AMERICAN ACADEMY OF OPHTHALMOLOGY®

Fragility of Results in Ophthalmology Randomized Controlled Trials

A Systematic Review

Carl Shen, MD,[1] Isabel Shamsudeen, BHSc,[1] Forough Farrokhyar, PhD,[2] Kourosh Sabri, MD[1,3,4]

Purpose: Evidence-based medicine is guided by our interpretation of randomized controlled trials (RCTs) that address important clinical questions. Evaluation of the robustness of statistically significant outcomes adds a crucial element to the global assessment of trial findings. The purpose of this systematic review was to determine the robustness of ophthalmology RCTs through application of the Fragility Index (FI), a novel metric of the robustness of statistically significant outcomes.

Design: Systematic review.

Methods: A literature search (MEDLINE) was performed for all RCTs published in top ophthalmology journals and ophthalmology-related RCTs published in high-impact journals in the past 10 years. Two reviewers independently screened 1811 identified articles for inclusion if they (1) were a human ophthalmology-related trial, (2) had a 1:1 prospective study design, and (3) reported a statistically significant dichotomous outcome in the abstract. All relevant data, including outcome, *P* value, number of patients in each group, number of events in each group, number of patients lost to follow-up, and trial characteristics, were extracted. The FI of each RCT was calculated and multivariate regression applied to determine predictive factors.

Results: The 156 trials had a median sample size of 91.5 (range, 13–2593) patients/eyes, and a median of 28 (range, 4–2217) events. The median FI of the included trials was 2 (range, 0–48), meaning that if 2 non-events were switched to events in the treatment group, the result would lose its statistical significance. A quarter of all trials had an FI of 1 or less, and 75% of trials had an FI of 6 or less. The FI was less than the number of missing data points in 52.6% of trials. Predictive factors for FI by multivariate regression included smaller *P* value ($P < 0.001$), larger sample size ($P = 0.001$), larger number of events ($P = 0.011$), and journal impact factor ($P = 0.029$).

Conclusions: In ophthalmology trials, statistically significant dichotomous results are often fragile, meaning that a difference of only a couple of events can change the statistical significance. An application of the FI in RCTs may aid in the interpretation of results and assessment of quality of evidence. *Ophthalmology 2017;■:1–7 © 2017 by the American Academy of Ophthalmology*

Supplemental material available at www.aaojournal.org.

In modern-day evidence-based medicine, randomized controlled trials (RCTs) represent a cornerstone on which we base our clinical decisions.[1] Statistical methods permeate the ophthalmic literature[2] and are meant to aid in our interpretation of the results of these trials as to whether or not true differences exists. *P* values are one such widely used statistic, but recent literature has highlighted the potential pitfalls in their interpretation.[3–6] This has been underscored by the difficulty in replicating "statistically significant" findings through reanalysis, let alone in repeating trials.[7,8] A natural extension that has emerged from this junction of *P* values and hypothesis testing is a question of the robustness of the data from which our conclusions are drawn. In other words, how might the conclusions have differed with slight changes in the results?

The Fragility Index (FI) is a recently described metric intended to assist in assessing the robustness of statistically significant dichotomous outcomes.[9] The FI is defined as the minimum number of patients whose status would have to change from a non-event to an event in the treatment group until statistical significance is lost. Events refer to the occurrence of dichotomous outcomes defined by the trials, such as graft failure, best-corrected visual acuity >20/40, etc.

For example, a plausible RCT investigating a medical (group A) vs. surgical (group B) management of glaucoma may report a superiority in group B if the primary outcome of failure occurred in 30 of 100 vs. 46 of 100 patients in group A ($P = 0.029$, by Fisher exact test). However, if 2 more patients in group B had met criteria for failure instead, statistical significance would have been lost ($P > 0.05$). This would equate to an FI of 2. In practical terms, these 2 events could represent any missing data, such as losses to follow-up or missed appointments, or simply reflect chance group imbalances.

https://doi.org/10.1016/j.ophtha.2017.11.015
ISSN 0161-6420/17

It has been shown that even in large trials from widely cited journals (*The New England Journal of Medicine*; *JAMA: The Journal of the American Medical Association*; *The Lancet*, etc.), the statistical significance of results is often fragile.[9] We hypothesized that RCTs in ophthalmology, where large trials represent only a small portion of the literature, would demonstrate similar fragility. Our primary objective was to explore the robustness of results in RCTs in ophthalmology through application of the FI. Our secondary objective was to identify factors associated with FI in ophthalmology.

Methods

Eligibility Criteria

We performed a search of the MEDLINE database between January 1, 2005, and October 30, 2016, to identify ophthalmology-related trials. We included trials that (1) were human RCTs, (2) had a 1:1 parallel 2-arm study design, and (3) reported at least 1 statistically significant dichotomous outcome in the abstract. Studies were excluded if the dichotomous outcome was analyzed as a time-to-event variable. Nested cohort, cross-sectional, or post hoc studies done on RCT data sets were excluded. Noninferiority trials were included if they subsequently conducted superiority analysis that satisfied inclusion criteria.

Search Strategy

A search strategy developed in conjunction with an academic librarian independently by 2 authors was used to identify RCTs with ophthalmology-related medical subject headings (MESH) headings and keywords in high-impact general medicine journals (*The New England Journal of Medicine*, *The Lancet*, *Journal of the American Medical Association*, *British Medical Journal*, *Canadian Medical Journal*) and any RCTs in high-impact general ophthalmology journals (*Ophthalmology*, *American Journal of Ophthalmology*, *JAMA Ophthalmology/Archives of Ophthalmology*, *British Journal of Ophthalmology*, *Survey of Ophthalmology*) based on Journal Citation Reports impact factor (accessed October 2016). The search strategy results were combined and duplicates removed. Studies were restricted to those reported in English. Full search strategy is available in Supplemental Materials (available at www.aaojournal.org).

Study Selection and Data Extraction

Two reviewers (C.S. and I.S.) independently screened titles and abstracts for eligibility criteria in duplicate. Full-text manuscripts were reviewed to determine final inclusion in the study. Duplicate articles were excluded. Any discrepancies were resolved by consensus decision.

Data were extracted using electronic data forms (C.S. and I.S.). Variables that were recorded consisted of date of publication, journal, funding source, single/multicenter, if a power calculation was performed and if target sample size was achieved, features of the outcome of interest (type [primary or secondary], *P* value, statistical test used, number of events), and missing patient data.[10] The Cochrane Risk of Bias Tool was used to grade systematic bias through assessment of randomization, allocation concealment, blinding, incomplete outcome assessment, selective reporting, and other sources of bias.

For each included RCT, we extracted data related to 1 statistically significant dichotomous outcome from the abstract. If more than 1 eligible outcome was reported, we chose the primary outcome if possible, or the most patient-relevant non−primary outcome as previously reported with Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology.[11]

Statistical Methods

The FI of each trial was calculated as originally described.[9] In brief, the results of each selected outcome were configured into a 2×2 contingency table as reported in the original trial. The *P* value of the outcome was then recalculated using a 2-sided Fisher exact test. Patients/eyes were then iteratively moved in the group with the fewest number of events from a non-event to an event, keeping overall sample size constant. The smallest number of patients/eyes required for the recalculated *P* value by Fisher exact test to be greater than or equal to 0.05 was recorded as the study's FI. An FI was recorded as zero if recalculation of the *P* value using Fisher exact test from the original statistical test used was greater than or equal to 0.05 without moving any events. FI was summarized based on study characteristics. Stepwise multivariable linear regression was performed to determine if study factors including journal impact factor, centers, funding source, outcome, intervention type, power calculation, *P* value, sample size, and number of events were associated with high FI. Coefficients with the corresponding confidence intervals (CI) with *P* values for the predicting factors and R^2 as measures of goodness of fit were reported. *P* values of < 0.05 were considered statistically significant. Statistical analysis was performed using SPSS version 20 (IBM Corporation, Armonk, NY).

Results

Trial Selection

The literature search identified 1811 potentially eligible studies. Selection of studies is outlined according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines in Figure 1. After review of abstracts, 1570 studies were excluded. On full text review, a further 85 studies were excluded resulting in 156 RCTs in our sample. The primary reason for exclusion during review was a non-1:1 parallel arm design.

Trial Characteristics

Trial characteristics are summarized in Table 1. Median total sample size in included RCTs was 91.5 (interquartile range [IQR] 60−221) and the median total number of events was 28 (IQR 14.25−65.75). The median number of missing patient data was 4 (IQR 0−16). Nearly all included RCTs were from ophthalmologic journals (155/156; 99.4%). Distribution of number of centers (single [59%] vs. multi [41%]), outcome (primary [45.5%] vs. secondary [54.5%]), and type of intervention (surgical [46.8%] vs. drug [43.6%]) were fairly even. Most studies were funded by government/nonprofit organizations (50/156 [32.1%]), although 45 trials (28.8%) did not report their funding. A power calculation was reported or referenced in 112 of 156 (71.8%) included trials.

Cochrane Risk of Bias

Cochrane Risk of Bias assessment is summarized in Table 2. The majority of trials were assessed as low risk of bias in randomization (69.2%), outcome assessor masking (51.3%), selective reporting (92.3%), and other sources of bias (70.5%). Most included trials were deemed high risk of bias in patient masking (42.3%) and surgeon masking (72.4%). An even number of trials were