The Fragility Index

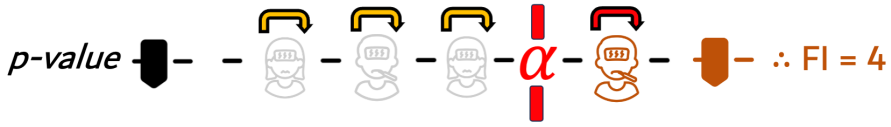
Key Features:	Key considerations:
- Fragility is the number of participants needed to switch a statistical outcome. - e.g., how many N can switch groups until significant ($p < 0.05$) $\mapsto$ n.s. ($p > 0.05$)	- higher FI is better! - the alpha criterion ($p < 0.05$, 0.01?) - what counts as a clinical event? - the sample size

by Garston Liang, Alexander Thorpe, & Quentin F. Gronau

1 Tipping the scales

If one doesn't count their local grocer, perhaps the most recognisable set of scales are held by Lady Justice, or Justitia. The Roman goddess has travelled long and far to sit atop courthouses across the world and, most remarkably, all without the help of her eyes. In our modern consciousness, her blindfold is what makes the scales trustworthy. It symbolises impartiality across wealth, power, and status for each and for all. Yet, could even this famous blindfold be just another fashion trend?

Fragility Index = How many people need to change places?



Atop the Old Bailey courthouse in London, Lady Justice does not wear a blindfold nor do her depictions upon early Roman coins. In fact, historians have traced the trend to be a 16th century addition before which, the scales had always been well within Justitia's view. And surely this makes for sounder judgement. Not only can one feel whether the scales are even, but only without the blindfold can one see the quality of the evidence on trial. A touch-up here with a chance event there, and the entire balance may no longer feel so secure.

In this primer, we examine the Fragility Index (FI) that measures whether a p-value-based conclusion would be revised should one or two pieces of evidence change places. Often, the p-values of RCT's are put to trial where the FI is used to understand whether the substantial resource commitment has yielded reliable outcomes. Perhaps, when armed with such a sword, it could be prudent to take off the blindfold and ask, when does the balance tip?

2 The Fragility Index & Fragility Quotient

Imagine an RCT finds that Treatment A is superior to Treatment B with a $p = 0.04$, but should a *single* person have swapped outcomes, then $p > 0.05$. That's fragile!

The Fragility Index calculates the number of participants needed to change the 'significance' of a p-value-based test, i.e., $p < 0.05$ to $p > 0.05$ or vice-versa. The FI is calculated by re-evaluating a p-value after an individual's clinical outcome is re-coded from one clinical category (e.g., recovered after treatment) to another (e.g., still sick). After each change, the p-value is compared against a nominal alpha value (typically, $\alpha = 0.05$) and so, the FI is the total number of participant-changes to cross to the other side of the alpha criterion.

As a metric, the Fragility Index is directly interpretable as a number of participants. A lower FI indicates that the conclusions of the study hinges upon a small number of participants whereas a higher FI indicates that the conclusions are robust against many changes in clinical outcomes.

There are a number factors that affect the FI but prime amongst these is the sample size. How one interprets the fragility of an FI = 4 will be different if N = 20 vs. 200. To address this gap, the Fragility Index can be converted into a Fragility Quotient (FQ) which is FI/N. In effect, this normalisation over sample size permits a comparison between studies with different *N's*. FQ values are constrained to values between 0:1, as compared to integer FI values, and higher FQ values indicate the p-value-based conclusions require a greater proportion of the sample to change places. As an example with $FI = 4$, Study A with $N = 20$ has a $FQ = 4/20 = 0.2$, whereas Study B with $N = 200$ has a $FQ = 4/200 = 0.02$.

The Fragility Index has enjoyed a recent resurgence in the medical literature as an aid for understanding the outcomes of RCT's. The typically large-N alongside statistical corrections for confounding factors are why conclusions from an RCT provide the clearest and largest pool of evidence for a treatment's efficacy. Surprisingly, however, many RCT's have low FI's. Recently, Xing and Lin (2023) surveyed clinical studies in the Cochrane Library and found that 20% of statistically significant results hinged on the outcome of a single participant $(FI = 1)$. This stark contrast between methodological strength and statistical fragility in RCT's adds recent calls for better statistical understanding in the medical field.

3 What is fragility?

Although the Fragility Index is numerically easy to interpret, where 'fragility' precisely begins and ends is hard to know. It may be relatively easy to see an $FI = 1$ and agree that much of the interpretation hinges upon a single outcome. However, should we learn that this clinical outcome is known *a priori* to have an extremely low prevalence, it becomes harder to imagine how many rare events would make a convincing FI.

Critics have argued that dichotomising fragility mirrors the problems of p-values by placing undue weight upon 'significance'. In fact, the comparison to significance testing is more than skin-deep. Statistically, the Fragility Index is inversely correlated with p-values and suffers from the same malaise. Sufficiently large samples will yield statistical significance even without a meaningful difference. Along a similar vein, the FI is not a measure of effect size and so a change in an individual's outcome status (i.e, from recovered to sick, or vice-versa) also ignores the magnitude of the person's symptoms or recovery. Like the p-value, the Fragility Index should be considered as only one index of robustness amongst many. Below, we detail a range of FI-related factors to consider when deciding whether to trust the robustness of a study's conclusions.

4 Factors affecting fragility

Outcome types & effect sizes A Fragility Index is straightforward for binary, discrete outcomes. For example, RCT's measuring the incidence of death in control vs. treatment arms can easily re-code an outcome from one category to another, bearing in mind the FI is entirely silent on all the shades of grey between death and health. For continuous outcomes, such as scores in a psychometric assessment, the FI can be applied with minor adjustments. One continuous version adjusting the measured outcomes, would be to discretise the continuous outcome variable, e.g., scores below 5/10 are unhealthy, scores higher are healthy). This option requires justification of why one chose a particular threshold and the trade-offs in information loss when collapsing the continuum. A second continuous option adjusting group membership is to randomly select a single participant to switch groups and calculate the FI. The same process is then bootstrapped and repeated many times, and across increasing numbers of participants. In effect, this provides a 'fragility distribution' where for each numeric participant-change, a distribution of p-values is obtained that can be compared to α. In either case, it is worth noting the FI focuses on primary outcomes and, so, secondary outcomes of RCT's are ignored.

Alpha, choice of statistical test, & reverse fragility The FI is tied to the choice of alpha criterion. While convention is that $\alpha = 0.05$, different α would yield a different number of outcome-changes to alter the outcome's significance. This is worth considering when comparing FI/FQ across RCT's with different thresholds. Similarly, the exact choice of statistical test can affect what the p-value that is derived. P-values for discrete outcomes could be generated from Fisher's exact test, Chi-square tests, or odds/relative-risk ratios. In fact, when the same data are converted from one format to another, it is possible to have $FI = 0$ where *only* one test yields $p < 0.05$.

As a final note, the Fragility Index is not exclusively tied to RCT's with 'significant' outcomes. One can imagine the reverse application for non-significant results, momentarily putting aside the overtones of post-hoc power calculations. Speculatively, one could examine the number of outcome-changes that would have produced a significant result. For the literature more broadly, the transparency requirements of funding agencies has likely bolstered the number of published studies once destined for the file drawer. Such work could be suited for a *reverse* Fragility Index.

Recommendations Just like with p-values, interpreting only a single metric is risky. Our recommendation is to examine the context contributing to fragility, such as the data-generating process or how the RCT classifies outcomes, to understand better the quality of the evidence on trial.

5 Further reading

- Large-scale fragility index assessment of clinical RCT's in Cochrane Library:
  Xing, A., & Lin, L. (2023). Empirical assessment of fragility index based on a large database of clinical studies in the Cochrane Library, *Journal of Evaluation in Clinical Practice*, 29, 359-370. Click for URL
- The Fragility Index and specific case examples for RCT's in different specialisations:
  Tignanelli, C. J., & Napolitano, L. M. (2019). The fragility index in randomized clinical trials as a means of optimizing patient care. *JAMA surgery*, 154, 74-79. Click for URL
- Reply to #2 highlighting shortcomings of FI: Acuna, S. A., Sue-Chue-Lam, C., & Dossa, F. (2019). The fragility index-P values reimagined, flaws and all. *JAMA surgery*, 154, 674-674. Click for URL