The Modified-Arm Fragility Quotient: An Improved Metric for Assessing Robustness in Clinical Trials

Version 2.0

Thomas F. Heston[1,2]

[1]Univ. of Washington School of Medicine, Seattle, WA, USA

[2]Elson S. Floyd College of Medicine, Washington State Univ., Spokane, WA, USA

ORCID: 0000-0002-5655-2512

November 2025

Version History

Version 2.0 (November 2025): Modified MFQ formula from FI/($2n_{\text{mod}}$) to (fragility count)/n_{mod} for 0–1 normalization (Neutrality Boundary Framework). For balanced trials using FI, MFQ/2 = FQ. All results updated accordingly.

Version 1.0 (October 2025): Initial publication.

Abstract

The Fragility Index (FI) quantifies the robustness of statistically significant results in randomized controlled trials by identifying the minimum number of outcome reversals required to change significance. The Fragility Quotient (FQ) normalizes FI by total sample size to enable cross-study comparisons. However, neither metric accounts for unequal randomization ratios, which are increasingly common in clinical trials. We introduce the Modified-arm Fragility Quotient (MFQ), defined as the fragility count divided by the sample size of the arm receiving outcome toggles during fragility calculation. Through mathematical analysis and simulation studies across 96 trial configurations (10,000 replications each), we demonstrate that MFQ maintains stability across allocation ratios (range 0.065–0.163 across ρ=0.5 to 3.0, 2.5-fold variation) while FQ varies more dramatically (range 0.019–0.109, 5.7-fold variation). We prove that MFQ is unbiased for allocation-invariant fragility while FQ introduces systematic bias under unequal allocation, though FQ achieves lower variance. Simulations demonstrate MFQ's practical superiority across realistic trial parameters. MFQ normalizes to a 0–1 scale consistent with the Neutrality Boundary Framework. For balanced trials using the Fragility Index as the fragility count, MFQ/2 = FQ, enabling straightforward conversion while maintaining standardized interpretation across allocation ratios. MFQ is a more accurate metric of fragility in trials employing unequal randomization.

Keywords: Fragility index, Randomized controlled trials, Unequal randomization, Clinical trial design, Statistical robustness

1 Introduction

Statistical significance testing via p-values remains the primary framework for evaluating randomized controlled trial (RCT) results[1], yet p-values alone provide limited information about result robustness[2]. The Fragility Index (FI), introduced by Walsh et al.[3], quantifies robustness by identifying the minimum number of outcome changes required to reverse statistical significance. This intuitive metric has been widely applied across medical specialties[4,5].

To enable comparisons across trials of different sizes, Ahmed et al.[6] proposed the Fragility Quotient (FQ), defined as FI divided by total sample size. While FQ successfully normalizes for sample size, it does not account for allocation ratios. Modern RCTs increasingly employ unequal randomization (e.g., 2:1, 3:1) to maximize patient access to novel therapies, reduce costs, or improve recruitment[7–9]. Under unequal allocation, FQ may provide misleading interpretations because it dilutes the modified-arm fragility signal with the unmodified arm size.

We address this limitation by introducing the *Modified-arm Fragility Quotient* (MFQ), defined as the fragility count divided by the sample size of the arm receiving outcome toggles during fragility calculation. The MFQ framework uses the Fragility Index as the standard fragility count, though it accommodates current or future modifications to the FI (such as label-invariant variants). MFQ normalizes to a 0–1 scale, enabling integration with the Neutrality Boundary Framework (NBF)—a unified system for measuring geometric distance from therapeutic neutrality (e.g., RR=1, Δ=0) across different data types[10]—for cross-domain robustness comparison.

Our contributions are threefold. First, we establish the mathematical relationship between FI, FQ, and MFQ, demonstrating their quantitative relationship under balanced allocation and analyzing their bias-variance tradeoff under unequal allocation (Section 2). Second, we demonstrate through simulations that MFQ maintains stable interpretation across allocation ratios while FQ varies systematically with randomization imbalance (Section 4). Third, we provide practical guidance for metric selection based on trial design characteristics (Section 5).

2 Mathematical Framework

2.1 Notation and Setup

Consider a two-arm RCT with binary outcome:

Define the allocation ratio: $\rho = n_I/n_C$.

Standard designs:

- Balanced: $\rho = 1$ (1:1 randomization)
- Unbalanced: $\rho \neq 1$ (e.g., $\rho = 2$ for 2:1)

	Event	No Event	Total
Intervention	a	b	$n_I = a + b$
Control	c	d	$n_C = c + d$
Total	$a + c$	$b + d$	$N = n_I + n_C$

2.2 Fragility Metrics

Definition 1 (Fragility Index)**.** The Fragility Index (FI) is the minimum number of outcome reversals (event $\leftrightarrow$ no event) in one arm required to change statistical significance, determined by Fisher's exact test with $\alpha = 0.05$ (two-sided).

By convention, reversals occur in the arm with fewer events; if tied then in the arm with smaller sample size[3]. Row totals remain fixed while column totals change during FI calculation.

Definition 2 (Fragility Quotient)**.**

$$\text{FQ} = \frac{\text{FI}}{N} = \frac{\text{FI}}{n_I + n_C} \tag{1}$$

Definition 3 (Modified-arm Fragility Quotient)**.** Let n_{mod} denote the sample size of the arm receiving outcome toggles during fragility calculation (the arm with fewer events, or if tied, the smaller arm), and let "fragility count" denote the number of outcome toggles required to reverse significance. Then:

$$\text{MFQ} = \frac{\text{fragility count}}{n_{\text{mod}}} \tag{2}$$

The standard fragility count is the Fragility Index (FI), though the MFQ framework accommodates current or future modifications to the FI (such as label-invariant variants). The 0–1 normalization enables integration with the Neutrality Boundary Framework.

2.3 Relationship Between Metrics

Theorem 1 (Relationship to FQ)**.** *Under balanced randomization ($\rho = 1$, $n_I = n_C = n$, $N = 2n$), when FI is used as the fragility count:*

$$MFQ = \frac{FI}{n} = \frac{2 \cdot FI}{2n} = \frac{2 \cdot FI}{N} = 2 \cdot FQ \tag{3}$$

Equivalently: $MFQ/2 = FQ$

Proof. When $n_I = n_C = n$, we have $N = 2n$ and $n_{\text{mod}} = n$. Therefore:

$$\text{MFQ} = \frac{\text{FI}}{n_{\text{mod}}} = \frac{\text{FI}}{n} = \frac{\text{FI}}{N/2} = 2 \cdot \frac{\text{FI}}{N} = 2 \cdot \text{FQ} \tag{4}$$

□

Theorem 2 (Behavior Under Unequal Allocation)**.** *For any allocation ratio ρ, when FI is used as the fragility count:*

$$FQ = \frac{n_{mod}}{N} \cdot MFQ \tag{5}$$

Specifically:

- *If the intervention arm is modified ($n_{mod} = n_I$):*

$$FQ = \frac{n_I}{n_I + n_C} \cdot MFQ = \frac{\rho}{1+\rho} \cdot MFQ \tag{6}$$

- *If the control arm is modified ($n_{mod} = n_C$):*

$$FQ = \frac{n_C}{n_I + n_C} \cdot MFQ = \frac{1}{1+\rho} \cdot MFQ \tag{7}$$

Which arm is modified depends on event frequencies, not allocation ratio. Under unequal allocation, FQ systematically underestimates or overestimates modified-arm fragility depending on which arm has fewer events.

Proof. By definition:

$$\text{FQ} = \frac{\text{FI}}{N} \tag{8}$$

$$\text{MFQ} = \frac{\text{FI}}{n_{\text{mod}}} \tag{9}$$

Therefore:

$$\text{FQ} = \frac{\text{FI}}{N} = \frac{n_{\text{mod}}}{N} \cdot \frac{\text{FI}}{n_{\text{mod}}} = \frac{n_{\text{mod}}}{N} \cdot \text{MFQ} \tag{10}$$

When $n_{\text{mod}} = n_I$: $\frac{n_I}{N} = \frac{n_I}{n_I+n_C} = \frac{n_I}{n_I+n_I/\rho} = \frac{\rho}{1+\rho}$
When $n_{\text{mod}} = n_C$: $\frac{n_C}{N} = \frac{n_C}{n_I+n_C} = \frac{n_I/\rho}{n_I+n_I/\rho} = \frac{1}{1+\rho}$ □

2.4 Bias-Variance Properties of MFQ and FQ

We now analyze the statistical properties of MFQ and FQ as estimators of allocation-invariant fragility.

Theorem 3 (Bias and Variance of FQ and MFQ)**.** *Let the target of inference be the allocation-invariant fragility*

$$\theta^\star := \frac{\text{FI}}{n_{\text{mod}}} \in [0,1], \tag{11}$$

where FI *is the fragility index and n_{mod} is the size of the arm to which toggles are applied. Define*

$$r := \frac{n_{\text{mod}}}{n_{\text{other}}} \geq 1, \qquad n = n_{\text{mod}} + n_{\text{other}}, \tag{12}$$

where n_{other} is the size of the other arm. Then:

1. *MFQ is unbiased:* $\text{Bias}(\text{MFQ}) = 0$
2. *FQ is biased for $r \neq 1$:* $\text{Bias}(\text{FQ}) = -\theta^\star/(r+1)$
3. *FQ has lower variance than MFQ:* $\text{Var}(\text{FQ}) < \text{Var}(\text{MFQ})$ *for* $r > 1$

Proof. Link both quotients to $\theta^\star$:

$$\text{MFQ} = \theta^\star \tag{13}$$

$$\text{FQ} = \frac{\text{FI}}{n} = \frac{\text{FI}}{n_{\text{mod}}} \cdot \frac{n_{\text{mod}}}{n} = \theta^\star \cdot \frac{1}{1 + n_{\text{other}}/n_{\text{mod}}} = \theta^\star \cdot \frac{1}{1 + 1/r} \tag{14}$$

Bias:

$$\text{Bias}(\text{MFQ}|\theta^\star) = 0 \tag{15}$$

$$\text{Bias}(\text{FQ} \mid \theta^\star, r) = \mathbb{E}[\text{FQ} - \theta^\star] = \theta^\star \left(\frac{1}{1 + 1/r} - 1 \right) = -\theta^\star \frac{1}{r+1}, \tag{16}$$

which equals 0 only at $r = 1$ and is strictly negative for $r > 1$.

Variance: Treat FI as a random integer under the data-generating process. Using scaling,

$$\text{Var}(\text{MFQ}) = \frac{\text{Var}(\text{FI})}{n_{\text{mod}}^2} \tag{17}$$

$$\text{Var}(\text{FQ}) = \frac{\text{Var}(\text{FI})}{n^2} = \frac{\text{Var}(\text{FI})}{n_{\text{mod}}^2 (1 + 1/r)^2} \tag{18}$$

Therefore

$$\frac{\text{Var}(\text{FQ})}{\text{Var}(\text{MFQ})} = \frac{1}{(1 + 1/r)^2} = \frac{r^2}{(r+1)^2} < 1 \tag{19}$$

for all $r > 1$, with limit 1 as $r \to \infty$.

Thus FQ has lower variance than MFQ, but is biased, while MFQ has higher variance but is unbiased. □

Corollary 4 (Bias-Variance Tradeoff)**.** *Under equal allocation ($r = 1$), $\text{MFQ} = 2 \cdot \text{FQ} = \theta^\star$ when FI is the fragility count, and both estimators are unbiased with proportional variances. For any unequal allocation ($r \neq 1$), MFQ remains unbiased while FQ becomes biased, but FQ achieves lower variance. The practical superiority of MFQ depends on whether bias reduction outweighs the variance increase in realistic trial settings.*

Remark 1. Theorem 3 establishes that MFQ and FQ represent different points on the bias-variance frontier. **The choice of $\theta^\star = \text{FI}/n_{\text{mod}}$ as the estimand reflects our position that allocation-invariant fragility—normalized to the modified arm—is the appropriate target for cross-trial comparison.** Given this target, MFQ is unbiased while FQ introduces systematic bias proportional to allocation imbalance. However, FQ achieves lower variance.

Across 96 simulated trial configurations (Section 4), MFQ demonstrates superior stability: 2.5-fold variation (0.065–0.163) versus 5.7-fold for FQ (0.019–0.109). This empirical evidence supports MFQ as the preferred metric for comparing trials with different allocation ratios, as the bias reduction consistently outweighs the variance increase in realistic settings.

Corollary 5 (FQ Divergence Under Imbalance)**.** *For any unbalanced trial ($\rho \neq 1$), when FI is the fragility count, FQ differs from MFQ by:*

- *Factor of $\frac{\rho}{1+\rho}$ if intervention arm modified*
- *Factor of $\frac{1}{1+\rho}$ if control arm modified*

The divergence increases as ρ moves further from 1 in either direction.

3 Interpretation and Clinical Meaning

3.1 MFQ Interpretation

MFQ measures fragility on a 0–1 scale consistent with the Neutrality Boundary Framework, enabling cross-domain comparison with other robustness metrics.

For balanced trials ($\rho = 1$) using FI as the fragility count: MFQ $= 2 \cdot$ FQ, providing straightforward conversion to traditional fragility quotient values.

For unbalanced trials: MFQ focuses on the modified arm (where toggles occur) while maintaining the 0–1 normalization that enables comparison across different allocation ratios and integration with other NBF metrics.

Example: MFQ = 0.10 indicates that 10% of the modified arm switching outcomes would eliminate significance. For a balanced trial using FI, this corresponds to FQ = 0.05 (5% of total sample).

3.2 Comparison Across Metrics

Consider a trial with $n_I = 100$, $n_C = 50$, experimental arm has 30 events, control arm has 10 events, FI = 5:

Metric	Value	Interpretation
FI	5	5 patients must switch
FQ	$5/150 = 0.033$	3.3% of total sample
MFQ	$5/50 = 0.10$	10% of modified arm (0–1 NBF scale)

Key insight: MFQ provides a normalized 0–1 measure that accounts for which arm receives toggles. For balanced trials, MFQ/2 enables direct comparison to traditional FQ values. Theorem 3 proves this normalization is statistically optimal.

4 Simulation Study

4.1 Design

We conducted Monte Carlo simulations to evaluate fragility metric behavior across trial designs.

Factors:

- **Allocation ratios**: $\rho \in \{0.5, 1.0, 2.0, 3.0\}$ (1:2, 1:1, 2:1, 3:1 randomization)
- **Intervention arm sizes**: $n_I \in \{50, 100, 200\}$
- **Baseline event rates**: $p_C \in \{0.2, 0.3, 0.4\}$
- **Effect sizes**: Odds ratio $\in \{1.5, 2.0, 3.0\}$
- **Replications**: 10,000 per condition

Note: Scenarios with $n_I = 50$ and OR $= 1.5$ were not run, yielding 96 conditions instead of the full $4 \times 3 \times 3 \times 3 = 108$ grid.

Total conditions: 96

Software Simulations were run in R 4.5.1 (R Foundation for Statistical Computing) using RStudio. Fisher's exact tests used `stats::fisher.test`.

Procedure:

1. Generate a 2×2 table from binomial distributions
2. Apply two-sided Fisher's exact test at $\alpha = 0.05$ (R `stats::fisher.test`)
3. If $p \leq 0.05$, compute FI, FQ, MFQ using canonical toggle rules (fixed row totals)
4. Analyze metric distributions

Terminology mapping In the simulation output, MFQ appears as `nIFQ`; thus MFQ = nIFQ = FI/n_{mod}. A legacy column `IFQ` corresponds to the conventional FQ.

4.2 Results

4.2.1 Metric Stability Across Allocation Ratios

Table 1 shows mean metric values across allocation ratios for $n_I = 100$, $p_C = 0.3$, OR $= 2.0$.

Table 1: Fragility metrics by allocation ratio

Allocation	FI	FQ	MFQ
1:2 ($\rho = 0.5$)	32.6	0.109	0.163
1:1 ($\rho = 1.0$)	6.5	0.033	0.065
2:1 ($\rho = 2.0$)	3.5	0.023	0.070
3:1 ($\rho = 3.0$)	2.5	0.019	0.075

Key finding: Across allocation ratios, FQ varies 5.7-fold (0.019 to 0.109), while MFQ varies only 2.5-fold (0.065 to 0.163).

For the canonical 1:1 allocation, MFQ $= 2 \cdot$ FQ ($0.065 = 2 \times 0.033$), confirming Theorem 1. As allocation becomes increasingly unbalanced, FQ varies systematically due to dilution by total sample size. For example, at $\rho = 0.5$ (1:2 randomization with more control patients),

FQ=0.109 reflects the large total sample, while MFQ=0.163 provides a normalized measure focused on the modified arm. At $\rho = 3.0$ (3:1 randomization), FQ=0.019 is depressed by the large total sample, while MFQ=0.075 maintains consistent 0–1 scale interpretation.

The stability of MFQ across allocation ratios (2.5-fold range vs 5.7-fold for FQ) demonstrates its superiority for comparing trials with different randomization schemes, empirically validating the bias-variance analysis in Theorem 3.

5 Discussion

5.1 When to Use Each Metric

Use MFQ when:

- Unequal randomization ($\rho \neq 1$)
- Comparing trials with different allocation ratios
- Standardized fragility reporting across diverse trial designs
- Meta-analysis of fragility across studies
- Unbiased estimation of allocation-invariant fragility is prioritized (Theorem 3)
- Integration with Neutrality Boundary Framework metrics

Use FQ when:

- Historical convention matters for comparison to prior literature using traditional FQ scale
- Total sample fragility perspective is specifically desired

Note: For balanced trials using FI as the fragility count, MFQ/2 = FQ enables direct conversion between metrics.

5.2 Practical Recommendations

For trial reporting, we recommend:

1. Always report FI (primary metric)
2. Report MFQ for all trials (balanced and unbalanced)
3. For balanced trials, note that MFQ/2 equals traditional FQ for historical comparison
4. Use MFQ as the standard for meta-analyses across different allocation ratios

The unbiasedness for allocation-invariant fragility established in Theorem 3, combined with empirical evidence of superior stability, provides strong justification for preferring MFQ in contemporary trial designs.

5.3 Integration with the Neutrality Boundary Framework

The 0–1 normalization range enables MFQ to integrate with the Neutrality Boundary Framework (NBF), a unified system for measuring geometric distance from therapeutic neutrality across different data types[10]. Within this framework, MFQ joins other 0–1 bounded metrics including Risk Quotient (RQ), Meaningful Change Index (MeCI), Distance to Independence (DTI), and ANOVA Neutrality Distance ($\text{ANOVA}_{\text{NBF}}$). This standardization facilitates comparison of evidence robustness across binary, continuous, and correlation-based outcomes.

The NBF emphasizes that robustness metrics should measure geometric distance from neutrality (e.g., RR=1, Δ=0) independently of study design choices like sample size or allocation ratio. MFQ achieves this by normalizing to the modified arm—the arm where outcome changes would move results toward or away from neutrality. This design-invariant approach enables cross-study comparison on a common 0–1 scale.

For balanced trials where FI is used as the fragility count, the relationship MFQ/2 = FQ provides a conversion factor for comparing to the traditional Fragility Quotient literature while maintaining NBF-consistent 0–1 scaling.

5.4 Limitations

MFQ measures fragility relative to the arm receiving toggles during fragility calculation. The identity of this arm depends on event frequencies, not study design. Trials where different arms would be modified under different outcome scenarios may require careful interpretation. The simulation was limited to binary outcomes; extension to survival and continuous outcomes requires future work. Additionally, our analysis assumes fixed row totals per canonical FI methodology; alternative toggle approaches may yield different metric behaviors.

6 Conclusion

The Modified-arm Fragility Quotient provides superior interpretability for unequally randomized trials while maintaining a quantitative relationship to FQ under balanced allocation (MFQ/2 = FQ when using FI). By normalizing to the modified arm on a 0–1 scale, MFQ enables direct comparisons across trials regardless of allocation ratio and integrates seamlessly with the Neutrality Boundary Framework. We have proven that MFQ is unbiased for allocation-invariant fragility while FQ introduces systematic bias under unequal allocation. Across 96 simulated trial configurations, MFQ demonstrates superior stability (2.5-fold variation) compared to FQ (5.7-fold variation), empirically validating its practical superiority. We recommend MFQ as the standard fragility measure for modern RCT designs employing any allocation ratio.

Data and Code Availability

R simulation code and complete results (CSV format) are available at Zenodo[11]: `https://doi.org/10.5281/zenodo.17386376`

References

[1] Goodman S. A dirty dozen: twelve P-value misconceptions. Seminars in Hematology. 2008 Jul;45(3):135-40. Available from: https://www.sciencedirect.com/science/article/pii/S0037196308000620.

[2] Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005 Aug;2(8):e124. Available from: http://dx.doi.org/10.1371/journal.pmed.0020124.

[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. Journal of Clinical Epidemiology. 2014 Jun;67(6):622-8. Available from: http://dx.doi.org/10.1016/j.jclinepi.2013.10.019.

[4] Tignanelli CJ, Napolitano LM. The Fragility Index in randomized clinical trials as a means of optimizing patient care. JAMA Surgery. 2019 Jan;154(1):74-9. Available from: https://doi.org/10.1001/jamasurg.2018.4318.

[5] Holek M, Bdair F, Khan M, Walsh M, Devereaux PJ, Walter SD, et al. Fragility of clinical trials across research fields: a synthesis of methodological reviews. Contemporary Clinical Trials. 2020 Oct;97:106151.

[6] Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility index? Critical Care Medicine. 2016 Nov;44(11):e1142-3. Available from: http://journals.lww.com/00003246-201611000-00043.

[7] Hey SP, Kimmelman J. The questionable use of unequal allocation in confirmatory trials. Neurology. 2014 Jan;82(1):77-9. Publisher: Wolters Kluwer. Available from: https://www.neurology.org/doi/abs/10.1212/01.wnl.0000438226.10353.1c.

[8] Berry DA. Adaptive bayesian clinical trials: the past, present, and future of clinical research. Journal of Clinical Medicine. 2025 Jan;14(15):5267. Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2077-0383/14/15/5267.

[9] Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. New England Journal of Medicine. 2017 Jul;377(1):62-70. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMra1510062. Available from: https://www.nejm.org/doi/full/10.1056/NEJMra1510062.

[10] Heston TF. The Neutrality Boundary Framework: Quantifying Statistical Robustness Geometrically. SSRN Electronic Journal. 2025 Oct. Available from: https://ssrn.com/abstract=5636470.

[11] Heston TF. Simulation Code and Results for "The Modified-arm Fragility Quotient: Normalizing Statistical Fragility for Unequal Randomization". Zenodo. 2025. Available from: https://zenodo.org/records/17386377.