Over-reliance on *P* Values in Urology: Fragility of Findings in the Hydronephrosis Literature Calls for Systematic Reporting of Robustness Indicators

Mandy Rickard, Armando J. Lorenzo, Jessica H. Hannick, Anne-Sophie Blais, Martin A. Koyle, and Darius J. Bägli

OBJECTIVE To review the robustness of hydronephrosis literature with the application of fragility index (FI) and fragility quotient (FQ) calculations.

METHODS A literature review was conducted using Pubmed, Medline, and Ovid for “hydronephrosis” and associated terms and we included all studies with at least 2 groups being compared. FI was calculated by populating study results into a 2-by-2 contingency table and generating a *P* value using Fisher’s exact test. Next, events were manually added to the group with the fewest events, while removing a nonevent from the same group and Fisher’s exact test repeated until the *P* value was >.05. FQ was calculated by dividing FI by the total sample size.

RESULTS The 130 included articles were published between 1986 and 2018 in 32 journals. Median citation count was 14 (0-252), 30% were RCTs and most papers originated in the United States (28%), Turkey(10%), and Canada(9%). Median FI was 2 (1-112), FQ was 0.023 (0.0010-0.55), and 60 papers (46%) had a FI of 1, indicating extremely fragile results. There was a significant difference in the FI between observational studies and RCTs (10 ± 17 vs 4 ± 5; *P* = .02); however, there was no difference in FQ (0.032 ± 0.030 vs 0.053 ± 0.080; *P* = .09) between them.

CONCLUSION Nearly half of studies in hydronephrosis literature reporting significant results are extremely fragile, requiring addition of only a couple of events in 1 treatment arm to significantly modify the results. As such, objective reporting of robustness of results should include FI and FQ which may help diminish over-reliance on *P* values as the main indicator of clinical significance in comparative studies. UROLOGY 133: 204–210, 2019. © 2019 Elsevier Inc.

When appraising the scientific literature, it is has been customarily accepted that a reported *P* value of <.05 indicates a significant finding, solely because the probability of the difference found between groups is above an arbitrary threshold of what could be encountered by chance alone.[1] As clinicians who are encouraged to practice evidence-based medicine, we are required to consume an exponentially growing body of literature, critically examine the results, and decide whether these warrant implementation into patient care or justify making a change in care pathways. Prospective observational and experimental studies—particularly randomized controlled trials (RCTs)—reporting significant differences between groups or interventions are considered models of high levels of evidence, therefore more likely to impact medical decision-making. A common example in Urology is the widespread use of continuous prophylactic antibiotics for vesicoureteral reflux (VUR), a time-honored practice that was ratified by the RiVUR trial, which has been widely quoted as clear evidence of benefit to their use.[2]

While commonplace, relying on *P* value interpretation to declare clinically-relevant differences between groups is perhaps one of the most common oversimplifications incurred during critical examination of a study. Compounding the problem is the quest for significance, sometimes even at the expense of methodological requirements to accurately select and carry out statistical analyses. Rather than a malicious process seeking to report spurious associations, this phenomenon may result from investigators lacking advanced biostatistical knowledge, coupled

From the Division of Urology, Hospital for Sick Children and Department of Surgery, University of Toronto, Toronto, Ontario, Canada; and the Division of Pediatric Urology, UH Rainbow Babies and Children’s Hospital, Cleveland, Ohio, USA

Address correspondence to: Mandy Rickard, M.N., N.P., Division of Urology, Hospital for Sick Children and Department of Surgery, University of Toronto, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada. E-mail: mandy.rickard@sickkids.ca

Submitted: January 30, 2019, accepted (with revisions): March 4, 2019

https://doi.org/10.1016/j.urology.2019.03.045
0090-4295

with consumers relying on authors, editors, and reviewers to interpret the results.[3] When interpreting a significant *P* value, readers may not routinely consider sample sizes, power analyses, or measures of effect (such as relative risks, odds ratios, risk differences, and 'number needed to treat')[4] to appraise *how* impactful the results truly are. For example, would readers accept significant results with more skepticism if they knew that, while "significant", the results of the study could be completely different if the outcome of interest occurred slightly more or less often in one of the treatment arms?

Enter the fragility index (FI), a relatively new measure aimed at objectively capturing this phenomenon. In essence, it's a calculation used to determine how many events would have to occur in an alternative treatment arm to change the *P* value from <.05 to >.05.[5-7] One drawback of the FI is that it can often be misleading in studies with large sample sizes, thus the fragility quotient (FQ) was subsequently introduced to adjust for this factor.[8] Herein, we evaluate the hydronephrosis (HN) literature, including all variations of HN and patient populations, to determine the robustness (or fragility) of this pool of literature with respect to FI. Furthermore, we also explore the utility of the FQ, an emerging measure in the medical literature and a relatively novel concept in the Urology specialty.

METHODS

A comprehensive literature review was conducted using PubMed, Medline, and Ovid for "hydronephrosis" and associated terms including: hydroureteronephrosis, urinary tract dilatation, megaureter, pelviectasis, caliectasis, ureteropelvic junction obstruction (UPJO), and ureterovesical junction obstruction. The search was limited to English language papers, human studies, and full text, excluding conference abstracts, basic science, and animal studies. We focused on papers reporting comparison of at least 2 groups, including clinical trials, observational, and cohort studies. This search resulted in 5892 relevant references that were uploaded into Covidence online systematic review software (www.covidence.org; Melbourne, Australia). After removal of 1504 duplicated references, 4388 remained for title and abstract screening. This was independently completed by 2 experienced urology clinicians, with conflicts being resolved by consensus.

Criteria used for selecting appropriate articles included the following: HN or related conditions known to cause HN (such as posterior urethral valves, stones, or VUR), the presence of at least 2 groups for comparison, a significant difference being reported between groups, and dichotomous variables available as outcomes. Of the titles and abstracts screened, 4045 failed to meet these eligibility requirements, which left 343 articles. Upon full text review, 130 appropriate articles remained for data extraction (Fig. 1).

During data extraction the following variables were collected: year of publication, name of publication, journal impact factor, citation count, *h*-index of corresponding author, country of origin, and methodological characteristics of each paper. The FI was calculated for each study following a previously described technique.[7,9] Briefly, the results of each study were populated into a 2-by-2 contingency table and a *P* value generated using Fisher's exact test. Next, hypothetical events (outcomes of interest from each study) were manually added to the group with the fewest number, while removing a nonevent from the same group, and Fisher's exact test repeated until the *P* value was >.05 (Table 1). The number of events required to make the *P* value transgress the .05 threshold is the FI[7]; the higher the value, the more robust the results of the study. We also specifically captured the number of individuals lost to follow-up (defined as subjects included in the study but no follow-up data could be obtained), as counts that exceed the calculated FI warn that the study's results are extremely fragile due to the possibility that events of interest occurred to these lost patients.

While the FI generates an objective measure of robustness, it is an absolute measure that fails to take into account study sample size, predisposing readers to infer that higher FI automatically means more robust results. For this reason, the FQ was also calculated for each trial. This ratio is a relative measure of the FI with consideration of the sample size of the study, calculated by dividing FI by the total sample size.[8]

Demographic characteristics of the included studies were presented as means (standard deviation), medians (range: minimum, maximum), counts, and percentages. We performed subgroup analyses to compare characteristics of RCTs vs observational studies. Results were considered statistically significant when *P* <.05. Data were analyzed using SPSS version 22.0 (SPSS Inc., Chicago, IL).

RESULTS

Characteristics of Included Studies

The 130 included articles were published between 1986 and 2018 in 32 journals, primarily the *Journal of Urology* (29%), *Journal of Pediatric Urology* (19%), *Urology* (12%), *Journal of Endourology* (5%) and *Pediatric Nephrology* (3%) (Fig. 2). The median citation count was 14 (0-252), *h*-index of corresponding author was 14 (0-83) and most papers originated in the United States (28%), Turkey (10%), Canada (9%), India (8%), and Sweden (5%) (Supplementary Fig. 1).

The majority of the included articles focused on pediatric patients (72%), and 24% were adult-based with the remaining 3% including both adults and children. Retrospective studies comprised most of the articles (56%) and only 30% of studies were classified as RCTs. While all studies pertained to HN, the main topics were VUR (36%), ureteropelvic junction obstruction (29%), stones (19%), and duplex anomalies (ie, ureteroceles) (6%). The median sample size was 105 (10-8543). The main outcomes of papers discussed surgical complications (36%), resolution of HN (32%), urinary tract infections (12%), and postoperative pain management (10%); 72% of studies reported these as their primary outcomes. Losses to follow-up were not reported in 81% of studies, greater than FI in 14%, and less than FI in 5%. Characteristics of the included papers can be reviewed in Supplementary Table 1.

Characteristics of RCTs

Of the 130 included articles, 39 (30%) were classified as RCTs. The characteristics of these RCT papers were similar to the overall paper characteristics including publishing journals, pediatric populations, topics discussed, and outcomes of interest. The countries of origin were different with most studies originating from Sweden (18%), India (15%), Turkey (10%), and the United States (8%). Most of the RCTs (54%) described the

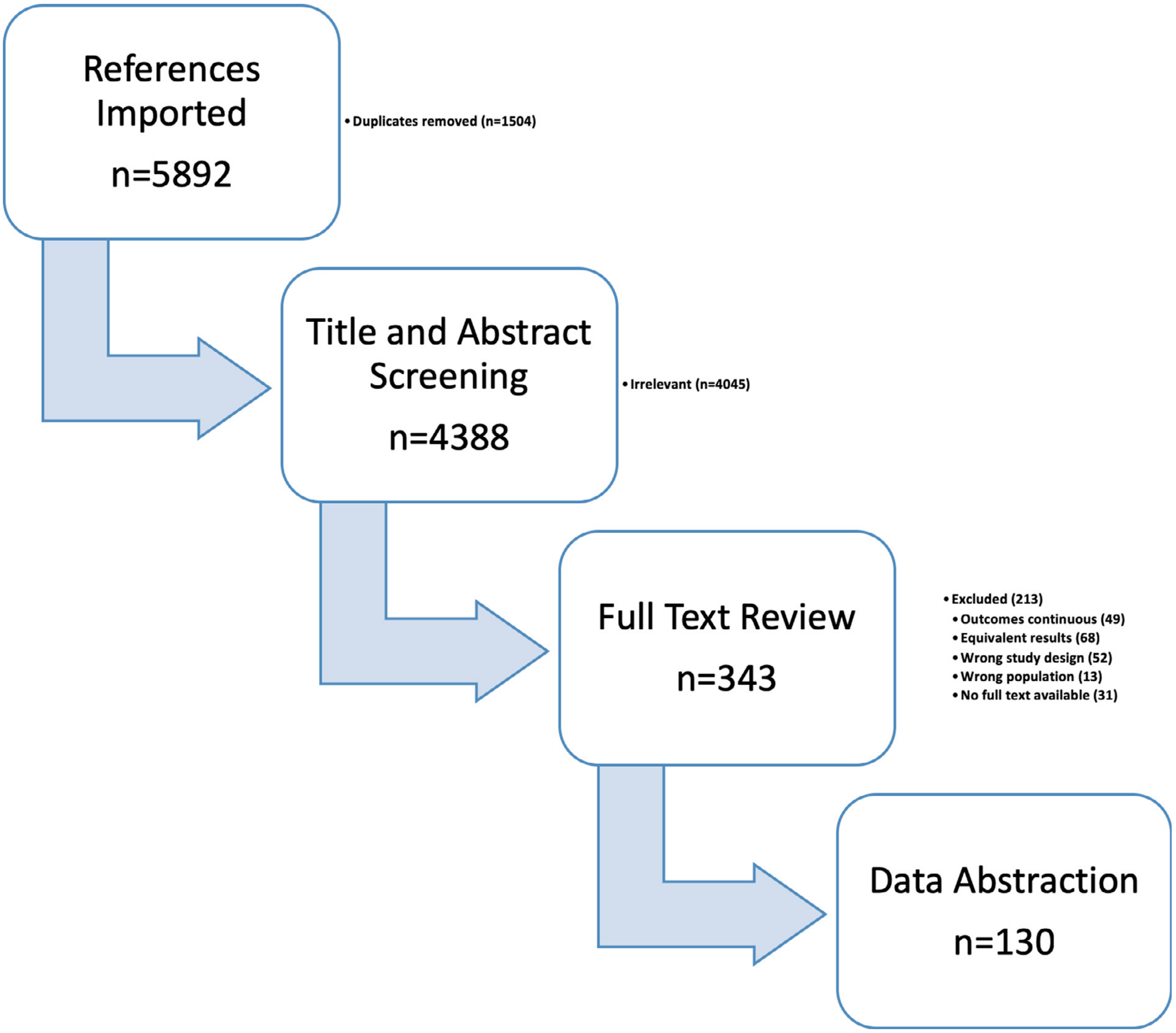
References Imported
n=5892
• Duplicates removed (n=1504)
Title and Abstract Screening
n=4388
• Irrelevant (n=4045)
Full Text Review
n=343
• Excluded (213)
• Outcomes continuous (49)
• Equivalent results (68)
• Wrong study design (52)
• Wrong population (13)
• No full text available (31)
Data Abstraction
n=130

Figure 1. Flow diagram of included studies. (Color version available online.)

method of randomization, however only 23%, 20%, and 31% described allocation concealment, blinding, and sample size justification, respectively. Most studies failed to report losses to follow-up (51%), and of those that did, 33% reported losses that were greater than the calculated FI. When comparing RCTs to observational studies, we found no difference in *h*-index of authors (18 ± 14 vs 18 ± 13; *P* = .88) and sample sizes (396 ± 1070 vs 137 ± 117; *P* = .13), respectively. However, we did note significant differences when comparing citation counts (24 ± 31 vs 41 ± 61; *P* = .04) and impact factor of publishing journals (3.11 ± 1.93 vs 5.70 ± 12; *P* = .05) for observational studies and RCTs, respectively.

Table 1. Method for calculating FI: Assuming Study Result group B reported a higher number of events

	Study Result		Fragility	
	Outcome	No Outcome	Outcome	No Outcome
Intervention A	A	B	A + *x*	B − *x*
Intervention B	C	D	C	D
	P<.05		*P*>.05	

Fragility Index = the smallest value of *x* required to change *P* <.05 to *P* >.05.
Fisher's exact test from <0.05 to >0.05.

Fragility Index and Fragility Quotient

The median FI of all articles was 2 (1-112) and FQ was 0.023 (0.0010-0.55) (Fig. 3). A total of 60 (46%) papers had a FI of 1, indicating extremely fragile results where a single event occurring in the group with the fewest events would be enough to show no difference between the groups. Overall, 65% of papers had a FI of 1-4, 10% had a FI of 5-9, and 25% had >10. The paper with the highest FI was by Legemate et al, "Characteristics and outcomes of ureteroscopic treatment in 2650 patients with impacted ureteral stones[10]" published in *World Journal of Urology* with an FI = 112, however the FQ was only 0.013. The median sample size of these papers was 88 (16-1000), FQ was 0.011 (0.001-0.063), and 57% were retrospective studies.

We noted a significant difference in the FI between observational studies compared to RCTs, with RCT results being more fragile with a lower FI (4 ± 5 vs 10 ± 17 for observational studies; *P* = .02), however there was no statistically significant difference in FQ (0.032 ± 0.030 vs 0.053 ± 0.080; *P* = .09). There

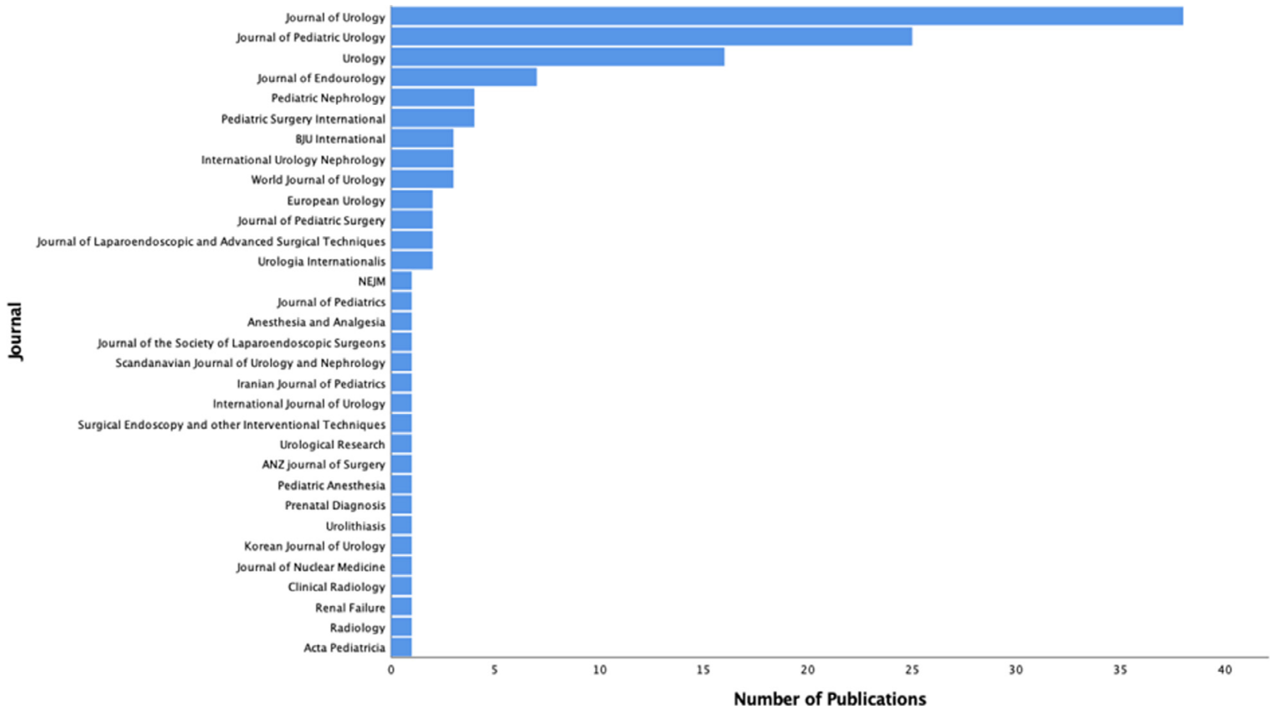

Figure 2. Number of publications by journal. (Color version available online.)

was also a difference in both FI and FQ between prospective and retrospective designs, as the FI for prospective studies was 4 ± 6 vs 11 ± 19 for retrospective (*P* <.01) and FQ was 0.032 ± 0.029 vs 0.057 ± 0.084 (*P* = .04) for prospective and retrospective studies, respectively. A total of 10 studies reported biostatistician support, however there was no difference in FI (10 ± 15 vs 8 ± 15; *P* = .72) or FQ (0.028 ± 0.037 vs 0.048 ± 0.70; *P* = .37) between those that did vs those that did not. The median year

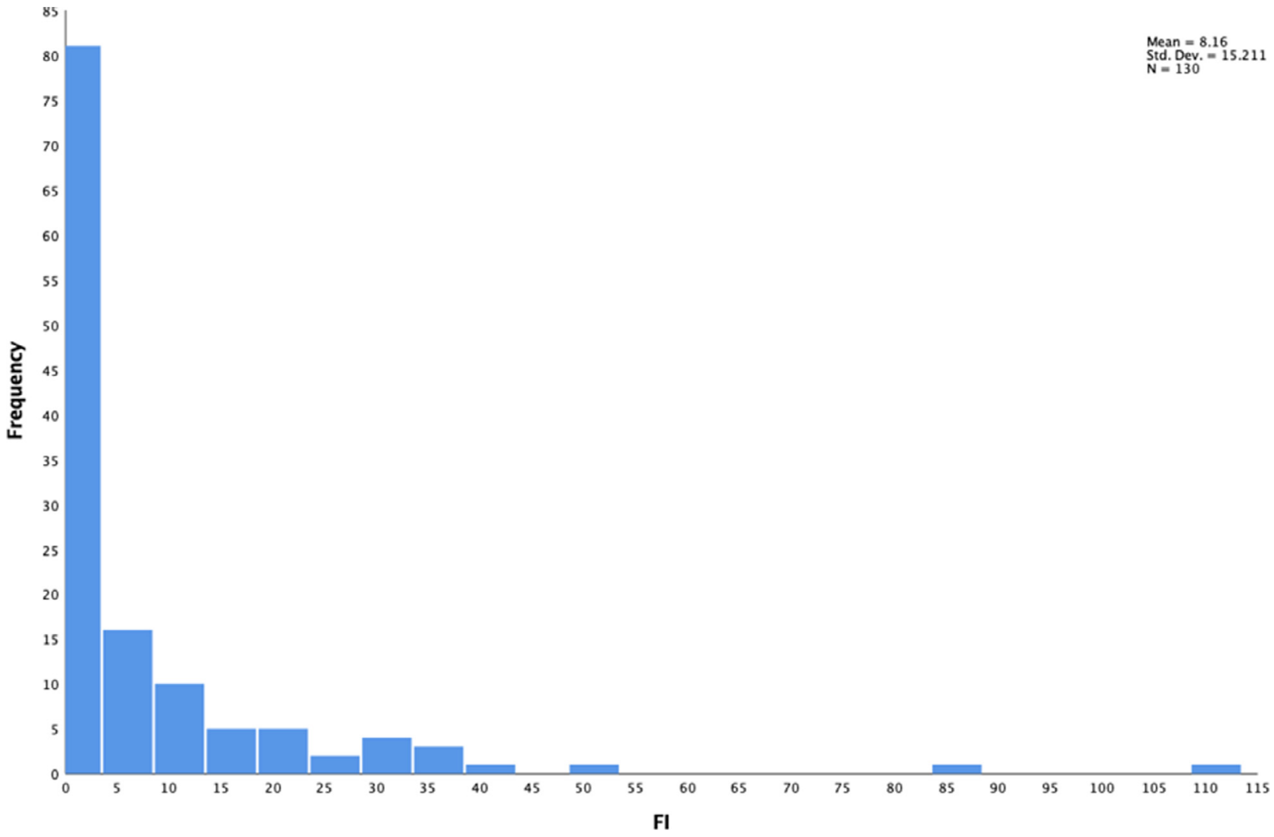

Figure 3. Histogram of FI scores. (Color version available online.)

of publication was 2010. We found no difference between papers published before and after 2010 for both FI (6 ± 8 vs 10 ± 19; $P = .11$) and FQ (0.050 ± 0.054 vs 0.044 ± 0.081; $P = .67$). We noted a significant positive correlation between FI and sample size ($r = 0.72$; $P <.01$) and FI and FQ ($r = 0.32$; $P <.01$).

DISCUSSION

In recent years there has been growing concern surrounding the indiscriminate reliance on P values, how they are reported and misinterpreted in medical and scientific literature, and the unfounded tendency to assume that a value less than the arbitrary time-honored cut-off of 0.05 indicates a clinically relevant difference between treatment arms.[11] The oversimplification of complex statistical calculations to a single number has prompted the American Statistical Association to develop a position statement on P values, cautioning investigators about exaggerating its meaning during their interpretation.[3] They acknowledge that many authors of scientific literature as well as those that consume it do not have formal training in statistics, which can lead to unintentional errors in calculating, reporting, and interpreting the results of these sophisticated analyses. Other papers have cautioned against the sole reliance on P values to draw clinically valid conclusions, suggesting that this misuse is partly responsible for the lack of reproducibility seen in some clinical trials.[12] It has been suggested that reporting confidence intervals in addition to P values would improve the interpretation of studies with significant results and improve understanding of "*how significant*" results actually are in literature that is used to inform clinical practice.[1,12]

This subject becomes more complicated when focusing on the irreproducibility of some study results, which has prompted Benjamin et al to propose that the significance level for scientific studies aiming to demonstrate differences between treatments be changed to <0.005.[13] They proposed that redefining the level of significance could help ensure that reported significant differences are indicative of true differences between groups and increase the likelihood of reproducibility. This proposed recalibration would be intended for initial trials testing interventions, with the aim of showing superiority over an alternative treatment. For subsequent observational studies confirming prior results, the traditional significance level of <0.05 could be preserved. Ultimately, whether other metrics are employed and/or significance level is adjusted, the current state of scientific literature remains the same. Standardized reporting criteria or changing of the P value does not address the main challenge: despite statistical significance between groups, FI begins to quantity whether a result may be of questionable clinical significance or relevance.

The concept of P value fragility was introduced in the 1990s, and has been explored by many medical specialties as a tool to appraise the quality of evidence and robustness of study results.[9,14-19] In the present study, we encountered findings in line with other studies,that is that the conclusions often hinge on a small number of events required to completely change the results. For example, Matics et al reviewed all clinical trials in high impact pediatric journals and found the median FI was 7, with losses to follow-up exceeding this number in 41% of studies.[16] In contrast to our study, they found no correlation between sample size and FI which could be explained by the small numbers of included papers as well as the known challenges with recruiting patients into trials.[20,21] Narayan et al reviewed urology RCTs in high impact journals and found the median FI to be 3 with 67% of studies reporting losses to follow-up greater that the FI.[15] Similarly, Shochet et al reviewed nephrology trials and found that the median FI was 3% and 41% of those studies also reported losses to follow-up greater than the FI. Such striking similarities in the overall lack of robustness of trials across various specialties supports the argument that the foundation of the medical literature and evidenced-based practice rely on extremely fragile results.

More recently, the concept of FQ has been introduced as a means to add context to the growing trend of FI reporting. While the FI is an absolute measure of robustness, the FQ generates a relative measure that adjusts the metric by taking into consideration the sample size of the study. Multiplying the FQ by 100 results in a calculation of the number of patients out of 100 that would be required to nullify the results of a study[22] (a FQ of 0.032 indicates that 3 patients out of 100 would be required to change the P value from <.05 to >.05). In the present study the median FQ was 0.023 indicating that on average 2/100 patients would be required to nullify significance. When considering the study with the highest FI in our series (FI = 112), the sample size was 8543, resulting in a FQ of 0.013. Applying this calculation would suggest that 1.3 patients out of 100 experiencing the alternative outcome would make these results nonsignificant. This finding is similar to Checketts et al who reviewed orthopedic literature and reported median FI of 2 and FQ of 0.022.[22] Tignanelli et al reviewed trauma trials and found a median FI of 3 and FQ of 0.016.[23] Overall, the robustness of the HN literature in our study is similar to the overall quality of the examined medical literature. This raises concerns regarding the value of evidence based on statistical significance alone and fuels the need for better strategies regarding study design, analysis, and declaring clinical significance.

We acknowledge that the literature cited herein concentrated on RCTs, and that FI and FQ were intended to measure the robustness of clinical trials, yet we expanded their use to other study designs in the present study. However, the concept of using the FI for analyzing observational studies has been previously explored.[24] We found a difference between FI when comparing observational studies and RCTs, with observational trials having a higher FIs. However, this could be explained simply by the higher numbers of observational studies compared to RCTs in our series, with this difference being less obvious had more clinical trials or less observational studies been included. Interestingly, there was no difference between observational studies and RCTs when considering FQ and many other clinical variables. RCTs are considered the gold standard of research methodologies with the assumption that randomization minimizes bias and balances confounders, therefore generating high-levels of evidence, supporting cause-and-effect conclusions, and better guiding of clinical practice. However, in the present study, important RCT characteristics were often overlooked or unreported, including description of randomization processes, allocation concealment, blinding and reporting of losses to follow-up, and suggesting suboptimal methodology. These problems have generated discussion about the utility of RCTs in medical research, questioning its value particularly when considering the associated difficulties with carrying them out.[25-27] Further, in designing RCTs, sample size calculation is carefully carried out with the intention of producing a difference between groups with P <.05 used as the threshold for this difference. As a result, properly calculated sample sizes may preclude large FIs which could arguably be considered unnecessary in a well powered trial. This may also further support the FQ concept which should be reflective of study robustness and accurate sample size calculation. As such, we propose that applying the concepts of FI and FQ to all study methodologies is a novel and interesting means of exploring the quality of all scientific literature.

The current review is not without its limitations. First, as discussed above, we have extended the FI and FQ concepts to study methodologies other than RCTs. Despite this, we feel that the concept of P value fragility can add transparency to routine reporting of comparative studies irrespective of their design. Next, we limited our search strategy to articles in English which may have eliminated appropriate articles demonstrating more robust study results reported in other languages. Finally, while not limited to our study, the current inability to calculate a FI for studies reporting equivalent results excluded an important number of studies. This is also true for papers reporting outcomes as continuous variables. Expanding FI methodology to these types of studies and outcomes would be beneficial to implementing FI and FQ in routine reporting in the medical literature.

Despite these limitations, we feel there are strengths to the present study. To our knowledge, this is the first study to explore the concept of FI in developmental urology and FQ in urology more broadly. We demonstrate that the robustness of the HN literature is similar to that of fragility assessments of most other such studies from other specialties. Next, we have demonstrated that when compared to studies employing other methodologies, RCTs have similar fragility of results to observational studies, suggesting the need for improvement in the quality of all studies, both observational and experimental.

CONCLUSION

We have demonstrated that many studies in the HN literature that report statistically significant differences between groups are extremely fragile and require only a few events in the alternative treatment arm to invalidate the results. This appears to be in line with similar reports from other medical and surgical specialties, indicating that evidence-based decision-making is often based on fragile findings. As such, in order to improve transparency and objective reporting, researchers should consider implementing FI and FQ in routine reporting of studies.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.urology.2019.03.045.

References

1. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337–350.
2. Hoberman A, Greenfield SP, Mattoo TK, et al. Antimicrobial prophylaxis for children with vesicoureteral reflux. *N Engl J Med*. 2014;370:2367–2376.
3. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129–133.
4. Tripepi G, Jager KJ, Dekker FW, et al. Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat. *Kidney Int*. 2007;72:789–791.
5. Feinstein AR. The unit fragility index: an additional appraisal of "statistical significance" for a contrast of two proportions. *J Clin Epidemiol*. 1990;43:201–209.
6. Grossman S, Zerilli T. Health and medication information resources on the world wide web. *J Pharm Pract*. 2013;26:85–94.
7. Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility index. *J Clin Epidemiol*. 2014;67:622–628.
8. Trimmel H, Voelckel WG. The authors reply. *Crit Care Med*. 2016;44:e1141–e1142.
9. Ridgeon EE, Young PJ, Bellomo R, et al. The fragility index in multicenter randomized controlled critical care trials. *Crit Care Med*. 2016;44:1278–1284.
10. Legemate JD, Wijnstok NJ, Matsuda T, et al. Characteristics and outcomes of ureteroscopic treatment in 2650 patients with impacted ureteral stones. *World J Urol*. 2017;35:1497–1506.
11. Braga LH, Lorenzo AJ, Suoub M, et al. Is statistical significance sufficient? Importance of interaction and confounding in hypospadias analysis. *J. Urol*. 2010;184:2510–2515.
12. Lu Y, Belitskaya-Levy I. The debate about p-values. *Shanghai Arch Psychiatry*. 2015;27:381–385.
13. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2:6–10.
14. Bowers A, Meyer C, Tritz D, et al. Assessing quality of randomized trials supporting guidelines for laparoscopic and endoscopic surgery. *J Surg Res*. 2018;224:233–239.
15. Narayan VM, Gandhi S, Chrouser K, et al. The fragility of statistically significant findings from randomised controlled trials in the urological literature. *BJU Int*. 2018;122:160–166.
16. Matics T, Khan N, Jani P, et al. The fragility index in a cohort of pediatric randomized controlled trials. *J Clin Med*. 2017;6:79.
17. Carter RE, McKie PM, Storlie CB. The fragility index: a p -value in sheep's clothing? *Eur Heart J*. 2016;38:346–348. ehw495.
18. Shochet LR, Kerr PG, Polkinghorne KR. The fragility of significant results underscores the need of larger randomized controlled trials in nephrology. *Kidney Int*. 2017;92:1469–1475.
19. Chase Kruse B, Matt Vassar B. Unbreakable? An analysis of the fragility of randomized trials that support diabetes treatment guidelines. *Diabetes Res Clin Pract*. 2017;134:91–105.

20. Walters SJ, Bonacho dos Anjos Henriques-Cadby I, Bortolami O, et al. Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom health technology assessment programme. *BMJ Open*. 2017;7: e015276.
21. Greenberg RG, Gamel B, Bloom D, et al. Parents' perceived obstacles to pediatric clinical trial participation: findings from the clinical trials transformation initiative. *Contemp Clin Trials Commun*. 2018;9:33–39.
22. Checketts JX, Scott JT, Meyer C, et al. The robustness of trials that guide evidence-based orthopaedic surgery. *J Bone Jt Surg*. 2018;100: e85.
23. Tignanelli CJ, Napolitano LM. The fragility index in randomized clinical trials as a means of optimizing patient care. *JAMA Surg*. 2019;154:74.
24. Bos D, Braga LH. *Evaluating the Robustness of the Pediatric Urology Literature: A Role for the Fragility Index*. Society of Pediatric Urology Fall Congress; 2018.
25. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. *Soc Sci Med*. 2018;210:2–21.
26. Cohen AT, Goto S, Schreiber K, et al. Why do we need observational studies of everyday patients in the real-life setting?: Table 1. *Eur Hear J Suppl*. 2015;17:D2–D8.
27. O'Brien T, Viney R, Doherty A, et al. Why don't Mercedes Benz publish randomized trials? *BJU Int*. 2010;105:293–295.