OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE
Breast

Fragility Index and Fragility Quotient in Statistically Significant Randomized Controlled Trials in Plastic Breast Surgery

Ron Skorochod, MD, MPH
Yoav Gronovich, MD, MBA

Background: The fragility index (FI) was conceived as an adjunct to the *P* value, signifying the strength of statistically significant results. The index states the minimal number of patients whose outcome must be changed from "event" to "nonevent" for the results to be statistically nonsignificant. The FI was applied in various medical specialties to assess the robustness of results presented in studies. We aim to assess the robustness of statistically significant results in studies on plastic surgery of the breast and determine factors correlated with studies deemed fragile.
Methods: A systematic literature review of PubMed databases using designated keywords was performed. Background characteristics were extracted from the studies, alongside the significance of outcomes. FI and fragility quotient were calculated for each analyzed outcome and correlated with various baseline characteristics.
Results: FI and fragility quotient were both significantly correlated only with the *P* value of the analyzed outcomes. However, grouping studies based on the *P* value into three categories did not demonstrate a difference in FI. Comparisons of fragile and robust studies did not demonstrate a statistically significant change in terms of baseline variables, except for the mean *P* value of the outcome.
Conclusion: Statistically significant results of randomized controlled trials in plastic surgery of the breast suffer from extensive fragility, and researchers should critically implement their conclusions in their practice. *(Plast Reconstr Surg Glob Open 2024; 12:e5916; doi: 10.1097/GOX.0000000000005916; Published online 20 June 2024.)*

INTRODUCTION

Statistical significance is a measure of the effect that is commonly reported in the context of assessing the efficacy of a certain intervention. The most common measure in the scientific community for statistical significance is the "*P* value," which signifies the probability that the study rejects the null hypothesis despite it being true, or a "type 1 error." The *P* value is typically set as 5%, and results smaller than that are considered statistically significant. Despite the wide recognition and popularity of the *P* value in the literature, it is the subject of great debate, mainly around the robustness of the results it signifies and its potential misuse by researchers and publishers. Several studies demonstrated the possible publication bias that is perpetuated by the *P* value, as articles with statistically significant results are more likely to be published.

Additionally, other studies demonstrated that funding of research was associated with a greater rate of statistically significant results, which can drive research to present mainly their significant results ("cherry picking") or to manipulate the data analysis until it produces statistical significance ("*P* hacking").

To overcome the inherent limitations of the *P* value, Walsh et al[1] proposed the fragility index (FI) for studies with binary outcomes. The FI quantifies the minimal number of patients whose outcome would have to change from a "nonevent" to an "event" to make statistically significant results—not significant. The fragility quotient (FQ) is the FI divided by the sample size of the study.

Since its introduction, the popularity of the FI as a means to interpret statistically significant results and assess the robustness of significant results has grown immensely in various medical fields. When researchers implemented the FI as an adjunct to the traditional *P*

From the Department of Plastic and Reconstructive Surgery, Shaare Zedek Medical Center; Hebrew University Faculty of Medicine, Jerusalem, Israel.

Received for publication March 5, 2024; accepted May 1, 2024.

DOI: 10.1097/GOX.0000000000005916

Disclosure statements are at the end of this article, following the correspondence information.

Related Digital Media are available in the full-text version of the article on www.PRSGlobalOpen.com.

values, the strength of evidence of a large portion of studies was discredited.[2–7] Accurate understanding of the quality of research findings is of utmost importance because randomized controlled trials (RCTs) are considered the highest level of scientific evidence and regularly serve as the cornerstone for clinical guidelines.[8]

Previous studies have implemented the FI to statistically significant outcomes of RCTs in the general subject of plastic surgery. However, due to the wide span of subjects and subspecialties in the field of plastic surgery, we found it critical to separate plastic surgery of the breast from the general overview provided in a previous study. We hoped that by analyzing the robustness of a large subspecialty within plastic surgery individually, we would be able to draw specific meaningful conclusions that are not impacted by confounding factors.

In this article, we aim to assess the robustness of statistically significant RCTs in the field of plastic and reconstructive surgery of the breast. We hope to further deepen our understanding of the breast surgery literature and reinforce critical thinking of researchers reading publications in the field.

Takeaways

Question: What is the robustness of evidence reported by randomized controlled trials in the field of breast surgery?

Findings: A systematic review of the literature was conducted. Relevant studies were defined, and background characteristics were extracted. The FI and FQ were calculated for each outcome and correlated with background characteristics. We found high fragility of studies and the FI and FQ to be significantly correlated with the *P* value.

Meaning: Statistically significant results of randomized controlled trials in plastic surgery of the breast suffer from extensive fragility, and researchers should critically implement their conclusions in their practice.

MATERIALS AND METHODS

Database Search

PubMed electronic database was searched using a dedicated search strategy utilizing keywords related to plastic and reconstructive breast surgery. The search strategy is detailed in the PRISMA flowchart (Fig. 1). Filters limiting

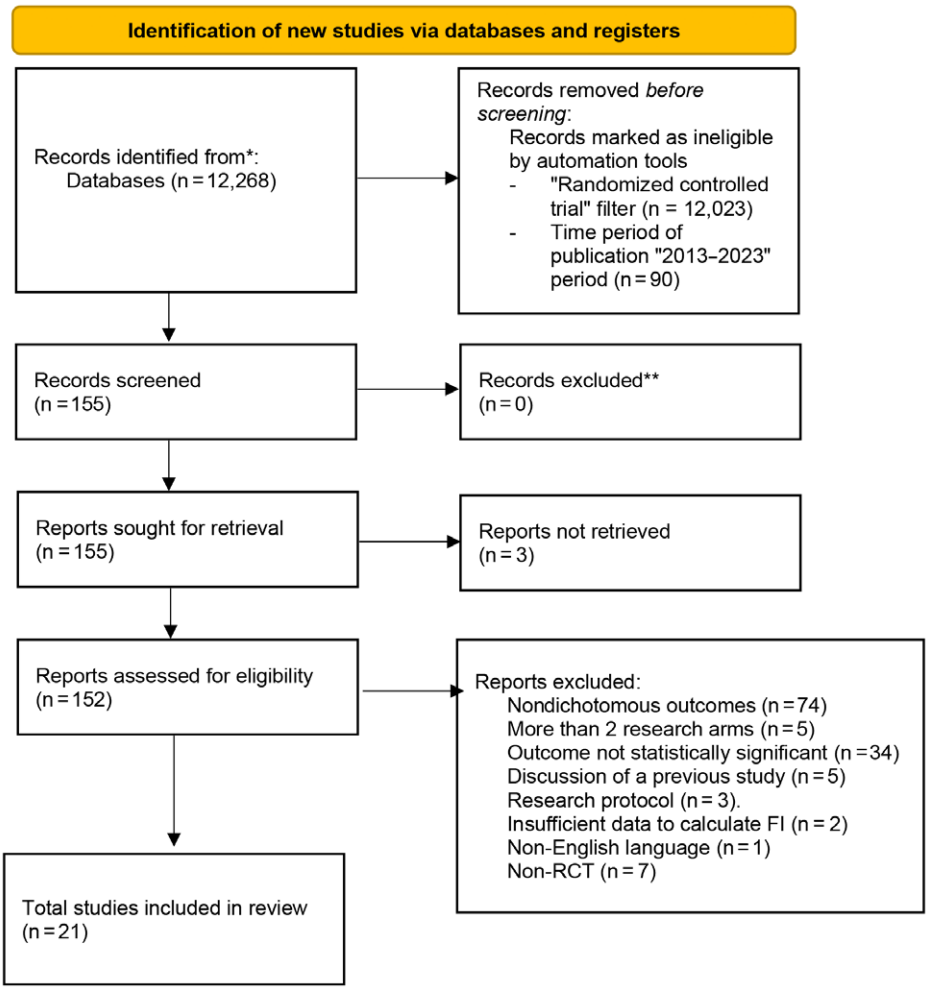

Fig. 1. PRISMA flow chart demonstrating the systematic search process.

the search to the period of 2003–2023 and to RCTs were applied.

The keywords and search strategy used for the review were "(((((((((‘breast reconstruction’[Title]) OR (‘breast reduction’[Title])) OR (‘breast augmentation’[Title])) OR (mammaplasty[Title])) OR (mastopexy[Title])) OR (‘post-mastectomy reconstruction’[Title])) OR (‘post-mastectomy breast reconstruction’[Title])) OR (‘Immediate breast reconstruction’[Title])) OR (‘autologous breast reconstruction’[Title])) OR (‘microvascular breast reconstruction’[Title]),” as seen in Supplemental Digital Content 1. (**See table, Supplemental Digital Content 1**, which displays keywords and phrases used to identify relevant studies. **http://links.lww.com/PRSGO/D302**.) Articles were included if they were RCTs, with dichotomous outcomes that were reported to be statistically significant.

Articles were excluded if they lacked full text, were not in the English language, and had nondichotomous outcomes or nonstatistically significant results. Additionally, articles that did not provide the raw data that are required for the calculation of the FI were also excluded. (**See table, Supplemental Digital Content 2**, which displays the list of excluded articles alongside the reason for exclusion. **http://links.lww.com/PRSGO/D303**.)

Data Extraction

The abstracts and full texts of all eligible articles were read, and relevant demographic information was extracted, alongside the results of statistically significant outcomes. (**See table, Supplemental Digital Content 3**, which displays variables extracted from each of the published articles included in this research article. **http://links.lww.com/PRSGO/D304**.)

Primary and secondary outcomes were extracted. The statistically significant outcome was chosen for analysis. In cases where the primary outcome was nonsignificant, the secondary outcome was chosen if it fulfilled the inclusion criteria. In instances where more than one eligible outcome was available, the primary outcome was chosen. If only the secondary outcomes were available for analysis, the outcome variable with the greatest sample size was chosen. In cases with equal sample size among the secondary outcomes, the most statistically significant outcome was chosen.

FI and FQ Calculation

FI and FQ were calculated for each of the chosen outcomes using the Fisher exact test, by changing one event to a nonevent at a time in the study group.

Statistical Analysis

Categorical variables were described using frequencies, and continuous variables using medians and ranges. Associations were tested with Pearson chi-square and Wilcoxon tests for categorical and continuous variables, respectively. Spearman correlation coefficients were used to test for intervariable correlations, and linear regression was used for predictions of continuous outcomes.

Table 1. Correlations of the FI and FQ with Study Characteristics Using Spearman Coefficient for Associations

Variable	Spearman rho	*P*
FI		
P	0.804 (–)	≤0.001
Loss to follow-up	0.235	0.306
Sample size	0.272	0.233
Citation count	0.340 (–)	0.132
Journal’s IF	0.005	0.132
No. events	0.208	0.366
FQ		
P value	0.718 (–)	≤0.001
Loss to follow-up	0.013 (–)	0.954
Sample size	0.478 (–)	0.29
Citation count	0.220 (–)	0.338
Journal’s IF	0.083 (–)	0.721
No. events	0.099	0.699

RESULTS

One hundred fifty-five studies were retrieved from the initial search after the application of the above-mentioned filters. One hundred thirty-four studies were excluded because they did not meet this study’s inclusion criteria after careful inspection of the abstract and full texts. The search strategy is depicted in the PRISMA flow chart (Fig. 1).

The mean number of patients lost to follow-up was 1.4, and the mean sample size was 94.2. The mean total number of events was 21.2.

The mean FI was 3.6, ranging from 1 to 17, and the mean FQ was 0.056. The mean *P* value was 0.012, 10 (47.6%) of studies were funded, and the primary outcome was the analyzed outcome in nine (43%) studies. Correlation analysis was performed to determine the correlation coefficients of the included studies’ characteristics with the FI and FQ, and can be seen in Table 1.

In an attempt to further understand the association between the *P* value of the study outcome and its corresponding FI, we grouped *P* values into three categories: less than 0.01, 0.01–0.03, and 0.03–0.05. The FI was compared between the categories, using the analysis of variance test, yet it was found to be nonsignificant ($P = 0.139$).

Studies were then classified as “robust” if the FI was more than 3 or “fragile” if the FI was less than 3 or if the number of patients lost to follow-up was greater than the FI. The various study characteristics were compared between the two groups to detect differences. The groups did not differ in terms of the characteristics in question, except for the *P* value, which was significantly lower among the “robust” studies ($P = 0.0003$). The exact values and corresponding *P* values can be observed in Table 2.

DISCUSSION

RCTs are the cornerstone of medical research. They are commonly regarded as the highest level of evidence for establishing causal associations in clinical research and base management guidelines implemented by clinicians worldwide.[8]

Table 2. Comparison of Baseline Study Variables between "Fragile" Studies and "Robust" Studies Based on FI Scores

P	Robust Studies (n = 10)	Fragile Studies (n = 11)	
0.38	7.51	3.28	IF
0.92	28.4	26.8	Citations
0.83	5 (50%)	5 (45.4%)	Funding
0.21	118.1	72.5	Sample size
0.5	2	0.91	Loss to follow-up
0.26	3 (30%)	6 (54.5%)	Primary
0.0003	0.003	0.019	*P*
0.469	25	17.9	No. events

However, the results of RCTs have been put into question, specifically the robustness of statistically significant conclusions. Researchers claim that the commonly regarded threshold for statistical significance of a *P* value less than 0.05 is arbitrary and might not be clinically relevant.[9]

In an attempt to provide an answer to similar claims, the FI was conceptualized by Walsh et al to accompany the reporting of statistically significant results of RCTs. The authors reviewed nearly 400 trials from high-impact medical journals that reported statistically significant results. After calculating the FI, they concluded that a substantial portion of statistically significant results, hinge on a small number of events, and their results should be regarded accordingly.[1]

Over the years, the FI was studied in various medical fields, recognizing the substantial fragility of results obtained from highly regarded RCTs. Authors of these investigations, constantly warrant readers to consider the strength of evidence provided in RCTs and advocate for more robust evidence.[2–7] Khan et al[6] found that in cardiovascular RCTs that were published in six of the highest impact factor journals, the FI was smaller than the number of patients lost to follow-up in 30% of trials, indicating substantial fragility.

Shen et al[7] applied the FI for the results of RCTs in ophthalmology. After careful screening of 156 eligible trials, the authors found that 25% of all screened trials had an FI of 1 or less, and more than 50% of the studies had FI smaller than the number of missing data points.

Pascoal et al[10] found in their report that in gynecologic surgery RCTs, the FI was 0 in 14% of studies, indicating that, by applying the Fisher exact test for the calculations, the results would lose their statistical significance.

In our study, we found 21 eligible studies, reporting significant results of RCTs focusing on plastic and reconstructive breast surgery. After the calculation of FI and FQ, more than 50% of studies were concluded to have fragile results, with an overall mean FI greater than 3 in the entire cohort.

The proportion of fragile yet statistically significant outcomes of RCTs published in plastic and reconstructive surgery of the breast is similar to those reported in the literature of other medical specialties. However, the mean FI seems to be on the higher end of the medical literature spectrum.[11]

Several factors can contribute to this observation, primarily, the skewed distribution of studies. A small number of studies with high FIs can compensate for a high proportion of fragile studies and lead to a high mean FI. This possibility, of statistical outliers, can be a direct result of variability in study quality that can be expected from studies focusing on different study topics with varying methodology.

However, when directly comparing, side-by-side, results of studies that were deemed fragile and robust, no differences were noted in terms of the sample size, number of events and patients lost to follow-up, impact factor of publishing journal, citation count, and external funding. These variables were previously linked to the fragility of studies.[12–14] The lack of association seen in our research suggests homogeneity of studies in known critical domains and demonstrates the quality of evidence derived from research focusing on a specific topic, which in our case was plastic surgery of the breast.

Chin et al[15] performed a systematic analysis of RCTs reporting positive results in the entire spectrum of plastic surgery. The authors found that the mean FI was 1, with more than 25% of articles having an FI of 0. Therefore, the authors concluded that the results of studies in plastic surgery suffer from substantial fragility and warrant careful interpretation. The main criticism discussed in the article is the reliance on significant results on a small number of events, which led to the fragility of their conclusions and ease of rebuttal.

In our report, we focused on an important subspecialty of plastic and reconstructive surgery and aimed to draw in-depth and meaningful conclusions on this specific topic. Although we found similar results to the more general study discussed previously, we learned that none of the studies in breast surgery had an FI of 0, and the robustness of studies was not refuted solely based on the number of patients lost to follow-up. Interpretation of these results can lead us to think that although there is great room for improvement in the strength of evidence of RCTs in breast surgery reports, in the spectrum of plastic surgery literature, breast surgery provided research of acceptable quality, especially compared with other plastic surgery subspecialties. Furthermore, by focusing on a specific subspecialty of plastic surgery, we were able to neutralize the impact of potential external confounders, unrelated to our study, and increased our confidence in the results of this study and the conclusions we drew from it. It is important to recognize the limitations of our study, namely the limited use of scientific databases and the small number of studies that met the inclusion criteria.

In an attempt to provide concise and valuable data, we applied strict exclusion criteria, to ensure that our results would be as little confounded as possible. However, it comes with a toll in the form of a small sample size and low heterogeneity of studies. Furthermore, the extracted characteristics of the studies we chose to analyze are limited in their extent, and some variables could have further enriched our understanding of the contributing factors.

In conclusion, statistical significance, especially at the arbitrary threshold of 5%, is dependent on the sample size

used in the study. To an extent, every result will become statistically significant when a large enough sample size is analyzed. Therefore, it is true to assume that by increasing the number of events and the sample size, the robustness of results would increase. However, the scientific community must take into consideration the importance of RCTs that are published despite their small sample size, namely because it allows for quicker dissemination of hypotheses and ideas to the worldwide scientific community. Relying solely on large, multicenter studies inhibits us from learning about innovations in real-time and considering them in day-to-day practice. Therefore, although it is crucial to carefully interpret the results of small-sized trials and to take into consideration the need for more extensive evidence and experience, we believe they are crucial for setting the ground for larger and more robust studies

Ron Skorochod, MD, MPH
Department of Plastic and Reconstructive Surgery
Shaare Zedek Medical Center;
Hebrew University Faculty of Medicine
Jerusalem, Israel
E-mail: Ron.skorochod@mail.huji.ac.il

DISCLOSURE

The authors have no financial interest to declare in relation to the content of this article.

REFERENCES

1. 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. *J Clin Epidemiol.* 2014;67:622–628.
2. Placer-Galán C, Enriquez-Navascués JM, Lopetegui AE, et al. An analysis of randomized controlled trials on anal fistula conducted between 2000 and 2020 based on the fragility index and reverse fragility index. *Colorectal Dis.* 2023;25:1572–1577.
3. Topcuoglu MA, Arsava EM. The fragility index in randomized controlled trials for patent foramen ovale closure in cryptogenic stroke. *J Stroke Cerebrovasc Dis.* 2019;28:1636–1639.
4. Ridgeon EE, Young PJ, Bellomo R, et al. The fragility index in multicenter randomized controlled critical care trials. *Crit Care Med.* 2016;44:1278–1284.
5. Ruzbarsky JJ, Khormaee S, Daluiski A. The fragility index in hand surgery randomized controlled trials. *J Hand Surg Am.* 2019;44:698.e1–698.e7.
6. Khan MS, Ochani RK, Shaikh A, et al. Fragility index in cardiovascular randomized controlled trials. *Circ Cardiovasc Qual Outcomes.* 2019;12:e005755.
7. Shen C, Shamsudeen I, Farrokhyar F, et al. Fragility of results in ophthalmology randomized controlled trials: a systematic review. *Ophthalmology.* 2018;125:642–648.
8. Del Paggio JC, Tannock IF. The fragility of phase 3 trials supporting FDA- approved anticancer medicine: a retrospective analysis. *Lancet Oncol.* 2019;20:1065–1069.
9. Zabor EC, Kaizer AM, Hobbs BP. Randomized controlled trials. *Chest.* 2020;158(1S):S79–S87.
10. Pascoal E, Liu M, Lin L, et al. The fragility of statistically significant results in gynaecologic surgery: a systematic review. *J Obstet Gynaecol Can.* 2022;44:508–514.
11. Kampman JM, Turgman O, Sperna Weiland NH, et al. Statistical robustness of randomized controlled trials in high-impact journals has improved but was low across medical specialties. *J Clin Epidemiol.* 2022;150:165–170.
12. Hayes J, Zuercher M, Gai N, et al. The fragility index of randomized controlled trials in pediatric anesthesiology. *Can J Anaesth.* 2023;70:1449–1460.
13. Muthu S, Ramakrishnan E. Fragility analysis of statistically significant outcomes of randomized control trials in spine surgery: a systematic review. *Spine (Phila Pa 1976).* 2021;46:198–208.
14. Shamsudeen I, Farrokhyar F, Sabri K. Fragility of results in ophthalmology randomized controlled trials: a systematic review. *Ophthalmology.* 2018;125:642–648.
15. Chin B, Copeland A, Gallo L, et al. The fragility of statistically significant randomized controlled trials in plastic surgery. *Plast Reconstr Surg.* 2019;144:1238–1245.