Exploration of Digestive Diseases

Open Exploration

Open Access Review

Considering the fragility index in reporting trials on diverticular disease

Andrew P. Zbar[1]*, Nir Horesh[2,3]

[1]Department of Neuroscience and Anatomy, Parkville Campus, University of Melbourne, Parkville 3010, Australia
[2]Department of Colorectal Surgery, Cleveland Clinic Florida, Weston, FL 33331, USA
[3]Faculty of Medicine, Tel Aviv University, Tel Aviv 5224213, Israel

***Correspondence:** Andrew P. Zbar, Department of Neuroscience and Anatomy, Parkville Campus, University of Melbourne, Grattan Street, Parkville 3010, Australia. andrew.zbar@unimelb.edu.au; apzbar1355@yahoo.com

Academic Editor: Roberto Cirocchi, University of Perugia, Italy
Received: February 27, 2025 **Accepted:** May 20, 2025 **Published:** June 26, 2025

Cite this article: Zbar AP, Horesh N. Considering the fragility index in reporting trials on diverticular disease. Explor Dig Dis. 2025;4:100578. https://doi.org/10.37349/edd.2025.100578

Abstract

The changing management paradigm of acute complicated diverticulitis and the elective indications for surgery have evolved in the last decade based on reported evidence-based data. Recently, it has been demonstrated that randomized controlled trials (RCTs), the highest trial format in the hierarchy of evidence-based reporting, suffer from a 'crisis of replicability'. The development of a fragility index (FI) quantitatively defines the robustness of an RCT by shifting the number of participants in a trial into a different binary group in an effort to influence reported statistical significance (the lower the FI the greater the study fragility). The only available report on FI in diverticular management showed that in an eclectic range of RCT's comparing intervention and non-intervention, two-thirds of the studies had an FI ≤ 1 where statistical recalculation using Fisher's Exact test rendered one-quarter of previously positive studies non-significant. Comparisons between studies and units are still dependent upon sample sizes and the numbers lost to follow-up even when some of the FI progeny (including a reverse FI, a fragility quotient dividing the FI by the sample size, or other incidence or generalized FI metrics) are utilized in assessment. Future analyses need to define all comparisons rather than cherry-picking examples where a *p* value approaches significance. Despite the fact that no FI value defines the strength of a RCT, its use attempts to link the reported *p* value with the sample size and the statistical power of the study. Positive findings in diverticular trials are then considered not so much definitive as rather provocateurs encouraging further similarly designed studies in different environments. Minimizing patient loss in treatment arms and reporting the reasons for drop-out, strictly adhering to randomization, consistent blinding, and group allocation concealment can all improve the logistical running of an RCT initially designed to evaluate some potentially important new treatment.

Keywords

Fragility index, *p* value, statistical significance, diverticular disease, randomized controlled trials

Introduction

Coloproctologists increasingly rely on evidence-based approaches to address management dilemmas. Over the last decades, the surgical treatment of acute complicated diverticulitis and elective surgery for recurrent inflammation complicated by obstruction, fistulation, or paracolic sepsis has undergone a radical transformation. The Hartmann's procedure (resection without primary anastomosis) has been considered the gold standard against which other operations and interventions are compared for the advanced emergency presentations of complicated diverticulitis. The last 2 decades have seen the advancement of resection with primary anastomosis performed either open or via minimally invasive (MIS) techniques and the use of a range of non-resectional procedures. Given the morbidity and mortality of the Hartmann's approach (however performed), the decision-making for protective stomas and the problems encountered with stoma reversal as well as the fact that many patients never end up getting their stoma closed, a range of alternatives have been sought particularly in those patients presenting with purulent peritonitis [1]. In an environment where MIS confers a quicker overall recovery with less pain and reduced length of hospital stay this latter Hinchey grade III case has provoked the greatest level of controversy. Here there has been a move away from the surgical standard which has seen the rise of selective percutaneous drainage of localized collections, laparoscopic lavage, reduced antibiotic use, and a more observational stance compared with routine elective resection. This shift has in part been driven both by a better understanding of the natural history of acute inflammation when managed conservatively and by advances in MIS technology.

Recent data highlighted the morbidity associated with blanket stoma approaches, including impacts on quality of life (QoL), difficulties in stoma reversal (especially in older or comorbid patients), and subsequent functional impairment. High-resolution imaging, percutaneous technology, and MIS procedures including lavage and drainage, enabled a more tailored management of complex cases [2], leading to an overall downward trend in emergent surgery at the expense of an increased rate of elective colectomy [3]. For instance, patients initially managed with percutaneous drainage who subsequently present with recurrent diverticulitis tend to have a higher Hinchey grade and are twice as likely to require emergent surgery [4, 5].

In contrast, the controversy remains for Hinchey Grade III cases despite the conduct of a range of randomized controlled clinical trials examining alternatives to Hartmann's procedure for perforative diverticulitis (Grade III and IV Hinchey). In some of these trials, new questions concerning management were raised. In the Laparoscopic Peritoneal Lavage of Resection for Generalized Peritonitis for Perforated Diverticulitis (LADIES) trial, patients were managed with a protective stoma (loop ileostomy) at the discretion of the surgeon with more patients in the primary anastomosis group who were ultimately stoma free [6]. This study also showed a distinct disadvantage in terms of mortality, other serious adverse events, and the need for surgical reintervention in those patients managed with laparoscopic lavage when compared with those undergoing a sigmoidectomy performed either as a Hartmann's procedure or with a primary anastomosis. Similar conclusions on the management of Hinchey Grade III were reached by the results of both the DILALA trial comparing lavage with resection [7] although there was nearly a 50% reduction in the need for more operations in the laparoscopic group. This study reported, however, a small number of patients in each group and failed to show any reduction in either the rates of hospital readmission or mortality. Likewise, the SCANDIV trial comparing laparoscopic peritoneal lavage with primary resection in both Hinchey Grade III and IV cases [8] showed no difference in morbidity or mortality but there was a higher surgical reintervention rate in those managed with laparoscopy. Studies of this nature (and with expanded categories) need to be replicated. These studies gave a glimpse into the management of complex cases, providing the practising surgeon with a wider toolbox for these patients, which have been implemented into the ASCRS, SAGES, and ESCP published management recommendations [9–11] in recent years. The guidelines of these agencies have been formulated by committees under the approval of Boards of Governors with individual recommendations made periodically after systematic reviews and their assessment by multidisciplinary teams. Increasing emphasis has been placed on the post-surgical QoL of these patients where the impact of bowel function in particular has been better categorized in a way that can potentially influence surgical decision-making. In this respect, there has been a shift towards more of a case-based approach that takes into account the severity of an acute episode rather than offering a blanket procedure for younger patients who present with their second or third bout of acute diverticulitis [12].

In the elective setting, surgical decision-making is more tailored. Given that the severity of acute attacks of recurrent diverticulitis has a propensity to abate [13, 14] the stated aim of future surgery is to improve the health-related quality of life (HRQoL). In this regard, both the DIRECT [15, 16] and the LASER [17] compared conservative and surgical management (largely laparoscopic colectomy) in patients with recurrent acute diverticulitis showing an improvement in HRQoL in the operative groups. In the DIRECT trial, one-quarter of those managed without operation initially ended up opting for surgery by 6 months and almost half by 5 years. Although the HRQoL advantage was maintained in the operative group, this approach should be balanced against the overall complication rate (including anastomotic leakage) in those undergoing surgery. The findings in the LASER trial were similar however, this trial was prematurely terminated and reported comparatively small numbers in each group. While current evidence supports both emergent and elective surgical approaches under specific circumstances, the decision-making process still remains individualized.

How do we gain evidence and understand statistical significance: the fragility index in diverticular practice?

Given the infrequent nature of admissions of patients with acute complicated diverticulitis, most management data are retrospective involving small patient series. However, certain surgical environments enable inter-institutional cooperation for data pooling and the development of national guidelines [14, 18–20]. The application of individual Societal recommendations is based upon the pyramid hierarchy of study types reported with greatest attention paid to randomized controlled trials (RCTs) and where there is an inverse relationship between the quality of evidence assigned and the risk of inherent bias [21]. The pyramid of evidence places RCTs at the highest level because of their strengths in randomization, blinding, and minimization of bias and confounding.

Our views of diverticular management have largely been swayed by our impressions of the significance of studies (as represented by reported *p* values) in trials where particular interventions would then lead to a dichotomous outcome (e.g., infection or no infection, survival or death, the need for surgical reintervention and so on). It is evident, however, that reported *p* values are in fact only rough quantitative guides to the strength of available evidence that a null hypothesis (namely that no differences exist between treated and untreated groups) would be rejected. In the diverticular disease setting, particularly if the sample sizes of a study are relatively small, such reliance on a blanket *p* value to guide management is of limited value.

Historically, it may come as no surprise that the original designer of the *p* value concept, the geneticist RA Fisher (1890–1962) was unable to readily explain the variable, setting the somewhat arbitrary cut-off for *p* at < 0.05 and insisting that the idea of its 'significance' was only an invitation to perform further clinical experimentation for further validation of the finding rather than as definitive proof [22, 23]. Although *p* values should still be given (even when approaching or approximating 0.05) they should also be accompanied by confidence intervals (CI) and an appreciation that there are potentially better measures of the strength of available evidence [24, 25]. Such misconceptions concerning the *p* value and concept persist. It is of relevance to appreciate that the assessment of scientific data based upon *p* values relies on an accurate discernment of false positives, with the risk that a publication so reporting may on occasion not necessarily advance the available evidence base. If inaccuracies in a trial were to exist, differences might be far too small to detect where it should be noted that the *p* value is not actually a signifier of the magnitude of an effect. Equally, the importance of a significant *p* value may be limited for clinically irrelevant endpoints. Studies may be so heterogenous in design and recruitment that even when they present statistical significance, they may have limited, if any relevant clinical significance.

To address these limitations—and to help resolve the "crisis of replicability" in RCTs [26], Walsh and colleagues [27] devised the concept of the 'fragility index' (FI). This measure was designed to quantitatively describe the relative robustness of an RCT, defining it as the minimum number of study participants that would be required to change the significance of a reported outcome. Methodologically, the FI quantifies the minimum number of participant outcome changes (i.e., from event to non-event or vice versa) required to change a statistically significant result ($p < 0.05$) to a non-significant one using a two-sided Fisher's Exact test. A low FI indicates that a trial's outcome is highly susceptible to small changes, while a high FI suggests robustness. In those cases where no change in patient numbers is required to move from significance to insignificance, the FI = 0, and in many such trials (around 20%), which often consist of a small number of participants the use of a Fisher's Exact test over a chi-squared analysis may shift the significance of the trial [28]. In such a case where there is chi-square and Fisher's discordance since the chi-square analysis assumes larger sample sizes, trials can be rendered 'non-significant' without the need for changing the event/non-event ratios. The FI process involved therefore represents a form of sensitivity analysis delineating how many patients would be required to produce a different outcome.

Such FI analyses have been used extensively for systematic reviews of trauma and orthopaedic surgery assessing the outcomes of hip, shoulder, and knee arthroplasty [29–33] but also in urologic and gynaecological surgery [34, 35], renal transplantation [36], oncologic surgery of the head and neck [37] and in plastic surgery [38]. This approach has, however, achieved less success in the management of traumatic brain injury [39] and in specialized paediatric surgery [40].

It is accepted at the present time that no specific value of the FI is considered to confer robustness for any given RCT. In this regard, however, studies purporting to show significant differences but where the number of patients lost to follow-up exceeds the FI value must be viewed with suspicion. Application of the FI may be made in accordance with the structural elements of an RCT designed to determine causality where two identical groups have been fashioned and where one is subjected to the intervention under study. Observations pertaining to outcomes are then designed to be made between these experimental and control groups emphasizing the binary nature of results that lend themselves to the FI methodology.

In such trials, it must be conceded that the outcomes of unknown cases could seriously alter the results. In this regard about two-thirds of RCTs have an FI ≤ the total number of cases lost to follow-up [41], severely limiting study comparisons. Further, the assumptions we make concerning the p value of studies and their relationship to the FI can cause confusion with larger p values (i.e., those close to 0.05) which are considered fragile rather incorrectly, however, providing the impression that p is a measurement of the magnitude of any effect. This is an element over which p has no dominion [42, 43].

The mechanics of FI permit a conversion from significant to insignificant with fewer events. In an example, which could apply to diverticular management as well, if a treatment arm assessed events as adverse (e.g., infection) but as favourable in the control arm (e.g., hospital discharge) then a fragile study where the FI = 1 would just require one patient switch to change its significance and hence clinical importance. This issue becomes complicated if FI values are compared between studies because of the close connection between the FI and the sample size; an effect that is in part obviated by the use of the fragility quotient (FQ) where [44]:

$$FQ = \frac{FI}{Sample\ size} \times 100$$

The FQ operates to compare study FI values where sample sizes differ by assessing proportions rather than cardinal numbers, and by altering the magnitude of fragility. However, it does not suggest that sample sizes must always be increased in some trials where a low FI can still represent a well-enough powered study. This issue is complicated since the expansion of a sample size beyond the projected power can create ethical issues where more controls can be placed at unnecessary risk (if there is a positive benefit from intervention) as well as increase the cost of an unnecessary part of a trial [45]. It should be accepted that smaller studies, no matter how well designed, tend to result in larger effects than larger trials and may therefore present particular characteristics that will make the findings less replicable [46]. There is in fact, therefore nothing to specifically link FI to the quality management of an individual RCT.

Within this framework, sensitivity analyses are often reported in RCTs in order to examine changes in outcome with different assumptions about trial structure. This commonly compares an intention-to-treat analysis (where patients are assessed on the random allocation for a proposed treatment) as opposed to a per-protocol analysis (where patients are classified according to the treatment they actually received). The FI can be rightly criticized since it draws attention to one type of sensitivity analysis potentially at the expense of other logistic trial concerns and inherent flaws in the allocated populations at the commencement of any trial [47, 48]. This might suggest a misallocation of fragility to trials where it may not exist since in many studies power is not applied for secondary outcomes. An example of this type of FI misconception occurred in the recent World Hip Trauma Evaluation (WHITE5) trial comparing cemented to uncemented hemiarthroplasty for intracapsular hip fractures showing marked differences in HRQoL that favoured cemented prostheses but where there was no difference in perioperative mortality [49]. Given the low mortality in this trial in either group, power studies to assess this secondary endpoint would have required a study recruiting 4 times the number of patients actually enrolled to likely show any significant difference in death rates. This type of criticism of an RCT based upon FI reliance is as much a criticism of the null hypothesis method as anything else. Where events are uncommon (an anastomotic leak, an intra-abdominal collection, a fistula, etc.) a single change by one patient can have a profound effect on the impression of trial validity.

In the first assessment of the FI in colorectal surgery reported by Nelms et al. [50] using a series of MeSH terms for colorectal and anorectal surgery and the ASCRS Textbook of Colon and Rectal Surgery as a guideline, prospective RCTs between 2016–2018 were analyzed. The authors included trials with dichotomous outcomes in paradigmatic areas of coloproctology (perioperative/endoscopy, anorectal disease, malignant disease, benign disease, pelvic floor disorders, etc.) and exclusion of trials with non-inferiority design. Using the Walsh methodology of addition and subtraction of events to non-events and recalculating the *p* value with Fisher's testing until it reached non-significance, the authors showed that the majority of colorectal studies across a broad range of intra-specialty disciplines had a low FI (median FI = 3 derived from 90 separate trials). Furthermore, in over half the studies (57%) cases lost to follow-up exceeded the FI derived, throwing much of coloproctological literature into the realm of doubt concerning both its veracity and its replicability. Importantly, the probability of a false finding by any RCT (the false discovery rate or FDR) is not measured by the *p* value as such.

Currently, the only available report of the FI in diverticular management is a recent study by McKechnie et al. [51] examining RCTs of both surgical and medical management protocols conducted between 2010–2022, with parallel superiority formats that compared an intervention arm with either non-intervention controls or with placebo-managed groups. Such studies included comparisons of primary anastomosis vs. a Hartmann's procedure in Hinchey grade III and IV peritonitis, laparoscopic lavage vs. resection in acute diverticular peritonitis, MIS elective colectomy vs. observation for a first attack of acute complicated diverticulitis (and for cases of first recurrence of diverticulitis) and open vs. laparoscopic colectomy for symptomatic diverticulitis. In their report of this eclectic mix of diverticular papers, there were 15 RCTs incorporating 1,434 patients managed with both complicated and uncomplicated diverticular disease. Two-thirds of trials had an FI = 0 or =1 where the recalculation with Fisher's test rendered one-quarter of the studies (some of which used Kaplan-Meier methodology and log-rank regression analysis), completely insignificant. Moreover, in half the studies the number of patients lost to follow-up exceeded the modified FI value. There was no effect of the year of publication, the impact factor of the published journal, or any calculated risk of bias. The FI values in this study remained low in recently published analyses, those where the sample size exceeded 100, those where the drop-out or loss-to-follow-up rate was < 5%, those where there were > 30 events, and studies which were industry-run. This first and only fragility study in diverticular management RCTs means that most would only need one patient to switch group outcomes to lose significance and clearly proposes that much of the current diverticular literature concerning management is fragile. The only recommendations made by the group to improve the robustness of the trials were the need to increase sample sizes and to ensure the retention of trial participants. In the context of emergency surgery for perforated diverticulitis, particularly for Hinchey III disease, this might be another testimony to the slow adoption of novel surgical techniques such as primary resection with anastomosis or laparoscopic lavage compared with the Hartmann's procedure which is still by far the most common procedure for this condition.

Limitations of the FI: implications in coloproctology

While the FI provides a useful sensitivity analysis, several limitations must be recognized. Alternative indices include the reverse FI (rFI), the FQ, the incidence FI (FIq), and the generalized FI (GFIq) each of which has been developed to overcome some shortcomings of the original FI [52]. The rFI modified by Khan et al. [53] calculates the minimum number of participants needed to have a different outcome for the endpoint to change from insignificant to significant amongst the group with the least number of events and modifying the events to non-events in an algorithm that reverses that used by Walsh [27]. The FQ, previously discussed in the chapter [44], has been employed to obviate the effect of sample size where a low FQ reflects a robust trial. This provides a measure of study vulnerability and a means of comparing studies because of the standardized scale of the FQ, where groups undergoing the same intervention have widely differing sample sizes [54]. Tables 1 and 2 show the metrics and their basic definitions as well as examples of the effects and importance of small or large sample sizes on the different indices used. Table 3 shows the calculation method of the indices. The FIq permits the analysis of only sufficiently likely outcomes for specialized trials using a reverse FI methodology, although it is accepted that the line between likely and less likely outcomes imposes its own limitations, particularly if an outcome has probability one meaning that reversal of significance becomes impossible [55]. A GFIq is an attempt to generalize the FI beyond a 2 × 2 contingency table and to extend its use beyond dichotomous outcomes and into situations where the sample size is not fixed, or where there is no fixation of either the allocated management groups or outcomes. These attempts to extend the FI surrogate instrument beyond dichotomous outcomes to continuous scales (such as visual analogue scales or HRQoL) or to events that are time-dependent (such as survival analysis) have not been applied yet to diverticular management [56, 57].

Table 1. Basic definitions of fragility indices

Metric index	Basic definition
Fragility index (FI)	The minimum number of event reversals (from non-event to event) required to change a statistically significant result to non-significant ($p > 0.05$).
Reverse fragility index (rFI) [52]	The minimum number of event reversals needed to turn a non-significant result ($p \geq 0.05$) into a significant one ($p < 0.05$).
Fragility quotient (FQ) [44]	The FI is divided by the total sample size.
Incidence fragility index (FIq) [52]	The FI normalized by the control group event incidence rate, making fragility comparable across trials with different event rates.
Generalized fragility index (GFIq) [52, 57]	A composite fragility metric incorporating both increases and decreases in event occurrences, allowing a broader assessment of study robustness.

Table 2. Effects and importance of smaller and larger sample sizes on different fragility indices

Index	Small sample effect (example)	Large sample effect	Explanation and importance
Fragility index (FI)	A small RCT with 50 patients shows that Drug A is significantly better ($p = 0.04$). Changing 2 non-events to events makes $p > 0.05$ where FI = 2.	A large RCT with 1,000 patients shows that drug B is significantly better ($p = 0.04$). Changing 15 non-events to events makes $p > 0.05$ where FI = 15.	In the small trial (FI = 2), a tiny change in outcomes alters significance, meaning that the results are fragile. In the larger trial (FI = 15), the results are more robust since more changes are needed in order to affect significance.
Reverse fragility index (rFI)	A trial with 60 patients shows no significant difference ($p = 0.07$). Adding 3 new events to the treatment group makes $p < 0.05$ where rFI = 3.	A large trial with 1,200 patients shows no significant difference ($p = 0.07$). Adding 20 new events to the treatment group makes $p < 0.05$ where rFI = 20.	The small study (rFI = 3) is close to significance, requiring just 3 event reversals. The large study (rFI = 20) needs a much bigger shift, showing greater confidence that no real effect exists.
Fragility quotient (FQ)	A study with FI = 3 and sample size = 200 results in an FQ = 0.015 (1.5%; 3/200).	A study with FI = 20 and sample size = 2,000 results in an FQ = 0.01 (1%; 20/2,000).	In the small study only 1.5% of the participants changing their outcome can alter significance. In the larger study an

Table 2. Effects and importance of smaller and larger sample sizes on different fragility indices (*continued*)

Index	Small sample effect (example)	Large sample effect	Explanation and importance
			FQ = 1% is consistent with greater stability.
Incidence fragility index (FIq)	A trial with an FI = 5 and a control group event rate = 10% (15/150 patients), results in an FIq = 0.33 (33%; 5/15).	A larger trial with an FI = 30 and a control group event rate = 12% (360/3,000 patients), results in an FIq = 0.083 (8.3%; 30/360).	In the small study a significant portion of the control group must switch outcomes to change the significance. In the larger trial the result (8.3%) is more reliable across different event rates.
Generalized fragility index (GFIq)	A study with an FI = 4, an rFI = 3, a sample size = 250, and a control group event rate = 12% (30/250) uses both the FI and rFI to generate a GFIq.	A study with an FI = 18, an rFI = 15, a sample size = 5,000, and a control group event rate = 10% (500/5,000) results in a more stable GFIq.	The small study has a lower GFIq, meaning both FI and rFI are low, making the results less stable. The large study' has a higher GFIq demonstrating more resilience to statistical shifts.

Table 3. Summary of index calculation

Metric index	Calculation
Fragility index (FI)	With *a* events in Group 1 (total patients = *a* + *b*) and *c* events in Group 2 (total patients = *c* + *d*). Where hypothetically: Event Non-event **Group 1** *a* *b* **Group 2** *c* *d* When FI is positive patients are moved from a non-event to an event in Group 1. When FI is negative patients are moved from an event to a non-event in Group 1. Changing outcomes preserves the number of patients in each group. FI is based on the number of changes required to render the *p* value ≥ 0.05. Where the hypothetical data structure is as: Event Non-event **Group 1** *a + f1* *b – f1* **Group 2** *c + f2* *d – f2*
Reverse fragility index (rFI)	Choose the group that has the fewest number of events and then change the events to non-events to render statistical significance. The total number of outcome changes = the rFI.
Fragility quotient (FQ)	The FQ is a relative measurement calculated by division as: $\frac{Absolute\ FI}{Total\ Sample\ FI}$
Incidence fragility index (FIq)	FIq is such that any probability $q \in [1,2]$ = the minimum number of changes in patient outcome with a probability of at least q in order to reverse statistical significance. This permits only sufficiently likely modifications according to the likelihood threshold value of *q*. The minimum modifications are then: $\min_{fi,f2} \in Z\,\vert f1\vert + \vert f2\vert$ (based on hypothetical tables above)
Generalized fragility index (GFIq)	To generalize the data set: Where there are *n* samples (and Y_1–Y_n observations), a significant (< 0.05) *t*-test becomes: $\lim_{Y_1 \to \infty} \sqrt{n}\left(\overline{Y} - 0\right) \Big/ S = n^{-1} \Big/ \sqrt{\left(1 - n^{-1}\right) / \left(n\left(n - 1\right)\right)} = 1$ For all modifications with Y as the sample mean and *S* the SD. In each case, the one sample *t*-test =1 (i.e., it is not significant) at the *a* = 0.05 level for any sample size [52].

Those trials where a *p* value is set at 5% or less have a measurable FDR (what most refer to as the error rate) which is calculated as [58]:

$$\frac{No.\ of\ false\ positives}{No.\ true\ positives + No.false\ positives}$$

There is an inverse relationship between this FDR and the positive predictive value (PPV) since the FDR = (1 – PPV).

Concerning this point, David Colquhoun [59] has renamed this FDR the 'false positive risk', opining in 1971 that "the function of significance tests is to prevent you from making a fool of yourself, and not to make unpublishable results publishable". Low FI values could be offset by increasing the sample size beyond the requisite *p* value to account for unforeseen circumstances, creating in effect a fragility buffer which enhances the likelihood of reproducibility [60, 61]. Even well-designed RCTs can experience an unexpectedly high drop-out rate resulting in a high baseline risk of a falsely large effect since the FI obtained competes with the fidelity of patient blinding and the concealment of group allocation along with the level of participant retention [27, 62, 63]. In diverticular disease, for example, many emergency cases are performed outside routine workday hours which might impose a difficulty in recruiting patients due to a lack of resources. Attempts to resolve the fragility of *p* have been made by Quatto et al. [64] in introducing a strength index (SI) or a 'strength concept' where changes from an event to a non-event (and vice versa) would have an equivalent probability of occurring in this setting so that the greater the SI the greater the trial reliability (and hence the reliability of *p*), Clearly, however, the arbitrary nature of the *p* threshold defines the false positive and negative rates expected.

Those smaller trials where sample size is an issue could formulate a likelihood ratio (LR) or a likelihood FI (LFI) which could be described as the minimum number of conversions necessary in a small group that would permit the LR to reach or exceed the designated LR. In broad terms for future analyses of results, our approach needs to be based on Bayesian theorems rather than on current probability analyses. Although this is somewhat of an old argument concerning statistical significance of results, it has only recently been redressed in light of the appearance of the FI in its various guises. The Bayesian approach is an alternate statistical paradigm where given the observed results of a trial, a prediction could be made as to whether the hypothesis itself is likely to be true or false. In a sense, this ancient statistical argument then would favour the inverse of the traditional null hypothesis testing (the so-called frequentist view of probability) that coloproctologists are so used to reading. However, it remains to be seen whether this approach will create less skepticism than the customary *p* value [65].

Conclusions

Trials addressing specific questions in diverticular disease often face challenges in reproducibility and generalizability. Therefore, it is no surprise that trials concerning very specific questions as they pertain to diverticular disease can often be viewed with skepticism and may not effectively guide management because they are not in essence replicable. Factors such as incomplete randomization, inadequate blinding, and selective reporting can all contribute to fragile findings. These ideal worlds in publication given the pressures to publish (even when there are no nefarious elements introduced into the process) do not exist. At the very least, the likelihood that a difference in effect has been demonstrated as wrong occurs anywhere from between one-third to one-half of cases [58, 66, 67]. It behooves us to ensure that future studies are drawn from wider multi-institutional participation with the highest possible standards of randomization, increasing sample sizes where possible to enhance power and to try (at least) to diminish the false discovery rate. Despite the best will in the world, it can be appreciated that the pressure on selective reporting, can skew publications towards positive results and must affect our concepts of management [68].

When we use an FI we are reapplying some of the concepts we presume when using *p* values, except that in the case of *p* we examine the way that different outcomes statistically impact these distributions. With FI on the other hand we are semi-quantitatively examining the effect of different outcomes [52]. One statistical trick to enhance the value of a publication will be to reduce the arbitrary *p* value to a much more rigorous standard (e.g., $p < 0.001$). Another will be the avoidance of cherry-picking of attributes in a study from a range of variables that have been examined but where the *p* values approach significance. Here it is wiser to report everything in supplemental appendices along with the routine publishing of confidence intervals [69]. As suggested by Wasserstein in a position paper by the American Statistical Association, the greater the transparency of the reported data, the greater a healthy suspicion will be engendered in the data assessment [70].

The collaborative approach within countries of specific diverticular problems and their management will also make the future brighter for study inclusion and consideration within meta-analyses [63, 71]. In this regard, a group of Israeli colorectal surgeons began in 2009 to pool results from 6 major University-affiliated teaching hospitals initially to document the incidence of emergency diverticulitis surgery and the morbidity and mortality of Hartmann's reversal [20, 72, 73]. This group has expanded the organization to trial MIS approaches and to follow the optimal timing of stoma closure [19, 74, 75]. It is accepted that the FI is far from perfect and for the moment we are stuck with it. At worst it may be no better than a *p* value if we do not understand its meaning. At best, it is an intuitive modality designed to decipher how small some changes might be needed for the entire outcome of a trial to be adversely affected. The FI links the reported *p* value to the sample size and to the statistical power of a study, but only loosely. It does, however, reinforce the need to try the trial over again in one's home territory as it were. But we know that the FI too does not have a validated threshold that would define a study as robust. Indeed, as others have asked, how robust is robust? And how fragile is fragile?

If we are to use the FI we must understand its limitations and not simply transfer the abuse of *p* to a misapplication of FI. Large FI values do not always mean that results are decisive and small FI values do not always imply that results are inconsequential [76] or even that an RCT has been poorly conducted. For example, demonstrating an FI of 0 or 1 raises serious concerns that such a study needs to be repeated with larger numbers before influencing therapy even when other parameters like the 95% CI reported are relatively tight. But this issue can be complicated at times. It can be more useful to note the low FI of a particular study where the lower boundary of a CI is close to the null value regardless of a high sample size. In this circumstance, one can be more confident of a treatment effect notwithstanding some small number of participants when the CI is well away from the null value, and when the corroborative FI is high, the measured event is prevalent and the statistical power is appropriate for that event. Simulations of such trials can be useful as has been done in cardiovascular trials where Yusuf et al. [77, 78] have suggested that a minimum of 650 events would be needed to be confident about a true effect. Although its value is unproven, the FI is easily understood. The strength of an RCT is reliant upon its sample sizes, the magnitude of any effect, and its confidence intervals, rather than upon any single integer that may well have been derived from a hypothetical trial that has not occurred. In practical terms, an understanding of an FI as it pertains to a particular field of surgery or a discrete intervention will likely assist in the design of future RCTs as well as in discerning the value of Societal recommendations of management. For the jobbing surgeon, it is only in this sense that the FI adds a more questioning eye to the meaning of the *p*. Since there is a close relationship between the FI and the *p* value it is not surprising that those studies where *p* lies just below 0.05 are fragile. For these studies we need not necessarily apply an FI at all. But if we read studies which are not threshold significant, they need to be repeated perhaps with larger samples and with more rigour. An RCT that barely meets the power of the study with a borderline *p* value and a low FI are all indications taken in combination that permit us to adjudicate a trial more shrewdly and to reserve judgment on its value. All of this is a package of decision-making on the value of data. Nothing is the final integer of arbitration. Along with the *p* value and the FI one should examine the design of the RCT, the response rates, the sample sizes, the confidence intervals, the patient numbers, and any reasons for their dropout [79].

Future application of more quantitative approaches towards the FI could also prove more valuable. Grimes has recently described what he calls the 'ellipse of insignificance' using the geometry of chi-square testing to grade error and to directly describe the imperfections of sensitivity and specificity analyses [80]. Graphs are created as vectors of a simple right-angled triangle with resolution of these lines in experimental and control directions to provide a d_{min} which is a theoretical minimum number of miscodings needed to alter significance. The more manipulative interventions an outcome can withstand (both in the treatment and the control arms of a trial, without an effective change in the results), the greater the strength of the trial. Such an approach is an advantage over conventional FI measurement providing an analysis metric for very large data sets which can consider changes to both arms of a trial concurrently. Of course, even this new metric cannot comment on patients who are lost to follow-up during the trial or where extreme datapoints have been removed by researchers [54, 62, 81].

It is further appreciated that the very design of some RCT's is 'fragile' in the sense that new treatments may be limited in their exposure to participants in an effort to avoid harm. Whether by design or not, the initial studies of laparoscopic lavage in acute Hinchey Grade III peritonitis were small and needed to balance unknown safety outcomes with the size of the sample. In any RCT, many factors contribute to patient loss, but the more of these we can control, the greater the validity of the trial. The more dropouts we discover in an RCT, the greater the chance that some sort of effect for a given treatment might be erroneously found. In the future, RCTs in colonic diverticular disease could include loss of follow-up information that is stratified by the treatment arm [60] with a reverse FI utilized in those trials initially showing non-significant findings. The point here is that the logistics of the trials we examine really do matter. Most notably, the strict adherence to randomization, the blinding and concealment of group allocation, the efforts made to retain participants, and the like. As mundane as they may seem, nevertheless all these operational and governance details are important. As diverticular disease management has undergone such radical philosophical shifts in the last 2 decades or so, we can only consider trials as on par currently with best-recommended practice and nothing more than commensurate with published Societal principles and expert panel guidelines. In an age where the public trust in science has been somewhat shaken, we can use new parameters of comparison even if on occasion in this fluid environment we end up finding that today's best concept is tomorrow's flat earth doctrine.

Abbreviations

CI: confidence interval

FDR: false discovery rate

FI: fragility index

FIq: incidence fragility index

FQ: fragility quotient

GFIq: generalized fragility index

HRQoL: health-related quality of life

LR: likelihood ratio

MIS: minimally invasive

PPV: positive predictive value

QoL: quality of life

RCTs: randomized controlled trials

rFI: reverse fragility index

Declarations

Author contributions

APZ: Conceptualization, Writing—original draft, Writing—review & editing. NH: Conceptualization; Writing—review & editing. Both authors read and approved this revised submitted version.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

Not applicable.

Funding

Not applicable.

Copyright

Publisher's note

Open Exploration maintains a neutral stance on jurisdictional claims in published institutional affiliations and maps. All opinions expressed in this article are the personal views of the author(s) and do not represent the stance of the editorial team or the publisher.

References

1. Huo B, Ortenzi M, Anteby R, Tryliskyy Y, Carrano FM, Seitidis G, et al. Surgical management of complicated diverticulitis: systematic review and individual patient data network meta-analysis: An EAES/ESCP collaborative project. Surg Endosc. 2025;39:699–715. [DOI] [PubMed]
2. Gregersen R, Mortensen LQ, Burcharth J, Pommergaard HC, Rosenberg J. Treatment of patients with acute colonic diverticulitis complicated by abscess formation: A systematic review. Int J Surg. 2016; 35:201–8. [DOI] [PubMed]
3. Simianu VV, Strate LL, Billingham RP, Fichera A, Steele SR, Thirlby RC, et al. The Impact of Elective Colon Resection on Rates of Emergency Surgery for Diverticulitis. Ann Surg. 2016;263:123–9. [DOI] [PubMed] [PMC]
4. Devaraj B, Liu W, Tatum J, Cologne K, Kaiser AM. Medically Treated Diverticular Abscess Associated With High Risk of Recurrence and Disease Complications. Dis Colon Rectum. 2016;59:208–15. [DOI] [PubMed]
5. Jalouta T, Jrebi N, Luchtefeld M, Ogilvie JW Jr. Diverticulitis recurrence after percutaneous abscess drainage. Int J Colorectal Dis. 2017;32:1367–73. [DOI] [PubMed]
6. Vennix S, Musters GD, Mulder IM, Swank HA, Consten EC, Belgers EH, et al. Laparoscopic peritoneal lavage or sigmoidectomy for perforated diverticulitis with purulent peritonitis: a multicentre, parallel-group, randomised, open-label trial. Lancet. 2015;386:1269–77. [DOI] [PubMed]
7. Angenete E, Thornell A, Burcharth J, Pommergaard H, Skullman S, Bisgaard T, et al. Laparoscopic Lavage Is Feasible and Safe for the Treatment of Perforated Diverticulitis With Purulent Peritonitis: The First Results From the Randomized Controlled Trial DILALA. Ann Surg. 2016;263:117–22. [DOI] [PubMed] [PMC]
8. Schultz JK, Yaqub S, Wallon C, Blecic L, Forsmo HM, Folkesson J, et al.; SCANDIV Study Group. Laparoscopic Lavage vs Primary Resection for Acute Perforated Diverticulitis: The SCANDIV Randomized Clinical Trial. JAMA. 2015;314:1364–75. [DOI] [PubMed]

9. Hall J, Hardiman K, Lee S, Lightner A, Stocchi L, Paquette IM, et al.; Prepared on behalf of the Clinical Practice Guidelines Committee of the American Society of Colon and Rectal Surgeons. The American Society of Colon and Rectal Surgeons Clinical Practice Guidelines for the Treatment of Left-Sided Colonic Diverticulitis. Dis Colon Rectum. 2020;63:728–47. [DOI] [PubMed]
10. Francis NK, Sylla P, Abou-Khalil M, Arolfo S, Berler D, Curtis NJ, et al. EAES and SAGES 2018 consensus conference on acute diverticulitis management: evidence-based recommendations for clinical practice. Surg Endosc. 2019;33:2726–41. [DOI] [PubMed] [PMC]
11. Schultz JK, Azhar N, Binda GA, Barbara G, Biondo S, Boermeester MA, et al. European Society of Coloproctology: guidelines for the management of diverticular disease of the colon. Colorectal Dis. 2020;22:5–28. [DOI] [PubMed]
12. Lin M, Raman SR. Evaluation of Quality of Life and Surgical Outcomes for Treatment of Diverticular Disease. Clin Colon Rectal Surg. 2018;31:251–7. [DOI] [PubMed] [PMC]
13. Buchs NC, Konrad-Mugnier B, Jannot A, Poletti P, Ambrosetti P, Gervaz P. Assessment of recurrence and complications following uncomplicated diverticulitis. Br J Surg. 2013;100:976–9: discussion 979. [DOI] [PubMed]
14. Alexandersson BT, Stefánsson T. Incidence and recurrence rate of sigmoid diverticulitis in patients requiring admission to hospital in Iceland from 1985 to 2014: nationwide population-based register study. BJS Open. 2020;4:1217–26. [DOI] [PubMed] [PMC]
15. van de Wall BJM, Stam MAW, Draaisma WA, Stellato R, Bemelman WA, Boermeester MA, et al.; DIRECT trial collaborators. Surgery versus conservative management for recurrent and ongoing left-sided diverticulitis (DIRECT trial): an open-label, multicentre, randomised controlled trial. Lancet Gastroenterol Hepatol. 2017;2:13–22. [DOI] [PubMed]
16. Bolkenstein HE, Consten ECJ, van der Palen J, van de Wall BJM, Broeders IAMJ, Bemelman WA, et al.; Dutch Diverticular Disease (3D) Collaborative Study Group. Long-term Outcome of Surgery Versus Conservative Management for Recurrent and Ongoing Complaints After an Episode of Diverticulitis: 5-year Follow-up Results of a Multicenter Randomized Controlled Trial (DIRECT-Trial). Ann Surg. 2019; 269:612–20. [DOI] [PubMed]
17. Santos A, Mentula P, Pinta T, Ismail S, Rautio T, Juusela R, et al. Comparing Laparoscopic Elective Sigmoid Resection With Conservative Treatment in Improving Quality of Life of Patients With Diverticulitis: The Laparoscopic Elective Sigmoid Resection Following Diverticulitis (LASER) Randomized Clinical Trial. JAMA Surg. 2021;156:129–36. [DOI] [PubMed] [PMC]
18. van Dijk ST, Daniels L, Ünlü Ç, de Korte N, van Dieren S, Stockmann HB, et al.; Dutch Diverticular Disease (3D) Collaborative Study Group. Long-term effects of omitting antibiotics in uncomplicated acute diverticulitis. Am J Gastroenterol. 2018;113:1045–52. [DOI] [PubMed]
19. Horesh N, Zbar AP, Nevler A, Haim N, Gutman M, Zmora O. Early experience with laparoscopic lavage in acute complicated diverticulitis. Dig Surg. 2015;32:108–11. [DOI] [PubMed]
20. Russell B, Zager Y, Mullin G, Cohen M, Dan A, Nevler A, et al. Naples Prognostic Score to Predict Postoperative Complications After Colectomy for Diverticulitis. Am Surg. 2023;89:1598–604. [DOI] [PubMed]
21. Brighton B, Bhandari M, Tornetta P 3rd, Felson DT. Hierarchy of evidence: from case reports to randomized controlled trials. Clin Orthop Relat Res. 2003;413:19–24. [DOI] [PubMed]
22. Fisher RA: Statistical Methods for Research Workers. Oxford: Oxford University Press; 1958.
23. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45:135–40. [DOI] [PubMed]
24. Colquhuon D. The false positive risk: a proposal concerning what to do about the *p* value. Am Stat. 2019;73:192–201. [DOI]
25. Why p-values can't tell you what you need to know and what to do about it [Internet]. YouTube; c2018 [cited 2025 Apr 14]. Available from: https://www.youtube.com/watch?v=agC-SG5-Qyk
26. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533:452–4. [DOI] [PubMed]

27. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clin Epidemiol. 2014;67:622–8. [DOI] [PubMed]

28. Tignanelli CJ, Napolitano LM. The Fragility Index in Randomized Clinical Trials as a Means of Optimizing Patient Care. JAMA Surg. 2019;154:74–9. [DOI] [PubMed]

29. Khan M, Evaniew N, Gichuru M, Habib A, Ayeni OR, Bedi A, et al. The Fragility of Statistically Significant Findings From Randomized Trials in Sports Surgery: A Systematic Survey. Am J Sports Med. 2017;45:2164–70. [DOI] [PubMed]

30. Ruzbarsky JJ, Rauck RC, Manzi J, Khormaee S, Jivanelli B, Warren RF. The fragility of findings of randomized controlled trials in shoulder and elbow surgery. J Shoulder Elbow Surg. 2019;28: 2409–17. [DOI] [PubMed]

31. Ekhtiari S, Gazendam AM, Nucci NW, Kruse CC, Bhandari M. The Fragility of Statistically Significant Findings From Randomized Controlled Trials in Hip and Knee Arthroplasty. J Arthroplasty. 2021;36: 2211–8.e1. [DOI] [PubMed]

32. McCormick KL, Tedesco LJ, Swindell HW, Forrester LA, Jobin CM, Levine WN. Statistical fragility of randomized clinical trials in shoulder arthroplasty. J Shoulder Elbow Surg. 2021;30:1787–93. [DOI] [PubMed]

33. Zabat MA, Giakas AM, Hohmann AL, Lonner JH. Interpreting the Current Literature on Outcomes of Robotic-Assisted Versus Conventional Total Knee Arthroplasty Using Fragility Analysis: A Systematic Review and Cross-Sectional Study of Randomized Controlled Trials. J Arthroplasty. 2024;39:1882–7. [DOI] [PubMed]

34. 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;122:160–6. [DOI] [PubMed]

35. Pascoal E, Liu M, Lin L, Luketic L. The Fragility of Statistically Significant Results in Gynaecologic Surgery: A Systematic Review. J Obstet Gynaecol Can. 2022;44:508–14. [DOI] [PubMed]

36. Budhiraja P, Kaplan B, Kalot M, Alayli AE, Dimassi A, Chakkera HA, et al. Current State of Evidence on Kidney Transplantation: How Fragile Are the Results? Transplantation. 2022;106:248–56. [DOI] [PubMed]

37. Suresh NV, Go BC, Fritz CG, Harris J, Ahluwalia V, Xu K, et al. The fragility index: how robust are the outcomes of head and neck cancer randomised, controlled trials? J Laryngol Otol. 2024;138:451–6. [DOI] [PubMed] [PMC]

38. Wang A, Kwon D, Kim E, Oleru O, Seyidova N, Taub PJ. Statistical fragility of outcomes in acellular dermal matrix literature: A systematic review of randomized controlled trials. J Plast Reconstr Aesthet Surg. 2024;91:284–92. [DOI] [PubMed] [PMC]

39. Condon TM, Sexton RW, Wells AJ, To M. 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:e0237879. [DOI] [PubMed] [PMC]

40. Schröder A, Muensterer OJ, Oetzmann von Sochaczewski C. The fragility index may not be ideal for paediatric surgical conditions: the example of foetal endoscopic tracheal occlusion. Pediatr Surg Int. 2021;37:967–9. [DOI] [PubMed] [PMC]

41. Dettori JR, Norvell DC. How Fragile Are the Results of a Trial? The Fragility Index. Global Spine J. 2020; 10:940–2. [DOI] [PubMed] [PMC]

42. Dettori JR, Norvell DC, Chapman JR. *P*-Value Worship: Is the Idol Significant? Global Spine J. 2019;9: 357–9. [DOI] [PubMed] [PMC]

43. Dervan LA, Watson RS. The Fragility of Using p Value Less Than 0.05 As the Dichotomous Arbiter of Truth. Pediatr Crit Care Med. 2019;20:582–3. [DOI] [PubMed]

44. Ahmed W, Fowler RA, McCredie VA. Does Sample Size Matter When Interpreting the Fragility Index? Crit Care Med. 2016;44:e1142–3. [DOI] [PubMed]

45. Stern BZ, Poeran J. Statistics in Brief: The Fragility Index. Clin Orthop Relat Res. 2023;481:1288–91. [DOI] [PubMed] [PMC]

46. Hong C, Salanti G, Morton SC, Riley RD, Chu H, Kimmel SE, et al. Testing small study effects in multivariate meta-analysis. Biometrics. 2020;76:1240–50. [DOI] [PubMed] [PMC]

47. Carter RE, McKie PM, Storlie CB. The Fragility Index: a P-value in sheep's clothing? Eur Heart J. 2017; 38:346–8. [DOI] [PubMed]

48. Potter GE. Dismantling the Fragility Index: A demonstration of statistical reasoning. Stat Med. 2020; 39:3720–31. [DOI] [PubMed]

49. Fernandez MA, Achten J, Parsons N, Griffin XL, Png M, Gould J, et al.; WHiTE 5 Investigators. Cemented or Uncemented Hemiarthroplasty for Intracapsular Hip Fracture. N Engl J Med. 2022;386:521–30. [DOI] [PubMed]

50. Nelms DW, Vargas HD, Bedi RS, Paruch JL. When the p Value Doesn't Cut It: The Fragility Index Applied to Randomized Controlled Trials in Colorectal Surgery. Dis Colon Rectum. 2022;65:276–83. [DOI] [PubMed]

51. McKechnie T, Yang S, Wu K, Sharma S, Lee Y, Park LJ, et al. Fragility of Statistically Significant Outcomes in Colonic Diverticular Disease Randomized Trials: A Systematic Review. Dis Colon Rectum. 2024;67:414–26. [DOI] [PubMed]

52. Baer BR, Gaudino M, Charlson M, Fremes SE, Wells MT. Fragility indices for only sufficiently likely modifications. Proc Natl Acad Sci U S A. 2021;118:e2105254118. [DOI] [PubMed] [PMC]

53. Khan MS, Fonarow GC, Friede T, Lateef N, Khan SU, Anker SD, et al. Application of the Reverse Fragility Index to Statistically Nonsignificant Randomized Clinical Trial Results. JAMA Netw Open. 2020;3: e2012469. [DOI] [PubMed] [PMC]

54. Baer BR, Gaudino M, Fremes SE, Charlson M, Wells MT. The fragility index can be used for sample size calculations in clinical trials. J Clin Epidemiol. 2021;139:199–209. [DOI] [PubMed] [PMC]

55. FragilityTools. R package version 0.0.2 (2021) [Internet]. GitHub, Inc.; c2025 [cited 2025 Apr 14]. Available from: https://github.com/brb225/FragilityTools

56. Caldwell JE, Youssefzadeh K, Limpisvasti O. A method for calculating the fragility index of continuous outcomes. J Clin Epidemiol. 2021;136:20–5. [DOI] [PubMed]

57. Fragility index: Fragility index for dichotomous and multivariate results [Internet]. GitHub, Inc.; c2025 [cited 2025 Apr 14]. Available from: https://github.com/kippjohnson/fragilityindex/blob/master/vignettes/vignette.Rmd

58. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014;1:140216. [DOI] [PubMed] [PMC]

59. Colquhoun D. Lectures on biostatistics. London: Oxford University Press; 1971.

60. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005;365:1348–53. [DOI] [PubMed]

61. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online Patient Recruitment in Clinical Trials: Systematic Review and Meta-Analysis. J Med Internet Res. 2020;22:e22179. [DOI] [PubMed] [PMC]

62. Baer BR, Fremes SE, Gaudino M, Charlson M, Wells MT. On clinical trial fragility due to patients lost to follow up. BMC Med Res Methodol. 2021;21:254. [DOI] [PubMed] [PMC]

63. Vigden B, Yasseri T. P-values: Misunderstood and misused. Frontiers Physics. 2016;4:6. [DOI]

64. Quatto P, Ripamonti E, Marasini D. Beyond the Fragility Index. Pharm Stat. 2025;24:e2452. [DOI] [PubMed] [PMC]

65. Ruberg SJ. Détente: A Practical Understanding of P values and Bayesian Posterior Probabilities. Clin Pharmacol Ther. 2021;109:1489–98. [DOI] [PubMed] [PMC]

66. Schwen LO, Rueschenbaum S. Ten quick tips for getting the most scientific value out of numerical data. PLoS Comput Biol. 2018;14:e1006141. [DOI] [PubMed] [PMC]

67. Franco A, Malhotra N, Simonovits G. Social science. Publication bias in the social sciences: unlocking the file drawer. Science. 2014;345:1502–5. [DOI] [PubMed]

68. Niforatos JD, Zheutlin AR, Chaitoff A, Pescatore RM. The fragility index of practice changing clinical trials is low and highly correlated with P-values. J Clin Epidemiol. 2020;119:140–2. [DOI] [PubMed]

69. Borenstein M. The case for confidence intervals in controlled clinical trials. Control Clin Trials. 1994; 15:411–28. [DOI] [PubMed]

70. Wasserstein RL, Lazar NA. The ASA's statement on p values: Context, process and purpose. Am Stat. 2016;70:129–33.

71. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000;53:1119–29.

72. Horesh N, Wasserberg N, Zbar AP, Gravetz A, Berger Y, Gutman M, et al. Changing paradigms in the management of diverticulitis. Int J Surg. 2016;33:146–50. [DOI] [PubMed]

73. Horesh N, Lessing Y, Rudnicki Y, Kent I, Kammar H, Ben-Yaacov A, et al. Considerations for Hartmann's reversal and Hartmann's reversal outcomes-a multicenter study. Int J Colorectal Dis. 2017;32:1577–82. [DOI] [PubMed]

74. Horesh N, Rudnicki Y, Dreznik Y, Zbar AP, Gutman M, Zmora O, et al. Reversal of Hartmann's procedure: still a complicated operation. Tech Coloproctol. 2018;22:81–7. [DOI] [PubMed]

75. Horesh N, Lessing Y, Rudnicki Y, Kent I, Kammar H, Ben-Yaacov A, et al. Timing of colostomy reversal following Hartmann's procedure for perforated diverticulitis. J Visc Surg. 2020;157:395–400. [DOI] [PubMed]

76. Garcia MVF, Ferreira JC, Caruso P. Fragility index and fragility quotient in randomized clinical trials. J Bras Pneumol. 2023;49:e20230034. [DOI] [PubMed] [PMC]

77. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis. 1985;27:335–71. [DOI] [PubMed]

78. Thorlund K, Imberger G, Walsh M, Chu R, Gluud C, Wetterslev J, et al. The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis: a simulation study. PLoS One. 2011;6:e25491. [DOI] [PubMed] [PMC]

79. Andrade C. The Use and Limitations of the Fragility Index in the Interpretation of Clinical Trial Findings. J Clin Psychiatry. 2020;81:20f13334. [DOI] [PubMed]

80. Grimes DR. The ellipse of insignificance, a refined fragility index for ascertaining robustness of results in dichotomous outcome trials. Elife. 2022;11:e79573. [DOI] [PubMed] [PMC]

81. Grimes DR, Heathers J. The new normal? Redaction bias in biomedical science. R Soc Open Sci. 2021;8: 211308. [DOI] [PubMed] [PMC]