Article

AMERICAN ORTHOPAEDIC FOOT & ANKLE SOCIETY

Foot & Ankle Orthopaedics
2024, Vol. 9(4) 1–11

DOI: 10.1177/24730114241300160
journals.sagepub.com/home/fao

The Fragility of Statistically Significant Binary Outcomes for Treating Achilles Tendinopathy: A Systematic Review of Randomized Trials

Omkar S. Anaspure, BA[1], Shiv Patel, BA[1], Anthony N. Baumann, DPT[2,3], Andrew Newsom, BS[2], Albert T. Anastasio, MD[4], and Annunziato Amendola, MD[4]

Abstract

Background: Randomized controlled trials (RCTs) are the gold standard for treatment efficacy, but foot and ankle RCTs are often small or inconsistent. The Fragility Index (FI) evaluates the stability of significant findings. This study assessed the fragility of RCT outcomes for Achilles tendon pathology (ATP) interventions.

Methods: This systematic review queried PubMed up to May 14, 2024, for RCTs on ATP interventions. RCTs with significant binary outcomes were included. Two reviewers assessed eligibility, extracted data, calculated FIs, and evaluated risk of bias. Frequency-weighted means were used for narrative synthesis.

Results: Eleven RCTs with 4506 patients (mean cohort size: 409.64 ± 160.54) and a mean age of 36.97 ± 13.51 years (n = 4356; 96.67%) were included, covering 24 binary outcomes. The median FI across all outcomes was 3 (interquartile range 1-4; mean 3.92), indicating that changing the outcome of just a few patients could shift a study's results from statistically significant to nonsignificant. Trials having an FI ≤ 3 comprised 58.33%. Three outcomes (12.5%) had an FI of zero after recalculating *P* values using the two-sided Fisher exact test. Half of the outcomes were robust. No RCT reported FIs or adjusted significance for multiple testing. Most studies (81.82%) performed 2 or more statistical tests, with an average of 30.81 ± 41.28 *P* values reported per study. The overall risk of bias was low in 1 study (9.09%) and moderate in 7 (63.64%). Most studies had low risk of bias in randomization (72.73%) and missing outcome data (90.91%).

Conclusion: The FI assesses the fragility of statistically significant binary results, revealing that many ATP RCTs have fragile outcomes due to small sample sizes. A median FI of 3 means that changing the outcome of 3 patients could shift a study's results from statistically significant to nonsignificant.

Keywords: Achilles, randomized controlled trials, outcomes, statistical significance, Fragility Index

Introduction

The most reliable treatment evaluations and causal determinations come from well-powered randomized controlled trials (RCTs), yet orthopaedic surgery RCTs often yield inconsistent results.[2,4,9,16,17,26,28] Analysis of these RCTs has shown that the *P* value and effect size have largely been utilized as the primary forms of comparing the outcomes from different treatment arms.[28,44] However, relying solely on these 2 metrics can be misleading, as *P* values are often overemphasized and should be used alongside other tools

[1]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]College of Medicine, Northeast Ohio Medical University, Rootstown, OH, USA
[3]Department of Rehabilitation Services, University Hospitals, Cleveland, OH, USA
[4]Department Orthopedic Surgery, Duke University, Durham, NC, USA

Corresponding Author:
Omkar S. Anaspure, BA, Perelman School of Medicine, University of Pennsylvania, 3400 Civic Center Blvd, Philadelphia, PA 19140, USA.
Email: Omkar.Anaspure@pennmedicine.upenn.edu

for interpreting results.[7,8,44] In foot and ankle surgery, RCTs often have smaller sample sizes compared with other orthopaedic conditions. This raises concerns about the validity of findings, as altering the outcomes of just a few patients in a treatment arm could significantly impact or even reverse the trial's conclusions by nullifying the significance.[3,33,35,44]

The Fragility Index (FI) is a metric that aims to assess the robustness of statistically significant results to quantify such phenomenon. The FI is designed to be used in conjunction with *P* values to aid in a more comprehensive interpretation of RCTs.[7,14,51] The FI of a study is defined as the smallest number of patients in the trial group with fewer outcome events whose status must change from a "nonevent" to an "event" to alter a statistically significant result to a nonsignificant one.[7,43] A small FI indicates statistical fragility, relying on few events for significance, whereas a large FI raises confidence in treatment impact.[10]

Given the small sample sizes and few events in foot and ankle surgery trials, our objective was to assess the robustness of significant RCT results in Achilles tendon pathology (ATP). Achilles tendon ruptures, the most common in the lower extremity, occur at an annual rate of up to 40 per 100 000.[13,20,27] These injuries, including tendinitis, are often seen in athletes and overuse cases.[29,52] Treatments range from nonsurgical options (cast, boot, brace) to surgical procedures (reattachment, tendon transfer).[29,36,52,53] Given the prevalence of ATP, high-quality evidence is crucial for comparing surgical and conservative management.

Recently, a review by Fackler et al[10] sought to examine the statistical stability of studies comparing operative vs nonoperative management for Achilles tendon rupture. However, this review was limited to Achilles tendon ruptures and only included a search of the top 10 orthopaedic journals, limiting its impact on the broader ATP literature. Additionally, it included cohort studies, not just RCTs. In contrast, we defined ATP as a broad range of Achilles tendon conditions, including both ruptures and tendinopathy (insertional and noninsertional) to provide a comprehensive assessment of treatment outcomes and fragility, avoiding the limited scope of previous studies that focused solely on specific pathologies like ruptures. This study expands on the work of Fackler et al by examining the fragility of significant findings from all RCTs on ATP interventions, applying the FI, and assessing statistical corrections. Understanding this fragility is crucial for clinicians, as fragile findings may not be robust enough to guide ATP management confidently.

Methods

Study Creation and Initial Search

This study is a systematic review of the literature examining the fragility of significant binary outcomes of RCTs. All RCTs regarding ATP were searched in PubMed from database inception until May 14, 2024. Search terms used in each database were ("Achilles Tendon"[Mesh] OR Achilles OR "Achilles tendon" OR "calcaneal tendon") AND ("Randomized Controlled Trial"[Publication Type] OR "randomized controlled trial"). This study was performed under the guidelines of the most recent Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRIMSA) for proper data reporting. The study registration can be found within the Open Science Framework registry at osf.io/96qja.

Study Definitions

ATP was defined broadly in this study to include a range of conditions affecting the Achilles tendon, such as tendinopathy, partial and complete ruptures, and other related disorders. This inclusive definition was chosen to ensure a comprehensive search and assessment of RCTs in foot and ankle surgery. By encompassing both acute and chronic conditions, our goal was to capture the full spectrum of interventions and their respective outcomes, offering a more complete evaluation of statistical robustness in the literature.

In FI terms, "robustness" means stable, reliable RCT results, whereas "frailty" indicates vulnerability and low reliability. A high FI signals robustness; a low FI signals frailty. A study was "robust" if FI exceeded dropouts, or "fragile" if FI was less than dropouts.

Inclusion and Exclusion Criteria

Inclusion criteria were RCT that examined patients who sustained any ATP and reported at least 1 significant binary outcome (as defined by the individual study) comparing either treatment groups or comparing pre- and posttreatment change. Exclusion criteria were nonrandomized controlled studies, studies without ATP, and studies without statistically significant binary outcomes.

Article Screening Process

After the search algorithm was executed in each of the 4 databases for the initial search, all articles were uploaded into Rayyan, a public website used for systematic reviews.[39] One individual screener performed a manual deduplication of articles. Two independent reviewers performed article screening based on title and abstract, followed by full-text screening based on inclusion and exclusion criteria. Lastly, the references of each included article were manually searched for articles not initially captured. Any conflicts during the article screening process were resolved by the first author.

Data Extraction

Two authors extracted data on all significant binary outcomes, including journal name, publication year, sample size, follow-up losses, events per arm, *P* values, correction use, FI reporting, and relevant significant outcomes.

Article Risk Assessment

Risk of bias was assessed using the Cochrane Risk of Bias for Randomized Trials ROB-2 tool, which examines bias under the following categories: randomization process, deviations from intended intervention, missing data, measurement of the outcome, selection of the reported result.[10,46] Each article is assessed and assigned a score of low risk, some concerns, or high risk of bias for each domain.[10,46]

Statistical Analysis

This study used the Statistical Package for the Social Sciences (SPSS) version 29.0 (IBM Corp, Armonk, NY) for statistical analysis. Frequency-weighted means and other descriptive statistics were used to describe the data where no statistical significance could be calculated. We calculated the FI for each outcome using the Fragility Index Calculator by ClinCalc statistics.[21] The FI is a recognized and validated metric that quantifies the robustness of statistically significant results by determining how many event-to-nonevent outcome changes are required to shift the *P* value above the significance threshold.[7,21] The ClinCalc Fragility Index Calculator automates event-to-nonevent switching and recalculates the 2-sided Fisher exact test until the *P* value exceeds .05, determining the FI. FIs were calculated for reported significant binary outcomes. Additionally, raw binary outcomes without significance tests were analyzed with Fisher exact test to identify unreported significant outcomes.

Results

Initial Search Results

Our database query yielded 531 potential studies. After title and abstract screening, 39 articles were retrieved for full-text analysis. Thirty articles were excluded as they did not find any binary outcomes as significant. Only 9 RCTs reported at least 1 significant binary outcome and were ultimately included. An additional 2 articles were included by citation search, for a total of 11 articles further pursued for data extraction (Figure 1).[5,11,22,25,30,31,34,38,40,45,54]

Characteristics of Trials and Outcomes

Eleven studies were included that reported at least 1 statistically significant binary variable. A total of 24 significant binary outcomes were reported across the 11 studies, and FIs were calculated for each outcome. We found that 3 studies only reported conversion of both groups to a binary endpoint without performing any statistical testing. On calculating the 2-sided Fisher exact test for the 4 binary outcomes found from the 3 studies, all 4 outcomes were found to be significant (Table 1, asterisked outcomes). There was a total of 4506 patients treated among all 11 studies, and the frequency-weighted mean age was 36.97 ± 13.51 (n = 4356 patients, 96.67%). The mean sample size of the included trials was 409.64 ± 160.54 and the mean losses to follow-up was 6.18 ± 3.25 patients (ie, 1.51% of the patients were lost to follow-up across trials). Among the included trials, overall risk of bias was low in 1 study (9.09%) and moderate in 7 studies (63.64%). Eight studies (72.73%) had low risk of bias in the randomization process and 10 studies low risk of missing outcome data (90.91%) (Figure 2). Trends emerged, where clinically important outcomes, such as rerupture rates and tendon healing, demonstrated greater robustness, with FI values of 4 or higher. Conversely, patient-reported outcomes related to satisfaction and less critical endpoints, such as mild discomfort, exhibited greater fragility with FI values often at or below 1.

Fragility Index

The median FI across all outcomes for the 24 evaluated outcomes was 3 events (interquartile range [IQR] 1-4, mean 3.92) which means that adding 3 events to one of the trial's treatment arms eliminated would eliminate its statistical significance. Three outcomes (12.50%) had an FI of zero because they lost their statistical significance when the FI calculator recalculated their *P* values using the 2-sided Fisher exact test.[21] In total, 12 outcomes were found to be robust and 12 to be fragile. Of the 12 robust outcomes, 1 outcome was calculated from a study that did not provide statistical analysis. No RCT reported FIs as part of their own statistical analysis, and none adjusted the significance (eg, Bonferroni correction) to reduce the risk of type I errors. The mean total *P* values reported by each study was 30.81 ± 41.28 (range 1, 136). Overall, 81.82% of studies (n = 9) performed 2 or more significance tests as part of their analysis. Table 2 depicts the FI values according to subgroups based on outcome type, sample size for each arm, number of events, and losses to follow-up. Figure 3 depicts the distribution of FIs across the study.

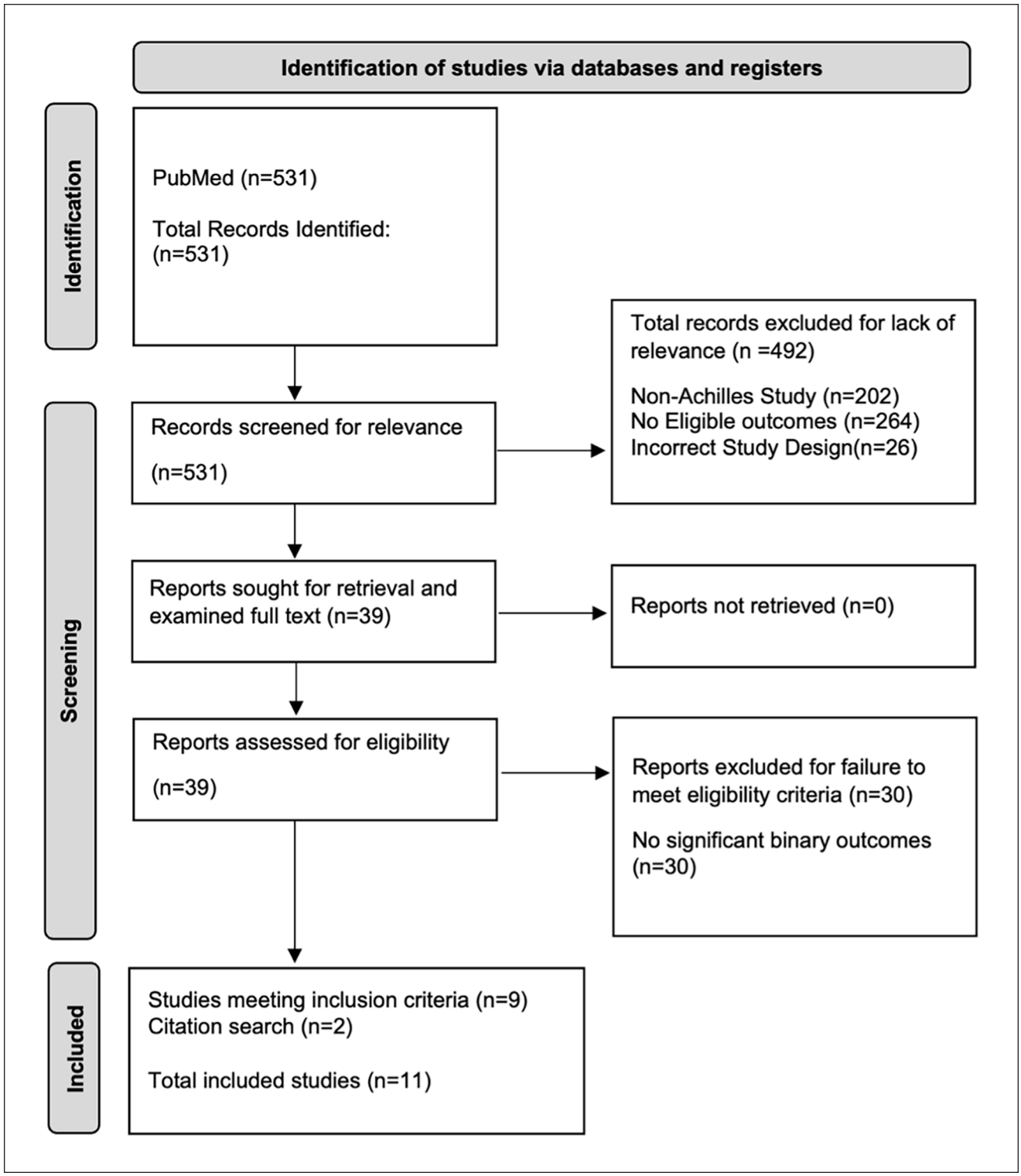

Figure 1. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram outlining the entire search progress, from initial search in 4 databases to final article inclusion.

Table 1. Study Demographics Table With Relevant Study Characteristics Such as Intervention Type, Procedure Type, Mean Patient Age (Unless Otherwise Reported as Median) and Lost-to-Follow-up.

Author	Year	Intervention	Outcome Category	Outcome	Groups	Age	Sex	Sample Size	Events	Loss to Follow-up	FI	Fragile/ Robust
Metz et al[30]	2008	Surgical	Complication	Skin-related complications*	Surgical	40 (median)	Male (n = 31) Female (n = 11)	42	2	0	4	Robust
					Nonoperative	41 (median)	Male (n = 35) Female (n = 6)	41	13			
Milgrom et al[31]	2003	Nonsurgical	Complication	Achilles tendinopathy	Summer training	18.7 ± 7	Male (n = 803) Female (n = 0)	803	29	0	25	Robust
					Winter training		Male (n = 697) Female (n = 0)	697	66			
Young et al[54]	2014	Nonsurgical	Questionnaire response/ patient-reported	No stiffness or minimal stiffness at 1 y	Weightbearing cast	–	–	32	26	17	4	Fragile
					Nonweightbearing cast			35	17			
Paoloni et al[40]	2004	Nonsurgical	Questionnaire response/ patient-reported	Asymptomatic with ADL at 24 wk	Topical glyceryl nitrate + rehabilitation	49	Male (n = 40) Female (n = 25)	36 tendons	28	7	3	Fragile
					Only rehabilitation			41 tendons	20			
Mafi et al[25]	2001	Nonsurgical	Questionnaire response/ patient-reported	Satisfaction after 12 wk of training	Eccentric training	48.1 ± 9.5	Male (n = 24) Female (n = 20)	22	18	0	3	Robust
					Concentric training	48.4 ± 8.3		22	8			
Keene et al[22]	2019	Nonsurgical	Complication	Mild discomfort or minor bleeding following injection*	Platelet-rich plasma	45.90 (13.74)	Male (n = 88) Female (n = 25)	113	22	1	4	Robust
					Placebo	45.16 (12.43)	Male (n = 84) Female (n = 32)	116	8			
Fischer et al[11]	2021	Surgical	Radiologic outcome	Normal tendon on ultrasonograph after 24 mo*	Open surgery	39.6 ± 7.3	Male (n = 81) Female (n = 9)	23	4	13	1	Fragile
					Minimally invasive surgery	39.3 ± 7.9	24	0				
				Generalized, diffuse inhomogeneity on tendon ultrasonograph after 24 mo*	Open surgery	39.6 ± 7.3	Male (n = 81) Female (n = 9)	23	0	15	1	Fragile
					Conservative treatment	45.2 ± 9.5	22	4				
Costa et al[5]	2020	Nonsurgical	Medications prescribed	Anticoagulant prescribed for VTE prophylaxis at 8 wk	Plaster cast	49.0 (13.9)	Male (n = 213) Female (n = 51)	266	191	0	13	Robust
					Functional splint	48.3 (13.8)	Male (n = 213) Female (n = 61)	274	162			
				Analgesics prescribed between 8 wk and 3 mo	Plaster cast	49.0 (13.9)	Male (n = 213) Female (n = 51)	266	29	0	3	Robust
					Functional splint	48.3 (13.8)	Male (n = 213) Female (n = 61)	274	14			
				Analgesics prescribed between 8 wk and 3 mo	Plaster cast	49.0 (13.9)	Male (n = 213) Female (n = 51)	266	5	0	1	Robust
					Functional splint	48.3 (13.8)	Male (n = 213) Female (n = 61)	274	0			
Möller et al[34]	2001	Surgical	Complication	Rerupture	Surgical	39.6	Male (n = 51) Female (n = 8)	59	1	0	4	Robust
					Nonsurgical	38.5	Male (n = 48) Female (n = 5)	53	11			

(continued)

Table 1. (continued)

Author	Year	Intervention	Outcome Category	Outcome	Groups	Age	Sex	Sample Size	Events	Loss to Follow-up	FI	Fragile/ Robust
Olsson et al[38]	2013	Surgical	Complication	Bracing complications	Surgical	39.8 ± 8.9	Male (n = 39) Female (n = 10)	44	13	6	5	Fragile
					Nonsurgical	39.5 ± 9.7	Male (n = 47) Female (n = 4)	50	2			
				Superficial infection	Surgical	39.8 ± 8.9	Male (n = 39) Female (n = 10)	44	6	6	2	Fragile
					Nonsurgical	39.5 ± 9.7	Male (n = 47) Female (n = 4)	50	0			
Silbernagel et al[45]	2001	Nonsurgical	Questionnaire response/ patient-reported	Satisfied with physical activity level at 1 year	Experimental	47 ± 14.7	Male (n = 17) Female (n = 5)	20	14	4	0	Fragile
					Control	41 ± 10.2	Male (n = 14) Female (n = 4)	16	6			
				Consider themselves fully recovered at 1 y	Experimental	47 ± 14.7	Male (n = 17) Female (n = 5)	20	12	4	1	Fragile
					Control	41 ± 10.2	Male (n = 14) Female (n = 4)	16	4			
				No pain during physical activity at 1 y	Experimental	47 ± 14.7	Male (n = 17) Female (n = 5)	19	14	6	0	Fragile
					Control	41 ± 10.2	Male (n = 14) Female (n = 4)	15	8			
				No pain after physical activity at 1 y	Experimental	47 ± 14.7	Male (n = 17) Female (n = 5)	19	11	6	0	Fragile
					Control	41 ± 10.2	Male (n = 14) Female (n = 4)	15	5			
			Symptoms	Pain walking stairs at 6 wk	Experimental	47 ± 14.7	Male (n = 17) Female (n = 5)	27 tendons	3	4	1	Fragile
					Control	41 ± 10.2	Male (n = 14) Female (n = 4)	26 tendons	9			
				Pain during activity at 6 wk in the experimental group	Pretreatment	47 ± 14.7	–	30 tendons	23	0	1	Fragile
					Posttreatment	47 ± 14.7	–	27 tendons	13			
				Pain during activity at 6 mo in the experimental group	Pretreatment	47 ± 14.7	–	30 tendons	23	0	4	Robust
					Posttreatment	47 ± 14.7	–	27 tendons	10			
				Pain during activity at 6 mo in the control group	Pretreatment	41 ± 10.2	–	27 tendons	15	3	4	Robust
					Posttreatment	41 ± 10.2	–	27 tendons	3			
				Asymptomatic at 6 mo in the experimental group	Pretreatment	47 ± 14.7	–	30 tendons	14	0	8	Robust
					Posttreatment	47 ± 14.7	–	27 tendons	25			
				Morning stiffness at 6 mo in the experimental group	Pretreatment	47 ± 14.7	–	30	25	0	2	Robust
					Posttreatment	47 ± 14.7	–	27	14			

Abbreviations: ADL, activities of daily living; VTE, venous thromboembolism.
*Statistically significant ($P < .05$).

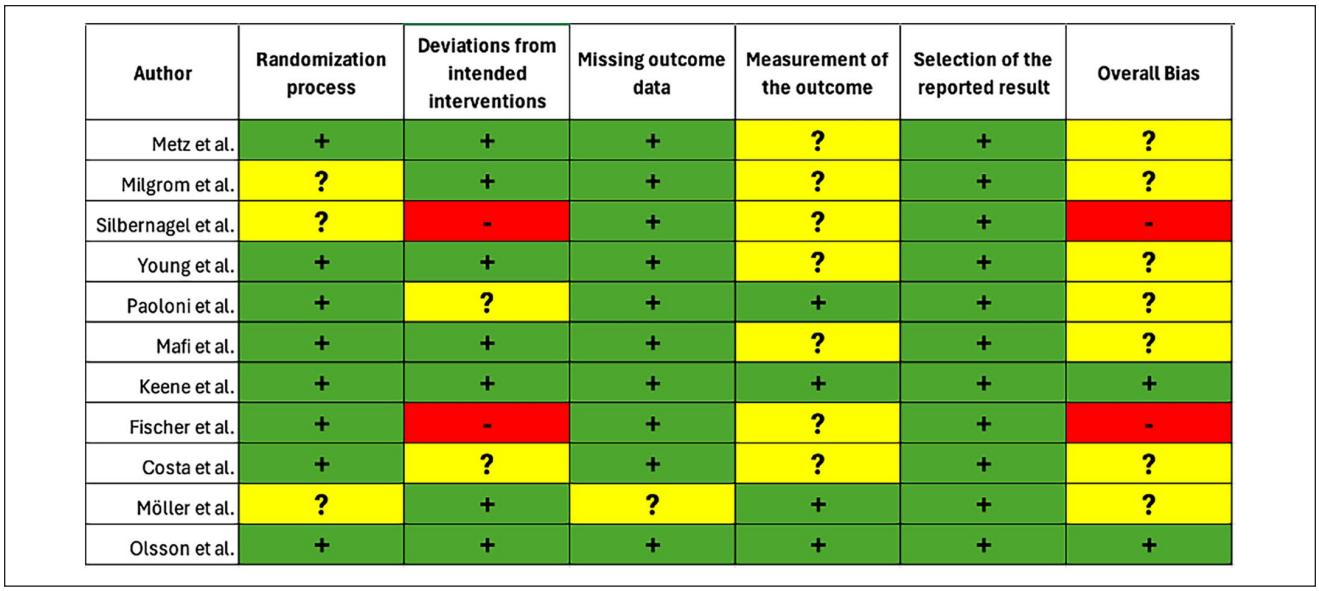

Figure 2. Outcomes of the Cochrane Risk of Bias 2.0 tool for randomized controlled trials; n = 11. The plus sign marks a low risk of bias, and the question mark indicates that there is some concern for bias.

Table 2. Fragility Indexes by Outcome Category.

Characteristic	Fragility Index
All outcomes (n = 24)	Median (IQR): 3 (1-4)
Questionnaire response/patient-reported (n = 7)	Median (IQR): 1 (0-3)
Satisfied with physical activity level at 1 y	0
Consider themselves fully recovered at 1 y	1
No pain during physical activity at 1 y	0
No pain after physical activity at 1 y	0
No stiffness or minimal stiffness at 1 y	4
Asymptomatic with ADL at 24 wk	3
Satisfaction after 12 wk of training	3
Complication (n = 6)	Median (IQR): 4 (4-4.75)
Skin-related complications	4
Achilles tendinopathy	25
Mild discomfort or minor bleeding following injection	4
Rerupture	4
Bracing complications	5
Superficial infection	2
Symptom (n = 6)	Median (IQR): 3 (1.25-4)
Pain walking stairs at 6 wk	1
Pain during activity at 6 wk in the experimental group	1
Pain during activity at 6 mo in the experimental group	4
Pain during activity at 6 mo in the control group	4
Asymptomatic at 6 mo in the experimental group	8
Morning stiffness at 6 mo in the experimental group	2
Medications prescribed (n = 3)	Median (IQR): 3 (2-8)
Anticoagulant prescribed for VTE prophylaxis at 8 wk	13
Analgesics prescribed between 8 wk and 3 mo	3
Analgesics prescribed between 8 wk and 3 mo	1
Radiologic outcome (n = 2)	Median (IQR): 1 (1-1)
Normal tendon on ultrasonograph after 24 mo	1
Generalized, diffuse inhomogeneity on tendon ultrasonograph after 24 mo	1

Abbreviations: ADL, activities of daily living; IQR, interquartile range; VTE, venous thromboembolism.

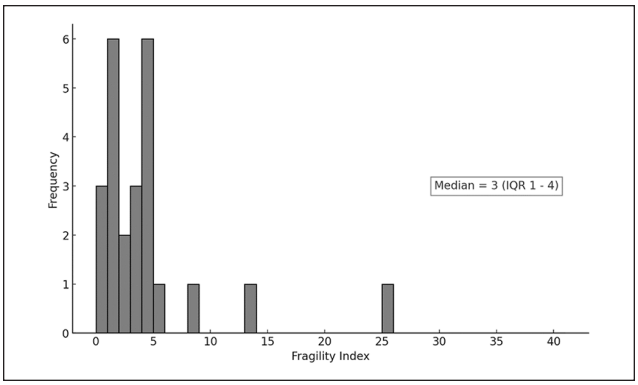

Figure 3. Frequency distribution of FI values from 11 trials showing 24 outcomes. The median number of patients whose status would have to change from a nonevent to an event to change a statistically significant result to a nonsignificant result was 3 (IQR 1-4). Overall, 50% of the FIs were deemed fragile and 50% were found to be robust. FI, Fragility Index; IQR, interquartile range.

Discussion

In this systematic review, we evaluated the fragility of statistically significant results in RCTs on ATP in foot and ankle surgery. By applying the FI, we assessed the stability of findings across various interventions for ATP. Among the 11 RCTs reviewed, as few as 3 outcome events could reverse the statistical significance of the treatment arm. We can encourage the incorporation of the FI into foot and ankle literature, as it could be highly beneficial in improving the stability of research findings, potentially influencing clinical practice.

This study expands on the findings of recent fragility analyses conducted by Parisien et al[41] and Fackler et al.[10] The initial review by Parisien et al[41] focused on comparative studies of Achilles tendon injuries and revealed that the outcomes were less statistically stable than previously thought, warranting cautious interpretation. Fackler et al's follow-up study on Achilles tendon ruptures also raised concerns about outcome stability. Both reviews included cohort and RCT studies, potentially confounding results and limiting clarity. Their search was restricted to the top 10 orthopaedic journals, further narrowing conclusions. In contrast, our study focused solely on RCTs for Achilles tendinopathy, without limiting the search to specific journals, providing a more comprehensive analysis.

The FI can be clinically relevant in ATP as it highlights the reliability of RCT outcomes. A low fragility indicates that results are unstable and easily reversed by a few additional events, suggesting the findings may not be robust enough for confident clinical decisions. In foot and ankle care, particularly with ATP, the FI exposes the vulnerability of conclusions from small or underpowered studies. For instance, a low fragility might suggest the effectiveness of a treatment, but minor changes in outcomes could negate its significance.[10] This may imply that clinicians should be cautious in interpreting these results and might need to consider additional factors or seek further evidence before altering their practice.[47,48] Variability in Achilles tendinopathy treatments increases outcome fragility, which a low fragility can highlight. Using FI with *P* values and CIs helps identify robust treatments, leading to better-informed decisions and more stable outcomes.[10,48] The Bonferroni correction in Achilles tendinopathy studies reduces type I errors but may increase type II errors, sparking debate as it can overly penalize studies with multiple hypotheses, limiting true effect detection.[1,42] Critics argue that although it controls for false positives, it may result in the dismissal of genuinely significant findings, suggesting that a balance is necessary when interpreting results from Achilles tendinopathy research.[1,42,49]

Overall, our findings align with those of Fackler et al and Parisien et al. The median FI for all reported outcomes (n = 24) in our study was 3 events, comparable to the 4 events reported by Fackler et al[10] and the average of 2.9 events reported by Parisien et al.[41] Categories like postoperative complications and prescribed medications showed the greatest variation in FI, although all categories had median values within a narrow range of 1 to 4. We also identified a mean sample size of 409 patients and a median of 30 events per outcome, smaller than the median sample size of 682 patients and 112 events per outcome reported by Walsh et al[51] in their analysis of 399 RCTs from high-impact medical journals. Our mean FI of 3 events was lower than the median FI of 8 events (range 3-18) reported by Walsh et al. Only 50% of reported outcomes in Achilles tendinopathy trials had robust dichotomous outcomes, raising concerns about the validity of outcomes in up to half of the ATP RCTs. These findings suggest that ATP RCTs, compared to those in other specialties, have smaller sample sizes, higher statistical frailty, and overall poorer quality.

Our study supports using the FI in evaluating ATP management. Although some critics view the FI as a "*P* value in disguise," others argue that RCTs with a priori power analysis are inherently fragile.[6] Evaluating the robustness or fragility of RCTs necessitates assessing uncertainty rather than solely focusing on statistical significance of dichotomous outcomes.[6] However, the growing body of literature in orthopaedic subspecialties that use the FI to evaluate the validity of dichotomous outcomes cannot be ignored.[9,10,12,15,17,28,32,41,44,50] Previous systematic reviews within adult reconstruction, shoulder, spine, foot/ankle, and hand surgery have all found median FIs ranging from 2 to 4, with a shoulder arthroplasty study reporting the highest FI within the orthopaedic literature (FI = 6).[9,10,15,28,41,44] The orthopaedic literature overall shows much lower FIs (FI = 2) and tends to have smaller cohort sizes when compared to high-impact medical

journals (FI = 8), with otolaryngology coming in at the second lowest (FI = 3).[7,9,37,51] Without threshold cutoffs, FIs must be contextualized within similar studies, so larger FIs cannot be evaluated in isolation. Given smaller sample sizes in orthopaedics, we redefined "fragile" and "robust" by comparing the FI with the dropout rate of the study group for better context as suggested by the literature.[18,19] We believe incorporating FI with *P* values would demonstrate a more comprehensive view of outcomes, leading to improved patient care. Alternatively, a focus on CIs would provide an alternative to relying solely on *P* values and their clear limitations. They serve as a valuable tool for assessing the precision of results, evaluating data compatibility with multiple hypotheses, and gaining deeper insights. CIs present a range of values consistent with the data, with the width indicating result precision and the spectrum of potential true outcomes.

It is worth noting that many of the fragile outcomes identified in this review were secondary, rather than primary, outcome measures. This is likely reflective of the broader state of ATP literature, where secondary measures are frequently used but may not receive the same level of rigorous validation as primary outcomes, nor are the sample size powered for the same level of confidence. The diminished robustness of these secondary measures underscores the need for further refinement in the design and reporting of RCTs in this field, particularly when it comes to defining and validating clinically meaningful primary endpoints. Although the FI highlights statistical vulnerability in many outcomes, its clinical applicability varies depending on the importance of the outcome itself. For example, outcomes like pain at 6 weeks, which are often fragile, may be less critical for guiding long-term treatment decisions. These fragile results, although valuable for patient comfort, may not necessarily indicate long-term treatment efficacy. In contrast, more robust outcomes, such as rerupture rates and tendon healing, have greater clinical relevance and should be prioritized in decision making. By distinguishing between these fragile and robust outcomes, clinicians can apply the FI more effectively, using it to focus on the most reliable endpoints when making treatment decisions for ATP.

The limitations of this study must be considered. Three outcomes had an FI of zero because their statistical significance was lost when recalculating *P* values using the 2-sided Fisher exact test. The Fisher exact test, a more conservative alternative to the Pearson χ^2 test for comparing proportions in a 2×2 contingency table, was used.[9,23,24] The Fisher exact test is suitable for all sample sizes and is the preferred method when sample sizes are small, or outcome events are uncommon.[7,9] However, because the FI is calculated using Fisher exact test, results may differ from those obtained with methods like the χ^2 test. The χ^2 test relies on an approximation suitable for large samples, whereas the Fisher exact test is precise, particularly for small samples.[7] Replacing Fisher exact test with another statistical method in small trials can result in a nonsignificant *P* value and an FI of 0, highlighting study fragility. Although only 12.5% of values were 0, this underscores the importance of consistent testing. Our reliance on a single database might miss relevant studies, potentially underestimating associations between FI and RCT outcomes, but our findings are consistent with existing literature.[7,9,10,15,17,28,41,44,50] FI is only applicable to dichotomous outcomes, limiting its use in analyzing continuous or time-to-event outcomes in RCTs. This underscores the need to use FI alongside other statistical methods and evaluate continuous outcomes separately. Our analysis included both primary and secondary outcome measures, which may vary in their clinical significance. Primary outcomes, such as rerupture rates, likely will carry greater weight, whereas secondary outcomes, such as mild discomfort, may contribute less to clinical decision making.

We also recognize that the inclusion of certain studies, such as Silbernagel et al,[45] may contribute more than other smaller studies to the overall number of FI calculations. This may impact the interpretation of the results, as a portion of the fragile outcomes in this review stem from secondary measures within these studies. Future analyses could benefit from a more detailed categorization of outcomes to assess whether primary measures consistently show greater or lesser fragility than secondary measures.

Conclusion

We found that studies on ATP management generally had low FI scores, with half of the outcomes still classified as fragile after adjusting for patient dropout rates. Outcomes from low-risk bias studies had FIs similar to those with some bias concerns.

Like the *P* value, the FI has limitations, and clinicians should be cautious when interpreting trials with low FI or *P* values for patient care. However, using the FI alongside other metrics can improve the evaluation of ATP trials by identifying studies with more robust outcomes.

Ethical Approval

Ethical approval was not sought for the present study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. Disclosure forms for all authors are available online.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Omkar S. Anaspure, BA, https://orcid.org/0000-0001-9135-0484
Anthony N. Baumann, DPT, https://orcid.org/0000-0002-4175-3135
Albert T. Anastasio, MD, https://orcid.org/0000-0001-5817-3826

References

1. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt*. 2014;34(5):502-508. doi:10.1111/opo.12131
2. Backmann M. What's in a gold standard? In defence of randomised controlled trials. *Med Health Care Philos*. 2017;20(4):513-523. doi:10.1007/s11019-017-9773-2
3. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*. 2010;303(12):1180-1187. doi:10.1001/jama.2010.310
4. Bhandari M, Guyatt GH, Swiontkowski MF. User's guide to the orthopaedic literature: how to use an article about a surgical therapy. *J Bone Joint Surg Am*. 2001;83(6):916-926. doi:10.2106/00004623-200106000-00015
5. Costa ML, Achten J, Wagland S, et al. Plaster cast versus functional bracing for Achilles tendon rupture: the UKSTAR RCT. *Health Technol Assess*. 2020;24(8):1-86. doi:10.3310/hta24080
6. Cote MP, Asnis P, Hutchinson ID, Berkson E. Editorial commentary: the statistical fragility index of medical trials is low by design: critical evaluation of confidence intervals is required. *Arthroscopy*. 2024;40(3):1006-1008. doi:10.1016/j.arthro.2023.10.010
7. Dettori JR, Norvell DC. How fragile are the results of a trial? The fragility index. *Global Spine J*. 2020;10(7):940-942. doi:10.1177/2192568220941684
8. Dettori JR, Norvell DC, Chapman JR. P-value worship: is the idol significant? *Global Spine J*. 2019;9(3):357-359. doi:10.1177/2192568219838538
9. 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(10):2188-2197. doi:10.1016/j.spinee.2015.06.004
10. Fackler NP, Karasavvidis T, Ehlers CB, et al. The statistical fragility of operative vs nonoperative management for Achilles tendon rupture: a systematic review of comparative studies. *Foot Ankle Int*. 2022;43(10):1331-1339. doi:10.1177/10711007221108078
11. Fischer S, Colcuc C, Gramlich Y, et al. Prospective randomized clinical trial of open operative, minimally invasive and conservative treatments of acute Achilles tendon tear. *Arch Orthop Trauma Surg*. 2021;141(5):751-760. doi:10.1007/s00402-020-03461-z
12. Forrester LA, McCormick KL, Bonsignore-Opp L, et al. Statistical fragility of surgical clinical trials in orthopaedic trauma. *J Am Acad Orthop Surg Glob Res Rev*. 2021;5(11):e20.00197. doi:10.5435/JAAOSGlobal-D-20-00197
13. Ganestam A, Kallemose T, Troelsen A, Barfod KW. Increasing incidence of acute Achilles tendon rupture and a noticeable decline in surgical treatment from 1994 to 2013. A nationwide registry study of 33,160 patients. *Knee Surg Sports Traumatol Arthrosc*. 2016;24(12):3730-3737. doi:10.1007/s00167-015-3544-5
14. Garcia MVF, Ferreira JC, Caruso P. Fragility index and fragility quotient in randomized clinical trials. *J Bras Pneumol*. 2023;49(1):e20230034. doi:10.36416/1806-3756/e20230034
15. Go CC, Maldonado DR, Go BC, et al. The fragility index of total hip arthroplasty randomized control trials: a systematic review. *J Am Acad Orthop Surg*. 2022;30(9):e741-e750. doi:10.5435/JAAOS-D-21-00489
16. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: study design: randomised controlled trials. *BJOG*. 2018;125(13):1716. doi:10.1111/1471-0528.15199
17. Herndon CL, McCormick KL, Gazgalis A, Bixby EC, Levitsky MM, Neuwirth AL. Fragility index as a measure of randomized clinical trial quality in adult reconstruction: a systematic review. *Arthroplast Today*. 2021;11:239-251. doi:10.1016/j.artd.2021.08.018
18. Heston TF. Statistical significance versus clinical relevance: a head-to-head comparison of the fragility index and relative risk index. *Cureus*. 2023;15(10):e47741. doi:10.7759/cureus.47741
19. Ho AK. The fragility index for assessing the robustness of the statistically significant results of experimental clinical studies. *J Gen Intern Med*. 2022;37(1):206-211. doi:10.1007/s11606-021-06999-9
20. Huttunen TT, Kannus P, Rolf C, Fellander-Tsai L, Mattila VM. Acute Achilles tendon ruptures: incidence of injury and surgery in Sweden between 2001 and 2012. *Am J Sports Med*. 2014;42(10):2419-2423. doi:10.1177/0363546514540599
21. Kane SP. Fragility index calculator. 2018. Accessed 30 May 2024. https://clincalc.com/Stats/FragilityIndex.aspx
22. Keene DJ, Alsousou J, Harrison P, et al. Platelet rich plasma injection for acute Achilles tendon rupture: PATH-2 randomised, placebo controlled, superiority trial. *BMJ*. 2019;367:l6132. doi:10.1136/bmj.l6132
23. Kuhn JE, Greenfield ML, Wojtys EM. A statistics primer. Statistical tests for discrete data. *Am J Sports Med*. 1997;25(4):585-586. doi:10.1177/036354659702500425
24. Lydersen S, Pradhan V, Senchaudhuri P, Laake P. Choice of test for association in small sample unordered r x c tables. *Stat Med*. 2007;26(23):4328-4343. doi:10.1002/sim.2839
25. Mafi N, Lorentzon R, Alfredson H. Superior short-term results with eccentric calf muscle training compared to concentric training in a randomized prospective multicenter study on patients with chronic Achilles tendinosis. *Knee Surg Sports Traumatol Arthrosc*. 2001;9(1):42-47. doi:10.1007/s001670000148
26. Matar HE, Platt SR. Overview of randomised controlled trials in orthopaedic research: search for significant findings. *Eur J Orthop Surg Traumatol*. 2019;29(6):1163-1168. doi:10.1007/s00590-019-02436-0
27. Mattila VM, Huttunen TT, Haapasalo H, Sillanpaa P, Malmivaara A, Pihlajamaki H. Declining incidence of surgery for Achilles tendon rupture follows publication of major RCTs: evidence-influenced change evident using the Finnish registry study. *Br J Sports Med*. 2015;49(16):1084-1086. doi:10.1136/bjsports-2013-092756

28. McCormick KL, Tedesco LJ, Swindell HW, Forrester LA, Jobin CM, Levine WN. Statistical fragility of randomized clinical trials in shoulder arthroplasty. *J Shoulder Elbow Surg*. 2021;30(8):1787-1793. doi:10.1016/j.jse.2020.10.028
29. Medina Pabon MA, Naqvi U. Achilles tendinopathy. In: *StatPearls*. StatPearls Publishing LLC; 2024.
30. Metz R, Verleisdonk EJ, van der Heijden GJ, et al. Acute Achilles tendon rupture: minimally invasive surgery versus nonoperative treatment with immediate full weightbearing—a randomized controlled trial. *Am J Sports Med*. 2008;36(9):1688-1694. doi:10.1177/0363546508319312
31. Milgrom C, Finestone A, Zin D, Mandel D, Novack V. Cold weather training: a risk factor for Achilles paratendinitis among recruits. *Foot Ankle Int*. 2003;24(5):398-401. doi:10.1177/107110070302400504
32. Miller EK, Neuman BJ, Jain A, et al. An assessment of frailty as a tool for risk stratification in adult spinal deformity surgery. *Neurosurg Focus*. 2017;43(6):E3. doi:10.3171/2017.10.FOCUS17472
33. Milto AJ, Negri CE, Baker J, Thuppal S. The statistical fragility of foot and ankle surgery randomized controlled trials. *J Foot Ankle Surg*. 2023;62(1):191-196. doi:10.1053/j.jfas.2022.08.014
34. Möller M, Movin T, Granhed H, Lind K, Faxen E, Karlsson J. Acute rupture of tendon Achillis. A prospective randomised study of comparison between surgical and non-surgical treatment. *J Bone Joint Surg Br*. 2001;83(6):843-848. doi:10.1302/0301-620x.83b6.11676
35. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA*. 2005;294(17):2203-2209. doi:10.1001/jama.294.17.2203
36. Myhrvold SB, Brouwer EF, Andresen TKM, et al. Nonoperative or surgical treatment of acute Achilles' tendon rupture. *N Engl J Med*. 2022;386(15):1409-1420. doi:10.1056/NEJMoa2108447
37. Naji L, Dennis B, Rodrigues M, et al. Assessing fragility of statistically significant findings from randomized controlled trials assessing pharmacological therapies for opioid use disorders: a systematic review. *Trials*. 2024;25(1):286. doi:10.1186/s13063-024-08104-x
38. Olsson N, Silbernagel KG, Eriksson BI, et al. Stable surgical repair with accelerated rehabilitation versus nonsurgical treatment for acute Achilles tendon ruptures: a randomized controlled study. *Am J Sports Med*. 2013;41(12):2867-2876. doi:10.1177/0363546513503282
39. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210. doi:10.1186/s13643-016-0384-4
40. Paoloni JA, Appleyard RC, Nelson J, Murrell GA. Topical glyceryl trinitrate treatment of chronic noninsertional Achilles tendinopathy. A randomized, double-blind, placebo-controlled trial. *J Bone Joint Surg Am*. 2004;86(5):916-922. doi:10.2106/00004623-200405000-00005
41. Parisien RL, Danford NC, Jarin IJ, Li X, Trofa DP, Vosseller JT. The fragility of statistical findings in Achilles tendon injury research: a systematic review. *J Am Acad Orthop Surg Glob Res Rev*. 2021;5(9):e21.00018. doi:10.5435/JAAOSGlobal-D-21-00018
42. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ*. 1998;316(7139):1236-1238. doi:10.1136/bmj.316.7139.1236
43. Potter GE. Dismantling the fragility index: a demonstration of statistical reasoning. *Stat Med*. 2020;39(26):3720-3731. doi:10.1002/sim.8689
44. Ruzbarsky JJ, Khormaee S, Daluiski A. The fragility index in hand surgery randomized controlled trials. *J Hand Surg Am*. 2019;44(8):698.e1-698.e7. doi:10.1016/j.jhsa.2018.10.005
45. Silbernagel KG, Thomee R, Thomee P, Karlsson J. Eccentric overload training for patients with chronic Achilles tendon pain—a randomised controlled study with reliability testing of the evaluation methods. *Scand J Med Sci Sports*. 2001;11(4):197-206. doi:10.1034/j.1600-0838.2001.110402.x
46. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. doi:10.1136/bmj.l4898
47. Sussmilch-Leitch SP, Collins NJ, Bialocerkowski AE, Warden SJ, Crossley KM. Physical therapies for Achilles tendinopathy: systematic review and meta-analysis. *J Foot Ankle Res*. 2012;5(1):15. doi:10.1186/1757-1146-5-15
48. Tarantino D, Mottola R, Resta G, et al. Achilles tendinopathy pathogenesis and management: a narrative review. *Int J Environ Res Public Health*. 2023;20(17):6681. doi:10.3390/ijerph20176681
49. VanderWeele TJ, Mathur MB. Some desirable properties of the Bonferroni correction: is the Bonferroni correction really so bad? *Am J Epidemiol*. 2019;188(3):617-618. doi:10.1093/aje/kwy250
50. Veronesi F, Borsari V, Martini L, et al. The impact of frailty on spine surgery: systematic review on 10 years clinical studies. *Aging Dis*. 2021;12(2):625-645. doi:10.14336/AD.2020.0904
51. 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(6):622-628. doi:10.1016/j.jclinepi.2013.10.019
52. Wang Y, Zhou H, Nie Z, Cui S. Prevalence of Achilles tendinopathy in physical exercise: a systematic review and meta-analysis. *Sports Med Health Sci*. 2022;4(3):152-159. doi:10.1016/j.smhs.2022.03.003
53. Yang X, Meng H, Quan Q, Peng J, Lu S, Wang A. Management of acute Achilles tendon ruptures: a review. *Bone Joint Res*. 2018;7(10):561-569. doi:10.1302/2046-3758.710.BJR-2018-0004.R2
54. Young SW, Patel A, Zhu M, et al. Weight-bearing in the nonoperative treatment of acute Achilles tendon ruptures: a randomized controlled trial. *J Bone Joint Surg Am*. 2014;96(13):1073-1079. doi:10.2106/JBJS.M.00248