false-positive, correct-positive, false-negative) separate because the impact on the results was marked in general.

However, in practice, it is stressed that the nature of the real effects will not be known. The results are, therefore, summarised here distinguishing only between the situations where the relevant overall test either is or is not (observed to be) statistically significant. The rationale for retaining this latter distinction in the results is the high chance that the overall test result (that is, whether the treatment groups differ significantly on the primary intention-to-treat analysis) will influence, at least to some degree, the extent to which subgroup analyses are conducted. Although it might have been more useful to combine the two separate situations (for example, false-positive and correct-positive) in a numerical fashion, this was not possible as the correct weighting of the two percentages was unknown (and, indeed, unknowable in practice). The results are, therefore, summarised here as a range across the possible situations.

The exception to this approach is in the case of the different types of differential effects, where the results are so dependent on the type of interaction that summarising the conclusions across them would be meaningless. To some extent, this makes the absolute figures in the relevant section artificial (since it relies on knowing the type of interaction), but researchers will have strong theoretical or prior observational expectations of the likely pattern in most cases in practice. Moreover, when this is not the case, there is a serious doubt about the wisdom of performing subgroup analyses in the first place.

Summary of results for the simplest case

No differential effects

Irrespective of the overall test result, the interaction test generally performed well in the sense that it yielded approximately 5% false-positives (*Figures 21* and *22*). The performance of the subgroup-specific tests was rather more erratic. Irrespective of the overall result, there were almost no instances where both tests were significant in opposite directions. However, if the overall finding was significant, the chance of finding just one subgroup significant could be as high as two in three (*Figure 22*). While there is an inevitable element of supposition as to how researchers might interpret this latter eventuality (i.e. as a differential effect or not), the finding of significance in just one subgroup would be quite likely to be misinterpreted. If the overall test result was non-significant (*Figure 21*), then the chance of just one subgroup-specific test being significant was still at least 7% and could be as high as 21% (depending on whether or not there was a true overall effect).

All these probabilities were calculated considering a single subgroup analyses and ignoring the issue of multiple testing, and it could, therefore, be argued that these risks relate to 'best-case' scenarios. As discussed in the simulation strategy section, it is emphasised that while the absolute values quoted here correspond to the nominal 5% threshold for statistical significance, they would be expected to show the same general influences on all levels of significance.

Differential effects

When a 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 effects. However, 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. In addition, the subgroup-specific analyses often suggested the wrong type of differential effect.

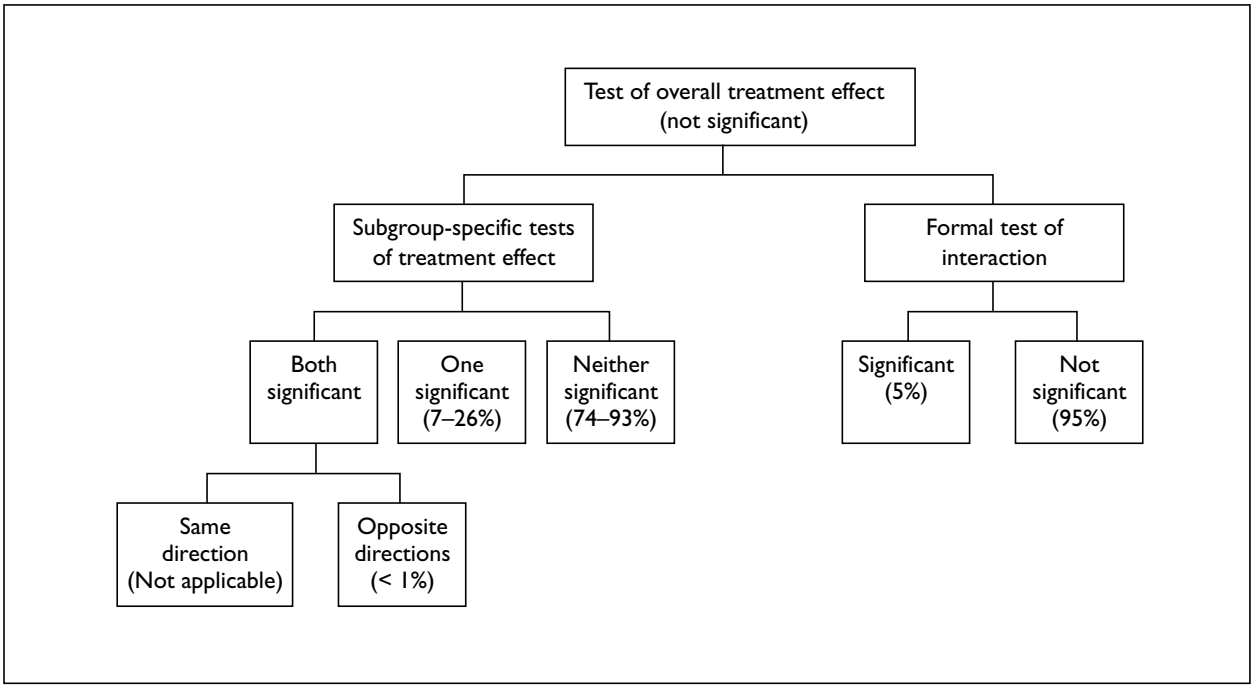


FIGURE 21 *Summary of results for the simplest case (overall test result not significant). This figure combines the results from data simulated with no overall treatment effect and with a true overall treatment effect detectable at nominal powers of 50, 80, 90 and 95%*

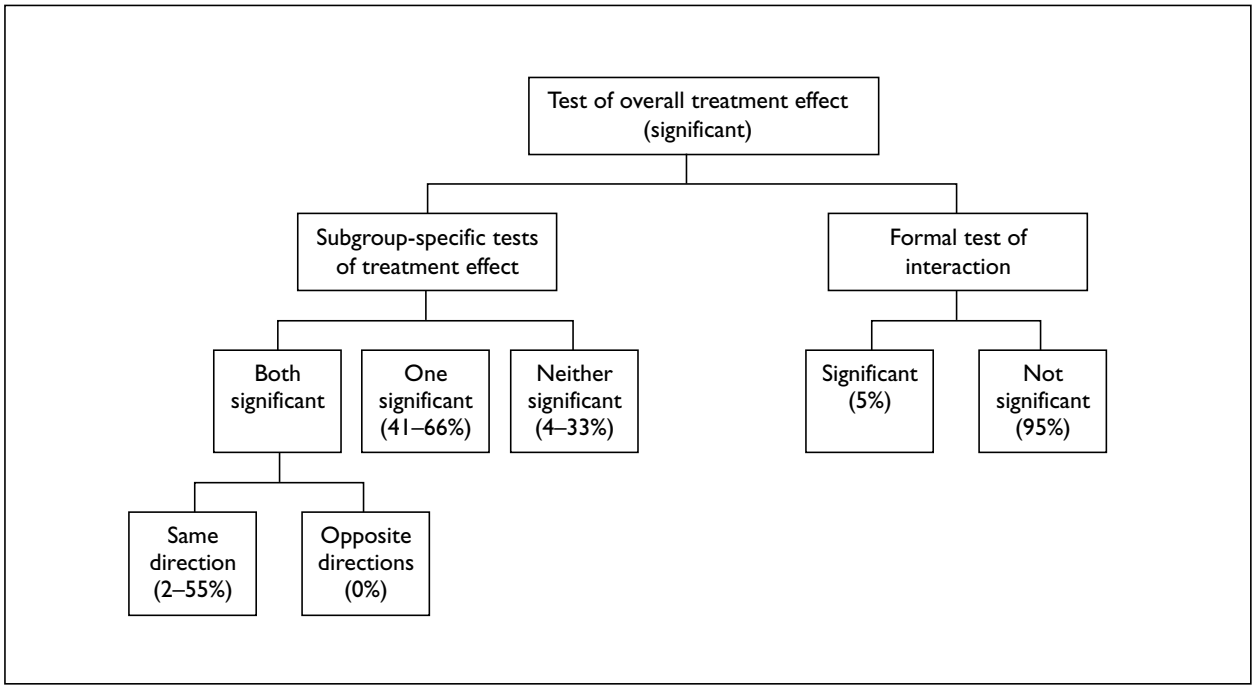


FIGURE 22 *Summary of results for the simplest case (overall test result significant). This figure combines the results from data simulated with no overall treatment effect and with a true overall treatment effect detectable at nominal powers of 50, 80, 90 and 95%*

Nominal power levels of 50 to 95% were considered for the overall treatment effect, and, in each case, the ability of 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 to occur in practice. This is demonstrated by the percentage of significant interaction tests given in *Figure 23* for an example 80% nominal power for the overall effect. For instance, an interaction of the same magnitude as the overall effect (which is still quite large) has only a 29% chance of being detected.

Also presented in *Figure 23* is the inflation factor as detailed in chapter 3, but here it is given on a log scale for interactions up to twice the magnitude of the overall effect. This is the factor required to increase the sample size to detect the interaction with the same power as the overall effect. As it can be seen, this varied with the size of the interaction but was independent of the power and sample size. For an interaction of the same magnitude as the overall effect, this inflation factor was 4. This increased dramatically to 100 or greater for more subtle interactions that were smaller than 20% of the overall effect.

Varying the trial specifications

Modifying the treatment group ratio had no noticeable effect on either the interaction or subgroup-specific tests, regardless of the overall test result. Although altering the subgroup ratio did not affect the performance of the interaction test, it did have an impact on the subgroup-specific test results. Specifically, as the ratio became more extreme the percentage of subgroup-specific tests where only one subgroup was significant increased among those with a significant overall test result.

Increasing the variance in one treatment group did not alter the findings. In contrast, a larger variance within one subgroup (much more likely in practice for an RCT) marginally reduced the percentage of significant interaction tests when the overall test was correctly non-significant or correctly significant and dramatically increased the percentage when the overall test was incorrectly non-significant or incorrectly significant. When the overall test was correctly non-significant, the results of subgroup-specific tests were not altered (apart from greater instability across sample sizes for relatively extreme differences in variances). When the overall test result was significant or incorrectly non-significant, the percentage of subgroup-specific analyses where only one subgroup was significant increased marginally with increasing differences in (subgroup-specific) variances.

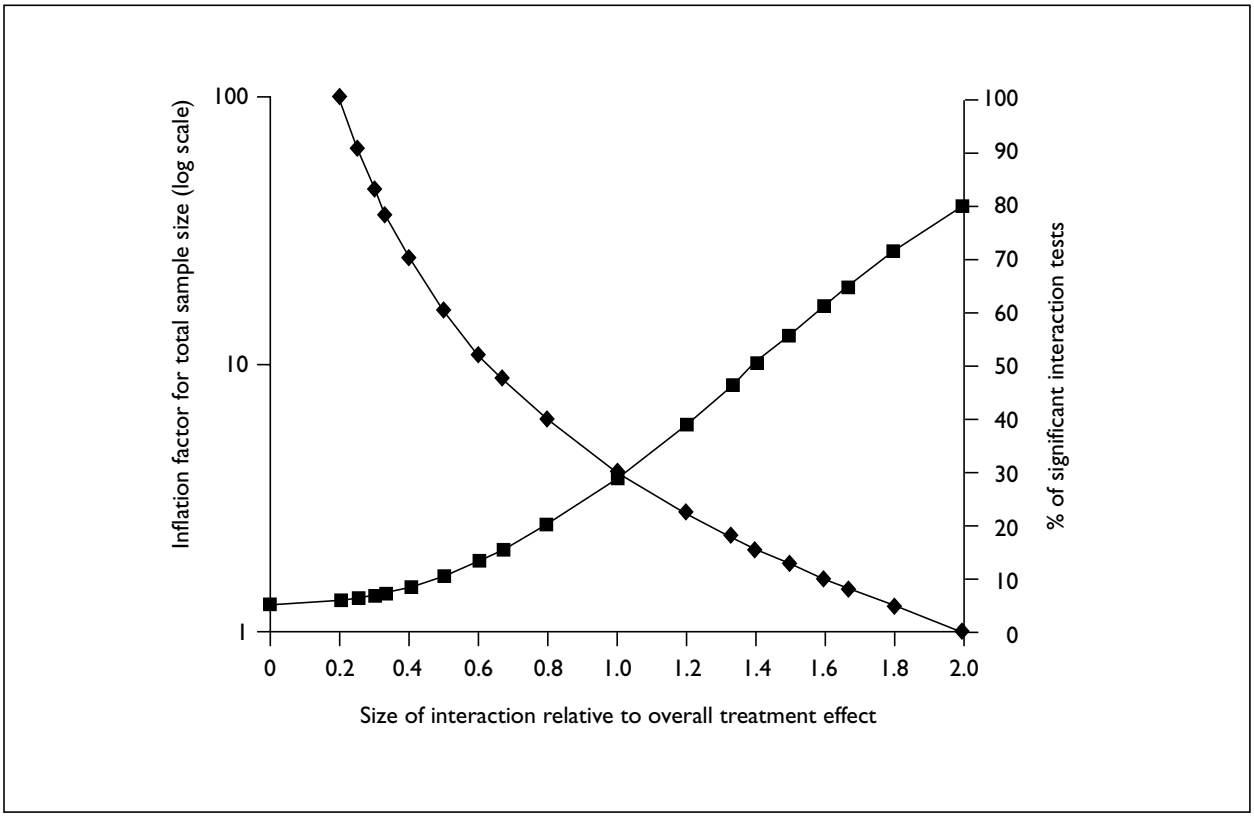


FIGURE 23 *Observed power for the interaction tests and the inflation factors required to increase sample sizes so that interaction tests have the same power as that for the overall treatment effect.* ◆*, Inflation factor;* ■*, % of significant interaction tests*

As the number of subgroups from a potential effect-modifier increased, the interaction test was unaffected. For subgroup-specific tests, the percentage finding one or more but not all subgroups significant increased with increasing numbers of subgroups.

How realistic are the scenarios covered by the simulations?

The scenarios covered by the simplest case (two equal-sized treatment groups and two equal-sized subgroups) and that of varying the subgroup ratio are perhaps the most realistic in the sense of being most likely to occur in practice. The only exception to this is possibly the assumption in these scenarios of equal variability, although, in practice, this is often reasonable even if transformations are necessary for it to hold. Unequal treatment (randomisation) groups may also occur in practice, but this did not affect the findings of subgroup analyses, either in the form of interaction or subgroup-specific tests in this study.

In terms of the extent of the variations applied to both the treatment group and subgroup ratios, altering these from 1:1 to 1:5 would certainly cover all practical situations with respect to the randomisation ratio. Moreover, changing the population in one of the subgroups from 50% to about 15% would cover most situations in which subgroup analyses would be considered. Likewise, including scenarios with between two and five subgroups covers most practical situations.

By definition, in the context of RCTs, differential variability is much more likely across subgroups than across the (randomised) treatment groups. In other study designs, both subgroup and treatment group variability may alter but only subgroup variability appears to affect the results of subgroup analyses. The alterations considered (up to a 25-fold ratio across treatment group/subgroup category) were quite extreme and much more than is likely in practice, and the results presented here should certainly encompass the 'worst-case' scenario. In practice, researchers might attempt to ameliorate such differences in variability using appropriate transformations, or alternative statistical methods may be implemented that do not make assumptions of equal variance.

Further investigations

While the coverage of the various scenarios employed for the above simulations has been both realistic and reasonably comprehensive (particularly for continuous outcomes), a number of further investigations beyond the immediate project would seem to be worthwhile. As already stated, not all of these extensions would be expected to have a major influence on the findings and, if anything, those already considered are likely to be best-case situations. The possibilities for further research are discussed in chapter 7.

Chapter 7
Recommendations

It is acknowledged that the following recommendations are a combination of established (prior) views on subgroup analyses and the findings from the simulations presented in this report. However, the main contribution of the current project has been to quantify the risks involved for different approaches and the summary of results in chapter 6 should be disseminated widely amongst trialists, readers of published trials and those attending courses on trial design and analysis. The conclusions are summarised under the headings of implications for design, analysis, presentation and interpretation.

Recommendations for future research

Existing scenarios could be covered even more comprehensively

As discussed briefly in the alternative analytical methods section, other methods of statistical data analysis could be considered in the same way as the three types of regression models covered here (namely multiple regression, logistic regression and Cox's proportional hazards regression). Examples of extensions are chi-squared tests for differences in proportions, tests for trends across ordered categories defined by the subgroups, Mantel–Haenzel tests and the log-rank test, but major changes to the findings in this report would not be anticipated in these cases. Clustering effects could also be incorporated, where the assumptions of independent observations within the simulated samples are modified.

In addition to the consideration of under- and over-dispersion, entirely different distributions for the outcome variables could be considered, including relaxing the imposed (exactly Gaussian) distribution for continuous outcomes, considering non-parametric tests and generalising the distribution for survival times to the Weibull distribution (of which the exponential distribution, covered here, is a special case) with survival outcome data. In the Weibull distribution, there is an additional (shape) parameter (equal to one in the case of the exponential). Concomitantly, the assumption for the Weibull case is that the hazard function increases or decreases monotonically through time, whereas that for the exponential is constant over time.

Finally, other types of outcome variable could be considered although, again, there is no reason to suppose that the results would differ markedly from those presented here. Two obvious additional outcomes are (a) counts of an outcome event for which Poisson regression would be appropriate or (b) categorical outcomes with more than two levels for which either the proportional odds or multinomial regression would be appropriate, depending on whether or not the (three or more) categories were ordered.

The same general approach could be applied to a wider set of contexts

The most obvious extension would be to cover observational studies by relaxing the 'allocation' of treatment groups from the presumption of randomisation involved in this project. While this is covered structurally by the simulations presented here, there is the additional complication of confounding. If residual confounding were low, either by chance, design (for instance, matching/stratification) and/or analysis (multivariable adjustment for confounders) then very similar results would be expected. However, this is unlikely in practise and the problem of unknown residual confounding is likely to remain. Where residual confounding is high, it would be expected that the situation would be worse than that for the RCT since any (observed) differential effects could be a consequence of (residual) confounders not taken into account in the analysis.

Secondly, the approach could be extended to quantitative synthesis of research evidence and, in particular, to meta-analyses of RCTs. This is especially worthwhile because it is often considered that there is more scope for subgroup analyses in the context of meta-analyses, since the available sample sizes are considerably greater. Specific issues that might be addressed are 'stratified' analyses where subgroups of subjects are analysed separately across trials; analyses of subsets of trials according to global assessments, such as trial quality; and the optimal use of meta-regression techniques.

An alternative general approach could be adopted

For the reasons given in the simulation strategy section, the present study concentrated on *p*-values rather than CIs. Further work could, therefore, redress this balance with more attention paid to CIs. In particular, attention could focus on their widths and the degree to which subgroup-specific intervals overlap. However, the identification of differential effects based on CIs is likely to have even more scope for misinterpretation than that based solely on *p*-values, and the results presented here are, therefore, likely to represent best-case scenarios. There are, of course, other approaches that could be adopted, such as Bootstrap simulations and Bayesian methods, although it is unlikely that the fundamental conclusions would alter for any of these alternative approaches or further investigations.

Implications for study design

- If possible, the study should be powered to cater for the subgroup analyses to be carried out, using *Figure 23* for instance.
- Ideally, subgroup analyses should be restricted to those proposed in advance of any data analysis, and the choices should be based on clinical interest and previous findings to avoid data dredging.
- The study should have as few subgroups within a potential effect-modifier as possible while continuing to make clinical sense.

Implications for data analysis

- Only formal tests of interaction should be performed – subgroup-specific tests should be avoided.
- The results portrayed in *Figures 21* and *22* should be disseminated widely across trialists and be appreciated generally, specifically in scenarios with no (true) differential effects. These results indicate that:
  - subgroup-specific tests are anti-conservative
  - the extent of this problem (i.e. where a differential effect may be incorrectly presumed) is exaggerated in situations with a significant overall treatment effect (in these cases, the treatment effect will be significant in just one subgroup up to 66% of the time)
  - even with a non-significant overall effect, just one subgroup will be significant in up to about 26% of occasions.
- Any differences in variability across subgroups should be investigated. If these are seen then attempts should be made to ameliorate them by appropriate transformations. If this is not possible then extreme caution should be used when interpreting the results of any test of differential effects. (At the same time, it should be recognised that any transformation of scale, such as the log, may affect the existence of differential effects.)

Implications for the presentation of subgroup analyses

- Any lack of differential effect should be interpreted with caution unless the trial was specifically powered with interactions in mind or the differences are expected to be substantial.
- Subgroup-specific effects and CIs can be helpful in interpreting differential effects but these should only be used after a formal test of interaction.
- Clear distinction should be made between subgroup analyses defined in advance and those identified, for whatever reason, once the main trial analyses have been performed.
- The findings of any subgroup analyses should not be over-emphasised. Unless there is a strong prior hypothesis for a given differential effect, any findings might be best viewed in the context of a hypothesis-generation exercise.
- In the presentation of RCTs, emphasis should almost always remain on the overall treatment effect rather than the subgroup analyses.

Interpretation of published subgroup analyses

- Published subgroup analyses should be interpreted with caution, especially those not accompanied by a formal test of interaction. Spurious apparent differential effects can be very common for subgroup-specific tests, especially if the overall test result is significant.
- Any subgroup analyses that have not been clearly proposed in advance should be viewed with extreme caution.
- Unless the study has been specifically powered to detect interactions, lack of statistical significance for interaction tests is a far from secure way of excluding differential effects (*Figure 9*).
- All subgroup effects should be interpreted in the full context of the literature with respect to both corroboration and biological plausibility.

Acknowledgements

This study was commissioned by the NHS R&D HTA Programme. The authors are very grateful to the referees for their constructive and helpful comments. The views expressed in this report are those of the authors, who are also responsible for any errors.

References

1. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med* 1987;**317**:426–32.
2. 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.
3. 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;**2**:349–60.
4. Mauri F, Gasparini M, Barbonaglia L, Santoro E, Grazia Franzosi M, Tognoni G, *et al.* Prognostic significance of the extent of myocardial injury in acute myocardial infarction treated by streptokinase (the GISSI trial). *Am J Cardiol* 1989;**63**:1291–5.
5. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *J Clin Epidemiol* 1996;**49**:395–400.
6. Senn S, Harrell F. On wisdom after the event. *J Clin Epidemiol* 1997;**50**:749–51.
7. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. On reaching the tunnel at the end of the light. *J Clin Epidemiol* 1997;**50**:753–5.
8. Davey Smith G, Egger M. Incommunicable knowledge? Interpreting and applying the results of clinical trials and meta-analyses. *J Clin Epidemiol* 1998;**51**:289–95.
9. Feinstein AR. The problem of cogent subgroups: a clinicostatistical tragedy. *J Clin Epidemiol* 1998; **51**:297–9.
10. Shah S, Peat JK, Mazurski EJ, Wang H, Sindhusake D, Bruce C, *et al.* Effect of peer led programme for asthma education in adolescents: cluster randomised controlled trial. *BMJ* 2001; **322**:583–5.
11. Peters TJ, Somerset M, Baxter K, Wilkinson C. Anxiety among women with mild dyskaryosis: a randomized trial of an educational intervention. *Br J Gen Pract* 1999;**49**:348–52.
12. Gruppo Italiano per lo Studio della Streptochinasi Nell Infarcto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet* 1986;**2**:397–402.
13. Byar DP. Assessing apparent treatment-covariate interactions in randomized clinical trials. *Stat Med* 1985;**4**:255–63.
14. Peters TJ. The design and analysis of randomized controlled trials of treatments for lower urinary tract symptoms. *BJU Int* 2000;**85**:3–9.
15. Moher D, Schulz KF, Altman D for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *JAMA* 2001;**285**:1987–91.
16. Senn S. Statistical issues in drug development. Chichester: Wiley; 1997.
17. Rockette HE, Caplan RJ. Strategies for subgroup analysis in clinical trials. *Recent Results Cancer Res* 1988;**111**:49–54.
18. Egger M, Davey Smith G. Bias in location and selection of studies. *BMJ* 1998;**316**:61–6.
19. Stallones RA. The use and abuse of subgroup analysis in epidemiological research. *Prev Med* 1987;**16**:183–94.
20. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med* 1990;**9**:657–71.
21. Bulpitt CJ. Subgroup analysis. *Lancet* 1988;**2**:31–4.
22. Oxman AD, Guyatt GH. A consumers guide to subgroup analyses. *Ann Intern Med* 1992;**116**:78–84.
23. Sterne JAC, Davey Smith G. Sifting the evidence – what's wrong with significance tests? *BMJ* 2001; **322**:226–31.
24. Hogg RV, Tanis EA. Probability and statistical inference. 3rd ed. New York: Macmillan; 1983.
25. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984;**13**:356–65.
26. Lubin JH, Gail MH. On power and sample size for studying features of the relative odds of disease. *Am J Epidemiol* 1990;**131**:552–66.
27. Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene-environmental interactions: comments on different approaches. *Am J Epidemiol* 1999;**149**:689–93.
28. Machin D, Campbell M, Fayers P, Pinol A. Sample size tables for clinical studies. 2nd ed. Oxford: Blackwell Science; 1997.
29. Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford: Blackwell Science; 1994.
30. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1996.
31. Elashoff JD. nQuery advisor version 3.0 user's guide. Los Angeles: nQuery; 1999.

Appendix 1

Generation and analysis of data

Continuous outcome data

Generation of data

The Box-Muller transformation[24] generates a variable from the Gaussian distribution by transformation of a random variable from the uniform distribution U (0,1). The random variable Y has a Gaussian distribution if its probability distribution function is defined by:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \quad -\infty < y < \infty$$

where μ (mean) and σ (standard deviation) are parameters satisfying $-\infty < \mu < \infty$ and $0 < \sigma < \infty$. Briefly we say that Y is $N(\mu, \sigma^2)$.

Methods of analyses

The regression models used to perform the subgroup-specific tests of treatment effect and the formal test of interaction are detailed in *Tables 30* and *31*.

Calculating treatment effect differences

Equal-sized treatment groups

In order to ascertain what treatment effect difference could be detected for different sample sizes at different powers, the following formula[28] was used (fixing sample size):

$$n_i = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2}{\delta^2} = \frac{z_{1-\alpha/2}}{4} \quad i = 1, 2$$

where n_i = size of each of the two treatment groups (assuming the two treatment groups are of equal size); $z_{1-\alpha/2}$ = significance level; $z_{1-\beta}$ = power; δ = clinically relevant treatment effect difference; σ = standard deviation (assumed to be the same in each treatment group, unless otherwise stated, in which instance a pooled estimate[29] was used).

TABLE 30 *Main effect one-way ANOVA model – subgroup-specific treatment effect tests – ($y_{ik} = \beta_0 + \beta_i + \varepsilon_{ijk}$)*

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS) = SS/df	F-statistics
Due to treatment	$\sum_i n_i(\overline{y}_{i.} - \overline{y})^2$	$r - 1$	MS (treatment)	$\frac{\text{MS (treatment)}}{\text{MSE}}$
Residual	$\sum_{ik} (\overline{y}_{ik} - \overline{y}_{i.})^2$	$n - r$	MSE	
Total	$\sum_{ik} (\overline{y}_{ik} - \overline{y})^2$	$n - 1$		

SS, sum of squares; df, degrees of freedom; MS, mean square; MSE, mean square error

TABLE 31 *Interaction two-way ANOVA model – formal test of interaction – ($y_{ijk} = \beta_0 + \beta_i + \beta_j + \beta_{ij} + \varepsilon_{ijk}$)*

Source of variation	SS	df	MS = SS/df	F-statistics
Due to treatment	$\sum_i n_{i.}(\overline{y}_{i..} - \overline{y}_{...})^2$	$r - 1$	MS (treatment)	$\frac{\text{MS (treatment)}}{\text{MSE}}$
Due to subgroup	$\sum_j n_{.j}(\overline{y}_{.j.} - \overline{y}_{...})^2$	$c - 1$	MS (subgroup)	$\frac{\text{MS (subgroup)}}{\text{MSE}}$
Due to interaction between treatment and subgroup	$\sum_{ij} n_{ij}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$	$(r - 1)(c - 1)$	MS (treatment x subgroup)	$\frac{\text{MS (treatment x subgroup)}}{\text{MSE}}$
Residual	$\sum_{ijk} (\overline{y}_{ijk} - \overline{y}_{ij.})^2$	$n_{..} - rc$	MSE	
Total	$\sum_{ijk} (\overline{y}_{ijk} - \overline{y}_{...})^2$	$n_{..} - 1$		

Unequal-sized treatment groups

When varying the treatment group ratio to give two unequal-sized groups, an adjustment is necessary.[30] Firstly, N was calculated as $n_1 + n_2$, using the above formula, assuming equal-sized treatment groups. A modified sample size N' was then calculated.

If $k = n_1/n_2$ is the ratio of the sample sizes in the two groups, then the required total sample size is:

$$N' = \frac{N(1+k)^2}{4k}$$

and the two sample sizes are given by $N' / (1 + k)$ and $kN' / (1 + k)$.

In the simulations varying treatment group ratio, N' remains the same as N whilst the ratio k varies between 2 and 5. Thus to calculate the adjusted treatment effect difference, a new N was calculated from N':

$$N = \frac{4kN'}{(1+k)^2}$$

Then $N/2$ is implemented as n in:

$$n_i = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2}{\delta^2} = \frac{z_{1-\alpha/2}}{4} \quad i = 1, 2$$

Binary outcome data

Generation of data

The random variable Y has a binomial distribution if its probability distribution function is defined by:

$$f(y) = \frac{n!}{y!\,(n-y)!} p^y(1-p)^{n-y},$$

$y = 0, 1, 2, ..., n$ = number of events

We say that y is $b\,(n, p)$, and each individual event, say x, comes from a Bernoulli distribution $b\,(1, p)$ and has the value 0 or 1 depending on whether an event has occurred or not.

$$f(x) = p^x(1-p)^{(1-x)}, \quad x = 0, 1$$

From n_{ij} random variates with a uniform distribution U (0, 1), n_{ij} $b\,(1, p_{ij})$ variates are then generated according to whether or not they exceed the specified p_{ij} for that category.

Methods of analyses

The likelihood function of Y from the binomial distribution is:

$$L = \prod_{i=1}^{y} p_i \prod_{i=y+1}^{n} (1 - p_i)$$

The logistic regression model, including a term for treatment, subgroup and interaction, can be written as:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

We can abbreviate the right-hand side of this model to $\alpha + \underline{\beta x}_i$ and it can be shown that:

$$p = \frac{e^{(\alpha + \underline{\beta x}_i)}}{1 + e^{(\alpha + \underline{\beta x}_i)}}$$

and $1 - p = \dfrac{1}{1 + e^{(\alpha + \underline{\beta x}_i)}}$

Therefore:

$$L = \prod_{i=1}^{y} \left(\frac{e^{(\alpha + \underline{\beta x}_i)}}{1 + e^{(\alpha + \underline{\beta x}_i)}}\right) \prod_{i=y+1}^{n} \left(\frac{1}{1 + e^{(\alpha + \underline{\beta x}_i)}}\right)$$

Before maximising this function, it is easier to use the log-likelihood function:

$$\log(L) = \sum_{i=1}^{y} (\alpha + \underline{\beta x}_i) - \sum_{i=1}^{y} \log(1 + e^{(\alpha + \underline{\beta x}_i)}) - \sum_{i=y+1}^{n} \log(1 + e^{(\alpha + \underline{\beta x}_i)}),$$

which equates to:

$$\log(L) = \sum_{i=1}^{y} (\alpha + \underline{\beta x}_i) - \sum_{i=1}^{n} \log(1 + e^{(\alpha + \underline{\beta x}_i)})$$

FORTRAN maximises this function. For an overall-effects model and subgroup-specific models, $\alpha + \underline{\beta x}_i$ becomes:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x_{1i}$$

Calculating treatment effect differences

The nQuery Advisor 3.0 software[31] was used to calculate p_2 (probability of an event within treatment group 2) that would lead to an odds ratio detectable with 80% power for different sample sizes (for the majority of cases p_1 was set to equal 0.5). This uses the formula:

$$n = \frac{\left[z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\right]^2}{(p_1-p_2)^2}$$

where the odds ratio $\varphi = \dfrac{p_2(1-p_1)}{p_1(1-p_2)}$

Unequal-sized treatment groups

The same adjustment is made as for the continuous outcome case detailed above.

Survival outcome data

Generation of data

Survival times were generated as detailed in the survival data section of chapter 2, using a mean survival time of 36 months. Data were simulated assuming a follow-up period of 60 months, thus any survival times generated to be greater than 60 months were taken to be censored due to the end of the follow-up period.

Methods of analyses

Maximum likelihood methods were used to maximise the partial log-likelihood function for the Cox-proportional hazards model.

$$\ln(L) = \sum_{j=1}^{D} \{ \sum_{k \in D_j} x_k\beta - d_j \ln [\sum_{i \in R_j} \exp (x_i\beta)]\}$$

where D is the total number of events, D_j is the set of observations that fail at time j (this will contain only one observation unless there are ties), d_j is the number of events at time j and R_j is the set of observations still at risk at time j (the risk pool).

Calculating treatment effect differences

The nQuery Advisor 3.0 software[31] was used to calculate λ_2 (hazard rate for treatment group 2) that would lead to an odds ratio detectable with 80% power for different sample sizes (λ_1 was set to equal 0.0278, that is 1/36 – a mean survival time of 36 months). This uses the formula:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(\varphi+1)^2}{(2-\pi_1-\pi_2)(\varphi-1)^2}$$

where $\varphi = \dfrac{\lambda_1}{\lambda_2}$ = hazard ratio and π_i = proportion experiencing an event by the end of follow-up (assumed in the simulations to be 60 months).

Methodology Group

Members

Methodology Programme Director
Professor Richard Lilford
Director of Research and Development
NHS Executive – West Midlands, Birmingham

Chair
Professor Martin Buxton
Director, Health Economics Research Group
Brunel University, Uxbridge

Professor Douglas Altman
Professor of Statistics in Medicine
University of Oxford

Dr David Armstrong
Reader in Sociology as Applied to Medicine
King's College, London

Professor Nicholas Black
Professor of Health Services Research
London School of Hygiene & Tropical Medicine

Professor Ann Bowling
Professor of Health Services Research
University College London Medical School

Professor David Chadwick
Professor of Neurology
The Walton Centre for Neurology & Neurosurgery
Liverpool

Dr Mike Clarke
Associate Director (Research)
UK Cochrane Centre, Oxford

Professor Paul Dieppe
Director, MRC Health Services Research Centre
University of Bristol

Professor Michael Drummond
Director, Centre for Health Economics
University of York

Dr Vikki Entwistle
Senior Research Fellow, Health Services Research Unit
University of Aberdeen

Professor Ewan B Ferlie
Professor of Public Services Management
Imperial College, London

Professor Ray Fitzpatrick
Professor of Public Health & Primary Care
University of Oxford

Dr Naomi Fulop
Deputy Director, Service Delivery & Organisation Programme
London School of Hygiene & Tropical Medicine

Mrs Jenny Griffin
Head, Policy Research Programme
Department of Health
London

Professor Jeremy Grimshaw
Programme Director
Health Services Research Unit
University of Aberdeen

Professor Stephen Harrison
Professor of Social Policy
University of Manchester

Mr John Henderson
Economic Advisor
Department of Health, London

Professor Theresa Marteau
Director, Psychology & Genetics Research Group
Guy's, King's & St Thomas's School of Medicine, London

Dr Henry McQuay
Clinical Reader in Pain Relief
University of Oxford

Dr Nick Payne
Consultant Senior Lecturer in Public Health Medicine
ScHARR
University of Sheffield

Professor Joy Townsend
Director, Centre for Research in Primary & Community Care
University of Hertfordshire

Professor Kent Woods
Director, NHS HTA Programme, & Professor of Therapeutics
University of Leicester

continued

HTA Commissioning Board

Members

Programme Director
Professor Kent Woods
Director, NHS HTA Programme, &
Professor of Therapeutics
University of Leicester

Chair
Professor Shah Ebrahim
Professor of Epidemiology of Ageing
University of Bristol

Deputy Chair
Professor Jon Nicholl
Director, Medical Care Research Unit
University of Sheffield

Professor Douglas Altman
Director, ICRF Medical Statistics Group
University of Oxford

Professor John Bond
Director, Centre for Health Services Research
University of Newcastle-upon-Tyne

Ms Christine Clark
Freelance Medical Writer
Bury, Lancs

Professor Martin Eccles
Professor of Clinical Effectiveness
University of Newcastle-upon-Tyne

Dr Andrew Farmer
General Practitioner &
NHS R&D Clinical Scientist
Institute of Health Sciences
University of Oxford

Professor Adrian Grant
Director, Health Services Research Unit
University of Aberdeen

Dr Alastair Gray
Director, Health Economics Research Centre
Institute of Health Sciences
University of Oxford

Professor Mark Haggard
Director, MRC Institute of Hearing Research
University of Nottingham

Professor Jenny Hewison
Senior Lecturer
School of Psychology
University of Leeds

Professor Alison Kitson
Director, Royal College of Nursing Institute, London

Dr Donna Lamping
Head, Health Services Research Unit
London School of Hygiene & Tropical Medicine

Professor David Neal
Professor of Surgery
University of Newcastle-upon-Tyne

Professor Gillian Parker
Nuffield Professor of Community Care
University of Leicester

Dr Tim Peters
Reader in Medical Statistics
University of Bristol

Professor Martin Severs
Professor in Elderly Health Care
University of Portsmouth

Dr Sarah Stewart-Brown
Director, Health Services Research Unit
University of Oxford

Professor Ala Szczepura
Director, Centre for Health Services Studies
University of Warwick

Dr Gillian Vivian
Consultant in Nuclear Medicine & Radiology
Royal Cornwall Hospitals Trust
Truro

Professor Graham Watt
Department of General Practice
University of Glasgow

Dr Jeremy Wyatt
Senior Fellow
Health Knowledge Management Centre
University College London

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

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