# The fragility of significant results underscores the need of larger randomized controlled trials in nephrology

see commentary on page 1319

Lani R. Shochet[1], Peter G. Kerr[1,2] and Kevan R. Polkinghorne[1,2,3]

[1]Department of Nephrology, Monash Medical Centre, Monashhealth, Clayton, Victoria, Australia; [2]Department of Medicine, Centre for Vascular Health, Monash University, Clayton, Victoria, Australia; and [3]Department of Epidemiology and Preventive Medicine, Monash University, Prahran, Victoria, Australia

**The Fragility Index is a tool for testing robustness of randomized controlled trial results for dichotomous outcomes. It describes the minimum number of individuals in whom changing an event status would render a statistically significant result nonsignificant. Here we identified all randomized controlled trials in five nephrology and five general journals from 2005-2014. A total of 127 randomized controlled trials reporting at least one dichotomous statistically significant outcome (p less than 0.05) were included and the Fragility Index was calculated. Twenty randomized controlled trials had a Fragility Index of zero and were excluded from further analysis. Linear regression was performed to assess factors associated with Fragility Indexes stratified by primary or secondary outcomes. The median sample size was 134 (range 2211506) with 36 (range 5–2743) total number of events. The median Fragility Index was three (range 1–166), indicating that in half the trials the addition of three events to the treatment with the lowest number of events rendered the result nonsignificant. For primary outcome studies a doubling in total event number and sample size significantly increased the geometric mean Fragility Index by 52% and 42%, respectively. Compared to a reported p value of 0.05 to 0.01, those reporting 0.01 to 0.001 or less than 0.001 had a significant 57% and 472% increase in the median Fragility Index, respectively. Forty-one percent had a Fragility Index less than the total loss to follow-up, indicating a potential to change a trial result had all individuals been accounted for. Thus, our study highlights the need for larger randomized controlled trials with accurate accounting for loss to follow-up to adequately guide evidence-based practice.**

*Kidney International* (2017) **92,** 1469–1475; http://dx.doi.org/10.1016/j.kint.2017.05.011

KEYWORDS: fragility score; nephrology; randomized controlled trials



**Correspondence:** *Kevan R. Polkinghorne, Department of Nephrology, Monash Medical Centre, 246 Clayton Road, Clayton, Melbourne, Victoria 3168, Australia. E-mail: kevan.polkinghorne@monash.edu*

**Received 24 November 2016; revised 17 April 2017; accepted 4 May 2017; published online 26 July 2017**

The efficacy of interventions aimed at improving health outcomes in a population are typically assessed by randomized trials. Statistical tests are applied to assess whether the effect of an intervention in a randomized controlled trial (RCT) is significant, which is arbitrarily reported as a *P* value of $<0.05$ (or a 95% confidence interval that does not contain the null hypothesis value). The *P* value is defined as the probability of obtaining a result that is equal to, or more extreme to that which was actually observed under the assumption of no effect or no difference (the "null hypothesis"). It measures how likely the observed differences seen with an intervention between 2 or more groups are due to chance.

To consider the reported *P* value in context, factors influencing it must be understood. The intervention or variables' effect size, the size of the sample, and the spread of the data (SD) will all affect the *P* value.[1] Importantly, statistical significance often does not translate to clinical importance. For example, in a study with a large sample size, even very small differences that are not of clinical importance can be highly statistically significant.

The Fragility Index (FI) is a method of assessing reported significant RCT results to provide further context to their interpretation.[2] The FI describes the minimum number of patients in whom changing (or "reassigning") an event status would alter a statistically significant result to a nonsignificant result. For example, a score of 3 implies that if the event status of 3 trial participants in the intervention arm with the fewest number of participants were different, the reported trial result would no longer be considered statistically significant when using the conventional *P* value cut-off of $<0.05$.

Compared with other specialties, RCTs are not as common in nephrology and are of low reported quality.[3] Given the importance placed on RCT results in the hierarchy of evidence to inform the development and conclusions of clinical practice guidelines, an assessment of the fragility of trial results in nephrology is warranted. We therefore aimed to assess the FI of reported RCT results in nephrology over a recent 10-year period. We hypothesized that a high proportion of RCTs in nephrology would be fragile, given the frequency of studies with small sample sizes, small event numbers, and large loss to follow-up, due to the nature of the study populations.

## RESULTS

### Characteristics of included trials

Of the 1233 potential studies identified, 127 met the inclusion criteria (Supplementary Table S1). Of these, 110 were published in nephrology journals with the remaining 17 spread among 4 general medical journals. Twenty trials had a calculated FI of 0, indicating that the trial result could not be reproduced using the Fisher exact test. Reasons for this varied; however, the majority were related to the original trial results by being a time-to-event analysis ($n = 11$), an adjusted analysis ($n = 2$), or an analysis of multiple events per participant ($n = 2$). Of the final 4 studies, 3 reported significant results using the chi-squared test that were not significant using Fisher exact test, and 1 study reported results of a 1-sided $P$ value chi-squared test as opposed to a 2-sided test (not specifically stated in the trial report). These 20 trials were excluded from further analysis.

Table 1 details the broad characteristics of the 107 included trials stratified by primary and secondary outcomes (Supplementary Table S1 lists details of the included individual RCTs). The median sample size was 134 (range 22–11,506) with the median total number of events 35 (range 5–2743, 75% of studies had <80 events). The majority of trials were blinded and analyzed the data by intention to treat; however, only a minority had adequate allocation concealment. Eighty-nine percent were published in specialist nephrology journals (94 of 107). Seventy-six percent of reported $P$ values were between 0.05 and 0.001. There were no statistically significant differences in trial characteristics according to whether the first reported positive outcome was primary or secondary.

**Table 1 | Summary of included trials stratified according to whether the reported positive outcome was primary or secondary**

| Characteristic | All studies ($N = 107$) | Primary outcome[a] ($n = 71$) | Secondary outcome[a] ($n = 36$) |
|---|---|---|---|
| Sample size | 134 (22, 11,506) | 132 (22, 9270) | 158 (26, 11,506) |
| Total number of events | 35 (5, 2743) | 40 (5, 1145) | 24 (8, 2743) |
| Allocation concealment | | | |
| None or unknown | 71 (66) | 47 (66) | 24 (67) |
| Yes | 36 (34) | 24 (34) | 12 (33) |
| Intention-to-treat analysis | | | |
| Yes | 80 (75) | 55 (77) | 25 (69) |
| No or unclear | 27 (25) | 16 (23) | 11 (31) |
| Blinding to treatment | | | |
| Yes | 76 (71) | 52 (73) | 24 (67) |
| No or unclear | 31 (29) | 19 (27) | 12 (31) |
| Journal | | | |
| *JASN* | 23 (22) | 16 (23) | 7 (19) |
| *cJASN* | 10 (9) | 5 (7) | 5 (14) |
| *AJKD* | 23 (22) | 15 (21) | 8 (22) |
| *NDT* | 23 (22) | 14 (20) | 9 (25) |
| *KI* | 15 (14) | 13 (18) | 2 (6) |
| General medical journals[b] | 13 (12) | 8 (11) | 5 (14) |
| Reported $P$ value | | | |
| 0.05 to >0.01 | 39 (36) | 22 (31) | 17 (47) |
| 0.01–0.001 | 43 (40) | 32 (45) | 11 (31) |
| <0.001 | 25 (23) | 17 (24) | 8 (22) |

AJKD, *American Journal of Kidney Diseases*; cJASN, *Clinical Journal of the American Society of Nephrology*; JASN, *Journal of the American Society of Nephrology*; KI, *Kidney International;* NDT, *Nephrology Dialysis Transplantation*.
Values are median (range) or n (%).
[a]No significant differences between primary and secondary outcomes studies for all comparisons $P > 0.05$.
[b]*New England Journal of Medicine* ($n = 7$), *The Lancet* ($n = 5$), *Journal of the American Medical Association* ($n = 3$), and *Annals Internal Medicine* ($n = 2$).

### Calculated FI of included trials

Figure 1 shows the distribution of the calculated FI and Table 2 summarizes the calculated FI of the included 107 studies grouped according to trial characteristics. The median FI was 3 (range, 1–166). Furthermore, 22% of trials had an FI of 1, indicating that the significance of the outcome was dependent on the event status of 1 participant.

The number lost to follow-up was not reported in 16 trials (15%); in the remainder, an average of 11 patients were lost to follow-up (median 1, range 0–317). In 31% of the trials with lost–to–follow-up data, the FI was less than the total number of participants lost to follow-up, indicating potential to change a trial result had all subjects been accounted for in the study and the FI was less than the "expected number of events" that could have occurred.

### Predictors of FI stratified by primary or secondary outcome

Figure 2 shows scatterplots of study sample size and event number versus FI stratified by whether the reported significant trial finding was a primary or secondary outcome. Results of the linear regression β coefficients are presented as analyzed on the log scale in Supplementary Table S2 and following transformation back from the log scale in Table 3. For each unit increase in a categorical and doubling of a continuous variable, Table 3 displays the percentage change in geometric mean (GM) fragility score and the 95% confidence interval (CI).

The relationship between total number of events and sample size was similar for both primary and secondary outcome studies with both significantly related to FI. For example, a doubling in the sample size of a primary outcome study was associated with a 42% increase in the GM of the FI (95% CI: 23–64; $P < 0.001$). Similarly $P$ values between 0.01 to 0.001 and those <0.001 were significantly associated with increased FI compared to $P$ values 0.05 to >0.01 for primary outcome studies and those <0.001 for secondary outcome studies.

When assessing study quality metrics, differences were seen depending on whether the study outcomes were primary or secondary. Adequate treatment allocation concealment was significantly associated with increased FI in primary outcome studies (113% change in GM FI, 95% CI: 28–253; $P = 0.004$) but not secondary studies. Primary outcome studies reporting on treatment analysis compared with intention to treat had lower FI; however, this was of borderline statistical