Internet Medical Journal
ISSN 1093-7935 · medjournal.com · Est. 1996

Statistical Fragility of Saline Nasal Irrigation for Rhinosinusitis Is Incomplete Without Robustness Assessment

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	07 MAY 2026
DOI	10.5281/zenodo.20074314
Article type	Commentary
Citation	Heston TF. Statistical Fragility of Saline Nasal Irrigation for Rhinosinusitis Is Incomplete Without Robustness Assessment. Internet Medical Journal. 2026;1:e20074314

Abstract

A recent systematic review of randomized controlled trials evaluating saline nasal irrigation (SNI) for rhinosinusitis correctly identifies moderate-to-high statistical fragility across eight trials but cannot determine whether non-significant results represent genuine nil effects or underpowered detections of real treatment benefits. Statistical classification of significance using p-values is the primary metric analyzed. The stability of this classification, as measured by the fragility index, is a derived, secondary metric of statistical significance. While the systematic review addressed both of these metrics, it failed to take into account robustness — the distance from therapeutic neutrality, accounting for variability. Robustness provides the missing complement needed to distinguish near-null effects from fragile-but-real detections. Reporting the statistical evidence triplet of significance, fragility, and robustness in future SNI reviews would materially improve the interpretation of evidence for clinical decision-making in rhinosinusitis.

Keywords

statistical fragility, saline nasal irrigation, rhinosinusitis, neutrality boundary, robustness, fragility index, fragility quotient, evidence quality

A recent systematic review of randomized controlled trials (RCTs) evaluating high-volume saline nasal irrigation (SNI) for rhinosinusitis demonstrates moderate-to-high statistical fragility using the fragility index (FI) and fragility quotient (FQ), yet these metrics address only one of the three orthogonal dimensions of evidence quality: they measure the stability of statistical significance classification (the first dimension) but not distance from therapeutic neutrality (the third dimension), where the relative risk equals one (1). The authors acknowledge the limitation of their fragility analysis by highlighting that the FI measures threshold sensitivity rather than study validity (2,3). However, it provides no metric to fill the gap. The missing dimension is robustness, and without it, fragility analysis of rhinosinusitis SNI trials yields partial evidence that cannot resolve the most consequential clinical question the literature poses: whether the observed trial results are genuinely near-null or represent real treatment effects that the available trials lacked power to confirm.

Fragility metrics and robustness metrics measure distinct, non-substitutable properties of statistical evidence. The FI and reverse fragility index (rFI) count the minimum number of outcome reversals required to cross the significance threshold; the FQ and reverse fragility quotient (rFQ) normalize that count to sample size, placing fragility on a proportion scale (4,5). Together, FI and FQ quantify classification stability — the sensitivity of the reported p-value to small perturbations in the underlying outcome data. Robustness, by contrast, quantifies the geometric distance from the neutrality boundary, accounting for variability. Therapeutic neutrality is the point at which both trial arms produce identical event rates; the intervention contributes nothing to the outcome; the treatment and outcomes are independent of each other (6). The dimensions of fragility and robustness are orthogonal because a result can be simultaneously stable and near-neutral, or unstable and far from neutral. A high FQ confirms that significance is unlikely to flip with minor data changes; it cannot confirm that the observed effect is meaningfully separated from nil. A high robustness score (nb) confirms the separation from nil but makes no claim about classification stability. Complete statistical evidence, therefore, requires a triplet of evidence: statistical significance as determined by the p-value, stability of this significance classification, and distance from therapeutic neutrality (7).

The most clinically consequential finding of the systematic review exposes precisely this gap. Five of eight principal dichotomous outcomes were non-significant, yielding a median reverse FI (rFI) of 6 and a median reverse FQ (rFQ) of 0.082. The rFI correctly communicates that each non-significant result is a median of six outcome reversals from statistical significance, but it cannot resolve whether these five outcomes represent genuine nil effects of SNI or underpowered detections of a real treatment benefit. Two trials with identical rFI values can represent fundamentally different clinical realities depending on where their observed effects sit relative to the neutrality boundary of therapeutic equivalence.

The three statistically significant principal outcomes evaluated by the systematic review compound the problem from the opposite direction: FQ values spanning from 0.0165 (8) to

0.2121 (9) confirm substantial heterogeneity in fragility, yet neither value addresses whether the observed risk differences are robustly separated from nil or are closely adjacent to the boundary of therapeutic equivalence.

The neutrality boundary framework (NBF) helps resolve these issues by adding the missing dimension of robustness. For binary outcomes in independent two-arm RCTs — the design used across all eight SNI trials analyzed — the risk quotient (RQ) is the appropriate robustness metric. It is defined as RQ = |ad − bc| / (N^2/4), where {a,b,c,d} represent the four cells of the standard 2x2 contingency table (6). An RQ near zero indicates the observed risk difference is essentially indistinguishable from the neutrality boundary of therapeutic equivalence, consistent with a near-null treatment effect regardless of p-value classification or fragility. An RQ near one indicates robust separation from nil. Applied to the non-significant principal outcomes in the Sethi et al. dataset, RQ would immediately distinguish those with a genuinely near-null signal — where SNI appears to produce no differential benefit worth pursuing further — from those representing real treatment effects that the trials lacked power to confirm statistically. This distinction directly informs the prioritization of future SNI trial investment and the confidence with which guideline developers can endorse or withhold recommendations. Because RQ requires only published event counts and sample sizes — the exact same data used by the FI — it can be calculated retrospectively from the 2×2 contingency tables underlying the 38 outcomes extracted by the systematic review, without additional data collection.

Three patterns in the SNI dataset illustrate the practical stakes. Among the five non-significant principal outcomes, rFI values of 4 to 8 indicate that each result is relatively close to the significance boundary; however, rFI is silent on whether these results are also close to the neutrality boundary, which is the clinically relevant question once statistical significance has already been ruled out. Among the significant outcomes, Pynnonen et al.'s result carries an FI of 2, requiring just two outcome changes to eliminate significance. Yet this FI value does not reveal whether the result represents a genuine near-boundary detection or a real SNI effect recorded with minimal statistical margin. Rabago et al.'s result (FQ = 0.2121) is the most stable by FQ, but whether it is also the most robust requires RQ to confirm. In two of eight principal outcomes, loss to follow-up exceeded the FI or rFI, compounding the interpretive gap: dropout-sensitive findings are precisely those for which robustness assessment is most needed, to determine whether the underlying effects are geometrically meaningful or artifactually near the threshold.

Future fragility analyses of rhinosinusitis SNI trials — and systematic fragility reviews across otolaryngology more broadly — should utilize the p–fr–nb triplet: reporting significance (p), percentile-normalized fragility (fr, computed via the modified-arm fragility quotient [MFQ] or global fragility quotient [GFQ] for binary outcomes), and robustness (nb, computed via RQ) for every dichotomous outcome. Because RQ requires only published event counts and sample sizes, datasets utilized to calculate the FI can simultaneously calculate the RQ. Guideline developers at the American Academy of Otolaryngology-Head and Neck Surgery who rely on these RCTs to inform SNI recommendations would benefit

from knowing not only how many outcome reversals threaten each finding, but how far each finding lies from the position of therapeutic equivalence. Statistical significance and fragility are necessary but insufficient for assessing evidence quality; robustness is the orthogonal complement that converts a partial evidence report into a complete one.

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. Ronit Sethi BS, Kaan Oral BA, Rahul Guda AB, Shiven Sharma JD, Sujay Ratna BS, Olivia First BA, et al. The Statistical Fragility of Saline Nasal Irrigation for Rhinosinusitis: A Systematic Review. The Laryngoscope. 2026 Apr 29;lary.70601. doi:10.1002/lary.70601

2. Andrade C. The Use and Limitations of the Fragility Index in the Interpretation of Clinical Trial Findings. J Clin Psychiatry. 2020 Mar 24;81(2). doi:10.4088/JCP.20f13334

3. Cote MP, Lubowitz JH, Rossi MJ, Matzkin E. The Fragility Index Is Typically Misinterpreted and of Low Value: Clinical Trials Are Designed to Be Fragile. Arthrosc J Arthrosc Relat Surg. 2024 Nov 1;40(11):2635–7. doi:10.1016/j.arthro.2024.08.003

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):622–8. doi:10.1016/j.jclinepi.2013.10.019

5. Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility index? Crit Care Med. 2016 Nov;44(11):e1142–3. doi:10.1097/CCM.0000000000001976

6. Heston TF. The Neutrality Boundary Framework: Quantifying Statistical Robustness Geometrically. arXiv. 2025 Nov 2;2511.00982. doi:10.48550/arXiv.2511.00982

7. Heston TF. Significance, Fragility, and Robustness in Clinical Trials: Stratifying Statistical Evidence. Cureus. 2025 Dec 31;17(12). doi:10.7759/cureus.100494

8. Pynnonen MA, Mukerji SS, Kim HM, Adams ME, Terrell JE. Nasal Saline for Chronic Sinonasal Symptoms: A Randomized Controlled Trial. Arch Otolaryngol Neck Surg. 2007 Nov 1;133(11):1115. doi:10.1001/archotol.133.11.1115

9. Rabago D, Zgierska A, Mundt M, Barrett B, Bobula J, Maberry R. Efficacy of daily hypertonic saline nasal irrigation among patients with sinusitis: a randomized controlled trial. J Fam Pract. 2002 Dec;51(12):1049–55. PubMed PMID: 12540331.