A Comparative Analysis of Unit Fragility and the Relative Risk Index

Thomas F Heston[1,2]

[1]Department of Family Medicine, University of Washington School of Medicine
[2]Department of Medical Education and Clinical Sciences, Elson S. Floyd College of Medicine, Washington State University

August 27, 2023

Abstract

BACKGROUND: In biostatistics, evaluating fragility is crucial for understanding their vulnerability to miscategorization. One proposed measure of statistical fragility is the unit fragility index (UFI), which measures the susceptibility of the p-value to flip significance with minor changes in outcomes. Although the UFI provides valuable information, it relies on p-values, which are arbitrary measures of statistical significance. Alternative measures, such as the fragility quotient (FQ) and the percent fragility index, have been proposed to decrease the UFI's reliance on sample size. However, these approaches still rely on p-values and thus depend on an arbitrary cutoff of $p < 0.05$. Instead of quantifying fragility by relying on p-values, this study evaluated the effect of small changes on relative risk. METHODS: Random 2x2 contingency tables associated with an initial p-value of 0.001 to 0.05 were evaluated. Each table's UFI and relative risk index (RRI) were calculated. A derivative of the RRI, the percent RRI, was also calculated along with the FQ. The UFI, FQ, RRI, pRRI, initial p-value, and sample size were compared. RESULTS: A total of 15000 cases were tested. The correlation between the UFI and the p-value was the strongest ($r = -0.807$), and the correlation between the pRRI was the weakest ($r = -0.395$). The RRI had the strongest correlation with the sample size ($r = 0.826$), and the UFI had the weakest correlation ($r = 0.3904$). The coefficient of variation for the average RRI was the smallest at 28.3%, and for the FQ, it was the greatest at 57.0%. The correlation between the UFI, FQ, and p-value is significantly greater than the correlation between the RRI, pRRI, and p-value (for all comparisons, $p < 0.001$). CONCLUSION: The RRI and pRRI are significantly less correlated with the p-value than the UFI and FQ, indicating relative independence of the RRI and pRRI from p-values.

A Comparative Analysis of Unit Fragility and the Relative Risk Index

Thomas F. Heston MD [1, 2]

[1] Department of Family Medicine, University of Washington School of Medicine

[2] Department of Medical Education and Clinical Sciences, Elson S. Floyd College of Medicine, Washington State University

Abstract

BACKGROUND: In biostatistics, evaluating fragility is crucial for understanding their vulnerability to miscategorization. One proposed measure of statistical fragility is the unit fragility index (UFI), which measures the susceptibility of the p-value to flip significance with minor changes in outcomes. Although the UFI provides valuable information, it relies on p-values, which are arbitrary measures of statistical significance. Alternative measures, such as the fragility quotient (FQ) and the percent fragility index, have been proposed to decrease the UFI's reliance on sample size. However, these approaches still rely on p-values and thus depend on an arbitrary cutoff of $p < 0.05$. Instead of quantifying fragility by relying on p-values, this study evaluated the effect of small changes on relative risk. METHODS:

Random 2x2 contingency tables associated with an initial p-value of 0.001 to 0.05 were evaluated. Each table's UFI and relative risk index (RRI) were calculated. A derivative of the RRI, the percent RRI, was also calculated along with the FQ. The UFI, FQ, RRI, pRRI, initial p-value, and sample size were compared. RESULTS: A total of 15000 cases were tested. The correlation between the UFI and the p-value was the strongest (r = -0.807), and the correlation between the pRRI was the weakest (r = -0.395). The RRI had the strongest correlation with the sample size (r = 0.826), and the UFI had the weakest correlation (r = 0.3904). The coefficient of variation for the average RRI was the smallest at 28.3%, and for the FQ, it was the greatest at 57.0%. The correlation between the UFI, FQ, and p-value is significantly greater than the correlation between the RRI, pRRI, and p-value (for all comparisons, $p < 0.001$). CONCLUSION: The RRI and pRRI are significantly less correlated with the p-value than the UFI and FQ, indicating relative independence of the RRI and pRRI from p-values.

KEYWORDS: Fragility index, Unit fragility index, Fragility quotient, percent fragility index, Relative risk index, percent relative risk index, statistical significance.

FUNDING INFORMATION: Self-funded, no external funding.

CONFLICT OF INTERESTS: No competing interests.

ETHICAL APPROVAL: This study did not involve human or animal research.

Introduction

The unit fragility index (UFI) was proposed in 1990 to quantify the fragility of a test of two proportions (1). The UFI looked at the effect of small outcome changes on the p-value. When evaluating a standard 2x2 contingency table, the UFI represented the number of unit changes (integer changes) in the cell counts. Using the standard labels for a 2x2 table (Table 1), the first step is to look at the p-value. If it is statistically significant ($p < 0.05$), one is added to the cell with the smallest count (2). For example, if "d" had the fewest cases, the table would be changed as follows: a+1, b-1, c-1, d+1. The resultant p-value would then be calculated. If the p-value flipped from significant to nonsignificant, then the UFI = 1. If the p-value remained significant, the process would increment by one unit again, such that the new table would be: a+2, b-2, c-2, and d+2. The p-value of this new table will be calculated. If the significance flipped, the process would end, and the UFI equal two. This is repeated until the p-value flips from < 0.05 to > 0.05, and the UFI is the number of increments required to do this.

Table 1. Standard nomenclature for a 2x2 contingency table

	Disease +	**Disease -**	
Exposure +	a	b	a + b
Exposure -	c	d	c + d
	a + c	b + d	a + b + c + d

The problem with the UFI is that it changes with increased sample size (3). An attempt to fix this issue was to calculate a Fragility Quotient (FQ), the UFI divided by the sample size (4). Another attempt to address this issue was incrementing by fractions instead of integer units, creating a percent fragility index (5). While these did help address the sample size issue, the reliance upon p-values remained.

P-values don't necessarily influence clinicians. What they want to know is significance. Specifically, should I use this medication or not? Should I do this test or not? The relative risk (RR) of a new test or medication, as quantified by a 2x2 contingency table, answers this question. Although still referred to as RR, the same formula is used to determine relative benefit (Table 2). If the RR is favorable, then the evidence favors doing the treatment or test. If unfavorable, then there is no benefit.

Table 2. A 2x2 table set up to evaluate the effectiveness of a new medication. The RR equals (a/(a+b)) / (c/(c+d)). For a 2x2 table set up in this manner, if the RR is greater than one, then the new medication is more likely than placebo to be of benefit.

	Benefit	**No Benefit**	
New Medication	a	b	a + b
Placebo	c	d	c + d
	a + c	b + d	a + b + c + d

The relative risk index (RRI) proposed here is a measure to quantify statistical fragility without relying upon p-values. While the UFI looks at the effect of incremental changes in

outcomes upon the p-value, the RRI looks at the effect of changing outcomes upon the RR. Instead of a threshold of flipping the p-value from significant to non-significant, the RRI quantifies how much of a change in outcomes is required to get the RR equal to one since when the RR equals one, it indicates that there is no value whatsoever for the new test or new treatment. The marginal totals remain fixed for the UFI and the RRI, and only the outcomes are slightly modified. Note that the UFI quantifies the effect of integer changes, whereas the RRI gives an exact number. If a "unit" RRI is desired, the RRI is rounded to the next integer.

The RRI is equal to (bc-ad)/(a+b+c+d). When using it for statistical purposes and for our calculations below, the absolute value is utilized. So when applied to the data, the RRI can be either positive or negative, but when used as a measure of fragility, the absolute value is used. The RRI results in the RR of a 2x2 contingency table being equal to one, at which point the new treatment or test has no benefit (or harm) (Table 3).

Table 3. This 2x2 table results in an RR equal to one.

	Benefit	**No Benefit**	
New Medication	a - RRI	b + RRI	a + b
Placebo	c + RRI	d - RRI	c + d
	a + c	b + d	a + b + c + d

The percent RRI (pRRI) is a derivative of the RRI that quantifies the percent change in the 2x2 table rather than the absolute change. Since the RRI is applied to each cell in the table, the pRRI equals (RRI/a + RRI/b + RRI/c + RRI/d)/4.

Methods

A Python program generated random 2x2 contingency tables with cell counts ranging from 15 to 250. Tables selected for analysis were limited to those associated with a p-value of less than 0.05 as determined by chi-square testing. The UFI, FQ, RRI, and pRFI were calculated for each table based on the initial p-value and outcomes changes to flip the p-value to > 0.05 or reach an RR of one. Fisher's z-transformation determined statistically significant differences in correlation coefficients. The Python program used is publicly available on GitHub (6).

Results

A total of 15,000 tables were created, with the lowest value in the 2x2 table being 15 and the highest value being 250. The lowest p-value was 0.001, and the highest p-value was 0.05 (Table 4)

Table 4. Variables input into the Python program

Variable	**Base Value**
Total Cases	15,000
Lowest Value	15
Highest Value	250
Lowest p-value	0.000999
Highest p-value	0.05

The average, standard deviation (sd), coefficient of variation (CV) for the variables of interest are shown in Table 5. Note that the CV for RRI and pRRI are significantly less than the CVs for the other variables.

Table 5. Average values of the variables across the entire 15000 cases evaluated.

Variable	**Average (sd)**	**CV**	**Minimum**	**Maximum**
P-Value	0.01460 (0.01362)	93.3%	0.001	0.050
Sample Size	307.0 (146.6)	47.8%	82	855
UFI	3.123 (1.729)	55.4%	1	9
FQ	0.01153 (0.00657)	57.0%	0.00116	0.04167
RRI	10.94 (3.10)	28.3%	4.91	22.99
pRRI	0.04010 (0.01205)	30.1%	0.01454	0.08789

The correlation coefficient with the p-value was strongest for the UFI and the FQ. The weakest correlation was with the pRRI, and the next weakest the RRI. Given a sample size of 15,000, it was found that these differences were all statistically significant, although, from a practical standpoint, the greatest differences found were between the UFI and FQ compared to the RRI and pRRI (Table 6). Note that the R-squared value for the UFI is 0.65, indicating that the p-value explains 65% of the variation in the UFI. The R-squared value for the pRRI is 0.16, indicating that the p-value explains only 16% of the variation in the pRRI.

Table 6. Correlation coefficients for measures of fragility and the p-value and sample size.

Variable	**P-Value**	**Sample Size**
UFI	-0.80706	0.39042
FQ	-0.7368	-0.43283
RRI	-0.49126	0.8261
pRRI	-0.39544	-0.77739

Discussion

Statistical fragility is poorly understood, with no consensus on cut-offs to determine whether a study is significantly fragile or robust. However, multiple studies have shown high fragility in the medical literature (3,7,8).

A weak correlation between RRI and p-values and between pRRI and p-values indicates relative independence of the RRI and pRRI; they do not appear to be colinear with p-values

like the UFI and FQ are. While the UFI and FQ strongly depend on the p-values, the RRI and pRRI do not. This test of 15000 cases shows that about 65% of the variation in the UFI is explained by the p-value, whereas only 16% of the variation in the pRRI is explained by the p-value. Thus, the RRI and pRRI provide new information beyond statistical significance. They add more information to the p-value than the UFI or FQ.

While the RRI and pRRI had the strongest correlation with sample size, the fact that the CV was significantly smaller than the CV for UFI and FQ indicates that, even though they are correlated with sample size, the RRI and pRRI only undergo minimal changes with varying sample sizes (i.e., the CVs for both were small). This tighter distribution of the RRI and pRRI means that values on the extremes are more meaningful.

The pRRI is particularly useful to clinicians responsible for the care of individual patients. A good, skeptical clinician will always question the fragility of a study. With the pRRI, clinicians can quantify just how large or small a change in outcomes is required to make the new treatment or test of zero benefit. Sometimes, even if a p-value is non-significant, it can still indicate that, more likely than not, the new treatment or test is beneficial in most cases (e.g., if the p-value = 0.06). This situation is nebulous. At what point is the test or treatment of no value? The pRRI answers this question. If there is a pRRI change in outcomes, the new test or treatment is useless.

Clinicians can use the pRRI in this manner: if a study shows a pRRI of 0.03, it means that if 3% of the outcomes were changed, there would be no benefit to the new test or treatment. Small pRRI values indicate fragility and large pRRI values indicate a robust study. As the

pRRI value increases, clinicians can become more confident that the new test or treatment will help their specific patient.

This study is limited by the fact that random 2x2 tables were generated. The data is not from research trials but a type of model. Testing with actual research findings is necessary to determine the practical utility of the RRI and the pRRI.

Conclusion

This study demonstrates that the relative risk index (RRI) and percent RRI (pRRI) provide measures of statistical fragility that are less dependent on p-values and sample size compared to existing methods like the unit fragility index (UFI) and fragility quotient (FQ). The RRI and pRRI correlate weakly with p-values, indicating they quantify different information beyond what is captured by p-values alone. The pRRI, in particular, allows a clinically valuable interpretation - the percent change in outcomes leading to no effect. The RRI and pRRI add nuance beyond significance testing for clinicians evaluating treatments and diagnostic tests. However, further validation using real-world data is needed. Overall, this initial investigation of the RRI and pRRI as alternatives to the UFI and FQ shows promise for quantifying fragility less reliant on arbitrary p-value cutoffs.

Bibliography

1. Feinstein AR. The unit fragility index: an additional appraisal of “statistical significance”

for a contrast of two proportions. J Clin Epidemiol. 1990;43(2):201–9.

2. Walter SD. Statistical significance and fragility criteria for assessing a difference of two proportions. J Clin Epidemiol. 1991;44(12):1373–8.

3. Herndon CL, McCormick KL, Gazgalis A, Bixby EC, Levitsky MM, Neuwirth AL. Fragility index as a measure of randomized clinical trial quality in adult reconstruction: A systematic review. Arthroplasty Today. 2021 Oct 11;11:239–51.

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

5. Heston TF. The Percent Fragility Index. International journal of scientific research. 2023 Jul 1;12(7):9–10.

6. Heston TF. A comparison of the relative risk index with unit fragility: software code [Internet]. GitHub. 2023 [cited 2023 Aug 14]. Available from: https://github.com/tomheston/A-Comparison-of-the-Relative-Risk-Index-with-Unit-Fragilty/blob/main/UFI_vs_RRI.ipynb

7. Doyle TR, Davey MS, Hurley ET. The Statistical Fragility of Management Options for Acute Achilles Tendon Ruptures - A Systematic Review of Randomized Control Trial with Fragility Analysis. J ISAKOS. 2022 Aug;7(4):72–81.

8. Vargas M, Marra A, Buonanno P, Coviello A, Iacovazzo C, Servillo G. Fragility Index and Fragility Quotient in Randomized Controlled Trials on Corticosteroids in ARDS Due to COVID-19 and Non-COVID-19 Etiology. J Clin Med. 2021 Nov 14;10(22).