Internet Medical Journal
ISSN 1093-7935 · medjournal.com · Est. 1996

Resampling Fragility, Perturbation Fragility, and Why the Global Fragility Index Is Not a P-Value in Disguise

Thomas F. Heston

Affiliation	Department of Family Medicine, University of Washington, Seattle, USA
Affiliation	Department of Medical Education and Clinical Sciences, Elson S. Floyd College of Medicine, Washington State University, Spokane, USA
ORCID	0000-0002-5655-2512
Published	22 Aug 2026
DOI	10.5281/zenodo.22059146
Article type	Perspective
Citation	Heston TF. Resampling Fragility, Perturbation Fragility, and Why the Global Fragility Index Is Not a P-Value in Disguise. Internet Medical Journal. 2026;1:e22059146

Abstract

The charge that the fragility index is a P-value in disguise rests on a strong correlation between the two across trials, and it overlooks a distinction between two distinct constructs derived from the same observed result. Resampling fragility asks how often the significance verdict would change if the trial were drawn again at the same size; it is a probability, and because it is based on the same observed result and a specified replication model, it is often strongly associated with the P-value. Perturbation fragility asks how many recorded outcomes must change before the observed table crosses the significance threshold; it is a count, the global fragility index, or a proportion, the global fragility quotient, and it measures the geometric distance of the observed table from the decision boundary, a quantity the P-value does not encode. Worked examples from published trials show tables with similar P-values and widely different perturbation fragility. The redundancy critique, whether made by resampling or by machine-learning models that use the P-value as a predictor, is correct about resampling fragility but does not address the fragility index.

Keywords

global fragility index, fragility quotient, statistical fragility, P-value, resampling, perturbation, p–fr–nb, clinical trials

Introduction

The characterization of the fragility index as "a P-value in sheep's clothing" reflects, in part, its strong inverse association with the P-value across simulated or observed trials [1,2]. However, this cross-trial association does not establish informational equivalence within an individual trial: at a given P-value, the fragility index may vary with sample size and event configuration. This distinction separates *resampling fragility*—the probability that statistical significance changes in a new sample—from *perturbation fragility*, which is the number of outcome changes required to move the observed table across the significance threshold. Treating a cross-trial correlation as a measure of within-trial perturbation fragility conflates these distinct concepts and overstates the critique.

The foundation is straightforward. A test statistic is computed from the observed contingency table. Given the null hypothesis and the test's specified reference distribution, the P-value is the probability of obtaining a test statistic at least as incompatible with the null hypothesis as the observed statistic [3]. Resampling fragility uses the observed arm-specific event rates and sample sizes to generate hypothetical replications, then estimates the probability that the significance classification changes. Because both quantities reflect how the observed data sit relative to the significance boundary, they will often be strongly associated. That association, however, does not make a resampling-based fragility measure identical to the P-value, nor does it establish redundancy with within-table perturbation fragility. It shows only that resampling behavior and statistical significance are related features of the same observed result.

The critical literature illustrates this pattern. The question of whether the fragility index merely restates the P-value was posed directly in cardiovascular trials, with a warning that the index can misrepresent a study's strength if interpreted without care [1]. A subsequent formal analysis concluded that the index lacks a probability-based motivation, can penalize small trials that reach the same significance level with fewer events, and may encourage an incorrect probabilistic interpretation; sensitivity analyses were recommended instead [4]. A third line of work retained the perturbation framework but restricted the index to sufficiently likely modifications, thereby incorporating the plausibility of the assumed outcome changes rather than treating all perturbations as equally credible [5]. The most recent contribution adds a machine-learning analysis: across 300,000 simulated trials, random forest and XGBoost models predicted fragility metrics with near-perfect accuracy. For the fragility index, the random-forest model achieved $R^2 = 1.0$, and the P-value accounted

for 79.8% of its split-based feature importance [6]. This result is unsurprising because the fragility index is deterministic given the four cell counts, while the P-value is itself derived from those counts. However, near-perfect prediction and split-based feature importance do not establish that the P-value and fragility index are informationally equivalent: with redundant predictors, feature importance describes how a fitted model reduces error, not the unique information contributed by one variable. The analysis therefore shows that the P-value is useful for predicting the fragility index in the simulated data, but it does not test whether the fragility index varies among trials with the same P-value—a question for which earlier simulation work found substantial variation [2].

The p–fr–nb framework separates three questions: statistical significance (p), classification stability (fr), and distance from therapeutic neutrality (nb). For two-arm binary outcomes, a global formulation defines fr as the global fragility quotient, GFQ = GFI/N, where GFI is the minimum number of path-independent cell-to-cell reallocations required to reverse the significance classification under a fixed test. The robustness coordinate, nb, is the distance to therapeutic neutrality; for a 2×2 table, it is measured by the risk quotient, RQ = $|ad - bc|/(N^2/4)$. Thus, the framework distinguishes one probability-based significance measure from two descriptive distances, each referenced to a different boundary [7,8]. It does not justify treating a resampling-based measure as identical to the P-value. Table 1 summarizes the two constructs.

Table 1. Resampling fragility compared with perturbation fragility

Source of doubt	What is perturbed	Measure
Sampling variability (future replications)	Entire new samples of size n	Resampling fragility (a probability)
Proximity of the observed table to the significance boundary (misclassification, late events, exclusions)	Individual outcome codes inside the observed table	Perturbation fragility (a count, or the quotient fr)

Worked Examples

Published trials show the two quantities coming apart. Each worked example is a pair of trials chosen for similar P-values. The first pair shows two significant trials, the second pair shows two nonsignificant trials, and the third shows similar P-values and sample size.

In a phase 2 trial of sustained remission in giant cell arteritis, 35 of 42 patients in one arm and 14 of 28 in the other met the endpoint (N = 70, p = .0039, GFI = 3, GFQ = .043) [9]. In a two-arm comparison from a cardiovascular outcomes trial in type 2 diabetes, 33 of 1,269 and 60 of 1,250 patients had the composite outcome (N = 2,519, p = .0041, GFI = 5, GFQ =

.0020) [10]. The P-values are similar, but the GFQ (perturbation fragility) differs 21-fold (Table 2).

Table 2. Comparison of the P-value with the GFQ in two significant trials

Study	N	P-value	GFI	GFQ
Cid et al., 2022	70	0.0039	3	0.043
Green et al., 2024	2519	0.0041	5	0.0020

The second example pairs two nonsignificant trials. In a trial of zoledronic acid after hip fracture, 33 of 1,062 and 23 of 1,065 patients sustained a new hip fracture (N = 2,127, p = 0.18, GFI 3, GFQ = 0.0014) [11]. In a dose-finding trial of intranasal midazolam for pediatric procedural sedation, 9 of 19 and 20 of 29 children achieved adequate sedation (N = 48, p = 0.23, GFI 2, GFQ = 0.042) [12]. The P-values are similar, but the GFQ differs 30-fold (Table 3).

Table 3. Comparison of the P-value with the GFQ in two non-significant trials

Study	N	P-value	GFI	GFQ
Lyles et al., 2007	2127	0.18	3	0.0014
Tsze et al., 2025	48	0.23	2	0.042

The third pair consists of two nonsignificant trials of similar size, reducing the likelihood that sample size alone explains the difference in GFQ. In a phase 2a dose-ranging trial of emodepside versus ivermectin for Strongyloides stercoralis infection, 19 of 25 and 22 of 25 participants were cured (N = 50, p = .46, GFI = 2, GFQ = .040) [13]. In a trial of adalimumab for steroid sparing in giant cell arteritis, 20 of 34 and 18 of 36 patients reached the remission endpoint (N = 70, p = .48, GFI = 6, GFQ = .086) [14]. The P-values and sample sizes are similar, but the GFQ differs 2-fold (Table 4).

Table 4. Comparison of the P-value with the GFQ in two trials of similar size

Study	N	P-value	GFI	GFQ
Taylor et al., 2025	50	0.4635	2	0.0400
Seror et al., 2014	70	0.4823	6	0.0857

Conclusions

The practical consequence of distinguishing resampling fragility from perturbation fragility is a change in what each number is asked to do. A resampling check addresses the stability of a significance classification under a specified replication model; it is related to, but not identical with, the P-value. A fragility index, such as the GFI, answers a question the P-value cannot: how many recorded outcomes separate the statistical significance verdict from its reversal. Here, three worked examples show that published clinical trials with a similar P-value can demonstrate widely different susceptibility to small perturbations in the data. Complete statistical evidence therefore reports probability (the P-value) together with fragility of the significance classification (fr, e.g. the GFQ) and distance from therapeutic neutrality (nb, e.g. RQ). Critics of statistical fragility are right that the fragility index is not a probability; that is what makes it worth reporting.

Declarations

Funding: This study did not receive any external funding.

Conflicts of Interest: The author reports no conflicts of interest.

Data Availability: Not applicable.

Research Ethics Statement: Not applicable. This perspective did not involve human subjects research, animal research, or protected health information.

AI Usage: Large language models were used for language editing and formatting assistance; the author reviewed, verified, and is fully responsible for all content.

References

1. Carter RE, McKie PM, Storlie CB. The Fragility Index: a P-value in sheep's clothing? Eur Heart J. 2017;38: 346–348. doi:10.1093/eurheartj/ehw495

2. Condon TM, Sexton RW, Wells AJ, To M-S. The weakness of fragility index exposed in an analysis of the traumatic brain injury management guidelines: A meta-epidemiological and simulation study. Pasin L, editor. PLOS ONE. 2020;15: e0237879. doi:10.1371/journal.pone.0237879

3. McHugh ML. The Chi-square test of independence. Biochem Medica. 2013; 143–149. doi:10.11613/BM.2013.018

4. Potter GE. Dismantling the Fragility Index: A demonstration of statistical reasoning. Stat Med. 2020;39: 3720–3731. doi:10.1002/sim.8689

5. Baer BR, Gaudino M, Charlson M, Fremes SE, Wells MT. Fragility indices for only sufficiently likely modifications. Proc Natl Acad Sci U S A. 2021;118: e2105254118. doi:10.1073/pnas.2105254118

6. Vivekanantha P, Son H, Kay J, Kotipalli S, Bouchard MD, Madden K, et al. Direct Comparison Between Loss-to-Follow-Up and Statistical Fragility Is Methodologically Inappropriate, and Fragility Reflects the *P* Value, Not Trial Robustness: A Simulation Analysis of 300,000 Randomized Controlled Trials. Arthroscopy. 2026; arj.70457. doi:10.1002/arj.70457

7. Heston TF. Significance, Fragility, and Robustness in Clinical Trials: Stratifying Statistical Evidence. Cureus. 2025;17. doi:10.7759/cureus.100494

8. Heston TF. The Global Fragility Index: A Path-Independent Measure of Statistical Fragility. SSRN Electron J. 2025; 5709162. doi:10.2139/ssrn.5709162

9. Cid MC, Unizony SH, Blockmans D, Brouwer E, Dagna L, Dasgupta B, et al. Efficacy and safety of mavrilimumab in giant cell arteritis: a phase 2, randomised, double-blind, placebo-controlled trial. Ann Rheum Dis. 2022;81: 653–661. doi:10.1136/annrheumdis-2021-221865

10. Green JB, Everett BM, Ghosh A, Younes N, Krause-Steinrauf H, Barzilay J, et al. Cardiovascular Outcomes in GRADE (Glycemia Reduction Approaches in Type 2 Diabetes: A Comparative Effectiveness Study). Circulation. 2024;149: 993–1003. doi:10.1161/CIRCULATIONAHA.123.066604

11. Lyles KW, Colón-Emeric CS, Magaziner JS, Adachi JD, Pieper CF, Mautalen C, et al. Zoledronic Acid and Clinical Fractures and Mortality after Hip Fracture. N Engl J Med. 2007;357: 1799–1809. doi:10.1056/NEJMoa074941

12. Tsze DS, Woodward HA, McLaren SH, Leu C-SS, Venn AMR, Hu NY, et al. Optimal Dose of Intranasal Midazolam for Procedural Sedation in Children: A Randomized Clinical Trial. JAMA Pediatr. 2025;179: 979. doi:10.1001/jamapediatrics.2025.2181

13. Taylor L, Many S, Jeanguenat H, Hattendorf J, Sayasone S, Keiser J. Efficacy and safety of ascending doses of emodepside in comparison with ivermectin in adults infected with Strongyloides stercoralis in Laos: a phase 2a, dose-ranging, randomised, parallel-group, placebo-controlled, single-blind clinical trial. Lancet Infect Dis. 2025;25: 1254–1264. doi:10.1016/S1473-3099(25)00255-5

14. Seror R, Baron G, Hachulla E, Debandt M, Larroche C, Puéchal X, et al. Adalimumab for steroid sparing in patients with giant-cell arteritis: results of a multicentre randomised controlled trial. Ann Rheum Dis. 2014;73: 2074–2081. doi:10.1136/annrheumdis-2013-203586