Title: Quantifying Clinical Uncertainty in times of Crisis: How Different Would a Sample have to be to Change an Inference?

Authors:

*Kenneth A. Frank, PhD, Measurement and Quantitative Methods, Education; Agriculture and Natural Resources, Michigan State University, kenfrank@msu.edu.

*Qinyun Lin, PhD, Center for Spatial Data Science, University of Chicago

*Spiro Maroulis, PhD, School of Public Affairs, Arizona State University

*Anna S. Mueller, Ph.D., Indiana University.

Ran Xu, PhD, Allied Health Sciences, University of Connecticut

Joshua M. Rosenberg, PhD, Education, Health and Human Sciences, University of Tennessee, Knoxville

Christopher S. Hayter, PhD, School of Public Affairs, Arizona State University

Ramy A. Mahmoud, M.D., MPH, Optinose, Inc.

Marynia Kolak, PhD. Center for Spatial Data Science, University of Chicago.

Thomas Dietz, PhD, Environmental Science and Policy, Sociology, Animal Studies, Michigan State University

Lixin Zhang, PhD, Epidemiology and Biostatistics, Microbiology and Molecular Genetics, Michigan State University

*equal authorship, listed alphabetically.

Corresponding author: Kenneth A. Frank, PhD, Measurement and Quantitative Methods, Education; Agriculture and Natural Resources, Michigan State University, kenfrank@msu.edu.

Quantifying Clinical Uncertainty in times of Crisis: How Different Would a Sample have to be to Change an Inference?

Abstract

Early evidence on the efficacy of a treatment often comes from single studies with a high degree of uncertainty. This is true even for well-designed and executed randomized controlled trials, as control and treatment groups can be imbalanced, unintentionally by an experimenter's action or simply by chance. Unfortunately, conventional methods for expressing that uncertainty -- standard errors and confidence intervals -- are statistical constructs notoriously prone to misinterpretation. This problem is amplified by the COVID-19 global pandemic, where it is crucial that a broad set of stakeholders have a common understanding of the strength of the inferences drawn from emerging research. In this paper, we present an approach for expressing the robustness of study inferences in terms of hypothetical changes to the underlying data. This generates statements such as "The inference would change if xx of the treatment patients who experienced a benefit were replaced by patients for whom there was no effect of the treatment." This characterizes the confidence of an inference in relatable terms that presume little statistical knowledge and, similar to the concept of fragility, can be particularly helpful in identifying studies where statistically significant results might not be particularly robust.

The COVID-19 pandemic is generating extraordinary demand -- from the public, policymakers, and medical practitioners alike -- for evidence-based strategies to save people's lives and facilitate a return to normalcy [1]. The demand for evidence in medicine is not new. Indeed, evidence-based medicine (EBM) is a cornerstone of current medical practice [2,3]. EBM stresses "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" [2].

The challenge for EBM during a rapidly evolving novel virus pandemic is that treatment strategies, research agendas, and even state-wide policies must move forward before definitive evidence can accumulate. For example, when lives are on the line, clinicians have an ethical obligation to end RCTs early and provide the potentially life-saving treatment to everyone in the clinical trial [4]. Similarly, when a patient is near death doctors may be torn between the potential of experimental treatments vs adhering to assignments of a randomized protocol (New York Times, Aug 5 The Covid Drug Wars that Pitted Doctor vs Doctor). Given this urgency, it is crucial that researchers leverage as many tools as possible to evaluate the robustness of evidence that shapes medical decision-making and our understanding of COVID-19 [1,5]. Furthermore, there is a need to present scientific evidence so that it is understandable to diverse stakeholders, including researchers, front line physicians, public officials, the media, and the public itself.

The crux of the challenge for EBM during a pandemic resides in how the uncertainty of the evidence is characterized. Even as RCTs are appreciated for their rigor, in any single RCT, those receiving the treatment may be slightly healthier than those receiving the control simply due to chance imbalances at baseline [6]. This is especially true for small trials. On subsequent trials, the imbalance may be in the other direction, where the control patients are slightly healthier than the treatment patients. Consequently, the comparison of outcomes for treatment

and control groups for any single trial may reflect baseline "imbalance" in health -- or confounding factors related to health -- between the groups. It is only over many trials and with growing size of the samples accumulated across studies that randomly unbalanced differences are expected to even out, supporting the "unbiasedness" of RCTs [6].

Standard errors are the main statistic used to characterize the inherent uncertainty due to potential imbalance between treatment and control groups for an estimated effect, which in turn informs how confident we are in an inference based on that estimate. Yet standard errors and their associated confidence intervals are theoretical statistical constructs notoriously prone to misinterpretation [7]. Thus, it is difficult to use standard errors or confidence intervals to convey uncertainty to broad audiences, including clinicians and policymakers without advanced statistical training. This scenario raises an opportunity for alternative methods for quantifying and conveying clinical uncertainty of inferences based on single RCTS as well as over the accumulation of trials [8].

In this paper, we advance the idea that quantifying the robustness of a study's inference(s) to changes in the underlying data can be used to augment interpretation of the uncertainty of inferences [9–11]. Our particular focus is on case replacement: We ask how many patients in a study sample would have to be replaced by patients for whom there was no effect of the treatment to change an inference based on statistical significance [10]? A robust inference would be one in which a large portion of the observed data would have to be replaced to change the inference, what we refer to as the Robustness of the Inference to Replacement (RIR). The RIR can show, for example, that in a small study even a finding with a small p-value (e.g., $p < 0.01$) might be overturned by the replacement of only a few cases, suggesting some uncertainty in the inference and caution regarding recommended clinical action. Case replacement has the

added advantage of being consistent with evidence that doctors and patients have an easier time making inferences from information presented in terms of natural frequencies (such as the number of patients who received the treatment that experienced a positive outcome) rather than probabilities [12,13].

In the application to dichotomous patient outcomes, such as mortality, the case replacement analysis can be executed in a way that maps closely to the existing concept of "Fragility" which has been gaining increasing attention in clinical epidemiology [8,9,11,14] with applications in oncology [15] and pediatrics [16]. The Fragility of a study asks how many patients would have to have different outcomes, or experience "event switches," to change an inference [11]. Recently, Walter, Thabane, and Briel [17] made important extensions of Fragility by developing a framework that considered minimally important differences as well as statistical significance.

In developing their framework, Walter Thabane and Briel [17] raised two important critiques of Fragility. First, they raised the concern that Fragility only uses statistical significance as a threshold for making an inference. Second, they noted that Fragility does not account for the relative prevalence of negative outcomes in the data – switching the outcome of a single case from a positive to negative outcome might be an extreme change in the data if negative outcomes are rare.

One benefit of viewing Fragility through the framework of case replacement is that it helps address the limitations identified by Walter et al. [17] . First, a case replacement framework explicitly represents thresholds for inference that include, but are not exclusive to, statistical significance. Second, case replacement is sensitive to the underlying rareness of the event. An additional benefit is that case replacement links Fragility to a more general framework

for assessing robustness of an inferences that can be applied to continuous outcomes and research designs other than RCTs [10].

Ultimately, our case replacement framework generates statements such as "The inference would change if xx of the treatment patients who experienced a benefit were replaced by patients for whom there was no effect of the treatment." Thus, our framework represents the robustness of an inference in the very relatable, tangible, terms of patient experiences. This informs debates about the bases for inferences and helps interpret the potential threat of sources of bias for an inclusive set of stakeholders.

In the following sections we briefly introduce the technical argument and motivation behind a case replacement approach to robustness and articulate its connection to Fragility for dichotomous outcomes. We then demonstrate how case replacement and Fragility can be used to examine the robustness of inferences in an emerging body of research, such as the efficacy of COVID-19 treatments, using two examples. The first is a small, preliminary RCT regarding the effect of hydroxychloroquine (HCQ) on pneumonia. The second is an application to a historical study-by-study accumulation of evidence, using a series of RCTs presented in a meta-analysis of the effects of antihypertensive treatments on stroke. Through these examples we show how quantifying the robustness of inferences in terms of case replacement can be applied to new, individual COVID-19-related RCTs as they become available, as well as to the accumulation of evidence across RCTs.

Methods: Expressing Uncertainty in terms of Changes in Outcomes

We characterize the uncertainty of an inference in terms of the changes to the data necessary to change the inference [9–11]. The specific data replacement approach we apply here draws on Frank et al. [10], which presented a non-parametric (no assumptions about theoretical statistical distributions) framework that can be applied to various types of outcome variables. To better introduce the case replacement framework, consider the idealized example in Figure 1 which compares the estimated effect to a threshold for making an inference. In Figure 1, the estimated treatment effect is 6 and just for demonstration purposes we have drawn the threshold for inference at 4. Different people may have different thresholds; researchers might employ a threshold based on statistical significance whereas clinicians might employ a threshold based on a minimally important difference [17–19]. But in all cases the threshold pragmatically links the evidence to recommended action [10,17]. That is, the threshold marks the point of indifference to the evidence. Any more evidence than the threshold and one would take an action favoring treatment A. Any less and one would not.

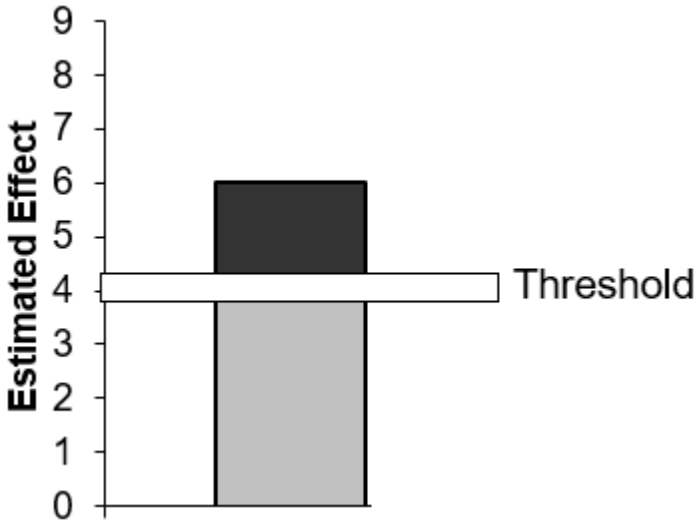

Figure 1. Estimated Effect and Threshold for Making an Inference

We can use the comparison of the estimated effect to the threshold in Figure 1 to characterize the strength of the evidence favoring the treatment. Specifically, one third of the estimated effect of 6 exceeds the threshold of 4. Correspondingly, one third of the estimated effect would have to be due to bias to change the inference [10]. Frank et al., [10] and Frank and Min [20] then demonstrate that one can interpret this difference in terms of "case replacement" between the observed sample and a hypothetical population where a null hypothesis of zero treatment effect held. Specifically, one would need to replace 1/3 of the observed cases with zero effect cases to reduce the estimated effect of 6 below the threshold for inference of 4. One could also think of this as reducing the treatment effect to zero in one third of the cases. The proportion of the cases that one must replace (or reduce the treatment effect to zero) to change the inference quantifies the robustness of the inference. The larger the proportion, the more robust the inference.

Formally, to calculate the changes in the data necessary to modify an estimated effect to a specific value, define the modified value ($\bar{\hat{\delta}}$) as a function of the observed estimated effect ($\hat{\delta}_o$) and the hypothesized effect in the unobserved replacement data (δ_u) [20–22]. Assuming the proportion of units receiving the treatment is the same in the observed and unobserved data, an expression for $\bar{\hat{\delta}}$ is:

$$\bar{\hat{\delta}} = (1-\pi)\hat{\delta}_o + \pi\delta_u, \quad (1)$$

where π represents the proportion of observed cases replaced by unobserved cases. Therefore $\bar{\hat{\delta}}$ is a mixture, according to π, of the estimate from the observed data ($\hat{\delta}_o$) with the effect in the unobserved data (δ_u).

To determine the conditions necessary to change an inference, we first assume a null hypothesis of zero effect holds exactly in the unobserved data: $\delta_u = 0$ [10,23,24]. For example, $\delta_u = 0$ holds exactly if the unobserved data are generated from a null hypothesis of zero effect and there is no sampling variability because there is no covariate imbalance. Assuming $\delta_u = 0$ yields:

$$\bar{\hat{\delta}} = (1-\pi)\,\hat{\delta}_o. \quad (2)$$

Next, set $\bar{\hat{\delta}} = \delta^{\#}$ where $\delta^{\#}$ defines the threshold for making an inference such as an estimate associated with an effect size of specific clinical significance [17,18] or with a p-value of .05:

$$\bar{\hat{\delta}} = \delta^{\#} = (1-\pi)\,\hat{\delta}_o. \quad (3)$$

Solving for π yields:

$$\pi = 1-\delta^{\#} / \hat{\delta}_o. \quad (4)$$

The expression in (4) allows one to calculate what percent of the cases (π) in the observed sample would have to be replaced with unobserved zero effect cases to reduce the combined estimate ($\bar{\hat{\delta}}$) below the threshold ($\delta^{\#}$) for making an inference [10]. For instance, in the simple example in Figure 1 where $\hat{\delta}_o = 6$ and $\delta^{\#} = 4$, $\pi = 1 - 4/6 = 1/3$, implying that to change the inference, 1/3 of the observed cases would have to be replaced with cases from a sample in which there was no effect of the treatment. This allows us to express uncertainty by conceptualizing how the existing data could be mixed with null hypothesis cases instead of in terms of the standard error – the theoretical standard deviation of estimated effects under the null hypothesis.

The general case replacement approach has been applied extensively across the social

sciences [25–27], physical sciences [28–30], and in policy [31,32], but it can also provide unique insight to dichotomous outcomes often used in health (e.g., “Improved” versus “Not Improved”; “Survived” versus “Deceased”). Adapting the general case replacement approach expressed in (4) to studies with dichotomous outcomes requires considering both the treatment status and outcome of the cases to be replaced. For example, one could replace cases from any or all the cells in a 2 x 2 contingency table that categorizes observations by treatment /control and “survived”/“deceased.”

In this paper we made two choices about the cases to be replaced that facilitate interpretation in a clinical context as well as comparison to related approaches in epidemiology [11]. First, we choose to replace cases primarily from the treatment “survived” category of observations, and define the Robustness of an Inference to Replacement (RIR) as the number of treatment success cases that would have to be replaced with zero effect cases to invalidate an inference given a particular threshold (the RIR could also be defined based on replacing cases in both the treatment and control rows). Second, we assume that all cases in the same treatment/outcome category are indistinguishable or exchangeable [33,34]. As a result, when replacing observed cases with unobserved zero effect cases, the only clear “changes” after replacement would be when successful cases were replaced with unsuccessful ones, and vice versa (i.e., situations in which success is replaced with success cannot be distinguished). This allows us to represent the hypothetical change in the sample not only in terms of replacement, but also in terms of “switching” of “improved” to “not improved” outcomes.

If one counts switches between events instead of replacements from a hypothetical sample, and uses statistical significance from zero as the threshold for inference, the result is the robustness measure of “Fragility” which has been recently reintroduced in epidemiology [9,11].

Walter, Thabane and Briel [17] recently critiqued the use of fragility on two fundamental grounds. First, fragility relies on statistical significance (page 35). Second, Walter et al [17] show that fragility is not directly sensitive to the underlying rareness of the event (page 36).

The RIR can directly address both of these critiques of Fragility. First, using the case replacement approach, as in Figure 1, indicates any threshold can be used as a basis for inference. Regarding the likelihood that a case will be replaced, consider the example in Table 1, following Walter et al [17]. These results are from a hypothetical experiment where 90/95 patients given Treatment A survived, 96/96 patients given treatment B survived, with a p-value of .029 (based on Fisher's exact test) leading to the inference that treatment B is more effective than treatment A. Walter et al. [17] note that the Fragility of this inference equals 1 – if one "Survived" case in treatment B were switched to a "Deceased" case, the success rate would change to 95/96 in treatment, with the corresponding p-value would change to 0.118. correspondingly, if one uses a threshold of p=.05, the one switch would lead one to an inference that there is no difference between treatments A and B. Walter et al note that the fallacy in this [the Fragility] argument is that the change from 0 to 1 death in treatment group B may actually be unlikely to occur.

	Deceased		**Survived**	**Total**
Treatment A	5		90	95
Treatment B	0	Fragility=1 ←	96 [RIR 38]	96
Total	5		186	191

Table 1. Robustness of Inference for hypothetical treatment and mortality. Example taken from Walter et al. [17]. Cells represent number of cases. Fragility = number of cases to switch to change the inference; RIR represents the robustness of the inference to replacement.

Walter et al's concern can be expressed by considering how switches are generated from

case replacement. In particular, we ask how many of the 96 Treatment B "Survived" cases would have to be replaced with zero effect cases to change the inference that Treatment B was more efficacious than Treatment A. We begin by assuming the distrubution for the replacement cases is defined by the mortality rate in the whole data – 5/191 or 2.6%. This is the expected mortality rate under the null hypothesis of no treatment effects, and therefore all the data are relavent for estimating the prevalence of mortality. Using the 2.6% mortality rate, for every 38 Treatment B "Survival" cases replaced, 37 would remain classified as "Survived," and 1 would be reclassified as "Deceased." Therefore, one would have to replace 38 Treatment B "Survived" cases with zero-effect cases to generate the one Treatment "Deceased" case necessary to change the inference ($p = 0.118$). RIR=38 while Fragility=1.

Formally, Fragility can be expressed as the expected number of replaced treatment successes multiplied by the overall observed probability of failure: Fragility = RIR x $\hat{p}$, where $\hat{p}$ is the observed probability of failure. This implies that RIR = Fragility/ $\hat{p}$. In the example, 38=1/.026. Thus, RIR is a funciton of $\hat{p}$, addressing Walter et al's critique of Fragility by incorporating the prevalence of outcomes in the data. [1] [2]

[1] Note that Fragility is defined only for results that are positive and statistically significant. For those results that are not statistically significant Fragility is technically undefined [11] although it would be possible to count the number of switches necessary to increase an estimated effect above a threshold for positive statistical significance, with a corresponding, Fragility= $\hat{p}$ x RIR, where = $\hat{p}$ would be based on the number of overall success rate instead of the failure rate.

[2] We could have estimated the mortality replacement rate ($\hat{p}$) from only the control group – 5/95, or about 5.2%, resulting in an RIR of 20. In this scenario the control group would represent the current state of mortality in the absence of the new treatment. In general, using the mortality rate from only the control group would be more conservative (fewer cases to replace indicating a less robust inference) in the event that a positive treatment effect was estimated. In the example we chose the mortality of the whole group to represent the null hypothesis of zero

Results: Using Robustness of the Inference to Replacement (RIR) to Express Uncertainty

Inference Regarding the effect of Hydroxychloroquine (HCQ) on Pneumonia

Consider one of the first reports of a randomized trial for the drug hydroxychloroquine (HCQ) [35]. Conducted at the Renmin Hospital of Wuhan University, 31of 62 patients were randomly assigned to receive HCQ in addition to the standard treatment. Pneumonia in 25 treatment patients "improved" moderately or significantly while 17 control patients "improved" moderately or significantly, resulting in a difference in success rates of 26 percentage points (25/31 - 17/31 = .26; $\chi^2 = 4.7$, $p = 0.03$; Table 2), and the conclusion that HCQ is efficacious. A critical challenge to this study is that it was not double-blinded [36]. Therefore, the researchers, physicians and patients could have been influenced in their behavior or labeling of the outcomes by knowledge of the treatment assignment, making the need to contextualize the uncertainty of the inference all the more necessary. The question we pose then is how many of the HCQ patients labeled as "improved" would have to be replaced to change the inference that HCQ reduced pneumonia?

	Exacerbated or Unchanged	**Improved (Moderate or Significant)**	**Total**
Conventional Treatment	14 {A}	17 {B}	31
Hydroxychloroquine	6 {C} ← Fragility=1	25 [RIR 3] {D}	31
Total	20	42	62

Table 2. Robustness of inference for hydroxychloroquine (HCQ) vs Conventional Treatments on Pneumonia. Data from Table 2 of Chen et al [35].[4] Cells represent number of cases. RIR represents the robustness of the inference to replacement with null hypothesis cases. Fragility = number of cases to switch to change the inference;

effect.

To answer the question in the preceding paragraph we consider replacing cases from HCQ "improved" with cases with probability of failure in the entire sample ($\hat{p}$ = 20/62 = 0.32). We replace cases until the $\chi^2 < .05$. Table 2 illustrates the result. If about three of the 25 HCQ "improved" cases were replaced with cases for which $\hat{p}$ =.32, we would expect one of those cases would "switch" from success to failure (.32 x 3), and the probability difference between HCQ and the control would drop to a magnitude that would no longer be statistically significant at the 5% level (24/31 - 17/31 = 23 percentage point difference, RIR = 3 , Fragility = 1, calculations conducted using http://konfound-it.com). The low number of replacements needed, whether using the RIR or Fragility, highlights the tenuous nature of the inference of an efficacious result despite its statistical significance.

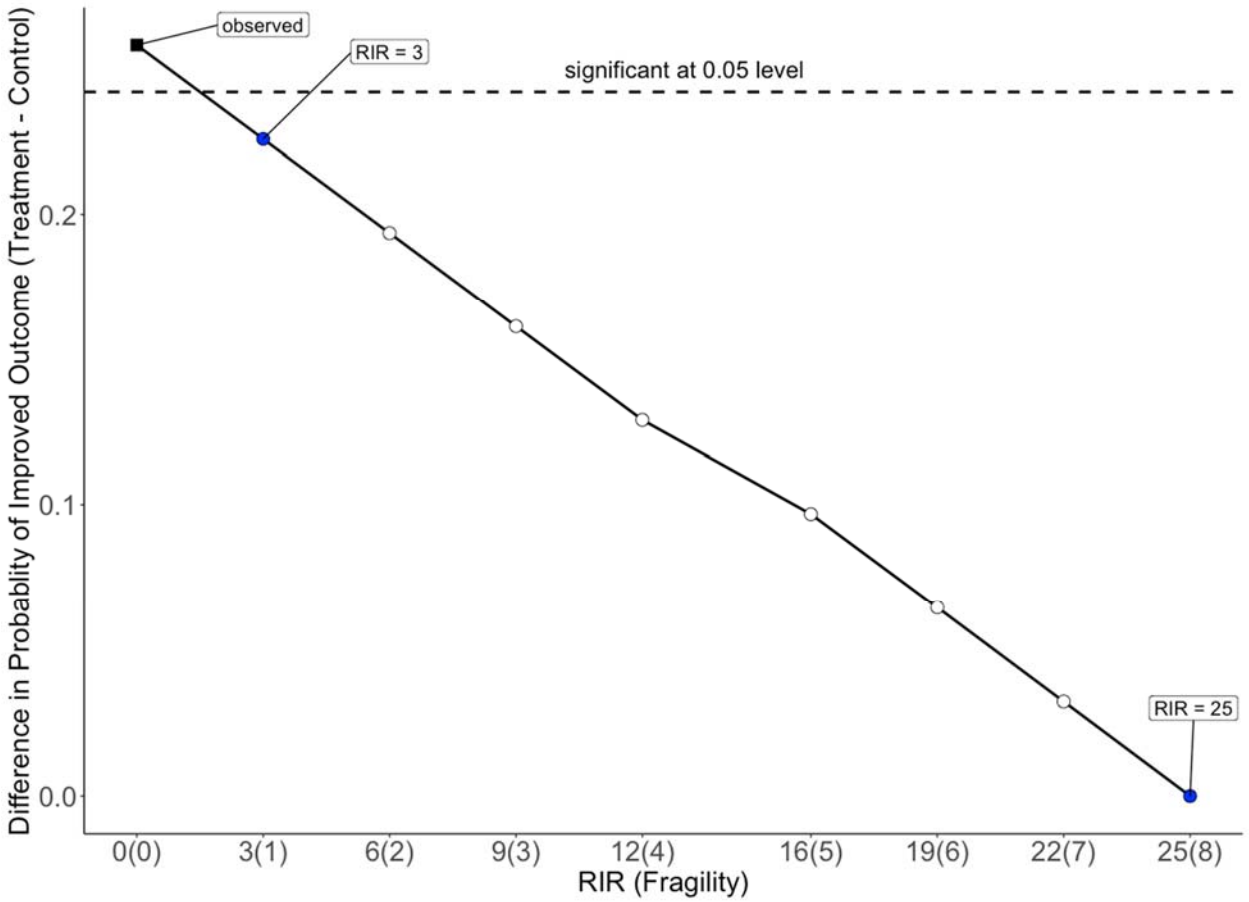

Figure 2. Difference in probability of improved outcome (treatment – control) after replacing observed cases in HCQ example. Black square, study estimate; dashed line, positive estimate significant at 5% level.

In Figure 2 we extend the HCQ example by plotting RIR against corresponding estimated effect sizes along a continuum to represent a broader potential set of thresholds [17]. Each data point represents the RIR to reduce the estimated effect in the HCQ example in Table 2 below a particular effect size. Consistent with Table 2, one would have to replace 3 of the observed treatment "improved" cases with cases for which $\hat{p}$ = .32 to reduce the estimated effect of .26 below the threshold (probability difference of .24) for statistical significance at the .05 level for a positive finding. But Figure 2 also shows an RIR of about 16 to reduce the initial probability difference of .26 to 0.10. That is, 16 of the treatment "improved" cases would have to be

replaced to reduce the estimated effect to 0.10. These 16 replacements would generate five switches – Fragility=5. More generally, Figure 3 represents the RIR with respect to any effect size, including effect sizes that define a minimal important difference [17–19]. [3]

Historical Example: Inference from Accumulation of Estimates of Antihypertensive Treatments.

The RIR can also be applied to inform uncertainty about estimates of treatment effects accumulated across a series of studies. This is an important complement to guidelines for characterizing the quality of evidence such as GRADE [37]. The extremely valuable guidelines describe the quality of a body of evidence in terms of aspects of the study design (e.g., non-random assignment to treatment, non-blinded assignment to treatment, differential attrition from treatment and control) that could cause bias. Our comparison of the strength of evidence relative to a pragmatic threshold can be particularly useful for evaluating accumulating evidence [8] especially in the context of the COVID-19 pandemic.

Online dashboards, such as the COVID-19 clinical trials registry (https://www.covid-trials.org/), provide near real-time tracking and categorization of findings accumulating across emerging research [38]. To illustrate what the accumulation of estimated effects of COVID-19 treatments might look like, we use a well-established example from the clinical trials literature of the effect of antihypertensive treatments on the probability of a stroke [39].

	Control		**Treatment**						
	Stroke	**No Stroke**	**Stroke**	**No Stroke**	**Total**	**Decrease in Stroke Probability for the**	**p value (Fisher)**	**RIR**	**Fragility**

[3] Generally, to reduce the difference in probability below any threshold $\delta^{\#}$, one can switch x cases according to: $x=D-(\delta^{\#}+B/(A+B))(D+C)$ where the letters refer to the cells in Table 2.

						Treatment Group			
Study 1: Wolff	1	41	2	43	87	-2.1	1	NA	NA
Study 2: VA II	20	174	5	181	380	7.6	0.003	61	4
Study 1 + Study 2	21	215	7	224	467	5.9	0.010	50	3

Table 3. Robustness of inferences to replacement (RIR) for antihypertensive treatment on stroke
Control, treatment, and total contain number of cases from Table II of Collins et al [39]. Remaining columns based on authors' calculations.

Table 3 presents the robustness of inferences to replacement (RIR) for effects of antihypertensive treatments on patient strokes for the first and second studies in Collins' et al [39] antihypertension meta-analysis. Study 1 concluded that antihypertensive treatments were not associated with a decrease in strokes with a p-value of 0.6. Therefore, we will not apply the RIR. The second study found a 7.6 percentage point *decrease* in stroke probability for the treatment group. This result is associated with a *p* value of .003. To change this inference, one would have to replace 61 (about 34%) of the treatment "no stroke" cases with cases for which $\hat{p}$ = 25/380 = .07. These 61 replacements would generate approximately 4 switches from treatment "no stroke" to treatment "stroke" (Fragility = 4). The RIR for the first two studies combined is 50 (Fragility = 3). Note that even though the number of total observations substantially increased, combining studies 1 and 2 results in a conclusion that is no more robust than the inference from Study 2 alone. This is in part because the estimate in Study 1, although not statistically significant, indicated a negative treatment effect for a relatively small sample.

As evidence from multiple RCTs accumulates, adding the RIR to meta-analyses of RCTs can help assess and visualize the robustness of inferences beyond reporting or examining p-values. For example, in Figure 3 we present a series of "robustness" updates as each study was

added in the hypertensive meta-analysis, where each subsequent point presents an updated estimated effect as well as corresponding RIR. Critically, the combined estimated treatment effect fluctuated by several percentage points until the 8th study (Year = 1979). As studies progressed, the estimated treatment effect stabilized and the number of replacements it would take to invalidate the inference increased substantially.[4] Continuous updates to an analogous figure using COVID-19 studies would present decision-makers with an up-to-date and intuitive characterization of combined estimates as well as the robustness of the inferences drawn from

[4] We also calculated RIR for each point in Figure 3 using fixed effects adjustments from cumulative meta-analysis [40]. Specifically, we first generated an implied two by two contingency table based on the estimated effect, standard error, as well as the sample size and the number of cases in the treatment group. Technically, there are always two solutions for a contingency table that satisfy these conditions. We intentionally picked the one with a similar overall success rate as the original table, as the overall success rate is essential for calculating RIR. Then we calculated RIR for the implied two by two contingency table. The results are similar to those in Figure 3: from the year 1967, the RIRs are 35, 38, 142, 149, 124, 216, 298, 1234, 1541, 1886, 2061, 2667, 6190, 6256, 6564, respectively.

scientific evidence.

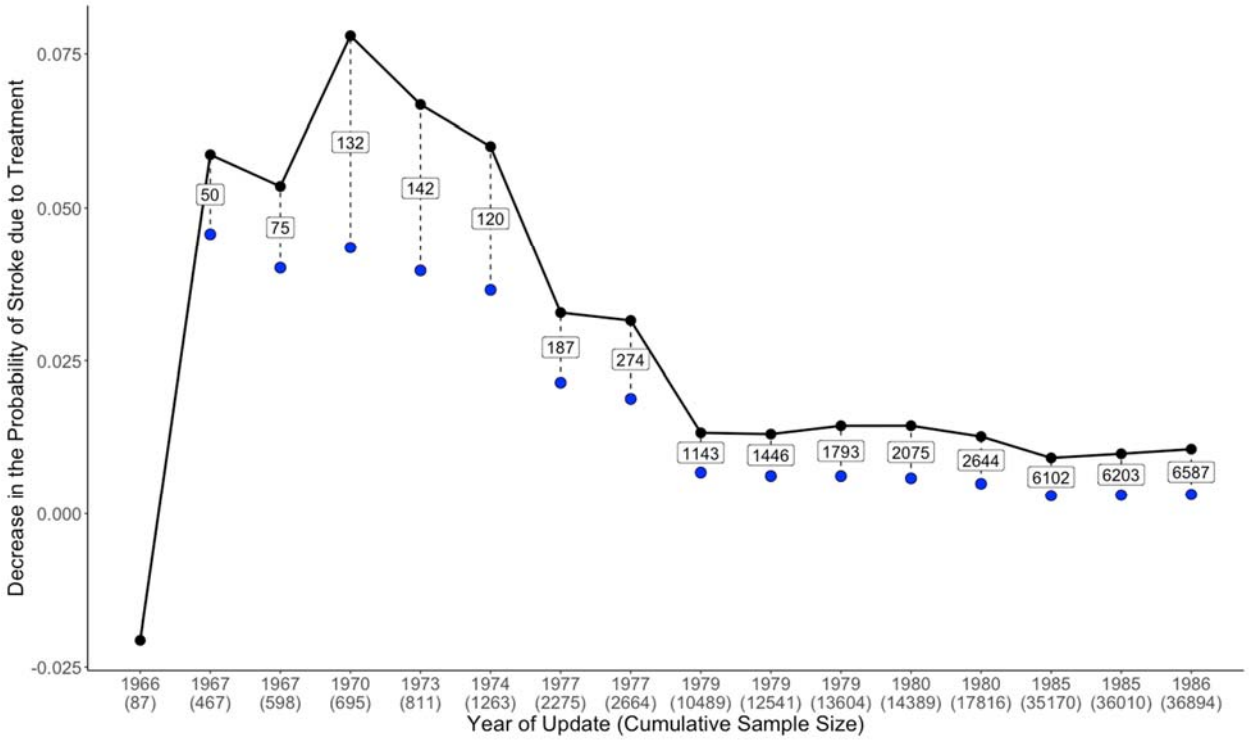

Figure 3. Robustness of Inferences to Replacement (RIR) as Evidence Accumulates: Historical Case of Antihypertension Treatment and Stroke. Black dots indicate the size of the estimated treatment effect based on all studies available up to that point in time; blue dots, the effect size just below statistical significance. Boxes label the corresponding RIR.

Discussion

While there is always a need to characterize the uncertainty of estimates and inferences generated from scientific research, that need is amplified during emergent crises like the COVID-19 pandemic [1]. Physicians and policymakers are under tremendous pressures to rapidly adapt best practices to protect public health and prevent mortality due to COVID-19, often with limited, emerging research. However, relying too heavily on results from single, early trials, even a well conducted RCT, can be problematic because in any single trial random assignment to control and treatment groups can be imbalanced, unintentionally by an experimenter's action or just by chance [6]. As a result, the uncertainty associated with the

conclusions drawn from a single trial should be quantified adding nuance to simple yes/no thresholds for statistical significance. The goal of this article was to offer the RIR as a way to express clinical uncertainty and to do so in clear terms of patient experiences. We also demonstrate how the RIR is particularly valuable when there is not yet a large accumulation of evidence, such as when there is only a single small or moderate sized RCT early in the trajectory of the accumulation of evidence (see Figure 3, above, particularly the left side).

Currently, the idea of statistical significance and the specific p-value are generally relied upon to convey the robustness of an inference. Consider the following quote from Anthony Fauci (as in Healio on April 29):

> The trial, which began Feb. 21 this year, compared Remdesivir with placebo in more than 1,000 patients. Remdesivir improved recovery from 15 days to 11 days, with a *P* value of 0.001….We would have normally waited several more days. The data may change, but the conclusion won't.

Similarly, Martin Landray, deputy chief investigator of a trial recently reported in the popular media regarding the use of dexamethasone to treat Covid-19: "That's a highly statistically significant result." In the former example, note the extremely small p-value is being used to convey the robustness of the conclusion. In the latter example, the language of "highly statistically significant" is seemingly in reference to a small p-value. In both cases, though technically correct, neither is an intuitive or precise description of the robustness of the inference. While some might argue that this difficulty is indicative of a deeper problem that requires the wholesale replacement of the null hypothesis significance testing paradigm [40,41], as illustrated in the quotes above p-values remain central to scientific discourse [9,42]. Quantifying robustness through case replacement can help mitigate the disadvantages of p-values

in several ways. First, instead of relying on an understanding of sampling distributions which may be unfamiliar to many, it generates statements such as "The inference would change if xx of the treatment patients who experienced a benefit were replaced by patients for whom there was no effect of the treatment." When the RIR is large, an accessible phrase might be "One way we know this answer is robust/stable is that to change our conclusions a large number (xx) of patients would have to be replaced with patients for whom the treatment had no effect." Thus, the RIR represents the robustness of the conclusion in the very relatable, tangible, terms of patient experiences. This can inform debates about the bases for conclusions and help interpret the potential threat of sources of bias for an inclusive set of stakeholders.

Second, in addition to being more accessible, a case replacement approach can quantify the robustness of an inference with greater precision than can be accomplished through using a p-value alone. To illustrate the concern with using only p-values to convey the robustness of a study's inference, consider the relationship between p-values and the RIR depicted in Figure 4 (assuming a p-value of .05 is used as a basis of a conclusion). As the p-value becomes smaller the RIR increases. More importantly, note that in this example even though it might be difficult to see or conceptualize the difference between *p* of .01 and *p* of .001, it is more direct to understand that the inference would change if 40 versus 100+ of the cases were replaced.

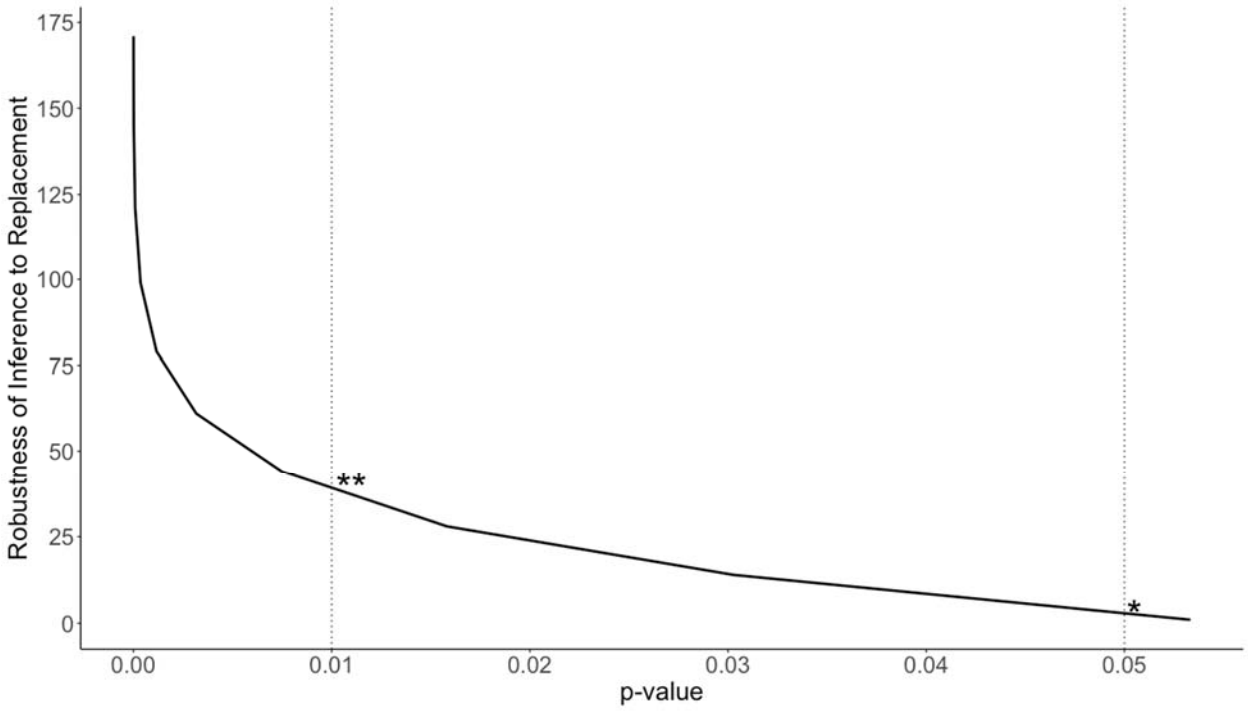

Figure 4. Robustness of Inference to Replacement (RIR) vs p-value. The curve shows functional relationship between RIR and p-value. Baseline table has 380 cases evenly divided between treatment and control with odds ratio favoring the treatment of 1.22. The steep slope on the left indicates that RIR conveys differences in uncertainty even when very small p-values are almost indistinguishable and difficult to interpret.

While not the focus of the examples in this paper, a case replacement approach to robustness can also provide benefit when concerns go beyond p-values into deeper concerns about bias [10]. For example, the RIR could be applied to observational studies which are being implemented during the early stages of the COVID pandemic [43]. In this scenario, the replacement cases are conceptualized as coming from counterfactual data in which those receiving the treatment would instead have received the control, and vice versa. The RIR then quantifies the robustness of the inference to uncertainty introduced by approximating the counterfactual with non-experiment techniques. In this application and others, the RIR can be

extended to replace treatment cases, control cases or a random sample of cases as in Frank et al [10]. In any application, the RIR helps weigh the strength of the evidence against concerns about violations of assumptions in the specific context of a given study [44]. But we emphasize the RIR is not a substitute for assessing the methodologic strength of a study – it is a helpful tool for understanding and communicating the stability or robustness of any given conclusion based on valid data.

No single sensitivity measure, including the RIR, is a panacea. But sensitivity measures can facilitate a common understanding among researchers, policymakers, journalists, clinicians, and the public about the strength of the evidence of potential interventions. This is crucial when, as a society, we must quickly weigh the expected benefits and harms of an intervention against the consequences of inaction.

References

[1] Djulbegovic B, Guyatt G. Evidence-based medicine in times of crisis. J Clin Epidemiol 2020. https://doi.org/10.1016/j.jclinepi.2020.07.002.

[2] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71–2. https://doi.org/10.1136/bmj.312.7023.71.

[3] Timmermans S, Berg M. The Gold Standard: The Challenge Of Evidence-Based Medicine. Philadelphia, PA: Temple University Press; 2003.

[4] Mueller PS, Montori VM, Bassler D, Koenig BA, Guyatt GH. Ethical Issues in Stopping Randomized Trials Early Because of Apparent Benefit. Ann Intern Med 2007;146:878–81. https://doi.org/10.7326/0003-4819-146-12-200706190-00009.

[5] London AJ, Kimmelman J. Against pandemic research exceptionalism. Science 2020;368:476–7. https://doi.org/10.1126/science.abc1731.

[6] Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Soc Sci Med 2018;210:2–21. https://doi.org/10.1016/j.socscimed.2017.12.005.

[7] Pocock SJ, Ware JH. Translating statistical findings into plain English. The Lancet 2009;373:1926–8. https://doi.org/10.1016/S0140-6736(09)60499-2.

[8] Atal I, Porcher R, Boutron I, Ravaud P. The statistical significance of meta-analyses is frequently fragile: definition of a fragility index for meta-analyses. J Clin Epidemiol 2019;111:32–40. https://doi.org/10.1016/j.jclinepi.2019.03.012.

[9] Feinstein AR. The unit fragility index: An additional appraisal of "statistical significance" for a contrast of two proportions. J Clin Epidemiol 1990;43:201–9. https://doi.org/10.1016/0895-4356(90)90186-S.

[10] Frank KA, Maroulis SJ, Duong MQ, Kelcey BM. What Would It Take to Change an Inference? Using Rubin's Causal Model to Interpret the Robustness of Causal Inferences. Educ Eval Policy Anal 2013;35:437–460.

[11] Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clin Epidemiol 2014;67:622–8. https://doi.org/10.1016/j.jclinepi.2013.10.019.

[12] Garcia-Retamero R, Hoffrage U. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. Soc Sci Med 2013;83:27–33. https://doi.org/10.1016/j.socscimed.2013.01.034.

[13] Whiting PF, Davenport C, Jameson C, Burke M, Sterne JAC, Hyde C, et al. How well do health professionals interpret diagnostic information? A systematic review. BMJ Open 2015;5:e008155. https://doi.org/10.1136/bmjopen-2015-008155.

[14] Tugwell P, Knottnerus A. Should the fragility index be routinely reported for systematic reviews? J Clin Epidemiol 2019;111:v–vi. https://doi.org/10.1016/j.jclinepi.2019.05.008.

[15] Forrester LA, Jang E, Lawson MM, Capi A, Tyler WK. Statistical Fragility of Surgical and Procedural Clinical Trials in Orthopaedic Oncology. JAAOS Glob Res Rev 2020;4:e19.00152. https://doi.org/10.5435/JAAOSGlobal-D-19-00152.

[16] Rickard M, Keefe DT, Drysdale E, Erdman L, Hannick JH, Milford K, et al. Trends and relevance in the bladder and bowel dysfunction literature: PlumX metrics contrasted with fragility indicators. J Pediatr Urol 2020:S1477513120303910. https://doi.org/10.1016/j.jpurol.2020.06.015.

[17] Walter SD, Thabane L, Briel M. The fragility of trial results involves more than statistical significance alone. J Clin Epidemiol 2020;124:34–41. https://doi.org/10.1016/j.jclinepi.2020.02.011.
[18] Angst F, Aeschlimann A, Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol 2017;82:128–36. https://doi.org/10.1016/j.jclinepi.2016.11.016.
[19] Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102–9. https://doi.org/10.1016/j.jclinepi.2007.03.012.
[20] Frank K, Min K-S. Indices of Robustness for Sample Representation. Sociol Methodol 2007;37:349–92. https://doi.org/10.1111/j.1467-9531.2007.00186.x.
[21] Cronbach LJ, Shapiro K. Designing evaluations of educational and social programs /. San Francisco: Jossey-Bass,; 1982.
[22] Fisher RA. Statistical methods for research workers (Darien, Conn.: Hafner Pub. Co.) 1970.
[23] Cinelli C, Hazlett C. Making sense of sensitivity: extending omitted variable bias. J R Stat Soc Ser B Stat Methodol 2020;82:39–67. https://doi.org/10.1111/rssb.12348.
[24] VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med 2017;167:268–74. https://doi.org/10.7326/M16-2607.
[25] Asensio OI, Delmas MA. Nonprice incentives and energy conservation. Proc Natl Acad Sci 2015;112:E510. https://doi.org/10.1073/pnas.1401880112.
[26] Dietz T. Altruism, self-interest, and energy consumption. Proc Natl Acad Sci 2015;112:1654. https://doi.org/10.1073/pnas.1423686112.
[27] Moumen N, Ben Othman H, Hussainey K. Board structure and the informativeness of risk disclosure: Evidence from MENA emerging markets. Adv Account 2016;35:82–97. https://doi.org/10.1016/j.adiac.2016.09.001.
[28] Carrico AR, Vandenbergh MP, Stern PC, Dietz T. US climate policy needs behavioural science. Nat Clim Change 2015;5:177–9. https://doi.org/10.1038/nclimate2518.
[29] Callen AL, Dupont SM, Pyne J, Talbott J, Tien P, Calabrese E, et al. The regional pattern of abnormal cerebrovascular reactivity in HIV-infected, virally suppressed women. J Neurovirol 2020. https://doi.org/10.1007/s13365-020-00859-8.
[30] Xu R. Statistical methods for the estimation of contagion effects in human disease and health networks. Comput Struct Biotechnol J 2020;18:1754–60. https://doi.org/10.1016/j.csbj.2020.06.027.
[31] Strunk KO, Goldhaber D, Knight DS, Brown N. Are There Hidden Costs Associated With Conducting Layoffs? The Impact of Reduction-in-Force and Layoff Notices on Teacher Effectiveness. J Policy Anal Manage 2018;37:755–82. https://doi.org/10.1002/pam.22074.
[32] Frank KA, Penuel WR, Krause A. WHAT IS A “GOOD” SOCIAL NETWORK FOR POLICY IMPLEMENTATION? THE FLOW OF KNOW-HOW FOR ORGANIZATIONAL CHANGE: What is a “Good” Social Network for Policy Implementation? J Policy Anal Manage 2015;34:378–402.
[33] Bernardo JM. The Concept of Exchangeability and its Applications n.d.:7.
[34] de Finetti B. Foresight: Its Logical Laws, Its Subjective Sources. In: Kotz S, Johnson NL, editors. Breakthr. Stat. Found. Basic Theory, New York, NY: Springer; 1992, p. 134–74. https://doi.org/10.1007/978-1-4612-0919-5_10.

[35] Chen Z, Hu J, Zhang Z, Jiang S, Han S, Yan D, et al. Efficacy of hydroxychloroquine in patients with COVID-19: results of a randomized clinical trial. MedRxiv 2020:2020.03.22.20040758. https://doi.org/10.1101/2020.03.22.20040758.
[36] Pacheco RL, Riera R. Hydroxychloroquine and chloroquine for COVID-19 infection. Rapid systematic review. J Evid-Based Healthc 2020;2. https://doi.org/10.17267/2675-021Xevidence.v2i1.2843.
[37] Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401–6. https://doi.org/10.1016/j.jclinepi.2010.07.015.
[38] Thorlund K, Dron L, Park J, Hsu G, Forrest JI, Mills EJ. A real-time dashboard of clinical trials for COVID-19. Lancet Digit Health 2020;2:e286–7. https://doi.org/10.1016/S2589-7500(20)30086-8.
[39] Collins R, Peto R, MacMahon S, Hebert P, Fiebach NH, Eberlein KA, et al. Blood pressure, stroke, and coronary heart disease. Part 2, Short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. Lancet Lond Engl 1990;335:827–38. https://doi.org/10.1016/0140-6736(90)90944-z.
[40] Rothman KJ. Special Article: Writing for Epidemiology. Epidemiology 1998;9:333–7.
[41] Harrington D, D'Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand S-LT, et al. New Guidelines for Statistical Reporting in the Journal. N Engl J Med 2019;381:285–6. https://doi.org/10.1056/NEJMe1906559.
[42] Goodman SN. Of P-Values and Bayes: A Modest Proposal. Epidemiology 2001;12:295–297.
[43] Miller A, Reandelar MJ, Fasciglione K, Roumenova V, Li Y, Otazu GH. Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study. MedRxiv 2020:2020.03.24.20042937. https://doi.org/10.1101/2020.03.24.20042937.
[44] Sampson RJ. After the experimental turn: A commentary on Deaton and Cartwright. Soc Sci Med 2018;210:67–9. https://doi.org/10.1016/j.socscimed.2018.04.013.