ORIGINAL CONTRIBUTION

When the *p* Value Doesn't Cut It: The Fragility Index Applied to Randomized Controlled Trials in Colorectal Surgery

David W. Nelms, M.D. • H. David Vargas, M.D. • Ryan S. Bedi, M.D.
Jennifer L. Paruch, M.D., M.S.

Department of Colon and Rectal Surgery at the Ochsner Clinic, New Orleans, Louisiana

BACKGROUND: The American Statistical Association, among others, has called for the use of statistical methods beyond $p \leq 0.05$. The fragility index is a statistical metric defined as the minimum number of patients for whom if an event rather than a nonevent occurred, then the *p* value would increase to ≥0.05. Previous reviews have demonstrated that many randomized controlled trials have a low fragility index, suggesting they may not be robust.

OBJECTIVE: The purpose of this study was to review the fragility indices of randomized controlled trials in colorectal surgery.

DATA SOURCES: A PubMed search was performed.

STUDY SELECTION: Colorectal surgery randomized controlled trials with a dichotomous primary outcome $p \leq 0.05$ and publication between 2016 and 2018 were systematically identified.

INTERVENTIONS: All procedural interventions related to colorectal surgery were included.

MAIN OUTCOME MEASURES: The main measures were the fragility index and the number of patients lost to follow-up for each trial. The percentage of trials with the number of patients lost to follow-up greater than the fragility index was calculated.

RESULTS: In total, 712 abstracts were reviewed, with 90 trials meeting the inclusion criteria. The median fragility index was 3 (interquartile range of 1 to 10). In 51 of the 90 trials (57%), the number of patients lost to follow-up was greater than the fragility index.

LIMITATIONS: The fragility index is only one measure of the robustness of a randomized clinical trial.

CONCLUSIONS: Most colorectal surgery randomized controlled trials have a low fragility index. In 57% of trials, more patients were lost to follow-up than would be required to change the outcome of the trial from "significant" to "nonsignificant" based on the *p* value. This emphasizes the importance of assessing the robustness of clinical trials when considering their clinical application, rather than relying solely on the *p* value. See **Video Abstract** at http://links.lww.com/DCR/B741.

Funding/Support: None reported.

Financial Disclosure: None reported.

Accepted as plenary abstract session VIII: Best of 2020 hybrid virtual meeting in San Diego, CA. Presentation # WP7.

Accepted as a plenary presentation at the American Society of Colon and Rectal Surgeons (ASCRS) meeting, June 6 to 10, 2020, in Boston, MA. Due to meeting cancellation from COVID-19, the presentation was submitted on the virtual ASCRS 2020 platform. Abstract #109.

Correspondence: Jennifer L. Paruch, M.D., M.Sc., 1514 Jefferson Highway, New Orleans, LA 70121. Email: Jennifer.paruch@ochsner.org. Twitter: @jlparuch

Dis Colon Rectum 2022; 65: 276–283
DOI: 10.1097/DCR.0000000000002146

CUANDO EL VALOR-P ES INSUFICIENTE: ÍNDICE DE FRAGILIDAD APLICADO EN ESTUDIOS ALEATORIOS CONTROLADOS EN CIRUGÍA COLORECTAL

ANTECEDENTES: La Sociedad Estadounidense de Estadística, entre otros, ha pedido el uso de métodos estadísticos más allá de p <0,05. El índice de fragilidad es una medida estadística definida como el número de desenlaces que podrían cambiar para revertir, o conseguir, la significación estadística, así el valor p aumentaría a ≥ 0,05. Las revisiones anteriores han demostrado que muchos estudios aleatorios controlados tienen un índice de fragilidad bajo, lo que sugiere que pueden poco sólidos.

OBJETIVO: El propósito de la présente investigación fué de revisar los índices de fragilidad de los estudios aleatorios controlados en cirugía colorrectal.

FUENTES DE DATOS: PubMed.

SELECCIÓN DE ESTUDIOS: Se identificaron sistemáticamente estudios aleatorios controlados de cirugía colorrectal con un resultado primario dicotómico, valor de p ≤ 0,05 y publicados entre 2016-2018.

INTERVENCIONES: Se incluyeron todas aquellas intervenciones con procedimientos relacionados con la cirugía colorrectal.

PRINCIPALES MEDIDAS DE RESULTADO: Las principales medidas fueron: el índice de fragilidad y el número de pacientes perdidos durante el seguimiento en cada estudio. Se calculó el el índice de fragilidad en porcentaje de estudios con el mayor número de pacientes perdidos durante el seguimiento mas prolongado.

RESULTADOS: En total, se revisaron 712 resúmenes con 90 ensayos que cumplieron con los criterios de inclusión. La mediana del índice de fragilidad fue de 3 (rango intercuartíl de 1 a 10). En 51 de los 90 estudios (57%), el número de pacientes perdidos durante el seguimiento fue mayor que el índice de fragilidad.

LIMITACIONES: El índice de fragilidad es solo una medida de la robustez de un estúdio clínico aleatorio.

CONCLUSIONES: La mayoría de los estudios aleatorios y controlados en cirugía colorrectal tienen un índice de fragilidad bajo. En el 57% de los estudios, se perdieron más pacientes durante el seguimiento de los que se necesitarían para cambiar el resultado del estudios de grado "significativo" a un grado "no significativo" según el valor-p. Este concepto enfatiza la importancia de evaluar la robustez de los estudios clínicos al considerar su aplicación verdadera aplicación clínica, en lugar de depender únicamente del valor-p. Consulte **Video Resumen** en http://links.lww.com/DCR/B741. *(Traducción—Dr. Xavier Delgadillo)*

KEY WORDS: Colorectal surgery; Fragility index; Randomized controlled trials; Research methodology; Statistical significance.

INTRODUCTION

As surgeons, we strive to provide our patients with the best care by utilizing the most credible scientific evidence. The highest form of clinical scientific evidence is the replicable randomized controlled trial (RCT). However, the replicability of clinical RCTs is plaguing the medical and surgical community,[1–6] and this problem has been termed the "replication crisis." While a variety of factors likely contribute to the high rates of nonreplicable trial results, reliance on a *p* value ≤0.05 as criterion for "statistical significance" is cited as a major contributor.[7–9]

Unfortunately, the *p* value and its assumptions are often misunderstood and misused.[7,10] Mounting evidence regarding the inappropriate use of the *p* value caused the American Statistical Association (ASA) to publish a 2016 statement on *p* values,[11] and in 2019, ASA published an entire supplemental journal issue devoted to statistical inference beyond the *p* value alone.[12] Some experts have even called for the entire concept of "statistical significance" to be abolished, including 800 signatories of a 2019 comment in *Nature*.[13]

Definition of the Fragility Index

Within this context, the fragility index (FI) emerged as a tool by which to judge the robustness of RCTs. First defined by Walsh et al in 2014,[14] the FI is the minimum number of patients for whom if an event rather than a nonevent occurred, the *p* value would no longer be <0.05. For example, an FI of 1 in a trial measuring surgical site infection (SSI) would mean if only one patient in the study had developed an SSI, the *p* value would increase to ≥0.05. The FI only applies to trials with dichotomous (ie, event versus nonevent) outcome variables. Trials with a large FI have been termed "robust," meaning the statistical significance of the trial can withstand many changes in patient outcomes, whereas trials with a small FI have been termed "fragile," meaning only a few changes in patient outcomes lead to loss of statistical significance.[14]

The Fragility Index of Published RCTs

Since inception by Walsh et al in 2014, the number of publications reviewing the FI of published RCTs in medical and surgical specialties has grown exponentially. Tignanelli and Napolitano published a 2019 study in *JAMA Surgery* in which they identified 25 trauma-related RCTs and found a median FI of only 3, with an interquartile range (IQR) of 1 to 8.[15] In their conclusion, they strongly recommended routine reporting of the FI for all trauma and surgery RCTs.[15]

No previous metaresearch of the FI of colorectal surgery–specific literature has been published. The purpose of this study was to determine the FIs of colorectal surgery RCTs. In addition, we sought to compare the FI of each trial with the number of patients lost to follow-up.

MATERIALS AND METHODS

Identification of Trial Abstracts

A PubMed search was performed for colorectal surgery–related abstracts using MeSH (Medical Subject Headings) terms. Colorectal disease related MeSH terms were generated using the index of *The ASCRS Textbook of Colon and*

Rectal Surgery, 3rd edition.[16] For each index entry in the ASCRS textbook, the PubMed MeSH database was used to search for an associated MeSH term. Redundancies in this list of MeSH terms were then eliminated by removal of terms that were subcategories of higher categories. A list of surgery/procedural-related MeSH terms was then also generated. A PubMed search expression was created by using Boolean operators such that each colorectal disease MeSH term was searched with the predicate that there was an associated surgery/procedure-related MeSH term. (See the Supplemental Appendix at http://links.lww.com/DCR/B742 for further details, including the complete list of MeSH terms and the final PubMed search expression.)

PubMed search filters were used to obtain abstracts for RCTs in only English and with human participants. The search was performed for articles with a true publication date from January 1, 2016, to December 31, 2018.

Manual Review of Abstracts for Inclusion Criteria

Each identified abstract was manually reviewed to determine whether the article met inclusion criteria. Included articles were 1) colorectal surgery related, 2) prospective RCTs (post hoc analyses of RCTs were excluded), 3) superiority design (ie, exclusion of noninferiority trials), 4) 2×2 factorial or two parallel arm design with a dichotomous outcome variable (>2 parallel arm and time-to-event analyses performed using Kaplan-Meier methods were excluded), and 5) those reported in the abstract to have a dichotomous primary outcome with a p value ≤ 0.05 or equivalent 95% confidence interval (CI), not including the null hypothesis except on the boundary. For all abstracts meeting inclusion criteria, the full text publication was obtained.

Data Extraction From Included Articles

For each included article, data were extracted for the primary outcome variable to generate a 2×2 contingency table of the two groups compared versus the two outcomes that occurred. The p value reported in the publication for this data table was recorded. Additionally, for each article, the number of patients lost to follow-up was recorded. "Lost to follow-up" was defined as the number of patients initially randomized for whom measurement of the primary outcome was not available as defined by Akl et al.[17] For papers in which loss to follow-up by this definition was not explicitly reported or inferable from the data, the number of patients lost to follow-up was recorded as 0.

Each included article was also categorized by topic according to the six pillars of colorectal disease (perioperative/endoscopy, anorectal disease, malignant disease, benign disease, pelvic floor disorders, and miscellaneous), as used to organize the ASCRS textbook. Additionally, the perioperative/endoscopy pillar was subdivided into individual categories of perioperative and endoscopy.

Calculation of the Fragility Index and Modified Fragility Index

The FI was calculated as previously reported by Walsh et al.[14] The group with the smaller number of events was iteratively added an event while simultaneously subtracting a nonevent (to keep the total group size the same) until the p value was ≥ 0.05 as calculated by a two-sided Fisher exact test. The number of iterations required to meet this criterion was the FI. If the p value calculated by the two-sided Fisher exact test was ≥0.05 for the original contingency table before changing any of the events, the FI was defined as 0.

Due to a concern that the Fisher exact test may be too conservative for many of the included trials, we also created a modified FI, which only differed from the original definition by calculating the p value with a χ^2 test rather than the Fisher exact test. All calculations, for both the FI and modified FI, were performed in Microsoft Excel 365 utilizing Microsoft Visual Basic for Applications version 7.1.

Statistical Analysis

A plot of the frequency distribution of FI was created. The overall median FI and interquartile range for all included trials were calculated in Excel. Additionally, the median FI and interquartile range were calculated for the categories of year, pillar of colorectal disease, most frequent journals, reported p value range, patient lost to follow-up range, and trial total sample size range. This was similarly performed for the number of patients lost to follow-up. The percentage of publications with an FI greater than or equal to the number of patients lost to follow-up was calculated. Finally, correlation analysis investigating the associations of FI with the p value and the FI with the sample size were performed using the Spearman rank correlation coefficient calculated with GraphPad Prism 9.0.

RESULTS

Study Selection

In total, we reviewed 712 abstracts. Of these, 90 studies met inclusion criteria (Fig. 1). Characteristics of the included trials are listed in Table 1. The p value was greater than 0.01 in 43% of included papers. The most frequent pillar of colorectal disease was perioperative/endoscopy. The included articles came from 52 different journals with the three most frequent journals being *Endoscopy*, *Annals of Surgery*, and the *American Journal of Gastroenterology*. The median sample size was 186, with an IQR of 91 to 313. The median number of events was 17.5, with an IQR of 4 to 48.

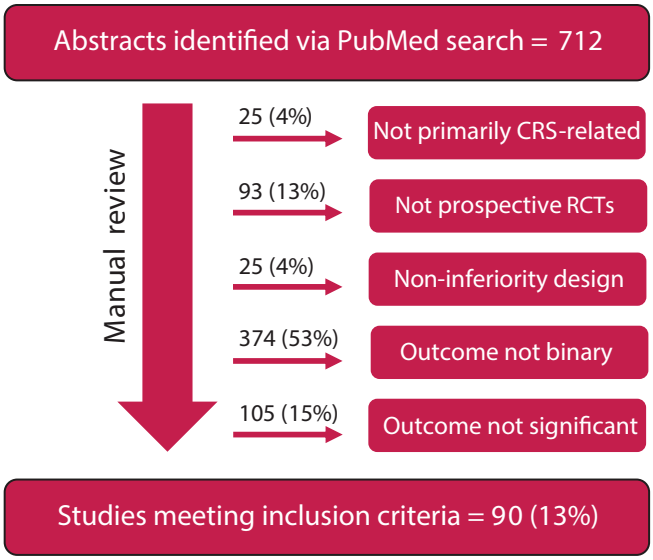

FIGURE 1. Flow diagram of the study selection process. CRS = colorectal surgery; RCT = randomized controlled trial.

Fragility Index and Modified Fragility Index

Figure 2 displays the number of studies per FI value. The median FI was 3, with an IQR of 1 to 10 (Table 1). The data are skewed, with most RCTs having low FIs and a few studies having large outlier FIs. The FI was less than or equal to the median of 3 in 57% (51 of 90) of studies. FI values of 0 occurred mainly because some studies reported a p value ≤0.05 using a statistical test such as the χ^2 test, but when recalculated with the Fisher exact test, the p value was ≥0.05 without changing any patient outcomes.

Analysis by subgroup is also shown in Table 1. The median FI was similar across the 3 years analyzed. The pillars of endoscopy and anorectal had the largest median FI of 4, while pelvic floor and miscellaneous each had the lowest values of 0, and each of these pillars only had one included study. The median FI was smallest in the subgroup of trials with the largest p values and largest in the subgroup of trials with the smallest p values.

The overall modified FI using the χ^2 test had a median of 3, with an IQR of 2 to 10. Per year, the modified FI had median values of 4, 3, and 3 in 2016, 2017, and 2018, respectively. In 50% of the trials, the modified FI was less than or equal to the median of 3.

Number of Patients Lost to Follow-up

Figure 3 graphs the number of studies per number lost to follow-up. The median number lost to follow-up was 3 (IQR 0 to 17). Table 1 shows the number lost to follow-up per subgroup analyzed. Figure 4 displays the net difference between lost to follow-up and the FI; in 57% of included studies, the number of patients lost to follow-up was greater than the FI for that study. This finding was consistent over the 3 years analyzed, with values of 67%, 50%, and 55% over 2016, 2017, and 2018, respectively. Using the modified FI, the percentage of studies with the number of patients lost to follow-up greater than the modified FI was 47%.

Relationship of FI to p Value and Sample Size

There is a strong inverse correlation between the FI and the p value with a Spearman rank correlation coefficient (r) of -0.93, with 99% CI of -0.96 to -0.88. There is a weak-to-moderate positive correlation between the FI and the total sample size, with r = +0.53 (0.30 to 0.71 99% CI). There is a similar weak-to-moderate direct correlation between the FI and the number of events, with r = +0.57 (0.34 to 0.73 99% CI). Scatter plots are included in the Supplemental Appendix at http://links.lww.com/DCR/B742.

DISCUSSION

Principle Findings and Context in the Literature

This study finds that the FI is low (median of 3) for most RCTs in colorectal surgery. In over half the trials, the FI was less than or equal to this median, so in most trials only a few changes in patient outcomes resulted in loss of "statistical significance." Additionally, we found in most trials (57%) that the number of patients lost to follow-up was greater than the FI. This implies that the unknown outcomes of these patients lost to follow-up could easily have caused the trials to lose "statistical significance" based on p value cutoff of 0.05. This suggests that many colorectal surgery RCT results are fragile and their conclusions may not be replicable.

Furthermore, to substantiate these findings, analysis was repeated using a modified FI calculated with the less conservative χ^2 test. Using the χ^2 test, the median modified FI remained the same as the FI, with a value of 3, and 50% of the trials still only required a few (≤3) changes in patient outcomes to lose statistical significance. The number of studies with the number of patients lost to follow-up greater than the modified FI remained high, at 47%.

Our results are similar to those of previous FI meta-research in other medical and surgical specialties.[14,15,18–21] There is no accepted FI cutoff for a robust versus fragile trial.[15] Narayan et al evaluated 41 RCTs in urology and found a median FI of 3 with an IQR of 1 to 4.5.[19] One of the highest median FIs found was the original publication by Walsh et al, which evaluated RCTs only in high-impact medical journals, and found a median FI of 8, with an IRQ of 0 to 109.[14] Others have also compared the FI to the number of patients lost to follow-up as a potential reference point and found high rates of the number lost to follow-up being greater than FI, such as 67.5% in the urology review by Narayan et al.[18,19]

TABLE 1. Included trial characteristics, FI, and LTF overall and by subgroup

Category	*n (%)*	*FI*	*IQR*	*LTF*	*IQR*
Overall	*90 (100%)*	*3*	*(1–10)*	*3*	*(0–17)*
Year					
2016	27 (30%)	3	(1–27.5)	6	(0–26)
2017	34 (38%)	3	(1–7)	3	(0–13.5)
2018	29 (32%)	2	(1–7)	0	(0–13)
Pillar of colorectal disease					
Perioperative/endoscopy	45 (50%)	3	(1–15)	2	(0–13)
Perioperative	14 (16%)	1	(0–3)	0.5	(0–5)
Endoscopy	31 (34%)	4	(1.5–25.5)	3	(0–15.5)
Anorectal disease	8 (9%)	4	(2–9.5)	7	(0–16)
Malignant disease	20 (22%)	2	(1–29)	5.5	(0–14)
Benign disease	15 (17%)	3	(2–5.5)	17	(0.5–14)
Pelvic floor disorders	1 (1%)	0	N/A[a]	0	N/A[a]
Miscellaneous	1 (1%)	0	N/A[a]	0	N/A[a]
Most frequent journals					
Endoscopy	7 (13%)	1	(1–10)	0	(0–4)
Ann Surg	5 (10%)	5	(2–9)	2	(0–7)
Am J Gastroenterol	4 (8%)	21.5	(3–49)	12	(0–29)
Br J Surg	4 (8%)	2	(2–2)	12	(11.5–14)
Gastrointest Endosc	4 (8%)	3	(1–10)	0	(0–2)
Int J Colorectal Dis	4 (8%)	2.5	(1.5–3)	9.5	(4.5–29.5)
Medicine (Baltimore)	4 (8%)	17	(4–31.5)	8	(2–41.5)
Surg Endosc	4 (8%)	4.5	(2.5–6)	1	(0–3)
Clin Gastroenterol Hepatol	3 (6%)	7	(4.5–20.5)	20	(10–42.5)
Dis Colon Rectum	3 (6%)	7	(5–7)	2	(1–15.5)
Lancet	3 (6%)	2	(1.5–5.5)	33	(27.5–37)
Reported *p* values					
$0.01 < p$ value ≤ 0.05	39 (43%)	1	(0–2)	1	(0–10)
$0.001 < p$ value ≤ 0.01	22 (24%)	3	(2–6)	7.5	(0–18.5)
p value ≤ 0.001	29 (32%)	18	(7–35)	7	(0–32)
Patients lost to follow-up					
0–2	44 (49%)	2	(1–7)	N/A	
3–16	23 (26%)	3	(1–13)	N/A	
17 or more	23 (26%)	4	(2–31.5)	N/A	
Trial total sample size					
1–90	22 (24%)	2	(0–3)	1	(0–16.5)
91–185	23 (26%)	1	(1–2)	0	(0–9)
186–312	22 (24%)	6.5	(2.5–24)	4.5	(0–11.25)
313 or more	23 (26%)	15	(2.5–36)	13	(1–36.5)

FI = fragility index; IQR = interquartile range; LTF = lost to follow-up; N/A = not applicable.
[a]The IQR was N/A because there was only 1 datum.

Strengths and Weaknesses of This Study

To our knowledge, this is the first study of the FI in RCTs specifically related to colorectal surgery. This study assesses only one metric of fragility and quality of the reviewed RCTs. Therefore, the finding that most colorectal RCTs may be fragile is not a judgment of the overall quality of colorectal surgery research. Our findings highlight how the general problem of the replication crisis and the use of statistics are worthy of discussion within colorectal surgery.

Meaning of the Findings

Conceptually, a low FI is not inherently bad. If the goal of a trial is to provide evidence for a p value <0.05, then an FI of 1 could be interpreted as an efficiently performed RCT, utilizing near the least number of patients necessary to meet this criterion.[20] However, a p value <0.05 alone is a poor criterion upon which to base any conclusion because this finding alone can be associated with a high false discovery rate (estimates of 14%–50%[7,22–24]) and inability to replicate findings.[1,2,6,7,9,11,25,26] Specifically, judging solely

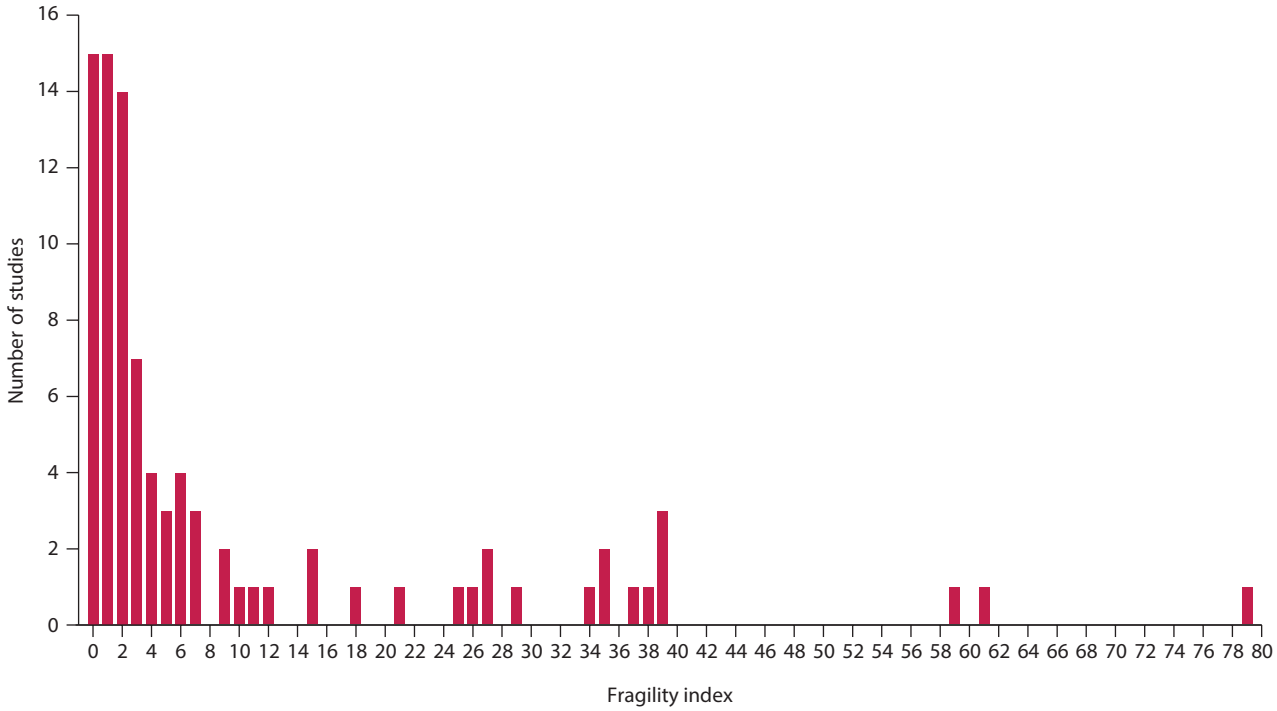

FIGURE 2. Frequency distribution of the number of studies per fragility index.

on *p* value ignores the importance of statistical power (and sample size) in evaluating trials with a "significant" result.[25] The probability of a false finding (false discovery rate) is not given by the *p* value, but instead is related to the *p* value, the statistical power, and the pretrial probability through Bayes' theorem.[10]

The FI has been shown to be directly associated with power and sample size.[20] Therefore, a low FI can mean

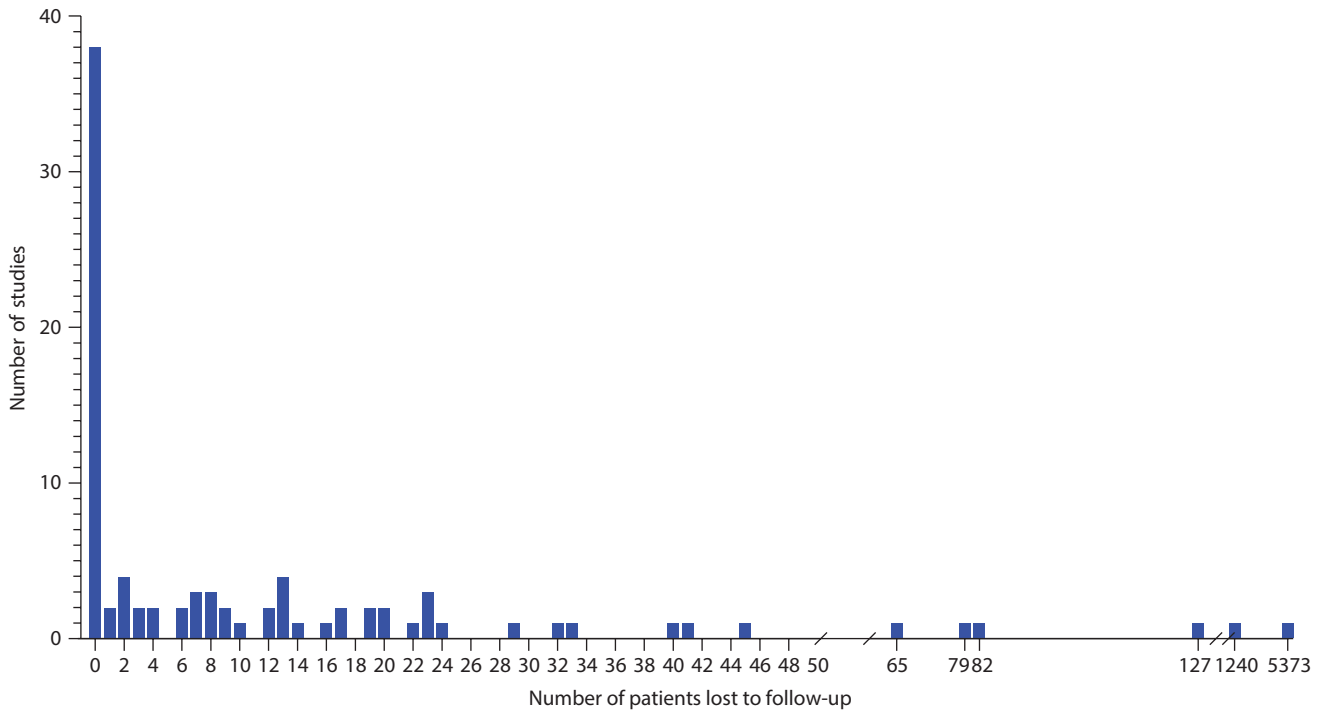

FIGURE 3. Frequency distribution of the number of studies per number of patients lost to follow-up.

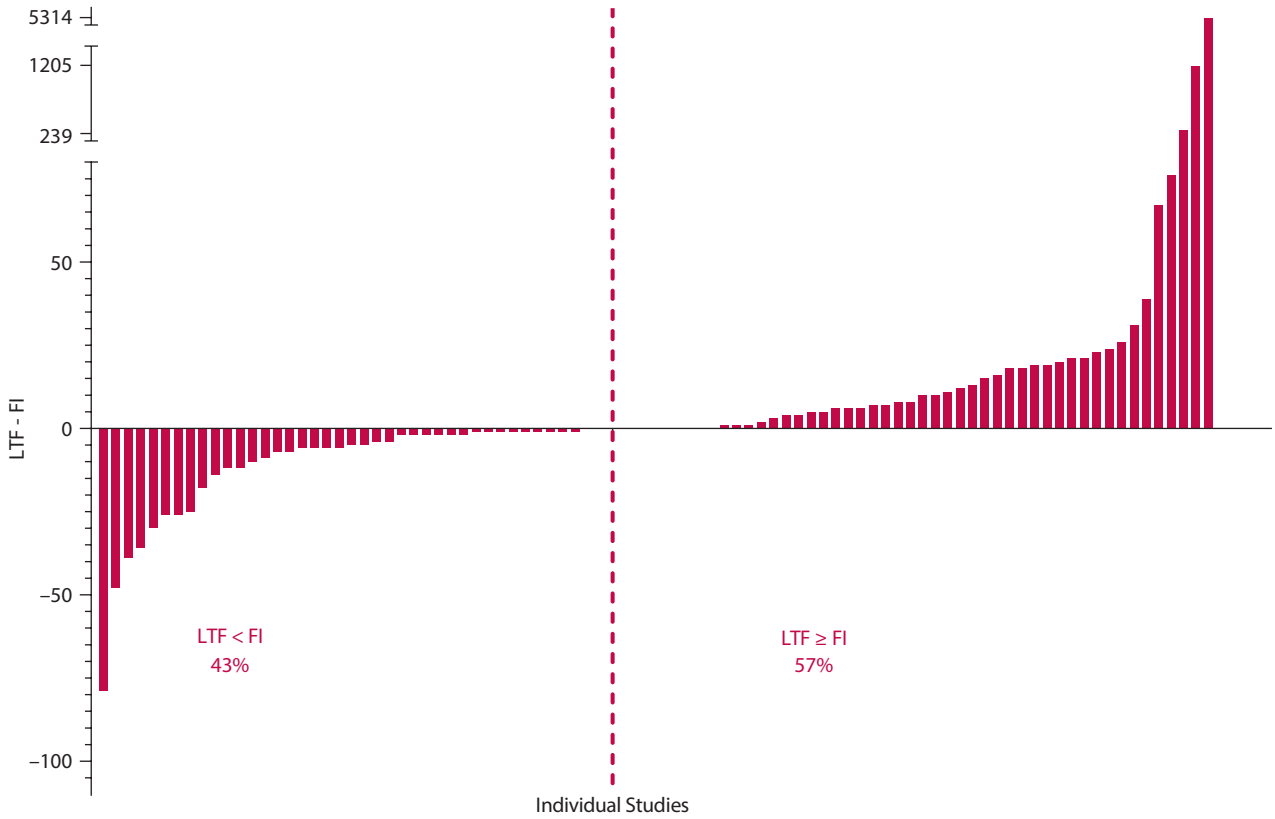

FIGURE 4. Net difference between the number of patients lost to follow-up (LTF) and the fragility index (FI). The net difference is positive (ie, the LTF ≥ FI) in 57% of the trials.

the *p* value of a trial is near 0.05 and/or it has low power (sample size), translating to a possible high false discovery rate and inability to replicate results rather than an efficient true discovery. This fits with the intuitive interpretation of the FI that a result that hinges on only a few patient outcomes may not be reliable.

Practical Conclusions From the Findings

From a practical standpoint, our findings highlight that the results of a clinical trial should not be interpreted based on the *p* value alone. The utility of the FI is that it is an intuitive way for a clinician to see how small changes in outcomes effect the *p* value. The *p* value should not be judged as a fixed exact number for a given trial, and in fact, it has been shown that wide variation in the *p* value is expected with repetition of a trial with statistical power <90%.[25] The FI contains information about the interrelated *p* value, sample size, and statistical power of a trial,[20] and as such can assist in interpretation of a trial.

However, the FI does have several limitations. First, it can only be used for dichotomous outcomes, so it is not generalizable for reporting in all RCTs. Second, there is no acceptable FI cut-off by which to judge an individual trial as fragile or robust. The lost to follow-up number of a trial may be a useful comparison, but with the large number of variables that contribute to lost to follow-up, this is certainly not a perfect criterion. Next, the FI contains the arbitrary significance level of 0.05, which we do not believe deserves further emphasis.

Implications for Future Research

Future research is specifically needed on how routine reporting of the FI would affect clinicians' ability to interpret RCTs.[20,27] Consideration should be given to routine reporting of the FI for colorectal surgery RCTs with binary outcomes. Lost to follow-up missing data also need to be minimized in RCTs to allow complete intention-to-treat interpretation.[28] A variety of other suggestions have been made regarding how to improve RCT interpretation, and complete review is beyond the scope of this article. The ASA statement on the *p* value provides an excellent review of the principles for appropriate interpretation of the *p* value.[11] The ASA also provides resources regarding alternative methods to the *p* value.[12]

CONCLUSIONS

The FI of most colorectal surgery RCTs is low. For the majority of trials (57%), the number of patients lost to follow-up is larger than the FI. Therefore, many "significant" trial results in colorectal surgery may not be robust or replicable. This finding is not unique to colorectal surgery but highlights the need for continued discussions of statistical methods in colorectal surgery research.

The FI is a useful tool to help clinicians interpret RCTs. As statisticians and experts call for the use of statistical methods beyond the *p* value alone, it is important for colorectal surgeons to build a statistical armamentarium. Randomized controlled trials must be judged as a whole, within the context of previous evidence, and not upon any single metric.

ACKNOWLEDGMENTS

We dedicate this work to the deceased David A. Margolin, M.D., F.A.C.S., F.A.S.C.R.S., who was director of colorectal research at the Ochsner Clinic, professor of surgery at the University of Queensland and Ochsner Schools of Medicine, and past president of the American Society of Colon and Rectal Surgeons (2018-2019). Dr. Margolin was instrumental in the conception of this project.

REFERENCES

1. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218–228.
2. Makel MC, Plucker JA, Hegarty B. Replications in psychology research: how often do they really occur? *Perspect Psychol Sci*. 2012;7:537–542.
3. Niven DJ, McCormick TJ, Straus SE, et al. Reproducibility of clinical research in critical care: a scoping review. *BMC Med*. 2018;16:26.
4. Prasad V, Gall V, Cifu A. The frequency of medical reversal. *Arch Intern Med*. 2011;171:1675–1676.
5. Prasad V, Vandross A, Toomey C, et al. A decade of reversal: an analysis of 146 contradicted medical practices. *Mayo Clin Proc*. 2013;88:790–798.
6. Lindsay DS. Replication in Psychological Science. *Psychol Sci*. 2015;26:1827–1832.
7. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
8. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ*. 2001;322:226–231.
9. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004;96:434–442.
10. Vidgen B, Yasserie T. P-values: misunderstood and misused. *Front Phys*. 2016;4:1–5.
11. Wasserstein RL, Lazar NA. The ASA Statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129–133.
12. *The American Statistician*. 2019;73(Supplement 1).
13. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567:305–307.
14. 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.
15. Tignanelli CJ, Napolitano LM. The fragility index in randomized clinical trials as a means of optimizing patient care. *JAMA Surg*. 2019;154:74–79.
16. Steele SR, Hull TL, Read TE, et al. *The ASCRS Textbook of Colon and Rectal Surgery*. 3rd ed. New York, NY: Springer Nature; 2016.
17. Akl EA, Briel M, You JJ, et al. LOST to follow-up Information in Trials (LOST-IT): a protocol on the potential impact. *Trials*. 2009;10:40.
18. Mazzinari G, Ball L, Serpa Neto A, et al. The fragility of statistically significant findings in randomised controlled anaesthesiology trials: systematic review of the medical literature. *Br J Anaesth*. 2018;120:935–941.
19. Narayan VM, Gandhi S, Chrouser K, Evaniew N, Dahm P. The fragility of statistically significant findings from randomised controlled trials in the urological literature. *BJU Int*. 2018;122:160–166.
20. Reito A, Raittio L, Helminen O. Fragility index, power, strength and robustness of findings in sports medicine and arthroscopic surgery: a secondary analysis of data from a study on use of the Fragility Index in sports surgery. *PeerJ*. 2019;7:e6813.
21. Ridgeon EE, Young PJ, Bellomo R, Mucchetti M, Lembo R, Landoni G. The fragility index in multicenter randomized controlled critical care trials. *Crit Care Med*. 2016;44:1278–1284.
22. Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing: an explanation for new researchers. *Clin Orthop Relat Res*. 2010;468:885–892.
23. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci*. 2014;1:140216.
24. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*. 2014;15:1–12.
25. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. *Nat Methods*. 2015;12:179–185.
26. Moonesinghe R, Khoury MJ, Janssens AC. Most published research findings are false-but a little replication goes a long way. *PLoS Med*. 2007;4:e28.
27. Khan M, Evaniew N, Gichuru M, et al. The fragility of statistically significant findings from randomized trials in sports surgery: a systematic survey. *Am J Sports Med*. 2017;45:2164–2170.
28. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med*. 2012;367:1355–1360.