BMJ

BMJ 2012;344:e1553 doi: 10.1136/bmj.e1553 (Published 15 March 2012)

RESEARCH

Credibility of claims of subgroup effects in randomised controlled trials: systematic review

Xin Sun *assistant professor*[1 2], Matthias Briel *assistant professor*[2 3], Jason W Busse *scientist*[2 4], John J You *assistant professor*[2 5], Elie A Akl *associate professor*[2 6], Filip Mejza *research fellow*[7], Malgorzata M Bala *research fellow*[8], Dirk Bassler *associate professor*[9], Dominik Mertz *assistant professor*[2 5 10], Natalia Diaz-Granados *doctoral candidate*[2], Per Olav Vandvik *researcher*[11 12], German Malaga *associate professor*[13], Sadeesh K Srinathan *assistant professor*[14], Philipp Dahm *professor*[15], Bradley C Johnston *assistant professor*[2 16], Pablo Alonso-Coello *researcher*[17], Basil Hassouneh *research fellow*[2], Stephen D Walter *professor*[2], Diane Heels-Ansdell *statistician*[2], Neera Bhatnagar *librarian*[18], Douglas G Altman *professor*[19], Gordon H Guyatt *professor*[2]

[1]Center for Health Research, Kaiser Permanente Northwest, Portland, OR, USA; [2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, 1280 Main Street West, Hamilton, ON, Canada, L8S 4K1; [3]Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Switzerland; [4]Institute for Work and Health, Toronto, ON, Canada; [5]Department of Medicine, McMaster University, Hamilton, ON, Canada; [6]Departments of Medicine and Family Medicine, State University of New York at Buffalo, NY, USA; [7]Department of Pulmonary Diseases, Jagiellonian University School of Medicine, Krakow, Poland; [8]II Department of Internal Medicine, Jagiellonian University School of Medicine, Krakow, Poland; [9]Department of Neonatology and Center for Pediatric Clinical Studies, University Children's Hospital Tuebingen, Tuebingen, Germany; [10]Michael G DeGroote Institute for Infectious Diseases Research, McMaster University, Hamilton, Canada; [11]Norwegian Knowledge Centre for the Health Services, Oslo, Norway; [12]Department of Medicine, Innlandet Hospital Trust, Gjøvik, Norway; [13]Department of Medicine, Universidad Peruana Cayetano Heredia, Lima, Peru; [14]Department of Surgery, University of Manitoba, Winnipeg, MB, Canada; [15]Department of Urology, University of Florida, College of Medicine, Gainesville, FL, USA; [16]Department of Anaesthesiology and Pain Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada; [17]IberoAmerican Cochrane Centre, Biomedical Research Institute-CIBER of Epidemiology and Public Health, Barcelona, Spain.; [18]Health Sciences Library, McMaster University, Hamilton, ON, Canada; [19]Centre for Statistics in Medicine, University of Oxford, Oxford, UK

Abstract

Objective To investigate the credibility of authors' claims of subgroup effects using a representative sample of recently published randomised controlled trials.

Design Systematic review.

Data source Core clinical journals, as defined by the National Library of Medicine, in Medline.

Study selection Randomised controlled trials published in 2007. Using prespecified criteria, teams of trained reviewers independently judged whether authors claimed subgroup effects and the strength of their claims. Reviewers assessed each of these claims against 10 predefined criteria, developed through a search of existing criteria and a consensus process.

Results Of 207 randomised controlled trials reporting subgroup analyses, 64 (31%) made claims for the primary outcome. Of those, 20 were strong claims and 28 claims of a likely effect. Authors included subgroup variables measured at baseline in 60 (94%) trials, used subgroup variable as a stratification factor at randomisation in 13 (20%), clearly prespecified their hypotheses in 26 (41%), correctly prespecified direction in 4 (6%), tested a small number of hypotheses in 28 (44%), carried out a test of interaction that proved statistically significant in 6 (9%), documented replication of a subgroup effect with previous related studies in 21 (33%), identified consistency of a subgroup effect across related outcomes in 19 (30%), and provided a compelling indirect evidence for the effect in 14 (22%). In the 19 trials making more than one claim, only one (5%) checked the independence of the interaction. Of the 64 claims, 54 (84%) met four or fewer of the 10 criteria. For strong claims, more than 50%

Correspondence to: G H Guyatt guyatt@mcmaster.ca

Extra material supplied by the author (see http://www.bmj.com/content/344/bmj.e1553?tab=related#webextra)

Appendix 1: approaches to selecting the primary outcome
Appendix 2: selection of pairwise comparisons
Appendix 3: criteria for judging the strength of a subgroup claim

failed each of the individual criteria, and only three (15%) met more than five criteria.

Conclusion Authors often claim subgroup effects in their trial report. However, the credibility of subgroup effects, even when claims are strong, is usually low. Users of the information should treat claims that fail to meet most criteria with scepticism. Trial researchers should report the conduct of subgroup analyses and provide sufficient evidence when claiming a subgroup effect or suggesting a possible effect.

Introduction

Subgroup analysis in randomised controlled trials seeks to determine whether a treatment effect varies across subgroups defined by patient characteristics. The findings of subgroup analyses offer the promise of individualising patient care, and such analyses are common in randomised controlled trials, with 40-65% reporting them.[1-7] In particular, claims of subgroup effect, in which the authors convey a conviction or belief of a difference in treatment effects between patient subgroups, can have a substantial impact on clinical practice and policy decision. One study found that in 35 randomised controlled trials published in top general medical journals reporting subgroup analyses, 21 (60%) claimed subgroup effects.[1]

Clinical action based on a credible subgroup effect may enhance benefits and avoid harms and unnecessary use of health resources. The authors of trial reports, however, often do not prespecify hypotheses for subgroups, fail to carry out a statistical test for interaction, and undertake a large number of subgroup analyses.[1 5 8] Given these limitations, it is perhaps not surprising that many inferences from subgroup analyses have proved spurious.[9] Misguided claims of subgroup effects may result in patients being denied potentially beneficial care or receiving potentially harmful treatment. Industry funded trials may be at greater risk of misleading results: when the main results are not statistically significant, industry funded trials are more likely than non-industry funded trials to report subgroup analyses.[10] Furthermore, industry funded trials less often prespecify subgroup hypotheses and use the test of interaction, regardless of the significance of the main effect.[10]

The credibility of a putative subgroup effect infers that a difference in treatment effect between subgroups is real, and reflects a continuum ranging from extremely unlikely to highly plausible. Since the 1990s much work has been done on documenting the limitations of subgroup analysis and in developing criteria to guide clinicians, scientists, and health policy makers in making appropriate inferences about their credibility.[8 11-14] No studies, however, have systematically examined a representative sample of trials on the extent to which claims of subgroup effects adhere to those standards. We carried out a systematic review of randomised controlled trials to determine the frequency with which authors claim subgroup effects and the extent to which their claims of subgroup effects are consistent with existing criteria.

Methods

We have previously published our study protocol detailing the design and analysis,[15] and the results on reporting of subgroup analyses.[10] Briefly, we included any randomised controlled trial carried out in humans unless it focused on a subset of the original population enrolled (that is, only one subgroup was examined), was explicitly labelled as a phase I trial, was exclusively a pharmacokinetic study, or was reported as a research letter.

We applied a sensitive search strategy,[15] combining both MeSH terms and free texts, to the core clinical journals in 2007 in Medline using the Ovid interface. The core clinical journals defined by the National Library of Medicine, known as the *Abridged Index Medicus*, included 118 journals in 2007, covering all specialties of clinical medicine and public health sciences.[16] After removing duplicate articles, our search identified 3662 journal reports.

Sampling and data collection

We stratified journals into higher impact groups (*Annals of Internal Medicine*, *BMJ*, *Journal of the American Medical Association*, *Lancet*, and *New England Journal of Medicine*) and lower impact groups according to the total citations in 2007 defined by the Web of Science.[17] We randomly sampled, in a 1:1 ratio, study reports from higher and lower impact journals.

Eight pairs of reviewers trained in health research methodology used standardised, pilot tested forms with detailed written instructions to independently screen titles, abstracts, and full texts and to extract data.[15] Reviewers resolved discrepancies by consensus or, if any remained, through discussion with one of two arbitrators (XS, GHG). Reviewers conducted calibration exercises to ensure consistency. At the screening stage of the full text, the reviewers selected one primary outcome for each eligible study using prespecified criteria (see supplementary appendix 1), and identified one pairwise comparison if the studies included three or more study arms (see supplementary appendix 2).

For each eligible study, reviewers extracted data on study characteristics, including study sample size, number of study arms, funding sources, clinical area, and type of intervention. They also determined whether results for the primary outcome were statistically significant (P<0.05). Reviewers recorded whether authors reported subgroup analyses and claimed subgroup effects, and the number of claims of a subgroup effect for both any outcome and specifically for the primary outcome. For each study claiming a subgroup effect, reviewers judged the strength of the authors' claim by using prespecified criteria.

Strength of a subgroup claim

We classified articles as making a subgroup claim if, in the abstract or the discussion section of the trial report, the authors stated that the effects of intervention differed, or may have differed, according to the status of a subgroup variable. We used prespecified criteria (see supplementary appendix 3) to classify the strength of a claim into three categories: strong claim (the authors convey a conviction that the subgroup effect truly exists), claim of likely effect (the authors convey a belief that the subgroup effect possibly exists), and claim of a suggestion of possible effect (the authors suggest a subgroup effect but convey uncertainty about whether such an effect exists).

Criteria for assessing the credibility of a subgroup claim

The criteria used for this study were primarily based on the 11 used to assess the credibility of subgroup effect recently published in this journal.[14] Through a search of Medline we also identified published methodological articles dealing with the conduct and interpretation of subgroup analyses,[1 4 5 8 11-14 18-22] and noted additional potentially relevant items from these articles. The study group then discussed the merit of each item and reached consensus on the 10 most important criteria (box).

These 10 criteria consist of nine from the previously published criteria[14] and one criterion (was the subgroup variable a stratification factor at randomisation?) identified from other methodological discussions (box). Of the two criteria we

Ten criteria used to assess credibility of subgroup effect

Design

- Was the subgroup variable a baseline characteristic?
- Was the subgroup variable a stratification factor at randomisation?*
- Was the subgroup hypothesis specified a priori?
- Was the subgroup analysis one of a small number of subgroup hypotheses tested (≤5)?

Analysis

- Was the test of interaction significant (interaction P<0.05)?
- Was the significant interaction effect independent, if there were multiple significant interactions?

Context

- Was the direction of subgroup effect correctly prespecified?
- Was the subgroup effect consistent with evidence from previous related studies?
- Was the subgroup effect consistent across related outcomes?
- Was there any indirect evidence to support the apparent subgroup effect—for example, biological rationale, laboratory tests, animal studies?

*Item was not included in our previously published list of criteria for subgroup credibility

removed from the originally published 11, one is applicable to meta-analyses only (is this a between study or a within study comparison?) and we viewed the other (is the magnitude of subgroup effect large?) as overlapping excessively with the likelihood that chance explains the difference in apparent effects and too specific to the outcome and clinical condition for application in this context.

Using detailed written instructions, the reviewers evaluated whether a subgroup claim for the primary outcome met each criterion. Where possible, reviewers also recorded the P value of the interaction test and, for each subgroup, the numbers of events and participants, the effect measure, point estimate and associated 95% confidence interval, and P value.

Sample size

Our sample size calculation was based on one of our study objectives that examined, in a multivariable regression analysis, the association of six study characteristics (independent variables) comprising nine categories with claim of subgroup effects for any outcome.[15] Setting a criterion of 10 events (that is, claim of a subgroup effect) for each category resulted in a required total of 90 events. Our pilot study suggested that we required a total of 464 trials.[7 15]

Data analysis

For subgroup claims about the primary outcome, we calculated the proportion of claims meeting each criterion and the number of criteria met by each claim. Using the test for trend we examined whether stronger claims met more criteria. We used the χ^2 test or Fisher's exact test to compare, between claims in higher versus lower impact journals, the proportions meeting each of the criteria, and the rank sum test for the number of criteria met by each subgroup.

Several trials made multiple claims of subgroup effects for the primary outcome. To avoid a possible clustering effect, we selected the strongest subgroup claim from each trial; if two or more claims had the same strength of claim, we randomly chose one.

For trials that failed to report interaction P values of subgroup analyses, we calculated the interaction P values if authors reported in both subgroups the mean and any of the following dispersion parameters: 95% confidence intervals, standard error, standard deviation, and number of participants, or the P value for a continuous outcome; the number of events or participants of experimental and control groups, or the point estimate and 95% confidence interval or P value for a binary outcome; or hazard ratio and the associated 95% confidence interval or P value for a time to event outcome. We applied the method of Altman and Bland to calculate the interaction P value.[23]

Results

Of 469 included trials, 207 reported subgroup analyses (figure⇓). Of those 207 studies, 83 (41%) claimed subgroup effects and 64 (31%) claimed a subgroup effect for the primary outcome. The inter-rater agreement on determining subgroup claim was high (κ=0.82, raw agreement=0.92). Table 1⇓ shows that study characteristics of trials that did and did not claim subgroup effects were similar.

Among the 83 trials that claimed a subgroup effect, 46 (56%) made one claim, 22 (26%) made two claims, and 15 (18%) made three or more claims. Thirty three trials (40%) made strong claims, 30 (36%) claimed a likely effect, and 20 (24%) suggested a possible effect. In the 64 trials that claimed a subgroup effect for the primary outcome, 45 (70%) made one claim, 11 (17%) made two claims, and 8 (13%) made three or more claims. Twenty trials (31%) made strong claims, 28 (43%) made a claim of a likely effect, and 16 (25%) made a claim of a suggestion of possible effect.

In the 64 trials that claimed a subgroup effect for the primary outcome, the only criterion usually satisfied was that the subgroup variable was measured at baseline (table 2⇓). All other criteria were satisfied less than 50% of the time, irrespective of the strength of the claim (table 2). Of 32 (50%) studies that reported in the methods that they did a test of interaction, 20 reported the interaction P value or provided information that allowed the P value to be calculated. In the 64 trials that claimed a subgroup effect for the primary outcome, authors reported a median of 6 (interquartile range 3-12) subgroup analyses per trial.

Of the 64 claims, 54 (84%) met four or fewer of the 10 criteria and only five (8%) clearly prespecified the subgroup hypotheses and clearly presented a statistically significant interaction test (P<0.05). A gradient was observed in the number of criteria met by the three categories of subgroup claims—strong (median 3, interquartile range 2-4), likely effect (3, 2-3), and suggestion of a possible effect (2, 2-3), trend test P=0.016. The proportions meeting each of the criteria between claims in high versus lower impact journals did not differ significantly in the proportions

meeting each of the criteria and the number of criteria met by each claim: median 3 (interquartile range 2-4) *v* 3 (2-4), P=0.92.

Discussion

Of 207 representative randomised controlled trials reporting subgroup effects, 64 (30%) claimed a subgroup effect for the primary outcome. Although strong claims were more likely than weaker ones to meet criteria that would justify those claims, even strong claims failed to meet most criteria for a credible subgroup effect (table 2). Thus the credibility of most claimed differences of treatment effect (the subgroup effect) in randomised controlled trials was low.

About two thirds of subgroup analyses associated with claims were not clearly prespecified and were among a large number of subgroup analyses reported (median 6, interquartile range 3-12). Rather than looking for significance in each subgroup separately, investigators should test the hypothesis that effects differ between subgroups. Of 64 randomised controlled trials claiming a subgroup effect, only 32 (50%) reported that they undertook a test of interaction, and of these 32 only 20 reported the interaction P value or information that allowed calculation of the P value.

The test of interaction is a statistical test that examines whether treatment effects, usually measured in relative (for example, risk ratios) or absolute (for example, risk differences) terms, differ among patient populations, such as older versus younger people. The P value of interaction test indicates the probability that, were the true effect the same in all subgroups, differences as large or larger would occur between the observed treatment effects among patient subgroups on multiple repetitions of the experiment, by chance. As an example, one randomised trial[24] tested whether the sequential versus combined use of fluorouracil plus irinotecan or oxaliplatin improved overall survival in patients with a poor prognosis for advanced colorectal cancer. The overall results showed no statistically significant difference (hazard ratio 1.07, 95% confidence interval 0.95 to 1.19). The authors also examined the treatment effects in men and women, and found a non-significant difference of effects between men compared with women (1.02, 0.89 to 1.16 *v* 1.18, 0.97 to 1.44, interaction P=0.21).

Strengths and limitations of the review

Strengths of our review include the identification of a large cohort of representative randomised controlled trials through a systematic search, exploration of subgroup claims in journals of both higher and lower impact, use of standardised forms for screening and data extraction, and calibration exercises to enhance the consistency between reviewers.

Our study has several limitations. Firstly, we did not search all medical journals and therefore our findings may not be applicable to all journals. We did, however, include all core clinical journals, which cover all areas of clinical and public health. Secondly, our results reflect study reports; some authors may not report fully the information about the conduct and results of subgroup analyses, and contextual evidence, such as indirect evidence (often referred to as biological rationale) and direct external evidence. We would argue that for readers to judge the credibility of subgroup claims authors must report detailed information on the conduct and interpretation of subgroup analyses. Thirdly, our criteria for the credibility of a subgroup analysis drew on our own previous work,[14] as well as scholarly inquiry by other authorities. Compelling empirical support for these criteria is, however, lacking. Our inferences about the limitations of current practice would be stronger if consensus was formal or empirical justifications stronger. Our concerns are, however, supported by the compelling logic of the criteria we chose and the considerable consistency among those who have written on the topic.

Implications for trial researchers and journal editors

Subgroup analyses represent an important approach to exploring heterogeneity of treatment effects across patient subgroups in randomised controlled trials. Our study does not suggest that researchers should not undertake subgroup analyses. Rather, they should infer (that is, claim) a subgroup effect only when it meets most of the criteria. The problem with analysing subgroups is not the way the analysis is done but rather the misguided claims of subgroup effects that could result in patients being denied potentially beneficial care or receiving potentially harmful treatment.

Researchers should report fully the conduct of subgroup analyses and provide sufficient evidence when claiming a subgroup effect or suggesting a possible effect. A large proportion of trial reports included in our study failed to report important information about the conduct of the analysis, such as prespecification of hypothesis, and contextual evidence, such as consistency across studies. Ideally, authors making claims about subgroup effects should include a table in their trial report dealing with the credibility criteria and should provide the registered protocol or blinded statistical analysis plan containing details of the plan for subgroup analyses. The consolidated standards of reporting trials statement should consider introducing more detailed guidance about the reporting of subgroup analyses; the current guideline only includes a limited discussion about subgroup analyses.[25]

Subgroup analyses need not meet criteria to be useful: they may also generate new hypotheses.[6] Trial researchers may undertake retrospective subgroup analyses, and the number of subgroup analyses in such analyses may be large. Researchers should, however, alert their readers to view the exploratory nature of these analyses and should explicitly state that the findings should not guide clinical practice. The likelihood of misinterpretation would be further reduced by omitting such findings from the statement of conclusion in the discussion or abstract.

For instance, one randomised trial[26] examined the effect of supportive-expressive group therapy compared with usual care on survival in patients with metastatic breast cancer. In a retrospective hypothesis they suggested that the intervention would show a smaller effect in patients who were positive for the oestrogen receptor and were likely respond to hormonal therapy, and a larger effect in patients who were negative for the oestrogen receptor. In a regression analysis including interaction terms, the researchers found a statistically significant interaction between treatment and oestrogen receptor status (difference in median survival between treatment versus control 11.8 months in patients positive for the receptor versus 20.5 months in patients negative for the receptor, interaction P=0.002). Although this analysis met six criteria, including the use of a baseline characteristic, significant interaction, independence of the interaction, testing for a small number of hypotheses (n=4), supportive indirect evidence, and results consistent with a previous related study, it failed to prespecify subgroup hypotheses, one of the critical criteria, and failed to meet three other criteria. The investigators appropriately indicated that the finding should be treated as hypothesis generating and warrants further study and replication.

In a contrasting example, a randomised controlled trial[27] reported comparing the effect of immobilisation in external rotation versus internal rotation on recurrent dislocation or subluxation in patients with an initial anterior dislocation of the shoulder, and found an absolute risk reduction of 16% (26% external rotation versus 42% internal rotation, P=0.03). Without indicating whether the subgroup hypothesis was prespecified, investigators examined the treatment effect according to age. The investigators found a statistically significant effect in patients aged between 21 and 30 but not in other age groups. They then claimed in the conclusion of the abstract that the intervention effect was greater in this age group (in our classification, a strong claim). This subgroup analysis failed to meet eight out of the 10 criteria, including the prespecification of a hypothesis and demonstration of significant interaction. The credibility of this subgroup finding is low and the investigators' claim misguided.

Implications for users of information from randomised controlled trial reports

Our study suggests that users of information from randomised controlled trial reports, including clinicians, guideline developers, and health policy makers, should be aware that many subgroup claims are not credible and should look for key criteria for credibility being addressed (box).

Unfortunately, for the foreseeable future such users may often find that authors have failed to address the criteria, and the extent to which they do not should be reflected in the judgment of credibility.

In our study, the reasons for a non-credible subgroup claim included: the current unavailability of sufficient evidence (for example, no external studies to suggest consistent effect across studies, or small events to test for a significant interaction test); investigators inappropriately carrying out the analysis, such as test of significance of a subgroup rather than the interaction test; and current evidence disagreeing with the claim made by the authors, such as inconsistency of effect across external studies or the unlikelihood of a significant interaction. Users of randomised controlled trial reports making subgroup claims should be alert to these problems.

Users of reports should be aware that subgroup analyses often produce spurious findings because investigators may identify a chance finding from a large number of subgroup analyses tested, particularly when investigators search for different treatment effects across groups without prespecification. Users should base clinical care on a subgroup effect only if authors make a compelling case for the credibility of their subgroup claim and not when it meets only a small number of criteria. Additionally, users should view subgroup analyses in the context of available systematic reviews. All available evidence should be considered in the interpretation of subgroup results—as implied by the criterion, consistency of subgroup effect across studies, and related outcomes.

Applying subgroup criteria to guide clinical practice

The 10 subgroup criteria may vary in the importance for the credibility of subgroup effect. Among those, we considered the critical criteria to be use of subgroup variables measured at baseline, prespecification of subgroup hypotheses, and statistical significance of interaction test. If these criteria, and most if not all other criteria, are met then acting on the basis of the subgroup analysis may be considered. This is particularly true if controversy exists between competing alternatives that the trial has tested. For instance, a randomised trial[26] that dealt with the controversy over whether reamed or unreamed nailing minimises complications in patients with tibial fractures showed no overall difference between groups. A subgroup analysis meeting critical and most criteria suggested that the reamed approach was superior in closed fractures but the unreamed approach was superior in open fractures. Whether fractures are open or closed may then be a reasonable criterion for deciding which surgical procedure to undertake; in this case, even if the subgroup hypothesis was incorrect, practising according to the results would not be deleterious.

More often, prudence would dictate whether patients should be treated according to the estimated overall effects—that is, unless a subgroup effect had been replicated in multiple studies and summarised in a systematic review and (ideally using individual patient data) meta-analysis that met all, or almost all, criteria.

Conclusion

Authors of reported randomised controlled trials often claim subgroup effects; however, the credibility of subgroup effects, even when claims are strong, is usually low. Authors should uniformly register protocols including plans for any subgroup analyses they intend to undertake and clearly and comprehensively report the conduct and interpretation of subgroup analyses. Journal editors should insist on application of appropriate criteria for evaluating the credibility of subgroup analyses, and inferences regarding credibility should be consistent with the extent to which the analyses meet criteria. If critical criteria are not met, the results should be treated as hypothesis generating and requiring replication before they are used as a basis for clinical management. Users of trial reports should be aware of the danger of spurious subgroup claims and should treat claims that fail to meet most criteria, and particularly the critical criteria, with scepticism.

We thank Jessica Truong and Neil Dattani for screening part of the study abstracts and title; Monica Owen for administrative assistance; and Aravin Duraikannan for developing the electronic data abstraction forms.

Contributors: XS and GHG conceived the study, had full access to all of the data in the study, and take responsibility for the integrity of the data and the accuracy of the data analysis. GHG is the guarantor. XS, GHG, MB, JWB, EAA, SDW, and DGA designed the study. MB, EAA, JWB, ND-G, JJY, FM, MMB, DB, DM, POV, GM, SKS, PD, BCJ, PA-C, BH, XS, J Truong, N Dattani, and NB acquired the data. XS, GHG, SDW, and DH-A analysed and interpreted the data. XS drafted the manuscript. XS, JWB, MB, GHG, JJY, BH, EAA, DB, DM, POV, BCJ, PD, SDW, DGA, GM, FM, DH-A, ND-G, PA-C, SKS, MMB, and NB critically revised the manuscript. XS provided administrative, technical, and material support.

Funding: This study was supported by the National Natural Science Foundation of China (project No: 70703025). The funder had no role in the study design, writing of the manuscript, or decision to submit this or future manuscripts for publication. MB is supported by Santésuisse and the Gottfried and Julia Bangerter-Rhyner Foundation. JWB is funded by a new investigator award from the Canadian Institutes of Health Research and the Canadian Chiropractic Research Foundation. DB is supported by the European Union (grant award health-F5-2009-223060). DM is supported by a research scholarship from the Swiss National Science Foundation (PBBSP3-124436 and PASMP3-132571) and the Lichtenstein-Stiftung, Basel, Switzerland. PD is supported by a Dennis W Jahnigan Career Development Award by the American Geriatrics Society. BCJ holds a SickKids Foundation postdoctoral fellowship. PA-C is funded by a Miguel Servet contract by the Instituto de Salud Carlos III (CP09/00137). JJY is supported by a career scientist award from the

What is already known on this topic

Investigators often report subgroup analyses and claim differences of treatment effects among patient subgroups

Investigators often undertake a large number of subgroup analyses but fail to prespecify subgroup hypotheses and do appropriate statistical tests

Criteria exist to assess the credibility of putative subgroup effects in randomised controlled trials

What this study adds

The credibility of most subgroup claims in randomised controlled trials, even strong claims, is usually low

Investigators should clearly and comprehensively report the conduct and interpretation of subgroup analyses

Journal editors should insist on application of appropriate criteria for evaluating the credibility of subgroup analyses

Ontario Ministry of Health and Long-Term Care. DGA is supported by Cancer Research UK (grant No C5529).

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, and no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: No additional data available.

1 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.
2 Bhandari M, Devereaux PJ, Li P, Mah D, Lim K, Schunemann HJ, et al. Misuse of baseline comparison tests and subgroup analyses in surgical trials. *Clin Orthop Relat Res* 2006;447:247-51.
3 Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? *Am Heart J* 2006;151:257-64.
4 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.
5 Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007;357:2189-94.
6 Gabler N, Duan N, Liao D, Elmore J, Ganiats T, Kravitz R. Dealing with heterogeneity of treatment effects: is the literature up to the challenge? *Trials* 2009;10:43.
7 Bender R, Koch A, Skipka G, Kaiser T, Lange S. No inconsistent trial assessments by NICE and IQWiG: different assessment goals may lead to different assessment results regarding subgroup analyses. *J Clin Epidemiol* 2010;63:1305-7.
8 Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;365:176-86.
9 Guyatt G, Wyer PC, Ioannidis J. When to believe a subgroup analysis. In: Guyatt G, Rennie D, Meade MO, Cook DJ, eds. *User's guide to the medical literature: a manual for evidence-based clinical practice* . 2nd ed. AMA, 2008:571-83.
10 Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ* 2011;342:d1569.
11 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;116:78-84.
12 Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. *Med J Aust* 2004;180:289-91.
13 Fletcher J. Subgroup analyses: how to avoid being misled. *BMJ* 2007;335:96-7.
14 Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117.
15 Sun X, Briel M, Busse J, Akl E, You J, Mejza F, et al. Subgroup Analysis of Trials Is Rarely Easy (SATIRE): a study protocol for a systematic review to characterize the analysis, reporting, and claim of subgroup effects in randomized trials. *Trials* 2009;10:101.
16 National Library of Medicine. Abridged Index Medicus (AIM or "core clinical") journal titles. 2012. www.nlm.nih.gov/bsd/aim.html.
17 Thomson Reuters. ISI Web of Knowledge website. 2012. www.isiwebofknowledge.com.
18 Cui L, Hung HM, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. *J Biopharm Stat* 2002;12:347-58.
19 Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266:93-8.
20 Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet* 2005;365:1657-61.
21 Pocock SJ, Lubsen J, Wang R, Lagakos SW. More on subgroup analyses in clinical trials. *N Engl J Med* 2008;358:2076-7.
22 Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002;21:2917-30.
23 Altman DG, Bland JM. Statistics notes. Interaction revisited: the difference between two estimates. *BMJ* 2003;326:219.
24 Seymour MT, Maughan TS, Ledermann JA, Topham C, James R, Gwyther SJ, et al. Different strategies of sequential and combination chemotherapy for patients with poor prognosis advanced colorectal cancer (MRC FOCUS): a randomised controlled trial. *Lancet* 2007;370:143-52.
25 Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c332.
26 Spiegel D, Butler LD, Giese-Davis J, Koopman C, Miller E, DiMiceli S, et al. Effects of supportive-expressive group therapy on survival of patients with metastatic breast cancer: a randomized prospective trial. *Cancer* 2007;110:1130-8.
27 Itoi E, Hatakeyama Y, Sato T, Kido T, Minagawa H, Yamamoto N, et al. Immobilization in external rotation after shoulder dislocation reduces the risk of recurrence. A randomized controlled trial. *J Bone Joint Surg Am* 2007;89:2124-31.

Accepted: 19 January 2012

Cite this as: *BMJ* 2012;344:e1553

Tables

Table 1| Study characteristics of trials in which subgroup effects were or were not claimed. Values are numbers (percentages) unless stated otherwise

Study characteristics	Subgroup effects claimed (n=83)	Subgroup effects not claimed (n=124)	Total (n=207)
Median (interquartile range) sample size per study arm*	199 (72-541)	221 (87-531)	214 (81-511)
Journal type:			
High impact†	51 (61)	90 (73)	141 (68)
Lower impact	32 (39)	34 (27)	66 (32)
Source of funding:			
Industry	41 (49)	62 (50)	99 (49)
Non-industry‡	42 (51)	62 (50)	108 (52)
Study area:			
Non-surgical	68 (82)	107 (86)	175 (85)
Surgical	15 (18)	17 (14)	32 (16)
Main effect for primary outcome:			
Statistically significant	50 (60)	71 (57)	121 (59)
Statistically non-significant	33 (40)	53 (43)	86 (42)

*Sample size considered for selected comparison.

†*Annals of Internal Medicine, BMJ, Journal of the American Medical Association, Lancet,* and *New England Journal of Medicine*.

‡Governmental agencies, private not for profit organisations, explicit statement of no funding, or funding source not reported.

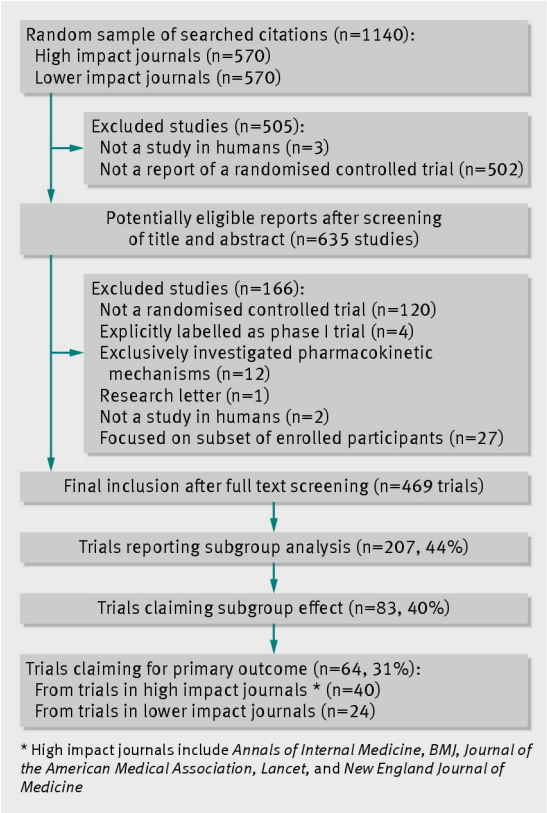
Table 2| **Proportion of claims meeting subgroup criteria for primary outcome**

Criteria	Strong claim (n=20)	Claim of likely effect (n=28)	Suggestion of possible effect (n=16)	Total (n=64)
Subgroup variable as a baseline characteristic	19 (95)	27 (96)	14 (88)	60 (94)
Subgroup variable a stratification factor at randomisation	5 (25)	3 (11)	5 (31)	13 (20)
Subgroup hypothesis specified a priori	10 (50)	11 (39)	5 (31)	26 (41)
A small number (≤5) of subgroup hypotheses tested	10 (50)	13 (46)	5 (31)	28 (44)
Significant interaction test (P<0.05)	2 (10)	4 (14)	0 (0)	6 (9)
Independence of interaction*	1 (25) (n=4†)	0 (0) (n=10†)	0 (0) (n=5†)	1 (5) (n=19†)
Direction of subgroup effect correctly prespecified	3 (15)	1 (4)	0 (0)	4 (6)
Subgroup effect consistency across studies	9 (45)	9 (32)	3 (19)	21 (33)
Subgroup effect consistent across related outcomes	7 (35)	9 (32)	3 (19)	19 (30)
Compelling indirect evidence	6 (30)	6 (21)	2 (13)	14 (22)

*19 trials claimed two or more subgroup claims.

†Number of trials having two or more subgroup claims.

Figure



Flow of study screening