OPEN Research Article

The Fragility of Statistical Findings in Achilles Tendon Injury Research: A Systematic Review

Robert L. Parisien, MD
Nicholas C. Danford, MD
Ian J. Jarin, MBS
Xinning Li, MD
David P. Trofa, MD
J. Turner Vosseller, MD

From the Harvard Medical School, Boston, MA (Dr. Parisien); the Columbia University Medical Center, New York, NY (Dr. Danford, Dr. Trofa); the Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ (Jarin); the Boston University Medical Center, Boston, MA (Dr. Li); and the Jacksonville Orthopaedic Institute, Jacksonville, FL (Dr. Vosseller).

Correspondence to Dr. Parisien: Robert.L.Parisien@gmail.com

Dr. Parisien or an immediate family member serves as a board member, owner, officer, or committee member of American Orthopaedic Society for Sports Medicine, Society of Military Orthopaedic Surgeons, and is on the Editorial or Governing Board of Arthroscopy, Sports Medicine, and Rehabilitation. Dr. Li or an immediate family member has received royalties from FH Ortho; serves as a paid consultant to FH Ortho; and is on the Editorial or Governing Board of Journal of Bone and Joint Surgery–American, American Journal of Sports Medicine, Journal of Medical Insight (JOMI), Orthopedic Reviews, and World Journal of Orthopaedics. Dr. Vosseller or an immediate family member has received royalties from New Clip Technics; serves as a board member, owner, officer, or committee member of AAOS, American Orthopaedic Foot and Ankle Society; is a member of a speakers' bureau or has made paid presentations on behalf of Arthrex, Inc; has received research or institutional support from Acumed, LLC; and is on the Editorial or Governing Board of Foot and Ankle Orthopaedics. 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. Danford, Dr. Jarin, and Dr. Trofa.

JAAOS Glob Res Rev 2021;5:1-6

DOI: 10.5435/JAAOSGlobal-D-21-00018

ABSTRACT

Introduction: Probability analysis with the reporting of *P* values is often used to determine the statistical significance of study findings in the Achilles literature. The purpose of this study was to determine the utility of applying a fragility analysis to comparative trials evaluating Achilles tendon injuries.

Methods: We identified all dichotomous outcome data for comparative studies of Achilles tendon injuries published in 11 orthopaedic journals from 2000 to 2020. The fragility index (FI) was determined by the number of event reversals required to change a *P* value from less than 0.05 to greater than 0.05, or vice-versa. The associated fragility quotient was determined by dividing the FI by the sample size.

Results: Of the 51,357 studies screened, 1,487 met the search criteria, with 51 comparative studies and 177 total outcome events included for analysis. The overall FI was only 4 with an associated fragility quotient of 0.048. One-half of the studies failed to report lost to follow-up data, with an additional 21.6% reporting loss to follow-up of greater than or equal to 4.

Conclusion: Our fragility analysis suggests that Achilles tendon injury outcomes are not as statistically stable as previously thought and should be interpreted with caution.

Evidence-based medicine aids orthopaedic surgeons by providing accurate data of treatment outcomes to properly inform clinical decision-making. To make appropriate data-driven decisions, clinicians must possess a comprehensive understanding of statistical findings.[1] Researchers often present statistical findings in the form of *P* values, a threshold below which the null hypothesis (H_0) is rejected in favor of the alternate hypothesis (H_1). By convention, the value for this threshold is 0.05, which means that the observed difference has a 5% probability of occurring by random chance.[2] *P* value interpretation is ubiquitous within orthopaedic research and acts as a guide for clinical decision-making through the determination of statistical significance.

The optimal treatment of Achilles tendon ruptures remain controversial because inconclusive findings across the literature have provided more questions than

answers with the persistence of a robust debate between nonaurgical and surgical management.[3-7] Further debate exists regarding the surgical management with some evidence suggesting that minimally invasive repair is superior to open repair, especially about wound complications.[8-10] As in most orthopaedic literature, *P* values are used to determine the statistical significance of such comparisons. However, the *P* value has received recent scrutiny and criticism, within the academic community, because it may not correlate with clinical significance. For example, the *P* value may be misinterpreted if the data sample from which it is generated contains substantial loss to follow-up, lacks sufficient statistical power, or contains confounding variables.[11-14]

One way to improve the interpretation of a *P* value is by using the statistical concept of fragility. The fragility of a given statistic is the change in outcome events necessary to alter the overall conclusion drawn from it. When applied to *P* value analysis, the fragility index (FI) provides the investigator with the number of outcome events required to change the value from less than 0.05 to greater than 0.05, thus altering the assessment of its significance. The FI therefore attempts to address the problem with a specific numeric threshold established by convention. The FI was initially proposed by Feinstein in 1990 and has helped inform a body of literature that emphasizes the statistical fragility of findings across various medical disciplines.[15-22] It has been applied across multiple orthopaedic subspecialties including spine, sport, trauma, and shoulder surgery.[18,23-25] To further account for the differences in sample size, Ahmed et al[26] proposed the concept of a fragility quotient (FQ), which is a measure of quantitative significance and is determined by dividing the FI by the sample size. In conjunction with the *P* value analysis, the FI and FQ aid in the interpretation of trial fragility and robustness. As such, studies that possess a low susceptibility to fragility are stronger in their conclusions than studies with high susceptibility to fragility.

The purpose of our study was to determine the statistical fragility of comparative studies in the Achilles tendon injury literature with utilization of FI and FQ analysis. We hypothesized a high susceptibility to fragility within the comparative literature of Achilles tendon injuries.

Methods

Comparative clinical studies of Achilles tendon rupture management published in 11 prominent orthopaedic journals from 2000 to 2020 were evaluated. These journals consisted of the *Journal of Bone and Joint Surgery, American Journal of Sports Medicine, Orthopaedic Journal of Sports Medicine, Foot & Ankle International, Journal of Foot & Ankle Surgery, Foot and Ankle Surgery, Bone & Joint Journal, Knee Surgery, Sports Traumatology, Arthroscopy, Clinical Orthopaedics and Related Research, Journal of ISAKOS,* and *Journal of the American Academy of Orthopaedic Surgeons*. The journals were selected for their relative prominence within the field of orthopaedic sport medicine and foot and ankle surgery with associated 2019 impact factors listed in Table 1.

The initial search was performed in PubMed with the search criteria "Achilles" and publication date between January 1, 2000, and June 31, 2020. Inclusion criteria were comparative studies reporting dichotomous categorical data and associated *P* values. Exclusion criteria consisted of studies reporting cadaveric data, animal data, in vitro data, nondichotomous data, and those with more than two treatment groups and systematic reviews. The following data from included studies were extracted: first author, journal title, year of publication, the number of study outcomes per group, primary versus secondary outcome, intervention, lost-to-follow up (LTF), *P* value, and the type of study (randomized controlled trial and nonrandomized controlled trial).

Fragility analysis was performed by manipulating the reported outcome events in a 2 × 2 contingency table until a reversal of significance was determined, with statistical significance defined as a *P* value of less than 0.05. For example, if a particular outcome was initially reported as significant, the number of outcome events required to increase *P* to greater than or equal to 0.05 was determined (Figure 1). If the outcome was initially nonsignificant, the number of outcome events required to decrease *P* to less than 0.05 was determined. The number of altered events required to overturn significance was recorded as the FI for a particular outcome. This was performed for each outcome event with the median value representing the median FI for the entire study. The FQ was calculated for each outcome event by dividing the FI by the corresponding sample size. Interquartile ranges (IQR) for FI and FQ were calculated as the difference between the 25th and 75th percentiles. Fragility analysis was performed for the following subgroups: primary versus secondary outcomes, rerupture, infection/wound complication, return to sport/activity, significant versus nonsignificant outcomes, randomized controlled trial (RCT) versus non-RCTs, studies published in the time periods of 2000 to 2004, 2005 to 2009, 2010 to 2014, and 2015 to 2020 (Table 2).

Table 1. 2019 Impact Factors of Included Journals

Journal	Impact Factor
American Journal of Sports Medicine	5.810
Journal of Bone and Joint Surgery	4.578
Clinical Orthopaedics and Related Research	4.329
Bone & Joint Journal	4.306
Knee Surgery, Sports Traumatology, Arthroscopy	3.166
Orthopaedic Journal of Sports Medicine	2.492
Foot & Ankle International	2.292
Journal of the American Academy of Orthopaedic Surgeons	2.286
Foot and Ankle Surgery	1.776
Journal of Foot & Ankle Surgery	1.043
Journal of ISAKOS	N/A

2019 journal impact factors obtained from InCites Journal Citation Reports.

Results

Of the 51,357 studies screened, 1,487 met the search criteria with 51 comparative studies included for analysis (Figure 2). There were 177 total outcome events with 32 initially reported as statistically significant ($P < 0.05$) and 145 initially reported as not statistically significant ($P \geq 0.05$). Of the 32 outcomes initially reported as statistically significant, the median number of events required to reverse significance (FI) was 2.5 (IQR, 1 to 5.5) (Table 2). The associated FQ for statistically significant outcomes was 0.026 (IQR, 0.014 to 0.059). Of the 145 outcomes initially reported as not statistically significant, the median number of events required to reverse significance (FI) was 4 (IQR, 3 to 6). The associated FQ for initially nonsignificant outcomes was 0.057 (IQR, 0.025 to 0.097). Therefore, statistically significant outcomes were 37.5% more fragile than nonsignificant outcomes. Of the 177 total outcome events, 9.6% (17) consisted of primary outcomes, whereas the remaining 90.4% (160) consisted of secondary outcomes. Primary outcomes were found to be slightly more stable than secondary outcomes, with a FI of 7 (IQR, 3.5 to 10) and 4 (IQR, 2 to 6), respectively. The associated FQ was 0.063 (IQR, 0.030 to 0.09) and 0.048 (IQR, 0.017 to 0.093), respectively. A subanalysis evaluating 30 outcome events relating to tendon rerupture demonstrated a FI of 3 (IQR, 2 to 6) and associated FQ of 0.027 (IQR, 0.010 to 0.063).

Table 2. Fragility Data Based on Trial and Outcome Characteristics

Characteristic	Events	Fragility Index (Interquartile Range)	Fragility Quotient (Interquartile Range)
All trials	177	4 (2-6)	0.048 (0.021-0.097)
Outcomes			
Primary	17	7 (3.5-10)	0.063 (0.030-0.09)
Secondary	160	4 (2-6)	0.048 (0.017-0.093)
Rerupture	30	3 (2-6)	0.027 (0.010-0.063)
Infection/wound complication	21	3 (2-4)	0.024 (0.013-0.106)
Return to sport/activity	17	5 (3.5-6)	0.078 (0.037-0.112)
Reported *P* value			
$P < 0.05$	32	2.5 (1-5.5)	0.026 (0.014-0.059)
$P \geq 0.05$	145	4 (3-6)	0.057 (0.025-0.097)
Comparative trial			
RCT	63	4 (2-8)	0.063 (0.033-0.085)
Non-RCT	114	4 (3-6)	0.041 (0.013-0.098)
Year of publication			
2000-2004	15	4 (3-5)	0.042 (0.030-0.097)
2005-2009	71	3 (2-6)	0.038 (0.013-0.122)
2010-2014	25	4 (2-6)	0.058 (0.012-0.010)
2015-2020	66	4 (3-7)	0.062 (0.031-0.085)

Figure 1

	Outcome A	Outcome B
Treatment A	1	20
Treatment B	6	12
P Value		0.04

	Outcome A	Outcome B
Treatment A	1	20
Treatment B	*5*	*13*
P Value		*0.08*

Demonstration of reversal of statistical significance with resultant Fragility Index (FI) = 1

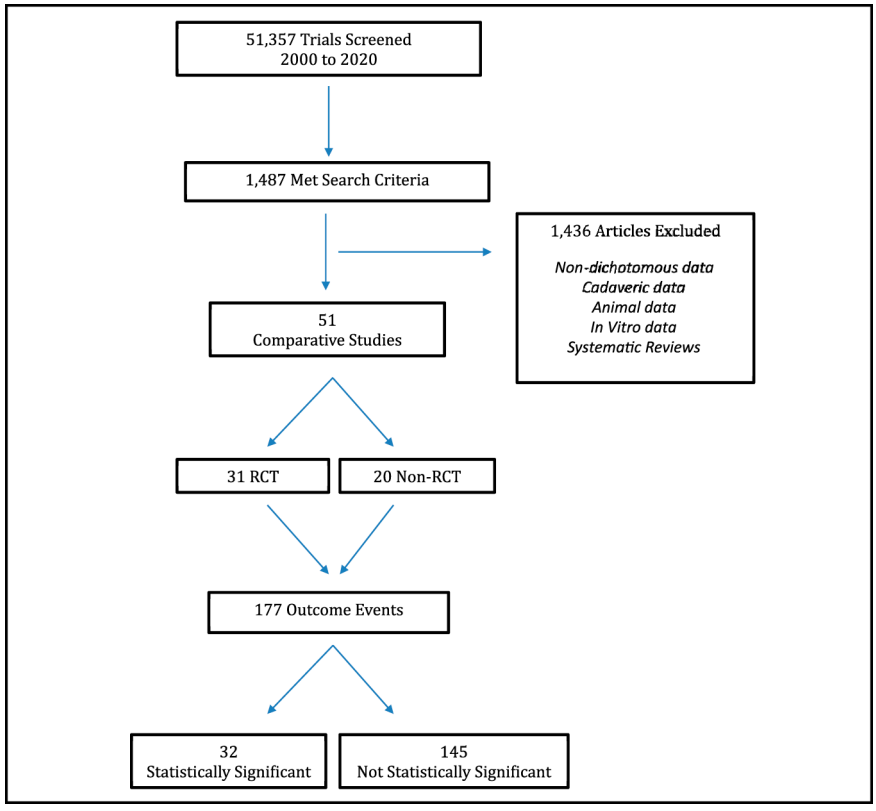
Accounting for sample size, the FQ for rerupture translates to a reversal of significance through a change of 2.7% of outcome events. Outcomes relating to infection/wound complication (21 events) demonstrated a similar level of fragility with a FI of 3 (IQR, 2 to 4) and FQ of 0.024 (IQR, 0.013 to 0.106). In addition, return to sport/activity outcomes (17 events) demonstrated a FI of 5 (IQR, 3.5 to 6) and FQ of 0.078 (0.037 to 0.112). There was no difference realized in the analysis of comparative trial type with both RCTs and non-RCTs demonstrating an identical FI of four. Further fragility subanalysis per year of publication identified a FI of four from 2000 to 2004, a FI of three from 2005 to 2009, a FI of four from 2010 to 2014, and a FI of four from 2015 to 2020, thus demonstrating consistent statistical fragility over the 20-year period (Table 2).

The overall FI, incorporating 177 outcome events from all 51 comparative studies, was only 4 (IQR, 2 to 6). The overall FQ was 0.048 (IQR, 0.021 to 0.097), indicating the reversal of only 4.8 patients of 100 is required to alter study significance of all included RCTs and non-RCTs when accounting for sample size. Of the 51 included studies, 49% (25) failed to report LTF data with an additional 21.6% (11) reporting LTF of ≥ 4. Therefore, 70.6% (36) of all included studies either did not report LTF data or reported an LTF greater than or equal to the overall FI of 4.

Discussion

This is the first study to report the FI and FQ for Achilles tendon injury data. Our hypothesis of a high susceptibility

Figure 2



Study identification flowchart

to statistical fragility within the literature of Achilles tendon injury was validated. Regardinlg the statistically significant results, a median of only 2.5 outcome events were required to reverse statistical significance. For nonsignificant results, a median of only four outcome events was required to grant a particular result as being statistically significant. Our results therefore reveal the statistical fragility of outcome data for Achilles tendon injuries and the statistical fragility of subgroups including primary and secondary outcomes, rerupture, infection/wound complication, return to sport/activity, significant and nonsignificant outcomes, and outcomes from four distinct time intervals over the course of 20 years. We further demonstrate the critical importance of accurate reporting of LTF data because 70.5% of studies either failed to report their LTF data or reported a LTF of greater than the overall study FI of four, thus suggesting the realization of a reversal of significance by simply maintaining a study follow-up.

P value analysis of comparative studies involving Achilles tendon injuries is fragile and therefore needs to be interpreted with caution. Similarly, *P*-values from other subspecialties within orthopaedic surgery have proven fragile as well. In evaluation of 339 outcome events across 102 comparative studies in the sport medicine literature, Parisien et al[24] demonstrated a FI of five. Fragility analysis by Khan et al[23] of 48 sport medicine and arthroscopic surgery RCTs revealed a FI of only two. Similarly, in an evaluation of statistical fragility in both the spine and orthopedic oncology literature, a FI of just two was reported.[18,27] Regarding the orthopaedic trauma literature, Parisien et al[25] analyzed 198 studies consisting of 775 total outcome events and demonstrated a FI of five and associated FQ of 0.046, representing just 3.8% of the total study population. Furthermore, the evaluation of RCTs in the pediatric orthopaedic literature were found to have a FI of only three.[28] In addition, in the fragility analysis of 72 clinical trials cited as strong evidence in the *American Academy of Orthopaedic Surgeons* Clinical Practice Guidelines, Checketts et al found a FI of only two with an associated FQ of 0.022. Furthermore, simple application of the Fisher exact test in evaluation of statistical significance nullified significance in 16 (22%) of all included studies, producing a FI of zero. This suggests that the significance of some studies may rest in the particular method of statistical analysis used, representing fragile data.[29] These previous fragility analyses have applied the concept of the FI to an orthopaedic subspecialty or to data from multiple subspecialties. Our study is unique in that we investigated the statistical fragility of a specific orthopaedic pathology, Achilles tendon injury. We support the use of both FI and FQ when presenting dichotomous results with associated *P* values, especially for studies investigating pathologies whose optimal treatment modalities are not definitively agreed on.

Strengths of this study include the large sample size and comprehensive evaluation of all comparative studies of Achilles tendon injuries over the past 20 years in 11 prominent orthopaedic surgery journals. Our inclusion of both primary and secondary outcomes, rerupture, infection/wound complication, return to sport/activity analysis, significant and nonsignificant *P* values, and non-RCTs, in addition to RCTs, represents additional study strengths as the robustness of such data often influence clinical treatment trends. Although non-RCTs may carry an increased risk of bias and confounding, the inclusion of these trials, in addition to RCTs, provides a more comprehensive evaluation of the existing Achilles tendon injury literature. Furthermore, our inclusion of FQ analysis allows for the interpretation of fragility relative to study sample size. However, utilization of FI analysis has limitations because it applies only to dichotomous data with reported *P* values and cannot be determined for study outcomes reporting continuous data, such as a visual analogue scale or a Likert scale evaluating varying degrees of agreement. Furthermore, given the relative lack of fragility analyses in the peer-reviewed comparative literature, specific fragility thresholds have yet to be determined and remain the focus of future study. In addition, LTF data are often evaluated as a factor contributing to or detracting from study strength. The 49% of studies failing to report LTF data may actually possess minimal LTF; however, we do not know because the authors of those studies failed to report that data. Our reporting of the failure of nearly half the studies to report LTF data is to highlight the lack of standardization in the way in which data are reported in the peer-reviewed Achilles literature. Furthermore, FI does not take into account the time at which events occur.[30] Nevertheless, the factor of time is less important in surgical trials where success rates of a procedure are of interest but may be more valuable in oncological trials looking at time-to-event endpoints such as increased survival time.[31]

Conclusion

The *P* value of 0.05 is a well-established cutoff indicating statistical significance, but a similar threshold does not exist for the FI and FQ. Therefore, understanding of how

the FI and FQ influences clinical decision-making requires further investigation. Our fragility analysis demonstrates that reported *P* values for Achilles tendon injuries lack statistical stability and should therefore be interpreted in the context of additional data. As such, we recommend the triple reporting of a FI, FQ, and *P* value to provide a comprehensive understanding of the robustness of statistical findings in comparative trials in the Achilles literature.

References

1. Djulbegovic B, Guyatt GH: Progress in evidence-based medicine: A quarter century on. *Lancet* 2017;390:415-423.

2. 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.

3. Keating JF, Will EM: Operative versus non-operative treatment of acute rupture of tendo achillis: A prospective randomised evaluation of functional outcome. *J Bone Joint Surg Br* 2011;93:1071-1078.

4. Willits K, Amendola A, Bryant D, et al: Operative versus nonoperative treatment of acute achilles tendon ruptures: A multicenter randomized trial using accelerated functional rehabilitation. *J Bone Joint Surg Am* 2010;92:2767-2775.

5. Stavenuiter XJR, Lubberts B, Prince RM, Johnson AH, DiGiovanni CW, Guss D: Postoperative complications following repair of acute achilles tendon rupture. *Foot Ankle Int* 2019;40:679-686.

6. Khan RJK, Fick D, Brammar TJ, Crawford J, Parker MJ: Interventions for treating acute Achilles tendon ruptures, *Cochrane Database Syst Rev*. 2004;(3):CD003674.

7. Lantto I, Heikkinen J, Flinkkila T, et al: A prospective randomized trial comparing surgical and nonsurgical treatments of acute achilles tendon ruptures. *Am J Sports Med* 2016;44:2406-2414.

8. Grassi A, Amendola A, Samuelsson K, et al: Minimally invasive versus open repair for acute Achilles tendon rupture meta-analysis showing reduced complications, with similar outcomes, after minimally invasive surgery. *J Bone Joint Surg Am Vol* 2018;100:1969-1981.

9. Karabinas PK, Benetos IS, Lampropoulou-Adamidou K, Romoudis P, Mavrogenis AF, Vlamis J: Percutaneous versus open repair of acute Achilles tendon ruptures. *Eur J Orthop Surg Traumatol* 2014;24:607-613.

10. Kołodziej L, Bohatyrewicz A, Kromuszczyńska J, Jezierski J, Biedroń M: Efficacy and complications of open and minimally invasive surgery in acute Achilles tendon rupture: A prospective randomised clinical study: Preliminary report. *Int Orthop* 2013;37:625-629.

11. Greenland S, Senn SJ, Rothman KJ, et al: Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur J Epidemiol* 2016;31:337-350.

12. Sterne JAC, Smith GD, Cox DR: Sifting the evidence—what's wrong with significance tests? *BMJ* 2001;322:226.

13. Pocock SJ: Current issues in the design and interpretation of clinical trials. *Br Med J (Clin Res Ed)* 1985;290:39-42.

14. Ioannidis JP: Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;294:218-228.

15. Matics TJ, Khan N, Jani P, Kane JM: The fragility of statistically significant findings in pediatric critical care randomized controlled trials. *Pediatr Crit Care Med* 2019;20:e258-e262.

16. Matics TJ, Khan N, Jani P, Kane JM: The fragility index in a cohort of pediatric randomized controlled trials. *J Clin Med* 2017;6:79.

17. Shen C, Shamsudeen I, Farrokhyar F, Sabri K: Fragility of results in ophthalmology randomized controlled trials: A systematic review. *Ophthalmology* 2018;125:642-648.

18. 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.

19. Noel CW, McMullen C, Yao C, et al: The fragility of statistically significant findings from randomized trials in head and neck surgery. *Laryngoscope* 2018;128:2094-2100.

20. Tignanelli CJ, Napolitano LM: The fragility index in randomized clinical trials as a means of optimizing patient care. *JAMA Surg* 2019; 154:74-79.

21. Meyer C, Heavener TE, Vassar M: The fragility index applied to liver-related trials. *Indian J Gastroenterol* 2019;38:456-459.

22. 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.

23. 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.

24. Parisien RL, Trofa DP, Dashe J, et al: Statistical fragility and the role of P values in the sports medicine literature. *J Am Acad Orthop Surg* 2019;27:e324-e329.

25. Parisien RL, Dashe J, Cronin PK, Bhandari M, Tornetta P: Statistical significance in trauma research: Too unstable to trust? *J Orthop Trauma* 2019;33:e466-e470.

26. Ahmed W, Fowler RA, McCredie VA: Does sample size matter when interpreting the fragility index? *Crit Care Med* 2016;44: e1142-e1143.

27. Forrester LA, Jang E, Lawson MM, Capi A, Tyler WK: Statistical fragility of surgical and procedural clinical trials in orthopaedic oncology. *J Am Acad Orthop Surg Glob Res Rev* 2020;4:e19.00152.

28. Khormaee S, Choe J, Ruzbarsky JJ, et al: The fragility of statistically significant results in pediatric orthopaedic randomized controlled trials as quantified by the fragility index: A systematic review. *J Pediatr Orthop* 2018;38:e418-e423.

29. Checketts JX, Scott JT, Meyer C, Horn J, Jones J, Vassar M: The robustness of trials that guide evidence-based orthopaedic surgery. *J Bone Joint Surg Am* 2018;100:E85.

30. 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.

31. Bomze D, Meirson T: A critique of the fragility index. *Lancet Oncol* 2019;20:e551.