

Users' Guides to the Medical Literature

How to Use a Subgroup Analysis

Users' Guides to the Medical Literature

Xin Sun, PhD; John P. A. Ioannidis, MD, DSc; Thomas Agoritsas, MD; Ana C. Alba, MD; Gordon Guyatt, MD, MSc

Clinicians, when trying to apply trial results to patient care, need to individualize patient care and, potentially, manage patients based on results of subgroup analyses. Apparently compelling subgroup effects often prove spurious, and guidance is needed to differentiate credible from less credible subgroup claims. We therefore provide 5 criteria to use when assessing the validity of subgroup analyses: (1) Can chance explain the apparent subgroup effect; (2) Is the effect consistent across studies; (3) Was the subgroup hypothesis one of a small number of hypotheses developed a priori with direction specified; (4) Is there strong preexisting biological support; and (5) Is the evidence supporting the effect based on within- or between-study comparisons. The first 4 criteria are applicable to individual studies or systematic reviews, the last only to systematic reviews of multiple studies. These criteria will help clinicians deciding whether to use subgroup analyses to guide their patient care.

JAMA. 2014;311(4):405-411. doi:10.1001/jama.2013.285063

 Supplemental content at jama.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Gordon H. Guyatt, MD, MSc, Departments of Clinical Epidemiology and Biostatistics and Medicine, Faculty of Health Sciences, McMaster University, 1200 Main St W, Room 2C12, Hamilton, Ontario, L8N 3Z5, Canada (guyatt@mcmaster.ca).

Clinical Scenario

You are a physician working at a regional trauma center. Your unit's committee, which is responsible for standardization of care, is considering using tranexamic acid to treat trauma patients arriving 3 hours after injury. Almost all the information on this topic is derived from a single, blinded trial that randomized trauma patients to tranexamic acid or placebo. The original publication reported that 99% of the enrolled patients were followed up and there was a reduction in all-cause mortality (relative risk [RR], 0.91; 95% CI, 0.85-0.97) with no apparent subgroup effect. A subsequent publication focused on an additional analysis addressing death from bleeding and reported a powerful subgroup effect with a large benefit for patients treated within 3 hours of injury and possible harm if treated 3 or more hours after injury. The committee's mandate is to decide whether tranexamic acid should not be given to patients 3 hours or more after injury. The credibility you place on the subgroup analysis will determine your decision.

The Challenge of Subgroup Analysis

Clinicians making treatment decisions use evidence applying most closely to the individual patient and treatment under consideration. To address this issue, clinical trialists and systematic review authors frequently conduct subgroup analyses to identify groups of patients (ie, sicker patients) who may respond differently to treatment than other groups (ie, less sick patients), or find more and less effective ways of administering treatment (eg, intravenous vs oral).^{1,2} Although subgroup analyses may help individualize treatment, they may also mislead clinicians.

For example, the Second International Study of Infarct Survival (ISIS-2) investigators reported an apparent subgroup effect: patients presenting with myocardial infarction born under the zodiac

signs of Gemini or Libra did not experience the same reduction in vascular mortality attributable to aspirin that patients with other zodiac signs had (Table 1).³ Despite statistical the findings' reaching significance ($P = .003$ for interaction), the investigators did not believe the subgroup effect—they reported the results to demonstrate the dangers of subgroup analysis. The eTable (in the Supplement) lists 19 examples in which other randomized clinical trial (RCT) authors have, when faced with biologically more plausible effects, claimed subgroup effects unsupported by subsequent evidence.

Clinician scientists may underestimate the extent to which chance can create imbalances (see Box 1 for another illustration). In the situations we described, the investigators were either demonstrating (the ISIS-2 example) or being misled by (eTable in the Supplement) the play of chance. When treatment effects are similar across patient groups or across ways of administering treatments, subgroup analyses will sometimes reveal apparently compelling but actually spurious subgroup differences.

The challenge for readers of the medical literature is to distinguish credible from less than credible reports of subgroup effects. Clinicians cannot rely on study authors to do this for them. A systematic survey of 407 RCTs found 207 with subgroup analyses. Of these 207, authors claimed subgroup effects in their primary outcome in 64.⁵ In most instances, the claims did not stand up to widely used guidance for the credibility of subgroup analyses.^{6,7} Thus, the subgroup claims were potentially misleading.⁸

We will now discuss a number of relevant general issues, followed by recommendations for how to assess subgroup analyses. Although our discussion focuses on individual RCTs and systematic reviews, the principles in this guide also apply to observational studies.

The Interest Is in Relative, Not Absolute, Subgroup Effects

Consider a 45-year-old, white, nonsmoking woman without heart disease or diabetes, elevated serum total cholesterol level (200 mg/dL), decreased high-density lipoprotein level (40 mg/dL), and blood

Table 1. Subgroup Analysis of the Second International Study of Infarct Survival

	No./Total (%) of Patients		Relative Risk (95% CI)	Relative Risk Reduction or Increase, %
	Aspirin	Placebo		
Vascular mortality in all patients	804/8587 (9.4)	1016/8600 (11.8)	0.79 (0.73-0.87)	20.7 Reduction
Gemini or Libra	150/1357 (11.1)	147/1442 (10.2)	1.08 (0.87-1.34)	8.4 Increase
Other astrological signs	654/7228 (9.0)	868/7187 (12.1)	0.75 (0.68-0.82)	25.4 Reduction

pressure of 130/85 mm Hg who is not receiving blood pressure treatment. Her risk of major coronary events in the next decade is 1.4%.⁹ (To convert cholesterol from milligrams per deciliter to millimole per liter, multiply by 0.0259.)

Now consider a 65-year-old man, a smoker, without heart disease or diabetes, presenting with elevated total serum cholesterol level (250 mg/dL), decreased high-density lipoprotein level (30 mg/dL), and blood pressure of 165/90 mm Hg, not taking antihypertensive medication. His risk of major coronary events exceeds 38%.

These 2 individuals represent the extremes of low- and high-risk subgroups of candidates for lipid-lowering therapy. A systematic review and meta-analysis showed that statin therapy reduces the RR of major coronary events by approximately 30% consistently across subgroups.¹⁰ Thus, the 45-year-old woman could expect an absolute risk reduction of about 0.4% (her baseline risk of 1.4% × 30%) and the 65-year-old man could expect an absolute reduction of 10%. We would thus conclude that there is a large difference between low- and high-risk patients—a subgroup effect—in absolute, but not relative terms.

In general, relative effects (eg, risk ratio, odds ratio, hazard ratio) have proved similar across risk groups, whereas absolute effects (eg, absolute risk reduction, number needed to treat) have far greater variability.¹¹⁻¹³ Thus, the question in subgroup analysis is not whether differences exist in absolute effects—they almost always do—but in relative effects.

The Interest Is in Subgroups Identifiable at the Start of a Study

Subgroup analysis in RCTs should focus on variables defined at the time of randomization. Analyses based on features that emerge during follow-up violate principles of randomization and are less valid.

For example, an RCT of intensive vs standard glucose management in an intensive care unit (ICU) found similar mortality in patients randomized to the intervention and the control groups. Among patients remaining in the ICU for more than 3 days, there was an apparent reduction in death rates in the intensive glucose management group.¹⁴ Decisions regarding length of stay may have differed between the intervention and control groups and may have been related to patients' prognosis. For instance, because intensive glucose management might have caused episodes of transient hypoglycemia, patients in this group might have remained longer than did similar patients in the control group. If the patients in the intervention group who stayed longer represent a group with a good prognosis, the prognostic balance that randomization initially achieved would be lost, creating a spurious treatment benefit in this subgroup.

The balance between groups achieved by randomization exists only when assessing patients in the groups to which they were initially randomized. Dividing patients into subgroups by clinical char-

acteristics that emerge—potentially as a result of treatment—after randomization may demonstrate apparently statistically significant differences but those differences arise because the patients themselves are different (ie, treatment and control patients are prognostically different), not because of a treatment effect. Subgroup claims based on characteristics arising during a study's conduct rather than on characteristics present at randomization have only low credibility.

Subgroup Claims Are Only as Credible as the Studies From Which They Arise

Consider an RCT that failed to conceal randomization, failed to undertake any blinding, and failed to follow-up half the enrolled patients. Because of a very high risk of bias, clinicians would be wise to be skeptical of any subgroup claims from such a study.

Subgroup Effects Are Not All-or-Nothing Decisions

Debates about subgroup effects may be framed as absolute acceptance or rejection, yes or no with nothing in between. This approach is undesirable and destructive: it ignores the uncertainty that inevitably accompanies such judgments. It is more realistic to view the likelihood of a subgroup effect as being real on a continuum ranging from "certainly true" to "certainly false." It is better to understand where on this continuum a putative subgroup effect lies. Viewing subgroup analyses in terms of a continuum ranging from certainly true to certainly false—with the expectation that most of the time, the proper conclusion would be "probably true" or "probably false"—rather than a sharp division between true or false is the approach that we will use in this Users' Guide.

Guidelines for Interpreting Subgroup Analyses

Clinicians will encounter subgroup analyses in individual observational studies or RCTs and in systematic reviews and meta-analyses. Box 2 presents our criteria for deciding on the credibility of a subgroup analysis. Four of these criteria apply to both individual studies and systematic reviews, the fifth only to systematic reviews.

Can Chance Explain the Subgroup Difference?

We have emphasized the powerful and underappreciated potential for chance to mislead investigators and clinician readers. Statistical tests help determine the extent to which study results may be explained by chance alone.

Consider Figure 1, presenting the results of a hypothetical analysis of subgroups 1 and 2 and their pooled results. Assume that investigators separately test the hypothesis that chance can explain the differences between treatment and control in subgroups 1 and

Box 1. The Miracle of DICE Therapy

In an imaginative investigation, Counsell et al⁴ directed students in a statistics class to roll different-colored dice to simulate 44 independent clinical trials of fictitious therapies. Participants received the dice in pairs and were told that one die was an ordinary die representing control patients, whereas the other was weighted to roll either more or fewer 6s (6 representing a patient death) than the control. Dice were colored red, white, and green, with each color representing a different treatment. The investigators simulated trials of different size (numbers of time the pair of dice were rolled), methodological rigor (errors made in filling out the results form), and experience level of operators.

Subgroup analysis of the red dice found no statistically significantly excess mortality. When a subgroup was created by combining the white and green dice, excluding cases with errors in the forms and using data from skilled operators, there was a 39% ($P = .02$) relative risk reduction attributable to the treatments.

The participants, however, had been deliberately misled: the dice were not loaded. This study showed how a completely random phenomenon can yield statistically significant results in a subgroup analysis.

2. They will conclude the answer is yes for group 1 (confidence intervals overlap an RR of 1.0) and no for group 2 (confidence intervals exclude an RR of 1.0). Investigators might then conclude that they have shown a subgroup effect: treatment is effective in subgroup 2 but not in 1.

Such a conclusion would be misguided. Given that the point estimates are the same and thus the confidence intervals completely overlap with one another, it is likely that the treatment effect is very similar in subgroups A and B. Thus, the differences in width of the confidence intervals (overlapping no effect in 1 but not in 2) reflect differences in sample size (larger in group 2 than group 1) or number of events (more events in group 2 than in group 1). Although the example shows exactly the same point estimates of effect for subgroups 1 and 2, the reasoning would also apply if confidence intervals are substantially overlapping when point estimates differ considerably.

The null hypothesis for the appropriate statistical test is that the treatment effect is the same in the 2 subgroups. The results provide no evidence that would lead to rejecting that hypothesis. Indeed, given the identical point estimates in 1 and 2, the appropriate test, a test for interaction, would yield a P value of $>.99$. Having conducted the appropriate test for interaction and concluded that chance explains any differences between groups, investigators should focus on the overall trial results rather than on separate subgroups 1 and 2.

Investigators made this error in logic in an RCT of angiotensin-converting enzyme (ACE) inhibitor vs diuretic-based antihypertensive therapy when they concluded that the "initiation of antihypertensive treatment involving ACE inhibitors in older subjects, particularly men, appears to lead to better outcomes than treatment with diuretic agents."¹⁵ The investigators based their conclusion on the relative risk reductions of 17% (95% CI, 3%-29%) in men and 0% (95% CI, -20% to 17%) in women. The appropriate test of interaction for the subgroup effect of sex on the outcome asks the question: Can chance explain the difference between an apparent

Box 2. Credibility of Within- and Between-Study Comparisons**Possible explanations of difference in subgroups**

Between-Study Comparisons

- Hypothesized difference
- Chance
- Other patient differences
- Different cointerventions
- Different outcome measures
- Different risk of bias

Within-Study Comparisons

- Hypothesized difference
- Chance

17% relative risk reduction in men and the 0% relative risk reduction in women? The P value associated with the interaction test is .15, meaning that if there were no true difference, by chance alone, apparent differences of this magnitude or greater than 15% of the time would be observed. Although the difference between the ACE inhibitor and diuretic-based therapies was statistically significant in men but not women, when the 2 groups were compared directly to one another and an interaction test was performed, the data were consistent with the null hypothesis that the effect did not differ between sexes.

Contrast this with an RCT addressing the relative effect of reamed vs unreamed nailing on reoperation rates in patients with tibial fractures.¹⁶ Reamed nailing decreased reoperations in patients with closed fractures (RR, 0.67; 95% CI, 0.47-0.76), but increased reoperation in those with open fractures (RR, 1.27; 95% CI, 0.91-1.78). When investigators performed an test of interaction to address the hypothesis that reamed vs unreamed nailing had the same effect on reoperations in closed and open fractures, the P value was .01. Differences between groups as large or larger than observed in this study would occur by chance only 1% of the time. When chance alone is unlikely to explain subgroup differences, a subgroup effect may be present but clinicians should also consider the other criteria that we present in this article.

A variety of statistical techniques are available to explore whether chance alone explains apparent subgroup differences.^{6,17,18} When assessing the results of these tests of interaction, clinicians should note whether differences in effect are quantitative (ie, same direction but varying magnitude by treatment effects) or qualitative (ie, beneficial in one subgroup but harmful in another). Qualitative effects in subgroups are uncommon.

Clinicians should also consider that failure to show differences between subgroups does not mean that differences do not exist. An insufficient number of study participants could result in an inability to show that differences exist (ie, the test for interaction was underpowered). On other hand, if the results of an appropriate statistical test show that chance is an unlikely explanation for an apparent subgroup effect, it does not mean the effect is real. It does mean clinicians should take the possible effect seriously.

Is the Subgroup Difference Consistent Across Studies?

One may generate a hypothesis concerning differential response in a subgroup of patients by examination of data from a single study.

Replication in other studies increases its credibility, and failure to replicate diminishes its credibility. Readers of trial reports should look carefully in the discussion sections for references to subgroup results in similar trials. Because investigators tend to select references related to evidence supporting their positions, statements from authors regarding a systematic search for related evidence strengthens arguments in favor of the subgroup analyses' results. The online appendix provides examples in which failure to replicate subgroup analyses undermined the subgroup claim. Subgroup claims failing replication warrant considerable skepticism.

Was the Subgroup Difference One of a Small Number of a Priori Hypotheses in Which the Direction Was Accurately Prespecified?

Embedded within any large data set are a certain number of apparent but, in fact, spurious subgroup differences. As a result, the credibility of any apparent subgroup difference that arises from post hoc rather than a priori hypotheses is questionable.

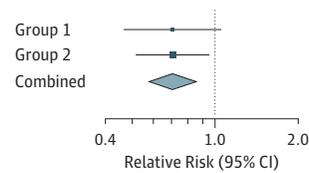
For example, in the first large trial of aspirin for patients with transient ischemic attacks, the investigators reported that aspirin had a beneficial effect in preventing stroke in men, but not in women with cerebrovascular disease.¹⁹ For many years, this led many physicians to withhold aspirin from women with cerebrovascular disease. The investigators, however, had stumbled across the finding in exploring the data rather than suspecting it beforehand. The apparent subgroup effect was subsequently found in other studies, and in a meta-analysis summarizing these studies, to be spurious.²⁰ Had clinicians been appropriately skeptical of this post hoc finding and demanded replication, they would not have missed the opportunity to prevent strokes in their female patients.

Even if investigators have prespecified their hypotheses, the strength of inference for confirmation of any hypothesis will decrease if a large number of hypotheses are tested. For example, investigators conducted an RCT of platelet-activating factor receptor antagonist in septic patients. For all 262 patients, results showed that the small benefit for therapy failed to meet the usual threshold $P < .05$ for statistical significance. A subgroup analysis of 110 patients with gram-negative bacterial infection was found to have a large, statistically significant advantage for platelet-activating factor receptor antagonist treatment.²¹

A subsequent, larger hypothesis-testing RCT involving 444 patients with gram-negative bacterial infection failed to replicate the apparent benefit observed in the subgroup analysis of the previous trial.²² The disappointed investigators might have been less surprised at the result of the second trial had they fully appreciated the limitations of their first subgroup analysis: the possible differential effect of platelet-activating factor receptor antagonist in gram-negative bacterial infection was 1 of 15 subgroup hypotheses that they tested.²³

The era of molecular medicine has increased the temptation for multiple hypothesis testing: the number of candidate subgroup analyses that can be performed for molecular analyses is enormous. Although gene-based information is often biologically fascinating, databases include information on many thousands or even millions of genetic or other molecular factors that are difficult to interpret. Testing large numbers of subgroup hypotheses will create some misleading results because of problems relate to multiple comparisons.²⁴

Figure 1. Inappropriate Statistical Comparison



The figure presents the results of a hypothetical analysis of subgroups 1 and 2 and their pooled results. Error bars indicate 95% confidence intervals. The size of the data markers (squares) reflects the amount that each group contributes to the pooled estimates.

For example, although many studies have identified pharmacogenetic markers for subgroups of patients with different responses to treatment or toxicity, only a handful of these differences have proved to be true when tested in additional data sets. Given the large number of genomic and other molecular markers, statistical significance thresholds are far more stringent when testing for subgroup differences. For example, in pharmacogenomics, for which millions of gene variants are tested, researchers and readers should pay little attention to claims of important findings unless the subgroup differences (eg, between patients carrying vs those not carrying 2 copies of a putative pharmacogenetic marker) are associated with P values lower than 10^{-8} .²⁵

A final issue in hypothesis testing is specification of the direction of the effect. In an RCT of vasopressin vs norepinephrine in 778 patients with septic shock, the investigators specified a priori a primary subgroup analysis: reduced mortality attributable to vasopressin over norepinephrine would be greater for patients with more severe septic shock.²⁶ In contrast to the investigators' expectations, vasopressin appeared to benefit only patients with less severe septic shock (RR, 1.04 in more severe vs 0.74 in less severe; interaction $P = .10$).

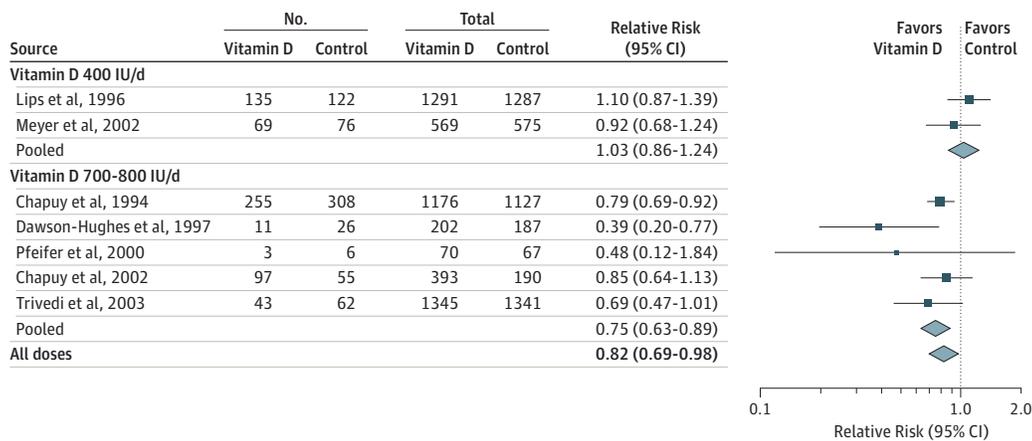
The investigators' failure to correctly identify the direction of the subgroup effect appreciably weakened any inference that vasopressin was superior to norepinephrine in the less severely ill patients. Clinicians should look for explicit statements on whether subgroup hypotheses, and their direction, were specified a priori.

Study reports often fail to clearly identify the extent to which a hypothesis arose before, during, or after the data were collected and analyzed or the number of subgroup hypotheses tested. If the investigators withhold this information, reporting only hypotheses that were statistically significant, the reader will be misled. When, however, the hypothesis has been clearly suggested by a different data set, and investigators replicate the finding in a new RCT, clinicians can be confident regarding a priori specification.

Is There a Strong Preexisting Biological Rationale Supporting the Apparent Subgroup Effect?

Subgroup claims are more credible if additional, external evidence (such as from laboratory studies or analogous situations in human biology) makes it plausible. Such evidence may come from 3 sources: studies of different populations (including animal studies), observations of subgroup differences for similar interventions, and results of studies of other related intermediary outcomes.

Figure 2. Meta-analysis of Studies Addressing the Effect of Vitamin D on Nonvertebral Fractures



The size of the data markers (squares) reflect the amount that each study contributes to the pooled estimates. This is based on Bischoff-Ferrari et al.³⁰

There is no shortage of biologically plausible explanations supporting almost any observation. One example of biologic evidence supporting a possible subgroup effect concerns an apparent effect described previously: a trial suggested that aspirin reduced stroke risk in men but not in women.¹⁹ Subsequent animal research provided a biologic rationale for the observed sex differences in aspirin's effects on stroke risk.²⁷ However, subsequent clinical trials found that there was no sex difference in stroke response to aspirin irrespective of the biological rationale found in laboratory animals.²⁸

One of the most useful roles of biologic rationale is to raise serious questions regarding an apparent subgroup effect that is inconsistent with our current understanding of biology. The apparent interaction between birth zodiac sign and the effect of aspirin in myocardial infarction (Table 1) provides an example in which the absence of a biological explanation seriously undermines the credibility of an apparent subgroup effect.

Subgroup Claims in Systematic Reviews and Meta-analyses: Within- vs Between-Trial Comparisons

Up to now, this Users' Guide has addressed individual studies. Making inferences about subgroup effects in systematic reviews requires application of the previously discussed 4 criteria and consideration of whether the comparison between subgroups is done within or between studies. In single trials, the comparison is always within: that is, the 2 groups of patients (eg, the older and younger) or the 2 alternative ways of administering the intervention (eg, higher and lower doses) were assessed in the same RCT. Within meta-analyses, this is not necessarily the case.

Consider the controversy regarding dose effects of vitamin D on fracture reduction.²⁹ One meta-analysis suggesting the benefit of higher doses examined the effect of vitamin D on nonvertebral fractures and reported on results of 2 studies of lower doses (400 IU) and 5 studies of higher doses (700-800 IU).³⁰ The pooled estimates from the low-dose studies suggested no effect on frac-

tures, while the higher dose studies suggest a 23% reduction in RR (Figure 2). The test for interaction, following the same principle as individual studies, addresses whether chance can explain the difference between the high and low dose studies and yielded a *P* value of .01.

The inference regarding the dose effect is, however, limited because this was a between- rather than a within-study comparison. As a result there are a number of competing explanations for the observed differences between the high- and low-dose studies.

Box 2 describes the generic competing explanations present in all between-study comparisons. Aside from the hypothesized effect of vitamin D dose in the vitamin D-fracture studies, explanations for the apparent differences in study results include the following: patients in the low-dose studies were exposed to adequate sunlight (and thus didn't need supplementation) whereas the high-dose study patients did not; the patients who received high dosages took calcium supplements and the patients receiving low dosages did not; the length of follow-up differed in the low- and high-dose studies; and the low-dose studies had a lower risk of bias than the high-dose studies.

Within-trial subgroup differences from well-designed and implemented RCTs leave only 2 likely explanations: chance and a real effect (Box 2). Most subgroup analyses from systematic reviews are limited by between-study comparisons.³¹ The exception is individual patient data meta-analyses in which most or all studies have included patients from each relevant subgroup. Investigators undertaking an individual patient data meta-analysis can conduct sophisticated analyses that compare the effects in subgroups within studies and then effectively pool across those studies.

Using the Guide

Returning to our opening scenario, your committee notes that almost all the data come from a single trial and thus reflect a within-trial comparison. The RR for death due to bleeding in patients receiving tranexamic acid of an 1 hour or less after injury is 0.68 (95% CI, 0.57-0.82), 1 and 3 hours from injury is 0.79 (95%

CI, 0.64-0.97), and more than 3 hours after injury is 1.44 (95% CI, 1.12-1.84). Chance appears a very unlikely explanation for the difference ($P < .001$). Trauma patients exhibit early fibrinolysis that could exacerbate bleeding; tranexamic acid inhibits fibrinolysis, providing a strong biological rationale for the treatment. The fibrinolysis may be largely resolved by 3 hours, thus providing a biological explanation for the absence of benefit after 3 hours. Prespecification is complex. The time-from-injury hypothesis was one of a small number of a priori hypothesis with a specified direction, but the analysis plan focused on all-cause mortality (in which the investigators found no subgroup effect). The analysis of cause-specific mortality represented a secondary exploration of the data. Ultimately, your committee decides that the subgroup effect, while far from completely secure, is sufficiently credible that your unit will not administer tranexamic acid to patients arriving more than 3 hours after their trauma.

Conclusions

The criteria for assessing subgroup analyses presented in this Users' Guide (Box 3) will help clinicians evaluating the credibility of claims of differential response to treatment in a definable subgroup of patients. These are intended as core criteria that clinicians can feasibly apply when evaluating a subgroup claim. More comprehensive criteria are available for readers seeking a deeper understanding of the nuances of assessing subgroup claims.⁷ Moreover, we have focused on data from randomized trials and their systematic reviews. Subgroup claims are increasingly based on observational data and these—like their estimates of effect in entire populations—warrant considerably greater skepticism.⁸

Box 3. Guidelines for Deciding Whether Apparent Differences in Subgroup Response Are Real

Issues for Individual Studies and Systematic Reviews

Can Chance Explain the Subgroup Difference?

Is the subgroup difference consistent across studies?

Was the subgroup difference one of a small number of a priori hypotheses in which the direction was accurately prespecified?

Is there a strong preexisting biological rationale supporting the apparent subgroup effect?

An Issue for Systematic Reviews Only

Is the subgroup difference suggested by comparisons within rather than between studies?

Applying these criteria, clinicians will sometimes find, at one extreme, relatively small interactions easily explained by chance and based on between-study differences generated by post hoc explorations. Less frequently, at the other extreme, they will find interactions with very small P values (for instance $< .01$), based on within-trial comparisons with consistent results following a limited number of subgroup hypotheses with a correctly specified direction. The former should be viewed with skepticism. The latter are more credible and can be used for clinical decision making.

Results between these extremes require consideration of a number of factors including the risks associated with administering or avoiding treatment and patient's values and preferences. Judgments about the credibility of subgroup claims, based on the criteria that we have suggested, are likely to play a key part in such decisions.

ARTICLE INFORMATION

Author Contributions: Dr Guyatt had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author Affiliations: Chinese Evidence-Based Medicine Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China (Sun); Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada (Sun, Agoritsas, Guyatt); Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, California (Ioannidis); Stanford Prevention Research Center, Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California (Ioannidis); Department of Statistics, Stanford University School of Humanities and Sciences, Meta-Research Innovation Center at Stanford (METRICS), Stanford, California (Ioannidis); Heart Failure and Transplantation Program, Toronto General Hospital, University Health Network, Toronto, Ontario, Canada (Alba).

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Guyatt reported institutional support for development of education presentations from UpToDate and Bristol-Myers Squibb. No other financial disclosures were reported.

Funding/Support: Dr Agoritsas was financially supported by the Fellowship for Prospective Researchers grant PBGEP3-142251 from the Swiss National Science Foundation. Dr Sun is supported by Young Investigators Award 2013SCU04A37 from Sichuan University in China.

Role of the Sponsors: The funders had no role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; in the preparation, review, or approval of the manuscript; or in the decision to submit the manuscript for publication.

Additional Contributions: We thank Diane Heels-Ansdell, MSc, Department of Clinical Epidemiology and Biostatistics, McMaster University, for creating Figure 2 and conducting the associated statistical analysis, for which she received no compensation, and we thank Deborah Cook, MD, from McMaster University in Canada, for conducting a meticulous and very helpful edit of the manuscript. She did not receive any payment for the editing effort.

Correction: The article was corrected on January 21, 2014, to add the subtitle to the front page.

REFERENCES

- 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(19):2917-2930.
- Sun X, Briel M, Busse JW, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ*. 2011;342. doi:10.1136/bmj.d1569.
- ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet*. 1988;2(8607):349-360.
- Counsell CE, Clarke MJ, Slattery J, Sandercock PA. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *BMJ*. 1994;309(6970):1677-1681.
- Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*. 2012;344. doi:10.1136/bmj.e155.
- Buyse ME. Analysis of clinical trial outcomes: some comments on subgroup analyses. *Control Clin Trials*. 1989;10(4)(suppl):187S-194S.
- 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.
- 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(21):2189-2194.
- Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: a report of the American

College of Cardiology/American Heart Association Task Force on Practice Guidelines [published online ahead of print November 12, 2013]. *Circulation*. 2013;doi:10.1016/j.jacc.2013.11.005.

10. Thavendiranathan P, Bagai A, Brookhart MA, Choudhry NK. Primary prevention of cardiovascular diseases with statin therapy: a meta-analysis of randomized controlled trials. *Arch Intern Med*. 2006;166(21):2307-2313.
11. Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. *Int J Epidemiol*. 2002;31(1):72-76.
12. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med*. 1998;17(17):1923-1942.
13. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med*. 2002;21(11):1575-1600.
14. Van den Berghe G, Wilmer A, Hermans G, et al. Intensive insulin therapy in the medical ICU. *N Engl J Med*. 2006;354(5):449-461.
15. Wing LM, Reid CM, Ryan P, et al; Second Australian National Blood Pressure Study Group. A comparison of outcomes with angiotensin-converting-enzyme inhibitors and diuretics for hypertension in the elderly. *N Engl J Med*. 2003;348(7):583-592.
16. Bhandari M, Guyatt G, Tornetta P III, et al; SPRINT Investigators. Study to prospectively evaluate reamed intramedullary nails in patients with tibial fractures (S.P.R.I.N.T.): study rationale and design. *BMC Musculoskelet Disord*. 2008;9:91.
17. Furberg CD, Morgan TM. Lessons from overviews of cardiovascular trials. *Stat Med*. 1987;6(3):295-306.
18. Schneider B. Analysis of clinical trial outcomes: alternative approaches to subgroup analysis. *Control Clin Trials*. 1989;10(4)(suppl):1765-1865.
19. The Canadian Cooperative Study Group. A randomized trial of aspirin and sulfapyrazone in threatened stroke. *N Engl J Med*. 1978;299(2):53-59.
20. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, I: prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *BMJ*. 1994;308(6921):81-106.
21. Dhainaut JF, Tenailon A, Le Tulzo Y, et al, BN 52021 Sepsis Study Group. Platelet-activating factor receptor antagonist BN 52021 in the treatment of severe sepsis: a randomized, double-blind, placebo-controlled, multicenter clinical trial. *Crit Care Med*. 1994;22(11):1720-1728.
22. Dhainaut JF, Tenailon A, Hemmer M, et al, BN 52021 Sepsis Investigator Group. Confirmatory platelet-activating factor receptor antagonist trial in patients with severe gram-negative bacterial sepsis: a phase III, randomized, double-blind, placebo-controlled, multicenter trial. *Crit Care Med*. 1998;26(12):1963-1971.
23. Natanson C, Esposito CJ, Banks SM. The sirens' songs of confirmatory sepsis trials: selection bias and sampling error. *Crit Care Med*. 1998;26(12):1927-1931.
24. Ioannidis JP. Microarrays and molecular research: noise discovery? *Lancet*. 2005;365(9458):454-455.
25. Panagiotou OA, Ioannidis JP, Genome-Wide Significance Project. What should the genome-wide significance threshold be? empirical replication of borderline genetic associations. *Int J Epidemiol*. 2012;41(1):273-286.
26. Russell JA, Walley KR, Singer J, et al; VASST Investigators. Vasopressin versus norepinephrine infusion in patients with septic shock. *N Engl J Med*. 2008;358(9):877-887.
27. Kelton JG, Hirsh J, Carter CJ, Buchanan MR. Sex differences in the antithrombotic effects of aspirin. *Blood*. 1978;52(5):1073-1076.
28. Collaborative overview of randomised trials of antiplatelet therapy, III: reduction in venous thrombosis and pulmonary embolism by antiplatelet prophylaxis among surgical and medical patients. *BMJ*. 1994;308(6923):235-246.
29. *Dietary Reference Intakes for Vitamin D and Calcium*. Washington, DC: Institute of Medicine; 2011.
30. Bischoff-Ferrari HA, Willett WC, Wong JB, Giovannucci E, Dietrich T, Dawson-Hughes B. Fracture prevention with vitamin D supplementation: a meta-analysis of randomized controlled trials. *JAMA*. 2005;293(18):2257-2264.
31. Contopoulos-Ioannidis DG, Seto I, Hamm MP, et al. Empirical evaluation of age groups and age-subgroup analyses in pediatric randomized trials and pediatric meta-analyses. *Pediatrics*. 2012;129(suppl 3):S161-S184.