Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses

Xin Sun,[1 2] Matthias Briel,[1 3] Stephen D Walter,[1] Gordon H Guyatt[1 4]

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada
[2]Center for Clinical Epidemiology and Evidence-Based Medicine, West China Hospital, Sichuan University, Chengdu, China
[3]Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Basel, Switzerland
[4]Department of Medicine, McMaster University, Hamilton, Canada

Correspondence to: Gordon H Guyatt, 1200 Main Street West, Rm 2C12, Hamilton, Ontario, Canada, L8N 3Z5
guyatt@mcmaster.ca

Accepted: 29 December 2009

Cite this as: *BMJ* 2010;340:c117
doi: 10.1136/bmj.c117

How can we tell the difference between spurious and real subgroup effects? This article identifies new criteria and proposes a checklist for judging the credibility of subgroup analyses

Subgroup analyses in randomised controlled trials (RCTs) or in meta-analyses of RCTs examine whether treatment effects vary according to patient group, way of giving an intervention, or approach to measuring an outcome. Subgroup analyses are common and often associated with claims of difference of treatment effects between subgroups—termed "subgroup effect", "effect modification", or "interaction between a subgroup variable and treatment".[1-3] A difference in effect between subgroups, if true, is likely to have important implications for clinical practice and policy making. Many subgroup claims are, however, subsequently shown to be false.[4] Thus, investigators, clinicians, and policy makers face the challenge of whether or not to believe apparent differences in effect.

Debates about subgroup effects may be framed in terms of absolute acceptance or rejection. For instance, in an intense academic debate,[5-11] one camp maintained that effects of propranolol on death differed in two groups of study centres, whereas the other remained highly sceptical. This "yes" versus "no" polarised approach is undesirable and destructive, mainly because it ignores the uncertainty that is inevitably part of such judgments. An approach that is more productive and more realistic is to place the likelihood that a subgroup effect is real on a continuum from "highly plausible" to "extremely unlikely", possibly by using a visual analogue scale. The question is then a decision of where on this continuum a putative subgroup effect lies.

SUMMARY POINTS

Seven existing criteria help clinicians assess the credibility of putative subgroup effects on a continuum from "highly plausible" to "extremely unlikely"

We suggest four additional criteria: subgroup definition on the basis of baseline characteristics, independence of the subgroup effect, a priori specification of the direction of the subgroup effect, and consistency across related outcomes

We propose a re-structured checklist of items addressing study design, analysis, and context

In 1991, Yusuf et al[12] discussed principles of analysing and interpreting subgroup effects, and stated that qualitative interactions (that is, when treatment is beneficial in one subgroup but harmful in another) are rare. They advocated a priori specification of subgroup hypotheses, completion of a small number of subgroup analyses, and use of an interaction test for analysing subgroup effects. In the subsequent year, Oxman and Guyatt[13] suggested seven criteria to guide inferences about the credibility of subgroup analyses. The greater the extent to which these criteria are met, the more plausible the putative subgroup effect is.

Since 1992, these seven criteria have been widely used to assess hypothesised subgroup effects,[14-23] and have undergone only minimal cosmetic revisions.[4] After years of use of the 1992 criteria, we had begun to perceive limitations. These limitations became vivid when deciding on the credibility of a subgroup hypothesis of a large multicentre randomised trial.[24] On the basis of this experience, a review of published methodological articles addressing subgroup analyses, and consultation with clinicians and epidemiologist colleagues, we identified four new criteria that could further aid differentiation between spurious and real subgroup effects. We now believe that failure to consider these criteria could result in misleading inferences about subgroup hypotheses. In this article, we describe these new criteria, use real-world examples to show how they influence the strength of inference of subgroup hypotheses, and discuss their implications. Finally, we propose a re-structured checklist of items addressing study design, analysis, and context.

Relative versus absolute effect in subgroup analyses

A crucial issue in subgroup analyses is that the effects should be examined with relative rather than absolute measures. By contrast with relative effects, which in most situations remain constant across varying baseline risks, absolute risk reductions will typically vary with baseline risk.

For example, consider the effect of statin therapy on major coronary events (that is, non-fatal myocardial infarction and coronary heart disease death) in patients with varying coronary risks. A 45 year old non-smoking woman without a family history of heart disease and without diabetes presents with a raised serum cholesterol (>5.2 mmol/L and a blood pressure of 130/85 mm Hg. Her risk of major coronary events in the next decade is

5%. Compare this woman to a 65 year old smoking male with a family history of heart diseases and diabetes, presenting with a raised serum cholesterol (> 6.2 mmol/l), and blood pressure of 160/90 mm Hg. His risk of major coronary events is 50%.

A meta-analysis showed that statin therapy could reduce the relative risk of major coronary events by 29.2%.[25] This relative effect was consistent across subgroups, including the determinants of coronary risk discussed in the previous paragraph. Because of the constant reduction in relative risk across subgroups (that is, we are confident that there is no subgroup effect for the relative effect measure), we can infer a reduction in absolute risk of major coronary events by 1.5% (from 5% to 3.5%) in the first patient and 14.6% (from 50% to 35.4%) in the second patient. If we were considering absolute risk reduction, an evident subgroup effect would exist (low risk patients, such as our female patient, have an absolute risk reduction of 1.5%, whereas high risk patients, such as our male patient, an absolute risk reduction of 14.6%).

This example shows how subgroup effects are often present when using the absolute risk reduction, but rarely present when using a relative effect measure. Indeed, in the presence of known prognostic factors that allow definition of groups at varying risk, if no subgroup effect is associated with these factors for relative measures of effect, a subgroup effect for absolute measures must exist. Our subsequent discussion, therefore, focuses exclusively on putative subgroup differences in relative effects.

The original seven criteria for subgroup analyses

The box shows the seven 1992 criteria,[13] in a re-structured checklist addressing design, analysis, and context of subgroup analyses in this paper. Inferences about subgroup effects are stronger, if, at the design stage, the comparison is made within rather than between studies, the subgroup hypothesis is specified a priori, and a small number of hypotheses are tested; if, in the analysis, the test for interaction between treatment and a subgroup variable (for example, age, sex, disease severity) suggests that chance is an unlikely explanation for apparent differences; and if, on the basis of the context, the difference in effect between subgroup categories is large and consistent across studies, and indirect evidence exists to support the difference (biological rationale).

Criteria to assess the credibility of subgroup analyses

Design

- Is the subgroup variable a characteristic measured at baseline or after randomisation?*
- Is the effect suggested by comparisons within rather than between studies?
- Was the hypothesis specified a priori?
- Was the direction of the subgroup effect specified a priori*
- Was the subgroup effect one of a small number of hypothesised effects tested?

Analysis

- Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?
- Is the significant subgroup effect independent?*

Context

- Is the size of the subgroup effect large?
- Is the interaction consistent across studies?
- Is the interaction consistent across closely related outcomes within the study?*
- Is there indirect evidence that supports the hypothesised interaction (biological rationale)?

*New criteria.

New criteria to judge the credibility of subgroup effects

1 Is the subgroup variable a characteristic measured at baseline or after randomisation?

Subgroups can be defined according to characteristics measured at baseline or after randomisation. Subgroups defined according to post-randomisation characteristics might be influenced by tested interventions; that is, the apparent difference of treatment effect between subgroups can be explained by the intervention itself, or by differing prognostic characteristics in sub-groups that emerge after randomisation, rather than by the subgroup characteristic itself. Thus, the credibility of subgroup hypotheses based on post-randomisation characteristics is severely compromised, and can be rejected simply on this criterion.

For instance, in a randomised trial of 1200 critically ill patients,[26] intensive insulin therapy, compared with conventional therapy, did not significantly reduce all-cause hospital mortality (37.3% *v* 40.0%, P=0.33). In 767 patients who stayed in the intensive care unit (ICU) for at least 3 days, the intensive insulin therapy group had a lower all-cause hospital mortality (43.0% *v* 52.5%, P=0.009), whereas in 433 patients who stayed in the ICU for less than three days, intensive therapy seemed to increase all-cause hospital mortality (26.5% *v* 18.9%, P=0.05). Because the subgroups were not selected on the basis of characteristics at baseline, the most likely explanation of the results is not that insulin therapy is harmful in those destined to stay in ICU for less than 3 days and beneficial in those destined to stay for more than three days, but rather that an effect of treatment was to create prognostic imbalance between groups in those who ultimately stayed less than three days or at least three days. Such post-randomisation subgroup analyses have very low credibility—in most cases, they can be readily dismissed.

2 Was the *direction* of the subgroup effect specified a priori?

Even if specified a priori, a putative subgroup effect is unlikely to be compelling if the investigator has little idea of the direction of the effect. A subgroup effect consistent with the *pre*-specified direction will increase the credibility of a subgroup analysis; failure to specify the direction—or worse yet, getting the direction wrong—weakens the case for a real underlying subgroup effect.

Users should look for explicit statements of a priori specification of subgroup hypothesis and subgroup

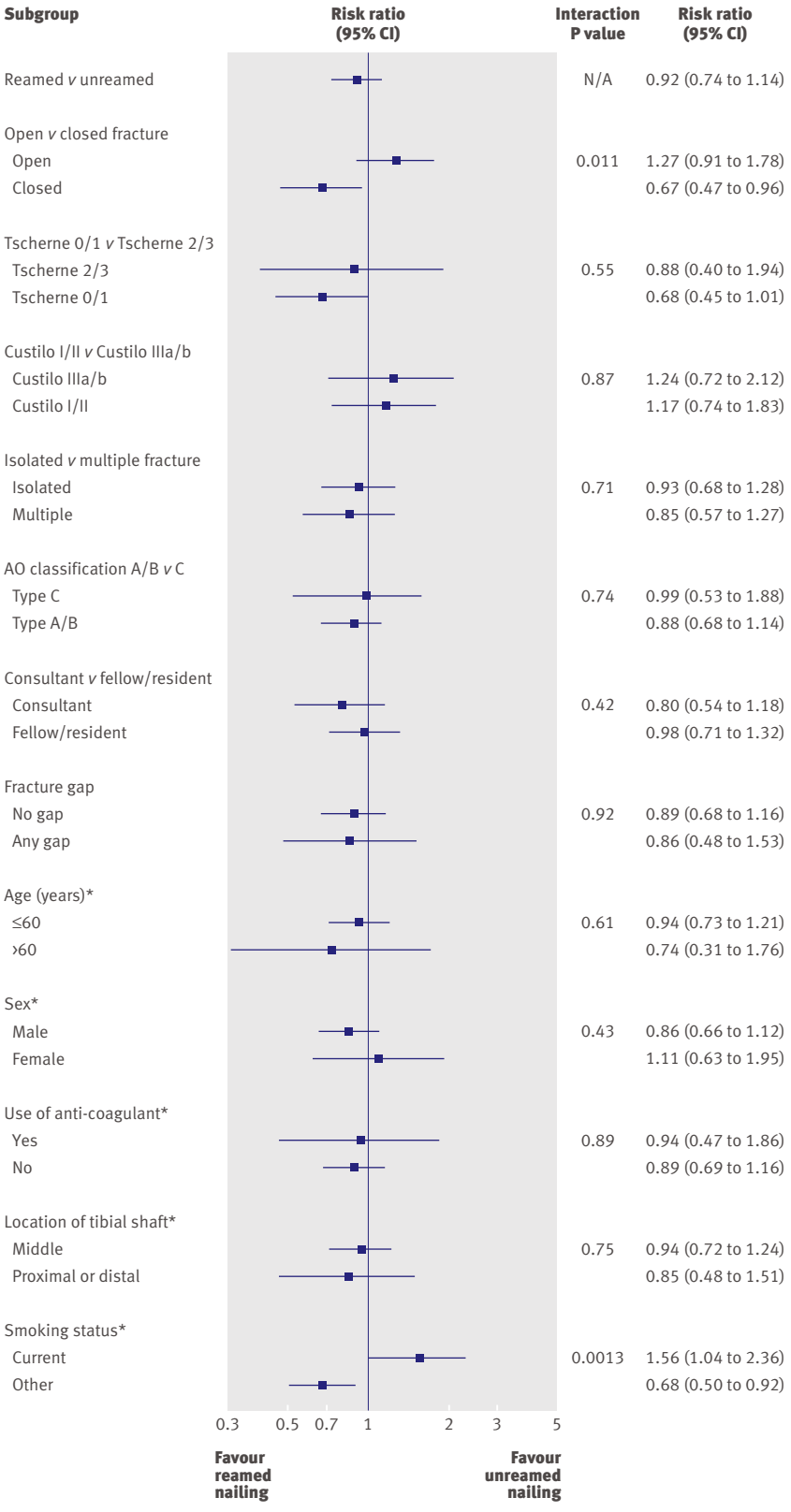

Fig 1 | Effect of reamed *v* unreamed nailing on re-operation in patients with fracture: a priori and post-hoc subgroup analyses. First point estimate and confidence interval indicates main effect. Subsequent pairs of point estimates and confidence intervals indicate effect of reamed *v* unreamed nailing on re-operation in categories of 12 subgroup variables. *Subgroup analyses done post hoc. Subgroup analysis by Tscherne type included patients with closed fracture only, and analysis by Gustilo type included open fracture only. In our analysis of significant and non-significant interactions, these two interactions were not included in regression model, resulting in ten interaction terms included in model

direction in the primary study reports. In view of emerging evidence of differences between protocols and study reports,[27] statements about what was included in registered or publicly available protocols finalised before the study or systematic review are desirable.

For instance, Russell et al[28] compared the effect of vasopressin versus norepinephrine infusion on 28-day mortality in a randomised trial of 778 patients with septic shock. As the primary subgroup analysis, the authors hypothesised a priori that the benefit of vasopressin over norepinephrine would be larger in patients with more severe septic shock. It turned out, however, that the benefit of vasopressin seemed to be greater in the patients with less severe septic shock (RR 1.04 in more severe *v* 0.74 in less severe septic shock, interaction P=0.10). The investigators' failure to correctly identify the direction of the subgroup effect appreciably weakens any inference that vasopressin is superior to norepinephrine in the less severely ill patients.

3 Is the significant subgroup effect independent?

When examining subgroup hypotheses, one must address the likelihood that the differences in effects can be explained by chance. The statistical approach that addresses this issue is called a test for interaction (the interaction meaning that the treatment effect differs across subgroup categories). The null hypothesis of the test for interaction is that no difference exists in the underlying true effect between subgroup categories. The lower the P value, the less likely it is that chance explains the apparent subgroup effect. Inevitably, the choice of a threshold for the P value involves subjective judgment. Rather than use of a threshold, a preferable way of assessing the P value is that as it gets smaller, the subgroup hypothesis becomes increasingly credible: we can be sceptical of any hypothesis with a P value of greater than 0.1, begin to consider the hypothesis if the P value is between 0.1 and 0.01, and take the hypothesis seriously when P values reach 0.001 or less.

When testing multiple hypotheses in a single study, the analyses might yield more than one apparently significant interaction. These significant interactions might, however, be associated with each other, and thus explained by a common factor. For instance, in a meta-analysis examining the effect of aspirin on the prevention of cardiovascular events, aspirin reduced the risk of stroke in women, whereas it had no apparent effect in men.[29] However, the men were generally younger than the women, suggesting that age, rather than sex, might explain the interaction.[30]

Expressing this in general terms, in a particular analysis, treatment effects apparently differ according to patients' status on variables A and B. A and B are statistically associated with each other. The difference of effects between patients in different categories with respect to A might, therefore, be explained by B (that is, the apparent effects of A on the size of treatment effect are due to confounding with B).

Another example comes from a trial of reamed versus unreamed nailing of tibial fractures.[24] Reamed and unreamed nailing produced no significant difference in

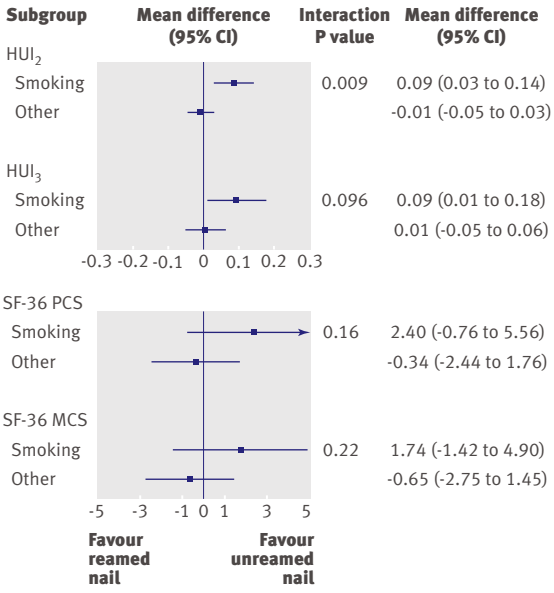

Fig 2 | Effect of reamed *v* unreamed nailing on Health Utility Index (HUI, 2a) and Short Form-36 (SF-36, 2b) in subgroups of smoking and other tibial fracture patients. PCS=physical component summary. MCS=mental component summary

the rate of re-operation (RR 0.92, 95% CI 0.74 to 1.14, fig 1). Analysis of seven a priori hypotheses suggested that reamed nailing had a lower re-operation rate in closed fractures (RR 0.64, 95% CI 0.47 to 0.96) while resulted in a higher re-operation rate in open fractures (RR 1.27, 95% CI 0.91 to 1.78, interaction P= 0.011, fig 1). We subsequently used the trial data to explore five additional hypotheses, one of which suggested that reamed nailing was superior in current smokers (RR 0.68, 95% CI 0.50 to 0.92) and unreamed nailing better in others (that is, ex-smokers and lifetime non-smokers) (RR 1.56, 95% CI 1.04 to 2.36, interaction P=0.001, fig 1).

We wondered if the apparently significant difference in treatment effect between smokers and non-smokers could be explained by fracture type (open *v* closed). In other words, one possibility was that the reason for the apparent smoking effect was that smokers tended to have open fractures and others tended to have closed fractures. In this case, the apparent association between preferred procedure (reamed or unreamed nailing) and smoking status might actually be due to confounding between smoking and fracture type (open and closed). To check for the independence of the interaction effect of smoking with procedure (reamed *v* unreamed), we included the interaction terms of treatment with smoking and treatment with fracture type in the same regression model. The analysis showed that the smoking interaction remained significant (P changed from 0.001 to 0.006) after adjusting for the interaction of fracture type with treatment. This suggests that the apparent smoking interaction cannot be explained by an association between smoking status and open versus closed fractures.

An additional check for independence of the association could include all significant and non-significant interactions in the regression model. Persisting significance of interaction terms strengthens the subgroup effect inference. In our analysis, this additional regression including both significant and non-significant hypothesised interactions (that is, the ten interactions between patient characteristics with treatment in fig 1) showed a persistent smoking interaction (P=0.008), thus providing further support for the independence of the smoking subgroup effect. A note of caution: adjustment for significant and non-significant interaction terms might be compromised by a limited sample size and small number of events,[31] providing a further rationale for prespecifying a limited number of important interactions.

4 Is the interaction consistent across closely related outcomes within the study?

If a subgroup effect is real, it is likely to manifest itself across all closely related outcomes. For example, in a randomised trial of 1692 patients with refractory non-small-cell lung cancer, Thatcher et al[32] compared the effect of gefitinib versus placebo on survival. The primary analysis showed a trend for a survival benefit with gefitinib over placebo (hazard ratio (HR) 0.89, 95% CI 0.77 to 1.02, P=0.087). Tests of a priori hypotheses indicated differential effects on survival in non-smokers (HR 0.67, 95% CI 0.49 to 0.92) and smokers (HR 0.92, 95% CI 0.79 to 1.06; interaction P=0.07). Secondary analyses on time to treatment failure showed similar differences of effects in non-smokers (HR 0.55, 95% CI 0.42 to 0.72) versus smokers (HR 0.89, 95% CI 0.78 to 1.01, interaction P=0.0015). The consistency of the subgroup effect across outcomes enhances its credibility.

In the trial of reamed versus unreamed nailing of tibial fractures,[24] unreamed nailing apparently reduced re-operations in current smokers while reamed nailing reduced re-operations in other patients (ex-smokers and lifetime non-smokers) (fig 1). To examine whether the difference existed in other outcomes, we tested the interactions between treatment and smoking status on quality of life measured by the Health Utility Index and short form-36 (fig 2). Results consistently suggested the superiority of unreamed nailing over reamed nailing in current smoking patients, and no or a small difference between unreamed and reamed nailing in other patients. This result strengthens the inference about an interaction with type of nailing and smoking status.

Discussion

Clinical and policy decision making always involves uncertainty. It is unlikely that a subgroup claim will meet either all or none of our criteria—in almost all instances, a subgroup claim will meet some but not all the criteria. Treating the likelihood that a subgroup effect is real as a continuum reflects the nature of the uncertainty. Judgment about its credibility will depend on how strongly clinicians and policy makers believe the subgroup effect is real. In other words, they will judge considering each criterion: the greater the extent to which criteria are met, the more likely the subgroup effect is real. When summarising the strength of the subgroup inferences, one can imagine—and possibly apply—a visual analogue scale with anchors of "highly plausible" and "extremely unlikely".

bmj.com: recent Research Methods & Reporting articles

- Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design (2010;340:c1066)
- CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials (2010;340:c332)
- CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials (2010;340:c869)
- Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes (2010;340:b5087)
- Economic impact of disease and injury: counting what matters (2010;340:c924)

For clinical practice and policy decision making, differences in prognosis, and the differences in absolute risk reduction that are associated with differences in prognosis, are far more important than relative subgroup effects for two reasons. First, identifiable and substantial differences in prognosis are fairly common, and one can be confident that potentially important differences in absolute effect across prognostic subgroups will occur. True subgroup differences in relative effects are, by contrast, fairly uncommon. Second, even if true differences in the effects of treatment across subgroups exist, those differences might not be large enough to mandate differences in management across those subgroups. This might be the case, for instance, if treatment is beneficial in all patients, but the size of treatment effect differs between subgroups. Assuming constant relative risk reductions, and using baseline risk to calculate absolute risk reductions for patient groups associated with validated differentiating prognostic characteristics, provides an optimum approach to trading off desirable and undesirable treatment results.[33]

We re-structured the checklist of items including the seven original and the four new criteria (table 1). This checklist is organised according to the design, analysis, and context of subgroup analysis.

The importance of these criteria varies, but the relative weight that should be applied to each criterion remains uncertain. If a credible weighting scheme could be established it might improve the efficiency and accuracy of judgments. One approach would be to develop a formal measurement instrument, allocating a specific weight to each criterion, and to validate the instrument by applying it to subgroup analyses that have been established to be real or spurious.

Contributors: All authors conceptualised the ideas in the manuscript and read and approved the manuscript. XS developed the first draft and incorporated comments from authors for successive drafts. GHG is the guarantor.

Funding: XS is supported by a grant from the National Natural Science Foundation of China (grant No. 70703025). MB is supported by Santésuisse and the Gottfried and Julia Bangerter-Rhyner Foundation.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

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 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.
4 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.
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 Altman DG. Within trial variation—a false trail? *J Clin Epidemiol* 1998;51:301-3.
7 Feinstein AR. The problem of cogent subgroups: a clinicostatistical tragedy. *J Clin Epidemiol* 1998;51:297-9.
8 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.
9 Horwitz RI, Singer BH, Makuch RW, Viscoli CM. Clinical versus statistical considerations in the design and analysis of clinical research. *J Clin Epidemiol* 1998;51:305-7.
10 Senn S, Harrell F. On wisdom after the event. *J Clin Epidemiol* 1997;50:749-51.
11 Smith GD, Egger M. Incommunicable knowledge? Interpreting and applying the results of clinical trials and meta-analyses. *J Clin Epidemiol* 1998;51:289-95.
12 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.
13 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;116:78-84.
14 Akl EA, Terrenato I, Barba M, Sperati F, Sempos EV, Muti P, et al. Low-molecular-weight heparin vs unfractionated heparin for perioperative thromboprophylaxis in patients with cancer: a systematic review and meta-analysis. *Arch Intern Med* 2008;168:1261-9.
15 Billingham LJ, Cullen MH. The benefits of chemotherapy in patient subgroups with unresectable non-small-cell lung cancer. *Ann Oncol* 2001;12:1671-5.
16 Bundy DG, Berkoff MC, Ito KE, Rosenthal MS, Weinberger M. Interpreting subgroup analyses: is a school-based asthma treatment program's effect modified by secondhand smoke exposure? *Arch Pediatr Adolesc Med* 2004;158:469-71.
17 Cranney A, Tugwell P, Wells G, Guyatt G. Meta-analyses of therapies for postmenopausal osteoporosis *I. Systematic reviews of randomized trials in osteoporosis: introduction and methodology. *Endocr Rev* 2002;23:496-507.
18 Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? *BMJ* 2001;322:989-91.
19 Hatala R, Keitz S, Wyer P, Guyatt G, for the Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine. 4. Assessing heterogeneity of primary studies in systematic reviews and whether to combine their results. *CMAJ* 2005;172:661-5.
20 Heckman GA, McKelvie RS. Necessary cautions when considering digoxin in heart failure. *CMAJ* 2007;176:644-5.
21 Kirpalani H, Barks J, Thorlund K, Guyatt G. Cooling for neonatal hypoxic ischemic encephalopathy: do we have the answer? *Pediatrics* 2007;120:1126-30.
22 Jaeschke R, O'Byrne PM, Mejza F, Nair P, Lesniak W, Brozek J, et al. The safety of long-acting beta-agonists among patients with asthma using inhaled corticosteroids: systematic review and metaanalysis. *Am J Respir Crit Care Med* 2008;178:1009-16.
23 Szczurko O, Cooley K, Busse JW, Seely D, Bernhardt B, Guyatt GH, et al. Naturopathic care for chronic low back pain: a randomized trial. *PLoS One* 2007;2:e919.
24 Bhandari M, Guyatt G, Tornetta P 3rd, Schemitsch EH, Swiontkowski M, Sanders D, et al. Randomized trial of reamed and unreamed intramedullary nailing of tibial shaft fractures. *J Bone Joint Surg Am* 2008;90:2567-78.
25 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;166:2307-13.
26 Van den Berghe G, Wilmer A, Hermans G, Meersseman W, Wouters PJ, Milants I, et al. Intensive insulin therapy in the medical ICU. *N Engl J Med* 2006;354:449-61.
27 Chan A-W. Bias, spin, and misreporting: time for full access to trial protocols and results. *PLoS Med* 2008;5:e230.
28 Russell JA, Walley KR, Singer J, Gordon AC, Hebert PC, Cooper DJ, et al. Vasopressin versus norepinephrine infusion in patients with septic shock. *N Engl J Med* 2008;358:877-87.
29 Berger JS, Roncaglioni MC, Avanzini F, Pangrazzi I, Tognoni G, Brown DL. Aspirin for the primary prevention of cardiovascular events in women and men: a sex-specific meta-analysis of randomized controlled trials. *JAMA* 2006;295:306-13.
30 Ridker PM, Cook NR, Buring JE. Use of aspirin as primary prevention of cardiovascular events. *JAMA* 2006;296:391.
31 Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004;66:411-21.
32 Thatcher N, Chang A, Parikh P, Rodrigues Pereira J, Ciuleanu T, von Pawel J, et al. Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer). *Lancet* 2005;366:1527-37.
33 Dans AL, Dans LF, Guyatt G. Applying results to individual patients. 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: 273-89.