Research methods and reporting

Thresholds for interpreting the fragility index derived from sample of randomised controlled trials in cardiology: a meta-epidemiologic study

Mohammad Hassan Murad [1,2] Abdalla Kara Balla,[2] Muhammad Shahzeb Khan,[3] Asim Shaikh,[4] Samer Saadi,[2] Zhen Wang [2]

10.1136/bmjebm-2021-111858

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjebm-2021-111858).

[1]Division of Public Health, Infectious Diseases and Occupational Medicine, Mayo Clinic, Rochester, MN, USA
[2]Evidence-based Practice Center, Mayo Clinic, Rochester, Minnesota, USA
[3]Division of Cardiology, Duke University School of Medicine, Durham, North Carolina, USA
[4]Department of Internal Medicine, Dow University of Health Sciences, Karachi, Pakistan

Correspondence to:
Dr Mohammad Hassan Murad, Public Health, Infectious Diseases and Occupational Medicine, Mayo Clinic, Rochester, Minnesota, 55905, USA; murad.mohammad@mayo.edu

To cite: Murad MH, Kara Balla A, Khan MS, *et al*. *BMJ Evidence-Based Medicine* 2023;**28**:133–136.

Abstract

The fragility index (FI) was proposed as a simplified way to communicate robustness of statistically significant results and their susceptibility to a change of a handful number of events. While this index is intuitive, it is not anchored by a cut-off or a guide for interpretation. We identified cardiovascular trials published in six high impact journals from 2007 to 2021 (500 or more participants and a dichotomous statistically significant primary outcome). We estimated area under curve (AUC) to determine FI value that best predicts whether the treatment effect was precise, defined as adequately powered for a plausible relative risk reduction (RRR) of 25% or 30% or having a CI that is sufficiently narrow to exclude a risk reduction that is too small (close to the null, <0.05). The median FI of 201 included cardiovascular trials was 13 (range 1–172). FI exceeded the number of patients lost to follow-up in 46/201 (22.89%) trials. FI values of 19 and 22 predicted that trials would be precise (powered for RRR of 30% and 25%; respectively, combined with CI that excluded risk reduction <0.05). AUC for meeting these precision criteria was 0.90 (0.86–0.94). In conclusion, FI values that range 19–22 may meet various definitions of precision and can be used as a rule of thumb to suggest that a treatment effect is likely precise and less susceptible to random error. The number of patients lost to follow-up should be presented alongside FI to better illustrate fragility.

Background

Due to many limitations and common misinterpretations of the p value,[1] the fragility index (FI) has been suggested as an easier, more intuitive way to communicate results to clinicians and other stakeholders.[2] The FI is defined as the minimum number of patients whose status would have to change from a non-event to event to turn a statistically significant result to a non-significant result. Thus, a randomised controlled trial (RCT) with statistically significant results that has an FI of 1 would lose significance even if one patient had the opposite outcome. FI was not intended to replace the p value, CI or precision judgements. Rather, it is intended to be a simple intuitive way to communicate findings to clinicians or the public.

A previous study evaluated cardiovascular RCTs with sample sizes over 500 participants that had a statistically significant primary outcome and showed a median FI of 13 (IQR, 5–26).[3] While intuitively one can think of an RCT outcome with FI of 1 or 2 to be less reliable, that is, susceptible to random error and erroneous misclassification of outcomes, it is not clear how to interpret FI of 5 or 6, for example. Thus, the lack of established cut-off or guide to aid in the interpretation of FI adds to some previously described[4–6] interpretational challenges.

Furthermore, modern frameworks of rating the certainty of evidence such as Grading of Recommendations, Development, Assessment and Evaluation (GRADE)[7] do not depend on statistical significance or the resultant calculation of FI. GRADE suggests that even if an estimate was statistically significant, it will not be considered precise (ie, robust or less prone to chance) unless it was derived from a body of evidence with a sample size that is adequate to detect a plausible relative risk reduction (RRR). GRADE suggests using RRR of 25%–30% for this estimation.[8] In addition to sample size considerations, GRADE suggests that judgements about precision should also consider whether the CI did not overlap a decision-making threshold that is considered to be trivial or unimportant.[8] Therefore, if the upper boundary of a relative risk is very close to the null or crosses a decision-making threshold, the results may still be considered imprecise despite statistical significance.

Considering the lack of anchors for FI and the lack of clarity about the relationship between FI and precision, we aimed to empirically evaluate FI in cardiovascular RCTs and study the association with precision. To date, this has not been studied and precision cannot be deduced from FI. Providing clinicians and other stakeholders with FI values that are likely to be associated with precise and reliable estimates can help them make judgements about certainty and trustworthiness of estimates.

Methods

This meta-epidemiological study follows the reporting guidance for methodology research.[9] A reporting checklist is provided in the online supplemental appendix. This study is a previously published protocol.[3] Since publicly available data were used, institutional review board approval was not applicable.

Data sources

Journals were selected for the present study based on a combination of the following features: impact factor, readership, specialisation in publication of cardiovascular RCTs and global recognition for consistent publication of influential RCTs over the last several decades. *The New England Journal of Medicine*, *The Lancet* and *Journal of the American Medical Association* were selected for having the highest impact factors in general medicine, while *Journal of the American College of Cardiology*, *European Heart Journal* and *Circulation* were selected for having the highest impact factors in the field of cardiovascular medicine. The rationale for targeting randomised trials with a sample size >500 and published in these specific journals was that we aimed to evaluate robustness in trials that were more likely to impact practice. We updated a previously published[3] search strategy through 13 September 2021. Details of the search strategy are available in the online supplemental appendix.

Study eligibility and data extraction

All RCTs were assessed for inclusion from the three cardiovascular journals whereas RCTs from the three non-cardiovascular journals were screened for determination of possible cardiovascular nature (if the interventions or outcomes were described as cardiovascular, such as those in the disciplines of heart failure, interventional cardiology, preventive cardiology, electrophysiology, cardiac imaging or stroke). Additional inclusion criteria were: (1) phase 3 or 4 RCT; (2) sample size ≥500 patients (an arbitrary cut-off to identify larger RCTs that are more likely to impact practice); (3) parallel arm study design and (4) at least one statistically significant binary outcome. Data were extracted in a pre-designed form. Study selection and data abstraction were performed by one reviewer (AKB) and verified by a second reviewer (SS). Discrepancies were reconciled by a third reviewer (MHM). Data were extracted using pre-defined forms that were pilot tested and included trials first author, year of publication, journal, impact factor, number of centres, country, a 2×2 table for the main outcome, number of patients lost to follow-up, funding, intervention type and control type.

Outcome measures

We evaluated the FI value that best predicts a precise treatment effect. Results were reported as the FI cut-off values and associated sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC).

Precision thresholds and definitions

Two methods were used to define a precise treatment effect: (1) whether an RCT was adequately powered for a RRR (25% or 30%); (2) whether the CI of the treatment effect was sufficiently narrow to exclude a small or trivial risk reduction of 0.05. The RRR thresholds of 25%–30% were recommended by the GRADE Working Group.[8] Precision guidance published in 2011 stated that although determining a threshold for adequate power is a matter of judgement and can change based on context, RRR 25%–30% can be considered a moderate or plausible RRR for most interventions, and can be used to determine whether a body of evidence had adequate sample size, assuming a type 1 error of 0.05 and a type 2 error of 0.20.[8] The thresholds for the second precision criteria of a CI boundary of RR of 0.05 was arbitrary. For the purpose of this analysis, we considered an RRR of less than 0.05 to be small or trivial, although we acknowledge that in a certain context such risk reduction may be relevant to some stakeholders.

Statistical analysis

Data from each RCT were presented in a 2×2 contingency table. The FI was calculated as described by Walsh *et al*.[2] Events were added to the smaller event group and non-events were simultaneously subtracted, while maintaining a constant patient population. The Fisher exact test was then used to recalculate the two-sided p value, while iteratively adding events until the p value reached or exceeded 0.05. The number of additional events required to reach a p value of ≥0.05 was defined as the FI. To determine whether an RCT had 80% power to detect a statistically significant difference using a χ^2 test with two-sided significance level of 0.05, we calculated the baseline risk for the control group and assumed a moderate RRR of 25% or 30% for the treatment group. We constructed ROC curves to predict FI values using a non-parametric model proposed by Pepe.[10] The sensitivity and specificity and the corresponding FI cut-off values were estimated using the minimised distance between the selected point on ROC curve and the perfect sensitivity and specificity.[11] The nearest to (0,1) method was used to find the cutpoint on the ROC curve closest to (0,1) (ie, the perfect sensitivity and specificity). We compared FI between trials that had FI less than the number of patients lost to follow-up, compared with trials that did not, using Mann-Whitney U test. We used the 'fragility' package, 'roctab' command and 'cutpt' package as implemented in Stata V.17.0 (StataCorp).

Results

Description of randomised trials

Database search identified 1365 potential citations from which 78 were included and added to trials identified in a previous study.[3] Therefore, we finally included 201 cardiovascular RCTs. The process of study selection is depicted in the online supplemental figure 1 and the list of RCTs is provided in online supplemental appendix table, along with their raw data, effect size and FI. Most RCTs were multicentred (93.3%). More than half of the RCTs (59.2%) had an active comparator and (62.7%) evaluated pharmacological interventions. The mean sample size of an RCT was 5234 participants (IQR: 1046–7046). The FI ranged 1–172 and had a median of 13 (IQR: 5–28). Eighteen RCTs (9%) had FI of 1. The description of included RCTs is provided in table 1.

FI cut-offs

Table 2 summarises FI cut-offs with highest AUC to predict precision based on whether the information size was sufficient (ie, the study had adequate power for RRR of 25% and 30%) or if the CI did not overlap an arbitrary decision-making threshold of 0.05. FI of 12 predicted that the RCTs would be powered for RRR of 25% or 30%. FI of 9 predicted that the CI excludes a risk reduction <0.05. FI of 19 predicted that RCTs would be powered for RRR of 30% and that the CI excluded a small risk reduction <0.05. FI of 22 predicted that RCTs would be powered for RRR of 25% and that the CI excluded a small risk reduction <0.05. AUC for meeting both of these two precision criteria was 0.90 (0.86–0.94).

Table 1 Description of 201 cardiovascular randomised trials

	Mean (range) or percentage
Multiple centres	93.0%
Active comparator	59.2%
Type of intervention	
Pharmaceutical	62.7%
Surgical	18.9%
Imaging	1.0%
Surgical	0.5%
Other	16.9%
Funding	
For profit	68.7%
Government	24.9%
Other	6.4%
Sample size	5234 (500–50,156)
Follow-up (months)	22.9 (1–118)
Loss to follow-up (%)	2.0 (0–26)
Fragility index	24.6 (1–172)
Sample size powered for 25% RRR	8274.2 (96–142 644)
Sample size powered for 30% RRR	5706.7 (68–96 236)

RRR, relative risk reduction.

FI exceeded the number of patients lost to follow-up in 46/201 (22.89%) trials. FI in this subset of trials was 40.33 (range 3–172); which was significantly higher than FI in trials that had FI equal or less than the number of patients lost to follow-up (FI 19.89, range 1–120, p value for the difference between the two FIs was 0.001).

Discussion

Studies that demonstrate statistically significant results provide evidence that rejecting the null hypothesis is less likely to be due to chance.[12] However, when such studies are underpowered, the possibility of both, type 1 and type 2 errors increases, and such results are labelled as fragile. Therefore, the FI was proposed as an intuitive and easy way to communicate statistically significant results to clinicians and other stakeholders including perhaps patients. This index has no known anchors or values at which the results would be considered adequate or robust. We evaluated the FI of modern and likely influential cardiovascular RCTs that enrolled 500 or more participants, published in high impact journals and had a statistically significant primary outcome. We report several key findings in this analysis. First, the current study has identified that FI values of 19–22 have the highest AUC (best combination of sensitivity and specificity) to predict that the estimates were precise. For decision-making purposes, RCTs with FI lower than this range are highly susceptible to chance and their results should be interpreted with caution.

A second important finding of this study is that many RCTs had FI of 1 and over half of them may not meet such precision cut-offs (median FI was 13). This means that if very few patients were re-classified in terms of having an event, the outcome would become statistically insignificant. Thus, the treatment effect of many cardiovascular RCTs remains fragile and susceptible to random error, despite their statistical significance. This finding of common fragility in trials has been observed in various fields such as cardiology, rheumatology, anaesthesiology, ophthalmology, critical care, spine surgery and sport medicine.[3 13–18] Lastly, almost 1 in 4 trials had FI that exceeded the number of patients lost to follow-up. Results of such trials are even more fragile and less robust because the patients lost to follow-up could be the patients who would have had a different outcome and would change the statistical significance of the difference between study arms. This finding provides a rationale for presenting the number of patients lost to follow-up alongside FI.

The implications of these findings to clinical practice are important. A well-known example of an RCT with FI of 1 that changed clinical practice was the one by Poldermans *et al*; which misleadingly suggested that perioperative beta blockers given to patients undergoing non-cardiac surgery reduce mortality. These findings were subsequently discredited and the routine implementation of the intervention has likely caused harm to many patients.[19] Evidence derived from trials with statistically significant results that are fragile should be labelled imprecise and warrant lower certainty. Low certainty should not lead to strong recommendations and universal implementation. In addition, FI values should be presented with additional information such as the number of patients lost to follow-up, as well as event rates and CIs.

Limitations and strengths

It is important to recognise that the current study has evaluated FI only as an intuitive way to present information to evidence users. It could be also used as a teaching tool. However, FI is certainly not a formal way to make judgements about imprecision and has limitations.[20–22] Imprecision judgements should be made using an established and rigorous approach based on CI and sample size considerations using context specific thresholds.[8] The thresholds we studied were arbitrary and may change based on the importance or nature of the outcome. Lastly, we anticipate that RCTs in lower tier journals may even have lower FI values because they will likely have smaller sample size. Lastly, decision making should depend on the totality of evidence synthesised in a systematic review, not an individual study.[23] FI is merely a way to present the finding of a single statistically significant RCT in a simplified way. FI does not change the binary view of hypothesis testing, but it adds nuance and communicate additional information beyond the binary view. For example, instead of saying: 'the results are significant', the FI index will inform stakeholders that

Table 2 Fragility index values predicting precision

Precision criteria	FI cutpoint	Sensitivity (%)	Specificity (%)	AUC	95% lower limit	95% upper limit
CI boundary closest to null ≥0.05	≥9	69.59	67.92	0.80	0.73	0.86
Sample size powered to RRR 25%	≥12	77.65	63.79	0.75	0.68	0.82
Sample size powered to RRR 30%	≥12	75.00	70.97	0.75	0.69	0.82
CI boundary ≥5% and RRR 25%	≥22	81.48	83.67	0.90	0.86	0.94
CI boundary ≥5% and RRR 30%	≥19	81.43	82.44	0.90	0.86	0.94

AUC, area under the receiver operating characteristic curve; FI, fragility index; RRR, relative risk reduction.

'the results are significant, but they would lose significance if two patients had a different outcome'.

Conclusions

The findings of this study demonstrate that FI values in the range of 19–22 can be used to suggest that a treatment effect is likely to be precise and less likely to be susceptible to random error. Contemporary cardiovascular RCTs with 500 or more participants that have statistically significant results have a median FI of 13. Thus, approximately half of them do not meet this proposed range of values. The findings also provide a rationale for presenting the number of patients lost to follow-up alongside FI.

Contributors MHM, ZW and AKB conceived the idea. AKB, MSK, AS and SS selected studies and extracted data. ZW conducted the analysis. MHM is the guarantor of this work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iDs

Mohammad Hassan Murad http://orcid.org/0000-0001-5502-5975

Zhen Wang http://orcid.org/0000-0002-9368-6149

References

1 **Amrhein V**, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;**567**:305–7.
2 **Walsh M**, Srinathan SK, McAuley DF, *et al*. The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility index. *J Clin Epidemiol* 2014;**67**:622–8.
3 **Khan MS**, Ochani RK, Shaikh A, *et al*. Fragility index in cardiovascular randomized controlled trials. *Circ Cardiovasc Qual Outcomes* 2019;**12**:e005755.
4 **Potter GE**. Dismantling the fragility index: a demonstration of statistical Reasoning. *Stat Med* 2020;**39**:3720–31.
5 **Carter RE**, McKie PM, Storlie CB. The fragility index: a P-value in sheep's clothing? *Eur Heart J* 2017;**38**:346–8.
6 **Porco TC**, Lietman TM. A fragility index: handle with care. *Ophthalmology* 2018;**125**:649.
7 **Hultcrantz M**, Rind D, Akl EA, *et al*. The grade Working group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017;**87**:4–13.
8 **Guyatt GH**, Oxman AD, Kunz R, *et al*. GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol* 2011;**64**:1283–93.
9 **Murad MH**, Wang Z. Guidelines for reporting meta-epidemiological methodology research. *Evid Based Med* 2017;**22**:139–42.
10 **Pepe MS**. *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press, 2003.
11 **Perkins NJ**, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol* 2006;**163**:670–5.
12 **Royall R**. The effect of sample size on the meaning of significance test. *Am Stat* 1986;**40**:313–5.
13 **Berti A**, Cornec D, Medina Inojosa JR, *et al*. Treatments for giant cell arteritis: meta-analysis and assessment of estimates reliability using the fragility index. *Semin Arthritis Rheum* 2018;**48**:77–82.
14 **Mazzinari G**, Ball L, Serpa Neto A, *et al*. The fragility of statistically significant findings in randomised controlled anaesthesiology trials: systematic review of the medical literature. *Br J Anaesth* 2018;**120**:935–41.
15 **Ridgeon EE**, Young PJ, Bellomo R, *et al*. The fragility index in multicenter randomized controlled critical care trials. *Crit Care Med* 2016;**44**:1278–84.
16 **Evaniew N**, Files C, Smith C, *et al*. The fragility of statistically significant findings from randomized trials in spine surgery: a systematic survey. *Spine J* 2015;**15**:2188–97.
17 **Khan M**, Evaniew N, Gichuru M, *et al*. The fragility of statistically significant findings from randomized trials in sports surgery: a systematic survey. *Am J Sports Med* 2017;**45**:2164–70.
18 **Shen C**, Shamsudeen I, Farrokhyar F, *et al*. Fragility of results in ophthalmology randomized controlled trials: a systematic review. *Ophthalmology* 2018;**125**:642–8.
19 **Bouri S**, Shun-Shin MJ, Cole GD, *et al*. Meta-Analysis of secure randomised controlled trials of β-blockade to prevent perioperative death in non-cardiac surgery. *Heart* 2014;**100**:456–64.
20 **Acuna SA**, Sue-Chue-Lam C, Dossa F. The fragility Index-P values Reimagined, flaws and all. *JAMA Surg* 2019;**154**:674.
21 **Tignanelli CJ**, Napolitano LM. The fragility Index-P values Reimagined, flaws and All-Reply. *JAMA Surg* 2019;**154**:674–5.
22 **Andrade C**. The use and limitations of the fragility index in the interpretation of clinical trial findings. *J Clin Psychiatry* 2020;**81**. doi:10.4088/JCP.20f13334. [Epub ahead of print: 24 03 2020].
23 **Murad MH**, Montori VM, Ioannidis JPA, *et al*. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. *JAMA* 2014;**312**:171–9.