The Fragility of Statistically Significant Findings From Randomized Trials in Sports Surgery: A Systematic Survey

Moin Khan, Nathan Evaniew, Mark Gichuru, Anthony Habib, Olufemi R. Ayeni, Asheesh Bedi, Michael Walsh, P.J. Devereaux and Mohit Bhandari

Am J Sports Med published online November 28, 2016
DOI: 10.1177/0363546516674469

AJSM PreView, published on November 28, 2016 as doi:10.1177/0363546516674469

Current Concepts

The Fragility of Statistically Significant Findings From Randomized Trials in Sports Surgery

A Systematic Survey

Moin Khan,*† MD, MSc, FRCSC, Nathan Evaniew,†‡ MD, PhD, Mark Gichuru,† BSc, Anthony Habib,† MD, Olufemi R. Ayeni,† MD, MSc, FRCSC, Asheesh Bedi,§ MD, Michael Walsh,‡‖ MD, PhD, FRCPC, P.J. Devereaux,†‖ MD, PhD, FRCPC, and Mohit Bhandari,†‡ MD, PhD, FRCSC
Investigation performed at McMaster University, Hamilton, Ontario, Canada

Background: High-quality, evidence-based orthopaedic care relies on the generation and translation of robust research evidence. The Fragility Index is a novel method for evaluating the robustness of statistically significant findings from randomized controlled trials (RCTs). It is defined as the minimum number of patients in 1 arm of a trial that would have to change status from a nonevent to an event to alter the results of the trial from statistically significant to nonsignificant.

Purpose: To calculate the Fragility Index of statistically significant results from clinical trials in sports medicine and arthroscopic surgery to characterize the robustness of the RCTs in these fields.

Methods: A search was conducted in Medline, EMBASE, and PubMed for RCTs related to sports medicine and arthroscopic surgery from January 1, 2005, to October 30, 2015. Two reviewers independently assessed titles and abstracts for study eligibility, performed data extraction, and assessed risk of bias. The Fragility Index was calculated using the Fisher exact test for all statistically significant dichotomous outcomes from parallel-group RCTs. Bivariate correlation was performed to evaluate associations between the Fragility Index and trial characteristics.

Results: A total of 48 RCTs were included. The median sample size was 64 (interquartile range [IQR], 48.5-89.5), and the median total number of outcome events was 19 (IQR, 10-27). The median Fragility Index was 2 (IQR, 1-2.8), meaning that changing 2 patients from a nonevent to an event in the treatment arm changed the result to a statistically nonsignificant result, or $P \geq .05$.

Conclusion: Most statistically significant RCTs in sports medicine and arthroscopic surgery are not robust because their statistical significance can be reversed by changing the outcome status on only a few patients in 1 treatment group. Future work is required to determine whether routine reporting of the Fragility Index enhances clinicians' ability to detect trial results that should be viewed cautiously.

Keywords: Fragility Index; randomized control trial; sports medicine; orthopaedic surgery

Although improving the methodologic quality of randomized controlled trials (RCTs) has received substantial attention in the sports medicine literature, little consideration has been directed toward the fact that the majority of RCTs in orthopaedics demonstrating statistically significant effects may be at risk of spurious findings or improbably large treatment effects due to small sample sizes and few outcome events.[2,4,5,19,23,28]

The Fragility Index is a recently described metric that informs evidence users about the fragility/robustness of statistically significant results.[10,27,31] It is defined as the minimum number of patients in a trial group with fewer events that would have to change status from a nonevent to an event to alter the results from statistically significant to nonsignificant. Events refer to the occurrence of any dichotomous event, such as the successful healing of a rotator cuff, presence of graft failure, recurrent dislocation, positive pivot test result, or the achievement of a certain functional score. Trials with a small Fragility Index suggest that the statistical significance of the results hinges on only a few events. A large Fragility Index should increase one's confidence in the observed treatment effect.

For example, consider a recent RCT in which 35 patients with large 2-tendon rotator cuff tears were randomized to rotator cuff repairs either with or without

The American Journal of Sports Medicine, Vol. XX, No. X
DOI: 10.1177/0363546516674469

acellular human dermal matrix (AHDM) augmentation and were assessed using gadolinium-enhanced magnetic resonance imaging postoperatively.[1] In this trial, 17 of the 20 patients who received augmentation with AHDM were found to have an intact rotator cuff tendon at 12 or 24 months, compared with 6 of 15 who did not.[1] This difference was statistically significant ($P < .01$), but if just 2 more patients in the nonaugmented group were found to have intact rotator cuffs, the trial result would not be statistically significant (ie, a result that is possible given the lack of blinding, losses to follow-up, and potential for prognostic imbalance between the groups due to the small sample size). In the examples above, the Fragility Index for the trial would be 2 events.

No studies to date have evaluated the Fragility Index for sports medicine and arthroscopy trials. Our primary objective was to evaluate the robustness of statistically significant results from sports medicine and arthroscopic trials by determining the Fragility Index. Our secondary objective was to identify any trial characteristics that are associated with Fragility Index values.

METHODS

Eligibility Criteria

We performed a systematic survey of all RCTs of orthopaedic sports or arthroscopic surgery published between January 1, 2005, and October 30, 2015. We included all trials that were randomized according to a 1:1 parallel 2-arm design, reported in their abstract at least 1 statistically significant dichotomous outcome (ie, a P value of $<.05$ under a null hypothesis that no difference existed or a 95% CI that excluded a null value), and examined a perioperative intervention in patients undergoing orthopaedic sports or arthroscopic surgery. We restricted our search to trials on human subjects reported in English.

Identification of Studies

We searched MEDLINE and EMBASE databases to identify all potentially eligible trials related to orthopaedic sports or arthroscopic surgery. We used medical subject headings (MeSH), as well as Emtree headings and subheadings in various combinations, and supplemented with free text to increase sensitivity (see Appendix 1, available online). We used the Cochrane sensitivity- and precision-maximizing search strategy for identifying RCTs in both MEDLINE and EMBASE.[16] The search strategy was adapted in PubMed to search for articles published online ahead of print. Two reviewers (A.H. and M.G.) independently screened the titles and abstracts of all trials independently for eligibility using piloted screening forms. Duplicate articles were manually excluded. Both reviewers reviewed the full-text version of all trials identified by title and abstract screening to determine final eligibility. All discrepancies were resolved by a consensus decision requiring rationale with the first author.

Data Extraction and Assessment of Risk of Bias

Data were extracted independently and in duplicate by both reviewers (A.H. and M.G.) using a piloted electronic data extraction form, and all extracted data were verified by the first author. For each RCT, we extracted data for 1 statistically significant dichotomous outcome that was identified from the abstract. When more than 1 eligible outcome was presented, we chose the primary outcome whenever possible or the most patient-important secondary outcome. This was done according to the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach for distinguishing between outcomes that are critical for decision making, important but not critical, or of low importance.[18]

The following data were extracted for each outcome: journal name, publication year, funding source, sample size for each arm, losses to follow-up for each arm, number of events for each arm, reported P value and statistical test used, and whether outcomes were primary or secondary. We also recorded the 2013 Thomson Reuters Journal Impact Factor and the most recent Science Citation Index for each trial. The Science Citation Index is a metric of citation frequency that reflects the cumulative number of citations to source items indexed within the Web of Science Core Collection.[32]

Two reviewers independently performed duplicate assessment of trial-level risk of bias using the Cochrane Collaboration's Risk of Bias Tool, which assesses allocation sequence generation and concealment; blinding of surgeons, outcome assessors, and patients; losses to follow-up and missing data; selective outcome reporting; and other potential sources of bias.[21]

Application of the Fragility Index

The Fragility Index for each outcome was calculated according to the method described by Walsh et al[31] using 2×2 contingency tables. The P value for each outcome was first recalculated using a 2-sided Fisher exact test. We then added events to the group with a smaller number of events while subtracting nonevents from the same group

*Address correspondence to Moin Khan, MD, MSc, FRCSC, Division of Orthopaedics, Department of Surgery, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4L8, Canada (email: moinkhanmd@gmail.com).
†Division of Orthopaedics, Department of Surgery, McMaster University, Hamilton, Ontario, Canada.
‡Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada.
§MedSport, Department of Orthopaedic Surgery, University of Michigan, Ann Arbor, Michigan, USA.
‖Department of Medicine, McMaster University, Hamilton, Ontario, Canada.

One or more of the authors has declared the following potential conflict of interest or source of funding: A.B. has received consulting fees from A3 Surgical and Smith & Nephew. O.R.A. has received consulting fees from Smith & Nephew. M.B. has received consulting fees from Amgen, Bioventus, Depuy, Eli Lilly, Smith & Nephew, Stryker, and Zimmer.

TABLE 1
Characteristics of Included Randomized Controlled Trials (N = 48)

Characteristic	No. of Studies (%)
Primary intervention	
Surgical	18 (38)
Shoulder	6 (13)
Knee	12 (25)
Hip	0 (0)
Anesthetic	19 (40)
Nonoperative	11 (23)
Outcome assessed	
Imaging	12 (25)
Perioperative comfort	10 (21)
Clinical assessment	13 (27)
Adverse events	13 (27)
Year of publication	
2005	5 (10)
2006	2 (4)
2007	5 (10)
2008	3 (6)
2009	2 (4)
2010	5 (10)
2011	7 (15)
2012	5 (10)
2013	5 (10)
2014	5 (10)
2015	4 (8)
Industry funding	
Yes	5 (10)
No/unclear	43 (90)
Journal	
Journal of Arthroscopy	14 (29)
American Journal of Sports Medicine	6 (13)
Knee Surgery, Sports Traumatology, Arthroscopy	5 (10)
Journal of Bone and Joint Surgery	4 (8)
Other	19 (40)

to keep the total number of participants constant. Events were added iteratively until the calculated P value became $\geq$.05. The smallest number of additional events required to obtain $P \geq .05$ was the Fragility Index for that outcome.

Statistical Analysis

We used descriptive statistics to summarize the Fragility Index for the sampled trials. Interobserver agreement for reviewers' assessments of study eligibility was calculated with Cohen's kappa coefficient.[15] Based on the guidelines by Landis and Koch,[24] a kappa (κ) of 0 to 0.2 represents slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, and 0.61 to 0.80 substantial agreement. A value above 0.80 is considered almost perfect agreement. We used the Pearson correlation coefficient to evaluate direct correlations. We planned a priori to evaluate the correlation between Fragility Index and sample size, outcome events, impact factor, citation index, and P values.[10,31] All tests of significance were 2 tailed, and $P < .05$ was considered significant. All analyses were performed using Microsoft Excel (Microsoft Corp; 2011) and SPSS version 21 (IBM Corp; 2012).

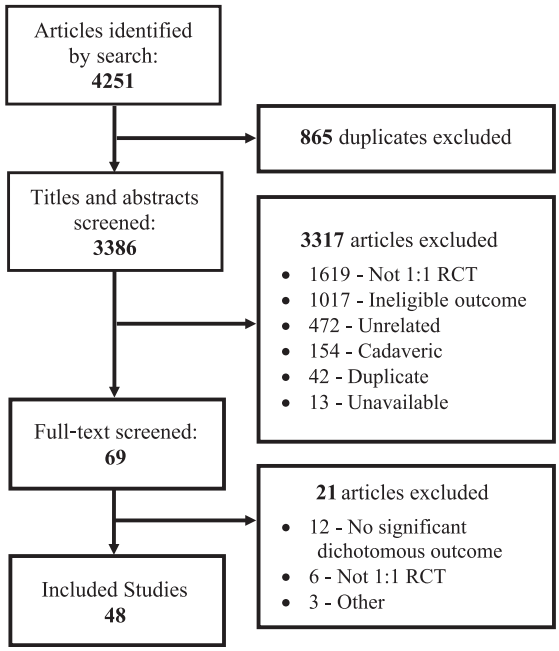

Figure 1. Flow of articles through screening and reasons for exclusion. RCT, randomized controlled trial.

RESULTS

Study Selection

The literature search identified 4251 potentially eligible studies. After exclusion of duplicate articles and application of exclusion criteria, 69 articles were eligible for full-text review. Of the 4203 excluded articles, 1619 were not randomized according to a 1:1 parallel design and 1017 did not present a statistically significant dichotomous finding. After full-text review, 21 articles were excluded, resulting in 48 orthopaedic sports medicine or arthroscopic surgical RCTs with 48 outcomes eligible for inclusion (Table 1 and Appendix 2, available online). The κ coefficient for the agreement between reviewers for title and abstract eligibility decision was 0.66 (95% CI, 0.58-0.75), indicating substantial agreement (Figure 1).

Characteristics of Trials and Outcomes

The median sample size of the included trials was 64 patients (interquartile range [IQR], 48.5-89.5), and the median number of total losses to follow-up was 2.5 patients (IQR, 0-10). Among the included trials, sequence generation was at a low risk of bias in 24 (50%) and allocation concealment was at a low risk of bias in 16 (33%). Investigators blinded surgeons in 7 (15%), patients in 15 (31%), and outcome assessors in 18 (38%) (Table 2). The median journal impact factor was 3.2 (IQR, 2.4-3.5), and the median Science Citation Index was 10 (IQR, 4-33).

Of the 48 outcomes, 27 (56%) outcomes were primary, 14 (29%) were secondary, and 7 (15%) were not specified. Trials reported dichotomous outcomes related to postoperative

TABLE 2
Risk of Bias of Included Randomized Controlled Trials

Item	Risk of Bias, n (%)		
	Yes	No	Unclear
Was the allocation sequence adequately generated?	24 (50)	13 (27)	11(23)
Was allocation adequately concealed?	16 (33)	4 (8)	28 (42)
Blinding surgeons: Was knowledge of the allocated interventions adequately prevented?	7 (15)	16 (33)	25 (52)
Blinding patients: Was knowledge of the allocated interventions adequately prevented?	15 (31)	7 (15)	26 (54)
Blinding outcome assessors: Was knowledge of the allocated interventions adequately prevented?	18 (38)	6 (13)	24 (50)
Were losses to follow-up (missing outcome data) accounted for?	27 (56)	18 (38)	3 (6)
Are reports of the study free of suggestion of selective outcome reporting?	26 (54)	4 (8)	18 (38)
Was the study free of any other potential bias (such as expertise bias)?	11 (23)	1 (2)	36 (75)

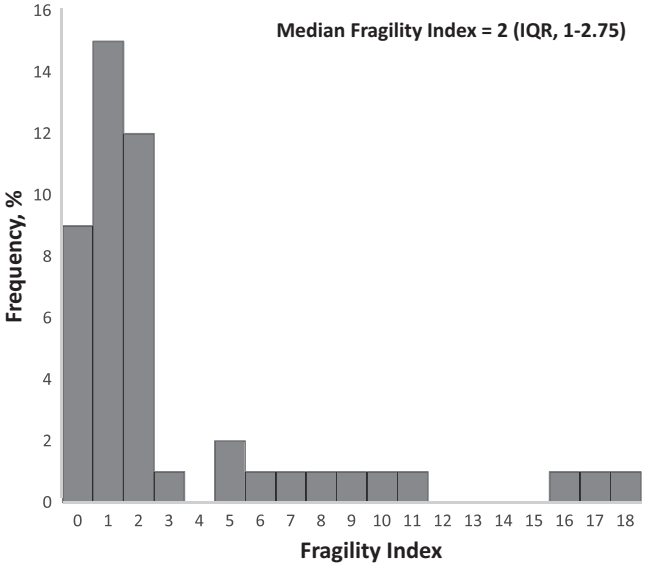

Figure 2. Distribution of Fragility Index values from 48 trials. The median number of patients whose status would have to change from a nonevent to an event to change a statistically significant result to a nonsignificant result was 2 (interquartile range [IQR], 1-2.8).

imaging in 12 (25%), perioperative comfort in 10 (21%), clinical assessment in 13 (27%), and presence of adverse events in 13 (27%). We found that 18 (38%) of the included trials evaluated surgical interventions, and the remainder evaluated perioperative anesthetic (19 [40%]) or nonsurgical (11 [23%]) interventions in patients undergoing arthroscopic surgery. Reported *P* values for each outcome were <.05 but ≥.01 for 31 (65%), <.01 but ≥.001 for 15 (31%), and <.001 for 2 (4%). The median number of events across both treatment groups for each outcome was 19 (IQR, 10-27).

Fragility Index

The median Fragility Index was 2 (IQR, 1-2.8) for the 48 evaluated dichotomous outcomes (Figure 2 and Table 3). Eight (17%) outcomes became nonsignificant when we recalculated the *P* value using the 2-sided Fisher exact test and, therefore, had a Fragility Index of zero. The Fragility Index was less than or equal to 3 events in 37 (77%) trials and less than or equal to the total number of patients lost to follow-up for 23 (48%) outcomes. Increasing Fragility index values correlated strongly with smaller reported *P* values ($r = 0.87$; $P < .01$) (Figure 3). Direct correlations with sample size, outcome events, impact factor, and citation index were not statistically significant (Table 4).

TABLE 3
Fragility Index by Subgroups Based on Trial or Outcome Characteristics[a]

Characteristic	Fragility Index, Median (IQR)
All trials (N = 48)	2 (1-2.8)
Outcome	
Primary (n = 27)	2 (1-6)
Secondary (n = 14)	1 (0.75-2)
Not reported (n = 7)	1 (0-2)
Sample size	
≤100 (n = 43)	2 (1-3)
>100 (n = 5)	0 (0-9)
Outcome events	
≤30 (n = 40)	1 (1-2)
>30 (n = 8)	5 (0.3-9.8)
Industry funding	
Yes (n = 5)	1 (0.5-10)
No/unclear (n = 43)	2 (1-3)
Reported *P* value	
<.05 to .01 (n = 31)	1 (0-2)
<.01 to .001 (n = 15)	5 (1-9)
<.001 (n = 2)	12 (7-12)

[a]IQR, interquartile range.

DISCUSSION

Key Findings

This systematic survey of the literature found that the median Fragility Index from RCTs in sports medicine and arthroscopic surgery reporting dichotomous outcomes was 2 (IQR, 1-2.8). This means that reversing the outcome

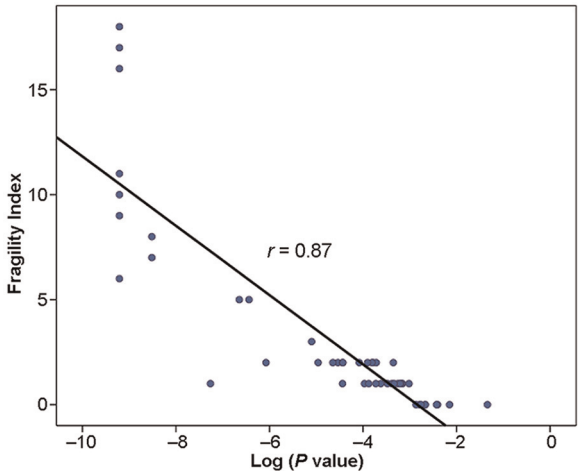

Figure 3. Increasing Fragility Index values correlated with decreasing *P* values.

status of 2 patients in 1 treatment group would completely change the results of a trial from statistically significant to nonsignificant. We identified that increasing Fragility Index values correlated strongly with smaller reported *P* values ($r = 0.87$; $P < .01$).

We found that for the outcomes evaluated, in 48% of trials, more patients were lost to follow-up than would be required to render the result nonsignificant based on the corresponding trial's Fragility Index. The trial results could have been different simply if the outcomes had been collected from some of those lost participants. Furthermore, we found that almost 80% of included trials had a Fragility Index value of ≤ 3. This finding suggests that there is considerable potential for findings to be significantly altered by factors such as simple data errors, loss to follow-up, early withdrawals, small imbalances in group prognosis, and biased evaluation. For this reason, it is critical that detailed information regarding the methodology of the trial and information related to potential risk of bias be reported to readers.

Despite the significant focus in the recent literature regarding the importance of methodological reporting in RCTs, we found a significant proportion of trials did not report important information relating to potential risk of bias in study results. For example, more than 50% of trials did not report blinding of surgeons, patients, or outcome assessors (Table 2). This is similar to other studies evaluating the quality of reporting in the orthopaedic literature.[3,6] For example, Chess and Gagnier[8] reviewed 232 RCTs published from January 2006 to December 2010 in top orthopaedic journals by impact factor and assessed 10 criteria relating to risk of bias. They found that 42% of trials failed to report on methodologic considerations that would have a significant effect on potential risk of bias present in the trial. Blinding is critical to ensure unbiased outcome adjudication; in trials with small Fragility Index values, the potential for bias in outcome assessment due to lack of blinding further highlights the potential fragility of the results.

TABLE 4
Direct Correlations Between Characteristics of Outcomes and Fragility Index Values

Variable	Correlation Coefficient (r)	Significance (P)[a]
P value	0.87	<.01
Sample size	0.11	.46
No. of events	0.23	.12
Impact factor	0.01	.93
Science Citation Index	0.17	.26
Risk of bias	0.05	.72
Loss to follow-up	0.05	.75

[a]A logarithmic transformation was applied to the *P* values.

Another finding that was identified in this study of trials in sports medicine and arthroscopic surgery was the lack of statistically significant findings from the majority of trials identified during the screening process. As has been identified in numerous reviews of the orthopaedic literature, many orthopaedic trials are not powered sufficiently to demonstrate a statistically significant difference between interventions.[11] Meta-analyses of many common orthopaedic interventions, including trials of sports medicine, also have found no difference between many comparator groups, potentially due to widespread deficiencies such as small sample sizes and potential risk of bias.[7,12,25]

Relationship to Previous Studies

This is the first study to evaluate the Fragility Index for randomized control studies in sports medicine and arthroscopic surgery. Previous studies evaluated the Fragility Index and its relationship to the spine literature, critical care literature, and studies published in high-impact medical journals.[10,27,31] The findings of this study are similar to the findings of the reviews by Evaniew et al[10] and Ridgeon et al,[27] who identified the median Fragility Index to be 2 (IQR, 1-3) in the spine literature and 2 (IQR, 1-3.5) in the critical care literature, respectively. Similar to the literature in spine surgery, trials in orthopaedic sports medicine and arthroscopic surgery have generally been of small sample sizes, have been at increased risk for bias, and generally have few outcome events.[13,19] The median sample size in our survey was 64 patients (IQR, 48.5-89.5), and the median total number of events for each outcome was 19 (IQR, 10-27). This contrasts significantly with the findings by Walsh et al[31] of 399 RCTs published in high-impact medical journals, which found a median sample size of 682 patients (range, 15-112,604) and a median number of events per outcome of 112 (range, 8-5142). The Fragility Index reported by them was 8 events (IQR, 3-18), which is higher than our reported median Fragility index of 2 events (IQR, 1-2.8). This highlights the significant discrepancy in the size of trials and outcome events and thus the confidence in the findings between research in both fields.

Strengths and Limitations

This study was methodologically rigorous and involved a comprehensive systematic search of medical databases and duplicate and independent reviewers for screening, data extraction, and outcome evaluation. A search of additional databases is unlikely to change the conclusions of this review. We had substantial agreement between reviewers for study eligibility. We identified a strong correlation of increasing Fragility Index in sports medicine and arthroscopic outcomes with decreasing *P* values; however, unlike previous studies, we were unable to demonstrate a statistically significant correlation between Fragility Index and sample size or number of outcome events.[10,31] This discrepancy may be due to a lack of statistical power related to the limited number of available RCTs in orthopaedic sports medicine. In addition, this may be due to the relatively small number of trials eligible for this review and clustering of data around very low Fragility Index values.

A limitation of the Fragility Index itself is its applicability only to trials performing 1:1 randomization and reporting statistically significant findings for dichotomous outcomes. Thus, it is not applicable to studies presenting data on continuous scales, such as the majority of functional outcome scores that constitute the majority of reported outcomes in the orthopaedic literature.[9,20,26,30] Another potential limitation may be the limited number of surgical outcomes evaluated in this review, as we included trials evaluating both surgical and nonsurgical interventions. A subgroup analysis of trials presenting only surgical outcomes did not, however, change the median Fragility Index score.

Implications

The Fragility Index delivers an innovative method to inform clinicians about the reliability and confidence they should have in the results of a study. Reliance on an arbitrary threshold for *P* values to interpret study results has been criticized due to its simplistic nature.[14,17,29] *P* values do not inform readers about the magnitude of the treatment effect or identify the ranges of possible true values consistent with the observed data.[5,22,31] Although the addition of the 95% CI can provide additional information and overcome some of these limitations, this does not address the potential for trials with small sample sizes and outcome events to produce spurious results.[10] As has been demonstrated by this review, the majority of trials presenting statistically significant *P* values are not at all robust; in fact, having 1 or 2 outcome events switch can lead a trial's results to be statistically nonsignificant. The Fragility Index, when used in conjunction with the *P* value and CIs, provides additional useful information for clinicians interpreting the results of a study.

When interpreting the results of a trial, clinicians should critically evaluate the size of the study and the number of outcome events in each arm. It is intuitive that studies that are larger have greater differences between the number of events in each arm, have lower *P* values, and are more likely to be robust. However, it is through the adoption of the Fragility Index that this can actually be quantified and easily interpreted. It is important for clinicians to pay particular attention to the number of patients lost to follow-up, as this can also affect the results of a trial, particularly when the number of patients lost to follow-up is greater than a trial's Fragility Index. Clinicians performing trials must be conservative in the selection of appropriate statistical tests to perform in the analysis of results. Our study demonstrated that 17% of trials lost statistical significance simply by selecting an alternative statistical test to evaluate statistical significance.

Clinicians can be confident that the results of a study with a large Fragility Index are more robust than are those from studies with smaller index values. Further research is required to define a threshold value for acceptable robustness, as well as potentially develop sample size calculations that can incorporate Fragility Index estimations.[10]

CONCLUSION

Most statistically significant RCTs in sports medicine and arthroscopic surgery are not robust because their statistical significance can be reversed by changing the outcome status on only a few patients in 1 treatment group. Future work is required to determine whether routine reporting of the Fragility Index enhances clinicians' ability to detect trial results that should be viewed cautiously.

REFERENCES

1. Barber FA, Burns JP, Deutsch A, Labbé MR, Litchfield RB. A prospective, randomized evaluation of acellular human dermal matrix augmentation for arthroscopic rotator cuff repair. *Arthroscopy*. 2012;28(1):8-15.
2. 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.
3. Bhandari M, Guyatt GH, Lochner H, Sprague S, Tornetta P III. Application of the Consolidated Standards of Reporting Trials (CONSORT) in the fracture care literature. *J Bone Joint Surg Am*. 2002;84(3):485-489.
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.
5. Bhandari M, Montori VM, Schemitsch EH. The undue influence of significant p-values on the perceived importance of study results. *Acta Orthop*. 2005;76(3):291-295.
6. Bhandari M, Richards RR, Sprague S, Schemitsch EH. The quality of reporting of randomized trials in the *Journal of Bone and Joint Surgery* from 1988 through 2000. *J Bone Joint Surg Am*. 2002;84(3):388-396.
7. Chaudhry H, Ayeni OR. The etiology of femoroacetabular impingement: what we know and what we don't. *Sports Health*. 2014;6(2):157-161.
8. Chess LE, Gagnier J. Risk of bias of randomized controlled trials published in orthopaedic journals. *BMC Med Res Methodol*. 2013;13:76.
9. El-Daly I, Ibraheim H, Rajakulendran K, Culpan P, Bates P. Are patient-reported outcome measures in orthopaedics easily read by patients? *Clin Orthop Relat Res*. 2016;474(1):246-255.

10. Evaniew N, Files C, Smith C, et al. The fragility of statistically significant findings from randomized trials in spine surgery: a systematic survey. *Spine*. 2015;15(10):2188-2197.
11. Evaniew N, Madden K, Bhandari M. Cochrane in CORR®: arthroplasties (with and without bone cement) for proximal femoral fractures in adults. *Clin Orthop Relat Res*. 2014;472(5):1367-1372.
12. Evaniew N, Simunovic N, McKee MD, Schemitsch E. Cochrane in CORR®: surgical versus conservative interventions for treating fractures of the middle third of the clavicle. *Clin Orthop Relat Res*. 2014;472(9):2579-2585.
13. Farrokhyar F, Karanicolas PJ, Thoma A, et al. Randomized controlled trials of surgical interventions. *Ann Surg*. 2010;251(3):409-416.
14. Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol*. 1998;51(4):355-360.
15. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 2nd ed. New York: John Wiley and Sons; 1981.
16. Glanville JM, Lefebvre C, Miles JNV, Camosso-Stefinovic J. How to identify randomized controlled trials in MEDLINE: ten years on. *J Med Libr Assoc*. 2006;94(2):130-136.
17. Goodman SN. Toward evidence-based medical statistics, 1: the P value fallacy. *Ann Intern Med*. 1999;130(12):995-1004.
18. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines, 2: framing the question and deciding on important outcomes. *J Clin Epidemiol*. 2011;64(4):395-400.
19. Harris JD, Cvetanovich G, Erickson BJ, et al. Current status of evidence-based sports medicine. *Arthroscopy*. 2014;30(3):362-371.
20. Hetaimish BM, Khan M, Crouch S, et al. Consistency of reported outcomes after arthroscopic management of femoroacetabular impingement. *Arthroscopy*. 2013;29(4):780-787.
21. Higgins JP. *Cochrane Handbook for Systematic Reviews of Interventions*. Vol 5. 5th ed. Chichester, UK: Wiley-Blackwell; 2008.
22. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
23. Karanicolas PJ, Bhandari M, Taromi B, et al. Blinding of outcomes in trials of orthopaedic trauma: an opportunity to enhance the validity of clinical trials. *J Bone Joint Surg Am*. 2008;90(5):1026-1033.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
25. Mundi R, Bhandari M. Cochrane in CORR(®): double-bundle versus single-bundle reconstruction for anterior cruciate ligament rupture in adults (review). *Clin Orthop Relat Res*. 2016;474(5):1099-1101.
26. Ramkumar PN, Harris JD, Noble PC. Patient-reported outcome measures after total knee arthroplasty: a systematic review. *Bone Joint Res*. 2015;4(7):120-127.
27. 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(7):1278-1284.
28. SPRINT Investigators, Bhandari M, Tornetta P, et al. (Sample) size matters! An examination of sample size from the SPRINT trial study to prospectively evaluate reamed intramedullary nails in patients with tibial fractures. *J Orthop Trauma*. 2013;27(4):183-188.
29. Sterne JA, Davey Smith G. Sifting the evidence: what's wrong with significance tests? *BMJ*. 2001;322(7280):226-231.
30. Sung J, Siegel J, Tornetta P, Bhandari M. The orthopaedic trauma literature: an evaluation of statistically significant findings in orthopaedic trauma randomized trials. *BMC Musculoskelet Disord*. 2008;9:14.
31. 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.
32. Web of Science Citation Index. http://ip-science.thomsonreuters.com/cgi-bin/jrnlst/jloptions.cgi?PC=K. Accessed April 3, 2016.