2008-1-29

The Orthopaedic Trauma Literature: An Evaluation of Statistically Significant Findings in Orthopaedic Trauma Randomized Trials

Sung, Jinsil

BioMed Central


http://hdl.handle.net/2144/2657

Boston University
The orthopaedic trauma literature: an evaluation of statistically significant findings in orthopaedic trauma randomized trials

Jinsil Sung¹, Judith Siegel¹, Paul Tornetta¹ and Mohit Bhandari*²

Address: ¹Department of Orthopaedic Surgery, Boston University, 818 Harrison Ave, Dawling 2 North, Boston Massachusetts, 02118-2393, USA and ²Department of Surgery, McMaster University, 293 Wellington Street N, Suite 110, Hamilton, Ontario, L8L 8E7, Canada

Email: Jinsil Sung - jinsil.sung@bmc.org; Judith Siegel - yoda@hotmail.com; Paul Tornetta - ptornetta@pol.net; Mohit Bhandari* - bhandam@mcmaster.ca

* Corresponding author

Abstract

Background: Evidence-based medicine posits that health care research is founded upon clinically important differences in patient centered outcomes. Statistically significant differences between two treatments may not necessarily reflect a clinically important difference. We aimed to quantify the sample sizes and magnitude of treatment effects in a review of orthopaedic randomized trials with statistically significant findings.

Methods: We conducted a comprehensive search (PubMed, Cochrane) for all randomized controlled trials between 1/1/95 to 12/31/04. Eligible studies include those that focused upon orthopaedic trauma. Baseline characteristics and treatment effects were abstracted by two reviewers. Briefly, for continuous outcome measures (ie functional scores), we calculated effect sizes (mean difference/standard deviation). Dichotomous variables (ie infection, nonunion) were summarized as absolute risk differences and relative risk reductions (RRR). Effect sizes >0.80 and RRRs>50% were defined as large effects.

Using regression analysis we examined the association between the total number of outcome events and treatment effect (dichotomous outcomes).

Results: Our search yielded 433 randomized controlled trials (RCTs), of which 76 RCTs with statistically significant findings on 184 outcomes (122 continuous/62 dichotomous outcomes) met study eligibility criteria. The mean effect size across studies with continuous outcome variables was 1.7 (95% confidence interval: 1.43–1.97). For dichotomous outcomes, the mean risk difference was 30% (95% confidence interval: 24%–36%) and the mean relative risk reduction was 61% (95% confidence interval: 55%–66%; range: 0%–97%). Fewer numbers of total outcome events in studies was strongly correlated with increasing magnitude of the treatment effect (Pearson’s R = -0.70, p < 0.01). When adjusted for sample size, the number of outcome events revealed an independent association with the size of the treatment effect (Odds ratio = 50, 95% confidence interval: 3.0–1000, p = 0.006).

Conclusion: Our review suggests that statistically significant results in orthopaedic trials have the following implications- 1) On average large risk reductions are reported 2) Large treatment effects (>50% relative risk reduction) are correlated with few number of total outcome events. Readers should interpret the results of such small trials with these issues in mind.

Published: 29 January 2008
Received: 6 September 2007
Accepted: 29 January 2008


This article is available from: http://www.biomedcentral.com/1471-2474/9/14

© 2008 Sung et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Background
Evidence-based medicine posits that health care research is founded upon clinically important differences in patient centered outcomes. Randomized trials continue to represent the reference standard for the comparison of surgical interventions [1-4]. Although fundamentally the most important for guiding clinical practice, few randomized trials are conducted in orthopaedic surgery. Current estimates suggest that less than 5% of the orthopaedic literature represent randomized trials [5-7]. Nevertheless, the impact of randomized trials, especially those with statistically significant findings, is large [8].

Statistically significant differences between two treatments may not necessarily reflect a clinically important difference. Although it is well known that orthopaedic studies with small sample sizes risk underpowered false negative conclusions (Beta-errors) [6,9,10], statistically significant findings in small trials can occur at the consequence of very large differences between treatments (treatment effect). It is not uncommon for randomized trials to report relative risk reductions larger than 50% when comparing one treatment with another [11-13].

Devereaux and colleagues caution the interpretation of small trials in cardiology [14]. For example, the peri-operative beta-blocker evidence suggests large treatment effects (i.e., relative risk reductions >75%) but these results are inconsistent with beneficial cardiovascular therapies established in trials with 10,000s of patients, which generally demonstrate moderate relative risk reductions in the order of 15 to 35% [14-16].

Our study had 2 objectives: 1) To determine the magnitude of treatment effects in a sample of orthopaedic randomized trials with statistically significant results and 2) to examine the association between the number of outcome events (a measure of study sample size) and the size of the treatment effect. We conducted a systematic review to identify randomized trials in orthopaedic trauma with the following hypotheses: 1) statistically significant studies would not always report large treatment effects and 2) studies with smaller sample sizes (and few outcome events) would be more likely to report larger treatment effects than those with larger sample sizes.

Methods
Eligibility Criteria
We included studies which met the following eligibility criteria: 1) published studies, 2) described as randomized trial, 3) involve the care of adult patients with fractures, either operative or conservative, 4) published in English and 5) contain sufficient outcomes information to calculate treatment effects for both dichotomous and continuous outcome measures. Our decision to focus upon trauma randomized trials was based upon two factors: 1) allowing comparison with previous studies evaluating this population of trials, and 2) practicality of limiting the number of trials to a sufficiently manageable number to optimize the efficiency of study completion and research resource utilization within our Departments.

Study Identification
We conducted a comprehensive search (PubMed, Cochrane database) for all randomized controlled trials between January 1, 1995 and December 31, 2004. We used the search terms "randomized controlled trial" and "fracture" and "surgery" with limits (adults 19+ years). The eligibility criteria were applied to potentially eligible study titles by two independent reviewers (JS, MB). One of the two reviewers was trained in health research methodology, while the other was an orthopaedic resident. Abstracts for those eligible study titles were retrieved by one of us. Following a second application of eligibility criteria to abstracts by independent reviewers, complete citations for those potentially eligible studies were retrieved. The methods section of each retrieved citation was reviewed by two of us to ensure all inclusion criteria were met. In addition to Medline searches, two of us performed a search of the NIH PubMed computerized database and one of us conducted a Cochrane Database search. For both searches, we used "fractures" and "randomized trials" as keywords.

Additional strategies to identify relevant citations included: 1) hand searches of the table of contents over the past 5 years of the Journal of Orthopaedic Trauma, Journal of Trauma, Clinical Orthopaedics and Related Research and Acta Orthopaedica Scandinavica, 2) review of the reference lists of eligible (included) studies to identify other potentially eligible studies, and 3) content experts' (traumatologist) review of the list of eligible studies to identify any missing studies.

Characteristics of Eligible Studies
Two reviewers independently abstracted general characteristics of each eligible study. These included, first author (surgeon/non-surgeon), geographic location, category of intervention, body region of focus (upper extremity, lower extremity, spine), number of participating centres, and funding (yes/no).

 Determination of Treatment Effects and Outcome Events
For dichotomous outcome measures (ie re-operation, infection), we calculated relative risks as percent (%) re-operation intervention group divided by percent (%) re-operation in comparison group. For ease of interpretation we converted relative risks to relative risk reductions [(1- Relative Risk) x 100]. We also calculated absolute risk differences. A relative risk reduction of 50% was interpreted
as Experiment treatment reduced the risk of an adverse outcome event by 50% compared to a control (comparison treatment).

For continuous outcome measures (i.e., functional scores) we calculated an effect size as described by Cohen [17]. 'Effect Size' is simply a way of quantifying the effectiveness of a particular intervention, relative to some comparison. It is easy to calculate, readily understood and can be applied to any measured outcome in surgical trials. It is the standardized mean difference between the two groups. We used the following formula for its calculation:

\[
\text{Effect Size} = \frac{\text{Mean of experimental group} - \text{Mean of Comparison group}}{\text{Pooled Standard Deviation}}
\]

For example, an effect size of 0.8 means that the score of the average person in the experimental group exceeds the scores of 79% of the comparison group. Cohen describes an effect size of 0.2 as 'small' and gives to illustrate it the example that the difference between the heights of 15 year old and 16 year old girls in the US corresponds to an effect of this size. An effect size of 0.5 is described as 'medium' and is 'large enough to be visible to the naked eye'. A 0.5 effect size corresponds to the difference between the heights of 14 year old and 18 year old girls. Cohen describes an effect size of 0.8 as 'grossly perceptible and therefore large' and equates it to the difference between the heights of 13 year old and 18 year old girls [17]. For each study, we documented outcomes measured deemed primary by the study authors. In cases in which primary outcomes were not specified by authors, two of us (trained orthopaedic surgeons) made judgments about the key outcomes based upon the study interventions.

Assessing Reviewer Agreement
Agreement in the application of study eligibility criteria, identification of study outcomes and study results (positive or negative) was quantified with the Kappa statistic with quadratic weighting. The kappa statistic, a measure of agreement among two or more individuals beyond chance, provided a measure of agreement among reviewers for titles, abstracts and methods sections of potentially relevant. In the context of inter-observer agreement studies, Fleiss and Donner provide persuasive arguments favoring the use of this statistic over other measures of agreement that have been proposed [18,19].

Data Analysis
We presented descriptive statistics about continuous variables with means, standard deviations and dichotomous variables as proportions. We calculated relative risks and 95% confidence intervals to describe treatment effects and compared relative risks in studies with few and many events with tests for interaction. Logistic regression provided methods for estimating the extent of association between the total number of events (i.e., endpoints driving termination) in the trial and the calculated treatment effect. We categorized the number of outcome events as follows: 1) 0–25 events, 2) 26–50 events, 3) 51–75 events, and 4) 76–100 events. We also categorized studies as having a relative risk reduction as follows: 1) <50% and 2) >50%. We expressed any associations using odds ratios (ORs) and associated 95% confidence intervals (CIs). Analyses were performed using SPSS version 13.0. We conducted a correlation analysis of the number of outcome events against the treatment effect with a Pearson’s R. We considered P < 0.05 as the level of statistical significance for all comparisons. All tests were two-tailed.

Results

Literature Search
We identified 433 potentially relevant study titles from the MEDLINE database search. (Figure 1). Application of the study eligibility criteria eliminated 171 titles, leaving 262 for further consideration (Table 1). Following complete review of 262 study abstracts, 94 were excluded, leaving 168 papers. In total, 168 studies appeared potentially eligible from study title and abstract alone and full manuscripts were retrieved for a detailed review. Agreement in the application of eligibility criteria to study titles and abstracts was substantial (Kappa = 0.80, 95% confidence interval: 0.74–0.86). Application of the eligibility criteria to the complete manuscripts of 168 trials eliminated 92 studies (Table 1). Thus, 76 papers that met all a priori eligibility criteria were included in the analyses (Appendix).

Study Characteristics
The 76 eligible trials were published across 9 different journals (Table 2). The majority of studies were published in Journal of Orthopaedic Trauma (21%), JBJS-British (18%), and JBJS-American (13%). Most of the studies were conducted in Europe (58%) followed by North America (32%) (Table 3). Seventy-three (96%) of the first authors were surgeons and the majority of studies were single center initiatives (83%). Funding for the conduct of the trial was reported in 17 studies (22%). A total of 9757 patients were randomized in the 76 trials. Moreover, study sample sizes ranged from 10 to 424 patients (mean = 77 patients, st.dev = 69).

Treatment Effects
We identified 62 dichotomous outcomes across our sample of 76 studies (Table 4). The mean sample size in studies that reported dichotomous outcomes was 81.2 ± 61.4 patients (median 55 patients) with a minimum 14 patients to a maximum 250 patients. For dichotomous outcomes, the mean risk difference was 30% (95% confidence interval: 24%–36%; range 2%–138%). The treatment effects (relative risks) averaged 0.39 (95%
confidence interval: 0.33–0.45; range: 0.03–1.0). This translated to a mean observed relative risk reduction of 61% (95% confidence interval: 55%–66%; range: 0%–97%). Thus, the average study reported large reductions (>50% relative risk reduction) in the risk of an adverse outcome event versus a comparative treatment; however, almost 1 in 2 study outcomes (47%) had RRRs<50%, and over 1 in 5 (23%) had RRRs<20%.

We identified 122 continuous outcomes across our sample of 76 studies (Table 4). The mean sample size in studies reporting continuous outcomes was 70 ± 69 patients (median = 50 patients) with a minimum 10 patients and a maximum 424 patients. The absolute mean difference between treatments was 81.2 (95% confidence interval: 49.2–113.2). The effect size averaged 1.7 (95% confidence interval: 1.43–1.97). Eighty four outcomes reported (69%) large effect sizes (ES>0.80).

**Association between Treatment Effect and the Number of Outcome Events**

The mean number of outcome events (Treatment group + Comparison Group) across studies with dichotomous outcomes was 24 ± 21 (median = 17). The total number of outcome events ranged from 1 event to 90 events. Fewer numbers of total outcome events in studies was strongly correlated with increasing magnitude of the treatment effect (Pearson’s R = -0.70, p < 0.01) (Figure 2). The relative risk reduction decreased as the number of outcome events increased from 0–25 events to 75–100 events (Rel. Risk Red = 73% vs 16%, respectively, P < 0.01) (Table 5) (Figure 3). In the logistic regression analysis, fewer than 50 outcome events was significantly associated with a relative risk reduction greater than 50% (Odds ratio = 21, 95% confidence interval: 2.1–200, p = 0.0–09). When adjusted for sample size, the number of outcome events continued to show independent association with the size of the treatment effect (Odds ratio = 50, 95% confidence interval: 3.0–1000, p = 0.006). The number of outcome events explained 32% of the variance in the regression model. In the 6 studies with greater than 50 total outcome events, 5 studies had modest to small relative risk reductions range (3%–33%).

**Discussion**

Our review of trials with statistically significant findings in orthopaedic trauma suggests the following: 1) trials have
small sample sizes 2) average treatment effects are large and 3) large treatment effects (>50% relative risk reduction) correlated with fewer number of total outcome events. Readers should interpret the results of such small studies with few events with caution.

**Strengths and Limitations**

Our study is strengthened by a comprehensive literature review including hand searches of major orthopaedic journal, careful study methodology and duplicate data abstraction. Our search strategy, although comprehensive may have missed studies related to fracture care due to errors in study indexing in Medline or those articles not indexed in PubMed. Our decision to identify only trials relating to orthopaedic trauma for pragmatic reasons limits our generalizability beyond this subspecialty. However, it remains plausible that the associations we observed are consistent throughout the orthopaedic literature in both English and Non-English trials. The lack of reporting of sufficient data to calculate treatment effect size for those studies with continuous outcome variables was another limitation. Fifty-two studies were excluded for this reason. We realize that excluding so many studies is a limitation. However, review of these studies suggests that they were similar in sample size, geographic location, funding status and number of centers to our included studies.

**Relevant Literature Review**

Whereas statistical significance means the likelihood that the difference found between groups could have occurred by chance alone, effect-sizes provide an estimate of the size of the treatment effect. Effect sizes are important because they facilitate the comparison of treatment effects across different studies. In most clinical trials, a result is statistically significant if the sample size of the study is large. For example, a large trial (N = 7601 patients) comparing the use effects of the angiotensin receptor blocker candesartan on cardiovascular outcomes reported a significant reduction in the development of atrial fibrillation with candesartan versus placebo (p = 0.048); however, the actual difference was 1.19% (6.74% vs 5.55%, respectively) [20].

The findings reported in some biomechanical studies should also be interpreted cautiously. For example, a biomechanical study that compared the compression effect of the 7.0-AO screw and the 6.5 mm Ideal Compression Screw (I.C.O.S.) screw in an in vitro subtalar arthodesis model. The authors reported that the AO screw led to a significantly increased mean contact force (p < 0.05); however, this increase was 7.6 N – the clinical significance of which is completely uncertain [21].

Clinical significance is a matter of judgment. However, clinically significant findings imply that the difference between treatment groups are large enough to be important to patients. We can argue that an absolute difference of 1.19% is not a clinically important difference. Alternatively, a study of 50 patients reporting a 20% (10/25, 40% vs 5/25, 20%, p = 0.12) absolute difference in atrial fibrillation rates between candesartan and placebo groups may be more compelling for clinical practice. However, the difference may not reach statistical significance.

There have been no studies in the surgical literature evaluating the association of treatment effect magnitudes and number of outcome events. However, investigators have examined inflated treatment effects in large medical trials stopped early for benefit at an early interim analysis [16]. For example, a trial that aimed to recruit 1000 patients may be terminated early if an interim analysis of 100 patients shows a very large benefit of the treatment over a comparison (ie relative risk reduction >50%). The decision to terminate a trial before it reaches its preplanned
Table 3: Characteristics of the Eligible Trials (N = 76 Trials)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. First Author:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surgeon:</td>
<td>73/76 (96.0%)</td>
<td>15/16</td>
<td>10/10</td>
<td>13/14</td>
<td>8/9</td>
<td>8/8</td>
<td>8/8</td>
<td>5/5</td>
<td>2/2</td>
<td>4/4</td>
</tr>
<tr>
<td>Non-Surgeon:</td>
<td>3/76 (4%)</td>
<td>1/16</td>
<td>0/10</td>
<td>1/14</td>
<td>1/9</td>
<td>0/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Epidemiologist:</td>
<td>0 (0%)</td>
<td>0/16</td>
<td>0/10</td>
<td>0/14</td>
<td>0/9</td>
<td>0/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>2. Epidemiology Affiliation:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes:</td>
<td>16/76 (21.0%)</td>
<td>0/16</td>
<td>2/10</td>
<td>10/14</td>
<td>0/9</td>
<td>1/8</td>
<td>3/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>No:</td>
<td>60/76 (78.9%)</td>
<td>16/16</td>
<td>8/10</td>
<td>4/14</td>
<td>9/9</td>
<td>7/8</td>
<td>5/8</td>
<td>5/5</td>
<td>2/2</td>
<td>4/4</td>
</tr>
<tr>
<td>3. Geographical Location:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North America:</td>
<td>24/76 (32%)</td>
<td>9/16</td>
<td>6/10</td>
<td>1/14</td>
<td>5/9</td>
<td>0/8</td>
<td>2/8</td>
<td>0/5</td>
<td>1/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Europe (includes England):</td>
<td>44/76 (58%)</td>
<td>7/16</td>
<td>4/10</td>
<td>10/14</td>
<td>4/9</td>
<td>8/8</td>
<td>2/8</td>
<td>4/5</td>
<td>1/2</td>
<td>4/4</td>
</tr>
<tr>
<td>Australia or New Zealand:</td>
<td>1/76 (1.3%)</td>
<td>0/16</td>
<td>0/10</td>
<td>1/14</td>
<td>0/9</td>
<td>0/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Asia:</td>
<td>6/76 (8%)</td>
<td>0/16</td>
<td>0/10</td>
<td>2/14</td>
<td>0/9</td>
<td>0/8</td>
<td>3/8</td>
<td>1/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Israel:</td>
<td>1/76 (1%)</td>
<td>0/16</td>
<td>0/10</td>
<td>0/14</td>
<td>0/9</td>
<td>0/8</td>
<td>1/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>4. Category of Intervention:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fracture Treatment:</td>
<td>70/76 (92.1%)</td>
<td>16/16</td>
<td>9/10</td>
<td>12/14</td>
<td>8/9</td>
<td>7/8</td>
<td>7/8</td>
<td>5/5</td>
<td>2/2</td>
<td>4/4</td>
</tr>
<tr>
<td>Other Treatment:</td>
<td>5/76 (6.6%)</td>
<td>0/16</td>
<td>1/10</td>
<td>1/14</td>
<td>1/9</td>
<td>1/8</td>
<td>1/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Vascular Treatment:</td>
<td>1/76 (1.3%)</td>
<td>0/16</td>
<td>0/10</td>
<td>1/14</td>
<td>0/9</td>
<td>0/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>5. Region of Body:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper Extremities:</td>
<td>16/76 (21.0%)</td>
<td>2/16</td>
<td>5/10</td>
<td>2/14</td>
<td>0/9</td>
<td>0/8</td>
<td>1/8</td>
<td>2/5</td>
<td>0/2</td>
<td>4/4</td>
</tr>
<tr>
<td>Lower Extremity</td>
<td>23/76 (30.3%)</td>
<td>9/16</td>
<td>1/10</td>
<td>4/14</td>
<td>2/9</td>
<td>1/8</td>
<td>5/8</td>
<td>1/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Long Bones:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spine:</td>
<td>2/76 (2.6%)</td>
<td>0/16</td>
<td>1/10</td>
<td>0/14</td>
<td>1/9</td>
<td>0/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Hip:</td>
<td>28/76 (36.8%)</td>
<td>4/16</td>
<td>2/10</td>
<td>6/14</td>
<td>6/9</td>
<td>6/8</td>
<td>2/8</td>
<td>2/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Foot and Ankle</td>
<td>4/76 (5.3%)</td>
<td>0/16</td>
<td>1/10</td>
<td>0/14</td>
<td>0/9</td>
<td>1/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Knee:</td>
<td>2/76 (2.6%)</td>
<td>1/16</td>
<td>0/10</td>
<td>1/14</td>
<td>0/9</td>
<td>0/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>Upper and Lower Extremity:</td>
<td>1/76 (1.3%)</td>
<td>0/16</td>
<td>0/10</td>
<td>1/14</td>
<td>0/9</td>
<td>0/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>0/4</td>
</tr>
<tr>
<td>6. Single or Multicentred:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single:</td>
<td>63/76 (83%)</td>
<td>15/16</td>
<td>6/10</td>
<td>13/14</td>
<td>6/9</td>
<td>5/8</td>
<td>8/8</td>
<td>5/5</td>
<td>2/2</td>
<td>3/4</td>
</tr>
<tr>
<td>Multicentred:</td>
<td>13/76 (17%)</td>
<td>1/16</td>
<td>4/10</td>
<td>1/14</td>
<td>3/9</td>
<td>3/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>1/4</td>
</tr>
<tr>
<td>7. Funding:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes:</td>
<td>17/76 (22%)</td>
<td>2/16</td>
<td>4/10</td>
<td>4/14</td>
<td>2/9</td>
<td>4/8</td>
<td>0/8</td>
<td>0/5</td>
<td>0/2</td>
<td>1/4</td>
</tr>
<tr>
<td>No:</td>
<td>27/76 (36%)</td>
<td>10/16</td>
<td>4/10</td>
<td>10/14</td>
<td>1/9</td>
<td>0/8</td>
<td>0/8</td>
<td>1/5</td>
<td>0/2</td>
<td>1/4</td>
</tr>
<tr>
<td>Unknown:</td>
<td>32/76 (42%)</td>
<td>4/16</td>
<td>2/10</td>
<td>0/14</td>
<td>6/9</td>
<td>4/8</td>
<td>8/8</td>
<td>4/5</td>
<td>2/2</td>
<td>2/4</td>
</tr>
</tbody>
</table>

* JBJS = Journal of Bone and Joint Surgery, JOT = Journal of Orthopaedic Trauma; CORR = Clinical Orthopaedics and Related Research

Sample size is based upon apriori statistical stopping rules (threshold of a p value is reached). If we extrapolate this to orthopaedic surgical trials, we can conceptualize these small trials (mean = 80 patients) with large treatment effects (relative risk reductions of >66%) as trials that were essentially “stopped early”. In doing so, the same play of chance positive findings can be extrapolated to this literature. Moreover, implications of stopped early trials can be explored from previous reports in the medical literature.
Several important lessons about early stopping of trials with large treatment effects exist in the literature [15,16]. For example, the preliminary results of the twelfth Medical Research Council acute myeloid leukemia trial ultimately revealed no evidence of a survival advantage for five courses of therapy compared to four courses in a randomized comparison involving 1078 patients (hazard ratio 1.09, 95% confidence interval [CI] 0.87–1.37, p = 0.4) [15]. However, large benefits of the 5 course therapy (53% and 45% reductions in the odds of death) in early interim analyses of fewer patients recruited were fortunately dismissed as “too good to be true” and implausible. The ultimate large trial prevented the wide adoption of a non-beneficial therapy with potential harms of more chemotherapy.

Another poignant example of the pitfalls of small trials and large effects is derived from Cardiology – the case of magnesium in acute myocardial infarction [22,23]. When initial small trials have claimed remarkably large benefits, subsequent trials typically demonstrated much more modest. For example, a meta-analysis of RCTs of magnesium in acute myocardial infarction demonstrated a statistically significant (P < 0.001) > 50% reduction in death with approximately 1,500 patients randomized [20]. However, the subsequent RCT of approximately 60,000 patients showed no benefit; in fact there was a trend toward excess mortality with magnesium (P = 0.07) [23]. Devereaux and colleagues argue that cardiovascular therapies rarely demonstrate relative risk reductions greater than 35%, because cardiovascular therapies usually only target a limited number of the multitude of pathogenic mechanisms underlying cardiovascular diseases, as is the case with peri-operative beta-blockers [24].

Take, for example, the RCT evaluating the efficacy of bisoprolol (Beta-blocker) in patients with a positive dobutamine echocardiography result and undergoing elective vascular surgery [25]. The trial was stopped when investigators had enrolled 112 patients of the pre-planned 266 patients. The relative risk reduction for the primary outcome (cardiac death or nonfatal myocardial infarction) was 91% (95% CI, 63%–98%). This very large treatment effect is implausible given that the authors anticipated a less beneficial effect of this drug. In fact, a subsequent trial of larger sample size and greater outcome events (496 patients) undergoing vascular surgery that reported no significant effect of Beta-blockers on cardiac death or non-fatal myocardial infarction [26].

Montori and colleagues recently conducted a systematic review to identify randomized trials stopped early for benefit [16]. Of 143 trials stopped early for benefit, the majority (92) were published in high-impact medical journals (New England Medical Journal, Journal of the American Medical Association, Annals of Internal Medicine). On average, trials recruited 63% of the planned sample and stopped after when a median of 66 total outcome events (experimental + control). The median relative risk reduction among truncated trials was 47%. Trials with fewer events yielded greater treatment effects (odds ratio, 28; 95% confidence interval, 11–73).

### Importance of Our Findings

If one accepts that our sample of orthopaedic randomized trials represents small trials with implausibly large benefits, then our findings are interesting and the implications of our study highly relevant and important. The average study in our review had a sample size of 81 patients but reported a large beneficial treatment effect (61% relative risk reduction). Surgeons should consider the plausibility of the magnitude of the treatment effect because chance effects do occur and happen more frequently than many of us realize [15]. Statistical simulation studies have shown that RCTs can overestimate the magnitude of the treatment effect depending on the timing (ie, the fraction of the total planned sample size or expected number of events) of the decision to stop [27].

Indirect evidence for the implausibility of treatment effects reported in small orthopaedic trauma trials with large reported treatment benefits is available. For example, early small randomized trials (sample sizes 48–141) comparing reamed versus non-reamed intramedullary
nailing identified relative risk reductions in nonunion with reamed nailing of 54% (but as high as 79%) [27]. If surgeons truly believed that reamed nails could reduce the risk of an important adverse event by over 50%, surely every surgeon would have adopted this relatively simple strategy [28]. However, surveys of surgeons conducted years after the conduct of these trials found that 42% of surgeons continued to use the non-reamed nail [29]. It remains plausible, then, that the surgical community believed that 54% risk reductions were "implausibly" high.

In another example, pooled analyses small randomized trials have reported large reductions in the risk of reoperation (74%) with plates compared to intramedullary nails in the treatment of humeral shaft fractures [30-32]. Again, if practicing surgeons believed these trials, they would abandon the use of nails. This is not so. The continued use of intramedullary nails for humeral shaft fractures suggests, at least in part, possible skepticism with small trials with large treatment effects.

**Recommendations**

Authors should cautiously interpret the positive findings of studies when sample sizes of the study are small and the total number of outcome events are few. As the number of outcome events increases, surgeons should have increasingly greater confidence in the reported magnitude of the treatment effect. For example, a trial claiming that reamed intramedullary tibial nails reduce the risk of revision surgery by 50% in 1000 patients with 200 outcome events is less likely influenced by the random chance than a similar study of 100 patients with 20 outcome events.

Wright and Gebhardt call for specialty orthopaedic societies to take action towards the conduct of multicenter ran-
domized trials [33]. Single center initiatives will rarely be sufficient to enroll sufficiently large numbers of patients efficiently.

Conclusion

Our review suggests that statistically significant results in orthopaedic trials have the following implications: 1) On average large risk reductions are reported 2) Large treatment effects (>50% relative risk reduction) are correlated with few number of total outcome events. Readers should interpret the results of such small trials with these issues in mind.

Appendix: list of 76 included RCTs


Acknowledgements
No funding was received in preparation of this manuscript. Dr. Bhandari is funded, in part, by a Canada Research Chair in Musculoskeletal Trauma, McMaster University.

References
Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2474/9/14/prepub