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. © 2014 The Authors. Published by Elsevier Inc. All rights reserved.

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

[☆] This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

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/$ - see front matter © 2014 The Authors. Published by Elsevier Inc. All rights reserved.
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