A Unified Framework for Fragility Metrics in 2×2 Trials

Thomas F. Heston, MD
Department of Family Medicine
University of Washington, Seattle, WA, USA
ORCID: https://orcid.org/0000-0002-5655-2512
Contact email: theston@uw.edu

Date: 2025-09-20
doi: https://doi.org/10.5281/zenodo.17167247

Abstract

Statistical fragility metrics quantify the robustness of binary trial results beyond the p-value. We introduce a unified definitional framework for fragility measures applicable to two-group studies with a single binary endpoint represented as a 2×2 contingency table. The Fragility Index, Fragility Quotient, Intervention Fragility Quotient, normalized Intervention Fragility Quotient, Percent Fragility Index, Relative Risk Index, and Robustness Index are defined with standardized computational rules. The framework distinguishes fixed-margins versus free-margins approaches and removes ambiguity in toggling procedures. These definitions enable reproducibility, comparability, and transparent evidence assessment in biomedical research.

Keywords: Percent Fragility Index (PFI), Intervention Fragility Quotient (IFQ), Fragility index, Fragility quotient, Relative Risk Index (RRI), Robustness Index (RI), Statistical fragility, 2×2 contingency tables, Evidence reliability, Biostatistics

1. Introduction

The reliance on p-values in biomedical research has led to longstanding concerns about reproducibility and interpretability (1–3). Fragility analysis addresses concerns about medical research reproducibility by quantifying how easily a statistical conclusion changes under small data perturbations. Here we standardize seven fragility metrics for two-group, binary outcome studies: the Fragility Index (FI), Fragility Quotient (FQ), Intervention Fragility Quotient (IFQ), normalized IFQ (nIFQ), Percent Fragility Index (PFI), Relative Risk Index (RRI), and Robustness Index (RI). We fix contingency-table structure, toggle rules, test selection, and tie handling to enable reproducible application across studies. These metrics address distinct but complementary dimensions of fragility, from integer toggles to continuous reallocations and proportional rescaling.

2. Framework and Definitions

2.1 Scope

Applicable to any 2×2 contingency table comparing two study arms with binary outcomes. Rows = study arms; columns = outcome A and outcome B. Cells {a, b, c, d} denote {arm A with outcome A, arm A with outcome B, arm B with outcome A, and arm B with outcome B}.

2.2 General rules:

- Statistical significance is defined as two-sided $p \leq 0.05$.
- Fisher's exact test applies to FI/FQ/IFQ/nIFQ; Pearson χ^2 applies to PFI/RI.
- The RRI does not utilize significance testing.
- When thresholds cannot be crossed, report metric as "not attainable."
- Polarity. Only FI and its quotients depend on outcome coding. PFI, RRI, and RI are invariant to row/column relabeling. Fix the outcome definition once; any redefinition may change results.

2.3 Fragility Index (FI)

Define a single target outcome (either outcome A or outcome B) as an "event" and define the other outcome as a "non-event" before analysis and hold it fixed. FI is polarity-sensitive: swapping what is defined as an event with what is defined as a non-event can change the FI. FI is the smallest non-negative integer of within-arm toggles (event ↔ non-event) needed to cross $\alpha = 0.05$ by two-sided Fisher's exact test, with row totals fixed. Apply toggles exclusively to the arm with fewer events; if both arms have the same number of events, then apply toggles to the arm with fewer total cases; if still tied, then the two rows are identical and toggling either arm yields an identical FI. Stop toggling at the first crossing of $\alpha = 0.05$ (4). If crossing cannot occur, report "not attainable." Calculation of the FI is shown in Tables 1 and 2.

Only FI and its quotients inherit this sensitivity to how an event versus a non-event is defined. For example, if outcome A is initially defined as an event and outcome B as a non-event; but later outcome B is defined as an event and outcome A defined as a non-event; then the FI will sometimes change. PFI, RRI, and RI are not sensitive to column relabeling (PFI uses fixed margins with $|4x|$; RRI uses $|ad-bc|$; RI depends on χ^2 which is unchanged by column swaps). Reported FI values are conditional on the chosen outcome definition.

Examine Tables 1 and 2. Note that if T is defined as outcome A then the FI would equal 3, but if T was defined as outcome B then the FI would equal 2, indicating a sensitivity to polarity.

Table 1. Baseline condition prior to event toggling

Study Arm	Event	Non-event	Total
Arm A	3	10	13
Arm B	9	2	11
Total	12	12	24

note: Fisher's exact test, two-sided, p-value = 0.0123 for the table {3, 10, 9, 2}

Table 2. Toggling intervention to calculate FI

Study Arm	Event	Non-event	Total
Arm A	6	7	13
Arm B	9	2	11
Total	15	9	24

note: Fisher's exact test, two-sided, p-value = 0.105. The FI intervention is applied to the arm with fewer pre-determined events resulting in an FI of 3 which is added to cell a and subtracted from cell b to keep the row totals fixed. Note that if the pre-determined definition of an event was changed to equal outcome B, then the column labels would change, and toggling would be applied to Arm B. In this case the FI would be 2 instead of 3, i.e. {3, 10, 9, 2} at baseline to {3, 10, 7, 4} after toggling.

2.4 Fragility Quotient (FQ)

FI divided by total sample size (FI/n). Expressed as a percentage of all participants whose outcomes would need to change to flip statistical significance (5). Note that the FQ depends on study arm allocation, and can change if the sample sizes of each arm are significantly different.

2.5 Intervention Fragility Quotient (IFQ)

FI divided by the size of the arm undergoing intervention by toggling. Indicates arm-specific fragility and uses the toggled-arm as the denominator; it is less sensitive to allocation (6). Use nIFQ for cross-study comparison. Under equal allocation, the IFQ = FQ x 2.

2.6 Normalized IFQ (nIFQ)

IFQ divided by 2, to normalize with the FQ to allow direct comparison. The nIFQ equals FQ under 1:1 allocation.

2.7 Percent Fragility Index (PFI)

Minimal fractional reallocation of outcomes (continuous toggles, fixed margins) needed to flip significance under χ^2. Reported as a percent of total participants. The PFI is based upon the concept of a Unit Fragility Index, which toggles cells by integer values but unlike the FI, is applied to both study arms and keeps both column and row marginal totals fixed (7). Solve for the minimal real x that flips Pearson chi-square. The PFI = |4x|/n, where x = smallest real number (by absolute value) such that table {a, b, c, d} replaced by {a + x, b - x, c - x, d + x} flips the χ^2 p-value from ≤0.05 to >0.05 or vice versa (8). Note that all cells must remain ≥ 0 and that row and column sums are unchanged. If no feasible x exists, then report "PFI: not attainable."

2.8 Relative Risk Index (RRI)

The RRI equals |ad − bc| / n, representing distance from therapeutic neutrality without toggling. Note that the RRI does not depend upon calculation of p-values and is not affected by polarity (9).

2.9 Robustness Index (RI)

Scaling factor k > 1 required to flip significance by proportionally inflating or deflating all cells. For studies with findings that are non-significant, multiply each cell by k until the findings become significant. For statistically significant studies, divide each cell by k until the findings become non-significant. The RI captures fragility relative to study size (10).

2.10 Reporting

There currently is no consensus regarding cut-off values defining fragile results (11). The following thresholds are the author's suggested conventions to aid interpretation; they are not validated decision rules at this time but a starting point. It is important to always report raw values first before giving fragility metrics.

Table 3. Reporting

Metric	Suggested Thresholds	Interpretation Guide	Notes
FI	fragile if FI ≤ number lost to follow-up	Smaller = more fragile	Polarity-sensitive; report raw FI.
FQ	none	Smaller % = more fragile	FQ = FI / n (total).
IFQ	none	Smaller % = more fragile	IFQ = FI / n(toggled arm); use nIFQ to compare with FQ.
nIFQ	≤ 5% fragile	Rescaled IFQ on a total-sample basis; equals FQ under 1:1 allocation.	nIFQ = IFQ / 2; reduces allocation sensitivity.
PFI	≤ 5% fragile	Fixed-margins, continuous reallocations .	Polarity-invariant; Pearson χ^2 used.
RRI	none	0 = neutrality; larger = farther away .	Polarity-invariant; used to help guide clinical decision-making.
RI	≤ 2 fragile	Factor change needed to flip significance.	Changes with absolute sample size even if proportions stay the same.

3. Worked Examples

Pearson χ^2 p-values were calculated using the built-in Google Sheets function, and Fisher's exact p-values calculated using a JavaScript function. The FI, exact UFI, and RI were calculated using manual iteration with evaluation of the p-value after each step. All other values were derived by application of the formulas given previously.

3.1 Example: Statistically Significant at Baseline

Consider the baseline case {10, 40, 2, 48}. This table represents an equal allocation trial of 50 cases per arm, with 10 vs 2 experiencing outcome A which the researchers identified as an event; outcome B was defined as a non-event. Fisher's exact at baseline = 0.0277. Toggling the arm with fewest events results in a FI of 1, i.e. Fisher's exact applied to {10, 40, 3, 47} flips significance (p = 0.0713). FQ = 1/100 = 1% indicating that toggling applied to 1% of total cases would flip significance. The IFQ = 2% indicating that toggling 2% of the arm affected by the FI intervention, would flip significance. The nIFQ = 1% = FQ indicating equal allocation of cases. Solving for the minimal real x that flips the Pearson χ^2 results in 0.816 {9.184, 40.816, 2.816, 47.184} which is associated with a PFI of |0.816 * 4| / 100 = 3.26%. The RRI = 4 (10 * 48 - 40 * 2), and the RI = 1.578.

Interpretation: although the baseline case was statistically significant, the results appear to be fragile based on a PFI ≤ 5%, an nIFQ ≤ 5% and a RI ≤ 2.

3.2 Example: Statistically Non-Significant at Baseline

Consider the baseline case {23, 8, 11, 6}. This table represents an unequal allocation trial of 31 cases in arm A and 17 cases in arm B. There were 23 vs 11 cases experiencing outcome A which the researchers identified as an event; outcome B was defined as a non-event. Fisher's exact at baseline = 0.5225. Toggling the arm with fewest events results in a FI of 4, i.e. Fisher's exact applied to {23, 8, 7, 10} flips significance (p = 0.0323). FQ = 4/48 = 8.3% indicating that toggling applied to 8.3% of total cases would flip significance. The IFQ = 23.5% and nIFQ = 11.8%. Note that the FQ and nIFQ are not equal, reflecting the fact that there was unequal allocation present. The the minimal real x applied to all cells, keeping marginal totals fixed, that flips significance = 1.91 which results in a PFI of |1.91 * 4| / 48 = 15.9%. The RRI = 1.0417 and the RI = 8.031.

Interpretation: the baseline case was statistically non-significant. This finding appears to be robust based on a PFI of 15.9%, an nIFQ of 11.8%, and a RI of 8.031.

4. Applications

4.1 Clinical research

Fragility metrics improve interpretation of RCT findings, particularly when sample sizes are small or results are marginally significant.

4.2 Evidence synthesis

Meta-analyses can report fragility metrics to contextualize trial reliability.

4.3 Methodology

Fragility metrics provide tools for teaching, reporting standards, and risk of bias assessment (12).

5. Discussion

This definitional framework advances fragility analysis by unifying rules and resolving ambiguities. FI, FQ, IFQ, and nIFQ quantify discrete toggling fragility. PFI offers a continuous fixed-margins perspective. RRI isolates effect divergence, while RI scales fragility relative to study size. Together, they create a reproducible toolkit for quantifying robustness.

5.1 Recommended Use

Prioritize nIFQ and PFI. nIFQ reports the fraction of the toggled arm (rescaled to total-sample basis) required to flip significance, so it compares cleanly across unequal allocations. PFI gives a fixed-margins, continuous perturbation that is stable under outcome relabeling and reads more naturally when outcomes are ambiguous. Use RI as a confirmation metric to quantify how much proportional scaling of the entire table would change the statistical conclusion; it isolates scale effects without changing proportions.

5.2 Polarity and Invariance

FI and its quotients are polarity-sensitive and depend on how the outcome is labeled; this is why they are less suitable here. PFI, RRI, and RI are invariant to pure relabeling of rows or columns. Fix the clinical definition once, but prefer metrics that are robust to coding choices when polarity is debatable.

5.3 RRI for Decision Framing

Encode the clinically beneficial outcome as the “success” column. The sign of $(ad - bc)$ gives direction (benefit vs harm). The magnitude $|ad - bc|/n$ is the distance from neutrality, independent of p-values, and is useful when data are uncertain and stakeholders need a “more likely than not” directional read. Pair RRI with effect size and its 95% CI; use RI alongside n to show how conclusions depend on study scale.

5.4 Limitations

The FI, FQ, IFQ, nIFQ, and RRI are restricted to binary outcomes and limited to 2×2 tables. The PFI and RI can potentially be applied more broadly but this remains not established and unverified.

6. Conclusion

We provide definitions for seven fragility metrics. This framework standardizes computation, enhances reproducibility, and supports transparent reporting in clinical research. Adoption of these definitions will allow consistent fragility reporting across studies and accelerate methodological progress.

How to Cite

Heston TF. A unified framework for fragility metrics in 2×2 trials. Zenodo. 2025 Sep 20. doi:10.5281/zenodo.17167247.

License

Declarations

Submission status: Original methodological contribution. Funding: No specific funding received. Conflicts of interest: None declared.

Bibliography

1. Lu Y, Belitskaya-Levy I. The debate about p-values. Shanghai Arch Psychiatry. 2015 Dec 25;27(6):381–5. DOI: 10.11919/j.issn.1002-0829.216027. PMID: 27199532. PMCID: PMC4858512.
2. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005 Aug 30;2(8):e124. DOI: 10.1371/journal.pmed.0020124. PMID: 16060722. PMCID: PMC1182327.
3. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013 May;14(5):365–76. DOI: 10.1038/nrn3475. PMID: 23571845.

4. 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):622-628. GSM 2024 h5-index 74. DOI: 10.1016/j.jclinepi.2013.10.019. PMID: 24508144.

5. Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility index? Crit Care Med. 2016 Nov;44(11):e1142-e1143. GSM 2025 h5-index 85. DOI: 10.1097/CCM.0000000000001976.

6. Heston TF. Adjusting fragility metrics for unequal trial randomizations. Autoimmun Rev. 2025 Sep;103935. DOI: 10.1016/j.autrev.2025.103935.

7. Feinstein AR. The unit fragility index: an additional appraisal of "statistical significance" for a contrast of two proportions. J Clin Epidemiol. 1990;43(2):201-209. GSM 2024 h5-index 74. DOI: 10.1016/0895-4356(90)90186-s. PMID: 2303850.

8. Heston TF. Redefining significance: robustness and percent fragility indices in biomedical research. Stats. 2024 Jun 17;7(2):537-548. GSM 2024 h5-index 16. DOI: 10.3390/stats7020033.

9. Heston TF. Statistical Significance Versus Clinical Relevance: A Head-to-Head Comparison of the Fragility Index and Relative Risk Index. Cureus. 2023 Oct 26;15(10):e47741. GSM 2024 h5-index 96. DOI: 10.7759/cureus.47741. PMID: 37899890. PMCID: PMC10602368.

10. Heston TF. The robustness index: going beyond statistical significance by quantifying fragility. Cureus. 2023 Aug 30;15(8):e44397. GSM 2024 h5-index 96. DOI: 10.7759/cureus.44397. PMID: 37791215. PMCID: PMC10542213.

11. Garcia MVF, Ferreira JC, Caruso P. Fragility index and fragility quotient in randomized clinical trials. J Bras Pneumol. 2023 Mar 17;49(1):e20230034. DOI: 10.36416/1806-3756/e20230034. PMID: 36946820. PMCID: PMC10171299.

12. Javier JV, Corvi M, Sabo G, Chari R, Yendluri A, Corvi J, et al. Randomized controlled trials evaluating carpal tunnel release are statistically fragile: A systematic review. Hand (N Y). 2025 Jun 27;15589447251348504. DOI: 10.1177/15589447251348505. PMID: 40576202.