ELSEVIER

Journal of Clinical Epidemiology 67 (2014) 622–628

Journal of Clinical Epidemiology

The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index

Michael Walsh[a,b,c,*], Sadeesh K. Srinathan[d], Daniel F. McAuley[e,f], Marko Mrkobrada[g], Oren Levine[b], Christine Ribic[a,b], Amber O. Molnar[h], Neil D. Dattani[i], Andrew Burke[g], Gordon Guyatt[a,b], Lehana Thabane[a], Stephen D. Walter[a,b], Janice Pogue[a,c], P.J. Devereaux[a,b,c]

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada, L8S4L8*
[b]*Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada, L8S4L8*
[c]*Population Health Research Institute, Hamilton Health Sciences and McMaster University, 237 Barton St East, Hamilton, Ontario, Canada, L8L2X2*
[d]*Department of Surgery, University of Manitoba, Health Sciences Centre, GE611 Sherbrooke St., Winnipeg, Manitoba, Canada, R3A1R9*
[e]*Centre for Infection and Immunity, Queen's University of Belfast, Health Sciences Building, 97 Lisburn Road, Belfast, BT97BL, UK*
[f]*Regional Intensive Care Unit, Royal Victoria Hospital, Victoria Hospital, 274 Grosvenor Road, Belfast, BT126BA, UK*
[g]*Department of Medicine, Western University, London Health Sciences Centre, University Hospital, 339 Windemere Road, London, Ontario, Canada, N6A 5A5*
[h]*Department of Medicine, University of Ottawa, Ottawa Hospital, Riverside Campus, 1967 Riverside Drive, Ottawa, Ontario, Canada, K1H7W9*
[i]*Faculty of Medicine, University of Toronto, Medical Sciences Building, 1 Kings College Circle, Toronto, Ontario, Canada, M5S1A8*

Accepted 7 October 2013; Published online 5 February 2014

Abstract

Objectives: A *P*-value <0.05 is one metric used to evaluate the results of a randomized controlled trial (RCT). We wondered how often statistically significant results in RCTs may be lost with small changes in the numbers of outcomes.

Study Design and Setting: A review of RCTs in high-impact medical journals that reported a statistically significant result for at least one dichotomous or time-to-event outcome in the abstract. In the group with the smallest number of events, we changed the status of patients without an event to an event until the *P*-value exceeded 0.05. We labeled this number the Fragility Index; smaller numbers indicated a more fragile result.

Results: The 399 eligible trials had a median sample size of 682 patients (range: 15–112,604) and a median of 112 events (range: 8–5,142); 53% reported a *P*-value <0.01. The median Fragility Index was 8 (range: 0–109); 25% had a Fragility Index of 3 or less. In 53% of trials, the Fragility Index was less than the number of patients lost to follow-up.

Conclusion: The statistically significant results of many RCTs hinge on small numbers of events. The Fragility Index complements the *P*-value and helps identify less robust results.

Keywords: Randomized controlled trials; Research methodology; Lost to follow-up

1. Introduction

In randomized controlled trials (RCTs), several factors influence our belief in whether a treatment has an effect. One influential factor is whether a hypothesis test demonstrates statistical significance by rejecting the null hypothesis at a particular threshold, most often a *P*-value less than 0.05. Statistical significance implies that the observed result, or a more extreme result, is unlikely to occur by chance alone and that the groups are therefore likely to truly differ.

The concept of a threshold *P*-value to determine statistical significance aids our interpretation of trial results. It allows us to distill the complexities of probability theory into a threshold value that informs whether a true difference

Conflict of interest: The authors have no conflicts of interest to declare. M.W. is supported by a New Investigator Award from the Kidney Research Scientist Core Education National Training (KRESCENT) Program. P.J.D. is supported by a Heart and Stroke Foundation of Ontario Career Investigator Award.

Ethical approval was not required for this study.

Funding: No funding was received for the conduct of this study and no sponsor was required.

* Corresponding author. Division of Nephrology, Marian Wing, St. Joseph's Hospital, 50 Charlton Ave E, Hamilton, Ontario L8N 4A6, Canada. Tel.: 905-522-1155; fax: +1-905-521-6153.

E-mail address: walshm@phri.ca (M. Walsh).

0895-4356
http://dx.doi.org/10.1016/j.jclinepi.2013.10.019

What is new

- Metrics exist, most notably p-values and 95% confidence intervals, to help determine how likely observed treatment effects are on the basis of chance.
- A shift of only a few events in one group could change typical hypothesis tests above the usual thresholds considered statistically significant.
- The Fragility Index helps identify the number of events required to change statistically significant results to non-significant results.
- The Fragility Index demonstrate results from randomized controlled trials in high impact journals frequently hinge on three or fewer events.

likely exists. However, the use of threshold *P*-values has received a great deal of criticism as an overly simple concept to determine whether a treatment effect is likely to truly exist. For example, readers may place a similar degree of belief in results with similar *P*-values irrespective of other factors such as the size of the trial or number of events in the trial. Furthermore, readers may have very different beliefs in the existence of a treatment effect on the basis of very small differences in *P*-values when one is above and one below the threshold value (eg, $P = 0.051$ and $P = 0.049$). Despite these limitations, the calculation, reporting, and interpretation of *P*-values and the wide acceptance of a $P < 0.05$ as significant persist. One approach to better communicate the limitations of *P*-value thresholds is to report an additional metric that demonstrates how easily significance based on a threshold *P*-value may be exceeded.

Consider a hypothetical example in which two RCTs at low risk of bias evaluate investigational drugs compared with placebo for the prevention of myocardial infarction. In the first trial, 100 patients are randomized to receive drug A and 100 patients to receive placebo. Fewer patients who receive drug A suffer a myocardial infarction (one vs. nine patients, $P = 0.02$ by Fisher's exact test). The second trial randomizes 4,000 patients to receive drug B and 4,000 patients to receive placebo. Fewer patients who receive drug B suffer a myocardial infarction (200 vs. 250 patients, $P = 0.02$).

As both trials were at low risk of bias and their results demonstrated nearly the same *P*-value, one's confidence in a true effect might be similar. However, the results from the first trial would be easily influenced by a small change in the numbers of events. If only one more patient experienced a myocardial infarction in the treatment group of the first trial, the *P*-value would change to 0.06. Despite the still impressive relative risk reduction of 78%, it would no longer be considered statistically significant. In contrast, adding one event to the treatment group in the second trial would have no meaningful impact on either the *P*-value, which would remain 0.02, or the point estimate of the relative risk reduction, which would remain 20%.

Knowing that statistical significance may be lost as a result of a few additional events may reduce confidence that a true treatment effect exists. The minimum number of patients whose status would have to change from a nonevent to an event required to turn a statistically significant result to a nonsignificant result could be used as an index of the fragility of the result (ie, a Fragility Index), with smaller numbers indicating a more fragile result. To explore the concept of fragility, we reviewed RCTs published in high-impact general medical journals and calculated the Fragility Index of results reported to have a $P < 0.05$.

2. Methods

We identified RCTs with a statistically significant result for at least one dichotomous outcome in the abstract published in high-impact general medical journals. We then calculated the Fragility Index for each of these trial results and summarized the Fragility Index as a function of trial characteristics.

2.1. Identification of trials

We used PubMed to identify RCTs published in the *New England Journal of Medicine*, the *Lancet*, the *Journal of the American Medical Association*, the *Annals of Internal Medicine*, or the *British Medical Journal* using the randomized controlled trial MeSH term. We drew a convenience sample determined by setting time limits of January 2004 to December 2010. Two reviewers independently screened all identified abstracts. We included trials that (1) were two parallel arm or two-by-two factorial design RCTs involving humans (ie, cluster RCTs, crossover RCTs, and >2 parallel arm designs were excluded), (2) allocated participants in a 1 to 1 ratio to treatment and control, and (3) in the abstract, reported at least one dichotomous or time-to-event outcome as significant ($P < 0.05$ or a 95% confidence interval (CI) that excluded the null value) under a null hypothesis that no difference existed. Statistically significant results for a noninferiority hypothesis were excluded.

2.2. Data

Two reviewers independently used standardized forms to abstract data from each trial. Abstracted data elements included details of the statistically significant outcome (type of outcome, whether it was the primary study outcome, use of adjustment, number of patients randomized to each group, number of patients analyzed in each group, and the number of patients who experienced an outcome in each group), trial design (method of allocation, adequacy of concealment, blinding, inclusion of all randomized patients

Trial Result		
	Event	No Event
Treatment A	a	b
Treatment B	c	d
Fisher's Exact Test p<0.05		

Calculated Fragility		
	Event	No Event
Treatmnt A	a+f	b-f
Treatment B	c	d
Fisher's Exact Test p≥0.05		

Fig. 1. Calculation of the Fragility Index in the scenario in which treatment A group has the fewest events. The smallest value of "*f*" that causes the Fisher's exact *P*-value to meet or exceed the 0.05 level is considered the Fragility Index. Higher values indicate less fragile results.

in the analysis according to the group they were allocated to), and the number of participants lost to follow-up. For trials with more than one significant result reported in the abstract, only the first reported result was considered. Disagreements between reviewers were adjudicated by a third reviewer.

2.3. Fragility Index calculation

The results of each trial were represented in a two-by-two contingency table using the patient sample that the authors used in the original analysis. For trials with a time-to-event outcome, the number of events in each group for the entire follow-up period was used as events to construct a two-by-two table. The Fragility Index was calculated by adding an event from the group with the smaller number of events (and subtracting a nonevent from the same group to keep the total number of patients constant) and recalculating the two-sided *P*-value for Fisher's exact test. Events were iteratively added until the first time the calculated *P*-value became equal to or greater than 0.05 (Fig. 1). The number of additional events required to obtain a *P*-value ≥0.05 was considered the Fragility Index for that trial result. We also performed a sensitivity analysis in which the Fragility Index was calculated using the same two-by-two table as the main analysis but was defined as the number of events required to obtain a 95% CI for the relative risk that included one, the null value.

Table 1. Characteristics of included trials

Characteristic	Number (*n* = 399)
Journal, *n* (%)	
New England Journal of Medicine	165 (41.3)
Lancet	112 (28.1)
Journal of the American Medical Association	48 (12.0)
Annals of Internal Medicine	33 (8.3)
British Medical Journal	41 (10.3)
Sample size, median (min–max)	682 (15–112,604)
Number of outcome events, median (min–max)	112 (8–5,142)
Reported *P*-value, *n* (%)	
<0.05–0.01	186 (46.6)
<0.01–0.001	168 (42.1)
<0.001	45 (11.3)
Included outcome, *n* (%)	
Primary	263 (65.9)
Composite	132 (33.1)
Time to event	206 (51.6)
Adjusted	35 (8.8)

Abbreviations: min, minimum; max, maximum.

2.4. Fragility Index by trial characteristics

We summarized the Fragility Index for the sampled studies using descriptive statistics. We described the Fragility Index of results based on whether they were primary outcomes, time-to-event outcomes, composite outcomes, whether the original result was adjusted for covariates, analysis was done by time-to-event methods, whether all patients were included in the analysis of the primary outcome (ie, the intention-to-treat principle), whether allocation was clearly concealed, the proportion of participants lost to follow-up, the magnitude of the original *P*-value reported, sample size, and the total number of events in the trial. We used linear regression models to evaluate associations between the Fragility Index (the dependent variable) and trial characteristics (the independent variables). For these models, we report the regression coefficients and the 95% CIs calculated by 1,000 bootstrap samples. We also determined the correlation between the Fragility Index and trial sample size and the total number of outcomes in the trial.

All calculations were performed using Stata MP, version 11 (StataCorp LP, College Station, TX, USA).

3. Results

We identified and reviewed the abstracts of 1,273 articles, of which 399 met our eligibility criteria (Appendix at www.jclinepi.com). Table 1 summarizes the included trials' characteristics. The median sample size was 682 patients (range: 15–112,604), with a median of 112 events (range:

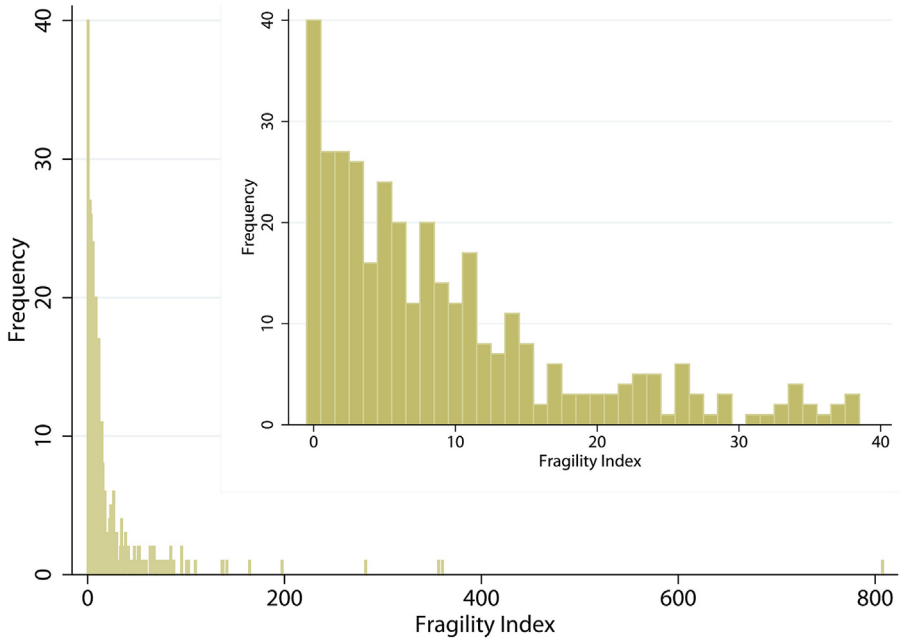

Fig. 2. Distribution of Fragility Index for all trials. Inset graph limited to trials with a Fragility Index ≤40.

8–5,142). The outcome evaluated was the primary outcome in 65.9% of the trials. Fifty-three percent of RCTs reported *P*-values <0.01. In 210 (52.6%) trials, all patients randomized were included in the analysis.

Fig. 2 reports the distribution of the Fragility Index for all included trials. The median Fragility Index was 8 (25th–75th percentile, 3–18; range, 0–808). Forty (10%) trials became nonsignificant when we applied the Fisher's exact test to their contingency table and therefore had a Fragility Index of zero. Of trials with a Fragility Index of zero, 28 (70%) were originally analyzed by a time-to-event analysis (of which, six were also adjusted for covariates), 3 (7.5%) used adjusted analyses, and the remaining 9 (22.5%) used the chi-square test, CI-based approaches, or imputed data for outcomes to determine statistical significance. One-quarter of the trials (100) had a Fragility Index less than or equal to 3. The sensitivity analysis using 95% CIs did not differ materially from the main analysis and demonstrated the same median and range of the Fragility Index for this sample of trials.

Table 2 reports the Fragility Index by subgroups of trial characteristics. All subgroups included trials with a Fragility Index of zero. Furthermore, the 25th percentile of the Fragility Index was ≤3 in all subgroups except for those with *P*-values <0.01, the half with the largest number of events (ie, >112 events), and the quarter with the largest sample size (ie, >2,522 participants). In linear regression, smaller reported *P*-values, larger numbers of events, and larger sample size were associated with less fragile results, and inadequate or unclear concealment was associated with more fragile results (Table 3). The linear regression did not identify differences in the Fragility Index on the basis of the outcome characteristics assessed (ie, primary outcome, time-to-event outcome, composite outcome). The Fragility Index was weakly correlated to total sample size ($r = 0.28$; $P < 0.001$) and moderately correlated with total number of events ($r = 0.64$; $P < 0.001$; Fig. 3).

In the 306 trials that clearly reported loss to follow-up, the median number of participants lost to follow-up was nine (25th–75th percentile, 2–39). The total number lost to follow-up exceeded the Fragility Index in 162 (52.9%) trials. When considering only the group in each trial with the fewest number of events (ie, the group to which the Fragility Index was added), the number of participants lost to follow-up still exceeded the Fragility Index in 132 (43.1%) trials.

4. Discussion

The Fragility Index was 3 or less in 25% of dichotomous outcomes from RCTs reported in high-impact general medical journals. Examples of trials with fragile results (ie, those with a small Fragility Index) were found across the distribution of sample sizes and number of events. Furthermore, in more than half the RCTs, more participants were lost to follow-up than would be required to make the result nonsignificant based on the corresponding trial's Fragility Index. Reporting the Fragility Index for statistically significant results may allow clinicians to draw appropriate inferences regarding their confidence in a putative treatment effect.

The magnitude of the *P*-value or the distance of the lower boundary of a CI from no effect, metrics that reflect potential random error, influences our beliefs in trial results. In addition, statistical significance may also be influenced by methodological limitations that increase systematic error (eg, losses to follow-up or inadequate blinding) [1]. Integrating

Table 2. Fragility Index by subgroups based on trial or outcome characteristics

Characteristic	Median Fragility Index (25th–75th percentile)
All trials (n = 399)	8 (3–18)
Outcome	
Primary (n = 263)	8 (3–21)
Not primary (n = 136)	7 (2–14)
Time to event (n = 206)	7.5 (3–17)
Not time to event (n = 193)	9 (3–21)
Composite (n = 132)	9 (3.5–22.5)
Not composite (n = 367)	7 (2–15)
Analysis	
Adjusted (n = 35)	6 (1–15)
Not adjusted (n = 364)	8 (3–19.5)
Intention to treat (n = 210)	8 (3–17)
Not intention to treat (n = 189)	7 (2–20)
Allocation concealment	
Adequate (n = 315)	8 (3–18)
Unclear or inadequate (n = 84)	10 (3–18)
Lost to follow-up	
≤1% (n = 144)	8 (3–17)
>1–5% (n = 88)	6 (3–13)
>5–10% (n = 34)	8 (2–14)
>10% (n = 40)	6 (2–19)
Not reported (n = 93)	10 (4–33)
P-value	
<0.05–0.01 (n = 186)	3 (1–9)
<0.01–0.001 (n = 168)	11 (5–21.5)
<0.001 (n = 45)	26 (11–47)
Number of events	
8–51 (n = 100)	3 (1–7)
52–112 (n = 100)	8 (3–15)
113–281 (n = 100)	9 (5–21)
282–5,142 (n = 99)	22 (6–52)
Sample size	
15–286 (n = 100)	4 (2–10)
287–682 (n = 100)	6 (3–16)
683–2522 (n = 100)	9 (2.5–18)
2523–112,604 (n = 99)	14 (6–52)

Table 3. Association between trial characteristics and the Fragility Index using linear regression

Characteristic	β Coefficient (95% CI)	*P*-value
Primary outcome	−8.8 (−23.2, 5.7)	0.23
Time-to-event outcome	−0.3 (−10.8, 10.1)	0.95
Composite outcome	5.7 (−7.8, 19.2)	0.41
Adjusted analysis	−2.7 (−15.5, 10.1)	0.68
Intention-to-treat analysis	−6.9 (−18.0, 4.2)	0.22
Allocation concealment unclear or inadequate	−9.8 (−17.6, −1.9)	0.02
Lost to follow-up		
≤1%	Referent	
>1–5%	4.6 (−7.5, 16.6)	0.46
>5–10%	−3.0 (−10.5, 4.6)	0.44
>10%	5.3 (−12.7, 23.2)	0.57
Not reported	15.9 (−2.7, 34.5)	0.09
Reported *P*-value		
<0.05–0.01	Referent	
<0.01–0.001	11.6 (4.0, 19.3)	0.003
<0.001	39.2 (4.9, 73.5)	0.03
Number of events		
8–51	Referent	
52–112	7.3 (4.5, 10.1)	<0.001
113–281	10.0 (6.7, 13.3)	<0.001
282–5,142	48.1 (27.7, 68.6)	<0.001
Sample size		
15–286	Referent	
287–682	6.6 (2.7, 10.6)	0.001
683–2,522	9.5 (4.0, 15.0)	0.001
2,523–112,604	39.5 (19.6, 59.3)	<0.001

Abbreviation: CI, confidence interval.

The β coefficient refers to the difference in the Fragility Index for trials with compared with without the characteristic or compared with the referent group.

these many factors is difficult. Work in cognitive psychology demonstrated that even experts in probability are intuitively poor in determining the relevance of these issues in the experimental design [2]. Many clinicians are unlikely to have substantial training in probability and statistics, and intuitive interpretation of *P*-values and CIs is likely limited. Treatment decisions often start with the decision of whether a treatment effect is believed to exist. A simple metric, such as the Fragility Index, may assist clinicians in determining the confidence they should have in the result.

The importance of fragility is underscored by the number of RCTs that initially reported statistically significant treatment effects but were later shown to be either ineffective (16%) or have effects that were substantially less than initially reported (16%) [3]. The only parameter associated with these findings is that the initial trial had a small sample size. However, even trials generally regarded as large may be fragile. Take, for example, the Leicester Intravenous Magnesium Intervention Trial (LIMIT-2), which tested the effect of intravenous magnesium on 28-day survival in patients with suspected acute myocardial infarction [4]. The trial was, by most accounts, large with 2,316 patients randomized, and it demonstrated a 24% relative risk reduction in mortality with a *P*-value of 0.04. Three years later, a trial of more than 58,000 patients demonstrated no benefit, and subsequent meta-analyses demonstrated that it was unlikely any true benefit exists [5]. The Fragility Index of the LIMIT-2 trial was 1. Had clinicians known this result hinged on one event, they may have been hesitant to believe and, therefore, act on it.

We are not the first group to suggest RCT results may be fragile. Pocock suggested that statistically significant results that required only a single event to change to cross the threshold of significance should be viewed with skepticism [7]. In a 1990 publication, Feinstein suggested a Unit Fragility Index that was computed by shifting one nonevent to an event in one group and one event to a nonevent in the second group [10]. Walter later expanded this concept and concluded that more empirical evidence was needed [11]. The concepts developed by Feinstein and Walter are similar to ours but more difficult to apply and interpret as they do not report the absolute number of events required to alter results (an intuitive metric) and they require decisions about the magnitude of effect that constitutes clinical significance.

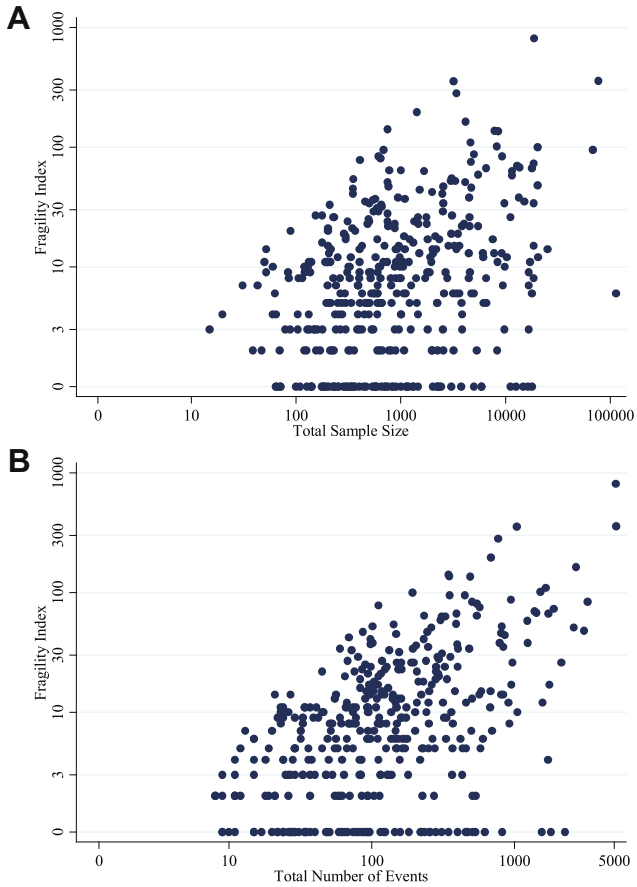

Fig. 3. Fragility Index by trial (A) sample size (all patients included in analysis) and (B) total number of events.

Reliance on *P*-values and their arbitrary thresholds to interpret study results has been criticized for several decades by statisticians and methodologists [6,7,8]. Despite this, *P*-values continue to be commonly reported and play an important part in the reporting and interpretation of RCTs. Given the *P*-value is unlikely to be wholly replaced by another statistic, addition of an equally simple metric such as the Fragility Index may improve our understanding of both trial results and *P*-values.

Others have suggested that the use of 95% CIs ameliorates the shortcomings of using *P*-values as a metric of statistical significance [9]. However, we believe that 95% CIs are typically thought of as dichotomous entities in that they either exclude the null value (signifying a statistically significant result) or they do not. The Fragility Index for *P*-values is consistent with the number of events required to change the bounds of a 95% CI from one that does not include the null value to one that does. Also, if one considers the example from the Introduction section in which a trial of 100 patients per group results in one vs. nine events, the relative risk reduction is 0.89 and the 95% CI is 0.14 to 0.99. The larger trial of 4,000 patients per group and 200 compared with 250 events results in a relative risk reduction of 0.20 and 95% CI of 0.04 to 0.33. Given the lower bound of the 95% CI is further from the null value in the small trial than the large trial, one may be more confident a treatment effect exists based on the results from the small trial. The Fragility Index of only 1, however, would suggest that caution is warranted in interpreting the results of the small trial. This suggests that 95% CIs suffer from many of the same limitations as *P*-values.

Yusuf et al. [12] suggested trials of at least 650 events were required to be sufficiently confident that true effects were identified in cardiovascular trials, and this was reinforced by simulation data [13]. Although this is appealingly simple, it primarily addressed the probability that a treatment effect that truly exists is missed by small trials (ie, statistical power). Furthermore, it is based on the concept of a risk in the control group of approximately 10% and small to moderate effect sizes. The Fragility Index addresses the issue of treatment effects that are detected but that may be unreliable. Both issues may, however, be largely a function of trials too small to reliably determine whether a treatment effect truly exists, and if it does, whether the estimated effect is likely close to the true effect. Although the requirement for 650 events may be the correct threshold for trials in many diseases/treatments, there are scenarios in which large treatment effects may require far less than 650 events for a robust assessment of the treatment. Conversely, there are scenarios in which there are more than 650 events, but the between-group difference is very small and therefore has a very small Fragility Index suggesting even 650 events may be insufficiently robust. The Fragility Index is useful in both these scenarios as shown in our data.

Although the Fragility Index can aid our understanding of trial results, it has limitations. It applies only to 1 to 1 randomization and only to binary data. Outcomes measured on continuous scales cannot, therefore, have a Fragility Index computed, and the Fragility Index will not benefit the interpretation of *P*-values in trials using continuous scales as outcomes. However, most patient-important outcomes are natural dichotomous (eg, death, stroke, myocardial infarction), and important continuous outcomes are frequently presented as dichotomies to aid interpretation. The use of the Fragility Index in time-to-event analyses may not always be appropriate. We found no material difference in the Fragility Index between time-to-event data and frequency data, which is consistent with the concept that most results are sensitive to the number of events in each group rather than the timing of the events. However, applying the Fragility Index to time-to-event data in which the numbers of events in both groups are similar but there is a clear difference in the timing of the events may be inappropriate and result in finding such trials inappropriately fragile. We were also limited to two-group comparisons, although our approach may be applied to pairs of groups in studies with more than two groups. These limitations aside, the Fragility Index will not identify all implausible or false-positive RCT results. However, the Fragility Index

has the merit that it is very simple and may help integrate concerns over smaller samples sizes and smaller numbers of events that are not intuitive.

We conclude that the significant results of many RCTs hinge on very few events. Reporting the number of events required to make a statistically significant result nonsignificant (ie, the Fragility Index) in RCTs may help readers make more informed decisions about the confidence warranted by RCT results.

Acknowledgments

The authors thank David L. Sackett for his advice in the development of the Fragility Index concept.

Contributors: M.W., David L. Sackett, G.G., and P.J.D. were responsible for the study conception. M.W., S.K.S., G.G., L.T., S.D.W., J.P., and P.J.D. contributed to study design. M.W., S.K.S., D.F.M., M.M., O.L., C.R., A.O.M., N.D.D., and A.B. contributed to data collection. M.W. performed analyses. M.W. and P.J.D. drafted the manuscript. M.W., S.K.S., D.F.M., M.M., O.L., C.R., A.O.M., N.D.D., A.B., G.G., L.H., S.D.W., J.P., and P.J.D. made critical revisions to the manuscript. M.W. is the guarantor. All authors approved the final version of the manuscript.

Appendix

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2013.10.019.

References

[1] Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. N Engl J Med 1979;301:1410−2.

[2] Kahneman D, Tversky A. On the psychology of prediction. Psychol Rev 1973;80(4):237.

[3] Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218−28.

[4] Woods KL, Fletcher S, Roffe C, Haider Y. Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2). Lancet 1992;339:1553−8.

[5] ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. Lancet 1995; 345:669−85.

[6] Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 1999;130:995−1004.

[7] Pocock SJ. Current issues in the design and interpretation of clinical trials. Br Med J (Clin Res Ed) 1985;290(6461):39−42.

[8] Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ 2001;322:226−31.

[9] Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol 1998;51:355−60.

[10] 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.

[11] Walter SD. Statistical significance and fragility criteria for assessing a difference of two proportions. J Clin Epidemiol 1991;44: 1373−8.

[12] 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(5):335−71.

[13] 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(10):e25491. [Electronic Resource].