Neurosurg Rev (2019) 42:9–14
DOI 10.1007/s10143-017-0870-8

REVIEW

The fragility of randomized controlled trials in intracranial hemorrhage

Yanfei Shen[1] · Xuping Cheng[1] · Weimin Zhang[1]

Received: 10 April 2017 / Revised: 17 May 2017 / Accepted: 6 June 2017 / Published online: 20 June 2017

Abstract Fragility of randomized controlled trials (RCTs) has been evaluated using a novel metric called fragility index (FI), which measures how many events the statistical significance of a dichotomous outcome depends on. This study aimed to evaluate the fragility of RCTs in intracranial hemorrhage. Literature search (PubMed/Embase) identified all RCTs of intracranial hemorrhage since 2006. The overall distribution of FI was evaluated. Subgroup and spearman correlation analyses were made to explore potential factors that may affect FI value. All the included RCTs were divided into two groups (positive and negative trials) according to the statistical significance of selected outcomes. Finally, 47 positive and 51 negative trials were included. Both the median FI ([2; IQR, 1–4] vs. [6; IQR, 4–9], $p < 0.001$) and the proportion of trials with FI ≤1 (2 vs. 18, $p < 0.001$) in positive trials were smaller than negative trials. In subgroup comparison within positive trials, sample size ([165; IQR, 87–200] vs. [83; IQR, 60–120], $p = 0.015$) and number of events ([35; IQR, 20–72] vs. [24; IQR, 11–32], $p = 0.015$) were higher in subgroup with FI >1 than the subgroup with FI ≤1. Weak positive correlations were found between FI and sample size and number of events. In the field of intracranial hemorrhage, trials reporting significant conclusions often depend on a small number of events. Compared to sample size, this phenomenon is more likely to be affected by statistical approach and trial methodology.

Keywords Fragility index · Randomized controlled trial · Intracranial hemorrhage · Review

Introduction

Rigorously conducted and adequately powered randomized controlled trials (RCTs) provide the most robust evidence for clinical practice. Conventionally, p value of 0.05 has been wildly used as statistic threshold to show the statistical significance for a specific intervention, which has also been frequently criticized for its simplicity and limitations in assessing clinical importance [2, 5, 9]. Statisticians indicated that when interpreting a positive outcome, issues such as sample size, magnitude of effect, outcome importance, trial methodology, potential random error, and conclusion application should all be considered in tandem with the p value itself [5, 8].

Despite all these criticisms, p value is still unlikely to be wholly replaced by another statistic; thus, an equally simple metric "fragility index (FI)" has been proposed as a complement of the p value of dichotomous outcomes [10]. For a trial reporting a statistically significant dichotomous outcome ($p < 0.05$), FI is defined as the smallest number of "non-events" that need to be changed to "events" to get the p value ≥0.05. In other words, a lower FI indicates less statistically robust result. Recent studies indicated that the median FI was only 2 (IQR, 1–3.5) of RCTs performed in critical area [6] and spinal surgery [3], which means adding only two events to one group of a trial could change the result from significant to non-significant.

However, the fragility of RCTs in field of intracranial hemorrhage has not been investigated. The primary objective of

Electronic supplementary material The online version of this article (doi:10.1007/s10143-017-0870-8) contains supplementary material, which is available to authorized users.

✉ Yanfei Shen
snow.shen@hotmail.com

[1] Department of Intensive Care Unit, Dongyang People's Hospital, No. 60, Wuning West Road, Dongyang, Jinhua, 322100 Zhejiang, People's Republic of China

this study is to assess the stability of significant dichotomous outcomes in RCTs with ICH patients, by systematically applying the FI. Our secondary objective was to make a comparison between positive and negative trials and identify potential factors which may affect FI value.

Methods

Study selection

PubMed and Embase were searched to identify all RCTs performed in ICH patients, from 1 January 2006 to now, with no limitation on the interventions and outcomes. Only trials designed as 1:1 two-group and reported at least one dichotomous outcome (either significant or non-significant) were included. Studies using multiple comparisons or using a quasi-randomized method were excluded. Assessment of the eligibility of studies was performed by two authors (Yanfei Shen and Weimin Zhang). Differences in opinions were resolved by discussion until consensus was reached.

Risk of bias

Two authors independently assessed the risk of bias and extracted data. Risk of bias was assessed using the Cochrane Collaboration Risk of Bias tool, which includes items of: sequence generation, allocation concealment, blinding of clinicians and outcome assessors, loss to follow-up and missing data. However, a large proportion of patients with ICH were unconscious; thus, blinding of patients was defined as unclear in this review. Selective reporting bias was difficult to detect and thus was not assessed.

Data extraction

All the included RCTs were divided into two groups. Trials that reported statistically significant dichotomous outcomes were defined as “positive trial,” otherwise defined as “negative trial.” For trials that reported more than one eligible outcome, we chose the primary outcome whenever possible and, otherwise, chose the most important secondary outcome using the following order: mortality, kinds of neurological symptom scores, and cerebral disorders. When the selected outcome was reported more than once, the earliest time point was used.

The following information was extracted: journal name, publication year, funding source, sample size for each arm, loss to follow-up for each arm, number of events for each arm, study period (defined as the study period until selected outcome was finished), reported p value, selected outcome and whether the outcome was primary or secondary, Jadad score, and the Thomson Reuters journal impact factor.

Calculation of FI

We calculated the FI of each outcome using a two-by-two contingency table according to the method described by Walsh et al. [10]. For RCTs with a significant dichotomous outcome, we first recalculated the p value using the two-sided Fisher exact test. We then iteratively added one event to the group with the lower event incidence while subtracting one non-event in the same group to preserve the total number. The smallest number of additional events required to eliminate the significance ($p \geq 0.05$) represented the FI of positive trial (positive FI). For RCTs with non-significant outcome, the calculation of FI is similar. We iteratively transferred a non-event to an event in the group with higher event incidence until the p value cross the 0.05 threshold, using the smallest number of additional events as negative FI. Both positive and negative FI were calculated using the intention to treat (ITT) method and the two-sided Fisher exact test. Thus, in trials using the per protocol (PP) method or the Pearson chi-square test, the statistical significance may be lost by simply changing the statistical method. In this condition, the FI was defined as zero. An example of FI calculation is shown in Online Resource 1.

Statistical method

All the continuous variables were skewed in the present study thus were presented as medians with interquartile ranges (IQR) and the Wilcoxon rank-sum test was used. Categorical variables were presented as a percentage and compared using the Fisher exact test. We also analyzed the correlation between FI and study characteristics using the Spearman correlation method. All statistical analysis was performed using the software STATA 11.2 (College Station, TX, USA). All tests were two-sided, and a significance level of 5% was used.

Results

Search results

Initial searching and screening returned 214 results plus 3 articles from cross referencing. Overall, 98 studies met criteria for analysis, including 47 positive trials and 51 negative trials. Details of the search strategy and the overall bias risk were present in Online Resource 1 (Table SI).

Comparison between positive and negative trials

The median FI in positive trials were significantly smaller than negative trials ([2; IQR, 1–4] vs. [6; IQR, 4–9], $p < 0.001$) (Table 1). Five (18.5%) positive trials had a FI of 0 as they lost their statistical significance when we simply recalculated their

Table 1 Comparison of study characteristics between positive and negative trials

Variables	Positive trials (*n* = 47)	Negative trials (*n* = 51)	*p*
Number of centers [median (IQR)]	1 (1–5)	1 (1–6)	0.860
Sample size [median (IQR)]	110 (72–184)	95 (48–238)	0.418
Number of events [median (IQR)]	31 (15–65)	24 (10–64)	0.330
Loss to follow-up [median (IQR)]	0 (0–3)	0 (0–6)	0.570
Impact factor [median (IQR)]	3.32 (1.28–4.95)	3.44 (1.41–5.79)	0.282
Study period			
Until hospital discharge	18 (38.2)	27 (52.9)	0.161
Less than 90 days	16 (34.0)	14 (27.5)	0.571
More than 90 days	13 (27.6)	10 (19.6)	0.475
All outcomes			
Primary outcome [*n* (%)]	23 (48.9)	32 (62.7)	0.222
Death [*n* (%)]	8 (17.0)	10 (19.6)	0.789
Neurological scores[a] [*n* (%)]	7 (14.9)	13 (25.5)	0.219
Cerebral disorder [*n* (%)]	20 (42.5)	27 (52.9)	0.320
Funded [*n* (%)]	23 (48.9)	20 (39.2)	0.416
FI [median (IQR)]	2 (1–4)	6 (4–9)	<0.001
FI ≤ 1 [*n* (%)]	18 (38.3)	2 (3.9)	<0.001
FI ≤loss to follow-up	14 (29.8)	10 (19.6)	0.242

[a] Neurological scores including activities of daily living score (*ADL*), glasgow coma scale (*GOS*), modified rankin scale (*MRS*) and NIH stroke scale (*NIHSS*). *FI* fragility index, *IQR* interquartile range

p values using the ITT method and the two-sided Fisher exact test (Fig. 1). Thirteen (27.7%) positive trials had a FI of 1. And the total proportion of FI ≤1 is significantly lower in negative FI group (2 vs. 18, $p < 0.001$). However, no significant difference was detected in study characteristics between positive and negative trials, including number of centers, sample size, number of events, loss to follow-up, study period, journal impact factor, and outcome categories (Table 1). Besides, the comparisons of positive and negative FI restricted to per disease and per selected outcome were also made in Online Resource 1 (Table SII and SIII), and the median (IQR) FI for SAH, intracranial hemorrhage, kinds of neurological scores, death and cerebral vasospasm are 2 (1–4), 3 (1–6), 3 (2–4), 2.5 (2–5) and 4 (1–4) in positive trials, respectively. However, the number of included studies became relatively small and may lose the statistical power to detect the significant difference.

Subgroups comparison (FI ≤1 and FI >1) within positive trials

The proportion of trials with FI ≤1 was 38.3% within positive trials (Table 2). Compared to FI ≤1 group, the sample size ([median, 16.5; IQR, 87–200] vs. [median, 83; IQR, 60–120], $p = 0.015$) and the number of events ([median, 35; IQR, 20–72] vs. [median, 24; IQR, 11–32], $p = 0.015$) were significantly higher in the FI >1 group. However, the other trial characteristics including study center, loss to follow-up, Jadad score, or publishing year did not show any statistical difference between these two subgroups.

Correlation between FI and trial characteristics

We found a weak positive correlation between FI and sample size and number of events both in positive ($r_s = 0.330$, $p = 0.023$; $r_s = 0.448$, $p = 0.001$, respectively) and negative trials ($r_s = 0.382$, $p = 0.008$; $r_s = 0.309$, $p = 0.005$, respectively) (Table 3). The *p* value was negatively correlated with FI value in positive trials ($r_s = -0.757$, $p < 0.001$), and positively correlated in negative trials ($r_s = 0.695$, $p < 0.001$). Number of study centers and journal IF were moderately correlated with FI value in negative trials; however, these correlations were not significant in positive trials. Other study characteristics showed no significant correlation with FI. An overview of included studies was presented in Online Resource 1 (Table SIV).

Discussion

Key findings

We systematically reviewed the literature to identify 1:1 designed RCTs in the field of intracranial hemorrhage reporting at least one important dichotomous outcome and found that the median FI of positive trials was only 2. 38.3% of positive trials had a FI ≤1. In addition, we also found that within the positive trials, the sample size and number of events were significantly lower in the subgroup with lower FI (FI ≤1).

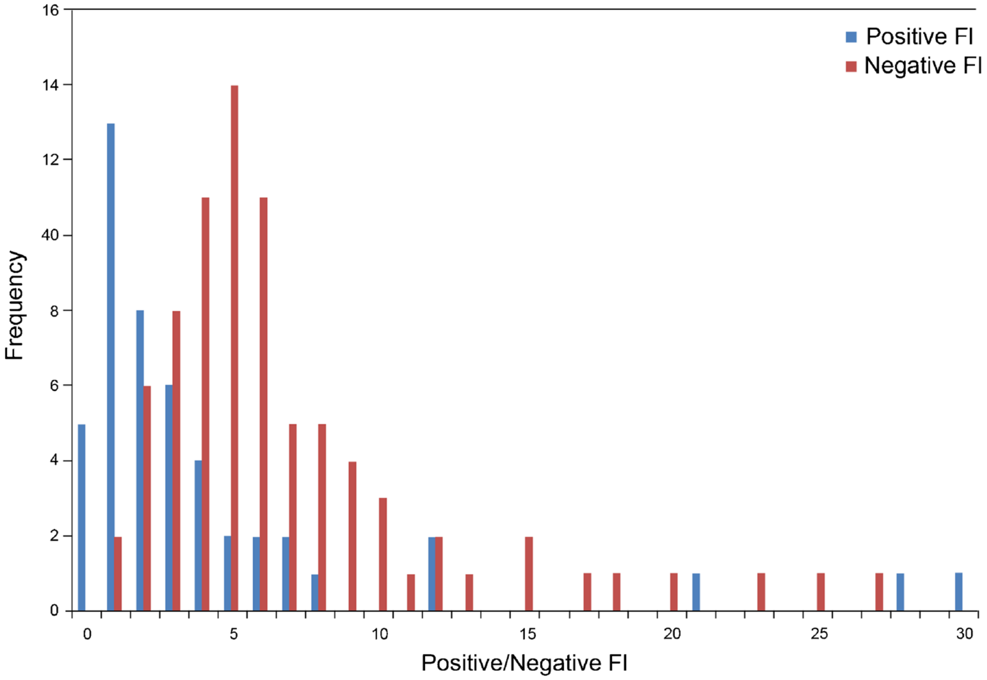

Fig. 1 Fragility index distribution of positive and negative trials

Finally, we found that both sample size and number of events were weakly positively correlated with the FI.

Positive FI and negative FI

Statistical analysis of biomedical research relies on *p* value to demonstrate significance, with a traditional threshold of 0.05. However, it has been criticized over decades [5, 9] for its simplicity and limitations as it can be easily affected by sample size, methodology bias, random error, and statistical approach [5, 8]. Given the *p* value is unlikely to be thoroughly replaced by another statistic, addition of an equally simple metric such as the FI was proposed as a complement of *p* value [10]. A recent study [6], covering multiple-center RCTs in critical area published in high-impact general journals, reported that the median FI was only 2 and greater than 40% of trials had a FI ≤1. This finding is quite similar to ours that in positive trials, the median FI was only 2, and 38.3% of positive trials had a FI ≤1. This is quite a frustrating finding as only one transition could eliminate the significance in these trials. Furthermore, we also noticed that FI of 14 positive trials (29.8%) was smaller than loss to follow-up, which is a very important issue both in clinical trials and in this review. In order to get more stable results, ITT approach is wildly adopted in superiority trials. And the conclusion may be less stable in superiority trials only using the PP method as we

Table 2 Subgroup comparisons between FI ≤1 and FI >1 within positive trials

Variables	FI ≤1 (*n* = 18)	FI >1 (*n* = 29)	*p*
Sample size [median (IQR)]	83 (60–120)	165 (87–200)	0.015
Number of events [median (IQR)]	24 (11–32)	35 (20–72)	0.029
Loss to follow-up [median (IQR)]	0 (0–0)	1 (0–3)	0.050
Journal impact factor [median (IQR)]	2.48 (1.06–3.78)	3.41 (2.12–5.78)	0.105
Reported *p* value	0.027 (0.014–0.037)	0.005 (0.002–0.010)	<0.001
Funded [*n* (%)]	10 (55.6)	13 (44.8)	0.556
Multiple centers	5 (27.8)	11 (37.9)	0.541
Year of publishing (≤2013)	9 (50)	15 (51.7)	1.000
Jadad score			
Random sequence generation	11 (61.1)	18 (62.1)	1.000
Allocation concealment	6 (33.3)	8 (27.6)	0.749
Blinding of investigators	2 (11.1)	7 (24.1)	0.449
Blinding of outcome assessors	8 (44.4)	14 (48.3)	1.000
Incomplete outcome data	13 (72.2)	20 (69.0)	1.000

Table 3 Correlation of fragility index with study characteristics

Study characteristics	r_s* for positive FI (n = 47)	p	r_s for negative FI (n = 51)	p
Total sample size	0.330	0.023	0.448	0.001
Total number of events	0.382	0.008	0.309	0.005
Number of loss to follow-up	0.261	0.076	0.152	0.286
Reported p value	−0.757	<0.001	0.695	<0.001
Study center	0.160	0.281	0.325	0.019
Journal Impact factor	0.206	0.164	0.299	0.033

*r_s: correlation coefficient using the Spearman correlation method

noticed that one study [4] loses the statistical significance when we simply change the PP to ITT method. On the other hand, even both ITT and PP analyses suggested a significant difference, there is still a possibility to get a real-world negative outcome, especially in trials with FI <loss to follow-up. Thus, caution should be raised when interpreting and applying these findings to patient care.

Despite two completely different scales of p value being used for significant and non-significant outcomes, we also evaluated the FI of RCTs with non-significant dichotomous outcomes in our review. Compared to that of positive trials, it is unsurprising that the FI of negative trials was larger as it had a broader p value range of 0.05 to 1. Noteworthy, only two negative trials (3.9%) had a FI ≤1 and other study characteristics were non-significant compared to those of positive trials (Table 1). Thus, we boldly speculated that this asymmetry of FI distribution (Fig. 1) may be largely caused by publication bias instead of other study characteristics, as the preponderance of positive findings has always unavoidably existed. In the present review, we noticed a large proportion of trials with low FI, five of which lost their significance when we simply changed the statistical approach. Furthermore, kinds of neurological scores were reported as main outcomes in the included studies. Thus, well-designed methodology, especially blinding [7], was critically important as these scores could be easily interfered by the unblinding of the outcome assessors. On the other hand, recording neurological scores as binary data (such as good outcome, poor outcome) was less statistically efficient than using the original ordered data [1] in these patients. Thus, we suggested that the statistical approach, interpretation of missing value, and the results we chose to submit for publication should be carefully reexamined to avoid potential biases or distortions, especially in trials with low FI.

Subgroup comparisons

To further explore the confounded factors of low FI, we performed a subgroup comparison within positive trials, divided by FI value (FI ≤1 or FI >1). Sample size and number of events were significantly lower while reported p value was higher in the FI ≤1 group, which was consistent with the outcomes in RCTs of spine surgery [3]. Besides, a weak positive correlation between FI and sample size and number of events was detected. For a real-world effective intervention, a larger sample size could lead to a more robust conclusion. However, trials with a large sample size may also have a low FI if the effectiveness of the intervention did not even exist [11]. Besides, large sample-sized RCTs are difficult to perform as these need a lot of manpower and material resources. Thus, we thought for positive trials with low FI; it is more reasonable to focus on the statistical approach and study methodology first before simply increasing the sample size.

Strength and limitations

The strength of this study lies in the systematic search approach, with no limitation on intervention and outcomes. The comparison between positive and negative trials made our results more comprehensive; however, it also brought a notable limitation. Under a null trial, it has an α chance of being significant, and under the alternate, the trial will have a β chance of being non-significant, and the probabilities of obtaining a false positive and false negative trial will not be the same. Thus, the higher FI of negative trials in our review is unsurprisingly and should be interpreted with caution. Besides, several other limitations should also be noted. First, this review yields a vast array of diseases and selected outcomes, which may be less useful for disease-specific population. However, the primary aim of this review is to provide an overall view of the fragility of RCTs in the field of intracranial hemorrhage and to provide clinicians with this cautionary tale. Second, we arbitrary set the time limitation for literature search. However, we thought a decade was long enough to reflect the overall updating of a specific field. Third, the reported dichotomous outcomes were diversified, and sometimes, it is difficult to choose (for example, both the mortality and neurological score may be reported in one trial as secondary outcomes). Under this condition, the outcome was selected using a predefined order as described above, and all the discrepancies were resolved by discussion.

Conclusion

In the field of intracranial hemorrhage, trials reporting positive dichotomous outcomes often depend on a small number of events. Compared to sample size, this phenomenon is more likely to be affected by publication bias, statistical approach, and trial methodology. Cautions should be raised when interpreting trials with a low FI. We suggest that FI should be wildly adopted as a complement of *p* values to allow easy identification of trials with less robust results.

Acknowledgements We thank editor Monica Liu of JECCM for her help in revision.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This is a systematic review and ethics committee approval was waived for this paper.

Informed consent None.

References

1. Bath PM, Gray LJ, Collier T, Pocock S, Carpenter J (2007) Can we improve the statistical analysis of stroke trials? Statistical reanalysis of functional outcomes in stroke trials. Stroke 38:1911–1915. doi:10.1161/STROKEAHA.106.474080
2. Chakkera HA, Schold JD, Kaplan B (2016) P value: significance is not all black and white. Transplantation 100:1607–1609. doi:10.1097/TP.0000000000001331
3. Evaniew N, Files C, Smith C, Bhandari M, Ghert M, Walsh M, Devereaux PJ, Guyatt G (2015) The fragility of statistically significant findings from randomized trials in spine surgery: a systematic survey. The spine journal : official journal of the North American Spine Society 15:2188–2197. doi:10.1016/j.spinee.2015.06.004
4. Li JY, Yuan LX, Zhang GM, Zhou L, Gao Y, Li QB, Chen C (2016) Activating blood circulation to remove stasis treatment of hypertensive intracerebral hemorrhage: a multi-center prospective randomized open-label blinded-endpoint trial. Chinese Journal of Integrative Medicine 22:328–334
5. Pocock SJ, Stone GW (2016) The primary outcome is positive—is that good enough? N Engl J Med 375:971–979. doi:10.1056/NEJMra1601511
6. Ridgeon EE, Young PJ, Bellomo R, Mucchetti M, Lembo R, Landoni G (2016) The fragility index in multicenter randomized controlled critical care trials. Crit Care Med 44:1278–1284. doi:10.1097/CCM.0000000000001670
7. Shepherd BE, Shaw PA, Dodd LE (2012) Using audit information to adjust parameter estimates for data errors in clinical trials. Clin Trials 9:721–729. doi:10.1177/1740774512450100
8. Thiese MS, Ronna B, Ott U (2016) P value interpretations and considerations. J Thorac Dis 8:E928-E931. doi:10.21037/jtd.2016.08.16
9. Thomas LE, Pencina MJ (2016) Do not over (P) value your research article. JAMA cardiology 1:1055. doi:10.1001/jamacardio.2016.3827
10. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, Molnar AO, Dattani ND, Burke A, Guyatt G, Thabane L, Walter SD, Pogue J, Devereaux PJ (2014) The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility index. J Clin Epidemiol 67:622–628. doi:10.1016/j.jclinepi.2013.10.019
11. Woods KL, Fletcher S, Roffe C, Haider Y (1992) Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2). Lancet (London) 339:1553-1558