Bidirectional fragility is a step forward but not far enough: the case for a global fragility index

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. Bidirectional fragility is a step forward but not far enough: the case for a global fragility index. Internet Med J. 2026;1(1):e19464166. doi:10.5281/zenodo.19464166

The bidirectional fragility index [1] represents a genuine improvement over the original unidirectional procedure [2] but stops short of the solution it claims to provide because it inherits two structural limitations that no directional refinement can resolve: an arbitrary constraint on which cells may be toggled and the complete absence of a robustness dimension that measures distance from therapeutic neutrality.

The original fragility index counts the minimum number of outcome toggles required to reverse statistical significance in a 2×2 contingency table using a two-sided Fisher's exact test [2]. This procedure is restricted to a single direction of change — toggling non-events to events in the arm with fewer events — which means it may overestimate the

perturbation needed to cross the significance boundary when the shortest path runs in the opposite direction or through the other arm. The bidirectional refinement corrects this asymmetry by searching across all possible directions of change within each arm [1,3]. Because it evaluates both arms and toggles in both directions, the bidirectional count is always less than or equal to the unidirectional count, thereby providing a tighter bound on classification instability.

The central limitation is one of scope. The bidirectional fragility index still constrains toggles to within-arm movements — an event in one arm switches to a non-event, or vice versa, but no observation may move between arms. This constraint fixes the row marginal totals of the contingency table while allowing column marginals to vary. The rationale for fixed row margins is that arm sizes are determined by study design, not by outcomes. However, the fragility index is a thought experiment — it asks what *would* happen if a small number of outcomes had been different, or if, e.g. an administrative error was made during data entry. In that counterfactual, patients do not leave the study; their outcomes change. If the thought experiment permits outcomes to change within an arm (altering column totals), there is no mathematical reason it cannot also permit outcomes to change across arms. The row-margin constraint is a modeling choice, not a logical necessity, and it restricts the perturbation space in ways that can overestimate the minimum number of changes required to flip significance.

The claim that the bidirectional approach "accounts for all possible directions of change" [1] is incomplete because it only accounts for changes in all directions within the

confines of fixed row margins. Fixing row marginal totals only is a subset of the full perturbation space. In contrast, the Global Fragility Index (GFI) searches across all admissible cell-to-cell reallocations in the contingency table — including cross-arm movements — without fixing any marginal totals, and identifies the path-independent minimum number of cell moves required to reverse significance [4]. The distinction is structurally meaningful: GFI is always less than or equal to the bidirectional FI, and the two can diverge substantially in tables with unequal allocation or asymmetric event distributions. If the goal is a minimal perturbation count — the data's fragility when subjected to small changes — then the GFI is the correct solution.

For context, the Unit Fragility Index (UFI) fixes both row and column margins [5,6]. The bFI fixes row margins only, and the GFI fixes neither, as the GFI allows toggling within and between study arms. While there may be an argument for keeping all marginal totals fixed or just row marginal totals fixed, these come at the expense of creating mathematical incompleteness. Mathematical incompleteness occurs when the metric cannot be calculated for every contingency table in the domain where significance can be both attained and reversed. The GFI is mathematically complete because it allows all possible rearrangements of the contingency table.

An exhaustive enumeration of every 2×2 table with N = 7 to 30 and at least one subject per arm — the domain in which significance can be both attained and reversed — shows that the UFI is not attainable in 5.6% of cases, the bFI is not attainable in 0.1%, but the GFI is attainable in all cases. The GFI < bFI 21.5% of the time, and GFI is never greater

than the bFI (data on Zenodo). One counterexample showing the mathematical incompleteness of both the bFI and UFI is the table {8, 2, 0, 1}. While this mathematical incompleteness of the bFI is small, it is nevertheless a proven limitation: in any cross-trial analysis, tables returning 'Not Attainable' must be excluded or imputed, either of which introduces bias that the GFI avoids entirely.

A secondary limitation concerns normalization. Raw fragility counts scale with sample size, and quotients that adjust for this depend on the choice of denominator. The Modified-arm Fragility Quotient (MFQ, also known as the intervention fragility quotient) normalizes to the arm actually toggled, which is well-defined for single-arm procedures [7]. The Global Fragility Quotient (GFQ) is divided by total N, which is unambiguous because the GFI does not privilege either arm [4]. The bFI, however, toggles both arms along a single path, leaving no principled denominator for normalization — dividing by N reproduces the same total-N quotient already known to distort under unequal allocation [7]. Of these quotients, the GFQ is structurally the most allocation-resistant: because the GFI searches an unrestricted perturbation space that does not privilege either arm, neither the numerator nor the denominator of GFQ depends on the allocation ratio. Whether this structural advantage translates into superior empirical performance across real trial distributions remains to be determined.

The bidirectional fragility index is presented as providing "a clearer reflection of the actual robustness of trial findings" [1]. This language deserves scrutiny. In the fragility literature, "robustness" is used informally to mean "not fragile," but if robustness is to be a

measurable property of statistical evidence rather than a synonym for a high toggle count, it must quantify something that fragility does not. Fragility — whether unidirectional, bidirectional, or global — measures classification stability: how many outcome changes are needed to flip the significance decision. A low FI count indicates classification instability (fragility), whereas a high count indicates classification stability. But classification stability is not the only dimension that matters. Clinicians routinely report effect sizes to quantify the magnitude of an observed difference, yet effect size alone does not indicate whether that magnitude is distinguishable from therapeutic neutrality given the study's sampling variability. A large trial with a barely significant p-value may have a high fragility quotient, indicating stable significance classification, yet the observed effect may sit trivially close to the point of therapeutic neutrality. Conversely, a small trial may have a low fragility quotient, indicating unstable classification, yet the observed effect may be geometrically far from neutrality. What is missing is a normalized metric that captures how far the result sits from neutrality relative to sampling uncertainty — a dimension that neither p-values, confidence intervals, effect sizes, nor fragility counts directly provide [8]. For 2×2 contingency tables, the Risk Quotient (RQ) formalizes this dimension as a [0, 1] metric quantifying geometric distance from the neutrality boundary [9]. Reporting significance (p), fragility (fr), and robustness (nb) together as a triplet captures the full evidence profile; reporting any fragility count alone — whether unidirectional or bidirectional — leaves the distance-from-neutrality question unanswered [10].

The bidirectional fragility index deserves recognition as a genuine correction to the original unidirectional procedure, and its publication in a leading methodology journal

underscores the growing consensus that classification stability matters. The next step is to move beyond directional refinements of partial evidence and adopt complete statistical evidence: significance, fragility, and robustness reported together. The tools to do this — p-values, GFQ, and RQ — are available, open-source, and applicable to every standard study design. The question is no longer whether fragility should be bidirectional; it is whether fragility metrics should be mathematically complete and allocation-resistant; and whether statistical significance and fragility alone, leaving out robustness of effect, are sufficient.

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. Kahana N, Perets M, Emile SH. Beyond One Direction: Revisiting the Fragility Index. J Clin Epidemiol. 2026; 112255-. doi:10.1016/j.jclinepi.2026.112255
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;67: 622–628. doi:10.1016/j.jclinepi.2013.10.019
3. Lin L, Chu H. Assessing and visualizing fragility of clinical results with binary outcomes in R using the fragility package. Gagniuc PA, editor. PLOS ONE. 2022;17: e0268754. doi:10.1371/journal.pone.0268754
4. Heston TF. The Global Fragility Index: A Path-Independent Measure of Statistical Fragility. Zenodo. 2025 [cited 4 Apr 2026]. doi:10.5281/zenodo.18078509

5. Feinstein AR. The unit fragility index: an additional appraisal of “statistical significance” for a contrast of two proportions. J Clin Epidemiol. 1990;43: 201–209. doi:10.1016/0895-4356(90)90186-s

6. Walter SD. Statistical significance and fragility criteria for assessing a difference of two proportions. J Clin Epidemiol. 1991;44: 1373–1378. doi:10.1016/0895-4356(91)90098-T

7. Heston TF. Adjusting fragility metrics for unequal trial randomizations. Autoimmun Rev. 2025;24: 103935. doi:10.1016/j.autrev.2025.103935

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

9. Heston TF. Redefining the Risk Quotient: A Generalized Framework for Fragility Analysis Across Study Designs. 2025 [cited 5 Oct 2025]. doi:10.5281/zenodo.17204904

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