■ ANNOTATION

What is a fragility index?

FRAGILITY AND REVERSE FRAGILITY INDEX FOR ASSESSING THE SIGNIFICANCE OF RESULTS FROM PUBLISHED RANDOMIZED CONTROLLED TRIALS

Cite this article: ***Bone Joint J*** **2024;106-B(4):319–322.**

N. Parsons,
M. R. Whitehouse,
M. L. Costa

From University of Warwick, Coventry, UK

The fragility index is an increasingly used tool for assessing the robustness of results obtained from randomized clinical trials (RCTs). Several recent papers have reported on the 'fragility' of conclusions from systematic reviews of RCTs in areas of trauma and orthopaedic surgery, including knee, shoulder, and hip arthroplasty, Achilles tendon rupture, and the use of platelet-rich plasma in the treatment of musculoskeletal conditions.[1-6] These systematic reviews base their critique of the existing RCT evidence on the apparent 'statistical fragility', 'fragility of statistically significance', or just plain 'fragility' of the trials. But what is a fragility index and are these criticisms legitimate? We describe the fragility index, show how it is calculated and interpreted, and provide some thoughts on its use and limitations and some guidance on reporting.

The fragility index is defined as the minimum number of participants in a trial reporting a positive result who would need to have had a different outcome for the results of the trial to lose statistical significance.[7] It is applied in trials that base their conclusions on binary outcomes, such as mortality or the need for revision surgery, where each trial participant can only have the answer 'Yes' or 'No' in terms of their outcome (i.e. the participant did or did not need revision surgery, or did or did not survive treatment). The fragility index is not applied to continuous or time-to-event outcomes. The fragility index can be calculated by adding one event (Yes) and subtracting one non-event (No) to the treatment group with the smaller number of events. The statistical test (e.g. Fisher's exact test) is then recalculated to determine if the result remains statistically significant. If it does, the process is repeated to add and subtract further events from the group with the smaller number of events until the p-value from the statistical test no longer shows significance (Table I). The fragility index is the number of participants that has to change from an event to a non-event for the threshold of statistical significance to be breached. In situations where it is not necessary to move a non-event to an event to change from a statistically significant to a non-significant result, the fragility index has a value of zero.[8] This might occur, for instance, if the incorrect statistical test had been used (e.g. if a chi-squared test were changed to Fisher's exact test, where the latter is preferred to the former for small sample sizes). The fragility index can therefore be considered to be a 'what if something different had happened' analysis with conclusions drawn on the basis of how many times that would need to happen to change the result.

A lower fragility index value is interpreted as higher fragility of the trial's results; for example, a fragility index equal to five means that if five study participants in the effective treatment arm had a different outcome, then the trial would no longer have a statistically significant result. By extension, the reverse fragility index calculates the minimum number of participants who would need to experience a different outcome for the binary endpoint to change a statistically insignificant association to an association that demonstrates statistical significance.[9]

It is important to note that the fragility index takes no account of the total study sample size. It has been suggested that a relative measure, the fragility quotient, defined to be the fragility index divided by the study sample size, should also be reported.[10] The fragility quotient (fragility index/n) is interpreted in a similar manner to the fragility index but presented on a standardized scale, allowing comparisons across clinical trials of different sizes to be made more directly. The fragility quotient is in part a means to provide a measure that is easier to interpret and characterize than the fragility index.

We do not believe that there is a specific value (threshold) or range of values for the fragility index or quotient that defines a RCT as fragile or robust, or that such a concept would be useful or advantageous.[8] However, the fragility index can be compared to reported data missingness and drop-out rates and patterns in a trial as a means of understanding how study inferences

Correspondence should be sent to M. L. Costa; email: matthew.costa@ndorms.ox.ac.uk

doi:10.1302/0301-620X.106B4.BJJ-2023-1043.R1 $2.00

Bone Joint J
2024;106-B(4):319–322.

Table I. The p-values for Fisher's exact test comparing the number of events for trials with decreasing numbers of treatment group events (from one to five additional events) for a notional trial with two intervention arms (n = 200 participants per arm) with trial significance set at the 5% level (p = 0.05). The FI = 5 because at the fifth iteration p = 0.056.

Trial	Events, n				p-value
	Treatment		Control		
	Yes	No	Yes	No	
Original trial	75	125	100	100	0.015
1	75 + 1	125 - 1	100	100	0.020
2	75 + 2	125 - 2	100	100	0.027
3	75 + 3	125 - 3	100	100	0.035
4	75 + 4	125–4	100	100	0.044
5	75 + 5	125–5	100	100	0.056

and conclusions might be changed under alternate scenarios in sensitivity analyses. Sensitivity analyses are secondary analyses to explore how outcome data may change under different assumptions during a trial, a common example being a per protocol analysis (analysis according to the treatment that a participant actually received) compared with the commonly used primary intention-to-treat analysis (where participants are analyzed according to the treatment to which they were randomly allocated).

There has been debate within both the trials and medical statistics communities on the merits of the fragility index (and reverse fragility index),[11] with evidence, via simulation studies, showing that the fragility index is strongly associated (highly correlated) with the p-value,[12] and as such we should be careful, as we are with a p-value, not to conflate the index, or fragility in general, with a measure of the strength of the treatment effect, independent of other aspects of the study. The fragility index, like the p-value, is ultimately related to the trial sample size,[13] and therefore we might argue the fragility index seems to offer little in addition to what we conventionally report with the standard Consolidated Standards of Reporting Trials (CONSORT) guidelines.[14] If we know the trial sample size, event rate, and p-value, the fragility index may seem superfluous.

In general, the fragility index can, and should, be reported when assessing the robustness of reported results from RCTs. However, we should aways bear in mind a number of important issues and be clear on these when reporting on the fragility and robustness of results of published RCTs. First, it is important to be clear and make a distinction between the fragility of binary outcomes that are the primary outcome of a trial, and binary outcomes that are reported as secondary outcomes. Most RCTs will routinely provide data on, for instance, mortality and report differences between intervention arms (e.g. chi-squared or Fisher's exact tests), irrespective of what the primary outcomes of the studies were. However, we should consider the hypotheses under test in a RCT and the inferences that we can legitimately make. RCTs will only (in almost all cases) be powered to detect a difference between intervention arms for the primary outcome. Therefore, analyses of secondary outcomes will often have low power, particularly binary outcomes that require large sample sizes. Therefore, for any RCT being assessed, it must be absolutely clear what the primary outcome was, because that affects how the fragility index result is interpreted. If the fragility is high for a RCT where mortality is the primary outcome, we should rightly be cautious about overinterpreting the results. However, if mortality is a secondary outcome, then we can often say little other than there was very little information on mortality from the study. It is very unlikely that a properly reported RCT would claim to be powered for a secondary outcome measure where this was not explicitly part of the a priori study design. In the context of binary outcomes included as secondary outcome measures, it is highly likely that the study team considered the appropriateness of inclusion of each of their outcome measures and have therefore reached the conclusion in advance that the binary outcome was not suitable as a primary or co-primary outcome measure. This may be due to the relative appropriateness of different outcomes to measure the difference in effect between intervention and control, which outcomes patients consider to be of primary importance, or consideration of attainable and meaningful samples sizes. The decision not to use a binary secondary outcome as the primary is therefore likely to have been strongly influenced by very sound methodological considerations, i.e. because the study would lack power and be 'fragile'.

As an example trial to elucidate a number of important issues, take the recently published World Hip Trauma Evaluation 5 (WHiTE5) trial comparing cemented to uncemented hemiarthroplasty for intracapsular hip fractures, in which we were involved, which reported 146 and 171 deaths in the cemented (n = 610) and uncemented (n = 615) groups, respectively.[15] Fisher's exact test gives p = 0.134, suggesting no evidence for a difference in death rates between cemented (23.9%) and uncemented (27.8%) groups. To attain statistical significance (to make the p-value < 0.05), only eight participants in the uncemented groups would need to provide a different outcome, i.e. reverse fragility index = 8, which has been reported previously as a median value for fragility index across a range of trials.[7] The relatively small number of participants that would need to have different outcomes for this study to have a different statistically significant result seems concerning, and may cause us to question the results more generally. However, we should not lose sight of the fact that the primary outcome for the WHiTE5 was not mortality, but health-related quality of life; therefore, mortality was analyzed and reported as a secondary outcome only (odds ratio (OR) 0.80 (95% confidence interval (CI) 0.62 to 1.05). Clearly there was little power in the study to detect differences in mortality, as the confidence interval for the OR is wide. However, the difference in death rates reported in the trial is consistent with the results from the primary outcome, which were significant. Therefore, can we just dismiss this as a difference that is 'suspect' or 'fragile' because there was little precision? The OR is likely to be an unbiased estimate of the treatment effect, under some weak assumptions about patterns of data missingness, despite the low precision, and suggests that mortality was lower in the cemented group, which is supported by the main study result. If the WHiTE5 trial had been designed to look for differences in mortality between cemented and uncemented groups, and they had known a priori that death rates were approximately 24% (cemented) and 28% (uncemented), then they would likely have needed to recruit at

least 5,000 patients to have 90% power to detect a difference in death rates of 4%, a more than four-fold increase in the trial size. Under the same assumed death rates as the WHiTE5 trial, this four-fold larger mortality-powered trial would have likely reported 584 (23.9%) and 684 (27.8%) deaths in the cemented (n = 2440) and uncemented (n = 2,460) groups respectively, with p = 0.002 suggesting that death rates in the cemented group were significantly lower than in the uncemented group. To make the p-value larger than 0.05, for this four-fold larger mortality-powered trial, requires 36 participants to change outcomes, i.e. fragility index = 36. This is now a large value of the index and suggests that the four-fold larger mortality-powered WHiTE5 trial is not fragile, but rather extremely robust. What this shows is that fragility measures, independent of other facts pertaining to the design and conduct of a RCT, must be treated with due caution, particularly when misapplied to a secondary outcome measure. It is important to know whether a trial is designed and powered based on a primary binary outcome, or whether a binary outcome is reported for routine, possibly ancillary, reasons such that directly linking the fragility of the outcome to the overall fragility of the trial will seem unduly harsh and overly critical. When undertaking systematic reviews of trials, we would advise that the fragility index and reverse fragility index should be used as a means of assessing fragility only for and between trials where the primary outcome is binary.

More generally, we would argue that taken on its own, the fragility index, like the p-value, rests too heavily on arbitrarily pre-stated simple thresholds for statistical significance and the frequentist approach to analysis, e.g. in Table I the result is significant at the 5% level if p = 0.044 but not so if p = 0.056. Much has been written about the null hypothesis statistical testing paradigm and its assumptions of a single test hypothesis and the use of a dichotomized approach to rejecting the null hypothesis on the basis of a single study.[16] Many of the reported critiques of trials based on the fragility index are really more of a general criticism of null hypothesis statistical testing, and indeed any kind of interpretation based on simple thresholding. If the number of events is small, then if one event is changed or missed from one or other intervention arm in a RCT, the inferences may change quite dramatically, and we can move from a position of statistical significance to one of non-significance. For this reason, we would advocate large RCTs with high event rates in order to generate robust and reliable evidence.[9] We should not use arguments based on fragility to assess the value of a RCT in isolation of other facts pertaining to the intent, conduct, and design of the RCT. Many factors must be considered, not just the fragility or robustness of a particular measure, when we critique a RCT. The fragility index clearly has value but can be legitimately criticized on the basis that it "highlights one minor sensitivity analysis at the expense of other features of trial design and quality assessment."[11] Good clinical, scientific, and statistical practice should be followed both in the conduct and reporting of RCTs, and we should always remember that no single index can ever replace scientific reasoning.[17]

Take home message

- The fragility index, the minimum number of participants required to change their outcomes for the results of a trial to lose statistical significance, is a useful metric for assessing the robustness of results from randomized controlled trials (RCTs).
- However, a fragility index, independent of other facts pertaining to the design and conduct of a RCT, should not be interpreted as a measure of the strength of an intervention effect and must be treated with due caution, particularly when applied to a secondary outcome measure.

References

1. **Ruzbarsky JJ**, **Rauck RC**, **Manzi J**, **Khormaee S**, **Jivanelli B**, **Warren RF**. The fragility of findings of randomized controlled trials in shoulder and elbow surgery. *J Shoulder Elbow Surg*. 2019;28(12):2409–2417.
2. **Ekhtiari S**, **Gazendam AM**, **Nucci NW**, **Kruse CC**, **Bhandari M**. The fragility of statistically significant findings from randomized controlled trials in hip and knee arthroplasty. *J Arthroplasty*. 2021;36(6):2211–2218.
3. **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.
4. **Parisien RL**, **Ehlers C**, **Cusano A**, **Tornetta P**, **Li X**, **Wang D**. The statistical fragility of platelet-rich plasma in rotator cuff surgery: a systematic review and meta-analysis. *Am J Sports Med*. 2021;49(12):3437–3442.
5. **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.
6. **Cordero JK**, **Lawrence KW**, **Brown AN**, **Li X**, **Hayden BL**, **Parisien RL**. The fragility of tourniquet use in total knee arthroplasty: a systematic review of randomized controlled trials. *J Arthroplasty*. 2023;38(6):1177–1183.
7. **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.
8. **Tignanelli CJ**, **Napolitano LM**. The fragility index in randomized clinical trials as a means of optimizing patient care. *JAMA Surg*. 2019;154(1):74–79.
9. **Khan MS**, **Fonarow GC**, **Friede T**, **et al**. Application of the reverse fragility index to statistically nonsignificant randomized clinical trial results. *JAMA Netw Open*. 2020;3(8):e2012469.
10. **Ahmed W**, **Fowler RA**, **McCredie VA**. Does sample size matter when interpreting the fragility index? *Crit Care Med*. 2016;44(11):e1142–e1143.
11. **Potter GE**. Dismantling the fragility index: a demonstration of statistical reasoning. *Stat Med*. 2020;39(26):3720–3731.
12. **Carter RE**, **McKie PM**, **Storlie CB**. The fragility index: a p-value in sheep's clothing? *Eur Heart J*. 2017;38(5):346–348.
13. **Porco TC**, **Lietman TM**. A fragility index: handle with care. *Ophthalmology*. 2018;125(5):649.
14. **Schulz KF**, **Altman DG**, **Moher D**, **CONSORT Group**. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c332.
15. **Fernandez MA**, **Achten J**, **Parsons N**, **et al**. Cemented or uncemented hemiarthroplasty for intracapsular hip fracture. *N Engl J Med*. 2022;386(6):521–530.
16. **Wasserstein RL**, **Schirm AL**, **Lazar NA**. Moving to a world beyond "p < 0.05". *Am Stat*. 2019;73.
17. **Wasserstein RL**, **Lazar NA**. The ASA Statement on *p*-values: context, process, and purpose. *Am Stat*. 2016;70(2):129–133.

Author information:

N. Parsons, PhD, Professor (Research Focused), Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK.

M. R. Whitehouse, PhD, Professor of Trauma & Orthopaedics, Bristol Medical School, University of Bristol, Beacon House, Bristol, UK.

M. L. Costa, PhD, Professor of Orthopaedic Trauma, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, UK.

Author contributions:

N. Parsons: Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft.
M. R. Whitehouse: Conceptualization, Methodology, Formal analysis, Visualization, Writing – review & editing.

M. L. Costa: Conceptualization, Methodology, Formal analysis, Visualization, Writing – review & editing.

Funding statement:
The authors received no financial or material support for the research, authorship, and/or publication of this article.

ICMJE COI statement:
N. Parsons reports payments from the National Institute for Health Research (NIHR), via Warwick Medical School, related to this study. M. R. Whitehouse reports funding from the NIHR via the Bristol Biomedical Research Centre, related to this study, a number of grants from the NIHR, unrelated to this study, royalties or licenses from Taylor & Francis, payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Heraeus, sits on or is chair of a number of Trial Steering Committees or Data Monitoring Committees for trials funded by NIHR, and reports leadership or fiduciary role in other board, society, committee or advocacy group, paid or unpaid for the British Hip Society, the British Orthopaedic Association, and the NIHR CRN, all of which are unrelated to this article. M. L. Costa reports grant funding, via the University of Oxford, from NIHR and Wellcome for research into Musculoskeletal Trauma, unrelated to this study.

Data sharing:
All data generated or analyzed during this study are included in the published article and/or in the supplementary material.

This article was primary edited by M. Hossain.