Research Article

Statistical Fragility and the Role of *P* Values in the Sports Medicine Literature

Robert L. Parisien, MD
David P. Trofa, MD
Jesse Dashe, MD
Patrick K. Cronin, MD
Emily J. Curry, BA
Freddie H. Fu, MD
Xinning Li, MD

From the Department of Orthopaedic Surgery, Boston Medical Center, Boston, MA (Dr. Parisien, Dr. Dashe, Dr. Curry, and Dr. Li), the Department of Orthopaedic Surgery, Columbia University Medical Center, New York, NY (Dr. Trofa), the Department of Orthopaedic Surgery, Massachusetts General Hospital & Harvard Medical School, Boston, MA (Dr. Cronin), and the Department of Orthopaedic Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA (Dr. Fu).

Correspondence to Dr. Li: xinning.li@gmail.com

Dr. Fu or an immediate family member is an employee of Medicrea and serves as a board member, owner, officer, or committee member of the American Academy of Orthopaedic Surgeons, the International Society of Arthroscopy, Knee Surgery, Orthopaedic Sports Medicine, and the World Endoscopy Doctors Association. None of the following authors or any immediate family member has received anything of value from or has stock or stock options held in a commercial company or institution related directly or indirectly to the subject of this article: Dr. Parisien, Dr. Trofa, Dr. Dashe, Dr. Cronin, Dr. Curry, and Dr. Li.

J Am Acad Orthop Surg 2018;00:1-6

DOI: 10.5435/JAAOS-D-17-00636

Abstract

Introduction: Comparative trials evaluating categorical outcomes have important implications on surgical decision making. The purpose of this study was to examine the statistical stability of sports medicine research.

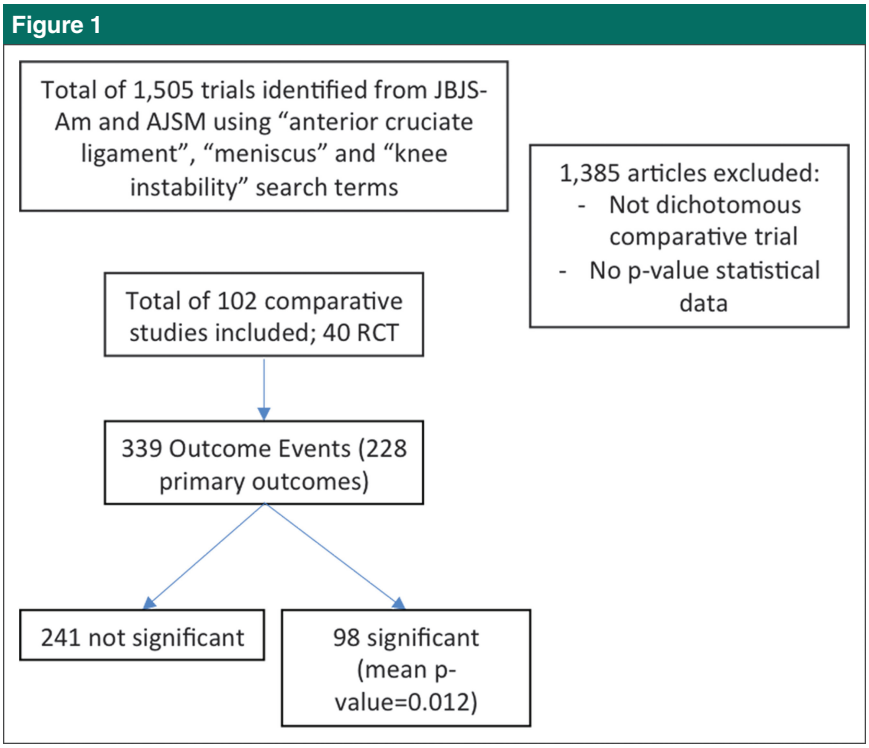
Methods: Comparative clinical sports medicine research studies involving anterior cruciate ligament, meniscus, and knee instability were reviewed in two journals between 2006 and 2016. The statistical stability for each study outcome was determined by the number of event reversals required to change the *P* value to either greater or less than 0.05. The number of patients lost to follow-up was also determined.

Results: Of the 1,505 studies screened, 102 studies were included for analysis, 40 of which were randomized controlled trials. There were 339 total outcome events, with 98 significant and 241 not significant. The Fragility Index, or the median number of events required to change the statistical significance of the overall study, was five (interquartile range, 3 to 8) or 5.4% of the total study population. In addition, the average number of patients lost to follow-up was 7.9, which is greater than the number needed to change the significance of each study arm and the entire study population.

Conclusion: Results in the comparative sports medicine literature may not be as stable as previously thought, with only a small percentage of outcome events needed to change study significance. Outcomes research based on a single discreet *P* value cutoff may be misleading.

The modern practice of evidence-based medicine (EBM) began with Archie Cochrane's[1] 1971 publication Effectiveness and Efficiency, highlighting the dearth of evidence behind many common medical practices. Cochrane's idea was subsequently expounded on by David M. Eddy,[2] MD, PhD, in his 1990 *JAMA* article, which evaluated medical practice standards. Numerous articles have since emphasized the principles of EBM leading to a drastic change in healthcare delivery. The practical application of EBM presents a substantial challenge with a new and particular skill set required by the practicing clinician to critically evaluate the rapidly growing body of literature. However, this system breaks down in the absence of high-quality reliable research.[3] This is of particular concern in the field of orthopaedics, in which high-quality research is substantially lacking. Among the top 100 most frequently cited studies in the orthopaedic literature, more than half are of level IV

Figure 1



Flowchart showing the total number of outcome events that comprised the study findings. AJSM = American Journal of Sports Medicine, JBJS-Am = Journal of Bone and Joint Surgery, RCT = randomized controlled trial

evidence (55.2%).[4] However, the quality of orthopaedic research is gradually improving, with the percentage of level I studies presented at the AAOS Annual Meeting increasing from 2% in 2001 to 7% in 2010.[5]

In the orthopaedic sports medicine literature, the best available evidence for clinical decision making is gathered from the critical evaluation of dichotomous comparative trials. R.A. Fisher statistical method is commonly used to assess the statistical significance between the two groups and evaluates the probability that the null hypothesis (H_o), a statement of equality between the two data sets, is accepted or rejected. The null hypothesis is rejected when the *P* value, the ratio of the observed difference between the two groups over the standard error of the difference, is below a set criterion. By convention, the a priori statistical cutoff is accepted as $P < 0.05$. Under these circumstances, it can be said that the observed difference has less than a 5% likelihood of being due to random chance. Using these principles, clinical decisions in orthopaedic sports medicine are routinely made based on this statistical *P* value cutoff. However, it has been noted across multiple disciplines that statistically significant findings may be unstable, hinging on relatively few events.[6-8] The purpose of this study was to conduct the first comprehensive examination of the sports medicine literature to determine statistical stability.

Methods

Study Identification

Comparative clinical sports medicine research studies published in the *Journal of Bone and Joint Surgery* (JBJS-Am) and the *American Journal of Sports Medicine* (AJSM) from 2006 to 2016 were retrieved using the following search terms, "anterior cruciate ligament," "meniscus," and "knee instability." Evaluation was focused specifically on these two journals based on their impact factors and relative prominence with regard to the published sports literature. According to the 2015 Science Citation Index, JBJS-Am is recognized as the top journal, with AJSM listed as the number 2 journal with regard to impact factor, 5.280 and 4.362, respectively. In addition, the SCImago Journal & Country Rank, a measure of scientific influence of scholarly journals that accounts for both the number of citations received by a journal and the importance or prestige of the journals, reports AJSM and JBJS as number 1 and number 5, respectively. Thus, JBJS-Am and AJSM represent the highest quality of published sports literature, and exhaustive analysis of 10 years of data within these two prominent journals likely demonstrates an accurate representative sample of the sports literature. Inclusion criteria were dichotomous comparative trials reporting categorical and *P* value statistical data. Overall, 1,505 trials were screened with the inclusion of 102 comparative studies consisting of 339 total outcome events (Figure 1).

Fragility Index

The statistical stability for each study outcome was determined by the number of events required to change the *P* value to either greater or less than 0.05, thus changing the study conclusions. This was performed by manipulating the reported outcome events, one event at a time, until a reversal of significance was appreciated. For example, if a particular outcome was initially reported as significant, the number of outcome events required to increase the

P value to ≥ 0.05 was determined. Conversely, if the outcome was initially reported as not significant, the number of outcome events required to decrease the P value to <0.05 was determined. The corresponding number indicated the number needed to reverse a particular outcome event and was recorded as the Fragility Index for that event (Figure 2). All event reversals were calculated in this manner with the median value representing the Fragility Index for the entire study population. Interquartile ranges (IQRs) were included where appropriate. In this manner, a comprehensive evaluation of all primary and secondary outcomes, as well as those findings that which were initially reported as significant ($P < 0.05$) along with those initially reported as insignificant ($P \geq 0.05$), was performed.

Statistical Analysis

The reported P value was recorded for each outcome event and verified for accuracy using the Fisher exact test. The total number of patients lost to follow-up was also determined for each trial. Statistical analyses were performed via the two-tailed Fisher exact test and chi-squared test with Yates correction when appropriate. IQRs were included when appropriate to provide a more comprehensive understanding and interpretation of reported median values.

Results

Of the 1,505 studies screened, 102 comparative studies were included for analysis, 40 of which were randomized controlled trials (RCTs). There were 339 total outcome events, of which 228 represented primary outcomes, with 98 reported as significant and 241 not significant. Of the 98 outcome events initially reported as having significant differences between two groups ($P < 0.05$), the average P value was 0.012. The median number of events to change these differences to insignificant findings was four (IQR, 2 to 12), with a total range of 1 to 145. Of the 241 outcome events initially reported as not significant ($P \geq 0.05$), the average P value was 0.59. The median number of events required to make these differences significant was six (IQR, 4 to 8), with a total range of 1 to 105. Therefore, the Fragility Index, or the median number of events required to reverse the statistical significance of the overall study, was only five (IQR, 3 to 8), with a total range of 1 to 145. This represents just 5.4% of the total study population. In addition, the average number of patients lost to follow-up was 7.9, which is greater than the number needed to change the significance of each arm of the study and the entire study population. No difference was appreciated between the randomized and nonrandomized trials. In evaluation of 140 outcome events in 40 RCTs, the median number of events required to reverse statistical significance was five (IQR, 3 to 7). In comparison, 199 outcome events were evaluated in 62 nonrandomized trials demonstrating a median number of six (IQR, 3 to 9) events required to reverse statistical significance.

Figure 2

	(-) Lachman	(+) Lachman
Hamstring Graft	2	25
Bone Patellar Tendon Bone Graft	8	17
P-Value		0.04

"The Flip"
$P < 0.05 \rightarrow P \geq 0.05$

	(-) Lachman	(+) Lachman
Hamstring Graft	*3*	*24*
Bone Patellar Tendon Bone Graft	8	17
P-Value		*0.09*

Table showing an example of the "flip" calculation done for each outcome event.

Discussion

Statistical interpretation of results in the orthopaedic sports medicine literature, which often informs clinical decision making, has relied heavily on P values and seems to be consistent with the greater published medical literature. The reporting of P values in journal abstracts increased from 7.3% in 1990 to 15.6% in 2014, with 33% of abstracts, 36% of meta-analyses, 39% of clinical trials, and 55% of RCTs reporting P value data in 2014.[8] Furthermore, medical journals have a propensity to publish statistically significant results with Chavalarias et al's[8] finding that 96% of abstracts and full-text articles in the biomedical literature reported at least 1 "statistically significant" result, with most reported P values between 0.05 and 0.001. However, the use of P values to inform conclusions in clinical research has been called into question because of inherent limitations that may be misleading and not well understood. Thus, it is critically important for the academic and clinical community to better understand the meaning of the P value: its strengths, limitations, and appropriate application for statistical inference. The Fragility Index has been proposed and used in the clinical epidemiology and biostatistics community as a useful metric for demonstrating how easily statistical significance based on a threshold P value may be overturned.[7] The P value has been further identified in the critical care literature as not providing statistically stable study results on which to base clinical decision making. In the evaluation of 56 multicenter RCTs, Ridgeon et al[9] advocate the "reporting of a Fragility Index for future trials in critical care to aid interpretation and decision making by clinicians."

Through our comprehensive evaluation of 102 dichotomous comparative studies, we identified a similar trend within orthopaedic sports medicine as the current body of literature is seemingly built on statistically unstable results with a Fragility Index of five. This means that only 5.4% of the total study population is needed to reverse trial significance. Because most study power analyses accept at least a 20% rate of patients being lost to follow-up, our findings clearly demonstrate that statistical significance and study stability hinge on relatively few events. As seen in our study, the average number of patients lost to follow-up was greater than the number needed to change the significance of each arm of the study (7.9 versus 5, respectively) and the entire study population. This finding suggests that simply ensuring improved patient follow-up could have caused a possible reversal of significance. As is distinctly evident, this creates concern when interpreting statistical findings for clinical practice because statistical significance does not infer clinical relevance. A P value of ≥ 0.05 only signifies that the evidence is not adequate to reject the null hypothesis.[10] This phenomenon does not necessarily imply the equivalence of two treatments; therefore, a clear distinction must be made between statistical significance and clinical relevance. The reader must possess the appropriate tools to aid in the accurate interpretation of clinical relevance because statistically significant differences based on isolated P values are seemingly too simplistic and misleading. Therefore, future efforts should be focused on encouraging the reporting of a Fragility Index in study results.

The evaluation and utilization of a Fragility Index have been described by several authors in other orthopaedic subspecialties and medical fields. Ridgeon et al[9] evaluated 862 trials from the critical care literature, of which 56 met the inclusion criteria, and found the Fragility Index to only be two (range, 1 to 3.5). In addition, greater than 40% of the trials were found to have a Fragility Index of less than or equal to 1, with 12.5% of the trials experiencing a loss to follow-up greater than their respective Fragility Index. Evaniew et al[11] analyzed the Fragility Index of randomized trials in the spine literature and found that in the 40 eligible trials with a sample size of 132 patients, the median Fragility Index was two (1 to 3). This means that adding two events to one of the trial's arms will eliminate its statistical significance. Furthermore, the Fragility Index was less than or equal to the number of patients lost to follow-up in 65% of the trials. Walsh et al[7] evaluated all the RCTs in high-impact medical journals including 399 eligible trials with a median sample size of 682 patients and a median of 112 events. They found the median Fragility Index to be eight (0 to 109), with 25% of trials demonstrating a Fragility Index of less than three. In 53% of all trials, the Fragility index was less than the number of patients lost to follow-up.

A recent systematic survey by Kahn et al[12] exclusively evaluated the fragility of statistically significant primary (or select secondary) outcomes from 48 RCTs published in 24 journals—JBJS-Am (4), *Knee Surgery, Sports Traumatology, Arthroscopy* (5), AJSM (6), Arthroscopy (14), and Other (19)—in the orthopaedic sports medicine and arthroscopy literature stating that "for each RCT, we extracted data for 1 statistically significant dichotomous outcome that was identified from the abstract. When more than 1 eligible outcome was presented, we chose the primary outcome whenever possible or the most patient-important secondary outcome." Through their analysis, they reported a Fragility Index of two and concluded that "most statistically significant RCTs in sports medicine and arthroscopic surgery are not robust because their statistical significance can be reversed by changing the outcome status on only a few patients in one treatment group." Although similar to our analysis in a few respects, obvious differences remain. In addition to 40 RCTs, we evaluated 62 nonrandomized trials, and, as such, our analysis can be more broadly applied to the greater body of sports medicine literature reporting on dichotomous comparative trials. In addition, as opposed to limiting our analysis to primary outcomes, we evaluated both primary and secondary outcomes within each study totaling 339 total outcome events as opposed to only 48. We further evaluated the fragility of each outcome event initially reported as nonsignificant in addition to those initially reported as significant. In comparison, Khan et al[12] limited their analysis to only outcome events initially reported as significant. Therefore, we feel that our median Fragility Index of five is more representative of all outcome data reported in the current literature with inclusion of nonrandomized trial data and both primary and secondary study outcomes, which are all vitally important when considering study validity and application of the fragility index. Most studies in the orthopaedic sports literature report multiple outcome events (ie, primary and secondary); therefore, each individual event, although time intensive, must be thoroughly evaluated. Of the 48 outcomes selected for evaluation by Khan et al,[12] almost half (44%) were either arbitrarily selected "patient-important secondary outcomes" (14, 29%) or "not specified" (7, 15%), with 17% of all originally reported outcome events losing statistical significance by simply selecting an alternative statistical test to calculate the initial reported

P value and thus demonstrating a Fragility Index of zero. Furthermore, Khan et al[12] report that 19 of their 48 included RCTs were published in "Other" journals of unknown, and likely lesser, impact factor. The remaining 29 studies were published in four notable journals with a final median journal impact factor of 3.2, as directly reported by Khan et al.[12] In comparison, the impact factor of the two journals from which we comprised our analysis was 5.280 (JBJS-Am) and 4.362 (AJSM). Thus, we feel that our analysis represents a more comprehensive critical evaluation of all reported outcome event data in the orthopaedic sports medicine literature.

All the authors mentioned earlier are consistent in their emphasis of caution when interpreting findings from dichotomous comparison trials reporting a low Fragility Index. Although it is true that study designs representing greater power through increased sample size will experience greater differences between outcome events and thus lower *P* values, we agree with Khan et al[12] in that the strength of dichotomous trials in the sports medicine literature can be most accurately quantified and easily interpreted through the appropriate inclusion and application of a Fragility Index. We therefore feel that it is essential for the Fragility Index to be reported in all future comparative trials to aid in the interpretation of study results and subsequent clinical decision making.

Strengths, Limitations and Bias

This study does have limitations. Our findings included specific topics related to the knee sports medicine literature, which does not allow us to make conclusions regarding the sports medicine literature in its entirety that includes shoulder and cartilage. Our search encompassed the analysis of 10 years of published trials in the two journals of highest impact and prominence in the orthopaedic sports literature. We feel this to be a strength because our average impact factor of 4.28 is substantially higher than that reported by the only other comparable analysis on this topic.[12] Furthermore, our analysis of both RCTs and nonrandomized trials represents a more comprehensive evaluation of the existing literature. For randomization to be successfully implemented, the randomization must be adequately concealed so that investigators and subjects are not aware of the upcoming intervention. The absence of adequate allocation concealment can lead to selection bias. Therefore, authors of randomized trials should provide enough details on how allocation concealment was achieved. We did not directly assess allocation concealment for each RCT included in our data set, so there remains a theoretical risk of inherent selection bias. However, we feel that our analysis of RCTs published in JBJS-Am and AJSM limits this potential source of bias compared with RCTs reported in journals of lesser impact in the orthopaedic sports medicine literature.

An additional strength lies in our evaluation of Fragility Indices of both primary and all secondary outcomes as opposed to limiting analysis to just primary or a single select secondary outcome. As stated previously, our study further evaluated Fragility Indices for each outcome reported as initially nonsignificant in addition to those outcomes initially reported as significant, thus demonstrating a more comprehensive analysis of study fragility compared with the methods used by the other meta-analysis on this topic.[12] For all potential sources of bias, it is important to consider the likely magnitude and direction of the bias. If all methodological limitations of studies were expected to bias the results toward a lack of effect, and the evidence demonstrates an effective intervention, it may be concluded that the intervention is effective even in the presence of these potential biases. As such, our method of analysis attempts to address the inherent and well-described outcome reporting bias present in randomized clinical trials in which statistically significant differences between the intervention groups are more likely to be reported than nonsignificant differences.[13] As with all meta-analyses, the quality of the analysis depends on the quality of data presented in each individual trial. As such, randomized trials, if appropriately designed and executed, prevent selection bias in allocating interventions to participants. Of our 102 trials analyzed, 62 represent nonrandomized trials which may be inherently prone to selection bias. However, we feel that a statistical analysis of dichotomous comparison trials would not be comprehensive with the select inclusion of only RCTs because a large majority of the orthopaedic sports literature consists of nonrandomized trials and that a broader analysis would prove consistent with our findings. In addition, we isolated our findings to clinical studies and excluded preclinical and translational research studies.

Conclusions

The results of comparative studies relying on categoric outcomes in the sports medicine literature may not be as stable as previously thought, with only a small percentage of outcome events required to change the significance of the entire study. The Fragility Index may be used as an effectual statistical complement because the clinical interpretation of outcomes

research, based on a single discreet *P* value cutoff, may be misleading. We thus recommend the thoughtful use and reporting of the Fragility Index, in addition to *P* value analysis, in the interpretation of statistical and clinical stability in the sports medicine literature.

References

References printed in **bold type** are those published within the past 5 years.

1. Cochrane A: Effectiveness and efficiency: Random reflections on health services. *BMJ* 1973;328:529.
2. Eddy DM: Clinical decision making: From theory to practice: Connecting value and costs: Whom do we ask, and what do we ask them? *JAMA* 1990;264: 1737-1739.
3. Parsons NR, Hiskens R, Price CL, Achten J, Costa ML: A systematic survey of the quality of research reporting in general orthopaedic journals. *J Bone Joint Surg Br* 2011;93:1154-1159.
4. Lefaivre KA, Shadgan B, O'Brien PJ: 100 most cited articles in orthopaedic surgery. *Clin Orthop Relat Res* 2011;469: 1487-1497.
5. Voleti PB, Donegan DJ, Baldwin KD, Lee GC: Level of evidence of presentations at American Academy of Orthopaedic Surgeons annual meetings. *J Bone Joint Surg Am* 2012;94:e50.
6. **Parisien RL, Dashe J, Cronin PK, Bhandari M, Tornetta P III: *Statistical Significance in Trauma Research: Too Unstable to Trust? AAOS Annual Meeting Abstract.* Orlando, FL, 2016.**
7. **Walsh M, Srinathan SK, McAuley DF, et al: The statistical significance of randomized controlled trial results is frequently fragile: A case for a fragility index. *J Clin Epidemiol* 2014;67:622-628.**
8. **Chavalarias D, Wallach JD, Li AH, Ioannidis JP: Evolution of reporting *P* values in the biomedical literature, 1990-2015. *JAMA* 2016;315:1141-1148.**
9. **Ridgeon EE, Young PJ, Bellomo R, Mucchetti M, Lembo R, Landoni G: The fragility index in multicenter randomized controlled critical care trials. *Crit Care Med* 2016;44:1278-1284.**
10. du Prel JB, Hommel G, Rohrig B, Blettner M: Confidence interval or *P*-value? Part 4 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* 2009;106: 335-339.
11. **Evaniew N, Files C, Smith C, et al: The fragility of statistically significant findings from randomized trials in spine surgery: A systematic survey. *Spine J* 2015;15: 2188-2197.**
12. **Khan M, Evaniew N, Gichuru M, et al: The fragility of statistically significant findings from randomized trials in sports surgery: A systematic survey. *Am J Sports Med* 2017; 45:2164-2170.**
13. Chan AW, Altman DG: Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *BMJ* 2005;330:753.