Fragility Index

Shengping Yang PhD, Gilbert Berdine, MD

I recently conducted a randomized clinical trial with two arms–a new drug treatment and a standard treatment–and the goal was to investigate whether patients treated with the new drug have lower in-hospital mortality. In each arm, there were 50 patients, and the numbers of deaths were 3 and 12 in the new and the standard treatment groups, respectively. A Fisher's exact test gives a p value of 0.023. I tend to conclude that patients treated with the new drug had lower mortality. Should I be confident with the conclusion?

In many clinical studies, a statistical test is used to determine whether there is a difference between two groups. Often a conclusion is made based on whether the p value from such a statistical test is smaller than a pre-set threshold value, usually 0.05. Although using such a threshold translates into one false conclusion in every 20 conclusions made, it is so widely accepted and used that 0.05 is almost becoming the magic number in data interpretation and reporting.

While using 0.05 as the cut-off value is convenient and a conclusion can be readily reached, a drawback is that such a p value based conclusion might be overly simplified. For example, if the p value is 0.0499, then we conclude that the two groups differ significantly; and if the p value is 0.0501, then we conclude that there is no significant difference between the two groups. As we can see, these two p values are very similar, but the conclusions are opposite, and this is due to the use of a hard cut-off value. P value is also associated with sample size. Consider two studies with very different sample sizes; even though the two p values obtained from the two studies are the same (for example to the 5th decimal place), they might have quite different implications. The Fragility Index (FI) was introduced as an attempt to use an additional metric to assess how reliable it is to make a conclusion in a two-arm randomized trial. Specifically, the FI is the minimum number of patients whose outcome would need to change from a non-event to an event to turn a statistically significant result into a non-significant one. For the trial mentioned above, because there were 3 deaths in the new treatment group and 12 deaths in the standard treatment group, a Fisher's exact test gives a p value of 0.023 (Table 1A). Now, if we keep the result for the standard treatment group (the group with higher event rate) unchanged, and change one patient in the new treatment group from alive to dead, then the p value from the Fisher's test would change to 0.054 (Table 1B). Based on these p values, in order to change the p value from less than 0.05 to greater than 0.05, it requires the change of status of one patient in the new treatment group from alive to dead, and thus, the FI is one.

Now, suppose that there is a much bigger study, in which each arm has 500 patients. Also, suppose that the numbers of deaths are 30 and 50 in the new and standard treatment groups, respectively. Then

Table 1. True and modified results

A. True result (p = 0.023)			B. Modified result (p = 0.054)		
	New treatment	Standard treatment		New treatment	Standard treatment
Expired	3	12	Expired	4	12
Alive	47	38	Alive	46	38

DOI: 10.12746/swrccc.v7i27.526

Table 2. True and modified results

A. True result (p = 0.026)			B. Modified result (p = 0.0364)		
	New treatment	Standard treatment		New treatment	Standard treatment
Expired	30	50	Expired	31	50
Alive	470	450	Alive	469	450

C. True result (p = 0.0495)			D. Modified result (p = 0.066)		
	New treatment	Standard treatment		New treatment	Standard treatment
Expired	32	50	Expired	33	50
Alive	468	450	Alive	467	450

the p value for comparing the two groups is 0.026 (Table 2A), which is very close to the p value in Table 1A. Again, if we keep the result for the standard treatment group unchanged, and change one patient in the new treatment group from alive to dead, then the p value would change to 0.0364 (Table 2B). In fact, if we change the status of another patient in the new treatment group from alive to dead, the p value is still less than 0.05. However, if we change three patients (in total) from alive to dead, then the p value would change to 0.066 (Table 2D), which is greater than 0.05. Therefore, for this study, the FI is three.

It is clear that the FI is in general greater for larger studies. However, a few other factors also affect the FI, such as the proportion of events and the true result absolute p value. For example, suppose that there are 500 patients in each arm, and the numbers of deaths are 200 (40%) and 236 (47.2%) for the new and standard treatment groups, respectively. Then the Fisher's test p value (0.026) is almost the same as the one in the Table 2A (note that the event rates are much higher). However, with the same sample size, even we change four patients in the new treatment group from alive to dead, the p value would still be less than 0.05 (Table 3B; 0.048). Furthermore, the true result p value can substantially affect the FI. Suppose that the p value from the Fisher's test is very small (e.g., 0.0004, Table 4A), then the FI can be quite large (>20).

While a smaller p value (below 0.05) is associated with stronger confidence in concluding that there is a significant difference, it is also true that the smaller the FI, the more fragile a trial's significant finding. Therefore, a negative correlation is expected between the p value and the FI. In fact, by using simulated data, some investigators suggested

Table 3. True and modified results

A. True result (p = 0.026)			B. Modified result (p = 0.048)		
	New treatment	Standard treatment		New treatment	Standard treatment
Expired	200	236	Expired	204	236
Alive	300	264	Alive	296	264

Table 4. True and modified results

A. True result (p = 0.0004)			B. Modified result (p = 0.048)		
	New treatment	Standard treatment		New treatment	Standard treatment
Expired	180	236	Expired	204	236
Alive	320	264	Alive	296	264

that the FI is simply a repackaging of the p value for a clinical trial. Although this suggestion is valid to a certain degree, it is still possible that the FI provides additional information should two studies have very similar p values.

The FI has a number of limitations; for example, it is limited to binary outcome with 1 to 1 randomization (nevertheless, trials with such a design are quite common). In fact, the FI is most suitable for two-group comparison; in order to apply the FI, trials with more than two groups might have to be compared two groups at a time, which could consequently introduce p value adjustments and much more complicated result presentation and interpretation. A more critical limitation of the FI is that there is no commonly agreed standard to interpret its value. Besides the commonly accepted notion that the smaller an FI, the more fragile a significant finding, the FI is often compared to the number of loss to follow-up. It is suggested that a trial's robustness should be questioned if the number of loss to follow-up patients exceeds the FI. However, the merit of this suggestion is subject to what disease is under study.

Fisher's exact test is known to be conservative. Therefore, it is possible that using the same data, the p value obtained from a certain test is significant, and from the Fisher's test is not. If this is the case, the FI for a trial is zero.

The FI is a simple and easy-to-use index that can help physicians/investigators to assess the fragility of a randomized trial. Many of the statistical properties of the FI are still unclear and thus warrant further investigations.

Keywords: fragility index, binary outcome, p value, Fisher's test

From: Department of Biostatistics (SY), Pennington Biomedical Research Center, Baton Rouge, LA; The Department of Internal Medicine (GB) at Texas Tech University Health Sciences Center in Lubbock, TX

Submitted: 1/7/2019
Conflicts of interest: none

References

1. Carter RE, McKie PM, Storlie CB. The Fragility Index: a P-value in sheep's clothing? European Heart Journal 2017;38: 346–48.
2. Mazzinari G, Ball L, Neto AS, etc. The fragility of statistically significant findings in randomised controlled anaesthesiology trials: systematic review of the medical literature. *British Journal of Anaesthesia* 2018;120(5):935–41.
3. Walsh M, Srinathan SK, McAuley DF, etc. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clinical Epidemiology 2014;67:622–287.