The Results of Randomized Controlled Trials in Emergency Medicine Are Frequently Fragile

Jamin Brown, DO*; Aaron Lane, DO; Craig Cooper, BS; Matt Vassar, PhD

*Corresponding Author. E-mail: jaminb@okstate.edu.

Study objective: Randomized controlled trials govern evidence-based clinical practice, and it is therefore critical that their results be robust. We aim to investigate the fragility of randomized controlled trials in emergency medicine by determining how often significance would be nullified with small changes in outcomes using the fragility index.

Methods: We conducted a methodological systematic review of randomized controlled trials in emergency medicine published in the top 10 general medicine journals and the top 10 emergency medicine journals. Inclusion criteria required that trials be emergency medicine studies structured with a 2-arm or 2-by-2 factorial design and report at least 1 statistically significant dichotomous outcome.

Results: A total of 180 trials met inclusion criteria. The median fragility index across all trials in emergency medicine was 4 (interquartile range [IQR] 2 to 10) and the median sample size was 140 (IQR 69.5 to 286). For trials from general medicine journals (n=32), the median fragility index was 9 (IQR 4 to 16.5) and the median sample size was 415.5 (IQR 219.5 to 901); for trials from emergency medicine journals (n=148), the median fragility index was 4 (IQR 1 to 9) and the median sample size was 119 (IQR 60 to 227.25). One third of all trials (62/180) had a loss to follow-up that was greater than or equal to the fragility index. There was a modest correlation between fragility index and total number of events (r=0.36; 95% confidence interval [CI] 0.23 to 0.48) and a weak correlation between fragility index and total sample size (r=0.26; 95% CI 0.12 to 0.39). There was no correlation between fragility index and either P value (r=−0.14; 95% CI −0.28 to −0.006) or Science Citation Index (r=0.07; 95% CI −0.08 to 0.22).

Conclusion: The statistical significance of the results of randomized controlled trials in emergency medicine was often contingent on a small number of events. Until frequentist interpretation of clinical trials is replaced with Bayesian analysis, the fragility index may have utility as a tool to aid clinicians in assessing the robustness of randomized controlled trials in emergency medicine when considered in conjunction with the fragility quotient and other reported metrics. [Ann Emerg Med. 2019;73:565-576.]

Please see page 566 for the Editor's Capsule Summary of this article.

Readers: click on the link to go directly to a survey in which you can provide **feedback** to *Annals* on this particular article.
A **podcast** for this article is available at www.annemergmed.com.

0196-0644/$-see front matter

https://doi.org/10.1016/j.annemergmed.2018.10.037

INTRODUCTION

Background

Modern medicine is informed by evidence-based clinical practice, and randomized controlled trials are the standard on which such evidence is frequently based. In emergency medicine, randomized controlled trials influence clinical decisionmaking and serve as a key component of the clinical policies established by the American College of Emergency Physicians (ACEP). Each of the 19 current clinical policies published by ACEP lists randomized controlled trials as class 1 evidence, the highest level of evidence, and many of the recommendations are underpinned by randomized controlled trials. For example, the recommendations to administer tissue plasminogen activator in a subset of patients with acute ischemic stroke,[1] perform bedside ultrasonography for unstable patients with blunt abdominal trauma,[2] administer barbiturates for refractory status epilepticus,[3] and obtain a noncontrast computed tomography scan of the head for certain patients with traumatic brain injuries[4] are all based on randomized controlled trial results. Given the significant role randomized controlled trials play in governing clinical practice, it is critical that their results be robust.

Historically, P values have been used to indicate statistical significance; however, this approach has significant limitations and has been heavily criticized for being overly simplistic, with frequent misapplication and misinterpretation.[5-10] Furthermore, such values provide little insight into the magnitude, or robustness, of a treatment effect. To account for this, we propose the use of an added metric, the fragility index, as a tool to describe the robustness of statistically significant results from randomized controlled trials in emergency medicine.

Editor's Capsule Summary

What is already known on this topic

The results of small studies may be unstable because a change in outcome for just a few patients might alter the study's conclusion.

What question this study addressed

This study used the fragility index to explore the stability of the conclusions of 180 research articles in emergency medicine.

What this study adds to our knowledge

The mean number of subjects was 140 and the fragility index was 4. In other words, not much has to change to alter these study's conclusions. Fragility was not strongly correlated with either sample size or P value.

How this is relevant to clinical practice

This article adds to the body of work demonstrating that readers must be cautious in placing too much confidence in a single study. It also provides further evidence that the frequentist approach to clinical trials may be suboptimal.

Importance

The fragility index is a novel tool developed to aid in assessing the robustness of statistically significant dichotomous outcomes from randomized controlled trials.[11] It is defined as the minimum number of patients whose status would have to change from a nonevent to an event to nullify a statistically significant result. A smaller fragility index indicates that the statistical significance of a given outcome hinges on only a few events, representing a relatively fragile outcome, whereas a larger fragility index indicates that the significance hinges on a greater number of events and suggests a more robust outcome. Consider, for example, a recent randomized controlled trial that investigated the effect of active compression-decompression cardiopulmonary resuscitation (CPR) with augmentation of negative intrathoracic pressure on survival rates with favorable neurologic function after out-of-hospital cardiac arrest.[12] Patients with out-of-hospital cardiac arrest were randomized to receive either active compression-decompression CPR or standard CPR. Seventy-five of the 840 patients who received active compression-decompression CPR survived to hospital discharge with favorable neurologic function, whereas only 47 of the 813 who received standard CPR were found to have that same outcome. The authors reported this outcome to be statistically significant, with a P value of .02. However, if the outcome of 4 patients in the control group were to change from nonevent to event, the significance of this result would be nullified (P=.05). The fragility index in this case would be 4. Although the fragility index alone yields useful information, a more complete understanding of study fragility can be had when the fragility index is considered in conjunction with the fragility quotient.

The fragility quotient is a metric that accounts for the fragility index in the context of sample size.[13] It is defined as the fragility index divided by the total sample size. The utility of the fragility quotient lies in its ability to provide an objective value to the subjective importance providers may assign to an outcome with a given fragility index in the setting of a certain sample size. Put simply, the fragility quotient assesses the robustness of the fragility index. To illustrate the utility of the fragility quotient, consider 2 trials, both with a calculated fragility index of 2. The first trial[14] had a total sample size of 30 and the second[15] a total sample size of 130. Many clinicians might argue that the fragility index of 2 with a sample size of 130 for the second trial represents a more fragile result than the fragility index of 2 with a sample size of 30 for the first trial. If these numbers were to be incorporated into the fragility quotient, the first trial would have a fragility quotient of 0.067 and the second a fragility quotient of 0.015. That is, the significance of the results for the first trial was contingent on approximately 7 events per 100 patients, whereas the significance of the results for the second trial was contingent on approximately 2 events per 100 patients. These results suggest that the second trial may indeed be "relatively" more fragile than the first, despite that both had the same fragility index.

Goals of This Investigation

Although the importance of randomized controlled trials to evidence-based clinical practice in emergency medicine is clear, the manner in which the significance of these trials is traditionally interpreted may be inadequate. Because the significance of an outcome can be contingent on just a few events, the fragility index could be a useful adjunct to aid in assessing the robustness of such trials. Here, we aim to investigate the fragility of randomized controlled trials in emergency medicine.

MATERIALS AND METHODS

This study was not subject to institutional review board oversight because it did not meet the regulatory definition of human subject research as defined in 45 CFR 46.102(d)

and (f) of the US Department of Health and Human Services' Code of Federal Regulations.[16]

We conducted a methodological systematic review of randomized controlled trials specific to emergency medicine that were published in 10 emergency medicine journals and 10 general medicine journals between 2006 and 2016. Requirements for inclusion necessitated that trials be randomized with 1:1 allocation of treatment to control, be structured as a 2-arm or 2-by-2 factorial design, and report at least 1 statistically significant dichotomous outcome. We also required that included studies be categorized as emergency medicine articles, which we defined as any trial conducted in the emergency department (ED) or based on an intervention relevant to the practice of emergency medicine in the out-of-hospital or ED setting. We excluded trials with crossover designs, trials with greater than 2 arms, cluster trials, noninferiority studies, trials with nondichotomous or continuous outcomes, and any trial outside the realm of emergency medicine. Subgroup analyses were not considered.

We included the top 10 emergency medicine journals and the top 10 general medicine journals, based on rankings obtained through Clarivate Analytics' Science Citation Index (previously a resource of Thomson Reuters). For both groups of journals, we elected to include only those that were general in scope. We excluded journals focused on surgery, trauma, toxicology, pediatrics, internal medicine, and family medicine, as well as those focused on article reviews and other secondary analyses. In accordance with these parameters, we included the following in the group of emergency medicine journals: *Resuscitation; Annals of Emergency Medicine; Academic Emergency Medicine; Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine; European Journal of Emergency Medicine; Injury; Emergency Medicine Journal; American Journal of Emergency Medicine; Emergency Medicine Australasia;* and *Canadian Journal of Emergency Medicine.* In the group of general medicine journals, the following were included: *New England Journal of Medicine, Lancet, Journal of the American Medical Association, BMJ, PLOS Medicine, BMC Medicine, Canadian Medical Association Journal, American Journal of Medicine, Medical Journal of Australasia,* and *Annals of Medicine.*

We conducted a series of searches through PubMed to identify randomized controlled trials specific to emergency medicine in the aforementioned journals. To select for emergency medicine articles within these journals, we used the Medical Subject Headings (MeSH) terms "emergency medicine" and "emergency service, hospital," and also applied "hospital emergency service" and "emergency department" to all fields. The final components of our search string selected for randomized controlled trials and articles published from 2006 to 2016. Using these search constraints, we performed the following search query for each journal, substituting *Journal Name* with the appropriate name of each journal: "*Journal Name*"[Journal] AND ("emergency medicine"[MeSH Terms] OR ("emergency service, hospital"[MeSH Terms] OR ("emergency"[All Fields] AND "service"[All Fields] AND "hospital"[All Fields]) OR "hospital emergency service"[All Fields] OR ("emergency"[All Fields] AND "department"[All Fields]) OR "emergency department"[All Fields])) AND Randomized Controlled Trial[ptyp] AND ("2006/01/01"[PDAT] : "2016/12/31"[PDAT]). Searches within the group of emergency medicine journals were conducted on September 25, 2017, whereas searches within the group of general medicine journals were conducted on October 2, 2017.

Two authors (J.B. and A.L.) met to discuss the inclusion criteria, and consensus was reached in regard to the abstract screening process. J.B. and A.L. then independently screened the abstracts, selecting for trials with any significant outcome regardless of whether that outcome was dichotomous. Included trials otherwise met the inclusion criteria detailed above. On screening completion, J.B. and A.L. reviewed and resolved discrepancies.

Data Collection and Processing

Two authors (J.B. and C.C.) met to discuss inclusion criteria, and consensus was reached in regard to the process for full-text review. Both J.B. and C.C. then performed an independent full-text review of records retained from the initial screening. They then reviewed and resolved discrepancies. J.B. extracted data for qualifying statistically significant dichotomous outcomes from included articles. For quality control, another author (M.V.) then independently verified a randomly selected proportion (25%) of the data extracted by J.B. previously. For trials reporting multiple eligible outcomes, the primary outcome was analyzed. If multiple primary outcomes were reported, a board-certified emergency physician (A.L.) conducted an independent review to identify the single most patient-important outcome according to the Grading of Recommendations Assessment, Development and Evaluation approach for selecting between outcomes that are critical for decisionmaking, important but not critical, or of low importance.[17] In applying this approach, A.L. ranked outcomes on a scale ranging from 1 (low importance for decisionmaking) to 10 (critical for decisionmaking), and the outcome with the highest ranking was included in our analysis. This methodology was also

applied to trials reporting multiple secondary or unspecified outcomes.

The following data were extracted from qualifying trials with a piloted electronic form: article name, journal name, publication year, number randomized for each group, number lost to follow-up for each group, a description of the statistically significant dichotomous outcome, the type of outcome (primary, secondary, or unspecified), number of patients analyzed per group, number of events per group, *P* value, statistical test used, and the Web of Science Citation Index. An example of the electronic form is shown in Appendix E1 (available online at http://www.annemergmed.com).

Primary Data Analysis

Characteristics of included trials were summarized with descriptive statistics that were calculated with Excel (version 2016; Microsoft, Redmond, WA). The fragility index for each study was calculated with a Web-based fragility calculator available through http://www.clincalc.com/Stats/FragilityIndex.aspx. As part of the calculation of the fragility index, the online calculator recalculated original *P* values based on the Fisher's exact test, using extracted data from included studies. Those values, which serve as a reference point in the calculation of the fragility index, were then recorded. The fragility quotient, defined as the fragility index divided by the sample size, was calculated for each trial with Excel. Instructions on how to screen articles, assess outcomes, and apply the fragility index and fragility quotient are detailed in Appendix E2 (available online at http://www.annemergmed.com).

RESULTS

We identified 996 records through our series of PubMed searches, 155 from general medicine journals and 841 from emergency medicine journals (Figure 1). After an initial screening of abstracts and subsequent full-text review of retained studies, 180 articles were included in our fragility analysis, 32 from general medicine journals and 148 from emergency medicine journals (Appendix E3, available online at http://www.annemergmed.com). Excluded during the initial screening of abstracts were 84 records from general medicine journals and 474 from emergency medicine journals. Records from general medicine journals were most frequently excluded during this initial screening for being cluster-based designs, whereas those from emergency medicine journals were most frequently excluded for being crossover designs. Of the 71 studies from general medicine journals that underwent full-text review, 39 were excluded, most commonly because they were not emergency medicine articles. Five journals from the general medicine group had no qualifying records. Within the group of emergency medicine journals, 367 articles underwent full-text review and 219 were subsequently excluded, with the most common reason being that the reported significant outcomes were nondichotomous. The remaining studies accounted for the 180 trials included in our fragility analysis.

Of all included records, more than half (102/180) were extracted from 3 emergency medicine journals (Table 1). Among the general medicine journals, most qualifying records were extracted from the *New England Journal of Medicine,* whereas most records from the emergency medicine journals were extracted from *Annals of Emergency Medicine*. The median sample size for trials from general medicine journals was 415.5 (interquartile range [IQR] 219.5 to 901), whereas the median sample size for trials from emergency medicine journals was 119 (IQR 60 to 227.25). For trials from general medicine journals, the median number of events was 94 (IQR 56.5 to 142.5), whereas the median for trials from emergency medicine journals was 43 (IQR 22 to 86). The median loss to follow-up for trials from general medicine journals was 2 (IQR 0 to 15.5) compared with 0 (IQR 0 to 6) for trials from emergency medicine journals. Trials from general medicine journals had a median Web of Science Citation Index of 69.5 (IQR 27.25 to 140.75), whereas the median for trials from emergency medicine journals was 12 (IQR 5 to 26). Recalculated *P* values ranged from 0 to .03 for trials from general medicine journals and 0 to .17 for trials from emergency medicine journals. The analyzed outcome was primary for 34% of trials from general medicine journals (11/32) and 42% of trials from emergency medicine journals (63/148). There were 38 articles for which we used the Grading of Recommendations Assessment, Development and Evaluation approach to select for the single most patient-important outcome in trials with multiple eligible outcomes. Four trials were stopped early.

The median fragility index across all included randomized controlled trials in emergency medicine was 4 events (IQR 2 to 10) (Table 2). Of the trials from general medicine journals, the median fragility index was 9 (IQR 4 to 16.5), whereas the median fragility index for trials from emergency medicine journals was 4 (IQR 1 to 9). The distribution of the fragility index across all trials is reported in Figure 2. Ten percent of trials (18/180), all of which were from emergency medicine journals, had results that became nonsignificant on application of the Fisher's exact test, resulting in a fragility index of 0. Thirty-six percent of trials (64/180) had a fragility index that was less than or equal to 2. For all included journals, *BMJ* had the highest

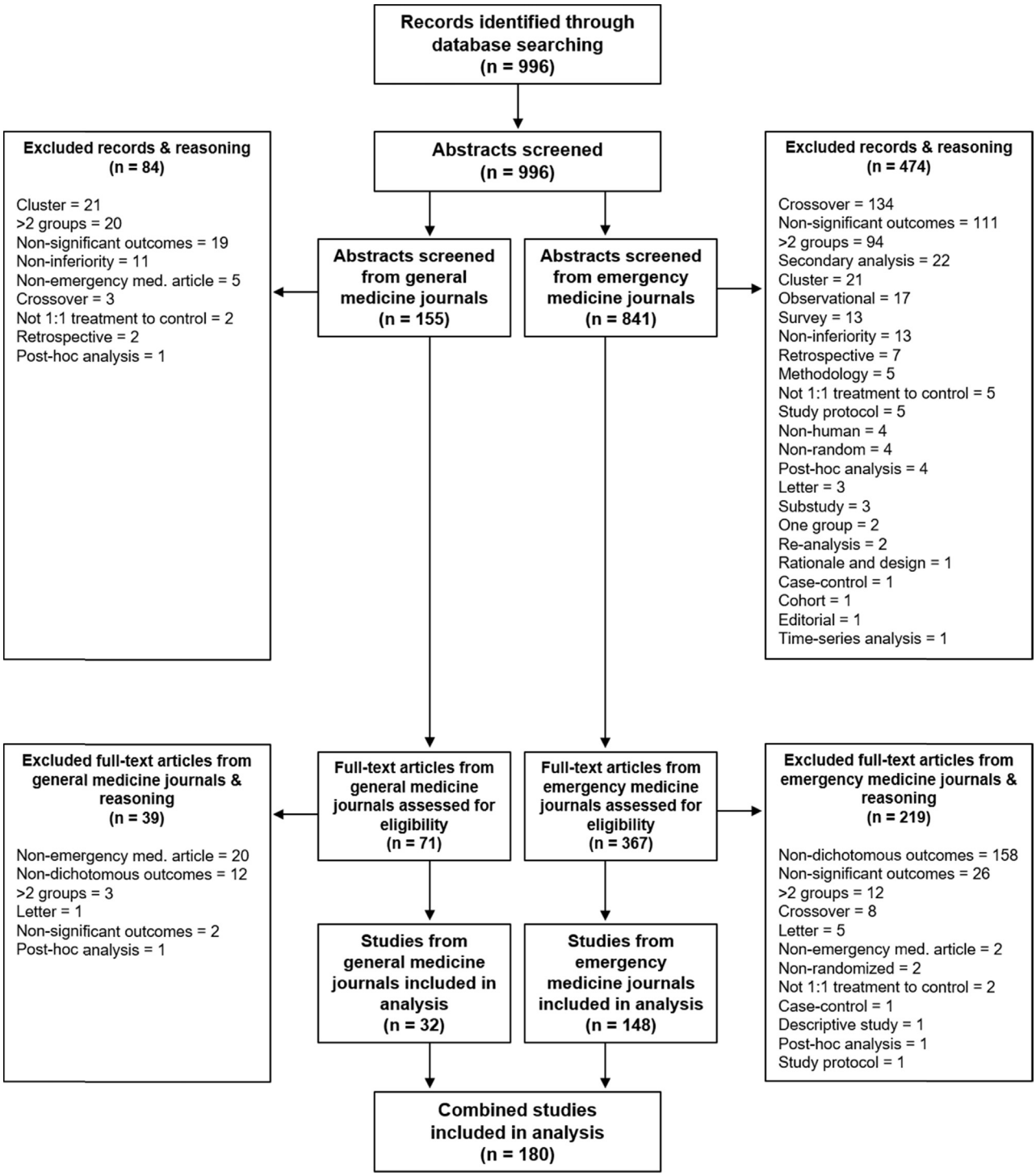


Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-analyses diagram detailing excluded records and reasoning.

median fragility index, at 16.5 (IQR 11.75 to 21.5), based on the results of 4 qualifying trials (Figure 3). Among journals with at least 8 qualifying trials, the *New England Journal of Medicine* had the highest median fragility index from the group of general medicine journals, at 9 (IQR 5 to 30), whereas *Annals of Emergency Medicine* had the highest median fragility index from the group of emergency medicine journals, at 5 (IQR 2 to 9.75). Among the 5 most cited studies (median Science Citation Index 325; IQR 304 to 397), the median fragility index was 23 (IQR 6 to 48). The fragility index varied by type of outcome, ranging from 5 (IQR 2 to 11.75) for primary outcomes to 3.5 for both

Table 1. Characteristics of qualifying randomized controlled trials.

Characteristic	All Trials Combined (n=180)	Trials From General Medicine Journals (n=32)	Trials From Emergency Medicine Journals (n=148)
Journal, general medicine, No. (%)			
New England Journal of Medicine	9 (5.0)	9 (28.1)	—
Journal of the American Medical Association	8 (4.4)	8 (25)	—
Lancet	6 (3.3)	6 (18.8)	—
Canadian Medical Association Journal	5 (2.8)	5 (15.6)	—
British Medical Journal	4 (2.2)	4 (12.5)	—
PLOS Medicine	0	0	—
BMC Medicine	0	0	—
American Journal of Medicine	0	0	—
Medical Journal of Australia	0	0	—
Annals of Medicine	0	0	—
Journal, emergency medicine, No. (%)			
Annals of Emergency Medicine	38 (21.1)	—*	38 (25.7)
American Journal of Emergency Medicine	36 (20.0)	—	36 (24.3)
Academic Emergency Medicine	28 (15.6)	—	28 (18.9)
Emergency Medicine Journal	18 (10.0)	—	18 (12.2)
Resuscitation	16 (8.9)	—	16 (10.8)
Canadian Journal of Emergency Medicine	5 (2.8)	—	5 (3.4)
Emergency Medicine Australasia	3 (1.7)	—	3 (2.0)
Scandinavian Journal of Trauma, Resuscitation, and Emergency Medicine	2 (1.0)	—	2 (1.3)
Injury	1 (0.6)	—	1 (0.7)
European Journal of Emergency Medicine	1 (0.6)	—	1 (0.7)
Sample size, median (IQR)	140 (69.5–286)	415.5 (219.5–901)	119 (60–227.25)
Number of events, median (IQR)	46.5 (24–104.5)	94 (56.5–142.5)	43 (22–86)
Loss to follow-up, median (IQR)	0 (0–8.25)	2 (0–15.5)	0 (0–6)
Web of Science Citation Index, median (IQR)	17 (5–41.5)	69.5 (27.25–140.75)	12 (5–26)
Recalculated *P* value, range	0–0.17	0–0.034	0–0.17
Primary outcome, No. (%)	74 (41.1)	11 (34.4)	63 (42.6)
Secondary outcome, No. (%)	66 (36.7)	21 (65.6)	45 (30.4)
Unspecified outcome, No. (%)	40 (22.2)	0	40 (27.0)

*Dashes represent content that does not apply to the corresponding characteristic.

secondary outcomes (IQR 1 to 9) and unspecified outcomes (IQR 2 to 7.5). The median fragility quotient for trials from general medicine journals was 0.026 (IQR 0.010 to 0.041) compared with 0.033 (IQR 0.010 to 0.069) for trials from emergency medicine journals. The total number lost to follow-up was greater than or equal to the fragility index in one third of all trials (62/180).

There was a weak correlation between fragility index and sample size (r=0.26; 95% confidence interval [CI] 0.12 to 0.39) (Figure 4) and a modest correlation between fragility index and total number of events (r=0.36; 95% CI 0.23 to 0.48). There was no important relationship between fragility index and either *P* value (r=–0.14; 95% CI –0.28 to 0.006) or Science Citation Index (r=0.07; 95% CI –0.08 to 0.22).

LIMITATIONS

There were a number of limitations to our study. First, the fragility index can be used only for randomized controlled trials that have a 1:1 allocation and at least one statistically significant dichotomous outcome.[11] A significant proportion of studies in emergency medicine were crossover-based manikin studies, which cannot be

Table 2. Fragility index and fragility quotient based on journal and outcome characteristics.

Characteristic	Median Fragility Index (IQR)	Median Fragility Quotient (IQR)
All trials (n=180)	4 (2–10)	0.033 (0.010–0.067)
Trials from general medicine journals (n=32)	**9 (4–16.5)**	**0.026 (0.010–0.041)**
New England Journal of Medicine (n=9)	9 (1–140)	0.031 (0.004–0.140)
Journal of the American Medical Association (n=8)	4.5 (2–50)	0.010 (0.003–0.177)
Lancet (n=6)	11 (3–36)	0.017 (0.002–0.083)
Canadian Medical Association Journal (n=5)	7 (3–11)	0.027 (0.008–0.141)
British Medical Journal (n=4)	16.5 (2–32)	0.048 (0.010–0.075)
Trials from emergency medicine journals (n=148)	**4 (1–9)**	**0.033 (0.010–0.069)**
Annals of Emergency Medicine (n=38)	5 (2–9.75)	0.029 (0.009–0.067)
American Journal of Emergency Medicine (n=36)	4 (2–9.25)	0.050 (0.016–0.084)
Academic Emergency Medicine (n=28)	3 (1–5.25)	0.035 (0.007–0.067)
Emergency Medicine Journal (n=18)	3.5 (1–9)	0.027 (0.013–0.036)
Resuscitation (n=16)	2.5 (1–6)	0.037 (0.012–0.072)
Canadian Journal of Emergency Medicine (n=5)	5 (0–14)	0.068 (0–0.123)
Emergency Medicine Australasia (n=3)	2 (0–7)	0.010 (0–0.140)
Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine (n=2)	7 (4–10)	0.048 (0.013–0.083)
Injury (n=1)	2 (2–2)	0.067 (0.067–0.067)
European Journal of Emergency Medicine (n=1)	0	0
Outcome type		
Primary (n=74)	5 (2–11.75)	0.039 (0.015–0.081)
Secondary (n=66)	3.5 (1–9)	0.026 (0.007–0.054)
Unspecified (n=40)	3.5 (2–7.5)	0.031 (0.010–0.067)

analyzed with the fragility index. Moreover, because the fragility index is amenable only to dichotomous outcomes, there were likely clinically important continuous outcomes that were excluded from our study. Our search strategy also limited results to studies from 2006 through 2016, and we used MeSH terms and other search constraints to generate a workable convenience sample and eliminate studies unrelated to emergency medicine. Comparing our search method with constraints with a general search of *Annals of Emergency Medicine*, we estimate that approximately 10% of relevant studies may have been missed with these constraints. Furthermore, our study included only randomized controlled trials from high-ranking journals according to Clarivate Analytics' Science Citation Index and may have excluded studies from less prominent journals that could have had smaller fragility indexes. Our convenience sample also returned a disproportionate number of trials from general medicine journals relative to emergency medicine journals, with far fewer trials from general medicine journals. Had more records from general medicine journals been included, our sample may have been composed of more trials with larger fragility indexes. Finally, there were 4 trials included in our study that were stopped early, and the calculated fragility indexes for the outcomes from those studies may not be consistent with those values that might have been obtained had the studies been completed. All of these issues affect the generalizability of our findings, and readers are encouraged to interpret our results with these limitations in mind.

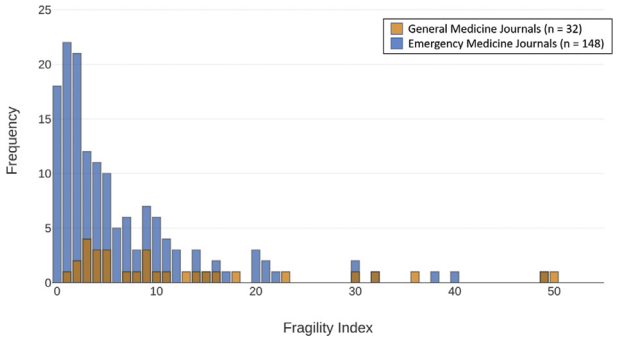


Figure 2. Distribution of fragility index across 32 trials from general medicine journals (median fragility index 9; IQR 4 to 16.5) and 148 trials from emergency medicine journals (median fragility index 4; IQR 1 to 9). We excluded 2 outlying data points to improve visualization of distribution. Excluded values: general medicine journal, fragility index 140; emergency medicine journal, fragility index 656.

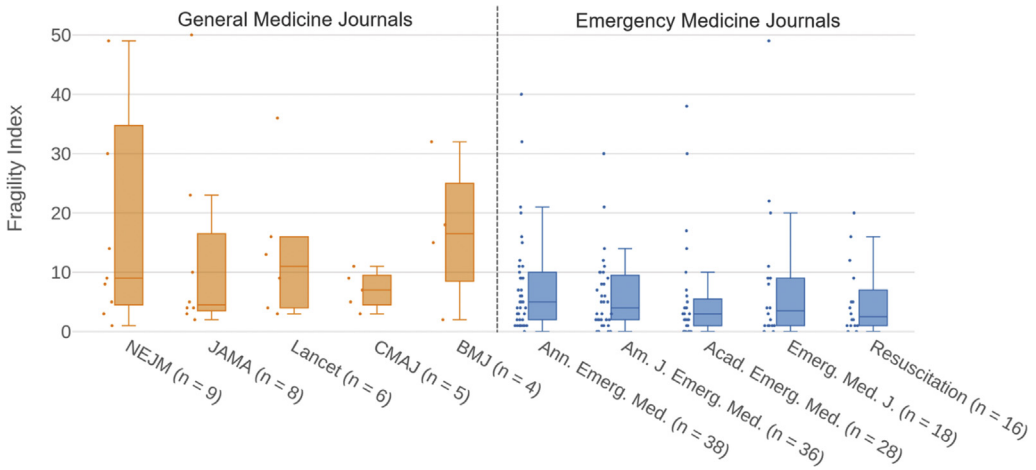


Figure 3. Distribution of fragility indexes for each of the top 5 general medicine and emergency medicine journals, based on total number of included trials. Upper and lower borders of boxes denote IQRs. Bolded horizontal lines contained within boxes denote medians. Whiskers represent ranges (or upper/lower fences when data include outliers). We excluded 2 outlying data points to improve visualization of distribution. Excluded values: *New England Journal of Medicine* fragility index 140; *American Journal of Emergency Medicine* fragility index 656.

DISCUSSION

This study investigated the fragility of randomized controlled trials in emergency medicine. Our findings demonstrate that the statistically significant dichotomous outcomes from such trials are frequently contingent on only a few events. Across all trials in emergency medicine, the median fragility index was 4. That is, if only 4 patients were to be moved from one outcome group to the other, statistical significance would be nullified. The median fragility index for trials published in general medicine journals was 9, more than double that of trials from emergency medicine journals, which had a median fragility index of 4. It is possible that these results reflect study quality. However, it may be more likely that they reflect some complex form of bias on the part of authors choosing a journal for their work and on the part of journals choosing what work they will publish. Regardless, these findings suggest that the results of randomized controlled trials published in emergency medicine journals were more fragile than those published in general medicine journals. However, we also found that trials from general medicine journals generally had much larger sample sizes compared with trials from emergency medicine journals. When considering the aforementioned fragility indexes in the context of sample size using the fragility quotient, we found the median fragility quotient for trials from general medicine journals to be 0.026, whereas that for trials from emergency medicine journals was 0.033. Put another way, the significance of results for trials from general medicine journals was typically contingent on slightly less than 2.5 events per 100 patients, whereas that for trials from emergency medicine journals was typically contingent on slightly greater than 3 events per 100 patients. These results suggest that trials from general medicine journals were actually "relatively" more fragile than those published in emergency medicine journals when considered in the context of sample size. Despite these main findings, there was some variability to our data, with examples of fragile results across the spectrum of sample sizes and throughout both groups of journals.

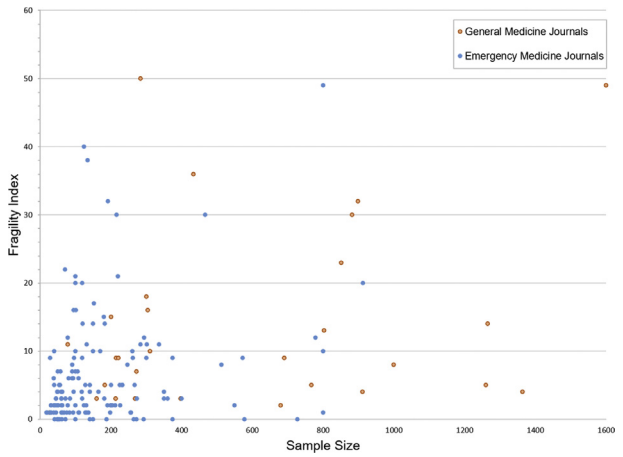


Figure 4. Fragility index by sample size for trials from general medicine journals (n=32) and emergency medicine journals (n=148). We excluded 5 outlying data points to improve visualization of distribution. Excluded values (sample size, fragility index): 2,323, 656; 2,470, 4; 4,855, 0; 5,718, 7; and 100, 140.

Ten percent of all trials had results that became nonsignificant on application of the Fisher's exact test, leading to a fragility index of 0 and representing the most

fragile records from our study. All were trials from emergency medicine journals. This loss of significance occurred largely in cases in which authors had used the χ^2 test for small sample sizes and subsequently obtained a *P* value of exactly or very near .05. The χ^2 test, though, is an approximation that is accurate only when applied to large sample sizes, whereas the Fisher's exact test is, as the name implies, an exact procedure that is accurate across both small and large sample sizes.[18] Thus, statistical significance was nullified in such cases on application of the more conservative Fisher's exact test. One example in which this occurred was a study that sought to compare the efficacy of tropisetron versus metoclopramide for the treatment of nausea and vomiting in undifferentiated ED patients.[19] The authors concluded that tropisetron resulted in a statistically significant reduction in vomiting compared with metoclopramide according to a *P* value of .05 that had been calculated with the χ^2 test. However, their trial had a sample size of just 100 patients, and this outcome was based on only 2 events in the tropisetron group and 9 in the metoclopramide group. When the Fisher's exact test was applied to their data, their results became nonsignificant, resulting in a fragility index of 0. Although we recorded no instances in which the fragility index was 0 for trials from general medicine journals, we did find numerous examples of studies that were still quite fragile. One study investigated the utility of icatibant versus standard therapy for treatment of angiotensin-converting enzyme inhibitor–induced angioedema.[20] The authors concluded that icatibant resulted in significantly more patients with complete resolution of edema within 4 hours of treatment. Their trial, though, had a sample size of only 27 patients, and this outcome was based on 5 events for the intervention group and 0 for the control group. Based on these data, the fragility index was calculated to be 1. Including these examples, 36% of all trials had a fragility index of 2 or less. Most of these were undermined by small sample sizes and low event totals, which may have influenced the fragile nature of their results. Moreover, one third of all trials had a total number lost to follow-up that was greater than or equal to the fragility index. Thus, many trials in emergency medicine hinged on just 1 or 2 events, and significance might be lost altogether if the trial were repeated. Furthermore, a large proportion of trials had as many missing data points as would be required to reverse significance. We suggest that clinicians pause when the results of a trial hinge on a small number of events or when as many data points are missing as would be required to reverse the significance of that trial.

Although trials in our study were frequently fragile, we did find examples of robust studies. One trial from a general medicine journal investigated whether a wait-and-see prescription for acute otitis media in children would reduce the use of antibiotics compared with a standard prescription.[21] The authors found that significantly more parents did not fill the antibiotic when this methodology was used, based on 82 events in the intervention group and 17 in the control group. Their study had a total sample size of 283. This outcome resulted in a fragility index of 50. In another trial from an emergency medicine journal, investigators evaluated the effect of an ED observation syncope protocol on hospital admission rate.[22] They found that significantly fewer patients with unexplained syncope were admitted to the hospital when this protocol was used, based on 9 events in the intervention group and 57 in the control group, with a total sample size of 124. This resulted in a fragility index of 40. Such robust studies were relatively rare, though, with only 15% of trials boasting a fragility index of greater than or equal to 15. Furthermore, we found no factors that could reliably predict the robustness of a given trial. The fragility index did correlate weakly with total sample size and modestly with total number of events, and primary outcomes had a higher median fragility index than other outcome types. However, these findings were inconsistent. We generally found that the fragility index varied widely across sample size, number of events, and outcome type. Furthermore, we found no correlation between the fragility index and Science Citation Index, indicating that even heavily cited randomized controlled trials may be based on fragile results. Thus, these basic trial characteristics provide little insight into the robustness of the results for any given study, and we therefore suggest that clinicians consider the fragility index in the context of the study as a whole.

The concept of fragility is not new, and there is now limited evidence available on the fragility of randomized controlled trials across other disciplines. Feinstein[23] originally proposed the unit fragility index in 1990, and Walter[24] later expanded on this idea, but their concepts were difficult to apply and required further decisionmaking beyond just a calculation. Walsh et al[11] were the first to propose the fragility index as used in our study, in which the metric represented an absolute number required to nullify statistical significance. Their group was also the first to characterize the fragility for a sample of randomized controlled trials, reporting in their study a median fragility index of 8 and median sample size of 682 for trials from high-impact general medicine journals. We found emergency medicine–based randomized controlled trials from our group of high-ranking general medicine journals to be similarly fragile, with a median fragility index of 9. Others have since reported on the fragility of trials across

other disciplines. Matics et al[25] reported a median fragility index of 7 with a median sample size of 152 across a cohort of trials in pediatric critical care that were extracted from high-impact journals, whereas Ridgeon et al[26] found a median fragility index of 2 with a median sample size of 126.5 across all trials in critical care, although the only outcome considered for the latter was mortality. For trials in nephrology, Shochet et al[27] recorded a median fragility index of 3 and median sample size of 134. Evaniew et al[28] evaluated the fragility of trials in spine surgery and found a median fragility index of 2, along with a median sample size of 132. To our knowledge, the study by Edwards et al[29] is the only one to date to have investigated both the fragility index and fragility quotient for a cohort of trials, reporting a median fragility index of 5, median fragility quotient of 0.012, and median sample size of 400 for randomized controlled trials supporting venous thromboembolism treatment. In comparison, we found a median fragility index of 4 and median sample size of 140 across all trials in emergency medicine. Thus, although each of these studies, including our own, reported fragile results, there appears to be variability among findings. More recent studies, on the other hand, have recorded more robust results. Kruse and Vassar[30] recorded a median fragility index of 16 and median sample size of 2,548 for randomized controlled trials supporting diabetes treatment, and Docherty et al[31] investigated the fragility of trials referenced in heart failure management guidelines and found a median fragility index of 26, with a median sample size of 2,331. These results, when considered in the context of the aforementioned studies, may suggest that there is even more variability in study fragility than previously thought. Further research is required to determine how much variability truly exists for trials across different disciplines and to pinpoint what constitutes a significant difference between fragility indexes.

With an ever-expanding literature base,[32,33] randomized controlled trials are likely to remain an important piece of evidence on which clinicians base their practice in the ED. Despite this, there has been little evolution in the statistical methodology used to denote significance in such trials. Traditionally, *P* values have been used to indicate statistical significance, and this will likely remain common practice. However, this approach has been heavily criticized for being overly simplistic, with frequent misapplication and misinterpretation.[5-9] Furthermore, *P* values provide little insight into the magnitude of a treatment effect. CIs do provide a more complete picture of effect size[34] but still do not provide an absolute measure of the number of events required to reverse significance. The fragility index is simple in its application and addresses some of these shortcomings. We suggest that the fragility index be considered in the context of sample size using the fragility quotient and reported in conjunction with other metrics, including the *P* value, CI, and number needed to treat, as applicable. In combination, these measures may aid clinicians in determining whether statistically significant results from randomized controlled trials are truly clinically significant.

Although the fragility index and fragility quotient do provide a relative wealth of information when considered in conjunction with other metrics, we again emphasize the limitations of the fragility index itself and acknowledge that this tool does not address some of the more complex issues underlying common practices in statistical analysis. Because the fragility index relies on *P* values, it is essentially an extension of the frequentist approach to data analysis. For this reason, it is also subject to the many issues inherent to *P* values and null hypothesis significance testing in general. Null hypothesis significance testing assumes a single hypothesis and many potential observations, including those that were not actually observed.[35] Clinical research, though, yields a single observation in the form of experimental results. Thus, a more logical approach may be that of Bayesian analysis, which considers many possible hypotheses in the context of the observed data. Null hypothesis significance testing also emphasizes binary decisionmaking that necessitates that the investigator either reject or fail to reject a hypothesis when such a decision often cannot reasonably be made on the basis of a single study.[36] Moreover, it relies on perfect randomization, perfect measurement, no confounding, and samples that perfectly mirror the target population.[37] To more accurately report and analyze imperfect and biased data, the analysis necessarily must incorporate judgment. Bayesian analysis does just that in the form of prior distributions. Prior distributions express previous knowledge about probabilities across parameter values. Bayesian analysis then reallocates probabilities across parameter values in a manner that best accommodates the data. This type of analysis provides complete information about the joint distribution of credible probabilities for parameter values, including means, SDs, effect sizes, and normality.[38] Clinical research may benefit from this alternative approach because it emphasizes estimation of effect sizes rather than binary decisionmaking. However, as long as the research community insists on hypothesis-based testing, metrics such as the fragility index may represent important tools to help clinicians better interpret the robustness of statistically significant results from randomized controlled trials.

In conclusion, the results from randomized controlled trials in emergency medicine were frequently fragile. One third of all trials had a loss to follow-up that was greater than or equal to the fragility index, meaning as many data points were missing

as would be required to reverse significance. Clinicians should interpret with caution trials that have small fragility indexes or loss to follow-up greater than the fragility index. Trials from general medicine journals were generally more robust than those from emergency medicine journals. However, trials from emergency medicine journals were "relatively" more robust when considered in the context of sample size. The fragility index is a novel metric that may aid clinicians in assessing the robustness of randomized controlled trials in emergency medicine when considered in conjunction with the fragility quotient and other reported metrics.

Supervising editor: David L. Schriger, MD, MPH. Specific detailed information about possible conflict of interest for individual editors is available at https://www.annemergmed.com/editors.

Author affiliations: From the Oklahoma State University Center for Health Sciences, Tulsa, OK.

Author contributions: MV conceived the study, organized the research team, supervised data analysis, and provided statistical advice. JB performed the search queries and extracted the data. MV independently verified a randomly selected proportion of the data. JB and AL performed an independent initial screening of abstracts. JB and CC performed an independent full-text review of articles retained from the initial screening. JB, AL, and CC analyzed the data. JB and MV drafted the article. All authors contributed to an internal review of the final article. JB takes responsibility for the paper as a whole.

All authors attest to meeting the four ICMJE.org authorship criteria: (1) Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND (2) Drafting the work or revising it critically for important intellectual content; AND (3) Final approval of the version to be published; AND (4) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding and support: By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The authors have stated that no such relationships exist.

Publication dates: Received for publication December 22, 2017. Revisions received May 22, 2018, and October 9, 2018. Accepted for publication October 29, 2018.

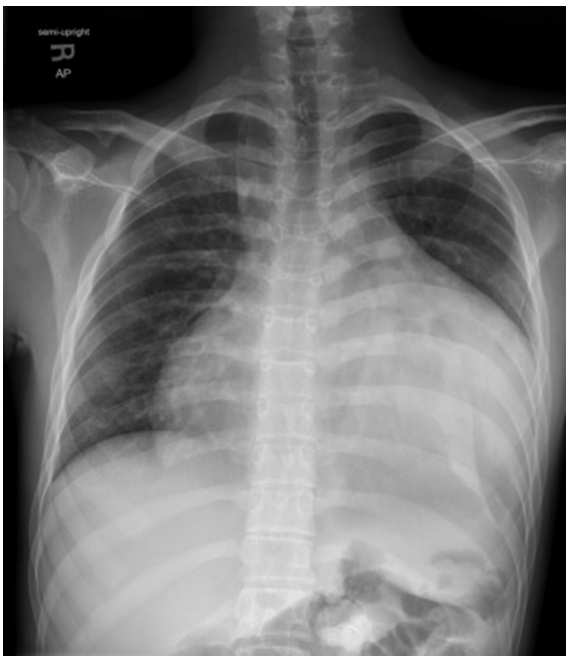
REFERENCES

1. Cantrill SV, Brown MD, Brecher D. Clinical policy: use of intravenous tissue plasminogen activator for the management of acute ischemic stroke in the emergency department. *Ann Emerg Med*. 2015;66:322-333.
2. Diercks DB, Mehrotra A, Nazarian DJ, et al. Clinical policy: critical issues in the evaluation of adult patients presenting to the emergency department with acute blunt abdominal trauma. *Ann Emerg Med*. 2011;57:387-404.
3. Fesmire FM, Bernstein D, Brecher D. Clinical policy: critical issues in the evaluation and management of adult patients presenting to the emergency department with seizures. *Emerg Med*. 2014;63:437-447.
4. Jagoda AS, Bazarian JJ, Bruns JJ Jr, et al. Clinical policy: neuroimaging and decisionmaking in adult mild traumatic brain injury in the acute setting. *J Emerg Nurs*. 2009;35:e5-e40.
5. Sterne JAC, Cox DR, Smith GD. Sifting the evidence—what's wrong with significance tests? another comment on the role of statistical methods. *BMJ*. 2001;322:226-231.
6. Rozeboom WW. The fallacy of the null-hypothesis significance test. *Psychol Bull*. 1960;57:416-428.
7. Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol*. 1998;51:355-360.
8. Rothman KJ. Significance questing. *Ann Intern Med*. 1986;105:445-447.
9. McIlroy D. Seduced by a P-value. *Anaesth Intensive Care*. 2014;42:551-554.
10. Ioannidis JPA. The proposal to lower P value thresholds to .005. *JAMA*. 2018;319:1429-1430.
11. 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.
12. Aufderheide TP, Frascone RJ, Wayne MA, et al. Standard cardiopulmonary resuscitation versus active compression-decompression cardiopulmonary resuscitation with augmentation of negative intrathoracic pressure for out-of-hospital cardiac arrest: a randomised trial. *Lancet*. 2011;377:301-311.
13. Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility index? *Crit Care Med*. 2016;44:e1142-e1143.
14. Gerhardt RT, Hermstad E, Crawford DM, et al. Postdischarge secobarbital after ED migraine treatment decreases pain and improves resolution. *Am J Emerg Med*. 2011;29:86-90.
15. Leung J, Duffy M, Finckh A. Real-time ultrasonographically-guided internal jugular vein catheterization in the emergency department increases success rates and reduces complications: a randomized, prospective study. *Ann Emerg Med*. 2006;48:540-547.
16. US Department of Health and Human Services. Protection of human subjects. Available at: https://www.hhs.gov/ohrp/sites/default/files/ohrp/humansubjects/regbook2013.pdf.pdf. Accessed October 2, 2017.
17. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol*. 2011;64:395-400.
18. Kim H-Y. Statistical notes for clinical researchers: chi-squared test and Fisher's exact test. *Restor Dent Endod*. 2017;42:152-155.
19. Chae J, Taylor DM, Frauman AG. Tropisetron versus metoclopramide for the treatment of nausea and vomiting in the emergency department: a randomized, double-blinded, clinical trial. *Emerg Med Australas*. 2011;23:554-561.
20. Baş M, Greve J, Stelter K, et al. A randomized trial of icatibant in ACE-inhibitor-induced angioedema. *N Engl J Med*. 2015;372:418-425.
21. Spiro DM, Tay K-Y, Arnold DH, et al. Wait-and-see prescription for the treatment of acute otitis media: a randomized controlled trial. *JAMA*. 2006;296:1235-1241.
22. Sun BC, McCreath H, Liang L-J, et al. Randomized clinical trial of an emergency department observation syncope protocol versus routine inpatient admission. *Ann Emerg Med*. 2014;64:167-175.
23. 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.
24. Walter SD. Statistical significance and fragility criteria for assessing a difference of two proportions. *J Clin Epidemiol*. 1991;44:1373-1378.

25. Matics TJ, Khan N, Jani P, et al. The fragility index in a cohort of pediatric randomized controlled trials. *J Clin Med Res*. 2017;6:79.
26. 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.
27. 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.
28. Evaniew N, Files C, Smith C, et al. The fragility of statistically significant findings from randomized trials in spine surgery: a systematic survey. *Spine J*. 2015;15:2188-2197.
29. Edwards E, Wayant C, Besas J, et al. How fragile are clinical trial outcomes that support the CHEST clinical practice guidelines for venous thromboembolism? *Chest*. 2018;154:512-520.
30. Kruse CB, Vassar MB. Unbreakable? an analysis of the fragility of randomized trials that support diabetes treatment guidelines. *Diabetes Res Clin Pract*. 2017;134:91-105.
31. Docherty KF, Campbell RT, Jhund PS, et al. How robust are clinical trials in heart failure? *Eur Heart J*. 2017;38:338-345.
32. Levine AC, Becker J, Lippert S, et al. Emergency Medicine Resident Association International Emergency Medicine Literature Review Group. International emergency medicine: a review of the literature from 2007. *Acad Emerg Med*. 2008;15:860-865.
33. Becker TK, Hansoti B, Bartels S, et al. Global emergency medicine: a review of the literature from 2016. *Acad Emerg Med*. 2017;24:1150-1160.
34. Lee DK. Alternatives to P value: confidence interval and effect size. *Korean J Anesthesiol*. 2016;69:555-562.
35. Schriger DL. Problems with current methods of data analysis and reporting, and suggestions for moving beyond incorrect ritual. *Eur J Emerg Med*. 2002;9:203-207.
36. Kruschke JK, Liddell TM. The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon Bull Rev*. 2018;25:178-206.
37. Kruschke JK. Bayesian estimation supersedes the t test. *J Exp Psychol Gen*. 2013;142:573-603.
38. Kruschke J. Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan. Cambridge, MA: Academic Press; 2015.

Images in Emergency Medicine

The *Annals* Web site (*www.annemergmed.com*) contains a collection of hundreds of emergency medicine-related images, complete with brief discussion and diagnosis, in 18 categories. Go to the Images pull-down menu and test your diagnostic skill today. Below is a selection from the Cardiovascular Images.



"Adolescent With Chest Pain" by Neal and Rempell, June 2017, Volume 69, #6, pp. 687, 713.