Statistical Significance in Trauma Research: Too Unstable to Trust?

Paul Tornetta III, MD[1]; Mohit Bhandari, MD, MSc, PhD[2]; ***Robert L. Parisien, MD[1]****;*
Jesse Dashe, MD[1]; Patrick Cronin[1];
[1]Boston University Medical Center, Boston, Massachusetts, USA;
[2]McMaster University, Hamilton, Ontario, Canada

Background/Purpose: Comparison trials are the most compelling evidence available for surgeons to make decisions. The outcomes of trials are based on hypothesis testing with an a priori statistical cutoff, which is generally accepted to be $P < 0.05$. This is to say that with 95% certainty one treatment is better than another and should therefore influence decision-making. However, when categorical outcomes are considered (such as nonunion, infection, etc), the statistical outcome of trials is dependent on the number of "outcome events", which are often a small percentage of the overall study population. We sought to examine how easily the statistical significance of comparison trials in fracture care would change if the number of events in one group were incrementally changed. By example, in one study, if 15 infections had occurred in one arm instead of 12, the *P* value would change from $P = 0.02$ to $P = 0.08$, changing its statistical significance (from <0.05 to >0.05) and likely how it would influence practice.

Methods: We screened all fracture care studies in the *Journal of Bone and Joint Surgery* and the *Journal of Orthopaedic Trauma* over a 20-year period. Inclusion criteria were comparison trials whose outcomes were categorical and had data included to be evaluated. Data on the number of patients in each arm of the trial, the number of events in each arm, and the number lost to follow-up were tabulated. Reported outcomes were considered "significant" if the *P* value was <0.05. For each study outcome we confirmed the *P* value that was reported and then we changed the number of events in one arm enough to "flip" the significance of the study. If the outcome was significant, then the required number of event changes to raise *P* to above 0.05 was determined, and if the outcome was not significant, the number of event changes that would drop *P* to <0.05 was determined. Analyses were performed using Fisher's exact test. The number of events as a percentage of the arm and the study population was calculated.

Results: Of 4040 studies, 198 met inclusion criteria and had 253 primary and 516 secondary outcomes. There were 118 randomized controlled trials (RCTs) and 80 retrospective studies. 230 outcomes were significant as reported and 539 were not. The median *P* value for significant studies was 0.003 (1.3E-09–0.049) and for nonsignificant studies was 0.6 (0.51-1). There were no differences in the findings for randomized versus nonrandomized trials so the data are presented together. The median number of patients in the studies was 95 (12-6000). The number of event changes in one arm for each outcome that would flip the significance is seen in Table 1 separated by the initial reporting of significant and nonsignificant results. The median number of events that were needed to flip the significance of the trials was only 5, which was on average 8.9% of one arm and 3.8% of the total study population. By comparison, the average lost to follow-up for the studies was 3%. Initially significant and nonsignificant studies were affected equally by event changes.

See pages 99 - 147 for financial disclosure information.

Table 1. Events Needed to Flip the Significance of RCTs

	Median No. of Events	Range	% of One Arm of Study	% of Study Population
Studies $P < 0.05$	4	1-340	7.8%	3.4%
Studies $P \geq 0.05$	5	1-40	9.1%	3.8%
All studies	5	1-340	8.9%	3.8%

Conclusion: The statistical outcomes of comparison trials that rely on noncontinuous variables such as infection, nonunion, secondary procedures, etc may not be as stable as previously thought. When evaluating trials that rely on events, small numbers of events may change the statistical significance of the trial. In evaluating 769 outcomes of 198 trials, we found that a median of only 4 events would flip studies with reported P values of <0.05 to over 0.05 and 5 events would make significant trials initially reporting $P \geq 0.05$. The overall percentage of the study population that would change significance was only 3.8% for all studies. Importantly, randomized trials faired no better than nonrandomized trials in this analysis. This highlights the need for readers to understand how P values relate to study findings and that using a discreet cutoff for P value in determining importance is likely not appropriate.

- The FDA has not cleared this drug and/or medical device for the use described in this presentation (i.e., the drug or medical device is being discussed for an "off label" use). For full information, refer to page 600.