How robust are clinical trials in primary and secondary ankle sprain prevention?

How robust are clinical trials in primary and secondary ankle sprain prevention?

Abstract

Objectives

Determine the statistical stability of RCTs examining primary and secondary prevention of ankle sprains.

Methods

Databases were searched to August 2023. We included parallel design RCTs, using conservative interventions for preventing ankle sprain, reporting dichotomous injury event outcomes. Statistical stability was quantified using Fragility Index (FI) and Fragility Quotient (FQ). Subgroup analyses were undertaken to test if FI varied based on by study objective, original approach to analysis (frequency vs time to event), follow-up duration, and pre-registration.

Results

3559 studies were screened with 45 RCTs included. The median number of events required to change the statistical significance (FI) was 4 (IQR 1-6). FI was similar regardless of study objective, original analysis, follow-up duration, and pre-registration status. Median (IQR) FQ was 0.015 (0.005-0.046), therefore reversing events <2 patients/100 would alter significance. In 80% of studies the number of patients lost to follow-up was greater than the FI.

Conclusion

RCTs informing primary and secondary prevention of ankle sprain are fragile. Only a small percentage of outcome event reversals would reverse study significance, and this is often exceeded by the number of drop outs. Robust reporting of dichotomous outcomes requires the use P values and key metrics such as FI or FQ.

Key words

P values; Statistical Fragility; Injury Prevention; Ankle sprain

Background

Lateral ankle sprain (LAS) is one of most prevalent injuries in physically active populations,(Gribble *et al.* 2016) and has the highest re-injury rate across all lower-limb musculoskeletal conditions.(Hootman *et al.* 2007; Gribble *et al.* 2016) Strategies for preventing LAS usually involves a combination of therapeutic exercise, ankle taping, or bracing. These strategies may be employed for primary prevention in a healthy population, or to reduce the risk of recurrent LAS (secondary prevention).

Randomised controlled trials (RCTs) are commonly used to quantify preventative effects, by comparing the counts/proportions of injured subjects across groups (intervention vs control); with conclusions informed by null hypothesis significance testing (NHST) and inferential probability (P values). However, over-relying on P value thresholds when making study conclusions can increase the risk of false-positive discovery (i.e. making an erroneously claim that the treatment is effective when it is not).(Colquhoun 2017) Feinstein (Feinstein 1990) and later Walsh (Walsh *et al.* 2014) promoted the use of the Fragility Index (FI) to help quantify the numerical stability (fragility) of observed differences in dichotomous data within clinical trials. FI represents the minimum number of outcome event reversals needed to overturn the trial results (i.e. where findings change from significant to nonsignificant and vice versa). A lower FI indicates less statistical robustness and compliments interpretation of P values. Although there is no critical cut-off for FI,(Khan *et al.* 2020) we can generate context by comparing it with the number of participant drop outs; our confidence in the significance of any effect should be reduced, if attrition approaches the FI number.(Walsh *et al.* 2015)

Scientific research can often be based on false-positive, non-replicable conclusions.(Ioannidis 2005) There is increasing empirical meta-research investigating the credibility of research practices in sport and exercise medicine research. A 2020 audit (Buttner *et al.* 2020) highlighted a propensity for questionable research practices in high-impact sport and exercise medicine journals, most commonly, this included hypothesizing after the results are known (HARKing) and p-hacking. Recent audits of RCTs published in the orthopaedic literature reported a mean FI of around 5,(Parisien *et al.* 2021; Xu *et al.* 2022; Fackler *et al.* 2022) and in 23-78% of outcomes, the numbers lost to follow-up exceeded the FI.

The purpose of this study was to determine the statistical stability of experimental research examining primary and secondary prevention of ankle sprains. Our key objectives were to calculate FI in each included study, and to compare the absolute FI with the number of patients lost to follow-up. Our

second objective was to evaluate if FI is influenced by study objective (primary vs secondary prevention), original analysis (frequency data vs time-to-event), sample size, follow-up duration, and pre-registration.

Methods

Study selection

In August 2023, 2 authors (CB and JW) conducted an electronic search on MEDLINE, EMBASE and on the Physiotherapy Evidence Database (PEDro). (**Supplementary file 1**). In PEDro, we ran 3 separate searches for clinical trials using the terms “re-injur$”, “reinjur$”, “recurren$”, “instability”, “sprain”, limiting each to ‘clinical trials’ AND the ‘foot or ankle’. Citation tracking was also undertaken on recent meta-analyses in this field.(Kemler *et al.* 2011; Schiftan *et al.* 2015; Taylor *et al.* 2015; Doherty *et al.* 2017; Bellows and Wong 2018; de Vasconcelos *et al.* 2018; Wagemans *et al.* 2022) No date or language restrictions were applied. Because this review is based on publicly available data and did not involve patients, an institutional review board approval was not sought. All studies generated from the electronic search were transferred onto Raayan software, where the titles and abstracts were independently screened by two authors (CB and AS), using the following predefined inclusion criteria: (1) 2-arm RCT using 1:1 randomization to a conservative intervention or control arm (2) reported dichotomous outcomes for ankle sprain / re-sprain. Articles that were post hoc secondary analyses of previously reported RCTs were also included. We excluded studies using surgical interventions. As FI can only be applied to dichotomous outcomes,(Tignanelli and Napolitano 2019) we could not include RCTs that only reported continuous event variables (eg. injury incidence, or time to re-injury).

Data extraction and analysis

Data from all studies were extracted and checked independently by at least 2 authors, using a pre-defined data collection form. The main author (CB) extracted data from all studies, and 2 authors (JW and AS) independently extracted data from 50% of studies each. In case of any discrepancies, a consensus was reached among all three authors (CB, JW, AS). Primary data abstracted included, the study objective (primary or secondary prevention), sample size of each group, follow-up duration, number of participants lost to follow-up, number of participants in comparative groups that suffered an injury/re-injury event over the entire follow-up period, and the type of statistical analysis undertaken

(eg. Fisher exact, Chi^2, or other). (**Supplementary file 2**) Our primary outcome was the median (interquartile range [IQR]) FI at the P = 0.05 threshold. Secondary outcomes were the median fragility quotient (FQ) with IQR and the number of RCTs in which the number of participants lost to follow-up was greater than the FI. Each trial result was inputted to a two-by-two contingency table in line with the author's original analysis. When trials used a time-to event outcome, our contingency tables were based on the number of events in each group for the entire follow-up period.(Walsh *et al.* 2014) The P values reported for each study were verified for accuracy using the Fisher exact test.

To calculate the FI, we used the FragilityTools package for R statistical software (v4.0.4).(*Https://Github.com/brb225/FragilityTools.* ; Baer *et al.* 2021a) (**Supplementary file 3**) In brief, the FragilityTools algorithms manipulated the injury counts in each study, until the fewest number of outcome modifications necessary to reverse significance occurs. This method is described as the "exact" algorithm and overcomes limitations of the commonly used Walsh (Walsh *et al.* 2014) and Khan (Walsh *et al.* 2014; Khan *et al.* 2020) algorithms which only modify injury counts in the group with the fewest number of events. (Baer *et al.* 2021a) When the proportion of re-injuries was reported as being significantly different ($p<0.05$), then: 1) an injury event was added to the group with a lower number of events (whilst simultaneously a non-event was subtracted to keep the number of participants in this group constant) or, 2) subtracted from the group with a greater number of events (whilst simultaneously a non-event was added to keep the number of participants in this group constant). Whichever modification produced a greater change in P value was selected. This process was repeated until the Fisher exact test 2-sided P value became >0.05.(Parisien *et al.* 2019) The opposite approach was taken in studies reporting non significance ($p>0.05$), with injury events modified, until the Fisher exact test 2-sided P value became <0.05.(Khan *et al.* 2020) For example, if a total of five iterations (i.e., five changes in injury outcomes) were required to change a study form statistically significant ($P<0.05$) to non-significant ($P>0.05$), the FI would be 5.

A previous criticism of FI is that it does not consider the likelihood of an event reversal. In other words, the hypothetical outcome reversals used to compute FI may be realistic for studies in which clinical events are common, however, they may not represent a real-world scenario when clinical events are infrequent. For example, a hypothetical event modification from non-injured to injured would be less likely in a study in which the treatment group has a 2% injury rate, compared to a study in which the treatment group has a 25% injury rate. Incidence fragility indices were developed to address this issue,

such that multiple FI's are computed across a range of "sufficiently likely outcome modifications." This is done by imposing a series of constraints on what event modifications can be performed for FI computations, based on four different likelihood thresholds. These likelihood thresholds are computed for each study using the observed event and non-event rates in the treatment and control groups (see Baer et al for full details).(Baer *et al.* 2021a) Thus, this method allows one to assess how stable the FI is (i.e., does the FI change substantially if the likelihood of an event modification changes). Incidence fragility indices were computed FragilityTools package. (*Https://Github.com/brb225/FragilityTools.* ; Baer *et al.* 2021a) Studies were considered to have a stable FI if the FI remained consistent regardless of whether it was computed with the reported injury incidence in the control group or treatment group (generally, the two lowest likelihood thresholds used). Additionally, we compared the incidence fragility indices with the FI that was computed using the exact algorithm (which does not consider the likelihood of event reversal, as described in the previous paragraph).

To account for different sample sizes, and the influence that this could have on FI, we calculated the Fragility Quotient (FQ), which is the FI divided by the per protocol sample size; (Ahmed *et al.* 2016) again, smaller values indicate a less robust study. The number of participants lost to follow-up was compared with the FI for each trial, and we also calculated the proportion of RCTs with an FI that was ≤1% of the total sample size.(Khan *et al.* 2020) All analyses were performed using R statistical software, with cross-checking using Excel version 14.1.3 (Microsoft Corp) and an online fragility index calculator (ClinCalc.com).(*Https://Clincalc.com/stats/FragilityIndex.aspx.*)

Results

The electronic search generated 3559 potentially relevant studies; 63 were retrieved for full-text evaluation, and 45 RCTs met the eligibility criteria. (**Supplementary file 3**)The reasons for exclusion from full-text were: wrong study design (n=4), insufficient data reported (n=8), duplicate data (n=4), wrong type of intervention (n=1), participants not eligible (n=1). There was an aggregate of 24122 included participants across the 45 included RCTs, with almost equal numbers focusing on primary (n=23) and secondary prevention (n=22). Only 20% (9/45) of trials were pre-registered.

Study characteristics

The median (IQR) sample size was 221 (92-765). All studies recorded injury events over a defined follow-up period. In most studies, (n=36) this was quantified by months of follow-up (median 10 months (IQR: 6-12), with the remaining trials following-up over 1 or 2 athletic/sports playing seasons. The median (IQR) total number of injury events per trial was 27 (13 - 46), with a median of 16 (10-30) events in the control groups and 7 (3-19) events in the intervention groups. 21/45 RCTs (46.6%) had statistically significant differences in injury counts ($p<0.05$), the remaining 53.3% (24/45) had P values >0.05. 80.0% of studies (36/45) detailed the nature of their between-group statistical analysis; most (n=21) used Pearson chi-squared (X^2) or Fischer exact test to compare injury counts across groups. The remainder (n=14) reported injury counts, but their primary data analysis was based on incidence rates or a time to event (most commonly Cox Regression methods), either with or without covariates. In n=10 trials the number of participants lost to follow-up was not reported. In the remaining n=35 trials, the median number of participants lost to follow-up was 15 (IQR 6.5 to 40.5).

Fragility Index and Fragility Quotient

The median (IQR) FI of the 45 trials was 4 (1-6). This indicates that a median of 4 events was required to reverse the significance of LAS injury outcomes. The FI in trials reported to have a statistically significant effect (n=21) was 2 (1-6), and those which had non-significant effects (n=24) had an FI of 4 (3-6).

Of the 45 trials analyzed, n=35 reported attrition data. Of the 10 trials that did not report attrition data, 8 reported statistically significant effects. FIGURE 1, which includes the 35 trials that provided adequate data on attrition, shows that in 80.0% (28/35), the numbers lost to follow-up exceeds the FI. In the 13 studies that reported a statistically significant effect and had adequate attrition data, only one (Mohammadi 2007) had a FI exceeding the number of drop outs (n=0), however the significant effect was fragile (FI=1).

In our secondary objective, we found that FIs (median [IQR]) were similar in primary vs secondary prevention trials (4 [2-5.5] vs 3.5 [1-5.75]), registered vs unregistered trials (3 [1-4] vs 4 [2-6]), and there was also little difference in FI in trials that based their analysis on frequency data (eg. Pearson chi-squared (X^2) or Fischer exact test) compared to those using time-to-event analyses ((4 [2-5] vs 5

[1.5-7.5]). The bubbles in FIGURE 1 represent study sample size; the chart shows that trials in which FI exceeded drop out, were typically smaller (median sample size of 40; [IQR 28-156], n=7) compared to trials where drop out exceeded FI (median sample size of 209 [IQR 92-661], n=28). Trials that failed to report adequate data on drop out also tended to be larger (median samples size 745 [IQR 358-1566], n=10).

The median (IQR) FQ of the 45 trials was 0.015 (0.005-0.046), and in 20 RCTs, the FI was ≤1% of the study sample. Studies which reported a statistically significant effect had median FQ of 0.009 (0.005-0.015) and those which reported a non-significant effect had a median FQ of 0.033 (0.007-0.077).

Incidence Fragility Indices

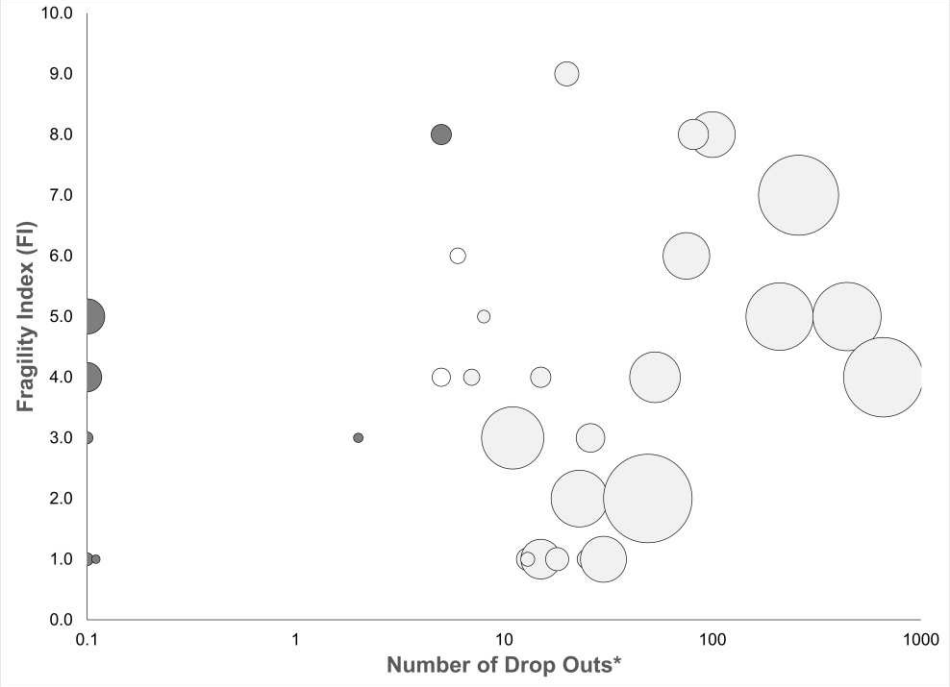
Incidence fragility indices were stable across the two lowest likelihood thresholds (corresponding to the event rates in the treatment and control groups), for all 19 studies which reported a significant effect, and for 15 of the 21 studies which reported a non-significant effect. In one case, (Heidt *et al.* 2000) the incidence fragility indices were highly unstable (i.e., the FI would be highly dependent on the likelihood of an event modification). For that case, a lower likelihood event modification (4.8%) would produce an FI=5 for a significant treatment beneficial effect, whereas a higher likelihood (8.1%) of event modification would produce an FI=21 for a significant treatment detrimental effect.

insert Figure 1 here

FIGURE 1

Fragility Index v Drop Out (Bubble area = Study sample size)

(n=35 RCT†)



† n=10 RCTs did not provide sufficient information on drop outs and were not included in the graph

Bubble area represents sample size

*x axis is on log10 scale for visualisation

Dark grey fill: FI>Drop Out (more robust findings)

Light grey fill: FI <Drop Out (less robust findings)

Discussion

Conclusions from dichotomous comparison trials are usually informed by null hypothesis significance testing (NHST) and inferential probability (P values). Over relying on P value thresholds can increase the risk of false positive or false negative conclusions,(Colquhoun 2017) and FI was introduced to help clinicians assess the stability of clinical trials data. In the 45 RCTs included, we calculated a median FI of 4. It is noteworthy that trials reporting a statistically significant effect had a lower median FI (2), and either did not report attrition data or had dropout rates which exceeded the FI. This suggests that much of the data informing ankle sprain prevention are fragile, as adding (or removing) a small number of injury events to one of the trial's arms, changes its statistical significance (and thus, influences clinical decision-making).

Meta-research investigating the fragility of data in sub-specialities of medicine have variable findings. Audits of the spinal and critical care literature found a high level of statistically fragility, based on a

mean FI of just 2. (Evaniew *et al.* 2015; Ridgeon *et al.* 2016) More statistically stable trials (mean FI of 8) have been identified in high impact medical journals, (Walsh *et al.* 2014) with others reporting mean FI values as high as 13 and 26 in cardiology(Murad *et al.* 2022) and heart failure(Docherty *et al.* 2017) research trials respectively. The conclusions from studies examining fragility in musculoskeletal research are comparable to our current findings. Interventional trials involving patients with Achilles tendinopathy or Achilles rupture had mean FI values of 4.5(Xu *et al.* 2022) and 5,(Fackler *et al.* 2022) respectively, and a large audit of 102 trials published in highly indexed journals from the orthopaedic sports medicine literature (Journal of Bone and Joint Surgery (JBJS-Am) and the American Journal of Sports Medicine (AJSM)), reported a mean FI of 5 (interquartile range 3-8).(Parisien *et al.* 2019)

We also presented the relative fragility of trials using the fragility quotient (FQ). This is another simple, but important addition because FI is an absolute measure of stability and therefore unaffected by trial size.(Ahmed *et al.* 2016) Our median FQ of 0.015 is also low, suggesting that reversing events in less than 2 patients out of every 100 would alter significance. This is less robust than other fields of medicine and fragility analyses of the orthopaedic and surgical literature found median FQ's ranging from 0.022(Checketts *et al.* 2018) to 0.082(Doyle *et al.* 2022)

The reporting of P values in medical journal abstracts continues to increase, but few include supplementary effect size or uncertainty metrics. (Chavalarias *et al.* 2016) Interpreting statistically significant differences based on isolated P values is misleading. We have previously reported on the high risk of false-positive claims of treatment effectiveness in physiotherapy research,(Bleakley *et al.* 2021) which was largely underpinned by an over reliance on all-or-nothing hypothesis significance testing when interpreting clinical outcomes. Our current findings raise further questions about the integrity of evidence-based practice in this field, suggesting that higher standards of reporting and data interpretation are required. NHST remains central to determining treatment effectiveness, but it is most efficient in the context of long-run repeated testing. (Szucs and Ioannidis 2017) Clinicians who do not fully understand statistical concepts, such as power to detect estimate differences and attrition, are more likely to base their conclusions solely on P values. (Khan *et al.* 2020) Currently, one of the most highly cited papers(Hewett *et al.* 1999) in the sports medicine literature has a FI of 1. This means that a single event reversal in the treatment group (which recorded 8/463 non-contact knee injuries) or the

control group (0/366 non-contact knee injuries), would shift their results from significant (p=0.011) to nonsignificant. (p=0.086).

Fragility tends to be higher in RCTs with small samples sizes, or where the event of interest is rare. Completing a sample size calculation in the initial phases of RCT design (which also accounts for participant attrition) will generate more robust effect estimates. A recent audit of sports science research found that just 10% of experimental studies (12/120) included a formal sample size estimation.(Abt *et al.* 2020) In frequentist research, sample size is typically underpinned by power or the precision of effect estimates, although FI based calculations are now freely available. (Baer *et al.* 2021b)

The gold standard is that researchers clearly describe the flow of participants through each phase of a trial. (Schulz *et al.* 2010) Adherence to this recommendation varies, and some fields of medical research report that less than 60% of published studies adequately describe the numbers of participants receiving the intervention, or the numbers included in the final analysis. (Hopewell *et al.* 2011) Expectations for attrition can depend on many factors, including the study population, duration of follow-up, and the event rate.(Schulz and Grimes 2002) We found that inadequate reporting of attrition was most likely to occur in large multisite trials focusing on primary prevention of LAS. It may be more difficult to establish and implement optimum procedures for minimising losses in such designs. Others suggest that participant retention is more likely when studies incorporate clear and transparent details in consenting documents, with study coordinators maintaining regular and consistent contact with participants and care providers. (Bedlack and Cudkowicz 2009) Poor retention may also be more likely if a treatment intervention has no perceived benefit (eg. a passive control), suggesting that where possible, RCTs should incorporate an active control or usual care. (Page and Persch 2013)

Attrition bias in Physical Therapy research is often quantified by the absolute loss to follow-up rate. (*PEDro.*) Although an acceptable rate for drop out is unclear, some suggest that a rate exceeding 20% significant challenges study validity.(Sackett, DL Straus, SE Richardson, WS Rosenberg, W Haynes, RB 2000) To gain further context on the stability of each trial we compared attrition with FI. In most trials, the numbers lost to follow-up exceeded the FI. This pattern raises concern, as those lost to

follow-up, could have been the patients with a different outcome, thus changing the statistical significance of the difference between study arms. Our data align with audits of sports medicine journals(Parisien *et al.* 2019) where the average loss to follow-up per trial (n=7.9) exceeded the average FI of 5. These patterns highlight the importance of presenting participant attrition figures, alongside FI. Adding a FIDO chart (Figure 1), which plots Fragility Index against Drop Out, is a simple way to summarise the robustness of dichotomous outcome data, and could be a useful addition for systematic reviews.

Limitations

FI has no known thresholds at which the results would be considered robust. We considered RCT's with a low FI to be less stable. This is based on the logic that a study is more susceptible to random error and erroneous misclassification of outcomes, if a small number of event reversals would alter its statistical significance. Although the relationship between FI and effect precision is also unclear, a recent meta-epidemiological study suggested that FI values <19 are highly susceptible to chance and should be interpreted with caution.(Murad *et al.* 2022) We also acknowledge that there are many other trial conditions affecting the validity of conclusions, including lack of registration, HARKING, p-hacking or false discoveries relating to other multiplicity issues (eg. analysis of multiple outcomes).(Li *et al.* 2017)

Some of the included RCTs analysed their LAS counts using time to event techniques, either with or without covariates. As our calculations were based on the more basic Fisher exact test, it is possible that the FI may be overly fragile, in trials where the events are similar in each group, but the timing of events is different. (Khan *et al.* 2020) To check this, we undertook a sensitivity analysis, and found no differences between included trials using time-to-event primary end points vs frequency data. We also calculated incidence fragility indices for each trial, and confirmed FI stability in all but 1 case.

FI has been criticized for being a restatement of a P value(Carter *et al.* 2017) and we acknowledge that P values and fragility indices are highly correlated (negatively).(Khan *et al.* 2020) P values and FI are both measures of evidence against the null hypothesis, but as the former are presented in units of probability, they are commonly mis-interpreted by clinicians, researchers, and patients. As the FI unit is 'patients', it is immediately interpretable for a clinical audience and gives a clearer metric to inform one's confidence in a study's results. Consistent reporting of FI could help to improve research culture,

by highlighting the relative drawbacks of using P value thresholds in isolation.(Khan *et al.* 2020) As most clinical research will continue to be underpinned by frequentist approaches, FI remains a simple and intuitive way to communicate findings to clinicians or the public, and represents an excellent adjunct to P values, confidence intervals and related precision judgements.(Murad *et al.* 2022) Finally this study was not preregistered; the decision to include one of the secondary objectives (examining if FI was influenced by the type of analysis: frequency data vs time-to-event) emerged during the analysis, rather than a priori.

Conclusion

Level 1 evidence informing the prevention of ankle sprains is fragile and susceptible to random error. The median FI of RCTs in this field was just 4, meaning that the statistical significance of most studies would be altered by a very small percentage of event reversals. Of further concern is that in three quarters of studies, the number of drop outs exceeded the FI. Reporting P values in conjunction with FI, provides a more intuitive method for interpreting the clinical stability of study data. Additionally, we recommend the use of incidence fragility indices, since these account for the likelihood of an event reversal.

Funding statement: No funding

References.

Https://Clincalc.com/stats/FragilityIndex.aspx. Available at: https://clincalc.com/Stats/FragilityIndex.aspx

Https://Github.com/brb225/FragilityTools. Available at: https://github.com/brb225/FragilityTools [Accessed November 2022].

PEDro.

Abt, G., Boreham, C., Davison, G., Jackson, R., Nevill, A., Wallace, E. and Williams, M. (2020) Power, precision, and sample size estimation in sport and exercise science research. *Journal of Sports Sciences,* 38(17), 1933-1935.

Ahmed, W., Fowler, R.A. and McCredie, V.A. (2016) Does sample size matter when interpreting the fragility index? *Critical Care Medicine,* 44(11), e1142-e1143.

Baer, B.R., Gaudino, M., Charlson, M., Fremes, S.E. and Wells, M.T. (2021a) Fragility indices for only sufficiently likely modifications. *Proceedings of the National Academy of Sciences of the United States of America,* 118(49), 10.1073/pnas.2105254118.

Baer, B.R., Gaudino, M., Fremes, S.E., Charlson, M. and Wells, M.T. (2021b) The fragility index can be used for sample size calculations in clinical trials. *Journal of Clinical Epidemiology,* 139, 199-209.

Bedlack, R.S. and Cudkowicz, M.E. (2009) Clinical trials in progressive neurological diseases. recruitment, enrollment, retention and compliance. *Frontiers of Neurology Neuroscience,* 25, 144-151.

Bellows, R. and Wong, C.K. (2018) The effect of bracing and balance training on ankle sprain incidence among athletes: A systematic review with meta-analysis. *International Journal of Sports Physical Therapy,* 13(3), 379-388.

Bleakley, C.M., Matthews, M. and Smoliga, J.M. (2021) Most ankle sprain research is either false or clinically unimportant: A 30-year audit of randomized controlled trials. *Journal of Sport and Health Science,* 10(5), 523-529.

Buttner, F., Toomey, E., McClean, S., Roe, M. and Delahunt, E. (2020) Are questionable research practices facilitating new discoveries in sport and exercise medicine? the proportion of supported hypotheses is implausibly high. *British Journal of Sports Medicine,* 54(22), 1365-1371.

Carter, R.E., McKie, P.M. and Storlie, C.B. (2017) The fragility index: A P-value in sheep's clothing? *European Heart Journal,* 38(5), 346-348.

Chavalarias, D., Wallach, J.D., Li, A.H. and Ioannidis, J.P. (2016) Evolution of reporting P values in the biomedical literature, 1990-2015. *Jama,* 315(11), 1141-1148.

Checketts, J.X., Scott, J.T., Meyer, C., Horn, J., Jones, J. and Vassar, M. (2018) The robustness of trials that guide evidence-based orthopaedic surgery. *The Journal of Bone and Joint Surgery.American Volume,* 100(12), e85.

Colquhoun, D. (2017) The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science,* 4(12), 171085.

de Vasconcelos, G.S., Cini, A., Sbruzzi, G. and Lima, C.S. (2018) Effects of proprioceptive training on the incidence of ankle sprain in athletes: Systematic review and meta-analysis. *Clinical Rehabilitation,* 32(12), 1581-1590.

Docherty, K.F., Campbell, R.T., Jhund, P.S., Petrie, M.C. and McMurray, J.J.V. (2017) How robust are clinical trials in heart failure? *European Heart Journal,* 38(5), 338-345.

Doherty, C., Bleakley, C., Delahunt, E. and Holden, S. (2017) Treatment and prevention of acute and recurrent ankle sprain: An overview of systematic reviews with meta-analysis. *British Journal of Sports Medicine,* 51(2), 113-125.

Doyle, T.R., Davey, M.S. and Hurley, E.T. (2022) The statistical fragility of management options for acute achilles tendon ruptures - A systematic review of randomized control trial with fragility analysis. *Journal of ISAKOS : Joint Disorders & Orthopaedic Sports Medicine,* 7(4), 72-81.

Evaniew, N., Files, C., Smith, C., Bhandari, M., Ghert, M., Walsh, M., Devereaux, P.J. and Guyatt, G. (2015) The fragility of statistically significant findings from randomized trials in spine surgery: A systematic survey. *The Spine Journal : Official Journal of the North American Spine Society,* 15(10), 2188-2197.

Fackler, N.P., Karasavvidis, T., Ehlers, C.B., Callan, K.T., Lai, W.C., Parisien, R.L. and Wang, D. (2022) The statistical fragility of operative vs nonoperative management for achilles tendon rupture: A systematic review of comparative studies. *Foot & Ankle International,* , 10711007221108078.

Feinstein, A.R. (1990) The unit fragility index: An additional appraisal of "statistical significance" for a contrast of two proportions. *Journal of Clinical Epidemiology,* 43(2), 201-209.

Gribble, P.A., Bleakley, C.M., Caulfield, B.M., Docherty, C.L., Fourchet, F., Fong, D.T., Hertel, J., Hiller, C.E., Kaminski, T.W., McKeon, P.O., Refshauge, K.M., Verhagen, E.A., Vicenzino, B.T., Wikstrom, E.A. and Delahunt, E. (2016) Evidence review for the 2016 international ankle consortium consensus statement on the prevalence, impact and long-term consequences of lateral ankle sprains. *British Journal of Sports Medicine,* 50(24), 1496-1505.

Heidt, R.S.J., Sweeterman, L.M., Carlonas, R.L., Traub, J.A. and Tekulve, F.X. (2000) Avoidance of soccer injuries with preseason conditioning. *The American Journal of Sports Medicine,* 28(5), 659-662.

Hewett, T.E., Lindenfeld, T.N., Riccobene, J.V. and Noyes, F.R. (1999) The effect of neuromuscular training on the incidence of knee injury in female athletes. A prospective study. *The American Journal of Sports Medicine,* 27(6), 699-706.

Hootman, J.M., Dick, R. and Agel, J. (2007) Epidemiology of collegiate injuries for 15 sports: Summary and recommendations for injury prevention initiatives. *Journal of Athletic Training,* 42(2), 311-319.

Hopewell, S., Hirst, A., Collins, G.S., Mallett, S., Yu, L. and Altman, D.G. (2011) Reporting of participant flow diagrams in published reports of randomized trials. *Trials,* 12, 253-253.

Ioannidis, J.P. (2005) Why most published research findings are false. *PLoS Medicine,* 2(8), e124.

Kemler, E., van de Port, I., Backx, F. and van Dijk, C.N. (2011) A systematic review on the treatment of acute ankle sprain: Brace versus other functional treatment types. *Sports Medicine (Auckland, N.Z.),* 41(3), 185-197.

Khan, M.S., Fonarow, G.C., Friede, T., Lateef, N., Khan, S.U., Anker, S.D., Harrell, F.E. and Butler, J. (2020) Application of the reverse fragility index to statistically nonsignificant randomized clinical trial results. *JAMA Network Open,* 3(8), e2012469.

Li, G., Taljaard, M., Van den Heuvel, E.R., Levine, M.A., Cook, D.J., Wells, G.A., Devereaux, P.J. and Thabane, L. (2017) An introduction to multiplicity issues in clinical trials: The what, why, when and how. *International Journal of Epidemiology,* 46(2), 746-755.

Mohammadi, F. (2007) Comparison of 3 preventive methods to reduce the recurrence of ankle inversion sprains in male soccer players. *The American Journal of Sports Medicine,* 35(6), 922-926.

Murad, M.H., Kara Balla, A., Khan, M.S., Shaikh, A., Saadi, S. and Wang, Z. (2022) Thresholds for interpreting the fragility index derived from sample of randomised controlled trials in cardiology: A meta-epidemiologic study. *BMJ Evidence-Based Medicine,*

Page, S.J. and Persch, A.C. (2013) Recruitment, retention, and blinding in clinical trials. *The American Journal of Occupational Therapy : Official Publication of the American Occupational Therapy Association,* 67(2), 154-161.

Parisien, R.L., Ehlers, C., Cusano, A., Tornetta, P., Li, X. and Wang, D. (2021) The statistical fragility of platelet-rich plasma in rotator cuff surgery: A systematic review and meta-analysis. *The American Journal of Sports Medicine,* 49(12), 3437-3442.

Parisien, R.L., Trofa, D.P., Dashe, J., Cronin, P.K., Curry, E.J., Fu, F.H. and Li, X. (2019) Statistical fragility and the role of P values in the sports medicine literature. *The Journal of the American Academy of Orthopaedic Surgeons,* 27(7), e324-e329.

Ridgeon, E.E., Young, P.J., Bellomo, R., Mucchetti, M., Lembo, R. and Landoni, G. (2016) The fragility index in multicenter randomized controlled critical care trials. *Critical Care Medicine,* 44(7), 1278-1284.

Sackett, DL Straus, SE Richardson, WS Rosenberg, W Haynes, RB. (2000) Edinburgh: Churchill Livingstone.

Schiftan, G.S., Ross, L.A. and Hahne, A.J. (2015) The effectiveness of proprioceptive training in preventing ankle sprains in sporting populations: A systematic review and meta-analysis. *Journal of Science and Medicine in Sport,* 18(3), 238-244.

Schulz, K.F., Altman, D.G. and Moher, D. (2010) CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology & Pharmacotherapeutics,* 1(2), 100-107.

Schulz, K.F. and Grimes, D.A. (2002) Sample size slippages in randomised trials: Exclusions and the lost and wayward. *Lancet (London, England),* 359(9308), 781-785.

Szucs, D. and Ioannidis, J.P.A. (2017) When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience,* 11, 390.

Taylor, J.B., Ford, K.R., Nguyen, A.D., Terry, L.N. and Hegedus, E.J. (2015) Prevention of lower extremity injuries in basketball: A systematic review and meta-analysis. *Sports Health,* 7(5), 392-398.

Tignanelli, C.J. and Napolitano, L.M. (2019) The fragility index in randomized clinical trials as a means of optimizing patient care. *JAMA Surgery,* 154(1), 74-79.

Wagemans, J., Bleakley, C., Taeymans, J., Schurz, A.P., Kuppens, K., Baur, H. and Vissers, D. (2022) Exercise-based rehabilitation reduces reinjury following acute lateral ankle sprain: A systematic review update with meta-analysis. *PloS One,* 17(2), e0262023.

Walsh, M., Devereaux, P.J. and Sackett, D.L. (2015) Clinician trialist rounds: 28. when RCT participants are lost to follow-up. part 1: Why even a few can matter. *Clinical Trials (London, England),* 12(5), 537-539.

Walsh, M., Srinathan, S.K., McAuley, D.F., Mrkobrada, M., Levine, O., Ribic, C., Molnar, A.O., Dattani, N.D., Burke, A., Guyatt, G., Thabane, L., Walter, S.D., Pogue, J. and Devereaux, P.J. (2014) The statistical significance of randomized controlled trial results is frequently fragile: A case for a fragility index. *Journal of Clinical Epidemiology,* 67(6), 622-628.

Xu, A.L., Ortiz-Babilonia, C., Gupta, A., Rogers, D., Aiyer, A.A. and Vulcano, E. (2022) The statistical fragility of platelet-rich plasma as treatment for chronic noninsertional achilles tendinopathy: A systematic review and meta-analysis. *Foot & Ankle Orthopaedics,* 7(3), 24730114221119758.

Supplementary files

1. **Search strategy**
2. **Data file**
3. **PRISMA flow**