Beyond fragility: what the fragility index cannot measure

Author: Thomas F. Heston, MD, MSc

Affiliations: Department of Family Medicine, University of Washington, Seattle, WA; Department of Medical Education and Clinical Sciences, Washington State University, Spokane, WA

ORCID: 0000-0002-5655-2512

Citation: Heston TF. Beyond fragility: what the fragility index cannot measure. Internet Medical Journal. 2026;1(1):e19465222. doi:10.5281/zenodo.19465222

The fragility index (FI) answers a question that *p* values cannot — how many outcome changes would reverse the significance classification — but it leaves unanswered the equally critical question of whether the observed result is geometrically separated from no effect, and recent editorials in major journals have now cataloged this limitation with unusual clarity. Two editorials published in early 2026 — one accompanying a systematic review of spinal cord stimulation trials in Anesthesiology [1] and another accompanying a fragility analysis of mechanical circulatory support trials in Journal of Cardiac Failure – Intersections [2] — arrive at strikingly convergent conclusions: the FI is sample-size

dependent, correlated with the p-value, restricted to dichotomous outcomes in its classic form, and unable to measure the strength of an effect.

The observation that trials are "fragile by design" because sample size calculations target the minimum enrollment needed to detect significance is particularly important, as it reframes low FI values not as evidence of unreliable findings but as an expected consequence of efficient trial design. These critiques are well-founded, and they point directly to a structural gap in the current evidence assessment toolkit that fragility metrics alone cannot close.

The foundational insight behind the FI — that the number of outcome changes required to flip a p-value across the 0.05 threshold reveals information about classification stability that the p-value itself conceals — remains sound [3]. A trial in which three outcome toggles reverse significance is less stable than a similarly sized trial requiring thirty toggles, and this distinction has genuine interpretive value. The fragility quotient (FQ), which normalizes the raw count to sample size, partially addresses the scaling problem [4]. Allocation-corrected variants such as the modified-arm fragility quotient (MFQ) further reduce bias from unequal randomization, and continuous extensions allow the concept to be applied beyond binary outcomes [5,6]. These refinements preserve the model-free property that makes fragility metrics practical: they require only published summary statistics and the exact test geometry that generated the original p-value, with no distributional reconstruction or simulation.

Yet even with these corrections, fragility metrics address only one of two evidence dimensions orthogonal to *p* values. Fragility quantifies how close a result sits to the significance decision boundary — whether a small perturbation could reverse the classification. It does not quantify how far the observed result sits from therapeutic neutrality, the point at which treatment and control are indistinguishable. A trial may have high classification stability — its p-value is robust to outcome changes — and yet show an effect barely distinguishable from no effect at all. Conversely, a trial may be fragile in the classification sense while describing an effect that is geometrically far from neutrality, suggesting an underpowered detection of a real treatment difference. Neither the FI nor any of its derivatives can distinguish between these scenarios, and this is the specific gap that the recent editorials identify but fail to resolve.

The p–fr–nb framework addresses this gap by defining complete statistical evidence as a triplet: significance (p), fragility (fr), and robustness (nb) [7]. The significance dimension (p) describes probability- specifically, the probability that the observed findings would occur by chance under the null hypothesis; it then classifies the findings as significant or nonsignificant based upon an arbitrary threshold of $p = 0.05$. The fragility dimension (fr) quantifies how stable the significance classification is under small perturbations to the data. The robustness dimension (nb) quantifies geometric distance from the neutrality boundary using the general formula $nb = |T - T_0| / (|T - T_0| + S)$, where T is the observed test statistic, T_0 is its value under neutrality, and S is a design-appropriate scale parameter [8]. All robustness metrics — including those for binary, continuous, diagnostic, ordinal, survival, and correlation designs — output a single 0–1 value, where 0

indicates the result lies on the neutrality boundary (where treatment is completely independent of the outcome) and values approaching 1 indicate maximal separation, indicating strong evidence of a true effect of treatment on outcomes. Critically, *nb* is orthogonal to both *p* and *fr*: it measures something that neither significance testing nor classification stability can capture, namely the strength of evidence that a non-zero effect exists, independent of whether the result crosses the conventional significance threshold.

The clinical value of this distinction is well illustrated by the cardiogenic shock data presented in the companion article to the JCF–Intersections editorial. The DanGer Shock trial showed an FI of 4 for 180-day mortality — meaning that four outcome changes would reverse the significance finding — but an FI of 17 for vascular adverse events [9]. Under the current framework, both results are reported as significant, yet the fragility analysis suggests the mortality result is more vulnerable to perturbation. What neither the FI nor the *p* value can determine is whether the mortality result reflects a treatment effect that is geometrically separated from neutrality or one that sits near the boundary of no difference. The *nb* metric would resolve this ambiguity directly.

In an empirical validation of the complete evidence framework across 100 pharmaceutical trials, 18% of statistically significant results exhibited the pattern of significance with low fragility and low robustness — a 13.4-fold elevation over the null expectation — representing trials where the *p* value classification was fragile, the effect was near neutrality, and the overall evidence quality was poor despite formal statistical significance [10]. Half of all trials showed discordance between their *p* value classification

and their complete evidence assessment, confirming that reporting *p* values alone yields systematically incomplete evidence.

The convergence of these Tier 1 editorials on the limitations of fragility-only assessment presents an opportunity for the field. The specific deficiencies identified — sample-size dependence, correlation with *p* values, absence of accepted thresholds, and inability to measure effect strength — are each addressed by complementing *p* values with fragility (*fr*) and robustness (*nb*) dimensions. Reporting the *p–fr–nb* triplet alongside effect size does not replace existing metrics but completes them. Future fragility analyses in cardiology and other fields should consider supplementing probability-based significance classification (*p*) with classification stability (*fr*) and robustness (*nb*), moving the evidence assessment from a single-axis evaluation to the three-dimensional framework that the underlying statistical questions require.

Declaration

The author reports no conflicts of interest. This study did not receive any external funding. Large language models were used for language editing and formatting assistance; the author reviewed, verified, and is fully responsible for all content.

References

1. Bhatia A. Understanding Stability by Evaluating Fragility (Indices). Anesthesiology. 2026;144: 764–766. doi:10.1097/ALN.0000000000005932
2. Kosyakovsky LB, Garan AR. Frailty, Thy Name is Significance: Examining the Utility of the Fragility Index in Cardiovascular Clinical Trials. J Card Fail - Intersect. 2026;0. doi:10.1016/j.yjcafi.2026.02.013

3. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clin Epidemiol. 2014;67: 622–628. doi:10.1016/j.jclinepi.2013.10.019

4. Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility index? Crit Care Med. 2016;44: e1142–e1143. doi:10.1097/CCM.0000000000001976

5. Heston TF. The Modified-Arm Fragility Quotient: An Improved Metric for Assessing Robustness in Clinical Trials. Zenodo. 2025. doi:10.5281/zenodo.17014261

6. Heston TF. Meaningful Change Index: A P-Value Independent Metric for Assessing Robustness and Fragility in Continuous Outcomes. Zenodo. 2025. doi:10.5281/zenodo.17212383

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 Neutrality Boundary Framework: Quantifying Statistical Robustness Geometrically. arXiv. 2025; 2511.00982. doi:10.48550/arXiv.2511.00982

9. Albert CL, Pampori A, Guglin M. Fragility Index Analysis of Contemporary Temporary Mechanical Circulatory Support Devices in Cardiogenic Shock. J Card Fail - Intersect. 2026; S3050661126000250. doi:10.1016/j.yjcafi.2025.12.014

10. Heston TF. Fragility Metrics Toolkit. Zenodo; 2026. doi:10.5281/zenodo.17254763