Empirical assessment of fragility index based on a large database of clinical studies in the Cochrane Library

Running title: Empirical assessment of fragility index

Aiwen Xing, PhD[1,2]; Lifeng Lin, PhD[3,*]

[1]Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, NJ, USA

[2]Department of Statistics, Florida State University, Tallahassee, FL, USA

[2]Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, USA

[*]Correspondence: Lifeng Lin, 1295 N. Martin Ave., Drachman Hall A240, Tucson, AZ 85724, USA; Email: lifenglin@arizona.edu

Conflict of interest: None.

Funding/Support: This research was supported in part by the US National Institute of Mental Health grant R03 MH128727 and the US National Library of Medicine grant R01 LM012982. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Word count: 4000 (abstract: 299); No. of tables: 5; No. of figures: 4; No. of supplemental file: 1

Abstract

Rationale aims and objectives: The fragility index (FI) and fragility quotient (FQ) are increasingly used measures for assessing the robustness of clinical studies with binary outcomes in terms of statistical significance. The FI is the minimum number of event status modifications that can alter a study result's statistical significance (or nonsignificance), and the FQ is calculated as the FI divided by the study's total sample size. The literature has no widely-recognized criteria for interpreting the fragility measures' magnitudes. This article aims to provide an empirical assessment for the FI and FQ based on a large database of clinical studies in the Cochrane Library.

Methods: We explored the overall empirical distributions of the FI and FQ based on five common methods (Fisher's exact test, chi-square test, risk difference, odds ratio, and relative risk) for determining statistical significance of binary outcomes in clinical research. We also considered three different scenarios for the FI calculation and evaluated the relationship between P-values and FIs or FQs using Spearman's ρ. Finally, we summarized empirical thresholds based on the overall distributions of the FI and FQ to facilitate their interpretations in future research.

Results: For about 20% of studies with significant results, the statistical significance was changed after modifying the event status of only one participant. Studies with significant results were considered slightly fragile if the significance hinged on the statuses of about five events. Studies were extremely fragile if FI≤1 or FQ≤0.01. The FIs were strongly correlated with P-values for significant studies, while Spearman's ρ varied according to the total sample sizes of studies.

Conclusions: The statistical significance of clinical studies could be changed after modifying a few events' statuses. Many studies' findings are fairly fragile. The distributions of the FI and FQ provide insights for appraising the robustness of evidence in clinical decision-making.

Keywords: Cochrane Library; fragility index; healthcare; robustness; sensitivity analysis.

1. INTRODUCTION

P-values are reported in nearly all clinical studies to assess their results for evidence-based medicine and clinical decision-making. The threshold of 0.05 is traditionally used to indicate the statistical significance of the results.[1] However, the P-value has been frequently criticized because it may cause researchers to overlook some essential factors such as the sample size and the number of patients lost to follow-up; it may also be misinterpreted as an effect measure.[2-7] Moreover, studies with a P-value less than the conventional threshold of 0.05 may be more likely to be reported.[1, 8] Such misuses of the P-value may cause a lack of research reproducibility and replicability.[9-11]

To better facilitate the interpretation of clinical studies, Walsh et al.[12] proposed a simple metric, called the fragility index (FI), as a supplement to the P-value. It assesses the robustness of the results of clinical studies in terms of statistical significance. The FI is defined as the minimum number of event status modifications needed to alter the significance of the results. Studies with a smaller value of the FI are regarded as more fragile. This index has been used in various research fields.[13-19] In addition, as the FI is an absolute value to measure fragility while the study size could affect the assessment of fragility, a relative metric fragility quotient (FQ), calculated as the absolute FI number divided by the total sample size, was proposed to account for the sample size.[20] These two metrics can be used to evaluate how easily the statistical significance may be changed based on the prespecified P-value threshold. Therefore, FI and FQ may be used together with P-value for assessing research reproducibility and replicability.

Nevertheless, many unknowns about the FI remain to be investigated. For example, the FI has been criticized for providing little information in addition to the P-value, depending on the threshold of statistical significance, and correlating with the sample size.[21, 22] Also, it may inappropriately penalize smaller studies for using fewer events than larger studies to achieve the same significance level.[23] Although the FQ addresses some of these concerns, the relative measures of fragility have not proven clinically meaningful.[24] Moreover, there are no widely-recognized criteria for interpreting the magnitudes of the fragility measures.[21, 25, 26] Consider two studies with equal FQs of 1%; the first one has a total sample size of 100 with FI=1, while the second one has a total sample size of 1000 with FI=10. It is unclear whether the FQ of 1% is small enough to declare that both studies are equally fragile, or the study with FI=1 is considerably more fragile than another with FI=10. Essentially, the literature has no consensus on defining a study to be "fragile." In addition, the evaluation of the FI has not been systematically assessed using real-world data. The lack of these criteria and reliance on an absolute value alone is of great concern when promoting the FI to clinicians. Last, Walsh et al.[12] originally considered studies with a 1:1 allocation ratio and restricted event modifications within a single arm. However, some applications of FI are not exactly for studies with balanced groups; the restrictions during the FI calculation may not guarantee to yield minimal modifications of event status. A more efficient and well-designed method has been proposed to evaluate the fragility.[27, 28]

For deriving a practical and evidence-based guideline for clinical researchers, we explore the empirical distributions of the FI and FQ using a large database of clinical studies with binary outcomes from the Cochrane Library. We consider five frequently used statistical methods for determining statistical significance, i.e., Fisher's exact test, chi-square test, risk difference (RD), odds ratio (OR), and relative risk (RR). The aims of this article are four-folded. First, we investigate the performance of the FI and FQ for different statistical significance levels. Second, based on empirical distributions of the FI and FQ, we propose rules of thumb for interpreting the magnitudes of the fragility measures. Third, we explore different scenarios of the FI calculation to validate FI algorithms' efficiency and accuracy. Fourth, we evaluate correlations between the fragility measures and P-values.

2. METHODS

2.1. FI and FQ

Consider an illustrative study with a binary outcome, which compares two treatments, denoted by 0 (control) and 1. Its results are represented in a 2×2 table (Table 1). The event statuses of both groups can be potentially modified to alter the significance for deriving the FI. The FI is derived by iteratively adding or subtracting events in either group and recalculating the P-value until the significance or nonsignificance of the result is altered. The total number of modified event statuses in group i can be denoted by $|f_i|$ ($i = 0, 1$). Increasing ($f_i > 0$) or decreasing ($f_i < 0$) event counts within groups may be determined based on experts' opinions. The minimum number of modifications on event status needed to alter the significance of results is defined as the FI, i.e., $\mathrm{FI} = \min(|f_0| + |f_1|)$.

Because the FI is an absolute measure,[29, 30] a "relative" measure fragility quotient (FQ) was proposed to account for the study's total sample size.[20] The FQ is calculated as the absolute FI value divided by the total sample size, i.e., $\mathrm{FQ} = \mathrm{FI}/(n_0 + n_1)$. As such, the FQ is useful for comparing fragility across studies, especially when their sample sizes differ greatly.[27] Like the FI, a greater value of the FQ indicates the study result is more robust.

2.2. Directions of altering significance

The FI originally proposed by Walsh et al.[12] was defined for a study with a statistically significant result. It is the minimum number of event status changes where the statistical significance hinges on, i.e., the minimum number of participants to have a non-event changed to an event that would result in a nonsignificant result.[21] Similarly, the concept of fragility can be extended to a study with a nonsignificant result, where the FI becomes the minimum number of events needed to change clinical findings from nonsignificant to significant.[31, 32] Of note, it is possible that the nonsignificant study can never be altered to be significant, even if all events or non-events (i.e., $f_i = e_i$ or $n_i - e_i$) have been modified to non-events or events.

2.3. Statistical methods for assessing treatment effects

The association between treatments and outcomes could be different due to different adopted statistical methods.[33] The various statistical methods used to produce a P-value can change the FI value and thus affect the interpretation of study results' fragility. Researchers may misunderstand the statistical strength of evidence for a study's result if inappropriate statistical methods are adopted.

Walsh et al.[12] calculated the FI based on a two-sided Fisher's exact test. The P-value calculated by Fisher's exact test is based on the hypergeometric distribution under the null hypothesis of no association. Although it is frequently used for small sample sizes, it can be used for all sample sizes.[34] Nevertheless, Fisher's exact test may cost more time for large sample sizes than alternative methods, such as the chi-square test, particularly if an iterative algorithm is used for deriving the FI. The P-value of the chi-square test is based on the asymptotic chi-square distribution. Traditionally, the chi-square test is recommended for contingency tables with the minimum expected count >5, while this method may be invalid for studies with small counts of events or non-events.[35]

In addition, the RD, OR, and RR are frequently used for quantifying treatment effects of binary outcomes. The P-value may be produced based on these effect measures directly. Of note, the OR and RR are analyzed based on the logarithmic scale; a continuity correction of 0.5 is usually applied to all data cells in the contingency table in the presence of zero counts.

2.4. Treatment group with event status modifications

Walsh et al.[12] originally proposed the FI by restricting modifications among the treatment group with the fewest events (say group 0). After iteratively switching event statuses in group 0 and recalculating the two-sided P-value for Fisher's exact test, the number of events in group 0 is modified until the change of significance. In the balanced setting, as modifications within the group with the fewest events may have a greater impact, this procedure would be reasonable and efficient for deriving the FI. However, the sample sizes in the two groups may not be exactly equal in practice. Restricting to a single group may not guarantee achieving the minimum number of event status modifications, and thus the study's fragility would be underestimated.

We consider three different ways of event status modifications and assess the differences in their resulting FI values (Table 1). They are detailed as follows.

- Scenario 1. We modify event status in both directions (adding or subtracting events) among both groups in each iteration until the change of significance. The total number of modified event statuses is the FI.
- Scenario 2. We first restrict modifications in both directions in a single group (say group 0) during the whole iterations until the change of significance. The total number of modified event statuses is

denoted as $|f_0|$. Then, we repeat this process separately for group 1 and obtain $|f_1|$. The FI is the smaller value among $|f_0|$ and $|f_1|$.

- Scenario 3 (Walsh et al.[12]). We modify event status in both directions only among the treatment group with the fewest events.

In scenario 1, the significance can be changed after modifying $|f_0|$ and $|f_1|$ events in groups 0 and 1, respectively, for the illustrative study (Table 1), then the FI is $|f_0| + |f_1|$. In each iteration, we recalculate the P-value after modifying the event status in the two groups separately. In scenario 2, we compute the FI by restricting modifications in a single group until the change of significance; then, we choose the smaller FI. In each iteration, we recalculate the P-value after restricting modifications among one group, say group 0. We conduct $|f_0|$ times of computation among group 0 until altering significance. Similarly, we repeat $|f_1|$ times computation in group 1. The FI is $\min(|f_0|, |f_1|)$. As such, this scenario is less computationally expensive than scenario 1. Scenario 3 corresponds to the original FI calculation by Walsh et al.[12] It assumes group 0 has the fewest events, and the significance would be changed until modifying the statuses of $|f_0|$ events. The FI is $|f_0|$.

2.5. Different P-value thresholds

The P-value threshold α, i.e., the significance level, determines the significance of clinical results. Therefore, it has a great impact on the FI. Conventionally, α is set to 0.05. Recently, due to the critiques of the misuse and misinterpretation of the P-value, some investigators proposed to lower the significance level.[4, 22, 36] Thus, this article additionally considers calculating the FI at α=0.01 and 0.001. The different choices of α permit us to investigate how the FI values vary with them and provide researchers with additional guidelines about the fragility measures when alternative significance levels are adopted.

2.6. Datasets and analyses

We collected the individual clinical studies from a large collection of meta-analyses with binary outcomes from the Cochrane Library, which provides valuable information on various healthcare-related topics to inform decision-making. We used the R software to iteratively collect all Cochrane reviews from 2003

Issue 1 to 2020 Issue 1. We only kept the latest version of a review if it had multiple updates during this period, and we excluded data from reviews that had been withdrawn.

In our primary analysis, we considered scenario 1 to derive the FI values using the five statistical methods with the three P-value thresholds regardless of the original methods and P-value thresholds used by the clinical studies. We calculated the FI and FQ values of each clinical study and derived their empirical distributions. Scenarios 2 and 3 were considered for the OR with a P-value threshold of 0.05. The analyses were implemented using the package "fragility" (version 1.0) in R (version 3.5.3).[37]

2.7. Empirical distributions

We summarized the overall empirical distributions of the FI and FQ values of the clinical studies from the Cochrane Library. They were categorized by the five statistical methods used for assessing fragility, i.e., Fisher's exact test, chi-square test, RD, OR, and RR. Of note, for Fisher's exact test, only clinical studies with a sample size of each group <1000 were considered due to a long computing time.

To quantitatively evaluate the fragility of one individual study compared to others, we proposed rules of thumb for the magnitudes of fragility based on the empirical distribution of FI and FQ through real-world data. Specifically, we determined the cutoffs based on the cumulative proportions (from 0 to a specific cutoff) for significant and nonsignificant studies separately, i.e., the percent of studies with FI/FQ ≤ the cutoff. In general, we roughly classified the magnitudes of fragility into five levels: extremely fragile, moderately fragile, slightly fragile, slightly robust, moderately robust, and extremely robust. As strict cutoffs may override important clinical information and may be misleading, we permitted overlapping ranges for the levels of fragility magnitudes, sharing a similar idea with the guidelines for interpreting the I^2 statistic for between-study heterogeneity.[38]

2.8. Correlation between fragility measures and P-values

The P-value used in insolation has been criticized for being simplistic[2]; as such, the FI and FQ were designed as supplements to aid interpretations of clinical findings. On the other hand, the FI has been criticized for being highly associated with the P-value and sharing similar information.[21-23, 39] As the sample sizes vary greatly in our database, we considered the total sample size as a potential confounder and investigated the relationship between the P-value and each of the FI and FQ across different sample sizes. The relationship was evaluated using Spearman's rank correlation coefficient ρ. The sample sizes were grouped into nine levels: <50, 50–100, 100–200, 200–300, 300–400, 400–500, 500–800, 800–1500, and >1500.

3. RESULTS

3.1. Basic characteristics

In total, we included 316,249 individual clinical studies with binary outcomes. Table 2 summarizes the characteristics of included studies. The median sample size was 123 patients (range: 2–2,212,432), with a median of 17 events (range: 0–212,070). Based on Fisher's exact test, about 20% of studies reported results with P<0.05.

3.2. FI and FQ based on different P-value thresholds

3.2.1. FI at the P-value threshold of 0.05

We derived empirical FI distributions based on each of the five methods at each of the three P-value thresholds. Figure 1 focuses on the distribution of the FI based on Fisher's exact test at the P-value threshold of 0.05 for significant studies; Figures S1–S4 show the distributions based on the other four methods. Figure 2 summarizes the distributions based on all five methods for both significant and nonsignificant studies. Based on the RD, 20,303 (26%) significant studies had FI=1, and the number was the most across all five methods. Similarly, the number of nonsignificant studies with FI=1 based on the RD was the most among the five methods, followed by the chi-square test. Table 3 shows the median and interquartile range (IQR). For significant studies based on Fisher's exact test, the median FI was 4 with an IQR of 2–9. Specifically, 13,729 (23%) had FI=1, and 28,737 (48%) had FI≤3.

Due to the smaller median FI values of significant studies based on all five methods at different thresholds (Table 3), studies with nonsignificant results generally seemed less fragile. Compared with nonsignificant studies, the wider IQRs of significant ones indicated distributions of the FI for significant studies were more dispersive.

3.2.2. FQ at the P-value threshold of 0.05

The median FQs of significant studies based on the five methods at the P-value threshold of 0.05 ranged from 0.0252 to 0.0270 with an IQR of about 0.01–0.06, indicating the significance of the results hinged on only fewer than 3 events per 100 participants. Similar to the conclusions based on the FI, nonsignificant studies seemed less fragile, with the FQ ranging from 0.0290 to 0.0426, indicating that the statistical significance of nonsignificant studies was contingent on more events per 100 participants than significant studies. Based on each method, about 70% of significant studies had FQ≤0.05, and about 89% had FQ≤0.1.

3.2.3. Alternative P-value thresholds

Figures S5–S31 give the empirical distributions for the P-value thresholds of 0.01 and 0.001. When the P-value threshold was lower, the median of FI increased, indicating studies were more robust and more events were needed to alter the significance. The median FI of nonsignificant studies based on the OR at the P-value threshold of 0.01 was 8 with an IQR of 6–11 and at the P-value threshold of 0.001 was 13 with an IQR of 10–17. At the P-value threshold of 0.001, 13% of nonsignificant studies based on the RD had FI≤3, while the proportions based on the OR and RR were both around 4%.

At the P-value threshold of 0.01, the median FQ based on Fisher's exact test was 0.0278 (IQR, 0.0121–0.0600) for significant studies and 0.0500 (IQR, 0.0233–0.0986) for nonsignificant ones. As such, the significance and nonsignificance of study results were contingent on approximately 3 and 5 events per 100 participants. Significant studies were more fragile according to the smaller median FQ. Among significant studies, 41% had FQ≤0.02 based on the chi-square test at the P-value threshold of 0.001.

3.3. Rules of thumb for interpreting magnitudes of fragility

Tables 4 and 5 present the rules of thumb for interpreting magnitudes of fragility for significant and nonsignificant studies based on the five statistical measures at the P-value threshold of 0.05. In general, we roughly classified the FI and FQ into six different levels, from extremely fragile to extremely robust, based on the cumulative proportions (from 0 to a specific cutoff) of the empirical distributions of the FI or FQ, i.e., the percent of studies with FI/FQ ≤ the cutoff. Significant studies might be extremely fragile if the FI=1. About 17% to 26% of significant clinical studies had FI values within this range. If the FI of a significant study was within 1–3 or 1–2, this study could be regarded as moderately fragile as it is only more robust than 18% (FI=1 based on the RR) to 48% (FI=3 based on Fisher's exact test) of studies. A significant study might be extremely robust if its FI≥27 (based on Fisher's exact test, the chi-square test, OR, and RR); it was more robust than about 90% of studies.

Similarly, nonsignificant studies might be extremely fragile if FI=1, and they were more robust than only a few studies (6%–12%). If the derived FIs were within 5–7 (by Fisher's exact test or chi-square test), 4–6 (by RD), 6–7 (by OR or RR), accordingly, they might be considered as slightly robust as they were more robust than at least 66% nonsignificant studies (based on the RR).

Additionally, studies with FQ≤0.01 can be considered extremely fragile (Table 5). The significance of about 20% of significant studies hinged on only 1 event per 100 participants (FQ≤0.01). If the FQ value was within 0.01–0.02, a significant study might be moderately fragile, as they were more robust than about 22%–43% of significant studies. The nonsignificant studies might be slightly fragile if the FQ values were within 0.02–0.04, 0.02–0.04, 0.02–0.03, 0.02–0.05, and 0.02–0.05, accordingly.

3.4. Fragility impacted by scenarios of event status modifications

Recall that the results above in our main analyses were based on scenario 1 of event status modifications. We also applied scenarios 2 and 3 based on the OR to evaluate their differences from scenario 1. Of note, we excluded 77,132 studies (24%) with equal numbers of modifications for the two groups (i.e., $|f_0| = |f_1|$) in scenario 2 from the original dataset for more accurate comparisons. Compared with scenarios 2 and 3, about 8% of studies had smaller FIs based on scenario 1, indicating that these studies had fewer events needed to alter the significance if there was no restriction in the modified groups. Moreover, 36.03% of studies had smaller FIs with modifications in both groups (scenario 1) compared with scenario 3. Restricting modifications within one group (scenarios 2 and 3) might underestimate the fragility of studies.

Additionally, we tried to examine whether it is reasonable to restrict event status modification within the group with fewer events or with a smaller proportion of events (i.e., e_i/n_i for group i). In scenario 2, FIs of 47% of studies (112,999 out of 239,117) were derived from groups with a smaller proportion of events. FIs of 53% of studies (126,319 out of 239,117) were derived from groups with fewer events. Therefore, the number of events or the proportion of events might not be the deterministic factors for the FI calculation.

3.5. Correlation between fragility measures and P-values

Figure 3 shows the relationship between the P-values and fragility measures at the P-value threshold of 0.05 based on Fisher's exact test. The median FIs increased as P-values decreased for significant studies, while it showed a reverse trend for nonsignificant studies. Nevertheless, having a P-value<0.005 or ≥0.5 did not exclude the possibility of having a very small FI (e.g., 1 or 2). The median FQs showed a similar trend.

For significant studies, there was a strong inverse relationship between P-values and fragility measures across most sample sizes (Figure 4). The correlations between P-values and FQs were relatively moderate. When the total sample size was small (<50), the correlations were relatively weak with smaller ρ in absolute magnitude. For nonsignificant studies, there were moderately positive correlations between P-values and FIs (Figure 2); the coefficients ρ were about 0.6 for all five methods. The scatter plots of P-values vs. FI can be found in Figures S32–S36; Figures S37–S64 present plots for other methods at alternative P-value thresholds.

4. DISCUSSION

4.1. Strengths

Researchers recently proposed various concepts of fragility, such as the reverse FI for nonsignificant studies[16] and the continuous FI for studies with continuous outcomes.[40] The FI was also extended to evaluate meta-analyses and network meta-analyses.[31, 32] However, their calculation methods were not exactly the same, while they share an essential idea of applying an iterative algorithm to derive the minimal data modifications needed to alter the significance. To find an efficient and accurate solution, we examined the impact of different scenarios of event status modifications on the resulting FI. When restricting modifications in a single group (scenario 2), our results showed that about half of the studies had fewer steps of event status modifications from groups with a smaller proportion of events (47%) or from groups with fewer events (53%) compared with scenario 1. This result indicated it might not be reasonable to focus on only a specific group during iterations. Nevertheless, computation time might be an important factor. In scenario 1, the computation time was doubled during each iteration because of modifications in both groups. To achieve the change of significance, we need to perform $2^{|f_0|+|f_1|}$ recalculations of P-values in total. The computation time is longer than scenario 2, but more accurate FIs may be found. In scenario 2, recall that the FI is $\min(|f_0|, |f_1|)$, so the total number of iterations is $|f_0| + |f_1|$, which is generally smaller than that of scenario 1. In general, users of the FI may consider balancing the computational efficiency and the result's accuracy.

4.2. Limitations

Our study has some limitations. First, the underpowered studies are more likely to yield low FI values.[41] Therefore, different types of study designs may provide various evidence for FI evaluation. The database used in this article included all types of clinical studies, such as case–control studies and cohort studies. However, the FI was originally proposed as a metric of fragility for randomized controlled studies, which usually have relatively small sample sizes. Randomized studies have been frequently questioned regarding their results' significance, which may be altered with small changes partially because of the small sample size and loss to follow-up.[30, 42] Considering the influence of sample sizes on FI, the empirical distributions based on the mixed database might reduce the interpretability of the FI.

Second, the method of the FI calculation used in this article was only applied to clinical studies with binary outcomes. An alternative FI-like method has been proposed for continuous outcomes[40]; it expands the applications of the fragility concept.

Finally, although we proposed the rules of thumb for evaluating the fragility, some important clinical factors should be considered throughout the decision-making process, such as the likelihood of event status modifications.[43] Additionally, our criteria of classifications were based on the cumulative proportions of empirical distributions of the FI and FQ, while there is no consensus on the cutoffs so far. Therefore, researchers should use our findings as supportive instead of deterministic evidence.

5. CONCLUSIONS

We have presented the empirical distributions of fragility measures based on a large Cochrane database of 316,249 studies. In general, for about 20% of studies with significant results, the statistical significance could be altered after modifying the event status of only one patient. Studies with significant results were considered slightly fragile if the significance hinged on about five events. Studies might be extremely fragile if their FI≤1 or FQ≤0.01. When using the RD, the robustness of studies might be underestimated, as the corresponding cutoffs from the rules of thumb were relatively smaller compared with other statistical methods for testing the association between treatments and outcomes. The FI was strongly correlated with the P-value for significant studies, while Spearman's ρ varied according to studies' total sample sizes. When the sample size was <50, the correlation was not very strong, and thus the FI may deliver valuable information in addition to the P-value.

References

1. Chavalarias D, Wallach JD, Li AHT, et al. Evolution of reporting *P* values in the biomedical literature, 1990-2015. *JAMA* 2016;315(11):1141-48. doi: 10.1001/jama.2016.1952
2. Sterne JAC, Davey Smith G. Sifting the evidence—what's wrong with significance tests? *BMJ* 2001;322(7280):226-31. doi: 10.1136/bmj.322.7280.226
3. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: context, process, and purpose. *The American Statistician* 2016;70(2):129-33. doi: 10.1080/00031305.2016.1154108
4. Ioannidis JPA. The proposal to lower *P* value thresholds to .005. *JAMA* 2018;319(14):1429-30. doi: 10.1001/jama.2018.1536
5. Goodman SN. Toward evidence-based medical statistics. 1: the *P* value fallacy. *Annals of Internal Medicine* 1999;130(12):995-1004. doi: 10.7326/0003-4819-130-12-199906150-00008
6. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nature Human Behaviour* 2018;2(1):6-10. doi: 10.1038/s41562-017-0189-z
7. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567(7748):305-07. doi: 10.1080/00031305.2018.1543137
8. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine* 2008;358(3):252-60. doi: 10.1056/NEJMsa065779
9. Nuzzo R. Scientific method: statistical errors. *Nature* 2014;506(7487):150-52. doi: 10.1038/506150a
10. Halsey LG, Curran-Everett D, Vowler SL, et al. The fickle *P* value generates irreproducible results. *Nature Methods* 2015;12(3):179-85. doi: 10.1038/nmeth.3288
11. Kvarven A, Strømland E, Johannesson M. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour* 2020;4(4):423-34. doi: 10.1038/s41562-019-0787-z
12. Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. *Journal of Clinical Epidemiology* 2014;67(6):622-28. doi: 10.1016/j.jclinepi.2013.10.019
13. Noel CW, McMullen C, Yao C, et al. The fragility of statistically significant findings from randomized trials in head and neck surgery. *The Laryngoscope* 2018;128(9):2094-100. doi: 10.1002/lary.27183
14. Topcuoglu MA, Arsava EM. The fragility index in randomized controlled trials for patent foramen ovale closure in cryptogenic stroke. *Journal of Stroke and Cerebrovascular Diseases* 2019;28(6):1636-39. doi: 10.1016/j.jstrokecerebrovasdis.2019.02.029
15. Docherty KF, Campbell RT, Jhund PS, et al. How robust are clinical trials in heart failure? *European Heart Journal* 2016;38(5):338-45. doi: 10.1093/eurheartj/ehw427
16. Khan MS, Fonarow GC, Friede T, et al. Application of the reverse fragility index to statistically nonsignificant randomized clinical trial results. *JAMA Network Open* 2020;3(8):e2012469. doi: 10.1001/jamanetworkopen.2020.12469
17. Del Paggio JC, Tannock IF. The fragility of phase 3 trials supporting FDA-approved anticancer medicines: a retrospective analysis. *The Lancet Oncology* 2019;20(8):1065-69. doi: 10.1016/S1470-2045(19)30338-9
18. Tignanelli CJ, Napolitano LM. The fragility index in randomized clinical trials as a means of optimizing patient care. *JAMA Surgery* 2019;154(1):74-79. doi: 10.1001/jamasurg.2018.4318
19. Lin L, Xing A, Chu H, et al. Assessing the robustness of results from clinical trials and meta-analyses with the fragility index. *American Journal of Obstetrics & Gynecology* 2022:In press. doi: 10.1016/j.ajog.2022.08.053
20. Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility index? *Critical Care Medicine* 2016;44(11):e1142-e43. doi: 10.1097/ccm.0000000000001976

21. Niforatos JD, Zheutlin AR, Chaitoff A, et al. The fragility index of practice changing clinical trials is low and highly correlated with *P*-values. *Journal of Clinical Epidemiology* 2020;119:140-42. doi: 10.1016/j.jclinepi.2019.09.029
22. Carter RE, McKie PM, Storlie CB. The fragility index: a *P*-value in sheep's clothing? *European Heart Journal* 2017;38(5):346-48. doi: 10.1093/eurheartj/ehw495
23. Potter GE. Dismantling the Fragility Index: a demonstration of statistical reasoning. *Statistics in Medicine* 2020;39(26):3720-31. doi: 10.1002/sim.8689
24. Niforatos JD, Zheutlin AR, Pescatore RM. Fragility measures: more limitations considered. *Annals of Emergency Medicine* 2019;73(6):696-97. doi: 10.1016/j.annemergmed.2019.01.035
25. Ho AK. The fragility index for assessing the robustness of the statistically significant results of experimental clinical studies. *Journal of General Internal Medicine* 2022;37(1):206-11. doi: 10.1007/s11606-021-06999-9
26. Dettori JR, Norvell DC. How fragile are the results of a trial? The fragility index. *Global Spine Journal* 2020;10(7):940-42. doi: 10.1177/2192568220941684
27. Lin L. Factors that impact fragility index and their visualizations. *Journal of Evaluation in Clinical Practice* 2021;27(2):356-64. doi: 10.1111/jep.13428
28. Abaid LN, Grimes DA, Schulz KF. Reducing publication bias through trial registration. *Obstetrics & Gynecology* 2007;109(6):1434-37. doi: 10.1097/01.AOG.0000266557.11064.2a
29. Choupoo NS, Das SK, Saikia P, et al. How robust are the evidences that formulate surviving sepsis guidelines? An analysis of fragility and reverse fragility of randomized controlled trials that were referred in these guidelines. *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine* 2021;25(7):773-79. doi: 10.5005/jp-journals-10071-23895
30. Acuna SA, Sue-Chue-Lam C, Dossa F. The fragility index—*P* values reimagined, flaws and all. *JAMA Surgery* 2019;154(7):674-74. doi: 10.1001/jamasurg.2019.0567
31. Atal I, Porcher R, Boutron I, et al. The statistical significance of meta-analyses is frequently fragile: definition of a fragility index for meta-analyses. *Journal of Clinical Epidemiology* 2019;111:32-40. doi: 10.1016/j.jclinepi.2019.03.012
32. Xing A, Chu H, Lin L. Fragility index of network meta-analysis with application to smoking cessation data. *Journal of Clinical Epidemiology* 2020;127:29-39. doi: 10.1016/j.jclinepi.2020.07.003
33. Agresti A. Categorical Data Analysis. Third ed. Hoboken, NJ: John Wiley & Sons 2013.
34. Kim H-Y. Statistical notes for clinical researchers: chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics* 2017;42(2):152-55. doi: 10.5395/rde.2017.42.2.152
35. Campbell I. Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine* 2007;26(19):3661-75. doi: 10.1002/sim.2832
36. Wayant C, Scott J, Vassar M. Evaluation of lowering the *P* value threshold for statistical significance from .05 to .005 in previously published randomized clinical trials in major medical journals. *JAMA* 2018;320(17):1813-15. doi: 10.1001/jama.2018.12288
37. Lin L, Chu H. Assessing and visualizing fragility of clinical results with binary outcomes in R using the fragility package. *PLOS ONE* 2022;17(6):e0268754. doi: 10.1371/journal.pone.0268754
38. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: John Wiley & Sons 2019.
39. Schröder A, Muensterer OJ, Oetzmann von Sochaczewski C. Meta-analyses in paediatric surgery are often fragile: implications and consequences. *Pediatric Surgery International* 2021;37(3):363-67. doi: 10.1007/s00383-020-04827-5
40. Caldwell J-ME, Youssefzadeh K, Limpisvasti O. A method for calculating the fragility index of continuous outcomes. *Journal of Clinical Epidemiology* 2021;136:20-25. doi: 10.1016/j.jclinepi.2021.02.023
41. Condon TM, Sexton RW, Wells AJ, et al. The weakness of fragility index exposed in an analysis of the traumatic brain injury management guidelines: a meta-epidemiological and simulation study. *PLOS ONE* 2020;15(8):e0237879. doi: 10.1371/journal.pone.0237879

42. Baer BR, Fremes SE, Gaudino M, et al. On clinical trial fragility due to patients lost to follow up. *BMC Medical Research Methodology* 2021;21(1):254. doi: 10.1186/s12874-021-01446-z
43. Baer BR, Gaudino M, Charlson M, et al. Fragility indices for only sufficiently likely modifications. *Proceedings of the National Academy of Sciences* 2021;118(49):e2105254118. doi: 10.1073/pnas.2105254118

Table 1. Illustration of a study with a binary outcome and three scenarios of deriving the fragility index.

Treatment group	Event	Non-event	Sample size
Original dataset			
Group 0	e_0	$n_0 - e_0$	n_0
Group 1	e_1	$n_1 - e_1$	n_1
Scenario 1: FI $= \min(\lvert f_0\rvert + \lvert f_1\rvert)$			
Dataset after altering significance (modifications in both groups)			
Group 0	$e_0 + f_0$	$n_0 - e_0 - f_0$	n_0
Group 1	$e_1 + f_1$	$n_1 - e_1 - f_1$	n_1
Scenario 2: FI $= \min(\lvert f_0\rvert, \lvert f_1\rvert)$			
Dataset after altering significance (modifications only group 0)			
Group 0	$e_0 + f_0$	$n_0 - e_0 - f_0$	n_0
Group 1	e_1	$n_1 - e_1$	n_1
Dataset after altering significance (modifications only group 1)			
Group 0	e_0	$n_0 - e_0$	n_0
Group 1	$e_1 + f_1$	$n_1 - e_1 - f_1$	n_1
Scenario 3 (Walsh et al.[12]): FI $= \min(\lvert f_0\rvert)$.			
Dataset after altering significance (modifications in the group with fewer events, say, group 0)			
Group 0	$e_0 + f_0$	$n_0 - e_0 - f_0$	n_0
Group 1	e_1	$n_1 - e_1$	n_1

Table 2. Characteristics of included clinical studies.

Characteristic	**Count or median**
Sample size, median (min to max)	123 (2 to 2,212,432)
Event count, median (min to max)	17 (0 to 212,070)
Fisher's exact test	
P<0.05	60,456 (20%)
P<0.01	36,482 (12%)
P<0.001	21,581 (7%)
Chi-square test	
P<0.05	58,822 (19%)
P<0.01	35,941 (11%)
P<0.001	21,616 (7%)
Risk difference	
P<0.05	78,382 (25%)
P<0.01	47,945 (15%)
P<0.001	29,087 (9%)
Odds ratio	
P<0.05	64,518 (20%)
P<0.01	35,808 (11%)
P<0.001	19,626 (6%)
Relative risk	
P<0.05	61,604 (19%)
P<0.01	33,125 (10%)
P<0.001	17,825 (6%)

Note: For Fisher's exact test, only studies with a sample size less than 1000 in both study groups were included due to considerations about computing time.

Table 3. Medians and interquartile ranges (IQRs, in parentheses) of the fragility index and fragility quotient based on the five methods at the P-value thresholds of 0.05, 0.01, and 0.001.

		Fragility index		Fragility quotient	
Threshold	**Method**	**Significant**	**Nonsignificant**	**Significant**	**Nonsignificant**
	Fisher's exact test	4 (2, 9)	4 (3, 6)	0.0267 (0.0116, 0.0585)	0.0373 (0.0169, 0.0714)
	Chi-square test	4 (2, 11)	5 (3, 6)	0.0252 (0.0103, 0.0571)	0.0387 (0.0167, 0.0779)
0.05	RD	3 (1, 9)	3 (2, 5)	0.0254 (0.0105, 0.0563)	0.0290 (0.0129, 0.0556)
	OR	4 (2, 11)	5 (4, 7)	0.0270 (0.0112, 0.0600)	0.0406 (0.0176, 0.0811)
	RR	5 (2, 11)	6 (4, 7)	0.0263 (0.0109, 0.0588)	0.0426 (0.0182, 0.0875)
	Fisher's exact test	4 (2, 11)	6 (4, 7)	0.0278 (0.0121, 0.0600)	0.0500 (0.0233, 0.0986)
	Chi-square test	5 (2, 14)	7 (5, 8)	0.0262 (0.0108, 0.0592)	0.0526 (0.0234, 0.1042)
0.01	RD	4 (2, 11)	5 (3, 7)	0.0267 (0.0111, 0.0595)	0.0408 (0.0189, 0.0769)
	OR	6 (2, 15)	8 (6, 11)	0.0281 (0.0111, 0.0615)	0.0638 (0.0276, 0.1286)
	RR	6 (2, 16)	9 (7, 12)	0.0259 (0.0104, 0.0583)	0.0682 (0.0286, 0.1446)
	Fisher's exact test	6 (2, 14)	8 (6, 10)	0.0286 (0.0120, 0.0608)	0.0686 (0.0328, 0.1310)
	Chi-square test	7 (3, 18)	10 (7, 12)	0.0270 (0.0107, 0.0600)	0.0739 (0.0335, 0.1452)
0.001	RD	5 (2, 13)	7 (5, 10)	0.0288 (0.0114, 0.0625)	0.0584 (0.0274, 0.1059)
	OR	7 (3, 20)	13 (10, 17)	0.0256 (0.0103, 0.0578)	0.0988 (0.0430, 0.1923)
	RR	7 (3, 20)	16 (11, 20)	0.0238 (0.0095, 0.0500)	0.1071 (0.0450, 0.2258)

Table 4. Proposed rules of thumb for assessing the magnitudes of the fragility of individual clinical studies based on the fragility index.

	Fisher's exact test	**Chi-square test**	**RD**	**OR**	**RR**
Significant studies					
Extremely fragile	≤1 (23%)	≤1 (20%)	≤1 (26%)	≤1 (19%)	≤1 (18%)
Moderately fragile	1–3 (23%–48%)	1–3 (20%–43%)	1–2 (26%–41%)	1–3 (19%–43%)	1–3 (18%–42%)
Slightly fragile	2–4 (38%–55%)	2–4 (34%–51%)	2–3 (41%–51%)	2–5 (32%–56%)	2–5 (31%–55%)
Slightly robust	5–14 (61%–84%)	5–14 (57%–80%)	4–11 (58%–80%)	6–14 (61%–81%)	6–14 (60%–80%)
Moderately robust	10–26 (78%–93%)	10–26 (73%–90%)	8–22 (74%–90%)	10–26 (74%–90%)	10–26 (73%–90%)
Extremely robust	≥27 (93%)	≥27 (90%)	≥23 (91%)	≥27 (91%)	≥27 (90%)
Nonsignificant studies					
Extremely fragile	≤1 (8%)	≤1 (6%)	≤1 (12%)	≤1 (6%)	≤1 (6%)
Moderately fragile	2–3 (19%–35%)	2–4 (15%–44%)	1–2 (12%–30%)	2–4 (13%–38%)	2–4 (12%–34%)
Slightly fragile	3–4 (35%–56%)	3–5 (26%–66%)	2–3 (30%–55%)	3–5 (24%–55%)	3–5 (22%–48%)
Slightly robust	5–7 (75%–87%)	5–7 (66%–84%)	4–6 (71%–83%)	6–7 (74%–85%)	6–7 (66%–83%)
Moderately robust	7–12 (87%–96%)	7–12 (84%–94%)	7–12 (87%–95%)	7–12 (85%–95%)	7–12 (83%–95 %)
Extremely robust	≥13 (97%)	≥13 (95%)	≥13 (96%)	≥13 (96%)	≥13 (96%)

Note: Each entry shows the cutoff values based on the cumulative proportions of the empirical distributions and the cumulative percentiles based on the corresponding cutoff values.

Table 5. Proposed rules of thumb for assessing the magnitudes of the fragility of individual clinical studies based on the fragility quotient.

	Fisher's exact test	**Chi-square test**	**RD**	**OR**	**RR**
Significant studies					
Extremely fragile	≤0.01 (22%)	≤0.01 (25%)	≤0.01 (24%)	≤0.01 (23%)	≤0.01 (23%)
Moderately fragile	0.01–0.02 (22%–41%)	0.01–0.02 (25%–43%)	0.01–0.02 (24%–43%)	0.01–0.02 (23%–41%)	0.01–0.02 (23%–42%)
Slightly fragile	0.02–0.03 (41%–54%)	0.02–0.03 (43%–55%)	0.02–0.03 (43%–55%)	0.02–0.03 (41%–53%)	0.02–0.03 (42%–54%)
Slightly robust	0.03–0.07 (54%–80%)	0.03–0.07 (55%–80%)	0.03–0.07 (55%–81%)	0.03–0.08 (53%–83%)	0.03–0.08 (54%–83%)
Moderately robust	0.04–0.15 (64%–95%)	0.04–0.15 (65%–95%)	0.04–0.16 (65%–96%)	0.04–0.16 (63%–95%)	0.04–0.15 (64%–95%)
Extremely robust	≥0.15 (95%)	≥0.15 (95%)	≥0.16 (96%)	≥0.16 (95%)	≥0.15 (95%)
Nonsignificant studies					
Extremely fragile	≤0.01 (13%)	≤0.01 (15%)	≤0.01 (20%)	≤0.01 (14%)	≤0.01 (14%)
Moderately fragile	0.01–0.03 (13%–43%)	0.01–0.03 (15%–42%)	0.01–0.02 (20%–38%)	0.01–0.03 (14%–40%)	0.01–0.03 (14%–39%)
Slightly fragile	0.02–0.04 (30%–53%)	0.02–0.04 (30%–52%)	0.02–0.03 (38%–52%)	0.02–0.05 (29%–57%)	0.02–0.05 (28%–56%)
Slightly robust	0.05–0.09 (62%–83%)	0.05–0.10 (60%–84%)	0.04–0.07 (63%–83%)	0.06–0.10 (64%–83%)	0.06–0.11 (62%–82%)
Moderately robust	0.07–0.14 (74%–94%)	0.07–0.16 (71%–94%)	0.05–0.11 (72%–95%)	0.08–0.16 (75%–94%)	0.08–0.19 (72%–94%)
Extremely robust	≥0.15 (95%)	≥0.17 (95%)	≥0.12 (96%)	≥0.17 (95%)	≥0.20 (95%)

Note: Each entry shows the cutoff values based on the cumulative proportions of the empirical distributions and the cumulative percentiles based on the corresponding cutoff values.

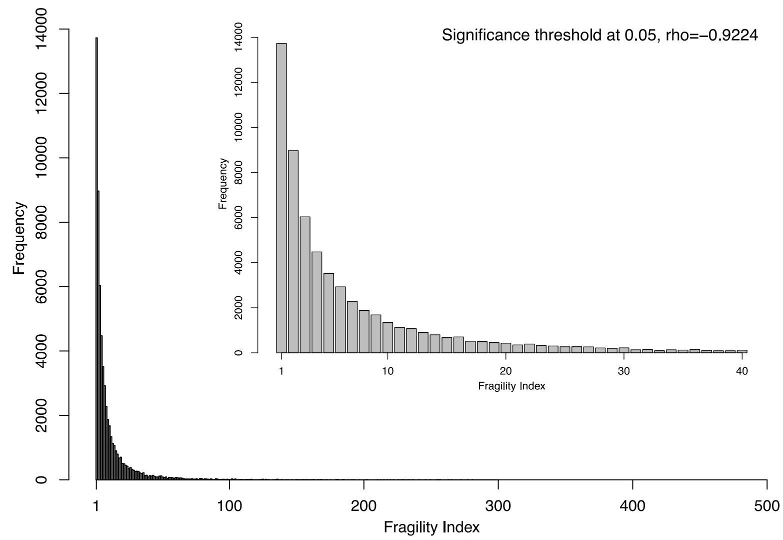

Figure 1. The empirical distribution of the fragility index based on Fisher's exact test at the P-value threshold of 0.05 for significant studies. Spearman's correlation coefficient ρ is used to assess the association between the P-value and fragility index.

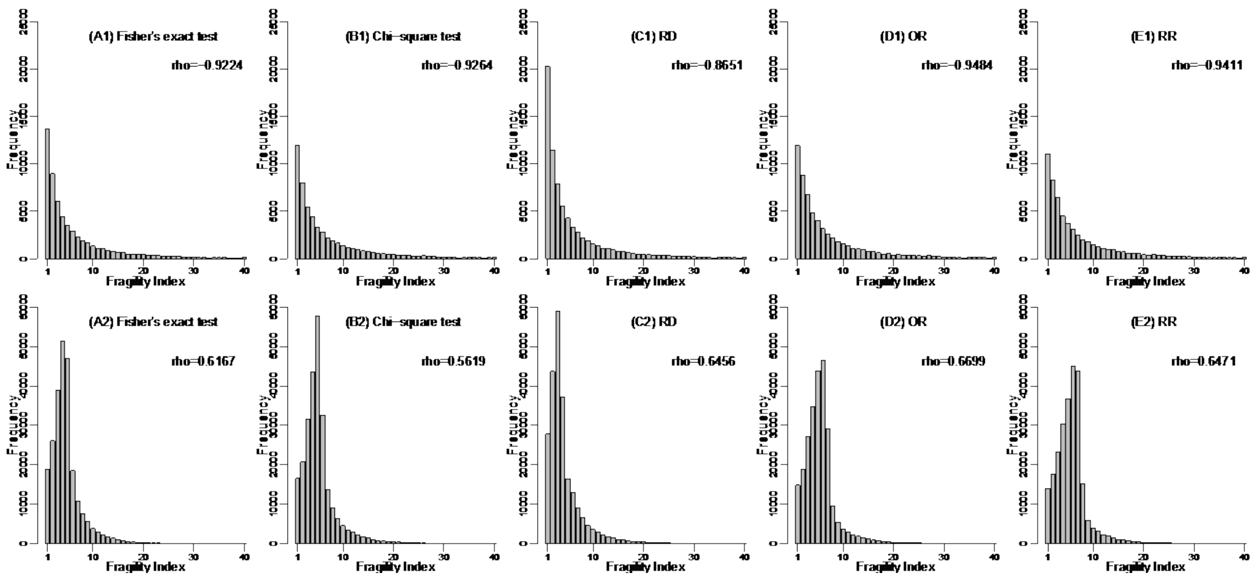

Figure 2. Empirical distributions of the fragility index based on the five statistical methods at the P-value threshold of 0.05. The upper panels (A1–E1) are for significant studies, and the lower panels (A2–E2) are for nonsignificant ones. The fragility index is truncated to 40. Spearman's correlation coefficient ρ is used to assess the association between the P-value and fragility index.

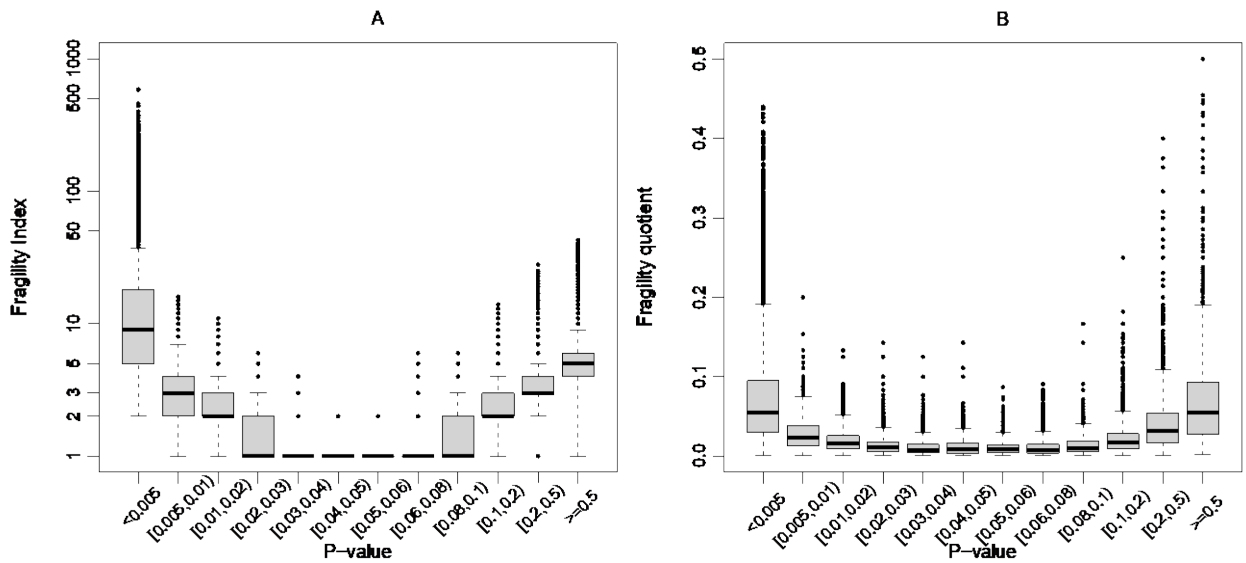

Figure 3. The fragility index (A) and fragility quotient (B) categorized by the P-value based on Fisher's exact test at the P-value threshold of 0.05. (A) The fragility index is presented on a logarithmic scale, and the included studies were truncated to those with FI<1000. (B) Included studies were truncated to those with FQ<0.5.

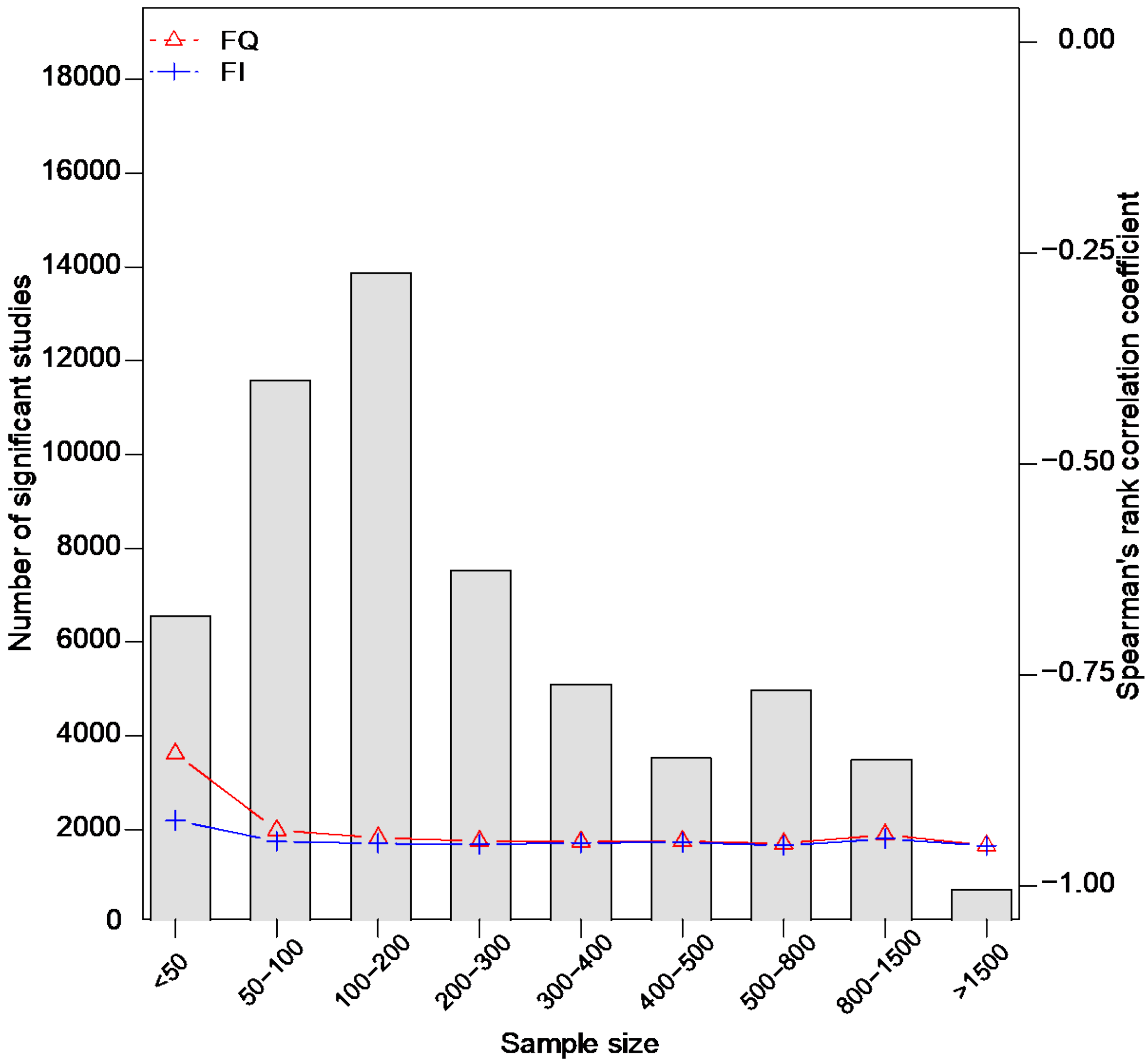

Figure 4. The number of significant studies and Spearman's rank correlation coefficient ρ between the fragility measures and P-value at the P-value threshold of 0.05 categorized by sample size based on Fisher's exact test.