Journal of Clinical Epidemiology 134 (2021) 150–159

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Hypothetical case replacement can be used to quantify the robustness of trial results

Kenneth A. Frank[a,#,*], Qinyun Lin[b,#], Spiro Maroulis[c,#], Anna S. Mueller[d,#], Ran Xu[e], Joshua M. Rosenberg[f], Christopher S. Hayter[c], Ramy A. Mahmoud[g], Marynia Kolak[b], Thomas Dietz[h], Lixin Zhang[i]

[a] *Measurement and Quantitative Methods, Education; Agriculture and Natural Resources, Michigan State University, East Lansing, MI*
[b] *Center for Spatial Data Science, University of Chicago, Chicago IL*
[c] *School of Public Affairs, Arizona State University, Phoenix, AZ*
[d] *Department of Sociology, Indiana University, Bloomington, IN*
[e] *Department of Allied Health Sciences, University of Connecticut, Storrs, CT*
[f] *Education, Health and Human Sciences, University of Tennessee, Knoxville*
[g] *Optinose, Inc. Yardley, PA*
[h] *Environmental Science and Policy, Sociology, Animal Studies, Michigan State University, East Lansing, MI*
[i] *Epidemiology and Biostatistics, Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI*

Accepted 8 January 2021; Available online 15 March 2021

Abstract

Objectives: We apply a general case replacement framework for quantifying the robustness of causal inferences to characterize the uncertainty of findings from clinical trials.

Study design and setting: We express the robustness of inferences as the amount of data that must be replaced to change the conclusion and relate this to the fragility of trial results used for dichotomous outcomes. We illustrate our approach in the context of an RCT of hydroxychloroquine on pneumonia in COVID-19 patients and a cumulative meta-analysis of the effect of antihypertensive treatments on stroke.

Results: We developed the Robustness of an Inference to Replacement (RIR), which quantifies how many treatment cases with positive outcomes would have to be replaced with hypothetical patients who did not receive a treatment to change an inference. The RIR addresses known limitations of the Fragility Index by accounting for the observed rates of outcomes. It can be used for varying thresholds for inference, including clinical importance.

Conclusion: Because the RIR expresses uncertainty in terms of patient experiences, it is more relatable to stakeholders than *P*-values alone. It helps identify when results are statistically significant, but conclusions are not robust, while considering the rareness of events in the underlying data.

Keywords: Robustness of findings; Randomized controlled trials; Fragility; Case replacement; Statistical significance; Clinical importance

In 2020, the COVID-19 pandemic generated extraordinary demand—from the public, policymakers, and medical practitioners alike—for evidence-based strategies to save people's lives and facilitates a return to normalcy [1]. But the priority placed on evidence-based medicine [2,3] and the need to assess the robustness of medical evidence as it emerges is not unique to times of crisis. Indeed, a cornerstone of medical practice and clinical decision-making is the use of evidence-based medicine (EBM), which 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 is that treatment decisions, research agendas, and even state-wide policies sometimes must move forward before definitive evidence can accumulate, particularly during any period of rapidly evolving science. When there are immediate implications of new findings—as there often are with the COVID-19 Pandemic

Conflicts of interest: We declare no competing interests and the works was not funded. No ethics committee review was needed for this commentary as it was an analysis of published results. We declare no competing interests and the works was not funded. No ethics committee review was needed for this commentary as it was an analysis of published results.

Equal authorship, listed alphabetically.
* Corresponding author. Tel.: 5172851587.
E-mail address: kenfrank@msu.edu (K.A. Frank).

https://doi.org/10.1016/j.jclinepi.2021.01.025
0895-4356/

What is new?

- We introduce a case replacement framework for sensitivity analysis of clinical trials.
- The framework supports statements such as "The inference would change if xx of the treatment patients who experienced positive outcomes were replaced by hypothetical patients who did not receive a treatment."
- The framework complements the Fragility Index by accounting for the rarity of negative outcomes. For example, large case replacement is required when the Fragility Index is small but negative outcomes are rare.
- The framework can be used for any threshold, including minimally important differences and statistical significance.
- The framework applies to a broad set of models and research designs.

[1,4]—it is crucial that researchers leverage as many tools as possible to evaluate the robustness of evidence that shapes clinical decision-making. Furthermore, for salient public health issues, 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. Therefore, in this paper, we advance a method for quantifying the robustness of a study's inference based on changes in the underlying data. This method can be used to quantify the robustness of inferences from many study designs [5]; however, in this paper we focus on application to Randomized Control Trials (RCT) in medicine.

Our approach to characterizing the robustness of an inference is based on case replacement. In an RCT, this involves asking how many patients randomly assigned to the treatment group who experienced a positive outcome would have to be replaced with hypothetical patients who did not receive a treatment in order to change an inference. This can support statements such as "The inference would change if xx of the n treatment patients who experienced positive outcomes were replaced by hypothetical patients who did not receive a treatment." This approach to quantifying the robustness of an inference is consistent with evidence that suggests that doctors and patients have an easier time making inferences from information presented in terms of natural frequencies (such as the number of treatment cases that experienced a positive outcome) rather than probabilities [6,7]. Furthermore, though we acknowledge recent and past critiques regarding the use of statistical significance as the threshold for clinical decision-making [8,9], some of our examples will use thresholds based on statistical significance to determine what it would take to change our inference. This is largely because statistical significance still dominates the public conversation (see the Discussion); however, our approach can be generalized to other thresholds, such as those identified by the minimally important difference, which is another standard for making inferences and informing clinical actions [10–12].

1. Characterizing the uncertainty of inferences

The crux of the challenge for EBM resides in how the uncertainty of the evidence is characterized. Even as RCTs are appreciated by many 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 [13]. 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 at baseline. 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 for large samples in a single trial or accumulated over many trials that randomly unbalanced differences are expected to even out, supporting the unbiasedness of RCTs [13].

Standard errors are the main statistic used to characterize the inherent uncertainty in a treatment effect estimate due to potential imbalance between treatment and control groups on covariates related to the outcome. Yet standard errors and their associated confidence intervals are theoretical statistical constructs notoriously prone to misinterpretation [14]. 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 [15].

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 [5,16,17]. Our particular focus is on case replacement: "To change an inference, how many patients from the treatment group who experienced a positive outcome would have to be replaced with hypothetical patients who did not receive a treatment [5]?" A robust inference, for example, would be one in which a large portion of the patients in the treatment group with positive outcomes 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 (eg, $P < 0.01$) might be overturned by the replacement of only a few patients, suggesting some uncertainty in the inference and caution regarding recommended clinical action. While the RIR helps characterize the inherent uncer-

tainty due to imbalance in sampling, it is worth noting that some RCTs suffer from issues that go beyond sampling imbalances, such as noncompliance, attrition, or problems in intervention implementation fidelity [13]. The RIR might have added value in such cases.

In the application to dichotomous patient outcomes, such as mortality, the RIR maps to the existing concept of "fragility" which has been gaining increasing attention in clinical epidemiology [6,15–17] with applications in oncology [18] and pediatrics [19,20]. The Fragility Index indicates how many patients from the treatment group would have to have different outcomes, or experience event switches, to change an inference [17].

Recently, Walter, Thabane, and Briel [10] raised two important critiques of the Fragility Index while extending the fragility framework. First, they raised the concern that the Fragility Index only uses statistical significance as a threshold for making an inference. Second, they noted that the Fragility Index does not account for the relative prevalence of negative outcomes in the data – switching the outcome of a single patient from a positive to negative outcome might be an extreme change if negative outcomes are rare.

One benefit of viewing fragility through the framework of case replacement is that it helps address these limitations [10]. First, the RIR explicitly represents thresholds for inference that include, but are not exclusive to, statistical significance. We will demonstrate this in the Methods and results sections and return to the issue in the Discusison. Second, RIR is sensitive to the underlying rareness of the event. An additional benefit is that case replacement links the concept of fragility to a more general framework for quantifying the robustness of inferences that can be applied to continuous outcomes and research designs other than RCTs [5].

Ultimately, our case replacement framework generates statements such as "The inference would change if xx of the n treatment patients who had positive outcomes were replaced by patients who did not receive a 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 quantify 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 the Fragility Index for dichotomous outcomes. We then demonstrate how RIR and the Fragility Index 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 that occurs in COVID-19 patients. The second is an application to a historical study-by-study emergence of evidence across 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 RIR can be applied to new, individual RCTs as they become available, as well as to the accumulation of evidence across RCTs.

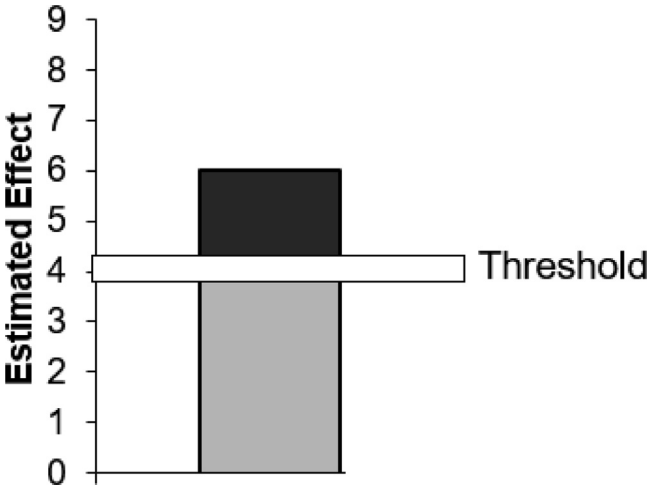

Fig. 1. Estimated effect and threshold for making an inference.

2. Methods: expressing uncertainty in terms of changes in data

We characterize the uncertainty of an inference in terms of the changes to the data necessary to change the inference [5,16,17]. Our approach to quantifying the robustness of an inference in terms of patient experiences is rooted in the tradition of the counterfactual. Frank et al [5] considered replacing cases with counterfactual cases in which those who received the treatment were hypothetically considered to have received the control, and vice versa. These replacements generate potential changes in the outcomes that we focus on here. Because the framework is nonparametric it is general and can be applied across a variety of outcomes and designs.

To better introduce the case replacement framework, consider the idealized example in Fig. 1 which compares an estimated effect to a threshold for making an inference. In Fig. 1, the estimated treatment effect is 6 and 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 [10–12]. But in all cases the threshold pragmatically links the evidence to recommended action [5,10]. 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.

We can use the comparison of the estimated effect to the threshold in Fig. 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, we would expect over many samples that one third of the estimated effect would have to be due to sampling uncertainty or bias to change the inference [5].

Frank et al, [5] and Frank and Min [21] demonstrate that one can interpret the difference between the estimated effect and the threshold in terms of case replacement between the observed sample and a hypothetical population where the treatment effect was zero (eg, there is no difference in mean outcomes between the treatment and control groups). 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. The proportion of the cases that one must replace 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 ($\overline{\hat{\delta}}$) as a function of the observed estimated effect ($\hat{\boldsymbol{\delta}}_o$) and the hypothesized effect in the unobserved replacement data (δ_u) [21–23]. Assuming the proportion of units receiving the treatment is the same in the observed and unobserved data, an expression for $\overline{\hat{\delta}}$ is:

$$\overline{\hat{\delta}} = (1-\pi)\hat{\delta}_o + \pi_u, \tag{1}$$

where π represents the proportion of observed cases replaced by unobserved cases. Therefore $\overline{\hat{\delta}}$ 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 there is no mean difference in outcomes between the treatment and control groups in the unobserved data: $\delta_u = 0$ [5,24,25]. For example, $\delta_u = 0$ holds exactly if the unobserved data are generated from a null hypothesis of zero average treatment effect and there is no sampling variability because there is no covariate imbalance. Assuming $\delta_u = 0$ yields:

$$\overline{\hat{\delta}} = (1-\pi)\hat{\delta}_o \tag{2}$$

Next, set $\overline{\hat{\delta}} = \delta^{\#}$ where $\delta^{\#}$ defines the threshold for making an inference such as an estimate associated with an effect size of specific clinical significance [10,11] or with a *P*-value of 0.05:

$$\overline{\hat{\delta}} = \delta^{\#} = (1-\pi)\hat{\delta}_o \tag{3}$$

Solving for π yields:

$$\pi = 1 - \delta^{\#}/\hat{\delta}_o \tag{4}$$

The expression in (4) allows one to calculate what proportion of the cases (π) in the observed sample would have to be replaced with unobserved cases (for which $\delta_u = 0$) to reduce the combined estimate ($\overline{\hat{\delta}}$) below the threshold ($\delta^{\#}$) for making an inference [5]. For instance, in the simple example in Fig. 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 unobserved cases for which $\delta_u = 0$. This allows us to express uncertainty by conceptualizing how the existing data could be mixed with unobserved 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 [26–28], physical sciences [29–31], and in policy [32,33], but it can also provide unique insight to dichotomous outcomes often used in health (eg, Improved vs. Not Improved; Survived vs. 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 2×2 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 [17]. First, we choose to replace cases from the treatment group who had a *positive outcome*, and define the Robustness of an Inference to Replacement (RIR) as the number of treatment cases that had positive outcomes that would have to be replaced to invalidate an inference. Second, we draw the replacement cases from a hypothetical population with the same rate of positive outcomes as in the control group, which represents the absence of a treatment.[1]

Next, we assume that all cases experiencing the same outcome (eg, improved vs. did not improve) are indistinguishable or exchangeable [34,35]. As a result, when replacing observed treatment cases that experienced positive outcomes, the only clear changes after replacement would be when cases that had positive outcomes were replaced with cases that did not have positive outcomes, and vice versa (ie, situations in which positive outcomes are replaced with positive outcomes cannot be distinguished). In other words, when *event* switches occur. This allows us to represent the hypothetical change in the sample not only in terms of replacement, but also in terms of switching of treatment cases that had positive outcomes to treatment cases that did not have positive outcomes. If one counts switches between events instead of just case replacements from a hypothetical sample, and uses statistical significance from zero as the threshold for inference, the result is the robustness measure called the Fragility Index which has been recently reintroduced in epidemiology [16,17].

The RIR directly extends the Fragility Index in two fundamental ways. First, using RIR as in Fig. 1, any threshold can be used as a basis for inference. We will provide a

[1] Another alternative would have been to use the positive outcome rate as estimated by the whole sample, including treatment and control. This would reflect the null hypothesis of no treatment effect.

Table 1. Robustness of inference for hypothetical treatment and mortality. Example taken from Walter et al [10]. Cells represent number of cases. Fragility Index= number of cases to switch to change the inference; RIR represents the robustness of the inference to replacement.

	Died		Alive	Total
Treatment A	5		90	95
		Fragility Index = 1		
Treatment B	0	←	96 [RIR = 19]	96
Total	5		186	191

demonstration of this in the Results section. Second, the RIR accounts for the likelihood that an outcome for a case will be switched. Consider the example in Table 1, drawn from Walter et al [10]. These results are from a hypothetical experiment where 90/95 patients given Treatment A survived (vs. died), 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 [10] note that the Fragility Index of this inference is 1 – if one alive case in Treatment B were switched to died, 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 $P = 0.05$, the one switch would lead to an inference that there is no difference between Treatments A and B. Walter et al [10] note that the "fallacy" in this [the Fragility Index] argument is that the change from 0 to 1 death in treatment group B may actually be unlikely to occur because of the rarity of death.

Walter et al's [10] 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 Treatment A cases to change the inference that Treatment B was more efficacious than Treatment A. We begin by drawing the replacement cases from a population represented by Treatment A with an estimated mortality rate of 5/95 or 5.3%. Using the 5.3% mortality rate, for every 19 Treatment B Survival cases replaced, we would expect 18 to remain classified as alive, and 1 to be reclassified as died. Therefore, we expect to have to replace 19 Treatment B alive cases to generate the one Treatment died case necessary to change the inference ($P = 0.118$). RIR = 19 out of 96 while the Fragility Index = 1.

Formally, the Fragility Index can be expressed as the expected number of replaced treatment cases with positive outcomes multiplied by the observed probability of negative outcomes in the control group: Fragility Index = RIR x $\hat{p}$, where $\hat{p}$ is the observed probability of a negative outcome in the control group. This implies that RIR = Fragility Index /$\hat{p}$. In the example, 19=1/.053. Thus, RIR is a funciton of $\hat{p}$, addressing Walter et al's [10] critique of the Fragility Index by incorporating the prevalence of positive and negative outcomes in the data.

While our focus is on quantifying the uncertainty of an inference deemed significant or clinically relevant, the RIR is flexible and can also quantify how far below the threshold for making an inference a positive but not significant treatment effect is. In this instance, we can ask how many cases in the treatment group that had *negative* outcomes would have to be replaced with cases from a hypothetical population with the rate of *positive* outcomes in the control group to change the inference.[2] This results in transferring cases from the treatment group with negative outcomes to the treatment group with positive outcomes.[3]

3. Results: using robustness of the inference to replacement (RIR) to express uncertainty

3.1. Inference regarding the effect of hydroxychloroquine (HCQ) on pneumonia

Consider one of the first reports of a randomized trial for the drug hydroxychloroquine (HCQ) on COVID-19 patients [36]. Conducted at the Renmin Hospital of Wuhan University, 31of 62 COVID-19 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 improvement 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. There are two challenges. One is that the small sample is more likely to have baseline

[2] An alternative would be to calculate how many treatment cases would have to switch from negative to positive outcomes to change the inference. From this one could estimate the positive outcome rate in the treatment group if the estimated treatment effect were at the chosen threshold for inference. This rate could then be used to calculate the RIR from the number of switches from treatment negative outcomes to positive outcomes.

[3] Note that the Fragility Index is defined only for results that are positive and statistically significant. For those results that are not statistically significant the Fragility Index is technically undefined [17] 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, RIR=Fragility Index/$\hat{p}$, where = $\hat{p}$ would be based on the positive outcome rate in the control group instead of the negative outcome rate.

Table 2. Robustness of inference for hydroxychloroquine (HCQ) vs conventional treatments on pneumonia. Data from Table 2 of Chen et al [36].[4] Cells represent number of cases. RIR represents the robustness of the inference to replacement with control cases. Fragility = number of cases to switch to change the inference.

	Exacerbated or Unchanged	Improved (Moderate or Significant)	Total
Conventional Treatment	14 {A}	17 {B}	31
Hydroxychloroquine	6 {C} ← Fragility = 1	25 [RIR = 2] {D}	31
Total	20	42	62

imbalance. Another is that the trial was not double-blinded [37]. 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" (by someone who might have known what treatment the patient had received) would have to be replaced to change the inference that HCQ reduced pneumonia?

To answer the question in the preceding paragraph we consider replacing cases from HCQ (treatment) improved with cases with estimated probability of exacerbated or unchanged in the conventional treatment (control) group ($\hat{p} = 14/31 = 0.45$). We then replace cases until the P value $< .05$. Table 2 illustrates the result. If about two of the 25 HCQ improved cases were replaced with cases for which $\hat{p}$=.45, we would expect one of those cases would switch from improved to exacerbated or unchanged (0.45×2), 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 = 2 (out of 25, or 8%), Fragility Index = 1, (calculations conducted using http://konfound-it.com). The low number of replacements needed, whether expressed as RIR or the Fragility Index, highlights the tenuous nature of the inference of an efficacious result despite its statistical significance.

In Fig. 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 [10]. 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 expect to have to replace 2 of the observed treatment improved cases with hypothetical cases with the rate of improved outcomes in the control group ($\hat{p} = .45$) 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 Fig. 2 also shows an RIR of about 11 to reduce the initial probability difference of .26 to 0.10. That is, one would expect to have to replace about 11 of the treatment improved cases (44%) to reduce the estimated effect to 0.10. These 11 replacements would generate five switches – Fragility Index =5. More generally, Fig. 3 represents the RIR with respect to any effect size, including effect sizes that define a minimal important difference [10–12]. [4] This can inform discourse about inferences made for different thresholds depending on the context and the participants.

3.2. Historical example: inference from accumulation of estimates of antihypertensive treatments

The RIR can 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 [38]. These extremely valuable guidelines describe the quality of a body of evidence in terms of aspects of the study design (eg, nonrandom assignment to treatment, nonblinded 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 [15] especially in the context of urgent crises, like the COVID-19 pandemic.

To illustrate what the accumulation of estimated effects of 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]. 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. 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 0.003.

[4] 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.

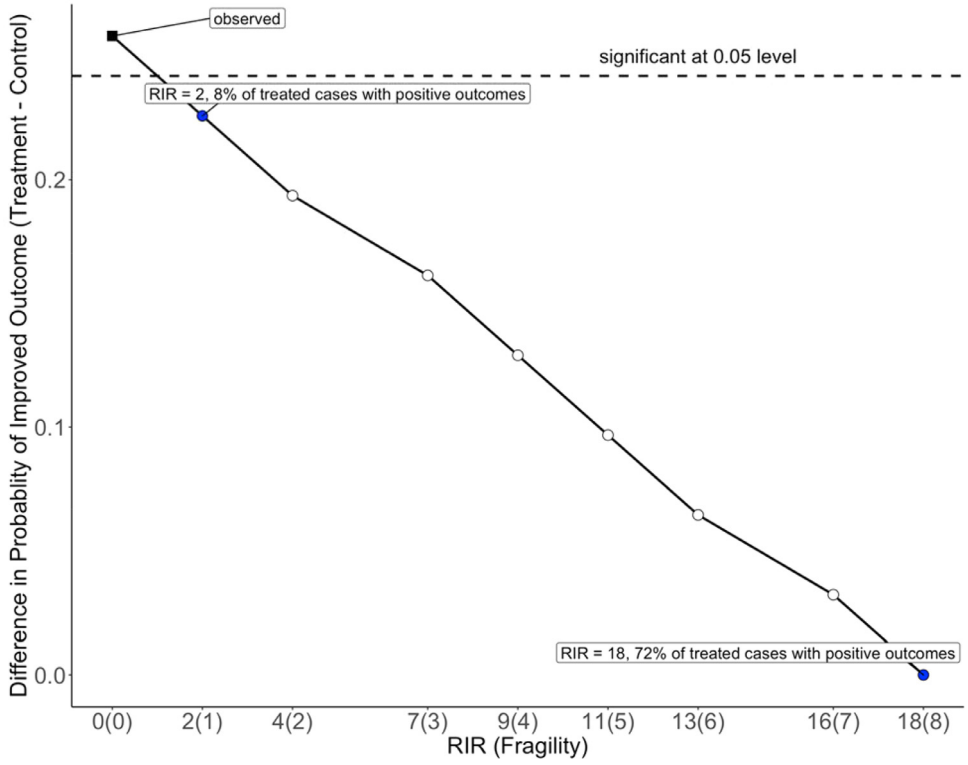

Fig. 2. Difference in probability of improved outcome (treatment – control) after replacing observed cases in HCQ example. Black square, study estimate; dashed line, positive estimate significant at 5% level.

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.

	Control		Treatment						
	Stroke	No stroke	Stroke	No stroke	Total	Decrease in stroke probability for the treatment group	*P* value (Fisher)	RIR	Fragility index
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	39	4
Study 1 + Study 2	21	215	7	224	467	5.9	0.010	34	3

To change the inference in the preceding paragraph, one would expect to have to replace 39 (about 22%) of the treatment "no stroke" cases with cases for which $\hat{p} = 20/194{=}0.10$ based on the control group. These 39 replacements would generate approximately 4 event switches from treatment "no stroke" to treatment "stroke" (Fragility Index = 4). The RIR for the first two studies combined is 34 (about 15%), with a Fragility Index of 3.

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 Fig. 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 RIR increased substantially.[5] Continuous updates to an analogous figure during an emergent health issue 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 scientific evidence.

4. 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]. At critical times, physicians and policymakers are under tremendous pressure to rapidly adapt best practices to protect public health and prevent mortality, often with limited, emerging research. However, relying too heavily on results from single, early

[5] We also calculated RIR for each point in Fig. 3 using fixed effects adjustments from cumulative meta-analysis [8]. The results are similar to those in Fig. 3 and are available from authors.

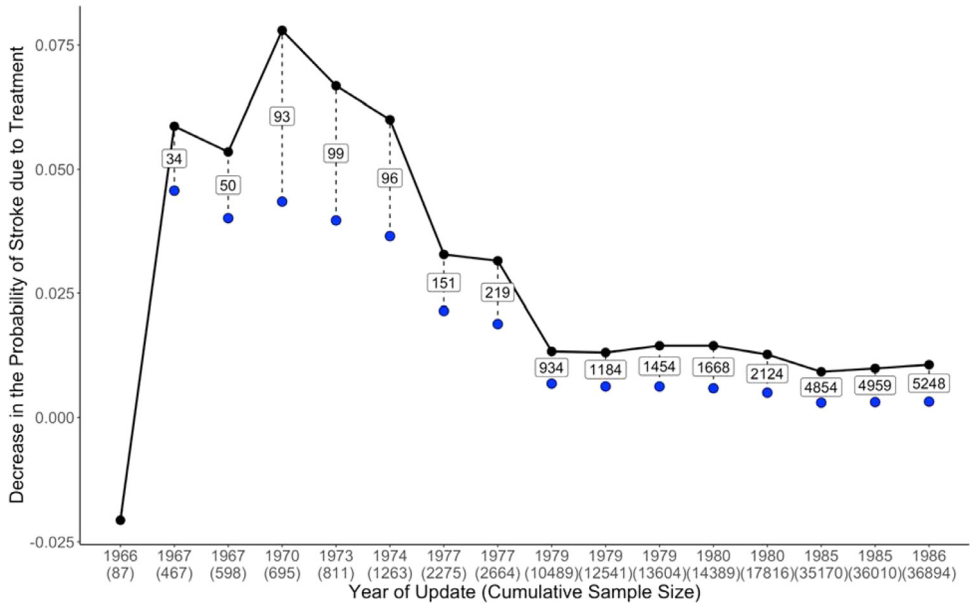

Fig. 3. Robustness of inferences to replacement (RIR) as evidence accumulates: Historical case of antihypertension treatment and stroke. Black dots indicate the size of the estimated treatment effect based on all studies available up to that point in time; blue dots, the effect size just below statistical significance. Boxes label the corresponding RIR.

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 [13]. Moreover, using *P*-values to communicate the inherent uncertainty and robustness of findings to broad audiences is difficult [15]. 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 [10]

In this study, we developed the Robustness of an Inference to Replacement (RIR) which, like the Fragility Index, can help identify and communicate when results are statistically significant but conclusions may not be robust. The RIR is based on a general case replacement framework for quantifying the robustness of causal inferences. The generality of the case replacement framework provides a connection to other research designs and statistical models, including those that use continuous outcomes. It also helps address known limitations of the Fragility Index by accounting for the observed rates of outcomes.

While some might argue that the difficulty with *P*-values is indicative of a deeper problem that requires the wholesale replacement of the null hypothesis significance testing paradigm [8,9], *P*-values remain central to discourse about scientific findings [16,40]. Use of robustness analysis approaches such as fragility and case replacement can help mitigate the disadvantages of *P*-values. Instead of relying on an understanding of sampling distributions which may be unfamiliar to many, fragility and case replacement characterize the robustness of conclusions in relatable and tangible terms. For example, the RIR supports statements such as "The inference would change if xx of the n treatment patients who experienced positive outcomes were replaced by hypothetical patients who did not receive a treatment." This can inform debates about the bases for conclusions and help an inclusive set of stakeholders interpret the potential threat of sources of bias.

One might raise the question about when the RIR is large enough to interpret an inference as robust. First, we encourage a comparison of the RIR with both the number of cases in the source cell (eg, treated cases that had positive outcomes) as well as the overall sample size. Second, given the generality of the RIR and the gap it and the Fragility Index fill in the characterization of uncertainty, we anticipate they will be employed extensively in the coming years. As this occurs, researchers and clinicians will develop norms about what values of RIR and the Fragility Index represent robust findings [5].

At its core, we emphasize that the RIR informs the conversation about the robustness of an inference, helping a broader community weigh in on the link between evidence and practice. In particular, the RIR provides more information that can convey confidence in findings more clearly and emphatically than the language of "highly statistically" significant. Compare the use of the RIR to the use of the *P*-value alone depicted in Fig. 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 0.01 and *P* of 0.001 on the horizontal axis, it is more direct to understand that the inference would change if 25 vs. 75+ of the cases were replaced on the vertical axis.

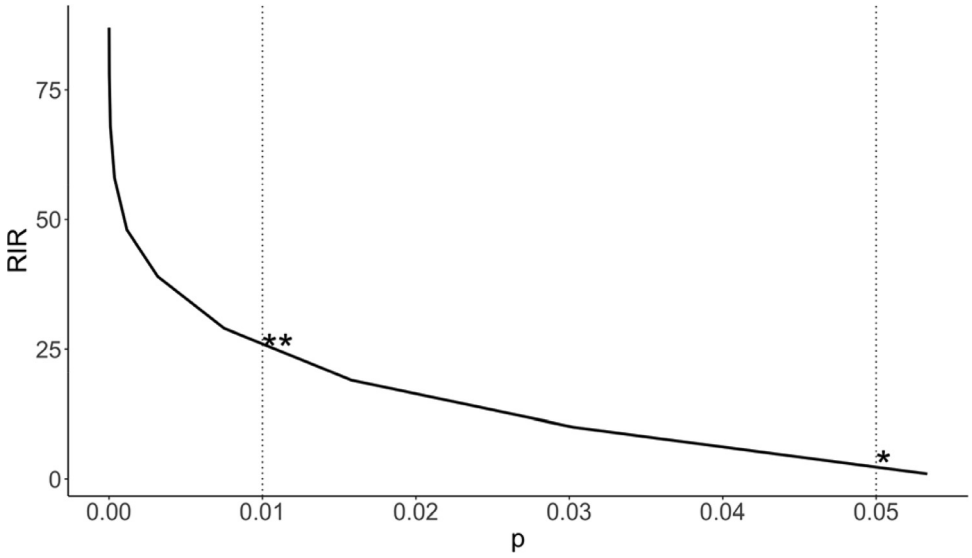

Fig. 4. Robustness of inference to replacement (RIR) vs *P*-value. The curve shows the functional relationship between the RIR and the *P*-value. Baseline table has 380 cases evenly divided between treatment and control with odds ratio favoring the treatment of 1.22. The steep slope on the left indicates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

While not the focus of this paper, a case replacement approach to robustness can also provide benefit when concerns go beyond *P*-values into deeper concerns about bias [5]. For example, the RIR could be applied to observational studies which are being implemented during the early stages of the COVID pandemic [41]. In this scenario, the replacement cases are conceptualized as quantifying the robustness of an inference drawn from a counterfactual comparison generated by nonexperimental techniques. 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 [42]. But we emphasize the RIR alone is not a substitute for assessing the comprehensive methodologic strength of a study – instead, it is a helpful tool for understanding and communicating the stability or robustness of any given conclusion based on data in the context of the study design.

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.

CRediT authorship contribution statement

Kenneth A. Frank: Conceptualization, Formal analysis, Methodology, Supervision, Visualization, Writing - original draft, Writing - review & editing. **Qinyun Lin:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - review & editing. **Spiro Maroulis:** Conceptualization, Methodology, Visualization, Writing - review & editing. **Anna S. Mueller:** Conceptualization, Writing - review & editing. **Ran Xu:** Conceptualization, Methodology, Writing - review & editing. **Joshua M. Rosenberg:** Conceptualization, Software, Writing - review & editing. **Christopher S. Hayter:** Writing - review & editing. **Ramy A. Mahmoud:** Writing - review & editing. **Marynia Kolak:** Writing - review & editing. **Thomas Dietz:** Writing - review & editing. **Lixin Zhang:** Writing - review & editing.

References

[1] Djulbegovic B, Guyatt G. Evidence-based medicine in times of crisis. J Clin Epidemiol 2020. doi: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. doi: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] London AJ, Kimmelman J. Against pandemic research exceptionalism. Science 2020;368:476–7. doi:10.1126/science.abc1731.
[5] 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;35:437–60.
[6] 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. doi:10.1016/j.socscimed.2013.01.034.
[7] 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. doi:10.1136/bmjopen-2015-008155.

[8] Rothman KJ. Special article: writing for epidemiology. Epidemiology 1998;9:333–7.
[9] 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. doi:10.1056/NEJMe1906559.
[10] Walter SD, Thabane L, Briel M. The fragility of trial results involves more than statistical significance alone. J Clin Epidemiol 2020;124:34–41. doi:10.1016/j.jclinepi.2020.02.011.
[11] 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. doi:10.1016/j.jclinepi.2016.11.016.
[12] 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. doi:10.1016/j.jclinepi.2007.03.012.
[13] Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Soc Sci Med 2018;210:2–21. doi:10.1016/j.socscimed.2017.12.005.
[14] Pocock SJ, Ware JH. Translating statistical findings into plain English. The Lancet 2009;373:1926–8. doi:10.1016/S0140-6736(09)60499-2.
[15] 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. doi:10.1016/j.jclinepi.2019.03.012.
[16] 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. doi:10.1016/0895-4356(90)90186-S.
[17] 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. doi:10.1016/j.jclinepi.2013.10.019.
[18] 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. doi:10.5435/JAAOSGlobal-D-19-00152.
[19] 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. doi:10.1016/j.jpurol.2020.06.015.
[20] Tignanelli CJ, Napolitano LM. The fragility index in randomized clinical trials as a means of optimizing patient care. JAMA Surg 2019;154:74–9. doi:10.1001/jamasurg.2018.4318.
[21] Frank K, Min K-S. Indices of robustness for sample representation. Sociol Methodol 2007;37:349–92. doi:10.1111/j.1467-9531.2007.00186.x.
[22] Cronbach LJ, Shapiro K. Designing evaluations of educational and social programs /. San Francisco: Jossey-Bass; 1982.
[23] Fisher RA. Statistical methods for research workers. Darien, Conn: Hafner Pub Co; 1970.
[24] Cinelli C, Hazlett C. Making sense of sensitivity: extending omitted variable bias. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2020;82(1):39–67.
[25] VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med 2017;167:268–74. doi:10.7326/M16-2607.
[26] Asensio OI, Delmas MA. Nonprice incentives and energy conservation. Proc Natl Acad Sci 2015;112:E510. doi:10.1073/pnas.1401880112.
[27] Dietz T. Altruism, self-interest, and energy consumption. Proc Natl Acad Sci 2015;112:1654. doi:10.1073/pnas.1423686112.
[28] 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. doi:10.1016/j.adiac.2016.09.001.
[29] Carrico AR, Vandenbergh MP, Stern PC, Dietz T. US climate policy needs behavioural science. Nat Clim Change 2015;5:177–9. doi:10.1038/nclimate2518.
[30] 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. doi:10.1007/s13365-020-00859-8.
[31] Xu R. Statistical methods for the estimation of contagion effects in human disease and health networks. Comput Struct Biotechnol J 2020;18:1754–60. doi:10.1016/j.csbj.2020.06.027.
[32] 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. doi:10.1002/pam.22074.
[33] Frank KA, Penuel WR, Krause A. What is a "good" social network for policy implementation? The flow of know-how for organizational change. J Policy Anal Manage 2015;34:378–402.
[34] Bernardo JM. The Concept of Exchangeability and its Applications n.d.:7.
[35] 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. doi:10.1007/978-1-4612-0919-5_10.
[36] 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. doi:10.1101/2020.03.22.20040758.
[37] Pacheco RL, Riera R. Hydroxychloroquine and chloroquine for COVID-19 infection. Rapid systematic review. J Evid-Based Healthc 2020;2. doi:10.17267/2675-021Xevidence.v2i1.2843.
[38] 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. doi:10.1016/j.jclinepi.2010.07.015.
[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. doi:10.1016/0140-6736(90)90944-z.
[40] Goodman SN. Of P-values and Bayes: a modest proposal. Epidemiology 2001;12:295–7.
[41] 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. doi:10.1101/2020.03.24.20042937.
[42] Oakes JM. The tribulations of trials: a commentary on Deaton and Cartwright. Soc Sci Med 2018;210:57–9. doi:10.1016/j.socscimed.2018.04.026.