THE
ORTHOPAEDIC
FORUM

Is a Subgroup Claim Believable?

A User's Guide to Subgroup Analyses in the Surgical Literature

By the Study to Prospectively Evaluate Reamed Intramedullary Nails in Tibial Fractures (SPRINT) Investigators*

Investigation performed at the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

Subgroup analyses are often reported in randomized controlled trials and meta-analyses. Apparent subgroup effects may, however, be misleading. Surgeons may therefore find it challenging to decide whether to believe a claim of subgroup effect (i.e., an apparent difference in treatment effect between subgroups of the study population). In the present study, we introduce seven widely used criteria to assess subgroup analyses in the surgical literature and include two examples of subgroup analyses from a large randomized trial to elaborate on the use of these criteria. Typically, inferences regarding subgroup effects are stronger if the comparison is made within rather than between studies, if the test for interaction suggests that chance is an unlikely explanation for apparent differences, if the subgroup hypothesis was specified a priori, if it was one of a small number of hypotheses tested, if the difference in effect between subgroup categories is large, if it is consistent across studies, and if there is indirect evidence supporting the difference (a biological rationale).

When testing the impact of surgical interventions, investigators may examine whether the effects differ between subgroups of patients or ways of administering an intervention—so-called subgroup analysis. For instance, in a randomized trial of removable splinting compared with casting for wrist buckle fractures in children, children with moderate injury (but not those with mild or severe injury) in the splint group had a larger change in scores on the Activity Scales for Kids than did the casting group[1]. In another example, a meta-analysis of sutures compared with staples for skin closure in orthopaedic surgery, the risk of a wound infection developing in patients with hip surgery (but not in other groups) was four times greater after staple closure than after suture closure[2].

Typically, the primary hypothesis of a randomized trial is to investigate the effect of treatment; subgroup analyses form secondary hypotheses in these studies. Such secondary subgroup analyses are common in randomized trials. A recent survey found that 38% (twenty-seven) of seventy-two surgical trials included subgroup analyses, and 57.4% (thirty-one) of fifty-four reported subgroup analyses claimed subgroup effects (that is, that the effect differed across subgroup categories)[3]. On the one hand, the real subgroup effects are important and informative, which allows targeting use of therapies to individual patients to achieve optimal treatments.

Disclosure: In support of their research for or preparation of this work, one or more of the authors received, in any one year, outside funding or grants in excess of $10,000 from the Canadian Institutes of Health Research (MCT-38140), National Institutes of Health (NIAMS-072; R01 AR48529), Orthopaedic Research and Education Foundation (American Academy of Orthopaedic Surgeons), Orthopaedic Trauma Association, Hamilton Health Sciences research grant, Zimmer, a Canada Research Chair in Musculoskeletal Trauma at McMaster University, and the National Natural Science Foundation of China (70703025). In addition, one or more of the authors or a member of his or her immediate family received, in any one year, payments or other benefits in excess of $10,000 or a commitment or agreement to provide such benefits from a commercial entity (Smith & Nephew).

*The Writing Committee included Xin Sun, PhD, Diane Heels-Ansdell, MSc, Sheila Sprague, MSc, Mohit Bhandari, MD, MSc, Stephen D. Walter, PhD, David Sanders, MD, Emil Schemitsch, MD, Paul Tornetta III, MD, Marc Swiontkowski, MD, and Gordon Guyatt, MD, MSc. Please see note preceding reference section for additional details regarding the authors and investigators.

On the other hand, the apparent differential effects among subgroups may be spurious, and the application of apparent subgroup findings to individual patients will be misleading. Therefore, an assessment of whether an apparent subgroup effect is spurious becomes an impetus. Many reasons may explain that an apparent subgroup effect is spurious. Some result from the fact that subgroup analysis is typically underpowered, involves multiplicity of testing, and often serves to generate a hypothesis. Some may be beyond the subgroup analysis itself. For instance, a poorly designed and conducted trial may result in misleading differential effects among subgroups of patients.

Typically, judging whether there is, or is not, a difference of effect between study subgroups involves uncertainty. Often, we cannot absolutely accept, or reject, a putative subgroup effect. What we can do is place the likelihood that a subgroup effect is real on a continuum from highly plausible to extremely unlikely.

In an effort to help clinicians assess the credibility of a putative subgroup effect, Oxman and Guyatt[4] 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 is the putative subgroup effect. Since then, these criteria have undergone rigorous assessment by methodologists and biostatisticians and have been used widely to assess the credibility of hypothesized subgroup effects in the medical literature[5-14]. We acknowledge that the assessment of subgroup credibility involves judgment, and may be fallible. Users such as clinicians who have limited training in research methodology, however, may find it useful as a guide to the interpretation of subgroup effects.

In this article, we present these seven criteria to the orthopaedic audience and illustrate their application to two possible subgroup effects from one of our orthopaedic trials (SPRINT[15]).

The SPRINT Trial: An Overview

The Study to Prospectively Evaluate Reamed Intramedullary Nails in Tibial Fractures (SPRINT) is a multicenter, blinded, concealed, randomized trial[15] in which 1319 skeletally mature patients with open (Gustilo Type I to IIIB) or closed (Tscherne Type 0 to 3) fractures of the tibial shaft underwent intramedullary nailing with either reamed or unreamed insertion. Randomization was stratified by center and type of fracture (closed compared with open). The primary outcome was a composite of reoperations to promote healing, treat infection, or preserve the limb within one year of follow-up. Reoperations included specific procedures for bone-grafting, implant exchange, dynamization of the fracture in the operating room or in the outpatient clinic, fasciotomy for intraoperative or postoperative compartment syndrome, drainage of a hematoma, and autodynamization. Of the 1319 randomized patients, 1226 (93%; 622 in the reamed intramedullary nailing group and 604 in the unreamed intramedullary nailing group) completed the one-year follow-up evaluation and were included in the analysis. There was no significant difference in the primary end point between reamed and unreamed intramedullary nailing (relative risk [RR], 0.92; 95% confidence interval [CI], 0.74 to 1.14).

Subgroup Analyses in the SPRINT Trial

Prior to the study, we specified seven subgroup hypotheses and used the test of interaction for each of the subgroup hypotheses on the basis of strong biological rationale and previous studies[16-19]. Our primary interest was to examine whether the treatment effects of reamed compared with unreamed nailing differ in closed compared with open factures (the results are described in our previous study[15]). Figure 1 shows the p value and treatment effect estimate for each subgroup and the p value for the test of interaction. Treatment effects differed significantly between open and closed fractures (test for interaction, $p = 0.011$). In open fractures, no significant difference was detected between reamed and unreamed nailing with respect to reoperations (RR, 1.27; 95% CI, 0.91 to 1.78). In closed fractures, reamed nailing reduced the risk of reoperations compared with unreamed nailing (RR, 0.67; 95% CI, 0.47 to 0.96).

After the analysis of those seven prespecified subgroup hypotheses, we additionally chose five variables, constituting five hypotheses, for subgroup analyses based on unexpected findings from an analysis for an advanced statistics course that examined the predictors to the reoperations (unpublished data). We found that treatment effects differed significantly between current smokers compared with nonsmokers or former smokers ($p = 0.0013$). In patients who were current smokers, reamed nailing had a higher risk of reoperation than unreamed nailing (RR, 1.56; 95% CI, 1.04 to 2.36). In nonsmokers or former smokers, reamed nailing had a lower risk of reoperation than unreamed nailing (RR, 0.68; 95% CI, 0.50 to 0.92; Fig. 1).

To explore the biological rationale supporting the apparent smoking subgroup effect, we surveyed four surgeon members of the SPRINT trial Steering Committee and six other orthopaedic experts, in a blinded fashion, regarding their expectations on the possibility of finding a subgroup effect in the five post hoc hypotheses. We first asked the members to rate the likelihood of the five post hoc subgroup hypotheses having a subgroup effect. We then presented a figure showing a significant subgroup effect stratified by the open and closed fracture type, which was from the smoking subgroup analysis. In the figure, we blinded the variable name (i.e., smoking) and coded the smoking subgroup categories as A and B. We asked the members to rate the possibility of those five hypotheses having the presented subgroup effect. In two blinded surveys, smoking was rated with the highest possibility.

We subsequently informed the committee members and the other orthopaedic experts that the presented subgroup effect was from the smoking hypothesis. The categories of smoking status (coded as A and B), however, remained blinded in relation to each subgroup. We urged them to judge which category of smoking status the presented subgroup was in and to provide the rationale for their

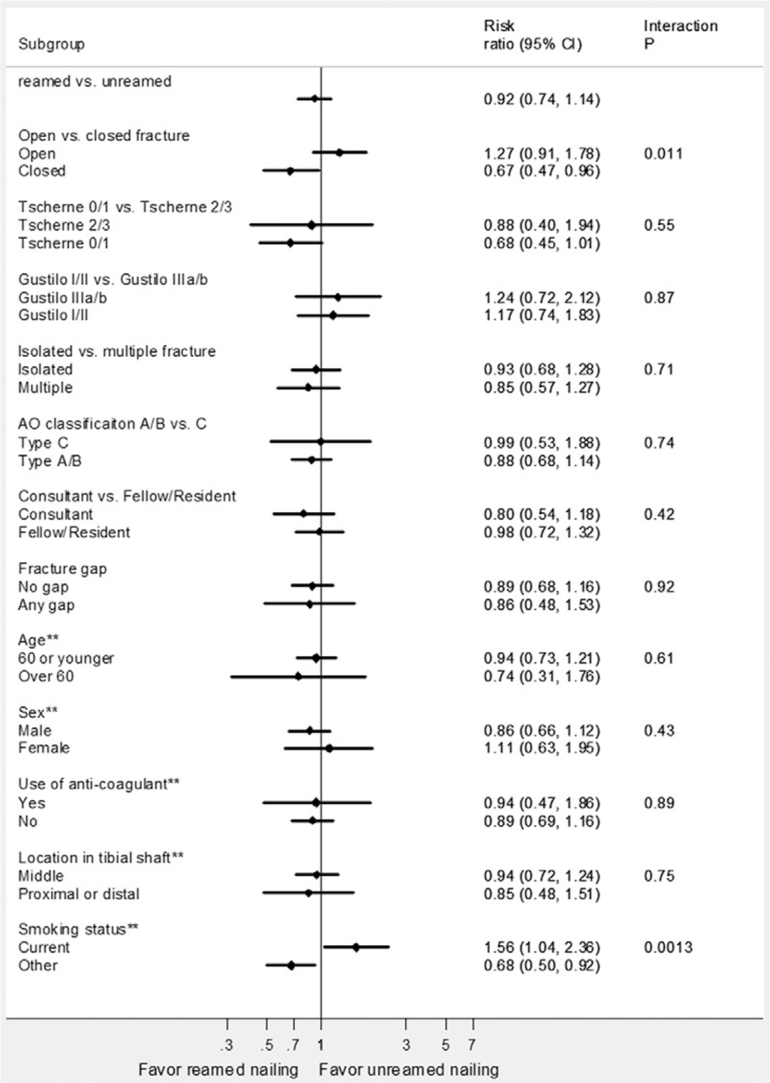

Fig. 1

Subgroup analyses for the primary outcome (i.e., reoperation). The first point estimate and confidence interval in this figure indicates the main effect. The subsequent pairs of point estimates and confidence intervals indicate the effect of reamed compared with unreamed nailing on reoperation in twelve subgroup variables. **The subgroup analyses were conducted post hoc. Subgroup analysis by the Tscherne type included patients with closed facture only, and analysis by Gustilo type included open fracture only. In our analysis that included significant and nonsignificant interactions, these two interactions were not included in the regression model, resulting in ten interaction terms being included in the model. (Reprinted, with modification, from: 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, with permission from BMJ Publishing Group Ltd.)

judgments. Nine effective responses were collected. Four provided a biological rationale that supported the apparent smoking subgroup effect, three provided a rationale against the finding, and two argued they could provide a biological rationale that supported or was against the apparent smoking subgroup effect.

Given what we know about the history of subgroup analyses—many apparent subgroup effects have proved spurious, and relatively few have been confirmed—can we trust that the apparent treatment effect differs in patients who smoke compared with those who do not? We applied seven criteria (Table I) to this analysis. We also contrasted it with

TABLE I Subgroup Analyses of Fracture Type and Smoking Status

	Subgroup Effect of Fracture Type*	Subgroup Effect of Smoking Status*
Is the difference suggested by comparisons within rather than between studies?	Yes	Yes
Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?	P = 0.011†	P = 0.0013†
Were the subgroup hypothesis and its direction specified a priori?	Yes	No
Is it one of a small number of subgroup hypotheses tested?	1 of 7 subgroup hypotheses	1 of 12 subgroup hypotheses
Is the magnitude of the subgroup effect large?	60% difference in RR reduction	88% difference in RR reduction
Is the observed differential effect consistent across studies?	No	No evidence
Is there indirect evidence supporting the hypothesized differential effects?	Yes	No

*RR = relative risk. †Test for interaction.

the subgroup effect in patients with open compared with closed fractures.

Whether the Subgroup Effect Is Believable: Seven Criteria for the Smoking and Open Fracture Compared with Closed Fracture Subgroup Analyses

Is the Difference Suggested by Comparisons within Rather Than Between Studies?

Inferences regarding differential effects on the basis of between-study comparisons are much weaker than those based on within-study comparisons. By within-study comparisons, we mean a situation in which patients in the subgroups under consideration were all enrolled in the same trial or trials. By between-study comparisons, we mean a situation in which patients in the subgroups of interest were enrolled in different trials, each of which addressed only one of the subgroups of interest.

Consider a situation in which Trial A enrolls exclusively patients with closed fractures and finds that reamed intramedullary nailing is superior to unreamed intramedullary nailing. Trial B enrolls only those with open fractures and reports results favoring unreamed intramedullary nailing. What are the possible explanations of the finding?

Although one might be tempted to attribute the difference to the different study samples—those with closed and open fractures—there are many other possibilities. Trial A may have minimized the risk of bias by concealed allocation, blinding outcome assessors, and achieving nearly complete follow-up. Trial B may have failed in all of these regards, and the different risk of bias may be responsible for the different results. There may be other differences in the patients' characteristics (age or degree of osteoporosis) that could explain the differences. Surgeons in Trial A may have been more experienced in the reamed procedure and surgeons in Trial B, in the unreamed procedure. Trial A may have measured only short-term outcomes, while Trial B followed patients for a longer period of time. Finally, chance may explain the apparent difference of treatment effects between subgroups. We define chance as an apparent difference in which the underlying truth is that there is no difference of effect.

What if, however, patients with open and closed fractures are participating in a single trial in which the reamed procedure appeared superior in those with closed fractures and the unreamed procedure appeared superior in those with open fractures? Trial methods, eligibility criteria, and surgeon expertise are likely to be identical for all patients. When, as is likely to be the case within a single trial, methods are identical for the patients in the subgroups of interest, we are left with only two compelling explanations: the subgroup effect is real, or chance is responsible for the apparent difference.

In our example, comparisons of treatment effects between closed and open fractures, and between current smokers and nonsmokers or former smokers, are made within a single study. This increases the credibility of the presence of the differential effects between current smokers and nonsmokers or former smokers and between open and closed fractures.

Does the Interaction Test Suggest a Low Probability That Chance Explains the Apparent Subgroup Effect?

When examining subgroup hypotheses, one must address the probability that the observed differences in effects can be explained by chance. The statistical approach that addresses this fundamental issue is called a test for interaction (the interaction meaning that the effect differs across subgroup categories such as patients with open fractures compared with patients with closed fractures)[20]. Typically, a test of interaction compares the estimated treatment effects—measured as relative risk, odds ratio, or difference in mean change—in subgroups of the study population, and addresses and presents the possibility of differences as great as or greater than those observed if there is truly no difference in effect between subgroups. An inappropriate, although frequently used, approach is to separately test the significance of the effect in each subgroup category. Such analyses fail to address the fundamental issue: can the difference in effect between subgroup categories be explained by chance?

The null hypothesis of the test for interaction is that there is no difference 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. Typically, investigators use the usual threshold p value of 0.05. Inevitably, the choice of threshold involves subjective judgment. An approach that avoids the arbitrariness of a single threshold is to consider that the larger the p value (e.g., >0.1), the more likely that chance explains the apparent difference in subgroup effects and the smaller the p value, the less likely that chance explains the apparent difference. The p value for the test of interaction is associated with sample size; the larger the sample size, the more likely the null hypothesis will be rejected if a subgroup effect exists. Sometimes investigators may examine post hoc the power of testing the subgroup hypotheses. This endeavor does not help in deciding on the credibility of the subgroup analysis: the issue of relevance is the possibility that chance could explain the observed findings.

Our analysis showed a small interaction p value for the test of the subgroup hypothesis by fracture type ($p = 0.01$). The p value is small enough to ensure that the subgroup effect is unlikely to be explained by chance; the p value for the test of the smoking hypothesis was even smaller ($p = 0.0013$). These results strengthen the inferences that the two subgroup hypotheses represent real effects.

Are the Subgroup Hypothesis and Its Direction Specified a Priori?

One may specify the subgroup hypothesis before or after the data are disclosed. Typically, a priori specification, which is driven by previous research evidence and/or biological rationale, represents careful consideration by the investigators regarding the possibilities of a significant interaction. At the other extreme, conducting subgroup analyses post hoc is likely to be data-driven: investigators highlight an apparent subgroup effect only after discovering it in the data, the so-called "data-fishing" approach. Accurate specification of the direction of the subgroup hypothesis a priori (for instance, specifying that reamed nailing will be superior in closed fractures and unreamed nailing in open fractures, rather than suggesting only that effects may differ in open and closed fractures) further strengthens the credibility of the subgroup inference (and lack of specification—or getting the direction wrong—undermines it). A desirable approach is for researchers to state explicitly in study protocols their subgroup hypotheses and the direction of the hypothesized subgroup effect.

Closed Compared with Open Fractures

Our subgroup hypothesis by fracture type (i.e., open and closed) was specified at the stage of trial design and was the subgroup hypothesis of primary interest. Not only was the hypothesis a priori, but the direction of the effect based on a compelling biological rationale was correctly specified. This enhances its credibility.

Smoking

The smoking subgroup hypothesis was specified only after the initial analysis was complete and would never have been explored had it not been part of an exercise for an advanced statistics course. At the time the smoking analysis was initially conducted, we had no hypothesis about its direction. In our blinded surveys, orthopaedic surgeons chose smoking as the most probable of a number of additional hypotheses conducted as part of the exercise for the statistics course. However, surgeons were split in choosing the direction that the effect should go (i.e., whether smokers would do better with reamed or unreamed nailing). The uncertainty about the direction of subgroup effect among expert orthopaedic surgeons suggests the absence of a compelling biological rationale. To the extent that one finds the presence of a compelling biological rationale important (some may not), uncertainty about the direction in effect would reduce the strength of inference regarding the presence of an underlying subgroup effect.

Is This One of a Small Number of Subgroup Hypotheses Tested?

Typically, one test of interaction carries a small risk of a false-positive finding. Multiple tests of interactions increase the possibility of a false-positive conclusion, and the more tests conducted, the greater the problem. Thus, a large number of tests of subgroup hypotheses may compromise the strength of a priori specification, and the credibility of significant subgroup effects decreases.

Closed Compared with Open Fractures

Our subgroup hypothesis by fracture type was one of the seven a priori hypotheses tested. It is also the subgroup hypothesis of primary interest as reflected in our stratification of randomization by open and closed fracture. This strengthens its credibility.

Smoking

After the data were disclosed and treatment allocation unblinded, we tested five post hoc subgroup hypotheses (beyond the seven we had generated a priori). These hypotheses were specified independently of the previous seven; the hypothesis that smoking might influence the magnitude of effect was one of the five post hoc subgroup hypotheses tested. One could view this hypothesis as one of twelve (i.e., seven a priori and five post hoc, and the relatively large number would weaken the subgroup inference) or as one of five post hoc hypotheses (in which case the relatively small number would not weaken the inference as much).

Is the Magnitude of Subgroup Effect Large?

The apparent treatment effect will inevitably differ among subgroup categories (e.g., open compared with closed and current smokers compared with nonsmokers or former smokers). Small differences in effects across subgroup categories are likely explained by chance; the larger the difference in effects between subgroup categories, the more likely the difference represents a true interaction. Large differences in the presence of small sample sizes may, however, occur by chance.

To determine the possibility that chance explains the apparent difference, an alternative to the statistical test of heterogeneity is to consider the confidence interval around the magnitude of subgroup effect[21]. For presenting the magnitude of the difference for a continuous variable, authors can use differences. If the outcome is a binary variable, they may present the ratio of relative risks (or ratio of hazard ratios if the outcome is time-to-event data). In the presence of a qualitative interaction (i.e., treatment is beneficial in one subgroup, whereas it is harmful in another), however, interpretation of a confidence interval around the magnitude becomes problematic. In this situation, we recommend considering a point estimate only.

Consider that a subgroup analysis shows that a treatment reduces the risk of pulmonary embolism by 58% in patients over sixty years old (RR, 0.42; 95% CI, 0.23 to 0.75) and by 20% in patients sixty years old or less (RR, 0.80; 95% CI, 0.68 to 0.95; test for interaction, p = 0.034). This indicates that the treatment effect (i.e., RR) on the reduction of pulmonary embolism is nearly twice as great in patients over sixty years old than in others (i.e., $0.80/0.42 \approx 2$); the 95% confidence interval around this ratio is 1.05 to 3.58.

Both the smoking and fracture type subgroup analyses yielded large and qualitative subgroup effects. Reamed intramedullary nailing reduced the relative risk of reoperation by 33% in closed fractures, but increased the risk by 27% in open fractures. Reamed nailing reduced the relative risk of reoperation by 32% in nonsmokers or former smokers, whereas it increased the risk by 56% in current smokers. These large differences of treatment effects across subgroups increase the credibility of the subgroup hypotheses.

Is the Observed Differential Effect Consistent Across Studies? Even small p values do not exclude the possibility that chance is the true explanation for an apparent subgroup effect; this is particularly true when investigators test multiple hypotheses. The more often, and more consistently, the subgroup effect is replicated in additional trials, the stronger the inference. Indeed, failure to reproduce an apparent subgroup effect has revealed the spurious nature of many previous subgroup claims. Ideally, a rigorous systematic review, which provides an overview of the subgroup findings across studies, will confirm or refute the consistency of subgroup effects. Sometimes, however, studies included in systematic reviews and meta-analyses may not provide sufficient data regarding results in the patient subgroups of interest to adequately address the issue. In such situations, meta-analyses can neither confirm nor refute the reproducibility of a subgroup analysis suggested by a single trial.

Closed Compared with Open Fractures

Our meta-analysis of five randomized trials[22-26] examined the relative impact of reamed compared with unreamed nailing on the reoperation rate in open and closed fractures (Fig. 2). This review described studies suggesting that reamed nailing was superior in both open and closed fractures. However, the previous studies were small and suffered from important limitations including lack of concealment, lack of blinding of the

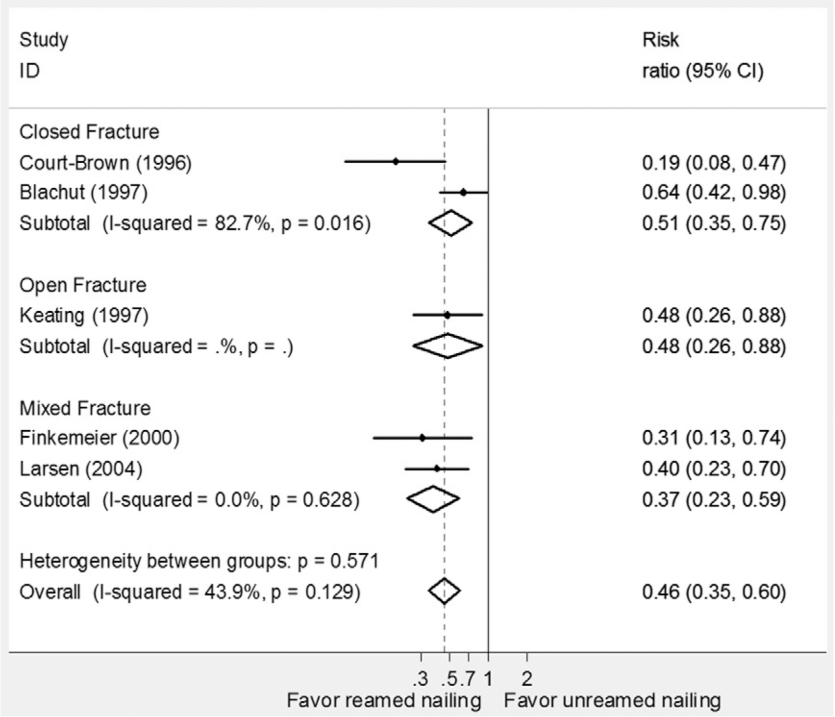

Fig. 2

Meta-analysis of randomized trials of reamed nailing compared with unreamed nailing in patients with a tibial shaft fracture[22-26].

outcome assessment, and substantial loss to follow-up. The data provide, however, no support for the subgroup hypothesis (i.e., the other studies fail to reproduce the subgroup effect of differences in the impact of reamed compared with unreamed nailing in closed compared with open fractures).

Smoking

We did not identify any observational or randomized trial evidence addressing the possibility of differential effects by smoking status. Other studies therefore fail to provide supporting evidence for our inferences regarding the smoking subgroup effect.

Is There Indirect Evidence Supporting the Hypothesized Differential Effects?

The presence of indirect evidence strengthens the beliefs of hypothesized subgroup effects. Typically, indirect evidence comprises several types of evidence, including basic science studies, physiological studies, and animal studies. Another way to describe this criterion would be: Is there a strong biological rationale for the putative subgroup effect?

The search for a biological rationale to explain an apparent subgroup effect is, given sufficient imagination, almost always successful. This limits the value of this criterion in providing compelling support for a subgroup hypothesis.

Closed Compared with Open Fractures

Intact or minimally damaged soft tissue and periosteum in the closed fractures might result in greater tolerability of reamed nailing. Thus, the added stability of reamed nailing might prove advantageous. On the other hand, devascularization in open fractures may render the bone vulnerable to the vascular compromise associated with reaming and may severely compromise the benefit of reamed nailing[18].

Smoking

Neither animal nor other relevant studies exist to support the smoking subgroup hypothesis. This, however, did not impair the ability of orthopaedic surgeons, blinded to the direction of the apparent effect, to generate a compelling biological rationale. Unfortunately, other surgeons generated an equally compelling rationale for an effect in the opposite direction.

Summary of the Credibility of Observed Subgroup Effects

Clinicians and investigators are often excessively ready to believe apparent subgroup effects. We believe that clinicians will make fewer mistakes if they err on the side of skepticism regarding explanations of heterogeneity in study results.

Nevertheless, the differential effect by reamed compared with unreamed nailing procedures for open compared with closed fractures may be real. The hypothesis has a strong biological rationale, is supported by a within-trial comparison, was our subgroup hypothesis of primary interest with a correctly hypothesized direction, presented a large difference of effects between open and closed fractures, and is unlikely to be explained by chance. Although the external evidence fails to support the differential effect, this subgroup analysis meets the other six criteria (Table I).

Inferences regarding the observed differential effect by smoking status are weaker. On the one hand, data supporting the hypothesis come from a within-study comparison, there are large differences between the subgroups, and the p value associated with the test of interaction is very low. On the other hand, the hypothesis was not supported by a compelling a priori specification, has no supporting data from other studies, and has a dubious biological rationale. Given some strong support, but failure to meet a number of important criteria, we cannot accept or clearly reject the possibility that the effects of reamed and unreamed nailing differ by smoking status.

In our analyses, we did not adjust the interaction p value for the multiple tests of interactions. While the strength of inferences decreases with multiple tests of interactions, altering the threshold p value to reflect the multiple testing remains controversial. In general, application of a multiple-comparisons correction is intended to control the overall type-I error rate for some set (or "family") of comparisons that are made. However, as Cox pointed out, "A probability of error referring to the simultaneous correctness of a set of statements seems relevant only if a certain conclusion depends directly on the simultaneous correctness of a set of the individual statements. . . . The fact that a probability can be calculated for the simultaneous correctness of a large number of statements does not usually make that probability relevant for the measurement of the uncertainty of one of the statements."[27] Others have supported this view in the specific context of clinical trials, arguing that "multiplicity adjustments may not be necessary if marginal, or separate, test results are *interpreted marginally* and have different implications. . ."[28]. By adopting well-defined hypotheses, one can claim that "we have the capacity to interpret test results marginally and to draw inferences accordingly."[28] This avoids the problem that by emphasizing the overall test error rate we may obscure and "lose focus of the clinical questions of main interest."[28]

Quite apart from the philosophical debate about the need to adjust for multiplicity or not, there is the practical problem that the relevant set of subgroup hypotheses under consideration may not be static. To adopt a strict policy of multiple-comparisons adjustment would imply the need to constantly modify the adjustment strategy as hypotheses are added or removed from the set. If the subgroup hypotheses are indeed well considered, and they represent separate scientific questions about the data, then the relevance of the result of one hypothesis test to the others seems limited. Instead, we prefer the notion of focusing on marginal interpretations of each separate hypothesis; therefore, we recommend against making multiple-comparisons adjustments to the hypothesis testing approach. At the same time, we acknowledge the issue of capitalizing on the play of chance. Inferences are stronger with a small number of a priori hypotheses than with a larger number.

One may learn from the SPRINT example that a convincing claim of a subgroup effect typically comes from a within-study comparison, has a significant interaction, is unequivocally specified a priori, is one of the small number of hypotheses tested, presents a large difference of effects, is supported by the external evidence, and has compelling biological rationale. The more criteria a claim meets, the more believable it is. Thus, we are here inclined to believe the subgroup effect by open and closed fracture as it satisfies all the important criteria. Conversely, we believe the smoking subgroup effect is uncertain mainly because it fails the criterion of a priori specification, which is key to the credibility of a subgroup effect.

Note: Details regarding the authors and investigators are provided below.

Author Contributions

Writing Committee: The Writing Committee (Xin Sun, PhD [Chair], Diane Heels-Ansdell, MSc, Sheila Sprague, MSc, Mohit Bhandari, MD, MSc, Stephen D. Walter, PhD, David Sanders, MD, Emil Schemitsch, MD, Paul Tornetta III, MD, Marc Swiontkowski, MD, and Gordon Guyatt, MD, MSc) assumes responsibility for the overall content and integrity of the manuscript.

Study concept and design: Gordon Guyatt, Xin Sun, Mohit Bhandari, Stephen Walter, Paul Tornetta III, Emil Schemitsch, David Sanders, Marc Swiontkowski.

Analysis and interpretation of data: Xin Sun, Gordon Guyatt, Diane Heels-Ansdell, Sheila Sprague, Stephen Walter, Mohit Bhandari, Paul Tornetta III, Emil Schemitsch, David Sanders, Marc Swiontkowski.

Drafting of the manuscript: Xin Sun, Diane Heels-Ansdell, Sheila Sprague, Mohit Bhandari, Gordon Guyatt.

Critical revision of the manuscript for important intellectual content: Mohit Bhandari, Gordon Guyatt, Stephen Walter, Paul Tornetta III, Emil Schemitsch, David Sanders, Marc Swiontkowski.

Statistical analysis: Xin Sun, Diane Heels-Ansdell, Stephen Walter, Gordon Guyatt, Mohit Bhandari.

Obtained funding: Mohit Bhandari, Gordon Guyatt, Marc Swiontkowski, Paul Tornetta III.

Study supervision: Mohit Bhandari, Gordon Guyatt, Stephen Walter, Paul Tornetta III, Emil Schemitsch, David Sanders, Marc Swiontkowski.

Role of Sponsors/Funders

The funding sources had no role in the design or conduct of the study; the collection, management, analysis, or interpretation of the data; or the preparation, review, or approval of the manuscript.

SPRINT Investigators

The following persons participated in the SPRINT Study:

Study trial co-principal investigators: Mohit Bhandari, Gordon Guyatt.

Steering Committee: Gordon Guyatt (Chair), Mohit Bhandari, David W. Sanders, Emil H. Schemitsch, Marc Swiontkowski, Paul Tornetta III, Stephen Walter.

Central Adjudication Committee: Gordon Guyatt (Chair), Mohit Bhandari, David W. Sanders, Emil H. Schemitsch, Marc Swiontkowski, Paul Tornetta III, Stephen Walter.

Steering / Adjudication / Writing Committee: Gordon Guyatt (Chair), Mohit Bhandari, David W. Sanders, Emil H. Schemitsch, Marc Swiontkowski, Paul Tornetta III, Stephen Walter.

SPRINT Methods Center staff:

McMaster University, Hamilton, Ontario: Sheila Sprague, Diane Heels-Ansdell, Lisa Buckingham, Pamela Leece, Helena Viveiros, Tashay Mignott, Natalie Ansell, Natalie Sidorkewicz.

University of Minnesota, Minneapolis, Minnesota: Julie Agel.

Data Safety and Monitoring Board (DSMB): Claire Bombardier (Chair), Jesse A. Berlin, Michael Bosse, Bruce Browner, Brenda Gillespie, Peter O'Brien.

Site Audit Committee: Julie Agel, Sheila Sprague, Rudolf Poolman, Mohit Bhandari.

Study Sites

London Health Sciences Centre / University of Western Ontario, London, Ontario: David W. Sanders, Mark D. Macleod, Timothy Carey, Kellie Leitch, Stuart Bailey, Kevin Gurr, Ken Konito, Charlene Bartha, Isolina Low, Leila V. MacBean, Mala Ramu, Susan Reiber, Ruth Strapp, Christina Tieszer.

Sunnybrook Health Sciences Centre / University of Toronto, Toronto, Ontario: Hans Kreder, David J.G. Stephen, Terry S. Axelrod, Albert J.M. Yee, Robin R. Richards, Joel Finkelstein, Richard M. Holtby, Hugh Cameron, John Cameron, Wade Gofton, John Murnaghan, Joseph Schatztker, Beverly Bulmer, Lisa Conlan.

Hospital du Sacre Coeur de Montreal, Montreal, Quebec: Yves Laflamme, Gregory Berry, Pierre Beaumont, Pierre Ranger, Georges-Henri Laflamme, Alain Jodoin, Eric Renaud, Sylvain Gagnon, Gilles Maurais, Michel Malo, Julio Fernandes, Kim Latendresse, Marie-France Poirier, Gina Daigneault.

St. Michael's Hospital / University of Toronto, Toronto, Ontario: Emil H. Schemitsch, Michael M. McKee, James P. Waddell, Earl R. Bogoch, Timothy R. Daniels, Robert R. McBroom, Robin R. Richards, Milena R. Vicente, Wendy Storey, Lisa M. Wild.

Royal Columbian Hospital / University of British Columbia, Vancouver, British Columbia: Robert McCormack, Bertrand Perey, Thomas J. Goetz, Graham Pate, Murray J. Penner, Kostas Panagiotopoulos, Shafique Pirani, Ian G. Dommisse, Richard L. Loomer, Trevor Stone, Karyn Moon, Mauri Zomar.

Wake Forest Medical Center / Wake Forest University Health Sciences, Winston-Salem, North Carolina: Lawrence X. Webb, Robert D. Teasdall, John Peter Birkedal, David Franklin Martin, David S. Ruch, Douglas J. Kilgus, David C. Pollock, Mitchel Brion Harris, Ethan Ron Wiesler, William G. Ward, Jeffrey Scott Shilt, Andrew L. Koman, Gary G. Poehling, Brenda Kulp.

Boston Medical Center / Boston University School of Medicine, Boston, Massachusetts: Paul Tornetta III, William R. Creevy, Andrew B. Stein, Christopher T. Bono, Thomas A. Einhorn, T. Desmond Brown, Donna Pacicca, John B. Sledge III, Timothy E. Foster, Ilva Voloshin, Jill Bolton, Hope Carlisle, Lisa Shaughnessy.

Wake Medical Center, Raleigh, North Carolina: William T. Ombremskey, C. Michael LeCroy, Eric G. Meinberg, Terry M. Messer, William L. Craig III, Douglas R. Dirschl, Robert Caudle, Tim Harris, Kurt Elhert, William Hage, Robert Jones, Luis Piedrahita, Paul O. Schricker, Robin Driver, Jean Godwin, Gloria Hansley.

Vanderbilt University Medical Center, Nashville, Tennessee: William Todd Obremskey, Philip James Kregor, Gregory Tennent, Lisa M. Truchan, Marcus Sciadini, Franklin D. Shuler, Robin E. Driver, Mary Alice Nading, Jacky Neiderstadt, Alexander R. Vap.

MetroHealth Medical Center, Cleveland, Ohio: Heather A. Vallier, Brendan M. Patterson, John H. Wilber, Roger G. Wilber, John K. Sontich, Timothy Alan Moore, Drew Brady, Daniel R. Cooperman, John A. Davis, Beth Ann Cureton.

Hamilton Health Sciences, Hamilton, Ontario: Scott Mandel, R. Douglas Orr, John T.S. Sadler, Tousief Hussain, Krishan Rajaratnam, Bradley Petrisor, Mohit Bhandari, Brian Drew, Drew A. Bednar, Desmond C.H. Kwok, Shirley Pettit, Jill Hancock, Natalie Sidorkewicz.

Regions Hospital, Saint Paul, Minnesota: Peter A. Cole, Joel J. Smith, Gregory A. Brown, Thomas A. Lange, John G. Stark, Bruce Levy, Marc F. Swiontkowski, Julie Agel, Mary J. Garaghty, Joshua G. Salzman, Carol A. Schutte, Linda (Toddie) Tastad, Sandy Vang.

University of Louisville School of Medicine, Louisville, Kentucky: David Seligson, Craig S. Roberts, Arthur L. Malkani, Laura Sanders, Sharon Allen Gregory, Carmen Dyer, Jessica Heinsen, Langan Smith, Sudhakar Madanagopal.

Memorial Hermann Hospital, Houston, Texas: Kevin J. Coupe, Jeffrey J. Tucker, Allen R. Criswell, Rosemary Buckle, Alan Jeffrey Rechter, Dhiren Shaskikant Sheth, Brad Urquart, Thea Trotscher.

Erie County Medical Center / University of Buffalo, Buffalo, New York: Mark J. Anders, Joseph M. Kowalski, Marc S. Fineberg, Lawrence B. Bone, Matthew J. Phillips, Bernard Rohrbacher, Philip Stegemann, William M. Mihalko, Cathy Buyea.

University of Florida at Jacksonville, Jacksonville, Florida: Stephen J. Augustine, William Thomas Jackson, Gregory Solis, Sunday U. Ero, Daniel N. Segina, Hudson B. Berrey, Samuel G. Agnew, Michael Fitzpatrick, Lakina C. Campbell, Lynn Derting, June McAdams.

Academic Medical Center, Amsterdam, The Netherlands: J. Carel Goslings, Kees Jan Ponsen, Jan Luitse, Peter Kloen, Pieter Joosse, Jasper Winkelhagen, Raphaël Duivenvoorden.

University of Oklahoma Health Science Center, Oklahoma City, Oklahoma: David C. Teague, Joseph Davey, J. Andy Sullivan, William J.J. Ertl, Timothy A. Puckett, Charles B. Pasque, John F. Tompkins II, Curtis R. Gruel, Paul Kammerlocher, Thomas P. Lehman, William R. Puffinbarger, Kathy L. Carl.

University of Alberta / University of Alberta Hospital, Edmonton, Alberta: Donald W. Weber, Nadr M. Jomha, Gordon R. Goplen, Edward Masson, Lauren A. Beaupre, Karen E. Greaves, Lori N. Schaump.

Greenville Hospital System, Greenville, South Carolina: Kyle J. Jeray, David R. Goetz, David E. Westberry, J. Scott Broderick, Bryan S. Moon, Stephanie L. Tanner.

Foothills General Hospital, Calgary, Alberta: James N. Powell, Richard E. Buckley, Leslie Elves.

Saint John Regional Hospital, Saint John, New Brunswick: Stephen Connolly, Edward P. Abraham, Donna Eastwood, Trudy Steele.

Oregon Health and Sciences University, Portland, Oregon: Thomas Ellis, Alex Herzberg, George A. Brown, Dennis E. Crawford, Robert Hart, James Hayden, Robert M. Orfaly, Theodore Vigland, Maharani Vivekaraj, Gina L. Bundy.

San Francisco General Hospital, San Francisco, California: Theodore Miclau III, Amir Matityahu, R. Richard Coughlin, Utku Kandemir, R. Trigg McClellan, Cindy Hsin-Hua Lin.

Detroit Receiving Hospital, Detroit, Michigan: David Karges, Kathryn Cramer, J. Tracy Watson, Berton Moed, Barbara Scott.

Deaconess Hospital Regional Trauma Center/Orthopaedic Associates, Evansville, Indiana: Dennis J. Beck, Carolyn Orth.

Thunder Bay Regional Health Science Centre, Thunder Bay, Ontario: David Puskas, Russell Clark, Jennifer Jones.

Jamaica Hospital, Jamaica, New York: Kenneth A. Egol, Nader Paksima, Monet France.

Ottawa Hospital–Civic Campus, Ottawa, Ontario: Eugene K. Wai, Garth Johnson, Ross Wilkinson, Adam T. Gruszczynski, Liisa Vexler.

Xin Sun, PhD
Diane Heels-Ansdell, MSc
Stephen D. Walter, PhD
Gordon Guyatt, MD, MSc
Department of Clinical Epidemiology and Biostatistics,
McMaster University, HSC 2C,
1200 Main Street West,
Hamilton, ON L8N 3Z5,
Canada.
E-mail address for X. Sun: sunx26@mcmaster.ca

Sheila Sprague, MSc
Mohit Bhandari, MD, MSc
SPRINT Methods Center,
Department of Clinical Epidemiology and Biostatistics,
McMaster University, 293 Wellington Street North,
Suite 110, Hamilton, ON L8L 2X2, Canada

David Sanders, MD
London Health Sciences Centre,
4th Floor, Westminster Tower,
800 Commissioners Road East, London, ON N6A 4G5, Canada

Emil Schemitsch, MD
St. Michael's Hospital, 55 Queen Street East,
#800, Toronto, ON M5C 1R6, Canada

Paul Tornetta III, MD
Department of Orthopaedic Surgery, Boston Medical Center,
850 Harrison Avenue, Dowling 2 North, Boston, MA 02118

Marc Swiontkowski, MD
Department of Orthopaedic Surgery, University of Minnesota,
2512 South 7th Street, Suite R200, Minneapolis, MN 55454

References

1. Plint AC, Perry JJ, Correll R, Gaboury I, Lawton L. A randomized, controlled trial of removable splinting versus casting for wrist buckle fractures in children. Pediatrics. 2006;117:691-7.

2. Smith TO, Sexton D, Mann C, Donell S. Sutures versus staples for skin closure in orthopaedic surgery: meta-analysis. BMJ. 2010;340:c1199.

3. Bhandari M, Devereaux PJ, Li P, Mah D, Lim K, Schünemann HJ, Tornetta P 3rd. Misuse of baseline comparison tests and subgroup analyses in surgical trials. Clin Orthop Relat Res. 2006;447:247-51.

4. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med. 1992;116:78-84.

5. Akl EA, Terrenato I, Barba M, Sperati F, Sempos EV, Muti P, Cook DJ, Schünemann HJ. 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.

6. Billingham LJ, Cullen MH. The benefits of chemotherapy in patient subgroups with unresectable non-small-cell lung cancer. Ann Oncol. 2001;12:1671-5.

7. 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.

8. Cranney A, Tugwell P, Wells G, Guyatt G; Osteoporosis Methodology Group and The Osteoporosis Research Advisory Group. Meta-analyses of therapies for postmenopausal osteoporosis. I. Systematic reviews of randomized trials in osteoporosis: introduction and methodology. Endocr Rev. 2002;23:496-507.

9. 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.

10. Hatala R, Keitz S, Wyer P, Guyatt G; 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.

11. Heckman GA, McKelvie RS. Necessary cautions when considering digoxin in heart failure. CMAJ. 2007;176:644-5.

12. Kirpalani H, Barks J, Thorlund K, Guyatt G. Cooling for neonatal hypoxic ischemic encephalopathy: do we have the answer? Pediatrics. 2007;120:1126-30.

13. Nakaoka H, Takahashi T, Akiyama K, Cui T, Tajima A, Krischek B, Kasuya H, Hata A, Inoue I. Differential effects of chromosome 9p21 variation on subphenotypes of intracranial aneurysm: site distribution. Stroke. 2010;41:1593-8.

14. Szczurko O, Cooley K, Busse JW, Seely D, Bernhardt B, Guyatt GH, Zhou Q, Mills EJ. Naturopathic care for chronic low back pain: a randomized trial. PLoS One. 2007;2:e919.

15. Bhandari M, Guyatt G, Tornetta P, 3rd, Schemitsch EH, Swiontkowski M, Sanders D, Walter SD. Randomized trial of reamed and unreamed intramedullary nailing of tibial shaft fractures. J Bone Joint Surg Am. 2008;90:2567-78.

16. Bhandari M, Guyatt GH, Swiontkowski MF, Schemitsch EH. Treatment of open fractures of the shaft of the tibia. J Bone Joint Surg Br. 2001;83:62-8.

17. Bhandari M, Guyatt GH, Tong D, Adili A, Shaughnessy SG. Reamed versus nonreamed intramedullary nailing of lower extremity long bone fractures: a systematic overview and meta-analysis. J Orthop Trauma. 2000;14:2-9.

18. Whiteside LA, Ogata K, Lesker P, Reynolds FC. The acute effects of periosteal stripping and medullary reaming on regional bone blood flow. Clin Orthop Relat Res. 1978;131:266-72.

19. SPRINT Investigators, Bhandari M, Guyatt G, Tornetta P 3rd, Schemitsch E, Swiontkowski M, Sanders D, Walter SD. Study to prospectively evaluate reamed intramedually nails in patients with tibial fractures (S.P.R.I.N.T.): study rationale and design. BMC Musculoskelet Disord. 2008;9:91.

20. Matthews JN, Altman DG. Statistics notes. Interaction 2: compare effect sizes not P values. BMJ. 1996;313:808.

21. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ. 2003;326:219.

22. Blachut PA, O'Brien PJ, Meek RN, Broekhuyse HM. Interlocking intramedullary nailing with and without reaming for the treatment of closed fractures of the tibial shaft. A prospective, randomized study. J Bone Joint Surg Am. 1997;79:640-6.

23. Court-Brown CM, Will E, Christie J, McQueen MM. Reamed or unreamed nailing for closed tibial fractures. A prospective study in Tscherne C1 fractures. J Bone Joint Surg Br. 1996;78:580-3.

24. Finkemeier CG, Schmidt AH, Kyle RF, Templeman DC, Varecka TF. A prospective, randomized study of intramedullary nails inserted with and without reaming for the treatment of open and closed fractures of the tibial shaft. J Orthop Trauma. 2000;14:187-93.

25. Keating JF, O'Brien PJ, Blachut PA, Meek RN, Broekhuyse HM. Locking intramedullary nailing with and without reaming for open fractures of the tibial shaft. A prospective, randomized study. J Bone Joint Surg Am. 1997;79:334-41.

26. Larsen LB, Madsen JE, Høiness PR, Øvre S. Should insertion of intramedullary nails for tibial fractures be with or without reaming? A prospective, randomized study with 3.8 years' follow-up. J Orthop Trauma. 2004;18:144-9.

27. Cox DR. A remark on multiple comparison methods. Technometrics. 1965;7:223-4.

28. Cook RJ, Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. J R Stat Soc Ser A Stat Soc. 1996;159:93-110.