Pharmaceutical Statistics

WILEY

MAIN PAPER OPEN ACCESS

Beyond the Fragility Index

Piero Quatto[1,2,3] | Enrico Ripamonti[2,4] | Donata Marasini[3]

[1]Department of Economics, Management and Statistics, University of Milan-Bicocca, Milan, Italy | [2]Milan Center of Neuroscience, University of Milan-Bicocca, Milan, Italy | [3]Bicocca Applied Statistics Center (B-ASC), university of Milan-Bicocca, Milan, Italy | [4]Department of Economics and Management, University of Brescia, Brescia, Italy

Correspondence: Enrico Ripamonti (enrico.ripamonti@unibs.it)

Received: 24 December 2023 | **Revised:** 13 September 2024 | **Accepted:** 24 October 2024

Funding: The authors received no specific funding for this work.

Keywords: fragility index | likelihood ratio | randomized clinical trial | strength index

ABSTRACT

The results of randomized clinical trials (RCTs) are frequently assessed with the fragility index (*FI*). Although the information provided by *FI* may supplement the *p* value, this indicator presents intrinsic weaknesses and shortcomings. In this article, we establish an analysis of fragility within a broader framework so that it can reliably complement the information provided by the *p* value. This perspective is named the *analysis of strength*. We first propose a new strength index (*SI*), which can be adopted in normal distribution settings. This measure can be obtained for both significance and nonsignificance and is straightforward to calculate, thus presenting compelling advantages over *FI*, starting from the presence of a threshold. The case of time-to-event outcomes is also addressed. Then, beyond the *p* value, we develop the analysis of strength using likelihood ratios from Royall's statistical evidence viewpoint. A new R package is provided for performing strength calculations, and a simulation study is conducted to explore the behavior of *SI* and the likelihood-based indicator empirically across different settings. The newly proposed analysis of strength is applied in the assessment of the results of three recent trials involving the treatment of COVID-19.

1 | Introduction

The analysis of fragility, adopted to assess the results of randomized clinical trials (RCTs), is controversial. Fragility can be conceived as a proxy to measure the robustness of a given procedure. When small variations in a dataset can easily subvert the significance of a *p* value, such a result can be interpreted as *fragile*. A fragility index (*FI*) was proposed as early as 1990 [1] and later revised by Walsh, Srinathan, and McAuley [2], who defined it as the minimal number of outcomes whose status must change from "nonevent" to "event" to transform a significant result into a nonsignificant one. It is calculated by virtually adding events to the group containing fewer events. Docherty et al. [3] extended *FI* to nonsignificant results. *FI* is defined as *forward* in the case of significance and as *reverse* in the case of nonsignificance [4]. In the latter case, it is defined as the minimal number of outcomes whose status must change from "event" to "nonevent" to transform a nonsignificant result into a significant one [5]. More recently, many variants of *FI* have been discussed [4, 6].

FI has objective weaknesses. First, there is no threshold above which the result can be considered robust according to the common terminology in the literature. Second, it has a negative correlation with the *p* value, so large *p* values are considered fragile and small *p* values are considered robust. Thus, the use of *FI* has been put under scrutiny. For instance, Condon et al. [7] suggested applying *FI* parsimoniously. Niforatos et al. [8] recommended caution in its interpretation. Schröder, Muensterer, and von Sochaczewski [9] and Potter [10] argued that the use of *FI* should be avoided.

Despite this criticism, in the biomedical sciences, *FI* is still widely used in many applied contexts [5, 11–19]. Moreover, beyond its inherent weaknesses, *FI* is in line with one of the

statements of the American Statistical Association (ASA), namely, "scientific conclusions and business or policy decisions should not be based only on whether a p value passes a specific threshold." [20] It is, therefore, worth reconsidering the analysis of fragility, setting it in a broader context than is currently adopted. In this article, we refer to this new framework as the *analysis of strength*.

In the following sections, we present the analysis of strength along two main directions, following the advice of Wasserstein, Schirm, and Lazar [21], who recommended developing possible suggestions for "supplementing or replacing p values." In Section 2, we propose a new strength index (*SI*) based on normal distributions together with an adaptive threshold. *SI*, inspired by *FI*, can be considered a valid diagnostic of the p value, as it is used to reduce both false positives and false negatives. Indeed, if the index is larger than a prefixed threshold, then the experimental result as expressed by the p value is confirmed. By contrast, if the index is not larger than the threshold, the experimental result can be questioned. In the former case, the index supports the p value, and in the latter case, it does not, thus reducing the rate of false positives or false negatives. The case of time-to-event outcomes is also discussed, and a new index, the survival fragility index (*SFI*) is proposed. In Section 3, we present a more general framework that can be adopted as an alternative to the p value, built upon the likelihood theory, which has been targeted as a possible inferential framework in clinical research [22] and, in particular, for pharmacovigilance [23, 24]. We define the fragility in the context of a likelihood-based approach, and then we propose a new likelihood strength index (*LSI*). This is structurally similar to *SI* but is used independently of the p value. We then discuss the problem of setting the threshold for these indices. In Section 4, the newly proposed strength indices *SI* and *LSI* are extended to one-sided hypotheses, which are often used in RCTs. In Section 5, we report the results of a simulation study concerning the behavior of *SI* and *LSI*. In Section 6, we illustrate our proposal in relevant medical settings regarding three trials conducted to assess drugs for COVID-19. A simulated setting to calculate *SFI* is also presented. Finally, in Section 7, we draw our conclusions and make recommendations for practitioners.

2 | Analysis of Strength for 2×2 Binomial Trials

2.1 | Measuring Strength

As a general premise, the analysis of strength will be discussed for contexts in which changes from an event to a nonevent and vice versa have the same probability of occurring [25]. Let us consider two independent binomial random variables representing the number of events x_i in n_i binomial trials performed under two different treatments G_i, $i = 1, 2$. Hereafter, we will refer to a general setting of comparing hypotheses concerning the unknown proportions θ_1 and θ_2 of events in the two treatment groups, namely,

$$\mathrm{H}_0: \theta_1 = \theta_2 \text{ vs. } \mathrm{H}_1: \theta_1 \neq \theta_2 \tag{1}$$

We will indicate with $\frac{x_1}{n_1}$ and $\frac{x_2}{n_2}$ the proportions of events verified in each group. Without loss of generality, we assume $\frac{x_1}{n_1} \leq \frac{x_2}{n_2}$. In particular, $\widehat{\theta} = \frac{x_2}{n_2} - \frac{x_1}{n_1}$ is an unbiased estimate of $\theta = \theta_2 - \theta_1$ and

$$\widehat{\sigma}^2 = \frac{x_1}{n_1^2}\left(1 - \frac{x_1}{n_1}\right) + \frac{x_2}{n_2^2}\left(1 - \frac{x_2}{n_2}\right) \tag{2}$$

is a suitable estimate of its variance. Assuming normal approximations for the sample proportions, we introduce a quantity Q such that

$$2\Phi\left(-\frac{\left|\frac{x_1}{n_1} - \frac{x_2}{n_2} + Q\right|}{\widehat{\sigma}}\right) = \alpha \tag{3}$$

where Φ is the standard normal cumulative distribution function.

In general terms, Q represents a rational number that is added to the difference of the proportions obtained in the sample that leads to a change in the p value with respect to the significance level α. With some algebra, we can explicitly write Q as a function of the p value. First, Equation (3) is equivalent to

$$\left|\frac{x_1}{n_1} - \frac{x_2}{n_2} + Q\right| = \widehat{\sigma} z_{1-\frac{\alpha}{2}} \tag{4}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of a standard normal random variable. Hence, for Q we provide the explicit formula

$$Q = \frac{x_2}{n_2} - \frac{x_1}{n_1} - \widehat{\sigma} z_{1-\frac{\alpha}{2}} = \left(z_{1-\frac{p}{2}} - z_{1-\frac{\alpha}{2}}\right)\widehat{\sigma} \tag{5}$$

where $z_{1-\frac{p}{2}} = \frac{\frac{x_2}{n_2} - \frac{x_1}{n_1}}{\widehat{\sigma}} = \frac{\widehat{\theta}}{\widehat{\sigma}}$ and $p = 2\Phi\left(-\frac{\widehat{\theta}}{\widehat{\sigma}}\right)$ is the two-sided p value for the null hypothesis (1).

If the experiment leads to a significant result, then Equation (5) takes a positive value, whereas it takes a negative value for a nonsignificant result. In the case of significance, from Equation (4) it follows that

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} + Q = \frac{x_1 + n_1 Q}{n_1} - \frac{x_2}{n_2} = \frac{x_1}{n_1} - \frac{x_2 - n_2 Q}{n_2}$$

and, in this case, Q represents the value that leads to a change of status when adding $n_1 Q$ events to x_1 or subtracting $n_2 Q$ from x_2. The case of nonsignificance can be treated in a similar way. In the case of significance, from Equation (5) we have

$$0 \leq Q \leq \frac{x_2}{n_2} - \frac{x_1}{n_1} \tag{6}$$

and for non-significance,

$$-\widehat{\sigma} z_{1-\frac{\alpha}{2}} \leq Q \leq 0 \tag{7}$$

Based on Q and Equations (5) to (7), we introduce the index *SI*, defined as

$$SI = \begin{cases} \dfrac{Q}{\frac{x_2}{n_2} - \frac{x_1}{n_1}} = 1 - \dfrac{\hat{\sigma} z_{1-\frac{\alpha}{2}}}{\frac{x_2}{n_2} - \frac{x_1}{n_1}} = 1 - \dfrac{z_{1-\frac{\alpha}{2}}}{z_{1-\frac{p}{2}}}, & p \le \alpha \\ \dfrac{Q}{-\hat{\sigma} z_{1-\frac{\alpha}{2}}} = \dfrac{\frac{x_2}{n_2} - \frac{x_1}{n_1}}{-\hat{\sigma} z_{1-\frac{\alpha}{2}}} + 1 = 1 - \dfrac{z_{1-\frac{p}{2}}}{z_{1-\frac{\alpha}{2}}}, & p > \alpha \end{cases} \tag{8}$$

SI is a normalized indicator that takes a value in (0,1). The greater its value, the stronger the reliability of the p value. It is straightforward to determine how *SI* varies as a function of the p value (see Supporting Information S1: Figure 1).

2.2 | Fixed Thresholds

Once introduced theoretically, *SI* needs to be compared with an appropriate threshold to be adopted in applications. For any value assumed by *SI*, without such a threshold, it is uncertain whether *SI* should be interpreted as a proxy for the fragility of the experiment or as a proxy for its strength. This is an inherent weakness of the concept of fragility for which, to our knowledge, no suitable thresholds have been proposed in the literature. The proposal by Murad et al. [11] in cardiology has, instead, been defined as a rule of thumb [26, 27].

An intuitive threshold value τ for *SI* in the interval (0,1) is $\tau = 1/2$. However, in the common case $\alpha = 0.05$, this would imply that, with a significant result, strength can be attributed to the p value if and only if $p \le 2\Phi\left(-2z_{1-\frac{\alpha}{2}}\right) \cong 0.0001$. This value appears rather conservative, at least for biomedical applications.

An alternative is to adopt the thresholds proposed in the context of the *analysis of credibility*, a Bayesian approach to assessing experimental results. In the case of a significant result, to attain the credibility threshold, Matthews [28] requires that $p \le 2\Phi\left(-\sqrt{\frac{1+\sqrt{5}}{2}} z_{1-\frac{\alpha}{2}}\right) \cong 0.0127$. Held [29] instead requires that $p \le 2\Phi\left(-\sqrt{2} z_{1-\frac{\alpha}{2}}\right) \cong 0.0056$. In the analysis of strength, we recommend adopting Held's more conservative threshold, thus obtaining

$$\begin{cases} SI \ge 1 - \dfrac{1}{\sqrt{2}} \iff \dfrac{\frac{x_2}{n_2} - \frac{x_1}{n_1}}{\hat{\sigma}} \ge \sqrt{2} z_{1-\frac{\alpha}{2}} \iff p \le 2\Phi\left(-\sqrt{2} z_{1-\frac{\alpha}{2}}\right), & p \le \alpha \\ SI \ge 1 - \dfrac{1}{\sqrt{2}} \iff \dfrac{\frac{x_2}{n_2} - \frac{x_1}{n_1}}{\hat{\sigma}} \le \dfrac{z_{1-\frac{\alpha}{2}}}{\sqrt{2}} \iff p \ge 2\Phi\left(-\dfrac{z_{1-\frac{\alpha}{2}}}{\sqrt{2}}\right), & p > \alpha \end{cases} \tag{9}$$

2.3 | Adaptive Thresholds

Adopting a fixed threshold is a general and easily reproducible procedure; however, in some applied scenarios, it may not be enough to control for false positives or false negatives. This is an inherent problem of adopting a fixed threshold and can be observed in simulations. To solve this problem, we propose an adaptive threshold for *SI*, which can be modeled based on suitable parameters. It depends on the total sample size n and two other parameters τ and ε, with $0 < \tau < 1$ and $\varepsilon > 0$. Such a threshold, denoted by $t_{\tau,\varepsilon}(n)$, should meet two desirable properties, namely, (i) it should be a positive strictly increasing function of the total sample size and (ii) it should satisfy the asymptotic condition

$$\lim_{n \to \infty} t_{\tau,\varepsilon}(n) = \tau \tag{10}$$

The first property ensures that small sample sizes are handled appropriately with respect to larger ones (see Section 2.4). The second property guarantees that as the sample size increases, there exists a value $\tau < 1$ indicating whether the p value is supported or not. Among the possible families satisfying (i) and (ii), we propose

$$t_{\tau,\varepsilon}(n) = \tau - \frac{1}{n^{\varepsilon}} \tag{11}$$

where ε determines the rate of convergence to the value τ. For instance, if we set the parameters to $\tau = 1/2$ and $\varepsilon = 1/2$, it follows that

$$t_n = t_{1/2,1/2}(n) = \frac{1}{2} - \frac{1}{\sqrt{n}} \tag{12}$$

and with $\varepsilon = 1/3$, we have

$$t_n = t_{1/2,1/3}(n) = \frac{1}{2} - \frac{1}{n^{1/3}} \tag{13}$$

which is positive (because the total sample size is larger than two) and strictly increasing, and

$$\lim_{n \to \infty} t_n = \tau = \frac{1}{2} \tag{14}$$

agreeing with Equation (10). Expressions (12) and (13) represent particularizations of the parametric family (11), which will be the subject of the simulation study (see Section 5 and the Supporting Information). Another example of a parametric family satisfying the desirable properties (i) and (ii) is presented in the Supporting Information. Note that, beyond these desirable properties, there is not stringent theoretical support for the choice of these thresholds. The support instead comes from numerical work, but other options may be possible.

It is worth observing that $\tau = 1/2$ has a natural interpretation in terms of the central value of *SI*. The values $\varepsilon = 1/2$ and $\varepsilon = 1/3$ have also been proposed as corrections to n for the so-called fragility quotient (*FQ*), a relative version of *FI* [30]. We also remark that in the case of Equation (12), it follows that $t_n \ge 1 - \frac{1}{\sqrt{2}}$ for $n > 23$, and in the case of Equation (13), for $n > 112$.

2.4 | Remarks and Properties

The quantity Q in Equation (5) allows us to identify a relation between the newly proposed *SI* and the *FI* by introducing $FI^* = n_1 Q$, which represents the number of changes (from event to nonevent and vice versa) in G_1 (the group with the smaller proportion of events) needed to subvert a significant or

nonsignificant p value. If $x_1 \leq x_2$, FI^* approximates FI based on the normal distribution, namely, $FI \cong n_1 Q$ for large sample sizes.

SI satisfies some desirable properties for a statistical procedure [10, 31, 32]; for example, it is easy to implement, it does not depend on the unknown intentions of the researcher through priors, and it leads different researchers to reach the same conclusions under the same conditions and with the same experimental results.

SI is not in agreement with Lindley and Scott's claim [33] that "significant at 5% depends on the sample size: it is more indicative of the falsity of the null hypothesis with a small sample than with a large one." This statement, in the case of significance, has been supported by Royall [33] and Berger and Sellke [34]. Thus, experiments conducted with fewer statistical units and leading to the same p value should be conceived as less fragile. Contrary to this assumption, *SI* is invariant to sample size when the p value is kept constant. This is illustrated in Table 1. One might argue that this is a limitation of *SI*, which, in other words, satisfies the so-called α-postulate, defined by Cornfield [35], which was described by Royall [36] as the statement that "equal p-values represent equal amounts of evidence [against the null hypothesis], at least approximately." This postulate has been deemed unreasonable [35] or wrong [36].

SI is a diagnostic tool and does not aim to measure the evidence for hypotheses since it is not an inferential procedure, and its role is to support or contradict the p value with an appropriate assessment. *SI* goes beyond the invariance thanks to the introduction of the adaptive threshold, which is a strictly increasing function of the sample size, for penalizing large samples compared with smaller ones, in agreement with Lindley and Scott's perspective [10, 33, 37, 38].

2.5 | Time-To-Event Outcomes

The analysis of strength can be extended to the case of time-to-event outcomes. To this end, as in the previous sections, we consider two groups G_i of sample sizes n_i $(i = 1, 2)$. For each group, we can construct the corresponding survival function $S_i(t)$ via the Kaplan–Meier method. We consider the null hypothesis of equality of the survival functions, namely, $H_0: S_1(t) = S_2(t)$. Let us focus on death as the event of interest. The survival time T represents the time at which death occurs; the censoring time C represents the time at which censoring occurs. At the end of the study period, for each individual we observe either T or C. If the event has occurred before censoring $(T \leq C)$, then we observe the survival time T. If censoring occurs before the event $(T > C)$, then we observe the censoring time. So, we can introduce an indicator variable taking the value 1 if $T \leq C$ (i.e., the event has occurred) or 0 if $T > C$ (i.e., censoring has occurred). To test the hypothesis of equality of the survival curves, we can adopt the log-rank test. Assume that the test led to a significant p value. We define *SFI*, which is calculated via the following algorithm. Focusing on the group with the least number of events, a 0 is transformed into a 1, that is, a censored unit is virtually considered as if it had the event. This change starts from the unit that has the shortest censoring time. Once this change has been made, the p value is recalculated. If this is subverted to nonsignificance, the algorithm stops, and the index is equal to 1. If the p value is still significant, the outcome of another censored unit (the one with the second longest censoring time) is transformed from 0 to 1. If the p value is not significant, the algorithm stops and $SFI = 2$. Otherwise, more outcomes are subverted until nonsignificance occurs. *SFI* can be defined as the minimum number of conversions from censoring to an event that leads to $p > \alpha$. If H_0 is supported by the p value, we again choose the group with the least number of events and, within that group, we select the unit with the longest survival time and event indicator equal to 1. For this unit, we change 1 into 0 and we re-evaluate the p value. If it is significant, the algorithm stops; otherwise, other steps will follow until the p value is subverted. Then, we define *SFI* as the negative of the minimum number of conversions from event to nonevent that causes $p \leq \alpha$.

We remark that *SFI* is a variant of the survival–inferred fragility index (*SIFI*) [39]. The latter has been introduced in the literature as a version of *FI* that is suitable for time-to-event data. The following algorithm is used to calculate *SIFI*. Let us start with the case of significance. The researcher considers the treatment group and, starting from time h (the last observation time before the end of follow-up), a unit is moved to the control group regardless of whether it is labeled with 1 or 0. If the p value becomes nonsignificant, *SIFI* is equal to 1; otherwise, the procedure continues until nonsignificance is reached. In the case of a nonsignificant result, the algorithm is reversed; that is, shifts of units (from 0 to 1 or from 1 to 0) only occur from the control group toward the treatment group until significance is reached [40].

There are two substantial differences between the two procedures. First, *SFI* focuses on the group with the lowest number of events, whereas *SIFI* focuses on the treatment group (in the case of significance). Second, *SFI* only transforms nonevents to events (the reverse in the case of nonsignificance) within the same group, whereas, with *SIFI*, both events and nonevents can be virtually shifted from one group to the other in the same analysis. Thus, the substantial difference is that *SIFI* moves units from one group to the other, whereas *SFI* only converts outcomes

TABLE 1 | For three different fictitious examples, we calculated *SI* with threshold t_n in Equation (13). We also calculated *FI* and *FQ*.

Trial name	G_1 event fraction	G_2 event fraction	p value	*SI*	*FI*	$FQ = FI/n$	$n = n_1 + n_2$	$t_{1/2,1/3}(n)$
A	160/400	314/600	0.0001	0.494	25	0.0025	1000	0.400
B	400/3000	507/3000	0.0001	0.494	51	0.0085	6000	0.445
C	250/4500	340/5500	0.0001	0.494	38	0.0038	10,000	0.454

without moving units. This has been argued as a shortcoming of *SIFI* since it does not respect the nature of the fragility concept [26]. Note, however, that the calculation procedure might be longer for *SFI* than *SIFI* (see Example 4 in Section 6). Finally, neither index has a threshold. We remark that *SFI* can be combined with *SI*. *SFI*, however, does not require the asymptotic approximation; this is essential for *SI*, which is obtained via Equation (8), and its application to the *p* value of the log-rank test (see Example 4).

3 | Analysis of Strength via Likelihood Ratios

3.1 | Likelihood Fragility Index

The SI presented in Section 2 can be applied in the case of normal approximations. Thus, at least medium or large sample sizes are required. Although this covers several applied scenarios, small trials are not uncommon in medical research. In this subsection, we present how to proceed with the analysis of strength via likelihood ratios (LRs). This procedure is the recommended practice for small trials. Adopting the two independent binomial random variables x_i $(i=1,2)$ and hypotheses (1), the likelihood function is given by

$$L(\theta_1,\theta_2)=P_{\theta_1,\theta_2}(X_1=x_1,X_2=x_2)=\prod_{i=1}^{2}P_{\theta_i}(X_i=x_i)$$
$$=\prod_{i=1}^{2}\binom{n_i}{x_i}\theta_i^{x_i}(1-\theta_i)^{n_i-x_i}$$

This attains its maximum at $\widehat{\theta}_i=\frac{x_i}{n_i}(i=1,2)$. Similarly,

$$L(\theta,\theta)=\prod_{i=1}^{2}\binom{n_i}{x_i}\theta^{x_i}(1-\theta)^{n_i-x_i}=\theta^{x_1+x_2}(1-\theta)^{n_1+n_2-x_1-x_2}\prod_{i=1}^{2}\binom{n_i}{x_i}$$

reaches its maximum at $\overline{\theta}=\frac{x_1+x_2}{n_1+n_2}=\sum_{i=1}^{2}p_i\widehat{\theta}_i$, with $p_i=\frac{n_i}{n_1+n_2}$. Based on this premise, we introduce the *LR*

$$LR(x_1,x_2)=\frac{\sup_{\theta_1=\theta_2}L(\theta_1,\theta_2)}{\sup_{\theta_1,\theta_2}L(\theta_1,\theta_2)}$$
$$=p_1^{n_1}p_2^{n_2}\left(1+\frac{x_2}{x_1}\right)^{x_1}\left(1+\frac{x_1}{x_2}\right)^{x_2}\left(1+\frac{n_2-x_2}{n_1-x_1}\right)^{n_1-x_1}$$
$$\left(1+\frac{n_1-x_1}{n_2-x_2}\right)^{n_2-x_2} \tag{15}$$

with $0<LR(x_1,x_2)\le 1$. It is straightforward to show from Equation (15) that the smaller $LR(x_1,x_2)$ is, the less support the experiment provides to the null hypothesis compared with the alternative hypothesis. Thus, we may establish a threshold $\lambda\in(0,1)$ under which the null hypothesis cannot be supported.

If H_0 is not supported, we propose a *likelihood fragility index* (*LFI*), defined as the minimum number of conversions from nonevent to event in the group with the smaller proportion of events that causes the *LR* to attain or exceed the threshold λ. Symmetrically, if H_0 is supported, we define *LFI* as the negative value of the minimum number of conversions from event to nonevent in the group with the smaller proportion of events that causes $LR<\lambda$.

We suggest adapting Royall's proposal [36] for the threshold λ, according to which, if a likelihood ratio is equal to 1, there is no evidence for the alternative hypothesis; if the ratio is less than $1/8=1/2^3$, there is fairly strong evidence in favor of the alternative hypothesis. A ratio of $1/32=1/2^5$ or less implies strong evidence.

On the one hand, focusing on G_1, if $LR(x_1,x_2)<\lambda$, we may add *LFI* events to G_1 until we attain $LR\ge\lambda$. On the other hand, if $LR(x_1,x_2)\ge\lambda$ (i.e., there is support for the null hypothesis), then to subvert the experimental result, a number equal to *LFI* events should be subtracted from the first group until $LR<\lambda$ is attained. Hence *LFI* assumes, in the context of likelihood, the same role as *FI* in the setting of the *p* value.

3.2 | Likelihood Strength Index

The analysis of strength via likelihoods represents a more general approach than *SI*, as the latter only works under the normality assumption. There is, however, a close relationship between the two perspectives. For large sample sizes, Equation (15) can be approximated through

$$LR(x_1,x_2)\cong e^{\frac{-\left(\frac{x_1}{n_1}-\frac{x_2}{n_2}\right)^2}{2\hat{\sigma}^2}} \tag{16}$$

As above, we assume that $\frac{x_1}{n_1}\le\frac{x_2}{n_2}$ without loss of generality. Since $FI\cong n_1Q$,

$$1-\sqrt{\frac{\log LR(x_1+FI,x_2)}{\log LR(x_1,x_2)}}\cong 1-\frac{\frac{x_2}{n_2}-\frac{x_1}{n_1}-Q}{\frac{x_2}{n_2}-\frac{x_1}{n_1}}=\frac{Q}{\frac{x_2}{n_2}-\frac{x_1}{n_1}}=SI \tag{17}$$

$$1-\sqrt{\frac{\log LR(x_1,x_2)}{\log LR(x_1+FI,\ x_2)}}\cong 1-\frac{\frac{x_2}{n_2}-\frac{x_1}{n_1}}{\frac{x_2}{n_2}-\frac{x_1}{n_1}-Q}$$
$$=\frac{-Q}{\frac{x_2}{n_2}-\frac{x_1}{n_1}-Q'}=\frac{Q}{-\hat{\sigma}z_{1-\frac{\alpha}{2}}}=SI \tag{18}$$

for $Q<0$. Thus, similarly to

$$SI\cong\begin{cases}1-\sqrt{\dfrac{\log LR(x_1+FI,x_2)}{\log LR(x_1,x_2)}}, FI\ge 0\\ 1-\sqrt{\dfrac{\log LR(x_1,x_2)}{\log LR(x_1+FI,x_2)}}, FI<0\end{cases}$$

we can define a *LSI* as

$$LSI=\begin{cases}1-\sqrt{\dfrac{\log LR\left(x_1+LFI,x_2\right)}{\log LR\left(x_1,x_2\right)}}, & LFI\geq 0\\ 1-\sqrt{\dfrac{\log LR\left(x_1,x_2\right)}{\log LR\left(x_1+LFI,x_2\right)}}, & LFI<0\end{cases} \tag{19}$$

LSI constitutes a normalized index of strength, valid for small sample sizes and small clinical trials, generalizing *SI* in the sense that, if $LFI\cong FI$ in (19), we obtain $LSI\cong SI$ through Equations (17) and (18).

To make the index operative, one may use either the fixed threshold (9) or the adaptive threshold (11). *LSI* has properties similar to those already presented for *SI*. Like *SI*, it is easy to implement, does not depend on the intentions of the researcher, and leads different researchers to reach the same conclusions under the same conditions.

4 | One-Sided Hypotheses

The indices *SI* and *LSI* can be extended to the case of a one-sided alternative hypothesis, namely,

$$\mathrm{H}_0:\theta_1=\theta_2 \text{ vs. } \mathrm{H}_1:\theta_1<\theta_2 \tag{20}$$

or

$$\mathrm{H}_0:\theta_1\geq\theta_2 \text{ vs. } \mathrm{H}_1:\theta_1<\theta_2 \tag{21}$$

For *SI*, from the equation

$$\Phi\left(\frac{\frac{x_1}{n_1}-\frac{x_2}{n_2}+Q}{\widehat{\sigma}}\right)=\alpha$$

for $\frac{x_1}{n_1}\leq\frac{x_2}{n_2}$, it can be shown that

$$Q=\frac{x_2}{n_2}-\frac{x_1}{n_1}-\widehat{\sigma}z_{1-\alpha}=\left(z_{1-p}-z_{1-\alpha}\right)\widehat{\sigma}$$

from which it follows that

$$SI=\begin{cases}\dfrac{Q}{\frac{x_2}{n_2}-\frac{x_1}{n_1}}=1-\dfrac{\widehat{\sigma}z_{1-\alpha}}{\frac{x_2}{n_2}-\frac{x_1}{n_1}}=1-\dfrac{z_{1-\alpha}}{z_{1-p}}, & p\leq\alpha\\ \dfrac{Q}{-\widehat{\sigma}z_{1-\alpha}}=\dfrac{\frac{x_2}{n_2}-\frac{x_1}{n_1}}{-\widehat{\sigma}z_{1-\alpha}}+1=1-\dfrac{z_{1-p}}{z_{1-\alpha}}, & \alpha<p<0.5\\ 1, & p\geq 0.5\end{cases} \tag{22}$$

We can use the same threshold already introduced for *SI*. From Equation (22), it can be seen that if $p\geq 0.5$, then the obtained experimental result provides maximum strength to the null hypothesis and, in general, we set $SI=1$ (see Supporting Information S1: Figure 2). For instance, this occurs if we test $\mathrm{H}_0:\theta_1=\theta_2$ or $\mathrm{H}_0:\theta_1\leq\theta_2$ versus $\mathrm{H}_1:\theta_1>\theta_2$ when $\frac{x_1}{n_1}\leq\frac{x_2}{n_2}$.

For either hypothesis (20) or (21), when $\widehat{\theta}_1\leq\widehat{\theta}_2$, the likelihood ratio

$$LR\left(x_1,x_2\right)=\frac{\sup\limits_{H_0}L\left(\theta_1,\theta_2\right)}{\sup\limits_{H_0\cup H_1}L\left(\theta_1,\theta_2\right)}=\frac{L\left(\overline{\theta},\overline{\theta}\right)}{L\left(\widehat{\theta}_1,\widehat{\theta}_2\right)}$$

coincides with Equation (15) and can be approximated with Equation (16) for large sample sizes (for details, see Appendix A).

In a similar way to

$$SI=\begin{cases}\dfrac{Q}{\frac{x_2}{n_2}-\frac{x_1}{n_1}}=1-\dfrac{\frac{x_2}{n_2}-\frac{x_1}{n_1}-Q}{\frac{x_2}{n_2}-\frac{x_1}{n_1}}\cong 1-\sqrt{\dfrac{\log LR\left(x_1+FI,x_2\right)}{\log LR\left(x_1,x_2\right)}}, & FI\geq 0\\ \dfrac{Q}{-\widehat{\sigma}z_{1-\alpha}}=1-\dfrac{\frac{x_2}{n_2}-\frac{x_1}{n_1}}{\frac{x_2}{n_2}-\frac{x_1}{n_1}-Q}\cong 1-\sqrt{\dfrac{\log LR\left(x_1,x_2\right)}{\log LR\left(x_1+FI,x_2\right)}}, & FI<0\end{cases}$$

we provide the normalized index (19) with the same thresholds already proposed. In addition, for $\mathrm{H}_0:\theta_1\leq\theta_2$ versus $\mathrm{H}_1:\theta_1>\theta_2$

$$LR\left(x_1,x_2\right)=\frac{\sup\limits_{\theta_1\leq\theta_2}L\left(\theta_1,\theta_2\right)}{\sup\limits_{\theta_1,\theta_2}L\left(\theta_1,\theta_2\right)}=\frac{L\left(\widehat{\theta}_1,\widehat{\theta}_2\right)}{L\left(\widehat{\theta}_1,\widehat{\theta}_2\right)}=1$$

follows from $\widehat{\theta}_1\leq\widehat{\theta}_2$, and hence we have $LSI=1$.

5 | A Simulation Study

To investigate the empirical behavior of *SI* and *LSI*, we performed a Monte Carlo study. Initially, we examined a simulated scenario in which the null hypothesis is supported. Within this context, we sampled x_1 and x_2 from binomial random variables, both characterized by a probability parameter $\theta_1=\theta_2=0.75$. In a putative clinical context, this outcome would suggest that the two treatments can be deemed interchangeable. We then extended the study by setting the equal probability parameters to 0.50, 0.25, or 0.10. As for the sample size, we set $n_1=n_2=500$, then 300, and 200. We also considered the case of unbalanced sample sizes. Under these conditions, we performed $N=10{,}000$ Monte Carlo simulations. For each pair $\left(x_1,x_2\right)$, we calculated the log odds ratio, the asymptotic confidence intervals (CIs) for the log odds, and the *p* value. We calculated the log odds and CIs for the log odds, instead of CIs for the difference of proportions, given that the latter has a true coverage probability less than nominal for proportions close to 0 or 1 [41]. The logarithmic transformation of the odds guarantees a faster convergence rate [42]. Since the simulation was carried out assuming the null hypothesis of equal proportions, we anticipated not rejecting H_0, meaning that zero would be within the CI for the log odds. However, due to random variation, there were instances in

TABLE 2 | Main results of the simulation study. Simulations were conducted under the null hypothesis $\theta_1 = \theta_2$. We adopted the adaptive threshold: $t_n = t_{1/2,1/3}(n) = \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$.

θ_1	θ_2	n_1	n_2	$0 \in CI$	$SI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$	$LSI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$	$0 \notin CI$	$SI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$	$LSI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$
0.75	0.75	500	500	9520	7621 (80.05)	7930 (83.29)	480	9 (1.87)	8 (1.66)
		300	300	9509	7769 (81.70)	8110 (85.28)	491	7 (1.42)	9 (1.84)
		200	200	9498	7796 (82.08)	8111 (85.39)	502	21 (4.18)	24 (4.78)
		500	300	9539	7651 (80.21)	7872 (82.52)	461	8 (1.73)	10 (2.17)
		500	200	9492	7692 (81.03)	7942 (83.67)	508	8 (1.57)	14 (2.75)
0.50	0.50	500	500	9518	7570 (79.01)	7822 (81.64)	482	13 (2.69)	8 (1.66)
		300	300	9545	7651 (80.15)	8011 (83.92)	455	14 (3.07)	14 (3.07)
		200	200	9502	7907 (83.21)	8259 (86.92)	498	24 (4.82)	16 (3.21)
		500	300	9490	7656 (80.67)	7938 (83.64)	510	8 (1.57)	7 (1.37)
		500	200	9560	7781 (81.39)	8055 (84.25)	440	8 (1.82)	9 (2.04)
0.25	0.25	500	500	9520	7621 (80.05)	7960 (83.61)	480	9 (1.87)	8 (1.66)
		300	300	9509	7769 (81.70)	7991 (84.03)	491	7 (1.42)	9 (1.83)
		200	200	9498	7796 (82.08)	8137 (85.67)	502	21 (4.18)	26 (5.18)
		500	300	9539	7651 (80.21)	7896 (82.77)	461	8 (1.73)	10 (2.17)
		500	200	9492	7692 (81.03)	7937 (83.62)	508	8 (1.57)	14 (2.75)
0.10	0.10	500	500	9501	7617 (80.17)	7890 (83.04)	499	14 (2.80)	14 (2.80)
		300	300	9553	7781 (81.45)	8139 (85.19)	447	9 (2.01)	11 (2.46)
		200	200	9564	7919 (82.80)	8419 (88.02)	436	9 (2.06)	19 (4.36)
		500	300	9508	7645 (80.41)	7985 (83.98)	492	7 (1.42)	20 (4.06)
		500	200	9525	7648 (80.29)	7956 (83.53)	475	7 (1.47)	28 (5.89)

Note: $0 \in CI$: CIs including 0 (reference value for the log(odds) ratio), the absolute number of cases are reported in the cell (percentage is given within parentheses). $0 \notin CI$: CIs not including 0 (reference value for the log(odds) ratio), absolute number (percentage). $SI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$: Strength index $> \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$, absolute number (percentage); $LSI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$: Likelihood strength index $> \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$, absolute number (percentage).

which the null hypothesis was erroneously rejected. In these cases, as well as those in which the null hypothesis was correctly upheld, we evaluated the performance of *SI*. We determined the proportion of cases in which *SI* added support to the experimental outcome while applying the adaptive threshold (13). We repeated the analysis for the adaptive threshold (12). Herein, we report the results for the threshold (13); other findings are presented in the Supporting Information, where additional scenarios with different parameters and sample sizes are also considered. We then extended the study to *LSI*. The results are shown in Table 2.

We observed that the obtained results are rather similar across different sample sizes. When the null hypothesis was (wrongly) rejected (with $\theta_1 = \theta_2 = 0.75$), *SI* provides strength to the experimental results in only 4.18% of the cases for a sample size of 200 and 1.42% for a sample size of 300. For the simulations with samples of size 500, we obtained strength in 1.87% of the cases. This implies that the number of false-positive changes from 480 when using the *p* value to 9 when adopting the strength index. Moreover, in about 80% of the cases, taking into consideration *SI* would lead to confirming the correct non-rejection of the null hypothesis. Similar considerations apply when varying the value of $\theta_1 = \theta_2$, and concerning *LSI*. We remark that *LSI* shows a behavior with that of *SI*.

We repeated the same simulation structure but setting $\theta_1 \neq \theta_2$. The results are shown in Table 3. For instance, we set $H_1: \theta_1 = 0.70$ and $\theta_2 = 0.60$. For equal sample sizes of 500, *SI* provides support to the *p* value in 56.07% of the cases, and *LSI* in 53.89% of the cases. Furthermore, in 81.94% and 78.13% of the cases, respectively, *SI* and *LSI* prevent a fallacy, namely, they do not confirm the experimental result that wrongly indicated non-rejection of the null hypothesis. We repeated all the previous simulations setting different values for the parameters θ_1 and θ_2 and, overall, we obtained quite similar results (see the Supporting Information).

By careful inspection of Tables 2 and 3, one may object that both *SI* and *LSI* could be rather selective in confirming the experimental result. In principle, this can be mitigated by changing the adaptive threshold or, in the case of *LSI*, by modifying the value of the parameter λ used for the likelihood ratio, for which only rules of thumb have been proposed in the literature. In the

TABLE 3 | Main results of the simulation study. Simulations were conducted under the alternative hypothesis $\theta_1 \neq \theta_2$. We adopted the adaptive threshold: $t_n = t_{1/2,1/3}(n) = \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$.

θ_1	θ_2	n_1	n_2	$0 \in CI$	$SI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$	$LSI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$	$0 \notin CI$	$SI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$	$LSI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$
0.70	0.60	500	500	814	147 (18.06)	178 (21.87)	9186	5151 (56.07)	4951 (53.89)
		300	300	2741	919 (33.53)	1013 (36.96)	7259	2681 (36.93)	2630 (36.23)
		200	200	4363	1890 (43.32)	2097 (48.06)	5637	1602 (28.42)	1599 (28.37)
		500	300	1186	458 (38.62)	570 (48.06)	8184	3679 (44.95)	3399 (41.53)
		500	200	2830	910 (32.15)	1154 (40.78)	7170	2546 (35.51)	2211 (30.84)
0.50	0.40	500	500	1064	222 (20.86)	257 (24.15)	8936	4616 (51.65)	4294 (48.05)
		300	300	3009	995 (33.07)	1134 (37.68)	6991	2410 (34.47)	2410 (34.47)
		200	200	4811	2215 (46.04)	2441 (50.74)	5189	1372 (26.44)	1172 (22.58)
		500	300	2109	579 (27.45)	719 (34.09)	7891	3125 (39.60)	3136 (39.74)
		500	200	3321	1173 (35.32)	1360 (40.95)	6679	1966 (29.43)	2151 (33.20)
0.30	0.20	500	500	434	66 (15.21)	80 (18.43)	9566	6452 (67.44)	6307 (65.93)
		300	300	1865	478 (25.63)	562 (30.13)	8135	3625 (44.56)	3664 (41.35)
		200	200	3624	1401 (38.65)	1613 (44.51)	6376	2064 (32.37)	2177 (34.14)
		500	300	1112	216 (19.42)	252 (22.66)	8888	4530 (50.96)	5068 (57.02)
		500	200	2147	567 (26.41)	637 (29.67)	7853	2837 (36.13)	3783 (48.17)
0.20	0.10	500	500	43	2 (4.65)	4 (9.30)	9957	8833 (88.71)	8825 (88.63)
		300	300	696	119 (17.09)	149 (21.41)	9304	5962 (64.08)	6116 (65.73)
		200	200	2018	571 (28.29)	678 (33.59)	7982	3618 (45.32)	4085 (51.17)
		500	300	264	36 (13.64)	40 (15.15)	9736	6879 (70.66)	7683 (78.91)
		500	200	826	134 (16.22)	151 (18.28)	9174	4744 (51.71)	6535 (71.23)

Note: $0 \in CI$: CIs including 0 (reference value for the log(odds) ratio), the absolute number of cases are reported in the cell (percentage is given within parentheses). $0 \notin CI$: CIs not including 0 (reference value for the log(odds) ratio), absolute number (percentage). $SI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$: Strength index $> \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$, absolute number (percentage); $LSI > \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$: Likelihood strength index $> \frac{1}{2} - \frac{1}{\sqrt[3]{n}}$, absolute number (percentage).

TABLE 4 | Summary of the main results emerged in the illustrative examples. *SI* and *LSI* have been calculated through Formula (8) and (19), respectively.

Study	n_1	n_2	*p* value	*Q*	*SI*	*FI*	*LFI*	*LSI*
Sterne et al. [43]	678	1025	0.0002	0.0407	0.467	28	26	0.448
Cao et al. [44]	99	100	0.32	−0.0568	0.495	−6	−6	0.541
RECOVERY [45]	3155	1561	0.16	−0.0074	0.277	−24	−27	0.309

Supporting Information, we show the results of the simulations conducted with the fixed threshold $1 - \frac{1}{\sqrt{2}}$ in Equation (9) or with the adaptive threshold (12).

It should be stressed that the analysis of strength simply offers possible confirmation of what is expressed by the *p* value, going beyond the main limitations of the fragility index. To adopt this procedure, an R package named `strength.analysis` has been uploaded to Github.[1] This provides researchers with the possibility of directly applying the analysis of strength to their data analysis. The reader will find the main functions introduced in this article already implemented in the package (see the Supporting Information for details).

6 | Illustrative Examples

In this section, we report three empirical applications of the analysis of strength discussed in the previous sections. These examples[2] address experimental treatments for COVID-19 during the acute phase of the pandemic in 2020. Example 1 concerns the administration of dexamethasone (an

anti-inflammatory drug), with both *SI* and *LSI* supporting the experimental treatment with the threshold (13). Example 2 concerns the administration of two antiviral drugs, lopinavir and ritonavir, with the null hypothesis supported by both *SI* and *LSI* with the threshold (13). Example 3 refers to the effect of administering hydroxychloroquine. The empirical result does not support the effect of the drug and is not confirmed by *SI* or *LSI* with the threshold (13). With the threshold (9), the null hypothesis is supported by *LSI*. For all these illustrative examples, we set a significance threshold $\alpha = 0.05$. As to the level λ, we set a value $\lambda = 1/8$. We also describe a fourth example in which *SFI* and *SIFI* have been calculated on simulated datasets.

Example 1. In 2020, Sterne et al. [43] (the WHO Rapid Evidence Appraisal for COVID-19 Therapies Research Group) published the results of a multicenter RCT conducted with 1703 subjects affected by COVID-19 and hospitalized. Of these, 678 were randomized to receive low-dose dexamethasone (the treatment group) and 1025 were randomized to receive the standard care (the control group). The primary outcome was mortality after 28 days; 222 patients died in the treatment group G_1 and 425 in the control group G_2. Referring to hypotheses (1), using the statistic $\widehat{\theta} = \frac{x_2}{n_2} - \frac{x_1}{n_1}$ and the estimate (2) of its variance, it was found that $\widehat{\theta} = 0.088$ and $\widehat{\sigma} = 0.0237$, so that $z_{1-\frac{p}{2}} = \frac{\widehat{\theta}}{\widehat{\sigma}} = 3.68$, which, being greater than $z_{1-0.025} = 1.96$, ensures $p = 0.0002$ and therefore significance. We first calculated the quantity Q; from Equation (5), we obtained $Q = 0.040$ and, by using Equation (8), it followed that $SI = 0.467$. By adopting the threshold (13), it followed that $t_{1,703} = 0.416$, and the experimental result was confirmed.

The sample sizes and the numbers of deaths in the two groups are large enough to justify the normal approximation. Thus, from $FI \cong n_1 Q$ we obtained $FI = 28$, which can be interpreted as the minimum number of deaths that should have occurred in addition to those observed in the treated group to subvert significance to nonsignificance.

We also applied the likelihood-based approach. Using Equation (11), we obtained $LR(x_1, x_2) = 0.00115$. As expected, considering the λ threshold stated above, such a low likelihood ratio would not support the null hypothesis. The minimum number of events to be added to x_1 to subvert the result is given by $LFI = 26$. Indeed, in this case, the likelihood ratio became $LR = 0.127 > \frac{1}{8} = 0.125$. We also calculated $LSI = 0.448$, which confirmed the experimental result by setting the threshold $t_{1,703} = 0.416$. The main calculations are summarized in Table 4.

Example 2. Cao et al. [44] published the results of an RCT on patients admitted to Jin Yin-Tan Hospital in Wuhan for COVID-19. Of the 199 patients observed, 99 (G_1) underwent treatment with lopinavir and 100 (G_2) underwent standard care. Considering mortality as an outcome, after 28 days, 19 deaths were reported in the treatment group and 25 in the control group. Using the normal approximation, we obtained $\widehat{\theta} = 0.058$ and $\widehat{\sigma} = 0.0587$, with $z_{1-\frac{p}{2}} = 0.99$. Given that $p = 0.32$, the experimental result was not significant. To assess the strength of the result, we first calculated the quantity Q. From Equation (5), we obtained $Q = -0.0568$. Using Equation (8), we calculated $SI = 0.495$, which, being greater than the threshold (13) given the value $t_{199} = 0.328$, provided strength to the experimental result. We also obtained $FI \cong -0.0568 \times 99 = -6$. This represents the number of deaths that would have to be subtracted from the 19 deaths to subvert the result from nonsignificance to significance. This can be obtained through a simple Fisher's test for the 2×2 table. The negative sign of *FI* is a convention to show the direction of the effect from nonsignificance to significance [46]. The analysis of strength can also be conducted through a likelihood-based approach. Using Equation (16), we obtained a value $LR(x_1, x_2) = 0.6135$. This was above the threshold $\lambda = 1/8$, thus supporting the null hypothesis. To subvert the experimental result, we need to subtract 6 events from the 19 events obtained in the treatment arm. It followed that $LR = 0.097$, $LFI = -6$, and $LSI = 0.541$, which strengthened the experimental result, being greater than $t_{199} = 0.328$.

Example 3. In June 2020, the results of the Randomized Evaluation of COVID-19 Therapy (RECOVERY) [45] study, relating to 4716 patients affected by COVID-19 and hospitalized, were made public. Of these patients, 1561 were treated with hydroxychloroquine and 3155 underwent standard care. The number of deaths after 28 days was 421 in the treated group (G_2) and 790 in the control group (G_1). The experimental results led to $\widehat{\theta} = 0.019$, $\widehat{\sigma} = 0.014$, $z_{1-p/2} = 1.42$, and $p = 0.16$, supporting the nonsignificance of the experimental result. However, having found $Q = -0.0074$, using Equation (8) we obtained $SI = 0.277$, which is below the threshold $t_{4,716} = 0.440$ and also below the fixed threshold (9) $1 - \frac{1}{\sqrt{2}} \approx 0.293$. The experimental result was not confirmed by the analysis of strength. It followed that $FI = -24$; that is, we would need to subtract 24 events from the 790 deaths to subvert the result from nonsignificant to significant. Adopting the likelihood approach, we obtained $LR(x_1, x_2) = 0.367$, which supports the null hypothesis. We calculated $LFI = -27$, implying that to subvert the experimental result, 27 events would need to have been discarded from the control group; this would have led to $LR = 0.121$. Finally, we calculated $LSI = 0.309$, which was not larger than the adaptive threshold $t_{4,716} = 0.440$ but was larger than the fixed threshold 0.293. To summarize, we found disagreement between the different statistical procedures used to assess the results of this experiment. This would support the idea that the question of whether hydroxychloroquine should be administered to patients with COVID-19 is open.

Example 4. A survival analysis was conducted for a simulated dataset containing 2000 observations (1000 each for the treatment and control groups). The two groups were created so that they would have small differences in median survival times (10 and 12 months, respectively), which were generated from a Weibull distribution. This led to a p value of the log-rank test that was not statistically significant ($p = 0.399$). This scenario represents a situation in which there is no difference in survival between the two groups. Since the initial p value was not significant, the *SFI* algorithm attempts to convert events (deaths) to censored observations (nonevents) in the group with fewer events, starting with those with the longest follow-up times. The goal is to assess how many such conversions are needed to make the p value significant. The number of conversions

required is reported as the *SFI* value. For *SIFI*, the algorithm moves subjects from the control group to the treatment group, beginning with those closest to the end of follow-up. The process continues until the p value is subverted, and the number of required iterations is reported as the *SIFI* value. In these conditions, we obtained $SFI = -56$ and $SIFI = 33$.

A second simulated setting had a more pronounced difference in median survival times between the two groups (10 vs. 16 months), resulting in an initial p value that was statistically significant ($p = 0.00012$). In this case, the *SFI* algorithm converts censored observations to events in the group with fewer events, starting with those with the shortest follow-up times. The process is repeated until the p value is no longer significant. The number of conversions required is recorded as the *SFI* value. For *SIFI*, the algorithm moves units from the treatment group to the control group, beginning with those with the longest follow-up times. This continues until the p value is no longer significant, and the number of iterations is reported as the *SIFI* value. In this case, we obtained $SFI = 63$ and $SIFI = 35$.

Finally, we repeated the simulation, setting the median survival times in the two groups to 10 and 18 months ($p = 0.00000052$). In this case, we obtained $SFI = 104$ and $SIFI = 57$. Note that, by invoking the normal asymptotic null distribution of the log-rank test statistic, in the three simulated settings, it is possible to calculate the strength index

$$SI = \begin{cases} 1 - \dfrac{z_{1-\frac{\alpha}{2}}}{z_{1-\frac{p}{2}}}, & p \le \alpha \\ 1 - \dfrac{z_{1-\frac{p}{2}}}{z_{1-\frac{\alpha}{2}}}, & p > \alpha \end{cases}$$

which was found to be 0.56, 0.49, and 0.61, respectively. The R code to reproduce these examples is available from the authors upon request.

7 | Conclusions

This article has presented a new approach, the analysis of strength, aimed at revising the traditional fragility analysis of RCTs from a new perspective. This has been done in the case of normal random variables by adopting *SI*, which can be compared with ad hoc thresholds. *SI* allows practitioners to establish a clear conclusion regarding the strength of an experimental result. Like the fragility index, *SI* has a predictable relationship with the p value and, therefore, one may wonder if *SI* might shed light on the suspicion [47] that the fragility index could be a "p value in sheep's clothing." [48] The answer for *SI* seems more optimistic than that for *FI* since *SI* thresholds can discern whether a "small" or "large" p value can validate the null hypothesis. This is not possible by simply adopting the fragility index. It should be noted that another substantial difference between *SI* and the p value is that the former is based on a counterfactual approach ("what would have happened if," or what-if), whereas the latter is based on a purely probabilistic approach. Thus, we may draw for *SI* the same conclusion drawn by Lin et al. [26] for *FI*: "this critique [that it may be strongly associated with the p-value and the sample size] is mostly from an undue conception that *FI* aims to replace the *P* value, although it does not. Alternatively, *FI* only serves as a supplemental measure to aid in the assessment of the robustness of statistical significance." In addition to binomial outcomes, we considered the case of time-to-event outcomes, and we proposed a new fragility index *SFI*, which may represent a possible substitute to the already adopted *SIFI*, as it may align more closely than *SIFI* with the original concept of fragility. In addition, *SFI* can be combined with *SI* (adopting the asymptotic approximation), thus allowing the p value of the log-rank test to be evaluated.

The analysis of strength has been also extended, more generally, through a likelihood approach [36]. This constitutes an alternative to the p value. Through the likelihood ratio, we showed how to quantify the number of events that would subvert the experimental result. This has been called *LFI*, and it has the same role as *FI* but in the context of the analysis of strength. The index *LSI* has the same function as *SI*, but the p value is substituted with a likelihood ratio and the threshold α is substituted with λ. *LSI* may or may not confirm what the likelihood ratio has shown regarding the null hypothesis, and it has the advantage of being useful for small sample sizes. In general, for a large sample size, $LSI \approx SI$.

FI and *LFI* can lead to qualitatively different conclusions, as can *SI* and *LSI*, being based on different metrics, as observed in the case of the drug hydroxychloroquine (Example 3). In Section 2, we recalled Lindley and Scott's claim and showed that *SI* is not in agreement with it. This was illustrated with an empirical example from which it emerges that *SI* is independent of sample size. This could seem an undesirable property of *SI*, but what matters is the presence of the adaptive threshold t_n, which is "small" for "small" values of the sample size. We proposed two instances of the parametric family (11), and two other thresholds are presented in the Supporting Information. The choice of a threshold cannot be given in absolute terms, as it depends on the operational context and should be selected to protect against false positives and false negatives as required. For example, in a context that stresses the need to rigorously protect against false positives, the choice of the threshold $t_n = \frac{1}{2} - \frac{1}{\log_e(n)}$ would be preferable to that of $t_n = \frac{1}{2} - \frac{1}{\log_{10}(n)}$, as can be seen from Supporting Information S1: Tables S5 and S7.

To summarize, before selecting a metric to evaluate the reliability of results, a researcher must first decide whether to use a p value or not. This decision is crucial and has been the subject of extensive debate in the recent literature. If the researcher chooses to use a p value, then diagnostics like *FI* and *SI* are appropriate. However, if a researcher opts against using a p value, or if the asymptotic approximation is not suitable, a likelihood-based approach is more appropriate. This is particularly relevant in small trials, such as those in rare disease research, in which sample sizes are often below 30. In such cases, it is more suitable to use metrics like *LFI* or *LSI*.

In conclusion, this study offers a practical solution to the problem of assessing p values and likelihood ratios in the context of RCTs

and, more generally, in comparisons of sample proportions. The proposed indices were corroborated by a simulation study. Such indices have the potential to constitute an objective, sensitive, and reliable marker to assess the strength of experimental results in comparing two proportions. They are easy to implement, and ready-to-use procedures [31] provide guidance for applied researchers in drawing statistical inferences. In future work, both *SI* and *LSI* may be appropriately reframed and applied to the case of adaptive trials, such as those with group sequential designs or designs with adaptive modification to the sample size.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

There is no original data associated with this submission. Secondary data can be retrieved from the cited sources. Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Endnotes

[1] See https://github.com/Binomial123/strength.analysis.git. The package will also be available on the CRAN website.

[2] The description of the studies reported in this section includes the main analyses without details about stratification variables such as gender, age, or variables linked to concomitant pathologies, or complementary ongoing treatments for Covid-19 such as the use of oxygen or invasive ventilation.

References

1. A. R. Feinstein, "The Unit Fragility Index: An Additional Appraisal of "Statistical Significance" for a Contrast of Two Proportions," *Journal of Clinical Epidemiology* 43, no. 2 (1990): 201–209.

2. M. Walsh, S. K. Srinathan, D. F. McAuley, et al., "The Statistical Significance of Randomized Controlled Trial Results Is Frequently Fragile: A Case for a Fragility Index," *Journal of Clinical Epidemiology* 67, no. 6 (2014): 622–628.

3. K. F. Docherty, R. T. Campbell, P. S. Jhund, M. C. Petrie, and J. J. V. McMurray, "How Robust Are Clinical Trials in Heart Failure?," *European Heart Journal* 38, no. 5 (2017): 338–345.

4. B. R. Baer, M. Gaudino, M. Charlson, S. E. Fremes, and M. T. Wells, "Fragility Indices for Only Sufficiently Likely Modifications," *Proceedings of the National Academy of Sciences of the United States of America* 118, no. 49 (2021): e2105254118.

5. M. S. Khan, G. C. Fonarow, T. Friede, et al., "Application of the Reverse Fragility Index to Statistically Nonsignificant Randomized Clinical Trial Results," *JAMA Network Open* 3, no. 8 (2020): e2012469.

6. B. R. Baer, M. Gaudino, S. E. Fremes, M. Charlson, and M. T. Wells, "The Fragility Index Can Be Used for Sample Size Calculations in Clinical Trials," *Journal of Clinical Epidemiology* 139 (2021): 199–209.

7. T. M. Condon, R. W. Sexton, A. J. Wells, and M. S. To, "The Weakness of Fragility Index Exposed in an Analysis of the Traumatic Brain Injury Management Guidelines: A Meta-Epidemiological and Simulation Study," *PLoS One* 15, no. 8 (2020): e0237879.

8. J. D. Niforatos, A. R. Zheutlin, A. Chaitoff, and R. M. Pescatore, "The Fragility Index of Practice Changing Clinical Trials Is Low and Highly Correlated With *P*-Values," *Journal of Clinical Epidemiology* 119 (2020): 140–142.

9. A. Schröder, O. J. Muensterer, and C. O. von Sochaczewski, "The Fragility Index May Not Be Ideal for Paediatric Surgical Conditions: The Example of Foetal Endoscopic Tracheal Occlusion," *Pediatric Surgery International* 37 (2021): 1–3.

10. G. E. Potter, "Dismantling the Fragility Index: A Demonstration of Statistical Reasoning," *Statistics in Medicine* 39 (2020): 3720–3731.

11. M. H. Murad, A. K. Balla, M. S. Khan, A. Shaikh, S. Saadi, and Z. Wang, "Thresholds for Interpreting the Fragility Index Derived From Sample of Randomised Controlled Trials in Cardiology: A Meta-Epidemiologic Study," *BMJ Evidence-Based Medicine* 28 (2022): 133–136.

12. J. C. Del Paggio and I. F. Tannock, "The Fragility of Phase 3 Trials Supporting FDA-Approved Anticancer Medicines: A Retrospective Analysis," *Lancet Oncology* 20, no. 8 (2019): 1065–1069.

13. C. J. Tignanelli and L. M. Napolitano, "The Fragility Index in Randomized Clinical Trials as a Means of Optimizing Patient Care," *JAMA Surgery* 154, no. 1 (2019): 74–79.

14. P. Budhiraja, B. Kaplan, M. Kalot, et al., "Current State of Evidence on Kidney Transplantation: How Fragile Are the Results?," *Transplantation* 106, no. 2 (2022): 248–256.

15. L. Lin, "Factors That Impact Fragility Index and Their Visualizations," *Journal of Evaluation in Clinical Practice* 27, no. 2 (2021): 356–364.

16. Y. Battaglia, A. Mantovani, R. Shroff, et al., "Online Haemodiafiltration and All-Cause Mortality: How Fragile Are the Results of the Studies Published So Far?," *Nephrology, Dialysis, Transplantation* 39, no. 6 (2024): 1034–1036, https://doi.org/10.1093/ndt/gfae003.

17. N. V. Suresh, B. C. Go, C. G. Fritz, et al., "The Fragility Index: How Robust Are the Outcomes of Head and Neck Cancer Randomised, Controlled Trials?," *Journal of Laryngology and Otology* 138, no. 4 (2024): 451–456, https://doi.org/10.1017/S0022215123001755.

18. M. A. Zabat, A. M. Giakas, A. L. Hohmann, and J. H. Lonner, "Interpreting the Current Literature on Outcomes of Robotic-Assisted Versus Conventional Total Knee Arthroplasty Using Fragility Analysis: A Systematic Review and Cross-Sectional Study of Randomized Controlled Trials," *Journal of Arthroplasty* 39, no. 7 (2024): 1882–1887, https://doi.org/10.1016/j.arth.2024.01.044.

19. A. Wang, D. Kwon, E. Kim, O. Oleru, N. Seyidova, and P. J. Taub, "Statistical Fragility of Outcomes in Acellular Dermal Matrix Literature: A Systematic Review of Randomized Controlled Trials," *Journal of Plastic, Reconstructive and Aesthetic Surgery* 91 (2024): 91–292, https://doi.org/10.1016/j.bjps.2024.02.047.

20. R. Wasserstein and N. Lazar, "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *American Statistician* 70 (2016): 129–133.

21. R. Wasserstein, A. Schirm, and N. Lazar, "Moving to a World Beyond '$p<0.05$', supplement," *American Statistician* 73, no. S1 (2019): 1–19.

22. H. M. J. Hung, J. Lawrence, and S. J. Wang, "*p*-Value, Hypothesis Testing, Strength of Evidence: Comment on 'The Role of *p*-Values in Judging the Strength of Evidence and Realistic Replication Expectations'," *Statistics in Biopharmaceutical Research* 13, no. 1 (2021): 30–31.

23. S. Chakraborty, A. Liu, R. Ball, and M. Markatou, "On the Use of the Likelihood Ratio Test Methodology in Pharmacovigilance," *Statistics in Medicine* 41, no. 27 (2022): 5395–5420.

24. Y. Ding, M. Markatou, and R. Ball, "An Evaluation of Statistical Approaches to Postmarketing Surveillance," *Statistics in Medicine* 39, no. 7 (2020): 845–874.

25. S. D. Walter, L. Thabane, and M. Briel, "The Fragility of Trial Results Involves More Than Statistical Significance Alone," *Journal of Clinical Epidemiology* 124 (2020): 34–41.

26. L. Lin, A. Xing, H. Chu, et al., "Assessing the Robustness of Results From Clinical Trials and Meta-Analyses With the Fragility Index," *American Journal of Obstetrics and Gynecology* 228, no. 3 (2023): 276–282.

27. A. Xing and L. Lin, "Empirical Assessment of Fragility Index Based on a Large Database of Clinical Studies in the Cochrane Library," *Journal of Evaluation in Clinical Practice* 29, no. 2 (2023): 359–370.

28. R. A. Matthews, "Moving Towards the Post $p < 0.05$ Era via the Analysis of Credibility, supplement," *American Statistician* 73, no. S1 (2019): 202–212.

29. L. Held, "The Assessment of Intrinsic Credibility and a New Argument for $P < 0.005$," *Royal Society Open Science* 6, no. 3 (2019): 18534.

30. B. R. Baer, M. Gaudino, S. E. Fremes, M. Charlson, and M. T. Wells, "Reassembling the Fragility Index: A Demonstration of Statistical Reasoning," *Journal of Clinical Epidemiology* 142 (2022): 317–318.

31. A. Ly, A. Stefan, J. van Doorn, et al., "The Bayesian Methodology of sir Harold Jeffreys as a Practical Alternative to the *p* Value Hypothesis Test," *Computational Brain & Behavior* 3, no. 2 (2020): 153–161.

32. E. Wagenmakers, "A Practical Solution to the Pervasive Problems of *p*-Values," *Psychonomic Bulletin & Review* 14, no. 5 (2007): 779–804.

33. R. Royall, "The Effect of Sample Size on the Meaning of Significance Tests," *American Statistician* 40, no. 4 (1986): 313–315.

34. J. O. Berger and T. Sellke, "Testing a Point Null Hypothesis: The Irreconcilability of *p* Values and Evidence," *Journal of the American Statistical Association* 82, no. 397 (1987): 112–122.

35. J. Cornfield, "Sequential Trials, Sequential Analysis and the Likelihood Principle," *American Statistician* 20, no. 2 (1966): 18–23.

36. R. Royall, *Statistical Evidence: A Likelihood Paradigm* (Boca Raton, FL: Chapman & Hall/CRC, 2000).

37. D. V. Lindley and W. F. Scott, *New Cambridge Statistical Tables* (Cambridge, UK: Cambridge University Press, 1984).

38. E. J. Wagenmakers and A. Ly, "History and Nature of the Jeffreys–Lindley Paradox," *Archive for History of Exact Sciences* 77, no. 1 (2023): 25–72.

39. D. Bomze, N. Asher, O. Hasan Ali, et al., "Survival-Inferred Fragility Index of Phase 3 Clinical Trials Evaluating Immune Checkpoint Inhibitors," *JAMA Network Open* 3, no. 10 (2020): e2017675, https://doi.org/10.1001/jamanetworkopen.2020.17675.

40. N. Horesh, D. Bomze, C. Lim, G. Markel, T. Meirson, and D. Azoulay, "Systemic Review of the Robustness of Randomized Controlled Trials for the Treatment of Cholangiocarcinoma in Three Domains: Survival-Inferred Fragility Index, Restricted Mean Survival Time, and the Spin Effect," *Hepatobiliary Surgery and Nutrition* 11, no. 6 (2022): 861–869, https://doi.org/10.21037/hbsn-21-118.

41. A. Agresti, *Categorical Data Analysis* (New York: Wiley, 2013).

42. J. M. Lachin, *Biostatistical Methods: The Assessment of Relative Risks* (New York: Wiley, 2009).

43. J. A. C. Sterne, S. Murthy, J. V. Diaz, et al., "Association Between Administration of Systemic Corticosteroids and Mortality Among Critically Ill Patients With COVID-19: A Meta-Analysis," *Journal of the American Medical Association* 324, no. 13 (2020): 1330–1341.

44. B. Cao, Y. Wang, D. Wen, et al., "A Trial of Lopinavir–Ritonavir in Adults Hospitalized With Severe Covid-19," *New England Journal of Medicine* 382, no. 19 (2020): 1787–1799.

45. RECOVERY Collaborative Group, "Effect of Hydroxychloroquine in Hospitalized Patients With Covid-19," *New England Journal of Medicine* 383, no. 21 (2020): 2030–2040.

46. B. R. Baer, S. E. Fremes, M. Gaudino, M. Charlson, and M. T. Wells, "On Clinical Trial Fragility due to Patients Lost to Follow Up," *BMC Medical Research Methodology* 21, no. 1 (2021): 254, https://doi.org/10.1186/s12874-021-01446-z.

47. R. Carter, P. McKie, and C. Storlie, "The Fragility Index: A *p*-Value in Sheep's Clothing?," *European Heart Journal* 38 (2017): 346–348.

48. J. Li and P. J. O'Connell, "The Fragility Index: The *p*-Value by Another Name?," *Transplantation* 106, no. 2 (2022): 239–240.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix A

To maximize the concave log-likelihood function

$$l(\theta_1, \theta_2) = \sum_{i=1}^{2} \left[x_i \log\theta_i + (n_i - x_i)\log(1 - \theta_i) \right]$$

subject to the linear constraint

$$h(\theta_1, \theta_2) = \theta_1 - \theta_2 \leq 0,$$

we apply the Karush–Kuhn–Tucker conditions, that is to say the first-order conditions

$$\frac{\partial l}{\partial \theta_i} = \frac{x_i - n_i\theta_i}{\theta_i(1 - \theta_i)} = \mu \frac{\partial h}{\partial \theta_i}$$

$(i = 1, 2)$ with the multiplier

$$\mu \geq 0$$

and the complementary slackness

$$\mu(\theta_1 - \theta_2) = 0$$

which implies that there are two cases:

(1) for

$$\mu = 0,$$

the maximum point is given by

$$\theta_i = \frac{x_i}{n_i} \ (i = 1, 2)$$

if

$$\frac{x_1}{n_1} \leq \frac{x_2}{n_2}$$

due to the constraint

$$\theta_1 \leq \theta_2;$$

(2) for

$$\theta_1 - \theta_2 = 0$$

the constrained maximum is attained at

$$\theta_1 = \theta_2 = \frac{x_1 + x_2}{n_1 + n_2}$$

if

$$\frac{x_1}{n_1} \geq \frac{x_2}{n_2}$$

because of $\mu \geq 0$.