Original Article

Comparative Studies in the Shoulder Literature Lack Statistical Robustness: A Fragility Analysis

Robert L. Parisien, M.D., David P. Trofa, M.D., Patrick K. Cronin, M.D., Jesse Dashe, M.D., Emily J. Curry, M.P.H., Josef K. Eichinger, M.D., William N. Levine, M.D., Paul Tornetta III, M.D., and Xinning Li, M.D.

Purpose: Evidenced-based decision-making is rooted in comparative clinical studies; however, a small number of outcome event reversals have the potential to change study significance. The purpose of this study was to determine the utility of applying fragility analysis to comparative studies in the published orthopaedic shoulder literature. **Methods:** Comparative clinical shoulder research studies reporting 1:1 dichotomous categorical data were analyzed in 6 leading orthopaedic journals between 2006 and 2016. Statistical significance was defined as a *P* value of less than .05. The fragility index (FI) 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, thus changing the study conclusions. The associated fragility quotient (FQ) was determined by dividing the FI by the total population comprising a particular outcome. **Results:** Of the 23,897 studies screened, 3,591 met search criteria, with 198 comparative studies ultimately included for analysis, 67 of which were randomized controlled trials. There were 357 total outcome events with 74 reported as significant and 283 as not significant. The FI was 4 (IQR 2-6) with an associated FQ of 0.066 (interquartile range [IQR] 0.038-0.102). There was no difference in statistical fragility between randomized and nonrandomized trials with both revealing a FI of 4 and FQ of 0.068 (IQR 0.044-0.107) and 0.065 (IQR 0.031-0.101), respectively. **Conclusions:** This current analysis reveals that comparative shoulder studies published in six leading orthopaedic journals are at risk of statistical fragility. As such, contemporary clinical shoulder literature may not be as robust as traditionally perceived with the reversal of only a few outcome events required to change study significance. Therefore, we advocate the reporting of both FI and FQ in addition to the *P* value as statistical complements to all comparative investigations to provide a more comprehensive understanding of trial stability and significance in the published shoulder literature. **Clinical Relevance:** Comparative study designs are commonly employed in shoulder research. Several studies in both the general medical and orthopaedic literature have identified a lack of statistical robustness through comprehensive fragility analysis. Our findings demonstrate the *P* value may be an inadequate independent statistical metric requiring the complement of a FI and FQ to aid in the interpretation and understanding of study significance for clinical decision-making.

The primary objective of evidence-based medicine is to produce meaningful and clinically relevant information to help guide medical decision-making.[1] The viability of evidence-based medicine depends on validated research findings and an informed readership. Within the shoulder surgery literature, randomized controlled trials (RCTs) and dichotomous comparison studies are frequently used toward this end.[2] Statistical findings are typically reported in terms of a threshold probability value (*P* value) below which the null hypothesis (H_o) is rejected. By convention, the statistical cutoff is set at the arbitrary α threshold of 0.05. For a

From Boston University Medical Center, Boston, Massachusetts (R.L.P., J.D., P.T., X.L); Columbia University Medical Center, New York, New York (D.P.T., W.N.L.); Harvard-Combined Orthopaedic Residency Program, Boston, Massachusetts (P.K.C.); Boston University School of Public Health, Boston, Massachusetts (E.J.C.); and the Medical University of South Carolina, Charleston, South Carolina (J.K.E.), U.S.A.

The authors report the following potential conflicts of interest or sources of funding: X.L and J.K.E. report other from FH Ortho, outside the submitted work. P.T. reports other from Smith & Nephew, outside the submitted work. Full ICMJE author disclosure forms are available for this article online, as supplementary material.

Received September 4, 2020; accepted August 30, 2021.

Address correspondence to Xinning Li, M.D., Boston University School of Medicine, 850 Harrison Avenue — Dowling 2 North, Boston, MA 02115. E-mail: Xinning.li@gmail.com

2666-061X/201490

https://doi.org/10.1016/j.asmr.2021.08.017

comparison that is calculated to have a *P* value of $< .05$, this means there is less than a 5% likelihood that the observed difference is due to random chance.[3] *P*-value analysis and hypothesis testing hold a prominent place in the orthopaedic lexicon to aid in the interpretation of trial outcomes and guide both clinical and treatment decision-making.

Despite its ubiquity, the use of the *P* value to determine statistical significance in dichotomous comparison trials has received substantial academic criticism. Such criticisms cite instances in which the *P* value may be overvalued without regard for important study design factors such as sample size, significant loss to follow up or lack of sufficient power.[4-6] Failure to consider these potential study limitations can lead to overconfidence in a study conclusion in which statistical significance, when taking these other factors into consideration, is a relatively fragile statistical measure. The concept of fragility encompasses the idea that a change in only a few outcome events may alter the overall conclusion of a trial. The reporting of a fragility index (FI) in research studies addresses the limitation of an a priori *P* value statistical significance threshold of .05 and serves as an additional statistical measure to help lend support to study conclusions. Studies with lower susceptibility to fragility are not affected by small alterations in results, thus lending more confidence to the conclusions. The identification of studies in which the results are fragile or susceptible to change is important. Rather than discounting or rejecting studies with potential fragility, the adoption of an understandable and easily conveyed statistical tool that directly examines statistical fragility may be considered. One proposed solution is the reporting of the FI for RCTs by determining the number of outcome events required to overturn, or flip, a statistically significant result to a non-significant result and vice-versa.[7] The concept of the FI was proposed by Feinstein in 1990 as the "unit fragility" and has since been applied to many disciplines of medicine with an alarming amount of statistical fragility noted across several specialties.[7-18] Although the FI represents a simplified and intuitive metric allowing for a clearer interpretation of study stability, it is an absolute measure that is dependent on study cohort size. Therefore, the fragility quotient (FQ) has been proposed by Ahmed et al.[19] as a relative measure of fragility that takes into account both the FI and sample size by dividing the FI by the total sample size. Reporting both the FI and FQ in conjunction with the *P* value may provide a more comprehensive and straightforward understanding of study stability in the context of the sample size.

The purpose of this study was to determine the utility of applying fragility analysis to comparative studies in the published orthopaedic shoulder literature in the top 6 orthopaedic journals. We hypothesized that the published orthopaedic shoulder literature in the top orthopaedic journals would prove relatively fragile with few outcome events required to reverse study significance.

Methods

Comparative clinical shoulder research studies pertaining to rotator cuff pathology, instability and arthritis in the *Journal of Bone and Joint Surgery* (JBJS; impact factor [IF] of 4.6), *American Journal of Sports Medicine* (AJSM; IF 6.1), *Arthroscopy* (IF 4.3), *Knee Surgery-Sports Traumatology-Arthroscopy* (KSSTA; IF 3.1), *Clinical Orthopaedics and Related Research* (CORR; IF 4.1), and the *Journal of Shoulder and Elbow Surgery* (JSES; IF 2.3) were retrieved from 2006 to 2016. These journals were selected due to their particular prominence in the published shoulder literature and recognized as 6 of the most impactful orthopaedic journals with impact factors of 4.6, 6.1, 4.3, 3.1, 4.1, and 2.3, respectively. Thus, analysis of 10 years of data within these 6 distinguished journals provides a well-represented sample of the shoulder literature. To identify a comprehensive list of relevant comparative studies, the following search terms were queried in the PubMed database: "rotator cuff," "shoulder arthritis," and "shoulder instability." These terms were combined in various permutations and combinations using Boolean operators to maximize the identification of relevant studies. Inclusion criteria included dichotomous comparative studies reporting categorical and *P* value statistical data. The type of outcome measure used (primary, secondary or unknown) also was documented as well as whether the particular outcome came from am RCT or non-randomized controlled trial (non-RCT).

Analysis was performed by manipulating the reported outcome events in a 2×2 contingency table until a reversal of significance was appreciated, with statistical significance defined as a *P* value of less than 0.05 (Fig 1).[13] For example, if a particular outcome was initially reported as significant, the number of outcome events required to raise *P* to $\geq .05$ was determined. Conversely, if the outcome was initially reported as not significant, the number of outcome events required to decrease *P* to $< .05$ was determined. The corresponding number indicates the number needed to reverse a particular outcome event and was recorded as the FI for that event. All overturned outcome events were calculated in this manner with the median value representing the FI for the entire study population. The FQ was simultaneously determined for each outcome event by dividing the FI by the sample size. In addition, the total FQ for all outcome events as well as the FQ for RCTs and non-RCTs was determined. The reported *P* value was recorded for each outcome event and verified for accuracy using the 2-tailed Fisher exact test. Interquartile ranges (IQRs) were calculated to provide a more comprehensive understanding and interpretation

	+ Instability	- Instability
Arthroscopic Repair	23	75
Open Repair	11	87
***P* Value**		0.04

	+ Instability	- Instability
Arthroscopic Repair	23	75
Open Repair	*12*	*86*
***P* Value**		*0.06*

Fig 1. Demonstration of the reversal of statistical significance with resultant fragility index (FI) = 1.

of the reported variability and dispersion as the difference between the 25th and 75th percentiles.

Results

Of the 23,897 studies screened, 3,591 met search criteria with 198 comparative studies included for analysis, 67 of which were RCTs (Fig 2). There were 357 total outcome events with 74 initially reported as statistically significant ($P < .05$) and 283 initially reported as not statistically significant ($P \geq .05$). Of the 74 outcomes initially reported as statistically significant, the median number of events required to reverse significance (FI) was 3 with a range of 1 to 96 (IQR 1-6) (Table 1). The associated FQ for statistically significant outcomes was 0.039 with a range of 0.005 to 0.333 (IQR 0.016-0.087). Of the 283 outcomes initially reported as not statistically significant, the median number of events required to reverse significance (FI) was 5 with a range of 1 to 14 (IQR 3-6). The associated FQ for initially non-significant outcomes was 0.071 with a range of 0.003 to 0.526 (IQR 0.045-0.105).

The final FI, or median number of events required to change the statistical significance of the overall study, was only 4 with a range of 1 to 96 (IQR 2-6). The final FQ, incorporating all 357 outcome events, was 0.066 with a range of 0.003 to 0.526 (IQR 0.038-0.102). Evaluation of 125 dichotomous outcome events evaluated in 67 RCTs also demonstrated a median number of 4 events required to reverse statistical significance with a range of 1 to 11 (IQR 3-6). Nonrandomized studies did not differ in statistical fragility in comparison to RCTs. Evaluation of 232 dichotomous outcome events in 131 non-RCTs revealed the median number of events required to reverse statistical significance (FI) as 4 with a range of 1 to 96 (IQR 2-6). The associated FQ for the RCTs was 0.068 with a range of 0.010 to 0.526 (IQR 0.044-0.107) and the FQ for the non-RCTs was 0.065 with a range of 0.003 to 0.333 (IQR 0.031-0.101)

Discussion

This study demonstrates a number of important findings regarding the statistical interpretation or fragility of shoulder-related investigations published over a 10-year period. Through comprehensive evaluation of 3,591 studies meeting our inclusion criteria, 198 comparative studies with dichotomous outcomes were identified. We found that the current body of shoulder-specific research (rotator cuff, shoulder arthritis, and shoulder instability) demonstrates study fragility with a median FI of 4. The FQ for all included studies was identified as 0.066, meaning that if only 6.6% of the patients in one arm of the trial were to experience an alternative outcome, the resultant effect would be the reversal of study significance. Furthermore, RCTs were found to be as statistically fragile as nonrandomized trials, with both demonstrating a FI of 4 and FQs of 0.068 and 0.065, respectively. In subgroup analysis of all 357 outcome events, those initially reported as significant ($P < .05$) were found to represent increased statistical fragility as compared to those initially reported as not significant ($P \geq .05$) with a FI of 3 versus 5 and a FQ of 0.039 versus 0.071, respectively.

There is a strong reliance on the interpretation of statistically significant results via *P*-value analysis to guide clinical decision-making in orthopaedic practice. Chavalarias et al.[10] found that 96% of abstracts and full-text articles published in the biomedical literature report a minimum of one "statistically significant" result ranging from $<.05$ to $<.001$, thus highlighting the pervasiveness of publication bias, or the tendency to report statistically significant findings. However, recent literature across multiple specialties has highlighted the inherent limitations of using *P* values to guide clinical decision-making.[8-18] For instance, a number of factors can influence the *P* value ,including the variables' effect size, sample size, and data dispersion.[11] As such, R.A Fisher, along with most statisticians, regard an α of 0.05 as completely arbitrary and rather prefer differing α values for given circumstances as opposed to a strict statistical cut-off of 0.05. Therefore, the FI has been proposed as a straightforward and useful metric for determining the stability or fragility of a *P* value in trials with dichotomous outcomes by identifying the minimum number of patients that, by reassigning an event status, would overturn a statistically or non-statistically significant result.[7-18] However, similar to the *P* value, the FI is an imperfect measure of trial stability in isolation as it provides an absolute measure of fragility without reference to sample size. Potter[20] has recently criticized the use of a FI in favor of sensitivity analysis to quantify the robustness of trial results. In her analysis, Potter correctly addresses the limitation of the FI in isolation as it may inappropriately penalize small trials

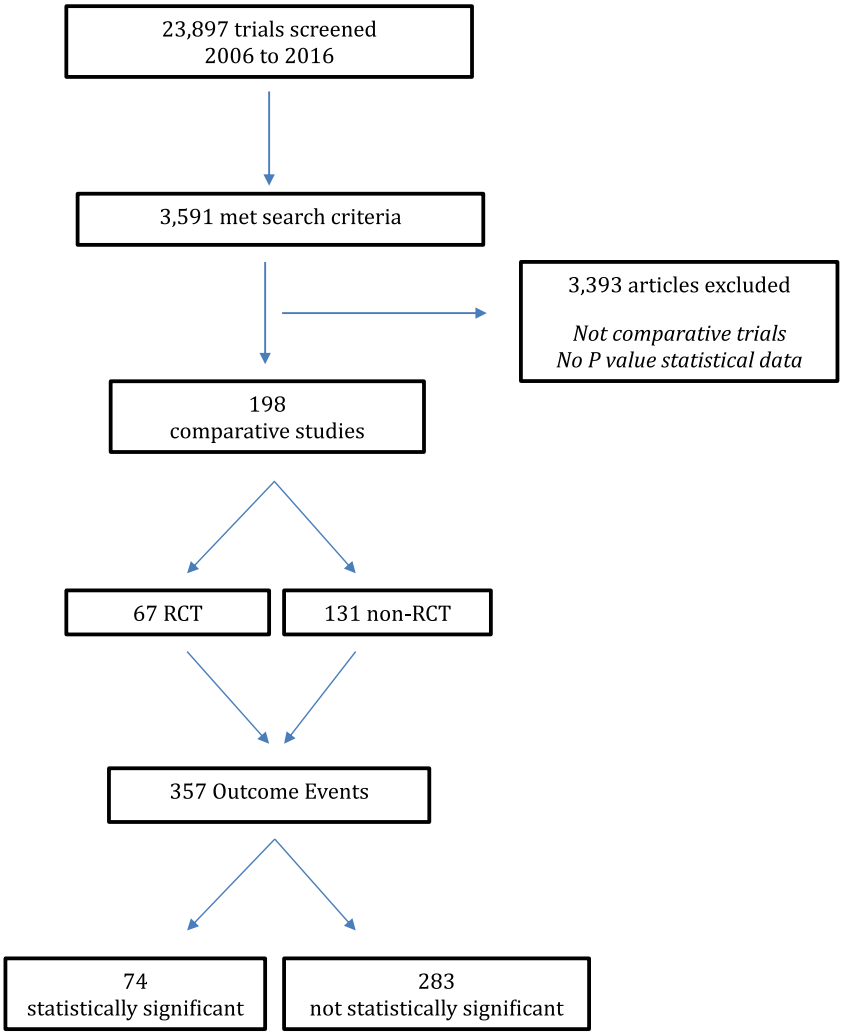

Fig 2. Study identification flowchart. RCT, randomized controlled trial.

for utilizing fewer events but fails to recognize the complementary value of the FQ, which accounts for sample size. Furthermore, confidence intervals (CIs) are often reported alongside *P* value analysis in an attempt to indicate probabilistic zones within which the true value is located while taking into account sample size which affects the CI range, with a larger sample size resulting in a smaller CI and vice versa. The CI attempts to account for all possible variations that may occur under the null hypothesis, yet similar to the α cut-off for *P* values, the CI is typically arbitrarily set at 95%. Furthermore, the FI is a single discreet value allowing for ease of reporting and interpretation while CIs may provide confusion for clinicians attempting to draw meaningful clinical conclusions.[21] As previously mentioned, some authors have suggested that the FI is an imperfect tool for evaluating statistical significance of trials with few outcome events.[20] As such, the purpose of the current investigation is not to criticize the utility of *P* values and CIs, but rather suggest additional reporting measures (FI and FQ) to augment *P* value interpretation in an effort to provide a more simplified and understandable determination of statistical stability and quantitative significance of trials in the orthopaedic shoulder literature.

As demonstrated previously, the inclusion of the FQ in statistical analysis provides a relative measure of

Table 1. Fragility Data Based on Trial and Outcome Characteristics

Characteristic	Events	Fragility Index (IQR)	Fragility Quotient (IQR)
All trials	357	4 (2-6)	0.066 (0.038-0.102)
RCT	125	4 (3-6)	0.068 (0.044-0.107)
Non-RCT	232	4 (2-6)	0.065 (0.031-0.101)
Outcome			
Primary	42	5 (2-7)	0.069 (0.045-0.111)
Secondary	80	5 (3-7)	0.066 (0.041-0.100)
Not specified	235	4 (2-6)	0.065 (0.033-0.103)
Reported *P* value			
$P < .05$	74	3 (1-6)	0.039 (0.016-0.087)
$P \geq .05$	283	5 (3-6)	0.071 (0.045-0.105)

IQR, interquartile range; RCT, randomized controlled trial.

stability and quantitative significance by taking into account sample size with direct reference to the FI. In isolation, a FI of 5 within a cohort of 535 may have vastly different clinical relevance as compared to the same FI value pertaining to a cohort of 35. Thus, the integration of the FQ aids in providing a more complete understanding of the *P* value and FI in determination of true trial significance by taking into consideration sample size and thereby taking into account the chance of Type II error within the study results. Ruzbarsky et al.[21] recently published their shoulder and elbow fragility analysis but limited their evaluation to RCTs (N = 30) and primary outcome events (N = 30), thus resulting in a significantly less comprehensive and robust analysis as compared to our current study.

The level of evidence in shoulder-specific research presented at the American Shoulder and Elbow Surgeons annual scientific meeting is increasing over time.[22] Comparative studies and RCTs evaluated in this current statistical fragility analysis of 6 of the highest impact journals in the orthopaedic shoulder literature are relatively unstable. The FI and associated FQ observed among these studies is concerning given the presented findings in these leading orthopaedic journals are considered among the best evidence available in the field of shoulder surgery and the findings of these studies are readily applied to clinical practice and treatment decisions. Establishing guidelines that encourage reporting of a FI and associated FQ, when dichotomous comparative outcomes are investigated, is one method that may strengthen the critical analysis and interpretation of clinical relevance among publications. This recommendation is not novel, as the evaluation of a FI has been described by a number of prior investigations supporting its routine use in comparative literature. For instance, Ridgeon et al.[14] identified a FI of 2 among 56 investigations meeting their inclusion criteria within the critical care literature. In the spine literature, Evaniew et al.[15] analyzed the FI of randomized trials finding that, among 40 eligible trials, the FI was 2 (range 1-3). Furthermore, among randomized controlled trials in high-impact medical journals inclusive of 399 eligible investigations, the FI was 8 (range 0-109), with 25% of trials demonstrating a FI of less than 3.[13] Similar data with regards to the use of a FI have been published in pediatrics, nephrology and ophthalmology with reported indices of 7, 3 and 2, respectively.[9,11,12] All of the aforementioned authors emphasize caution when interpreting statistical findings from dichotomous comparison studies that exhibit a low FI. In addition, Kahn et al.,[16] in their analysis of RCTs in the sports medicine literature, argued that the strength of dichotomous trials can be quantified and most easily interpreted through the appropriate inclusion and application of a FI, in addition to *P* value analysis. In further evaluation of 339 outcome events from 102 comparative trials in the sports medicine literature, Parisien et al.[17] came to the same conclusion, having demonstrated quantitative significance with FI of 5. Additional support for the reporting of both a FI and associated FQ in the orthopaedic literature has been demonstrated in close evaluation of 198 comparative studies consisting of 20 years of published orthopaedic trauma literature. Parisien et al.[18] reported a FI of 5 and FQ of 0.046 in evaluation of 775 outcome events and thus made a recommendation for the inclusion of a FI and associated FQ to aid in the evaluation and interpretation of isolated *P* values. Our findings demonstrate the *P* value may be an inadequate independent statistical metric requiring the complement of a FI and FQ to aid in the interpretation and understanding of study significance. We therefore advocate for triple reporting of the *P* value, FI and FQ in all comparative investigations to provide a more comprehensive understanding of trial stability and significance in the published shoulder literature including both RCTs and non RCTs.

Limitations

Given our search encompassing 10 years of published shoulder research in 6 prominent orthopaedic journals, it is unknown whether lower impact journals contain higher or lower rates of fragility and if the inclusion of additional journals would have significantly impacted our results. Furthermore, per convention, fragility analysis may only be applied to comparative studies reporting dichotomous outcomes. Given this relatively limited scope, fragility analysis may not be an appropriate statistical method for the remainder of studies in the shoulder literature. Finally, discreet FI and FQ threshold values or threshold ranges have yet to be determined. Future investigations are thus required to better understand the relationship of fragility analyses of statistical significance with clinical significance.

Conclusions

The current analysis reveals that comparative shoulder studies published in 6 leading orthopaedic journals are at risk of statistical fragility. As such, contemporary clinical shoulder literature may not be as robust as traditionally perceived with the reversal of only a few outcome events required to change study significance.

References

1. Shah HM, Chung KC. Archie Cochrane and his vision for evidence-based medicine. *Plast Reconstr Surg* 2009;124:982-988.
2. Gartsman GM, Morris BJ, Unger RZ, Laughlin MS, Elkousy HA, Edwards TB. Characteristics of clinical shoulder research over the last decade: A review of shoulder articles in The Journal of Bone & Joint Surgery from 2004 to 2014. *J Bone Joint Surg Am* 2015;97:e26.

3. Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing: An explanation for new researchers. *Clin Orthop Relat Res* 2010;468:885-892.
4. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;294:218-228.
5. Pocock SJ. Current issues in the design and interpretation of clinical trials. *Br Med J (Clin Res Ed)* 1985;290:39-42.
6. Sterne JA, Smith GD. Sifting the evidence-what's wrong with significance tests? *Phys Ther* 2001;81:1464-1469.
7. Feinstein AR. The unit fragility index: An additional appraisal of "statistical significance" for a contrast of two proportions. *J Clin Epidemiol* 1990;43:201-209.
8. Docherty KF, Campbell RT, Jhund PS, Petrie MC, McMurray JJV. How robust are clinical trials in heart failure? *Eur Heart J* 2017;38:338-345.
9. Matics TJ, Khan N, Jani P, Kane JM. The fragility index in a cohort of pediatric randomized controlled trials. *J Clin Med* 2017;6(8).
10. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA* 2016;315:1141-1148.
11. Shochet LR, Kerr PG, Polkinghorne KR. The fragility of significant results underscores the need of larger randomized controlled trials in nephrology. *Kidney Int* 2017;92:1469-1475.
12. Shen C, Shamsudeen I, Farrokhyar F, Sabri K. Fragility of results in ophthalmology randomized controlled trials: A systematic review. *Ophthalmology* 2018;125:642-648.
13. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, 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.
14. 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.
15. Evaniew N, Files C, Smith C, Bhandari M, Ghert M, Walsh M, et al. The fragility of statistically significant findings from randomized trials in spine surgery: A systematic survey. *Spine J* 2015;15:2188-2197.
16. Khan M, Evaniew N, Gichuru M, Habib A, Ayeni OR, Bedi A, 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.
17. Parisien RL, Trofa DP, Dashe J, Cronin PK, Curry EJ, Fu FH, et al. Statistical fragility and the role of P values in the sports medicine literature. *J Am Acad Orthop Surg* 2019;27:e324-e329.
18. Parisien RL, Dashe J, Cronin PK, Bhandari M, Tornetta P 3rd. Statistical significance in trauma research: Too unstable to trust? *J Orthop Trauma* 2019;33:e466-e470.
19. Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility index? *Crit Care Med* 2016;44:e1142-e1143.
20. Potter GE. Dismantling the fragility index: A demonstration of statistical reasoning. Stat Med 2020;39:3720-3731. Mittal N, Bhandari M, Kumbhare D. A tale of confusion from overlapping confidence intervals. *Am J Phys Med Rehabil* 2019;98:81-83.
21. Ruzbarsky JJ, Rauck RC, Manzi J, Khormaee S, Jivanelli B, Warren RF. The fragility of findings of randomized controlled trials in shoulder and elbow surgery. *J Shoulder Elbow Surg* 2019;28:2409-2417.
22. Kay J, Memon M, de Sa D, Simunovic N, Athwal GS, Bedi A, et al. Level of clinical evidence presented at the open and closed American Shoulder and Elbow Surgeons annual meeting over 10 years (2005-2014). *BMC Musculoskelet Disord* 2016;17:470.