

Guideline Evidence Audits Require Robustness Beyond the Fragility Index

Author(s)	Thomas F. Heston
Affiliation(s)	Department of Family Medicine, University of Washington, Seattle, USA
Affiliation(s)	Department of Medical Education and Clinical Sciences, Elson S. Floyd College of Medicine, Washington State University, Spokane, USA
ORCID	0000-0002-5655-2512
Published	19 APR 2026
DOI	10.5281/zenodo.19656227
Article type	Commentary
Citation	Heston TF. Guideline evidence audits require robustness beyond the fragility index. Internet Medical Journal. 2026;1:e19656227

© 2026 The Author(s). This article is distributed under the terms of the [Creative Commons Attribution 4.0 International License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Fragility index audits of clinical guideline evidence measure classification stability but not distance from therapeutic neutrality — leaving the most consequential question in evidence grading unanswered. Applied to the randomized controlled trials recently cited in the NCCN guidelines for gastric cancer, the fragility index cannot distinguish a low-powered detection of a real treatment effect from a near-null result that narrowly crossed the significance threshold, a distinction with direct implications for confirmatory trial prioritization and guideline strength ratings. The significance-fragility-robustness framework resolves this gap by adding robustness dimension — a geometrically derived, model-free measure of distance from therapeutic neutrality — as an orthogonal third statistical metric alongside significance and fragility, completing the evidence picture that guideline audits require.

Keywords

fragility index, statistical robustness, neutrality boundary, NCCN guidelines, gastric cancer, evidence quality, p-fr-nb, clinical trial methodology, guideline evidence, complete statistical evidence

Fragility index (FI) analyses of clinical guideline evidence establish only half the picture. As applied to randomized controlled trials cited in major oncology guidelines — including, most recently, the NCCN guidelines for gastric cancer (1) — the FI identifies how many outcome changes would reverse a trial's significance classification but does not measure whether the underlying effect is close to zero or meaningfully separated from it. The distinction matters: a trial with an FI of three may have detected a real, clinically important effect that was simply underpowered, or it may have produced a near-null result that narrowly crossed the significance threshold. A fragility audit that cannot separate these two cases is structurally incomplete, regardless of how many trials it analyzes or how carefully the index is calculated.

The FI rests on an established and useful principle. The number of outcome reversals required to change a trial's significance classification was formalized as a measure of classification stability in randomized controlled trials (2). A low FI indicates that a statistically significant result depends on very few patient events and would be reversed by a small perturbation to the data — a clinically meaningful property that alerts readers to results that may not replicate or withstand modest changes in follow-up ascertainment. The FI has accordingly been adopted as a supplementary reporting metric across surgery, cardiology, and now oncology guideline evaluation. Its adoption reflects a genuine demand: clinicians need tools that go beyond the p-value to assess the stability of trial-derived evidence. That demand, however, is only partially met by a metric that captures classification stability without addressing how far the observed effect is from zero.

The structural gap is most consequential when fragility audits are applied to guidelines that govern high-mortality diseases. Gastric cancer is among the five leading causes of cancer-related mortality worldwide, and the NCCN guidelines for gastric cancer are followed across dozens of countries, directly governing systemic treatment decisions in a disease where therapeutic selection carries significant morbidity and cost. A recent analysis applying the FI to randomized controlled trials cited in the NCCN guidelines (1) frames a high FI as evidence of robustness — conflating classification stability (fragility) with distance from therapeutic neutrality (robustness). These are orthogonal constructs: a trial can require many outcome changes to flip significance while its observed effect remains geometrically near zero, and a trial can be easily flipped while its effect is strongly separated from the null. Without directly measuring robustness (distance from therapeutic neutrality), the analysis cannot answer the question that matters most for guideline interpretation: are trials with low fragility indices detecting real effects poorly, or near-null effects adequately? Without a robustness dimension, the evidence-quality gradient within the guideline remains unmapped — and the appeal to robustness in the paper's title is unfulfilled by the analysis itself (3).

The p–fr–nb framework resolves this ambiguity by introducing a second orthogonal dimension in addition to fragility. Significance (p) quantifies the probability of obtaining the observed result under the null hypothesis. Fragility (fr) measures the stability of this p-value-derived significance classification: how much perturbation the data can absorb before the p-value crosses the significance boundary (typically an alpha of 0.05). Robustness (nb) measures an entirely different dimension: the geometric distance from therapeutic neutrality. This represents how far the observed data sits from the boundary at which treatment and control produce indistinguishable outcomes (3,4). Each of these three dimensions is necessary for a complete statistical analysis of the research data. Providing only a p-value can be misleading due to the binary and arbitrary nature of the 0.05 cutoff. A trial can be fragile yet robust — it detected a real separation from the null, but sparse data make the significance call unstable. Or it can be stable and weakly robust — the p-value is firmly below threshold, but the treatment effect barely exceeds the line of clinical indifference. For guideline evidence audits, this independence is not a statistical abstraction: it determines whether a low-fragility-index finding should prompt calls for a confirmatory trial (the effect may be real but underdetected) or for re-evaluating whether the effect is meaningful enough to anchor a treatment recommendation at all.

Empirical data support the need to report all three dimensions together. In a validated analysis of 129 clinical trials, 50% showed discordance between their p-value classification and their complete p–fr–nb evidence assessment (3), and demonstrated that even the addition of the FI to p-value reporting still mischaracterizes a substantial proportion of trial results when robustness is absent. This problem is not confined to gastric oncology: FI analyses of trials cited in endometriosis management guidelines (5) follow the same structural template — fragility without robustness — leaving the robustness dimension unmeasured across guideline domains.

Refinements to the FI are important, but ultimately incomplete. For example, the bidirectional refinement of the FI proposed by the same Kahana group in a companion publication (6) addresses the direction of significance transitions and represents a genuine methodological contribution, but it retains the same single-dimensional architecture by leaving out a robustness metric. Also, as with the original FI, the bidirectional FI remains mathematically incomplete (7). While refinements to the FI are important, unless a robustness dimension is present, it remains unclear if an observed effect is geometrically separated from the point of therapeutic neutrality.

Guideline evidence audits are high-value instruments that should be built on complete, rather than partial, statistical evidence. Future analyses of oncology guideline-supporting trials should report the p–fr–nb triplet alongside effect size, providing the three-dimensional picture of evidence quality, significance, stability, and distance from null that clinical decision-making requires. The NCCN gastric cancer trial cohort assembled by recent fragility analysis is a natural starting point: calculating neutrality boundary metrics alongside the

published fragility indices would immediately stratify low-fragility-index trials into those reflecting real-but-unstable effects and those reflecting near-null effects that narrowly achieved significance — a distinction that carries direct implications for the prioritization of confirmatory trials and the interpretation of guideline strength ratings. Establishing robustness as a distinct metric and a standard component of guideline evidence audits is the next logical step in the evolution of statistical methodology from evidence description to evidence grading.

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 commentary 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. Kahana N, Boaz E, Horesh N, Dourado J, Rogers P, Aeschbacher P, et al. Using the Fragility Index analysis for assessment of the robustness of evidence in randomized controlled trials cited in the NCCN guidelines for gastric cancer. *Surgery*. 2026 Apr 14;110176. doi:10.1016/j.surg.2026.110176
2. 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 Jun;67(6):622–8. doi:10.1016/j.jclinepi.2013.10.019
3. Heston TF. Significance, Fragility, and Robustness in Clinical Trials: Stratifying Statistical Evidence. *Cureus*. 2025 Dec 31;17(12). doi:10.7759/cureus.100494
4. Heston TF. Neutrality Boundary Fragility: A Geometric Framework for Statistical Stability. *SSRN Electronic Journal*. 2025 Oct 21. doi:10.2139/ssrn.5636470

5. Yagur Y, Horesh N, Levin G, Meyer R. Fragility Index Assessment of Randomised Trials Informing Endometriosis Management Guidance. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2026;133(6):1300–2. doi:10.1111/1471-0528.70185
6. Kahana N, Perets M, Emile SH. Beyond One Direction: Revisiting the Fragility Index. *Journal of clinical epidemiology*. 2026;112255-. doi:10.1016/j.jclinepi.2026.112255
7. Heston TF. Bidirectional fragility is a step forward but not far enough: the case for a global fragility index. *Internet Medical Journal*. 2026 Apr 8;1(1):e19464166–e19464166. doi:10.5281/zenodo.19464166