Journal of Clinical Epidemiology 124 (2020) 34−41

ORIGINAL ARTICLE

The fragility of trial results involves more than statistical significance alone

Stephen D. Walter[a,*], Lehana Thabane[a], Matthias Briel[a,b]

[a]Department of Health Research Methodology, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
[b]Department of Clinical Research, Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel and University of Basel, Basel, Switzerland

Accepted 27 February 2020; Published online 13 April 2020

Abstract

Objectives: The fragility of clinical trial findings has been previously defined as the number of changes in outcomes that are required to change their statistical significance. We show that reliance on statistical significance alone provides only a limited and potentially misleading perspective, and an enhanced approach is developed.

Methods: Clinical importance of trial results and their quantitative stability are incorporated into an enhanced framework to assess fragility.

Results: Examples show that the small data changes required to affect statistical significance may actually be unlikely to occur. Recognizing this limitation, and because statistical significance conveys no information about the treatment effect size, our approach additionally takes into account the clinical importance of the results and their quantitative stability. The interpretation of studies with various combinations of these features is described.

Conclusion: The concept of fragility should include clinical importance of trial findings and their quantitative stability, as well as statistical significance. Study results should be declared as stable only if they are statistically significant and quantitatively stable, but they can be either clinically important or unimportant; otherwise, the findings should be declared as unstable, or fragile.

Keywords: Fragility; Clinical trials; Statistical significance; Clinical importance; Data stability

1. Introduction

In a recent study [1], a sample of meta-analyses was evaluated to determine how often small changes in the data might change the statistical significance and associated interpretation of results. Specifically, the authors examined how many more outcome events or nonevents in one or more of the component studies there would need to be in order for the conclusions to change from being statistically significant to not statistically significant (or vice versa). Significance was defined in terms of whether or not the 95% confidence interval on the pooled treatment effect estimate included the null value. A similar survey of individual clinical trials was published earlier, but considered only changing statistically significant results into being not significant [2]. In both investigations, a Fragility Index was defined as the minimum number of changes in the patient outcome events that would be required to cause the statistical significance of a study (or meta-analysis) to change. The results of studies with smaller values of the index would be regarded as more fragile. The fragility concept has been used in a wide variety of clinical applications [3−8].

This approach, of relying exclusively on statistical significance to determine which study findings might be fragile, is in stark contrast with recent calls by the American Statistical Association (ASA) to abandon the term ''statistical significance'' [9,10]. The ASA has provided a 400-page special issue journal with numerous authors suggesting how the current practice of using *P*-values should be updated. Several prominent journals have issued editorial statements in response [11−13].

Funding: This study was partially funded by a grant from the Natural Sciences and Engineering Research Council (Canada).

Conflicts of interest: The authors declare they have no conflicts of interest.

Authors' contributions: S.D.W. conceived and coordinated the study and wrote the first draft of the article. All authors critically revised the article and approved the final version before submission.

* Corresponding author. Tel.: 905-525-9140; fax: 905-577 0044.
E-mail address: walter@mcmaster.ca (S.D. Walter).

https://doi.org/10.1016/j.jclinepi.2020.02.011
0895-4356/

What is new?

- "Fragility" is intended to measure the reliability of clinical study results.
- Statistical significance alone has limitations and can be misleading.
- Clinical importance and quantitative stability of results are also relevant.
- We propose an enhanced framework incorporating all three factors.

Given the current controversies about the use of statistical significance and *P*-values, it is appropriate to consider this issue in the context of assessing fragility in the results of clinical trials. We do this initially through an illustrative example, in which a minimal change in the data would cause a drastic change in its statistical significance, and hence, the study would be declared as "fragile" by the Fragility Index, but where this change might actually be unlikely to occur. We then move on to discuss two additional elements of fragility that seem relevant, in particular, the clinical importance of the observed treatment effect and the quantitative stability of that finding. A framework is then developed to more comprehensively assess the fragility of a study by using all three of these elements.

2. Methods

As noted previously, previous authors [1,2] have defined fragility through an index, this being the smallest number of changes in the patient outcome events that would result in a randomized trial or meta-analysis moving from being statistically significant to being not statistically significant, or vice versa. A low value of this Fragility Index is intended to convey the notion that the study conclusions are, in some sense, statistically less "solid," unreliable, unstable, or "*fragile*." Because the present study will be discussing additional factors that also influence one's impression of "fragility," we will instead refer to the Fragility Index method, which examines the impact of small changes in the data only on the classification of statistical significance, as dealing with "*significance stability*."

We now introduce a simple example to illustrate the impact in a hypothetical data set of making the smallest possible change in outcomes. This will highlight the limitations of statistical significance in this context and will indicate that evaluating fragility through an examination of statistical significance alone is inadequate and could actually be misleading.

2.1. Example: the small changes in the data used to assess the fragility index may be unlikely to occur

Implicit in the Fragility Index approach is the assumption that some defined minimal changes in the data could easily have happened by chance; by extension, these incremental changes are taken to represent counterfactual scenarios that can be compared with the observed data as plausible alternative study results. However, it is important to note that modifying the data by changing the outcome even for only one patient may correspond to alternative results that are actually quite unlikely.

To see this, consider the trial whose original data are shown in Table 1.

There have been five deaths out of 95 patients with treatment A, but no deaths out of 96 patients on treatment B. This result gives a two-sided *P*-value of 0.029 with Fisher's exact test, which would be declared statistically significant under the usual 5% criterion. Now, if we pursue a "significance only" approach to examine fragility, we can assess what would have happened if there had been a minimal change, so that there was instead one death out of the 96 patients on treatment B. This leads to a big change in the statistical significance, with a new *P*-value of 0.118, putting the study well into the "not statistically significant" category. Because the statistical significance has changed category after this smallest possible change in the data, the study would be declared by the use of the Fragility Index method as "fragile" (specifically, the Fragility Index = 1), or more precisely, as having "unstable significance." (Similar conclusions are reached if we [1] decrease the number of deaths in A by 1 or [2] increase the number of deaths in B and also decrease the number of deaths in A by 1 to keep the total number of deaths constant; in both these scenarios, the result is not statistically significant, with two-sided *P* values of 0.059 and 0.211, respectively.)

The fallacy in this argument is that the change from 0 to 1 death in treatment group B may actually be unlikely to occur. For example, suppose the data in Table 1 represent a trial that is comparing an experimental treatment A, and treatment B, which is usual care. Furthermore, suppose that it is known from previous studies that death is rare under usual care; specifically, suppose the true underlying death rate is 0.1%. Then the probability of observing no deaths in 96 patients is approximately 90.8%, compared with a probability of only 8.7% of observing exactly one death.

Table 1. Example where an unlikely small change in the data substantially changes statistical significance

	Original table		Modified table	
	Treatment A	Treatment B	Treatment A	Treatment B
Died	5	0	5	1
Alive	90	96	90	95
P-value	0.029		0.118	

So, in other words, we would here be assessing fragility by comparing the observed data with an alternative study outcome that is much less likely to be observed than what has been seen in the actual data. We would here be assessing significance fragility by comparing the observed data (with no deaths on B) with a hypothetical alternative (one death on B), which is far less likely to be seen in reality. In that sense, it seems unwise (on this basis alone) to regard the study findings as fragile.

2.2. The Fragility Index value does not tell us how likely the "smallest change" might be

A Fragility Index of 5, for instance, means that at least five changes of patient outcome status would be required to change the statistical significance of a study result. However, a corollary of the previous point is that we cannot tell if 5 (in this example) is a lot or a little in the specific context of the study in question. In other words, the Fragility Index tells us about the change necessary to alter the statistical significance of a study, but it does not tell us how likely it is for such a change to occur. This is essential because variability in the data is not taken into account.

2.3. Factors other than statistical significance should be taken into account

Having demonstrated that there are major problems with the approach of assessing fragility using the Fragility Index, which is based on significance stability alone, it is helpful at this point to refer back to some earlier literature. We will review work from the 1990s by Feinstein [14], who attempted to define the fragility in terms of the quantitative stability of study results relative to clinical importance, and by Walter [15], who suggested combining statistical significance with Feinstein's fragility measure. We then go on to propose an updated and integrated framework that uses several factors to define fragility of a particular study or meta-analysis.

In 1990, Feinstein [14] proposed a "Unit Fragility Index" $f = N/(n_1.n_2)$, where n_1 and n_2 are the sample sizes in the two treatment groups, and $N = n1 + n2$ is the total sample size in the study. It can be shown that f corresponds to the change in the risk difference induced by the smallest possible change in the data (increasing or decreasing the frequency in one cell of the usual 2×2 table of data, while keeping both the row and column totals of the table fixed). A large value of the Unit Fragility Index would imply that relatively substantial changes to the risk difference would occur with only a minimal change to the data, that is, that the findings are inherently unstable from the quantitative perspective.

An additional ingredient in Feinstein's approach is to evaluate g, being the difference between the treatment effect in terms of observed absolute risk difference, and a "quantitatively significant difference" δ. The quantity δ can be thought of as the smallest clinically important effect, sometimes referred to as the "minimally important difference," or the MID [16]. Note that the clinically important effect should be prespecified, and based on clinical considerations, but it would not necessarily correspond to the assumed value of the treatment effect (often denoted as "delta") used in sample size and power calculations at the planning stages of a study.

Various combinations of the observed values of f and g then lead to a characterization of the study results. First, if $g \geq 0$, Feinstein's method declares the result as "*quantitatively significant*," but not otherwise. Quantitative significance simply implies that the observed treatment effect exceeds the minimum effect that would be regarded as clinically important, but it does not directly take statistical variation into account. Again, to avoid confusion by the use of the term "significance" in this context, from this point onward, we will instead refer to Feinstein's "quantitative significance" as "*clinical importance*."

Second, if $g > f$, with Feinstein's approach, the results are declared to be "*quantitatively stable.*" This implies that the smallest possible change in the data would not result in the treatment effect falling below the clinically important difference.

Third, if $g < 0$, under Feinstein's method, one would conclude that the result is not clinically important because it fails to show an effect that exceeds the minimal clinically important effect; again, this conclusion takes no direct account of statistical variation. However, if $f \geq |g|$ in addition, Feinstein's Unit Fragility Index is greater than the observed treatment effect (in absolute value); here, we would infer that there is enough instability in the results that they could have become clinically important with only a minimal change in the observed data. On the other hand, if $f < |g|$ in addition to having $g < 0$, the study result would be classified as clinically unimportant and quantitatively sufficiently stable that a small change in the data would not alter that conclusion.

To be clear, and specifically to distinguish it from *significance stability* as in the Fragility Index method, we will refer to Feinstein's Unit Fragility analysis as dealing with *quantitative stability.* Thus, a study result would be taken by Feinstein as *quantitatively stable* if the observed treatment effect exceeds the clinically important difference, and if, in addition, the smallest possible change in the data would not alter that conclusion; or, that the observed treatment effect fails to exceed the clinically important difference, and a small change in the data would not alter that conclusion. In contrast, a small change in a quantitatively unstable finding could result in a clinically important result becoming clinically unimportant, or vice versa, so one remains uncertain about the clinical importance of the result.

Feinstein's framework was developed further [15], and we now update this approach, basing it on three essential elements: clinical importance, quantitative stability, and statistical significance. To summarize the combined impact of these three factors, we can declare the results of each type of study as overall *fragile* or *stable.* To operationalize this method, we now consider the inferences that can be

drawn from studies found with the various possible combinations of the three elements. For simplicity, we only consider superiority trials, where the significance testing is against a null hypothesis of no treatment effect. Some simple modifications would be required to assess the fragility of noninferiority or equivalence trials.

2.4. Three-factor framework to assess study fragility or stability

The framework developed here requires that one assesses the results of a given study according to the three factors: statistical significance, clinical importance, and quantitative stability. Statistical significance is evaluated in the usual way, declaring a study to be statistically significant or not, relative to an agreed criterion (traditionally, significance is declared at the 5% type I error rate level). Using a categorization of statistical significance alone in this way corresponds to the classification of studies by the Fragility Index method proposed earlier [1,2].

The second element, clinical importance, is determined according to whether or not the observed treatment effect exceeds the minimally important difference, as defined by the investigators, ideally based on perspectives of patients and clinical practitioners [17,18]. As discussed below, observers may disagree about what constitutes a clinically important effect [16,19], and hence, they may also disagree about the fragility or stability of study results. However, it seems important to incorporate clinical importance as an element of fragility because ultimately, clinicians, patients, and policy makers need to make a decision about whether or not to adopt a new treatment.

The third element, quantitative stability, can be introduced into the framework in a number of different ways. First, Feinstein's approach [14] would define it in terms of how (or if) a change of one patient outcome would affect the assessed clinical importance of the results; one evaluates if this smallest possible change to the data does or does not change one's decision about clinical importance. Second, and as a modification of Feinstein's method, one could use a more stringent criterion and examine the effect of a given, larger threshold number of data changes; for example, one could might find out if a change in five or more patient outcomes would change the conclusion on clinical importance. Third, one could retain the Fragility Index as proposed previously, along with a defined threshold value. Quantitative stability would then be declared with respect to the impact (if any) of such changes on statistical significance. As an example, if changes to 10 patient outcomes would be required to change the statistical significance, and an index value of 5 had been agreed as a threshold, one would declare the results to be quantitatively stable. But if fewer than five changes would alter the statistical significance, the results would be called quantitatively unstable.

A summary of how these factors can be applied is shown in Table 2.

There are eight possible combinations of the three factors, as we now review (Fig. 1).

For each case, we briefly describe the study interpretation and arrive at an overall declaration of stability or fragility of the study result. For simplicity, we will assume that Feinstein's approach is used to assess quantitative stability in each of these case scenarios, but we later discuss how their interpretation might differ if the Fragility Index method had been used instead.

2.4.1. Case 1

Study results are statistically significant, clinically important, and quantitatively stable. A clinically important effect has been found, it is statistically significant, and the conclusion would not be altered by small changes to the data. Overall: STABLE.

2.4.2. Case 2

Study results are statistically significant, clinically important, and quantitatively unstable. The best estimate is that there is a clinically important effect, and it is statistically significant. However, the conclusion could be affected by small changes to the data, such that the true treatment effect would then not be regarded as clinically important. Overall: FRAGILE.

2.4.3. Case 3

Study results are statistically significant, clinically unimportant, and quantitatively unstable. A clinically unimportant effect has been found, but it was statistically significant. In principle, this could happen with large sample sizes, leading to highly precise estimates of a small treatment effect. However, the result is quantitatively unstable, so there remains the possibility of a clinically important effect that would have been found in slightly different data. In practice, this scenario is relatively unlikely, and it would only occur if, despite the very precise estimates, the data happen to lie particularly close to a boundary case for this scenario, as in Fig. 1, so that even a small change in study outcomes would have changed the conclusion about clinical importance. Overall: FRAGILE.

2.4.4. Case 4

Study results are statistically significant, clinically unimportant, and quantitatively stable. There appears to be a reliable finding of a clinically unimportant effect, but nevertheless, one that is statistically significant. This conclusion would not be affected by small changes in the data. Overall: STABLE.

2.4.5. Case 5

Study results are statistically not significant, clinically important, and quantitatively stable. The results suggest a clinically important effect, a conclusion that would be unaffected by small changes in the data. However, more data

Table 2. Steps in assessing the fragility of study results through a three-factor framework

Step	Notes
1. Assess **statistical significance**	Takes sample variation into account.
2. Determine **clinical importance**	Investigators should prespecify a threshold for the smallest treatment effect that would be important to patients in their clinical practice (MID).
3. Evaluate **quantitative stability**	Quantitative stability can be defined in terms of whether or not a given threshold level of change in the data would change the apparent clinical importance of the results. This threshold could be with respect to changing a defined number of patient outcomes, with the minimum being a change of only one patient outcome. *Alternative*: examine if such changes in the data would affect the statistical significance of the results.
4. Declare the study results to be **stable** or **fragile**	Claim stability for the study if its results are statistically significant and quantitatively stable, remaining clinically important or unimportant. Otherwise, declare study fragile.

Abbreviation: MID, minimally important difference.

would be required to confirm this in terms of statistical significance. Overall: FRAGILE.

2.4.6. Case 6

Study results are statistically not significant, clinically important, and quantitatively unstable. This study provides weak evidence of a clinically important effect, as indicated by its point estimate. This conclusion about clinical importance could have been different with only small changes to the data, and lack of statistical significance also indicates uncertainty about the findings. Overall: FRAGILE.

2.4.7. Case 7

Study results are statistically not significant, clinically unimportant, and quantitatively unstable. Very little can be inferred. A clinically unimportant effect has been shown, but one that cannot be relied on, given its lack of statistical significance and its quantitative instability, meaning that only small changes in the data could lead to a different conclusion about clinical importance. Overall: FRAGILE.

2.4.8. Case 8

Study results are statistically not significant, clinically unimportant, and quantitatively stable. This study leans to a conclusion of a clinically unimportant treatment effect, but the lack of statistical significance does not support this. Further investigation would be warranted in this case (and in case 7) only if investigators remained convinced of a small effect, and that is what they wish to demonstrate more convincingly. Overall: FRAGILE.

By comparing these scenarios, we see that overall stability will be declared only if the study results are statistically significant and quantitatively stable. There will be an associated decision about the clinical importance of the results.

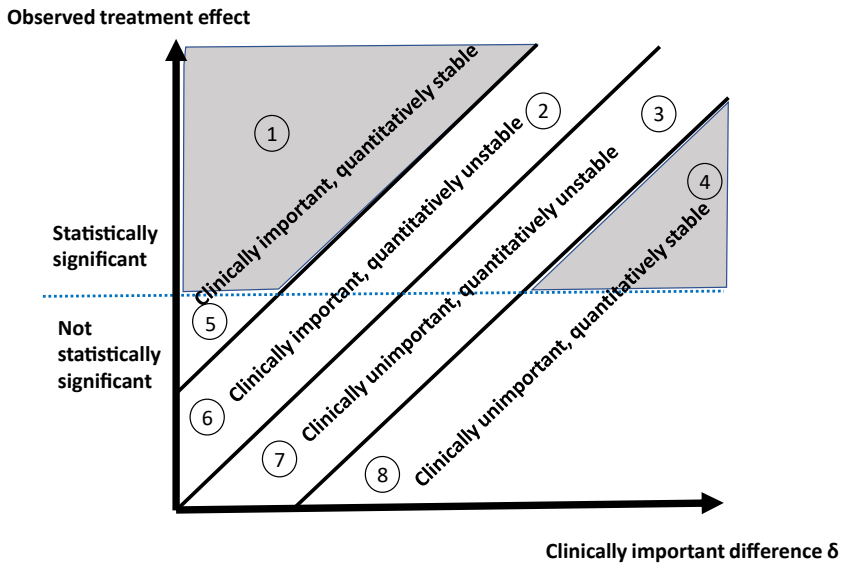

Fig. 1. Study fragility according to statistical significance, clinical importance, and quantitative stability. Shaded zones indicate stable studies.

Failing either or both of statistical significance and quantitative stability, the results will be declared to be fragile.

Fig. 1 shows how these cases are related, with respect to the observed treatment effect and the definition of the clinically important difference. Studies found toward the upper part of this plot correspond to results that show a larger observed treatment effect, are more statistically significant, and correspondingly of increasing clinical importance. Studies above the main 45-degree diagonal from the origin are those in which the observed treatment effect exceeds the definition of clinical importance, and those below the diagonal show clinically unimportant effects. If they are close to the main diagonal, the classification of results as clinically important or not becomes less certain, or quantitatively unstable.

The factors in this framework are logically related in many ways. First, we note, for instance, that strong statistical significance will more often be associated with quantitatively stable findings. Second, changing the agreed definition of the clinically important difference may affect the study classification as stable or fragile, even if the observed outcomes are unaltered; for example, requiring a more stringent definition of the clinically important difference could ''move'' the clinically important and stable results of a study initially in Case 1 to the right in Fig. 1, through the regions of quantitative instability (where the determination of clinical importance is uncertain), ultimately to become a ''Case 4,'' with quantitatively stable results showing a clinically unimportant treatment effect. Hence, other researchers may disagree in their interpretation of a study as fragile or stable because they have adopted different definitions of clinical importance.

Third, we remark that the relative frequencies of the various cases mentioned earlier will depend on sample sizes and the associated precision of estimated treatment effects. In particular, with large sample sizes and precisely estimated treatment effects, the central diagonal bands (in Fig. 1) with quantitatively unstable results will be narrow and less likely to occur. On the other hand, with less data and poorer precision, these bands will be wider, yielding more studies with quantitatively unstable results, and these studies will be declared as fragile.

Finally, requiring stronger statistical significance (e.g., changing the type I error rate from 5% to 1%) would move the horizontal boundary line for statistical significance upwards and thus reduce the number of studies that can declare stable findings of a clinically important treatment effect, and correspondingly more studies will be classified as fragile.

3. Discussion

We have seen how the use of the Fragility Index [1,2] can be misleading because it is exclusively based on statistical significance, whose *P*-values are inappropriate as a measure of evidence of a treatment effect. In addition, interpretation of a given value of the Fragility Index is difficult because sample variation is not taken into account; as a consequence, it does not tell us how likely or unlikely are the associated changes in the data that would lead to a change in the statistical significance of a study. Furthermore, given the general concerns currently being raised about *P*-values [9,10], it seems inadequate to use the Fragility Index in isolation because it is essentially telling us only about the uncertainty in those *P*-values.

Because the Fragility Index actually only deals with the fragility of statistical significance, it in no way incorporates any notion of clinical importance. Hence, one might use the index and declare a study to be stable but still have no idea if the results actually matter in clinical practice. Conversely, because the index does not tell us about the likelihood of the changes required to change the assessed statistical significance, there are situations where the index would declare a study to be fragile, but where in fact the conclusion about statistical significance would be unlikely to change as a result of the index frequency of changes in the data.

Furthermore, we need to bear in mind that *P*-values represent a different concept from statistical significance. The *P*-value tells us about how much evidence there is against the null hypothesis (but based on underlying model assumptions). Statistical significance (corresponding to the *P*-value being above or below a threshold defined by the chosen type I error rate) can be helpful if an inferential decision has to be taken about the effect being tested.

These considerations have led to the development here of the proposed, enhanced framework for assessing the fragility of study results. First, statistical significance has been retained as one of its components because that reflects sample variation in results, and, in particular, it tells us if a particular finding might have been because of chance, if, in fact, there was no treatment effect. Second, the framework also takes clinical importance of the study findings into account. The rationale for its inclusion is that clinicians will want to achieve a final decision about whether or not to adopt a new treatment. To ignore clinical importance seems deficient.

Third, the framework involves an assessment of the quantitative stability in the data. Here, one has to decide how much change in the data can be tolerated. Typically, the changes in the data can be defined by a threshold number of patients whose outcome status would need to change to alter the conclusions, either with respect to clinical importance or to statistical significance. Feinstein's proposal was for this threshold to be the minimum change of exactly one patient outcome, with an associated impact on clinical importance, but other thresholds might reasonably be higher, and based on empirical evidence. In surveys of trials and meta-analyses with statistically significant results [1,2], the median Fragility Index values were 8 and 12, respectively, as they affected statistical significance.

In this article, we have limited our attention to trials with a binary outcome; however, further work might suggest

extensions of our ideas to other types of outcomes, such as continuous or time-to-event data. Our framework is applicable in principle to meta-analyses too, but here one would have to consider statistical heterogeneity in addition; in our view, unexplained heterogeneity increases fragility, but this also needs to be confirmed by further work. Finally, the framework is limited by considering only single outcomes from a study, but it could potentially be extended to allow for multiple outcomes.

One also has to choose where the impact of small changes in the data will be assessed. In this article, we have primarily used quantitative stability to address their impact on the conclusion about clinical importance. One could alternatively use quantitative stability to assess impacts on statistical significance, in much the same way as originally proposed for the Fragility Index [2]. However, using statistical significance as the target would mean that two of the three factors (quantitative stability and statistical significance) in the assessment framework would be strongly related, and so one would essentially be dealing with significance fragility or trying to make some probabilistic statement about the reliability of *P*-values. Although this would conform with the original Fragility Index proposal, it seems to be somewhat a contorted logic, and an examination of the reliability of a finding of clinical importance seems much preferable.

A suggestion made by a reviewer was that the magnitude of change in the *P*-value might be more revealing than just assessing if statistical significance has changed above or below some threshold. We agree, but one would then still essentially be using a single dimension defined by the possible values of *P*. Furthermore, clinical investigators still largely rely on the significant/not significant dichotomy to interpret the usefulness of their trial results.

Practical use of the proposed framework requires investigators: to choose the level of the threshold to declare statistical significance, to define the minimum clinically important effect, and to select the criteria for quantitative stability and its threshold in terms of minimum data changes required. These tasks are not trivial, even for experienced methodologists. A review has shown that there is, for instance, a wide range of approaches that investigators have used to specify the target treatment effect in clinical trials, for the purpose of sample size calculation [20]. The original Fragility Index is relatively simple to calculate, but it is limited in scope and interpretability. It only deals with the stability of statistical significance, a feature that itself is now identified as something that is a poor measure of evidence. And it says nothing about target effect sizes or what would constitute a clinically important finding.

In contrast, the acquisition of the elements required in the currently proposed framework avoids the limitations of inference being essentially based on statistical significance alone and can provide greater interpretability of the data in terms of clinical importance, while maintaining some focus on statistical uncertainty in the findings. An additional advantage is that the framework has broader applicability, extending beyond trials with a dichotomous outcome (such as to studies with continuous or survival outcomes).

Once the three constituent factors have been established for the framework, the observed 2×2 data table can be suitably modified to establish which of the eight case scenarios applies. It may also be helpful to calculate the numerical locations of all the boundaries in Fig. 1 and plot where the study (or meta-analysis) in question lies.

The calculations for this framework are not difficult, and the insights they provide go well beyond what is available from assessing the stability of statistical significance with the Fragility Index in isolation. A more formal assessment of this issue might involve analysis of statistical robustness, by defining and simulating data generating models, for example [21−23]. A Bayesian approach may also be possible. However, we have here pursued the concept of fragility, as it has been framed in biomedical research, as a relatively simple or ad hoc concept, that may have greater appeal to clinical investigators.

In line with the current moves away from inappropriate uses of *P*-values alone, we suggest that examination of the additional elements in our proposed framework will yield considerable benefits.

References

[1] Atal I, Porcher R, Boutron I, Ravaud P. The statistical significance of meta-analyses is frequently fragile: definition of a fragility index for meta-analyses. J Clin Epidemiol 2019;111:32−40.
[2] Walsh M, Srinathan S, Mrkobrada M, McAuley DF, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results: a case for a Fragility Index. J Clin Epidemiol 2014;67:622−8.
[3] Shen Y, Cheng X, Zhang W. The fragility of randomized controlled trials in intracranial hemorrhage. Neurosurg Rev 2019;42:9−14.
[4] Ruzbarsky JJ, Khormaee S, Rauck RC, Warren RF. Fragility of randomized clinical trials of treatment of clavicular fractures. J Shoulder Elbow Surg 2019;28:415−22.
[5] Bertaggia L, Baiardo Redaelli M, Lembo R, Sartini C, Cuffaro R, Corrao F, et al. The Fragility Index in peri-operative randomised trials that reported significant mortality effects in adults. Anaesthesia 2019;74:1057−60.
[6] Topcuoglu MA, Arsava EM. The fragility index in randomized controlled trials for patent foramen ovale closure in cryptogenic stroke. J Stroke Cerebrovasc Dis 2019;28:1636−9.
[7] Khormaee S, Choe J, Ruzbarsky JJ, Agarwal KN, Blanco JS, Doyle SM, et al. The fragility of statistically significant results in pediatric qrthopaedic randomized controlled trials as quantified by the fragility index: a systematic review. J Pediatr Orthop 2018;38:e418−23.
[8] 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.
[9] Statistical inference in the 21st century: a world beyond $p < 0.05$. In: Wassertein RL, Schirm AL, Lazar NA, editors. American Statistician, 73 2019:1−401.
[10] Wasserstein RL, Lazar NA. The ASA's statement on p-values: context process, and purpose. Am Statistician 2016;70:129−33.
[11] Amrhein V, Greenland S, McShane B. Retire statistical significance. Nature 2019;567:305−7.

[12] Ioannidis J. The proposal to lower P value thresholds to .005. JAMA 2018;319:1429−30.
[13] Harrington D, Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand SL, et al. New guidelines for statistical reporting in the journal. N Engl J Med 2019;381:285−6.
[14] Feinstein AR. The unit fragility index: an additional appraisal of ''statistical significance'' for a contrast of two proportions. J Clin Epidemiol 1990;43:201−9.
[15] Walter SD. Statistical significance and fragility criteria for assessing a difference in two proportions. J Clin Epidemiol 1991;44:1373−8.
[16] Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407−15.
[17] Crosby RDK, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003; 56:395−407.
[18] Johnston BC, Ebrahim S, Carrasco-Labra A, Furukawa TA, Patrick DL, Crawford MW, et al. Minimally important difference estimates and methods: a protocol. BMJ Open 2015;5(10): e007953.
[19] Miller WR, Manuel JK. How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration. Drug Alcohol Rev 2008;27:524−8.
[20] Cook JA, Hislop J, Adewuyi TE, Harrild K, Altman DG, Ramsay C, et al. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. Health Technol Assess 2014;18:v−vi. 1-175.
[21] Huber PJ. Robust statistics. New York: Wiley; 2005.
[22] Clarke BR. Robustness theory and application. New York: Wiley; 2018.
[23] Farcomeni A, Ventura L. An overview of robust methods in medical research. Stat Methods Med Res 2010;21:111−33.