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

Resampling Fragility in the MUTTON-HF Randomized Clinical Trial of an Indigenous Food Is Medicine Intervention

Thomas F. Heston

Affiliation	Department of Family Medicine, University of Washington, Seattle, USA
Affiliation	Department of Medical Education and Clinical Sciences, Elson S. Floyd College of Medicine, Washington State University, Spokane, USA
ORCID	0000-0002-5655-2512
Published	15-Sept-2026
DOI	10.5281/zenodo.22784136
Article type	Perspective
Citation	Heston TF. Resampling Fragility in the MUTTON-HF Randomized Clinical Trial of an Indigenous Food Is Medicine Intervention. Internet Medical Journal. 2026;1:e22784136

Abstract

Statistical fragility is not a single quantity. Resampling fragility estimates the probability that a trial's significance classification would reverse in a new sample of the same size, while perturbation fragility measures the minimum change to the recorded data that reverses the classification in the observed table. The primary endpoint of the Medically Utilized Tailored Traditional Foods to Optimize Nutrition in Heart Failure (MUTTON-HF) randomized clinical trial carries a resampling fragility of 0.38 under a within-arm binomial replication model, and a global fragility quotient of 0.0097 (unstable). These fragility metrics indicate that the significance classification behind this clinically promising result is easily reversed and therefore warrants cautious interpretation.

Keywords

resampling fragility, perturbation fragility, statistical fragility, global fragility quotient, complete statistical evidence, heart failure, food is medicine

Introduction

Critics have questioned whether the fragility index provides information beyond the p-value [1,2]. Evaluating that criticism requires distinguishing what different fragility measures describe and whether those quantities can be recovered from the p-value alone. Resampling fragility is the estimated probability that a significance classification would reverse in a new sample of the same size under a specified replication model. Perturbation fragility is the minimum change to the recorded data, under a fixed analytical rule, that moves the observed result across the significance threshold (Heston, 2026). The first describes the chance of a different significance classification in a hypothetical repeat study; the second counts the changes to recorded outcomes needed to reverse the current classification, with the permitted changes, statistical test, and significance threshold fixed. These measures answer different questions, but that distinction alone does not establish whether either adds information beyond the p-value. The Medically Utilized Tailored Traditional Foods to Optimize Nutrition in Heart Failure (MUTTON-HF) trial, a randomized clinical trial of a culturally and medically tailored meal program for adults with heart failure in rural Navajo Nation, provides a current and instructive setting in which to compute both forms on the same primary endpoint [3].

The p-value quantifies the compatibility of the observed data with the null hypothesis under the specified statistical model; however, it does not measure the probability that the null hypothesis is true, the magnitude of an effect, or the practical importance of a result [4]. The fragility index introduced a complementary intuition for binary trials, counting the within-arm outcome changes needed to move a significant two-by-two table across the 0.05 boundary [5]. A redundancy critique soon followed, arguing that because fragility indices correlate strongly with p-values across trials, the index is the familiar quantity in new dress [1]. The reported correlation supports a close relationship, but correlation alone does not establish that the fragility index adds no information. The relevant question is whether the p-value alone determines how many recorded outcomes must change to reverse statistical significance. That question is distinct from estimating the probability that significance would reverse in a repeat study.

The taxonomy of statistical fragility resolves the ambiguity [6]. Resampling fragility may be estimated by bootstrap resampling or Monte Carlo simulation under a specified population model, with the analysis and threshold held fixed; its output is a probability, and its value depends on the chosen replication model. Perturbation fragility is computed directly from the observed counts under a fixed test; its output is a count or a bounded quotient, and it requires no model beyond the test itself. Within the p–fr–nb framework, complete statistical evidence is the triplet of the p-value measuring significance, a native

fragility quotient, fr, measuring classification stability, and a neutrality-boundary robustness metric, nb, measuring distance from therapeutic neutrality [7].

The MUTTON-HF primary endpoint illustrates both forms. Hospitalization or emergency department visit within 90 days occurred in 43 of 106 intervention patients (40.6%) versus 57 of 100 control patients (57.0%), a relative risk of 0.72 (95% confidence interval, 0.54–0.96) (Eberly et al., 2026). With the analytical rule fixed at a two-sided Fisher's exact test and a 0.05 threshold, the observed table yields p = 0.025, a significant baseline classification.

Calculating Resampling Fragility

The MUTTON-HF trial's primary outcome was hospitalization or emergency department visit (all-cause). In the intervention group (culturally and medically tailored meal program), 43 events occurred in 106 subjects; in the control group (usual dietary advice), 57 occurred in 100.

Resampling fragility was computed under a simple replication model: each arm is redrawn at its observed event rate, with arm sizes fixed, and the two-sided Fisher's exact test is recomputed for each replicate sample. Summing the probabilities of all replicate tables that fail to reach statistical significance gives a resampling fragility of 0.38. Under this model, approximately 38 of every 100 same-size replications would lose statistical significance, while 62 would retain it. The online calculator used for this computation is located at https://www.fragilitymetrics.org

Calculating Perturbation Fragility

Perturbation fragility speaks in the units of the observed data. The global fragility index (GFI), the minimum number of cell-to-cell reallocations that reverses the classification, equals 2; recording two additional intervention-arm patients as reaching the endpoint raises p to 0.051. Any cell-to-cell transfer is allowed, the two-sided Fisher’s exact test is recalculated after each transfer, and the GFI is the smallest number of transfers that carries the p-value across the 0.05 threshold.

For the MUTTON-HF trial, the baseline contingency table is {43, 63, 57, 43}. This corresponds to a baseline two-sided Fisher’s exact p-value of 0.025 (significant). Two transfers yield the GFI witness table {45, 61, 57, 43}, with a two-sided Fisher’s exact p-value of 0.051 (nonsignificant).

The global fragility quotient (GFQ), the GFI divided by the sample size, is 2/206 = 0.0097. Under the framework’s GFQ < 0.01 criterion, the significance classification is unstable.

Discussion

The two fragility assessments work together, and each answers a question clinicians already ask. Resampling fragility answers the replication question: if the same trial were run again, would the finding hold [8]? Here, nearly four in ten repeat trials would lose statistical significance. Perturbation fragility answers the data integrity question: how close did this trial come to a different statistical significance classification? Here, within two patients: had two intervention-arm outcomes been recorded differently, the result would have been nonsignificant. Specifically, reclassifying two intervention-arm patients from no event to an event would reverse statistical significance under the fixed test.

None of this diminishes the trial, which stands as a model of community-designed pragmatic research, from Native-sourced meals and local food-hub delivery to prespecified subgroup results consistent in direction. Fragility assessment measures evidence quality; it does not render a verdict. These calculations quantify the current finding's sensitivity to sampling variation and changes in recorded outcomes, supporting further confirmation alongside the expansion and economic analyses the investigators describe [3]. The reporting practice that makes such statements possible is simple: name the form of fragility, state the analytical rule, threshold, baseline classification, and permitted change, and for a resampling estimate state the sampling model as well. A reversal probability and an outcome-edit count answer different questions about the same table, and evidence assessment is better served by reporting both under their own names than by debating which one the fragility index really is.

Declarations

Funding: This study received no external funding.

Conflicts of Interest: The author reports no conflicts of interest.

Data Availability: Not applicable.

Research Ethics Statement: Not applicable. This perspective did not involve human subjects research, animal research, or protected health information.

AI Usage: Large language models were used for coding, language editing, and formatting assistance; the author reviewed, verified, and is fully responsible for all content.

References

1. Carter RE, McKie PM, Storlie CB. The Fragility Index: a P-value in sheep's clothing? Eur Heart J. 2017;38: 346–348. doi:10.1093/eurheartj/ehw495

2. Li J, O'Connell PJ. The Fragility Index: The P-Value by Another Name? Transplantation. 2022;106: 239. doi:10.1097/TP.0000000000003806

3. Eberly LA, George C, Sandman S, Bex D, Jones K, Yazzie A, et al. An Indigenous Food Is Medicine Intervention: The MUTTON-HF Randomized Clinical Trial. JAMA Intern Med. 2026;186: 1077. doi:10.1001/jamainternmed.2026.2879

4. Wasserstein R, Lazar N. The ASA Statement on p -Values: Context, Process, and Purpose. Am Stat. 2016;70: 129–133. doi:10.1080/00031305.2016.1154108

5. 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

6. Heston TF. Resampling Fragility, Perturbation Fragility, and Why the Global Fragility Index Is Not a P-Value in Disguise. Internet Med J. 2026;1: e22059146. doi:10.5281/zenodo.22059146

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

8. Van Zwet EW, Goodman SN. How large should the next study be? Predictive power and sample size requirements for replication studies. Stat Med. 2022;41: 3090–3101. doi:10.1002/sim.9406