

Chapter 11

Statistical Fragility and Robustness in Pharmaceutical Trials

Thomas F. Heston^{1,2*}

¹Department of Family Medicine, University of Washington, Seattle, Washington, USA.

²Department of Medical Education and Clinical Sciences, Washington State University, Spokane, Washington, USA.

*Corresponding Author.

Introduction

A statistically significant trial is supposed to mean something durable: the finding would stand if the experiment were run again. Over the past two decades, that assumption has failed repeatedly under direct examination [1,2]. Empirical replication programs confirmed the concern: when 100 published psychology experiments were repeated with high-powered designs, 97% of the original studies had reported statistically significant results, but only 36% of the replications did [3]. The pharmaceutical industry encountered the same problem from the inside, as systematic in-house attempts to reproduce published preclinical findings succeeded in only a minority of cases [4,5]. In clinical medicine and therapeutics, the stakes of evidentiary instability and medical reversal are concrete: regulatory authorizations, prescribing patterns, and clinical practice guidelines directly depend on the reliability and replication of published trials [6].

Statistical fragility bears on the reproducibility crisis in a direct, logical way. An examination of 399 randomized controlled trials in high-impact journals found that the median number of outcome changes needed to erase statistical significance, the fragility index, was 8; in a quarter of the trials it was 3 or fewer, and in 53% it was smaller than the number of patients lost to follow-up [7]. A trial whose significance classification hinges on a handful of outcomes lies near the decision boundary, and ordinary sampling variation is enough to push a repeat trial across it: under the normal approximation, a trial significant at $p = 0.049$ is expected to yield a nonsignificant result in roughly half of identical repetitions [8]. It follows that fragile findings are likely overrepresented among the findings that fail to replicate: not because such trials are biased, but because their significance classification was never stable to begin with. This connection is stated here as a deduction from the arithmetic of the significance threshold; its contribution to observed replication failure rates has not yet been quantified empirically.

Amyotrophic lateral sclerosis (ALS) supplies a recent, concrete instance. In the phase 2 CENTAUR trial, sodium phenylbutyrate–taurursodiol (AMX0035) slowed functional decline, and the result carried the drug to regulatory approval in the United States and Canada in 2022 [9,10]. The trial's composite secondary outcome of death, tracheostomy, or hospitalization was reported as not significantly different between groups, yet that comparison was one patient away from statistical significance: reclassifying a single placebo-group participant from no event to an event would have carried it across the 0.05 threshold, with the robustness effect estimate itself well separated from no difference. An endpoint read as a settled null was, in fact, an open question. The confirmatory phase 3 PHOENIX trial then showed no change in the primary functional outcome at 48 weeks, and the sponsor voluntarily withdrew the drug from the market [11]. As of October 2026, the registry record for PHOENIX lists no posted results (ClinicalTrials.gov, NCT05021536), leaving the evidence that removed the drug harder to examine than the evidence that won its approval. Marginal evidence converted into binary verdicts, a nonsignificant result read as no effect, replication left to settle what the first report could not, and incomplete reporting at the end: each element of this account is the subject of this chapter.

This chapter approaches the problem through Heston's p–fr–nb framework, which replaces the solitary p-value with complete statistical evidence: the p-value for significance, a fragility coordinate (fr) for the stability of the significance classification, and a robustness coordinate (nb) for the distance of the result from therapeutic neutrality [12]. The sections that follow first set out a taxonomy of statistical fragility, distinguishing four forms that are often conflated; then introduce statistical robustness as a second, independent axis of evidence quality; and finally assemble the three coordinates into complete statistical evidence and situate them against the familiar effect-size measures: confidence intervals, relative risk, absolute risk reduction, and the number needed to treat.

A Taxonomy of Statistical Fragility

Statistical fragility describes the stability of the significance classification: how easily a trial's verdict of significant or nonsignificant would change. The term covers several distinct constructs, and conflating them has generated avoidable controversy. The taxonomy is organized by a single question: which component of the study is allowed to vary while the classification is reassessed? Figure 1 shows the taxonomy of statistical fragility.

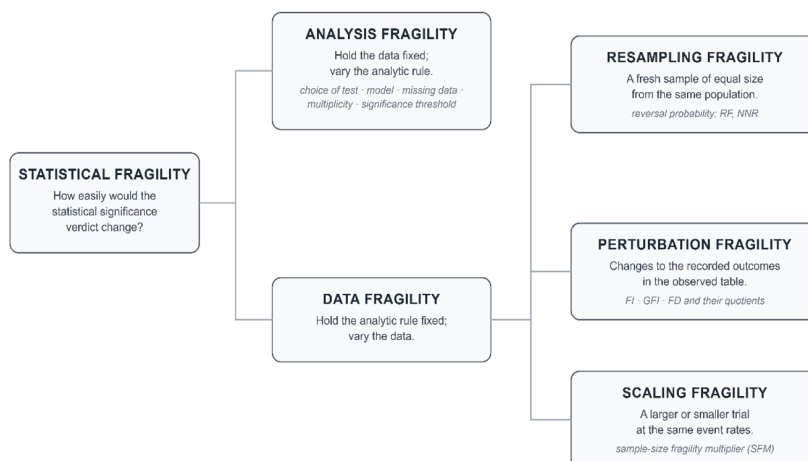


Figure 1: A taxonomy of statistical fragility

Legend: Statistical fragility has two branches: fragility of choice of analysis and fragility of the data. Data fragility has three components: resampling, perturbation, and scaling fragility.

The first division in the taxonomy separates analysis fragility from data fragility. In analysis fragility, the data stay fixed and the analytical rule changes: the choice of test, the adjustment set, the handling of missing data, the multiplicity correction, the stopping rule, or the significance threshold itself. A result that survives only one specific analytic path is unstable in a way no additional patient data can repair, which is why analysis fragility is logically the first form to examine. For a clinician, analysis fragility answers a practical question: would this trial still be positive if the statisticians had made different, equally defensible choices?

In data fragility, the analytical rule stays fixed, and the data vary, and the data can vary in three distinct ways [13].

1. Resampling fragility is the estimated probability that the significance classification would reverse in a new sample of the same size drawn from the same population; it is a probability, it concerns hypothetical replicate samples, and it requires a specified replication model.
2. Perturbation fragility counts the recorded outcomes in the observed table that must change before the classification flips; it is a count, or a proportion when normalized, and it is a property of the observed table alone.
3. Scaling fragility asks how much the amount of data would have to change, with the event proportions, allocation, and analytic rule held fixed, before the classification flips. These are four forms of statistical fragility, not four estimates of one quantity, and a trial can be stable under one form and unstable under another without contradiction.

The distinction carries practical weight. A recurring critique holds that fragility measures are a p-value in disguise because they correlate strongly with the p-value across trials. The critique applies to resampling fragility, which under the normal approximation is a function of the p-value alone, but not to perturbation fragility: published trials matched on their p-values have been shown to differ by factors of 2 to 30 in the global fragility quotient (GFQ), the proportion of the sample that must be reallocated to flip the classification [13]. The practical rule is therefore to name the form of fragility before reporting a value. Within the p-fr-nb framework, the fragility coordinate fr is always a native perturbation-fragility quotient: for a two-arm binary-outcome trial, the modified-arm fragility quotient (MFQ), which normalizes the outcome changes to the arm actually modified and thereby remains fair across unequal randomization ratios [14], or the GFQ for the general case. Higher values of fr indicate a more stable classification; lower values indicate fragility. For the reader of a trial report, fr answers a question the p-value does not: how many recorded patient outcomes stand between this conclusion and its reversal?

Robustness in Practice: Distance from Therapeutic Neutrality

Every trial design has a point of therapeutic neutrality, the value at which treatment and control do not differ: a relative risk or odds ratio of 1, a hazard ratio of 1, a mean difference of 0, a correlation of 0. Statistical robustness, in the sense used throughout this chapter, is the geometric distance of the observed effect estimate from that point. The definition of robustness in clinical trials departs deliberately from the purely mathematical definition of insensitivity to modeling assumptions or outliers [15,16]. Here robustness answers a clinical question: how far is this result from no effect at all? A result can be statistically significant yet lie close to neutrality, significant because the trial was large rather than because the effect is substantial.

The neutrality boundary framework makes the distance concrete. Each design family receives a robustness metric that expresses separation from therapeutic neutrality on a bounded scale from 0 to 1, and most take the form shown in the formula below:

Neutrality Boundary Formula

$$nb = \frac{|T - T_0|}{(|T - T_0| + S)}$$

Where, T is the observed summary statistic, T_0 is its value at therapeutic neutrality, and S is a fixed scale constant that does not depend on sample size. The normalization is design-specific: metrics for continuous, diagnostic, correlation, ordinal, and survival outcomes follow this form directly, whereas the two-by-two risk quotient normalizes the observed separation against its theoretical maximum.

Every such metric is bounded between 0 and 1, equals 0 exactly at neutrality, and is computed from the point estimate rather than from its precision, so nb carries no information about sampling uncertainty and adds none to the significance test.

For a two-by-two table, the robustness metric is the risk quotient (RQ), equal to $|ad - bc| / (N^2/4)$ for cell counts a, b, c, d and total sample size N . Simulation-derived bands classify results as near neutrality ($RQ < 0.075$), intermediate (0.075 to 0.227), or far from neutrality ($RQ \geq 0.227$), with the cutoffs stated as provisional conventions rather than validated thresholds. The construction parallels the number needed to treat: it adds no information beyond the effect estimate, but repackages that estimate onto a bounded, decision-ready scale.

Fragility and robustness are two separate axes, never opposite ends of one scale. The fragility coordinate is claim-invariant: a stable classification strengthens confidence in whatever verdict the trial reached, significant or not. The robustness coordinate is claim-dependent: a far result supports the claim that an effect exists, while a near result supports the claim that no meaningful effect exists. The combination matters most for nonsignificant trials. A nonsignificant result that is both stable and near neutrality is affirmative evidence of no effect; it is a different object, clinically, from an underpowered trial that merely failed to reach significance.

Complete Statistical Evidence

Conventional trial reporting treats the p -value, the confidence interval (CI), and an effect-size measure such as the relative risk (RR), the absolute risk reduction (ARR), or the number needed to treat (NNT) as a complete account of the evidence. The p -fr- nb framework defines this as partial evidence: it quantifies significance and magnitude but says nothing explicit about the stability of the significance classification or the standardized distance from neutrality. Complete statistical evidence is defined as the triplet p -fr- nb : the p -value for significance, the native fragility quotient fr for classification stability, and the neutrality-boundary robustness metric nb for distance from therapeutic neutrality. Each coordinate answers a question a skeptical reader asks. Is the result significant? Would a small change in the recorded outcomes reverse that verdict? Is the estimated effect meaningfully separated from no effect?

The p -value and the effect-size measures cannot carry all three answers. A p -value of 0.049 and a p -value of 0.0001 are both reported as statistically significant, yet they describe different evidential situations: at $p = 0.049$ about one in every two identical repeat trials is expected to reverse the verdict, while at $p = 0.0001$ roughly 37 repeats are expected per reversal. The confidence interval does not supply the missing information, because for the common Wald-type constructions the 95% CI excludes the null value precisely when $p < 0.05$; the interval restates the significance verdict and adds a range of magnitudes, which is valuable, but it does not state how many recorded outcomes separate the trial from the opposite conclusion, and trials matched on p differ manyfold on exactly that quantity, as shown above. The effect-size measures have the complementary limitation: an RR of 0.60 conveys magnitude, but not whether the significance classification resting on it would survive the reclassification of two patients, and an ARR or NNT inherits whatever instability the underlying counts contain. Three distinct questions cannot be answered by statistics constructed to answer one of them.

The triplet constitutes complete statistical evidence because its coordinates cover the three inferential dimensions of a trial result, while the conventional effect-size measures retain a separate and necessary role: nb captures relative, standardized magnitude on a bounded scale, whereas the ARR, the NNT, and gains in survival express absolute magnitude in clinical units. Complete statistical evidence plus the absolute effect size constitutes complete evidence for a result. Note that distance from the neutrality boundary (robustness) helps determine whether or not the findings represent a true difference; effect size is used to determine if the difference is clinically meaningful.

Empirical validation supports the stratification the triplet provides. Among 77 statistically significant trials in a 129-trial validation sample, 39.0% met concordant-positive criteria ($p \leq 0.05$, $MFQ > 0.10$, $RQ \geq 0.227$), a pattern that arose in only 1.4% of significant results under null simulation, a 27.9-fold enrichment; 31.2% showed the significant-unstable-near pattern, in which the significance classification was unstable and the separation from neutrality minimal [12]. For the clinician, the triplet converts a binary verdict into a graded one: a significant, stable, far result warrants a different level of confidence than a significant result that one patient's outcome would overturn.

Analysis Fragility

The choice of statistical test to analyze a pharmaceutical trial outcome can change the significance classification. This is analysis fragility. An example of this is a trial looking at the effect of adding methotrexate versus placebo to corticosteroid therapy in giant cell arteritis [17]. The completion of treatment analysis found that 7 out of 15 in the methotrexate group and 15 out of 18 in the placebo group suffered a relapse. This difference is significant by the two-sided Pearson chi-square test ($p = 0.026$) but nonsignificant by two-sided Fisher's exact test ($p = 0.061$). Here, with a single set of outcomes (7/15 vs 15/18), the significance classification differs depending upon which statistical test is selected for analysis.

Resampling Fragility

Resampling fragility (RF) is the estimated probability that the original significance classification would reverse in a new sample of the same size drawn under a specified data-generating process, with the analysis and significance threshold held fixed. Where perturbation fragility counts edits to the observed table, resampling fragility concerns hypothetical replicate samples: it is a probability, not an outcome-edit count, and it depends on the specified replication model. For this reason, it is recognized in the taxonomy of statistical fragility but sits outside the operational p -fr- nb triplet. Its companion, the number needed to reverse ($NNR = 1/RF$), restates the same probability as the number of identical repeat trials expected to produce one reversal of the significance classification — the replication analog of the number needed to treat. NNR is claim-symmetric: for a significant baseline it counts expected repeats until a nonsignificant result, and for a nonsignificant baseline the reverse.

Resampling Fragility and Number Needed to Reverse Formulas

Under the normal approximation, RF is a function of the p-value alone, at any sample size:

$$z = \Phi^{-1} \left(1 - \frac{p}{2} \right)$$

$$RF = \Phi(1.96 - z) - \Phi(-1.96 - z)$$

(significant baseline; the complement for a nonsignificant baseline)

$$NNR = \frac{1}{RF}$$

Where, p = the observed two-sided p-value of the baseline trial. ϕ = the standard normal cumulative distribution function; Φ^{-1} is its inverse. z = the standard normal deviate corresponding to p , $z = \phi^{-1}(1 - p/2)$. RF = resampling fragility, the probability that an identical repeat trial reverses the significance classification; the formula shown applies to a significant baseline, and its complement applies to a nonsignificant baseline. NNR = number needed to reverse, the expected number of identical repeat trials per one reversal. The constant 1.96 is the standard normal critical value at the two-sided 0.05 threshold. The closed form is a normal approximation; for counted outcomes under Fisher's exact test, it is approximate, with deviations confined to small sparse tables.

Resampling fragility can be visualized by considering taking a beaker of fluid (the sample) from a lake (the population). Repeated samples will vary slightly, and with regularity flip the underlying statistical significance (Figure 2). The frequency of flipping statistical significance is closely related to the underlying p-value.

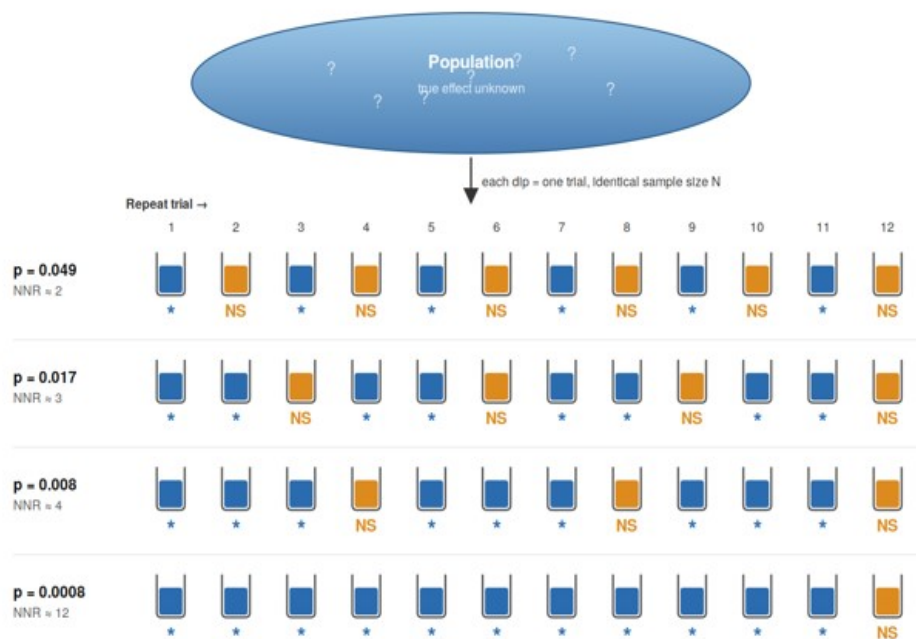


Figure 2: Resampling fragility: repeat trials as repeated samples from the same population

Legend: * = statistically significant. NS = nonsignificant. Each beaker is one trial of identical size N drawn from the same population; the true effect stays unknown to the investigator. One reversal per NNR repeat trials is expected on average. Reversals are drawn at regular intervals for clarity only: each repeat trial reverses with probability $RF = 1/NNR$ at random, not on a fixed schedule.

The frequency of reversal of the p-value significance classification varies in a manner closely tied to the base p-value (Table 1).

Table 1: Resampling Fragility

p-value	~ NNR	1	2	3	4	5	6	7	8	9	10	11	12
0.049	2	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
0.017	3	*	*	NS	*	*	NS	*	*	NS	*	*	NS
0.008	4	*	*	*	NS	*	*	*	NS	*	*	*	NS
0.0008	12	*	*	*	*	*	*	*	*	*	*	*	NS

Legend: * = statistically significant, NS = nonsignificant. Columns 1–12 represent identical repeated trials of the same size. Each row shows the expected frequency of reversal for the given baseline p-value: one reversal per NNR trials on average. The pattern is schematic; actual reversals occur at random with probability RF per trial, not at fixed intervals. NNR values are computed from the p-value under the normal approximation and rounded to the nearest integer.

Perturbation Fragility

Perturbation fragility quantifies the stability of a statistical significance classification by asking how much the observed data must change before that classification reverses. Given a completed trial summarized as a contingency table (Table 2), a perturbation metric counts the recorded outcomes that must be altered — toggled within an arm, reallocated between cells, or added to or removed from the table — to move the two-sided Fisher’s exact p-value across the 0.05 threshold. The count may be reported raw or normalized to a 0–1 quotient expressing the proportion of relevant data that must change. A result that reverses after modification of a single outcome rests on weaker evidence than one requiring many, even when both report identical p-values. Perturbation fragility is one of several distinct forms of statistical fragility — alongside analysis fragility (sensitivity to analytic choices), resampling fragility (probability of reversal in a replicate sample), and scaling fragility (sensitivity to sample size with proportions fixed) — and is the form measured by the metrics in this section. These metrics differ only in which perturbations they admit: the constraint set defines the metric. Tables 3 through 6 present the allowed moves for each, in order of progressively relaxed constraints, from the single-arm, single-direction toggling of the Walsh index to the unconstrained unit edits of the fragility distance.

Table 2 presents the standard contingency table for a two-arm trial with a binary outcome. The four cells record the observed counts: a and b are the events and non-events in the intervention arm, c and d are the events and non-events in the control arm, with total sample size $N = a + b + c + d$. All fragility metrics described below operate directly on these observed counts, and all assess statistical significance with the two-sided Fisher’s exact test.

Table 2: Contingency table for a two-arm binary-outcome trial

Group	Event	Non-event
Intervention	a	b
Control	c	d

Legend: a, b, c, and d are the observed cell counts; $N = a + b + c + d$.

The Walsh fragility index (Table 3) is the original 2014 metric and is defined procedurally: in a table significant at baseline, non-events are converted to events, one at a time, in the arm with fewer events, with the p-value recalculated after each toggle until $p \geq 0.05$. The index is the count that this prescribed procedure yields. It is defined only for significant baselines, restricts modification to a single arm and a single direction, and makes no claim of minimality.

Table 3: Allowed moves for the Walsh fragility index

Group	Event Non-event
Intervention	$a \leftarrow b$
Control	$c \leftarrow d$

Legend: The Walsh fragility index is defined only for tables that are statistically significant at baseline (two-sided Fisher’s exact $p < 0.05$). Toggles convert non-events to events, one at a time, in the arm with fewer events; the behavior when event counts are tied is not defined. Toggles are restricted to a single arm and a single direction. The p-value is recalculated after each toggle, and the process stops when the two-sided Fisher’s exact p-value is ≥ 0.05 . The index is defined procedurally, not as a minimum over all allowed toggles.

The Heston fragility index (Table 4) modifies the Walsh index in three respects: it is bidirectional, permitting toggles from non-event to event or the reverse; it is defined for both significant and nonsignificant baselines, measuring the distance to the 0.05 boundary from either side; and it is defined as a true minimum over all allowed toggles rather than as a procedural count. Modification remains restricted to a single arm — the arm with fewer events or, if tied, the smaller arm.

Table 4: Allowed moves for the Heston fragility index

Group	Event Non-event
Intervention	$a \leftrightarrow b$
Control	$c \leftrightarrow d$

Legend: The Heston fragility index is defined for both significant and nonsignificant baseline tables, flipping the significance classification in either direction. Toggles are allowed from a non-event to an event or vice versa (bidirectional). Toggles are restricted to a single arm — the arm with fewer events or, if tied, the smaller arm — but not to a single direction. The index is the minimum number of toggles that moves the two-sided Fisher’s exact p-value across the 0.05 threshold, defined as a true minimum rather than a procedural count.

The Lin fragility index (Table 5), as implemented in the fragility R package [18], relaxes the single-arm restriction: toggles remain within-arm, preserving both arm sizes, but may be distributed across either or both arms. The index is the minimum total number of toggles, summed across arms, required to move the p-value across the 0.05 threshold.

Table 5: Allowed moves for the Lin fragility index

Group	Event Non-event
Intervention	$a \leftrightarrow b$
Control	$c \leftrightarrow d$

Legend: The Lin fragility index is defined for both significant and nonsignificant baseline tables. Toggles are allowed from a non-event to an event or vice versa (bidirectional), within either or both arms, with both arm sizes preserved. Toggles are restricted to within-arm

changes but not to a single arm or a single direction. The index is the minimum total number of toggles, summed across arms, that moves the two-sided Fisher's exact p-value across the 0.05 threshold.

The global fragility index and the fragility distance (Table 6) remove the within-arm restriction entirely. The global fragility index is the minimum number of cell-to-cell reallocations — each decrementing one cell and incrementing another, holding N fixed — required to flip the significance classification; moves between arms are allowed. The fragility distance discards the fixed-N constraint as well: its unit is the addition or removal of a single subject in any cell, and it equals the minimum number of such unit edits required to cross the boundary. Each successive relaxation can only shorten the path to the boundary, yielding the ordering $FD \leq 2 \cdot GFI$ and $GFI \leq FI \leq \text{Walsh FI}$, where all are defined.

Table 6: Allowed moves for the global fragility index and fragility distance

Group	Event	Non-event
Intervention	$a \pm 1$	$b \pm 1$
Control	$c \pm 1$	$d \pm 1$

Legend: The global fragility index is the minimum number of cell-to-cell reallocations that moves the two-sided Fisher's exact p-value across the 0.05 threshold. Each reallocation decreases one cell by one and increases another by one, so the total sample size (N) remains fixed. Reallocations are not restricted to arm or direction, and between-arm moves are allowed. The fragility distance does not reallocate subjects but adds or removes one subject at a time from any cell, with N free to change; it is the minimum number of such unit edits that moves the two-sided Fisher's exact p-value across the 0.05 threshold. Both metrics are defined for significant and nonsignificant baseline tables.

Scaling Fragility

Scaling fragility asks the counterfactual question the other data-fragility forms leave open: with the event proportions, the allocation ratio, and the analytic rule all held fixed, how much would the amount of data have to change before the significance classification flips? Its metric is the sample-size fragility multiplier (SFM), the smallest factor greater than 1 by which the total sample size must be multiplied (for a nonsignificant baseline) or divided (for a significant baseline) to cross the threshold, scored by the Pearson chi-square test without continuity correction. Because scaling the sample at fixed proportions scales the chi-square statistic by the same factor, $SFM = \chi^2/3.841$ for a significant baseline and $SFM = 3.841/\chi^2$ for a nonsignificant baseline. When $\chi^2 = 0$, no finite increase in sample size can produce significance while the proportions remain fixed. For the reader of a trial report, the SFM states how oversized or undersized the trial was relative to the minimum needed for the classification it reached.

The SFM is a secondary metric, not a substitute for the fragility coordinate fr: it describes a hypothetical trial of a different size, not changes to the recorded outcomes in front of the reader. Its value lies in translation. An SFM of 377 reads as a study 377 times larger than significance required, a plain-language rendering of an extreme p-value; an SFM just above 1 marks a classification that crossed the threshold with almost no margin of sample size to spare.

The CENTURY trial illustrates the calculation [19]. The trial randomized 1028 patients with chronic coronary artery disease to comprehensive care or standard care, and 11-year all-cause mortality was 24 of 515 (4.7%) with comprehensive care versus 42 of 513 (8.2%) with standard care. Recalculation from these counts gives Pearson chi-square = 5.321, $p = 0.0211$, and Fisher's exact $p = 0.0222$. The SFM is $5.321/3.841 = 1.385$: the trial was approximately 1.39 times the size needed to reach the significance threshold. Equivalently, proportionally shrinking the trial from 1028 to approximately 742 participants, a 27.8% reduction with event proportions and allocation fixed, would place the mortality result at the 0.05 boundary. A greater proportional reduction would make the result nonsignificant.

Robustness in Clinical Trials

Statistical robustness, the nb coordinate of the triplet, is the geometric distance of the observed effect estimate from therapeutic neutrality, expressed on a bounded 0-to-1 scale. It is computed from the point estimate alone, not from its precision, so it carries no information about sampling uncertainty and is unaffected by sample size. For a two-arm binary-outcome trial the metric is the risk quotient (RQ), with results banded as near neutrality ($RQ < 0.075$), intermediate, or far from neutrality ($RQ \geq 0.227$). A far result describes an effect estimate well separated from no effect; a near result describes one lying almost on top of it.

Robustness matters because the p-value entangles effect separation with sample size, and the two failure modes this creates are opposite. A very large trial can declare significance for an effect sitting almost at neutrality, producing a statistically positive result of little clinical consequence. A small trial can return a nonsignificant p-value for an effect estimate far from neutrality, and that nonsignificance is then read, wrongly, as evidence of no effect. Because nb is independent of sample size, it separates these cases. Its interpretation is claim-dependent: a far nb supports the claim that an effect exists, a near nb supports the claim that none does, and a nonsignificant p-value paired with a far nb supports neither conclusion — it marks the trial as inconclusive.

A small trial illustrates the second failure mode [20]. A multicenter, double-blind, randomized trial compared etanercept with placebo in 17 patients with biopsy-proven giant cell arteritis and corticosteroid-related side effects. The primary outcome, the ability to withdraw corticosteroid therapy while controlling disease at 12 months, was reached by 4 of 8 patients (50%) on etanercept versus 2 of 9 (22.2%) on placebo, a nonsignificant difference (Fisher's exact $p = 0.335$). The robustness metric tells a different part of the story: $RQ = |4.7 - 4.2| / (17^2/4) = 0.277$, in the far band. The estimate lies well away from therapeutic neutrality, but 17 patients cannot stabilize a significance classification around it. The nonsignificant-far pattern marks the trial as inconclusive and consistent with underpowering, not as evidence of no effect — the same judgment the authors reached narratively in concluding that the limited number of patients precluded definitive conclusions and that etanercept warranted evaluation in a larger trial. The RQ formalizes that judgment as a number available to any reader.

The Complete Evidence Framework: a Worked Example

The DELI transition trial compared insulin degludec/liraglutide with multiple daily insulin injections in adults with type 2 diabetes after hospital discharge. Among 64 analyzed participants, hypoglycemia occurred in 2 of 32 patients receiving the combination treatment and 10 of 32 receiving multiple daily injections during four weeks of follow-up. The outcome was at least one episode lasting more than 15 minutes with glucose below 3.0 mmol/L, documented by continuous glucose monitoring [21].

Significance Classification

Table 7 shows the results from this two-arm binary outcome trial. The two-sided Fisher exact p-value = 0.022 and Pearson's χ^2 p-value = 0.010. Yates-corrected χ^2 p-value = 0.025.

Table 7: Patients experiencing at least one hypoglycemic episode

Group	Experienced at least one hypoglycemic episode	No hypoglycemic episodes
IdegLira	2	30
Multiple daily insulin injections	10	22

Legend: Values are patient counts during four weeks after hospital discharge. IDegLira = insulin degludec/liraglutide. Hypoglycemia was defined as glucose below 3.0 mmol/L for more than 15 minutes, documented by continuous glucose monitoring. Each group contained 32 patients.

All common significance tests for 2×2 contingency tables are concordant, indicating statistical significance. Analysis fragility is low.

Fragility

The fragility calculation asks how easily that classification changes. Reclassifying one treatment-group patient from no hypoglycemia to hypoglycemia changes the comparison to 3/32 versus 10/32 and raises Fisher's p-value to 0.0596 (Table 8). In this case, the Walsh, Heston, and Lin FI as well as the GFI all equal one. The GFQ = $1/64 = 0.0156$. By definition, since only one toggle is required to change the statistical significance classification, the significance classification is considered unstable.

Independent binomial resampling, using the observed event rates and two-sided Fisher's exact test, gives a resampling fragility of 0.346 and an NNR of 2.9. Under this model, about one in every three repeat trials of the same size would be classified as nonsignificant. This is consistent with Table 1, which shows that a p-value of 0.017 has an NNR of approximately 3. In this trial, both perturbation fragility (GFI) and resampling fragility (NNR) suggest the significance classification is unstable.

Table 8: Patients experiencing at least one hypoglycemic episode

Group	Experienced at least one hypoglycemic episode	No hypoglycemic episodes
IdegLira	3	29
Multiple daily insulin injections	10	22

Legend: Hypothetical perturbation of the observed counts in Table 7. One IDegLira patient is reclassified from no hypoglycemia to hypoglycemia; both group sizes remain unchanged. This single outcome toggle increases the two-sided Fisher's exact p-value from 0.0221 to 0.0596, reversing the significance classification at $\alpha = 0.05$.

Robustness

The neutrality boundary is where the relative risk (RR) = 1. In this case, the relative risk of the observed data = $(2/32) / (10/32) = 0.20$. For a 2×2 contingency table, quantifying distance from the neutrality boundary (robustness) is done by the risk quotient (RQ). $RQ = |2 \times 22 - 30 \times 10| / (64^2/4) = 4 * |44 - 300| / (64^2) = 4 * 256 / 64^2 = 1024 / 4096 = 0.25$. Based on a Monte Carlo simulation of 720,000 trials [12], an RQ of 0.25 places the observed effect in the far-from-neutrality band under the framework's provisional thresholds. Taken together, the p-fr-nb triplet is: significant-unstable-far.

Effect size

The absolute risk reduction (ARR) = $10/32 - 2/32 = 0.25$, or 25 percentage points, giving an estimated number needed to treat (NNT) = $1/0.25 = 4$ over four weeks. The effect size suggests a clinically meaningful result.

Complete evidence summary

Together, this analysis shows that the treatment (IdegLira) compared to standard care (multiple daily injections of insulin) is statistically significant, but unstable, as the significance classification could be flipped if the outcome of a single patient was changed. The results are far from the neutrality boundary, consistent with a true effect being present. The RQ helps distinguish an apparent treatment effect that is close to no difference from one that shows substantial separation between the groups. But, RQ does not indicate whether or not the effect is

clinically meaningful. Is the difference observed large enough to matter to patients? Looking at effect size, it appears that in this trial, the effect is clinically meaningful.

This example demonstrates why significance, fragility, and robustness are required for complete statistical evidence: the p-value gives a significance classification, a fragility metric indicates whether that classification is stable or unstable, and the robustness analysis indicates whether the estimated effect is likely real. Complete evidence requires a subjective component: whether or not the size of the effect is clinically meaningful. The triplet assesses the statistical evidence for an effect (i.e. it provides complete statistical evidence); the absolute effect size helps establish its clinical importance. The p-fr-nb triplet plus effect size is thus defined as complete evidence.

Summary

The problem this chapter set out is that conventional reporting treats the p-value, with its confidence interval and an effect-size measure, as the whole of the statistical evidence, when it answers only one of the three questions a trial result raises. A p-value states whether the result crossed the significance threshold. It does not state how easily that classification would reverse, and it does not state how far the estimated effect lies from therapeutic neutrality. Because trials sitting at the edge of the decision boundary reverse under ordinary sampling variation, evidence reported as a bare significance verdict feeds directly into the reproducibility problem the chapter opened with. The p-fr-nb triplet is the proposed solution: the p-value for significance, a native fragility quotient for classification stability, and a neutrality-boundary robustness metric for distance from no effect. Together these constitute complete statistical evidence, and with the absolute effect size in clinical units they constitute complete evidence for a result.

The taxonomy of statistical fragility makes the fr coordinate precise, and the precision matters. Analysis fragility, resampling fragility, perturbation fragility, and scaling fragility are four distinct constructs, not four estimates of one quantity, and each answers a different counterfactual: a different analytic choice, a repeat trial, altered recorded outcomes, or a different sample size. Naming the form before reporting a value dissolves the recurring objection that fragility is a p-value in disguise; that critique holds for resampling fragility, which is a transform of p, and fails for perturbation fragility, where trials matched on p differ manyfold. The fr coordinate of the triplet is always a perturbation measure, and it therefore adds information the p-value does not contain.

Robustness supplies the dimension that remains missing even after stability is settled. Because nb is computed from the point estimate and is independent of sample size, it separates the two opposite failures the p-value invites: the large trial whose significant result lies nearly on top of neutrality, and the small trial whose nonsignificant result lies far from it. Its interpretation is claim-dependent, which is what lets the triplet distinguish affirmative evidence of no effect from a trial that was merely too small to decide.

The incentive to adopt the triplet is practical. Every metric in the framework is model-free, computed from the published counts or summary statistics with no raw data, simulation, or refitting, so any reader of a trial report can complete the evidence the authors began. What the reader gains is a graded verdict in place of a binary one: a significant, stable, far result is solid evidence, with all three coordinates concordant. In contrast, a discordant triplet — a significant result that a single outcome would overturn and that lies close to the neutrality boundary — should be viewed skeptically, regardless of the p-value.

Complete statistical evidence in pharmaceutical trials is when significance, fragility, and robustness (the p-fr-nb triplet) are all reported together. The statistical analysis is incomplete otherwise.

Declaration

Acknowledgments: This work was undertaken in memory of Jon Bagby. Currently, amyotrophic lateral sclerosis remains a relentlessly progressive disease for which there is no cure. Jon's courage in facing ALS brought its human consequences into sharper focus and helped inspire this work.

References

- [1] Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Med.* 2005;2: e124. doi:10.1371/journal.pmed.0020124.
- [2] Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol Sci.* 2011;22: 1359–1366. doi:10.1177/0956797611417632.
- [3] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science.* 2015;349: aac4716. doi:10.1126/science.aac4716.
- [4] Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature.* 2012;483: 531–533.
- [5] Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov.* 2011;10: 712–712. doi:10.1038/nrd3439-c1.
- [6] Prasad V, Vandross A, Toomey C, Cheung M, Rho J, Quinn S, et al. A Decade of Reversal: An Analysis of 146 Contradicted Medical Practices. *Mayo Clin Proc.* 2013;88: 790–798. doi:10.1016/j.mayocp.2013.05.012.
- [7] 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.
- [8] Goodman SN. A comment on replication, P-values and evidence. *Stat Med.* 1992;11: 875–879. doi:10.1002/sim.4780110705.
- [9] Ketabforoush A, Faghihi F, Azedi F, Ariaei A, Habibi MA, Khalili M, et al. Sodium Phenylbutyrate and Tauroursodeoxycholic Acid: A Story of Hope Turned to Disappointment in Amyotrophic Lateral Sclerosis Treatment. *Clin Drug Investig.* 2024;44: 495–512. doi:10.1007/s40261-024-01371-1.

- [10] Paganoni S, Macklin EA, Hendrix S, Berry JD, Elliott MA, Maiser S, et al. Trial of Sodium Phenylbutyrate-Taurursodiol for Amyotrophic Lateral Sclerosis. *N Engl J Med*. 2020;383: 919–930. doi:10.1056/NEJMoa1916945.
- [11] Wen X, Lan T, Su W, Cao B, Wang Y, Chen Y. Latest progress and challenges in drug development for degenerative motor neuron diseases. *Neural Regeneration Research*. 2026;21: 1849–1863. doi:10.4103/NRR.NRR-D-24-01266.
- [12] Heston TF. Significance, Fragility, and Robustness in Clinical Trials: Stratifying Statistical Evidence. *Cureus*. 2025;17. doi:10.7759/cureus.100494.
- [13] 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.
- [14] Heston TF. Adjusting fragility metrics for unequal trial randomizations. *Autoimmun Rev*. 2025;24: 103935. doi:10.1016/j.autrev.2025.103935.
- [15] Box GEP. NON-NORMALITY AND TESTS ON VARIANCES. *Biometrika*. 1953;40: 318–335. doi:10.1093/biomet/40.3-4.318.
- [16] Huber PJ, Ronchetti EM. *Robust statistics*. 2nd ed. Hoboken (N.J.): Wiley; 2009.
- [17] Jover JA, Hernández-García C, Morado IC, Vargas E, Bañares A, Fernández-Gutiérrez B. Combined Treatment of Giant-Cell Arteritis with Methotrexate and Prednisone. *Ann Intern Med*. 2001;134: 106–114. doi:10.7326/0003-4819-134-2-200101160-00010.
- [18] 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.
- [19] Gould KL, Johnson NP, Roby AE, Kirkeeide R, Haynie M, Nguyen T, et al. Optimal medical care and coronary flow capacity-guided myocardial revascularization vs usual care for chronic coronary artery disease: the CENTURY trial. *Eur Heart J*. 2025;46: 3273–3286. doi:10.1093/eurheartj/ehaf356.
- [20] Martínez-Taboada VM, Rodríguez-Valverde V, Carreno L, Lopez-Longo J, Figueroa M, Belzunegui J, et al. A double-blind placebo controlled trial of etanercept in patients with giant cell arteritis and corticosteroid side effects. *Ann Rheum Dis*. 2008;67: 625–630. doi:10.1136/ard.2007.082115.
- [21] Gomez-Medina AM, Henao-Carillo DC, Villamil-Castaneda LP, Gomez-Quesada Y, Muñoz-Velandia OM, Yepes CA, et al. Insulin DEgludec/LIraglutide versus multiple daily insulin injections in the transition from hospital to outpatient management assessed by continuous glucose monitoring: the DELI transition trial. *Diabetologia*. 2025;68: 1899–1907. doi:10.1007/s00125-025-06446-y.