Quantifying Replicability and Consistency in Systematic Reviews

Iman Jaljuli\textsuperscript{a}, Yoav Benjamini\textsuperscript{b}, Liat Shenhav\textsuperscript{b}, Orestis A. Panagiotou\textsuperscript{c}, and Ruth Heller\textsuperscript{a}

\textsuperscript{a}Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel; \textsuperscript{b}Center for Studies in Physics and Biology, Rockefeller University, New York, NY; \textsuperscript{c}Department of Health Services, Policy & Practice, Brown University, Providence, RI

ABSTRACT

Systematic reviews and meta-analyses are important tools for synthesizing evidence from multiple studies. They serve to increase power and improve precision, in the same way that large studies can do, but also to establish the consistency of effects and replicability of results across studies. In this work we propose statistical tools to quantify replicability of effect signs (or directions) and their consistency. We suggest that these tools accompany the fixed-effect or random-effects meta-analysis, and we show that they convey important information for the assessment of the intervention under investigation. We motivate and demonstrate our approach and its implications by examples from systematic reviews from the Cochrane Library. Our tools make no assumptions on the distribution of the true effect sizes, so their inferential guarantees continue to hold even if the assumptions of the fixed-effect or random-effects models do not hold. We also develop a version of this tool under the fixed-effect assumption for cases where it is crucial and justified.

1. Introduction

In systematic reviews and meta-analyses, several studies that examine the same question are synthesized together. Viewing all the available information is extremely valuable for practitioners in the health sciences. Notable examples of systematic reviews are those conducted by the Cochrane Collaboration to estimate the effects of healthcare interventions (Higgins and Green 2011). Deriving conclusions about the overall health benefits or harms of a drug or an intervention from an ensemble of studies can be difficult, since the studies are never exactly the same and differences between studies may affect the inference.

There are many reasons to perform a meta-analysis, most notably to increase power and improve precision. Additional reasons are related to answering questions that cannot be addressed by individual studies. Such questions are often due to the fact that, as noted in the Cochrane Handbook for Systematic Reviews of Interventions (Deeks et al. 2019, sec. 9.1.3): “Primary studies often involve a specific type of patient and explicitly defined interventions. A selection of studies in which these characteristics differ can allow investigation of the consistency of effect and, if relevant, allow reasons for differences in effect estimates to be investigated”; and also in order “To settle controversies arising from apparently conflicting studies or to generate new hypotheses. Statistical analysis of findings allows the degree of conflict to be assessed formally, and reasons for different results to be explored and quantified.”

These last two reasons are directly related to the growing concerns in recent years about lack of replicability of results in medical research and in science at large (Nuzzo 2014; McNutt 2014; Collins and Tabak 2014; Gibson 2021). The discussions are about the issues that may hurt the published research of a study, such as, publication bias and unreported exploratory steps. Most of the contemplated solutions involve the single stand-alone study: its design, preregistration, conduct, analysis and report, all in a reproducible way.

Replicability of a result is a concept not well established. Relying on Fisher (1935), a result has been replicated if the \( p \)-value is lower than some small threshold in both the original and the replicating studies (this was Fisher’s motivation for introducing a threshold for the \( p \)-value). This definition was used by “The Psychology Reproducibility Project” (Open Science Collaboration 2015) and is still the most acceptable criterion to date. One negative aspect of this approach is the emergence of the replicability effort as a single one-shot effort, ending with a clear conclusion of “replicated” versus “not-replicated.” For this reason, a major concern in the design of replication studies is to guarantee a large enough sample size, in order to assure sufficient power for making such a conclusion, see, for example the report by the National Academies of Sciences, Engineering, and Medicine (2019).

Systematic reviews and meta-analyses offer a natural approach to assessing replicability. The studies are conducted by independent groups of researchers that try to follow similar protocols, but local deviations are unavoidable. Obviously, the different studies enlist different subjects, but furthermore they are often conducted on different populations from which subjects are drawn. The individual studies are also not necessarily of sufficient power, so that if they do not get a statistically significant result it cannot be concluded that there is a replica-
bility problem. Nevertheless, the effects from a number of small studies can be combined to assess whether the intervention effect has been replicated.

A rather simpler approach would be mere counting of studies with significant results in each direction, but this is only descriptive of the data and reflective of each study’s sample size and error. Meta-analysis looks at the entire set of studies as a single body of evidence. So does our quantification method, that combines evidence across the entire set of studies to add an inferential statement about the body of evidence as a whole. Our method can have more power than counting, since it can yield the replicability statement that at least a certain proportion of the studies have an effect (without identification of which studies these are) even when most (or all) studies are nonsignificant. This is based on pooling the evidence in the same direction from all the studies: if 10 studies all have (one-sided) \( p \)-values between 0.05 and 0.2, the evidence that there is an effect in at least two studies is strong (i.e., below the 0.05 level of significance). Our method also guards against stating replicability when the evidence hinges on one (or a small) portion of the studies: if two out of 20 studies have \( p \)-value close to 0.05, and all other \( p \)-values are randomly spread between 0.1 and 1, there is no evidence to establish that there is an effect in more than a single study (at the 0.05 level).

Unlike Fisher’s case of two studies, where the two statistically significant results imply that there is an effect in the two studies, for three or more studies, showing that there is an effect in a number of studies does not amount to counting the number of statistically significant studies. An inference approach that combines subsets of the studies’ \( p \)-values has been offered, and replicability analysis tools that make use of it in a structured way were developed in previous work by Benjamini and Heller (2008), Benjamini, Heller, and Yekutieli (2009), Heller (2011), and Wang and Owen (2019).

The last reason for conducting a meta-analysis relates to identifying not only lack of replicability but also inconsistency of results, and then possibly explaining their sources. Traditionally, in meta-analysis, such questions are addressed by the many methods to assess effect heterogeneity (Higgins and Thompson 2002; Riley, Higgins, and Deeks 2011; Aanagiotou and Trikalinos 2015; Borenstein et al. 2017; Deeks et al. 2019). A widely used measure is the \( I^2 \) which quantifies the percentage of variation in treatment effects that is due to study differences (heterogeneity) rather than chance (Higgins and Thompson 2002; Higgins, Thompson, and Spiegelhalter 2009). \( I^2 \) is commonly used for quantifying the degree of between-study heterogeneity. Because \( I^2 \) relates the magnitude of heterogeneity variance to the overall variance, a “large” value may be due to large heterogeneity but also due to large sample sizes (within-study variability); thus, an alternative and often more informative measure is the variance of treatment effects across studies, that is, \( \tau^2 \) (Turner et al. 2015).

However, the presence of effect heterogeneity does not convey any information as to whether the treatment effects are consistently in favor or against the intervention under study. For example, treatment effects in a meta-analysis may be heterogeneous but also consistently in favor of the intervention: they may all be positive (or all negative) but their confidence intervals (CIs) may not at all overlap. To demonstrate it, consider a heterogeneous case, now add a large enough constant to all of the studies’ effects. Heterogeneous effects may also exhibit inconsistency whereby the intervention is beneficial in some studies and harmful in others. This inconsistency, termed qualitative interaction (Gail and Simon 1985) may be of particular concern. Assuming the RE model, it is more likely to occur when the heterogeneity is large and the effect is small, but also for a fairly small true magnitude of heterogeneity when the number of studies considered is large.

In this work we argue that, in addition to reporting the estimated effect heterogeneity of a meta-analysis, it is important to quantify the consistency of the evidence in favor or against the intervention. We propose an inference method that quantifies replicability and consistency of treatment effects and provides a lower confidence bound on the number of studies with a treatment effect in the same direction. It allows meta-analysts to make statements such as “with 95% confidence, out of the five meta-analyzed studies, at least two have a positive effect” or “at least one study has a negative effect and at least one study has a positive effect,” see Section 4 for specific examples.

For a concrete example, consider the meta-analysis conducted by Hughes et al. (2020), studying the effect of antihypertensive agents on dementia or cognitive impairment. Combining the results of 12 randomized controlled trials (corresponding to 13 treatment comparisons) of blood lowering treatments versus control, they conclude that “Blood pressure lowering with antihypertensive agents compared with control was significantly associated with a reduction in cognitive decline” based on a pooled estimated odds ratio of 0.93 with 95% confidence interval ranging from 0.88 to 0.98. Using the tools we offer here, the analysis reported in Section 2 enables us to conclude that result replicability is not established, that is, there are not at least two comparisons finding that antihypertensives are better than placebo in lowering the risk of dementia or cognitive impairment; we will return to this example later in Section 7, Figure 9.

Meta-analyses are mainly performed to estimate an overall treatment effect from a group of studies that have been deemed similar enough (Borenstein et al. 2009). Synthesis can be performed under the fixed-effect (FE) or the random-effects (RE) models, and the approach we offer here is a useful inferential addition to both. The models differ in their assumption about effects heterogeneity. The FE model assumes there is a common effect across studies. If this assumption is true, the test of the null hypothesis of no treatment effect may be far better powered than the test in each of the individual studies. However, if in fact there is heterogeneity in effect direction, then the RE model is recommended since it assumes the treatment effects are different, though they are also assumed to be independent identically distributed samples from a distribution (usually Gaussian). Some researchers argue that the FE assumptions are implausible most of the time, and thus, suggest to always use the RE model (Higgins and Green 2011; Aanagiotou and Trikalinos 2015). Others make a choice based on clinical knowledge around the patient, intervention, and outcome attributes in meta-analyzed studies. The Cochrane Handbook for Systematic Reviews of Interventions (Deeks et al. 2019, sec 9.5.4) cautions against choosing a RE over FE meta-analysis based on a statistical test for heterogeneity. It is relevant to
note that the RE model was also recommended by Kafkafi et al. (2005) as the tool for assessing replicability across laboratories in animal phenotyping experiments. A major criticism of the RE model is that validating the distributional assumptions of the treatment effects is difficult (Deeks et al. 2019) and it is not generally possible to distinguish whether heterogeneity results from clinical, methodological, or other variability.

We develop our tool in two forms: with and without the common effect assumption (of FE meta-analysis). We recommend using the assumption free form regardless of the meta-analysis performed, unless the common effect assumption is justified. In that case we recommend using the form developed under these assumptions.

In Section 2, we explain our methodology for quantifying the evidence toward consistency in effect direction that tailors the existing general replicability analysis tools to meta-analyses in systematic reviews. In Section 3 we demonstrate via simulations that the available meta-analysis tools do not provide the insights into replicability that our proposed analysis does. In Section 4 we demonstrate how evaluation of replicability contributes to the assessment of the intervention effects in case studies from the Cochrane library. In Section 5 we summarize our evaluation of the extent of consistency (and inconsistency) in effect direction in the entire breast cancer domain of the Cochrane library. The special case of the common-effect assumption is discussed Section 6; in Section 7 we discuss alternative tools and in Section 8 we discuss the interpretations of high heterogeneity and conclude with some final remarks.

2. Replicability Analysis

We define replicability in terms of the true but unknown parameters and suggest a method for establishing replicability based on summary statistics. We do not assume a true common effect or a distribution of true effects in this section. In Section 6 we develop a version of this method under the common effect assumption of FE meta-analysis, using weighted means.

2.1. Replicability Framework

Let \( n \) be the number of studies available for meta-analysis, \( \theta = (\theta_1, \ldots, \theta_n) \) the unknown treatment effects vector. For study \( i \in \{1, \ldots, n\} \) the null hypothesis of no treatment effect is, without loss of generality, \( H_i : \theta_i = 0 \). For convenience and practicality, assume no effect is zero effect. The effect is decreased if \( \theta_i < 0 \), the effect is increased if \( \theta_i > 0 \). We use the terms decreased and increased in the sense of being smaller and greater than the value for no effect, respectively.

For a group of \( n \) studies and \( u \in \{2, \ldots, n\} \), \( u/n \) replicability of an increased effect is when at least \( u \) studies have an increased true effect, that is, if the true effects vector \( \theta \) is in the set

\[
A^u/n(R) = \left\{ \theta : \sum_{i=1}^{n} I(\theta_i > 0) \geq u \right\},
\]

where \( I(\cdot) \) is the indicator function. If \( \theta \in A^u/n(R) \), at most \( n - u \) studies have nonincreased effects; otherwise, at least \( n - u + 1 \) studies have nonincreased effects. Similarly we define \( u/n \) replicability of a decreasing effect if \( \theta \in A^u/n(L) = \left\{ \theta : \sum_{i=1}^{n} I(\theta_i < 0) \geq u \right\} \).

Thus, we have \( u/n \) replicability if at least \( u \) studies have an effect in the same direction,

\[
\theta \in A^u/n(R) \cup A^u/n(L).
\]

We define inconsistency if for \( u \neq 0 \), the \( u/n \) replicability is met for both an increased and a decreased effect. In particular, inconsistency is established if \( 1/n \) replicability is met for both the increased and the decreased effect directions.

Replicability and inconsistency rise to the argument only when there are at least two studies with nonzero effects, namely, \( \theta \in A^2/n(R) \cup A^2/n(L) \) or \( \theta \in A^{1/n}(R) \cap A^{1/n}(L) \). Otherwise, inconsistency of effects cannot be argued since we have no more than one effect declared as nonzero.

We have consistency if at least two studies show an effect in the same direction, but no studies show an effect in the opposite direction, that is, \( \theta \) is in the set

\[
\left\{ A^{2/n}(R) \setminus A^{1/n}(L) \right\} \cup \left\{ A^{2/n}(L) \setminus A^{1/n}(R) \right\}
\]

For establishing \( u/n \) replicability, counting estimates that are statistically significant in each direction does not provide valid inference. In Sections 2.2 and 2.3 we show how to establish \( u/n \) replicability in each direction and in overall, respectively.

2.2. Establishing \( u/n \) Replicability of Effect Direction

Let \( \hat{\theta}_i, \hat{\Delta}_i \) be the estimated effect size and its standard error for study \( i \). For testing \( H_i : \theta_i = 0 \), the test statistic is \( \hat{\theta}_i/\hat{\Delta}_i \) and the \( p \)-values for the left- and right-sided alternative hypotheses are \( p_i^L, p_i^R \), respectively.

In order to establish \( u/n \) replicability of increased effects using the aforementioned summary statistics, we test the composite null hypothesis that at most \( u - 1 \) studies have positive effects:

\[
H^{u/n}(R) : \theta \notin A^{u/n}(R),
\]

Let \( \Pi(u) \) denote the set of all \( n - (u - 1) \)-tuples from \( \{1, \ldots, n\} \), and let \( i \) be a tuple in \( \Pi(u) \). Testing the composite null hypothesis is possible by the key observation that it is rejected if and only if for every \( i := \{i_1, \ldots, i_{n-u+1}\} \in \Pi(u) \), the corresponding intersection hypothesis

\[
H_i^R : \theta_{i_1} \leq 0, \ldots, \theta_{i_{n-u+1}} \leq 0
\]

is rejected (Benjamini, Heller, and Yekutieli 2009). For example, for \( u = 2 \), at least two of the \( n \) studies have increased effects if and only if for each of the \( n \) subsets of \( n - 1 \) studies, at least one study has an increased effect.

A level \( \alpha \) test rejects \( H^{u/n}(R) \) if for every \( i \in \Pi(u) \), the intersection hypothesis \( H_i^R \) is rejected in favor of the alternative that there exists a \( j \in i \) with \( \theta_j > 0 \) by a \( \alpha \) level test. For each intersection hypothesis many statistical tests are available that combine the individual \( p \)-values or test statistics (Loughin 2004; Futschik, Taus, and Zehetmayer 2019). The preferred test depends on the (unknown) alternative, and there is no single test that dominates all others. Fisher's combining method aggregates the study \( p \)-values \( p_i^R, \ldots, p_{i_{n-u+1}}^R \) with the combining
function \( f(p_1^R, \ldots, p_{n-u+1}^R) = -2 \sum_{j=1}^{n-u+1} \log p_j^R \) (Fisher 1946; Littell and Folks 1971).

Fisher’s combining method is popular in various application fields (e.g., genomic research, education, social sciences) since it has been shown to have excellent power properties (Owen 2009). It is rarely used in meta-analyses of randomized clinical trials, where the focus is on effect sizes. We shall consider the following extension of this combining method, which is useful if the treatment effects are suspected to have mixed signs. In such a case, a potentially more powerful test is based on aggregation of the p-values that are below a predefined threshold \( t \) (Zaykin et al. 2002), and the null distribution is adjusted accordingly. Our test statistic for the intersection hypothesis \( H^R_k \) is therefore,

\[
C^R_k(t) = -2 \sum_{j\leq t} \log \left( \frac{p_j^R}{p_j^R|p_j^R \leq t} \right).
\]

The null distribution has a simple form. Using the computation method in Hsu, Small, and Rosenbaum (2013), the p-value for the intersection hypothesis is

\[
p^{R}(i) = \sum_{k=1}^{n-u+1} p_{n-u+1+\alpha}(k) \times \left[ 1 - F_k \left( -\log \left( \frac{C^R_k(i)}{\alpha/2} \right) \right) \right],
\]

where \( F_k(\cdot) \) is the cumulative gamma distribution with scale parameter equal to one and shape parameter \( k \), and \( p_{n-u+1+\alpha}(\cdot) \) is the cumulative Binomial distribution with \( n - u + 1 \) trials and probability of success \( \alpha \). The p-value for \( H^{u/n}(R) \) is

\[
r^R(u) = \max_{i \in \Omega(u)} p^{R}(i).
\]

Since \( C^R_k(i) \) is monotone in the p-values, \( r^R(u) \) can be computed efficiently in \( O(n \log n) \) computations by sorting the right-sided p-values. Then \( r^R(u) \) will be the p-value of the intersection hypothesis with indices corresponding to the \( n - u + 1 \) largest (i.e., least significant) p-values. Formally, denoting the sorted right-sided p-values by \( p_{1}^{R}(1) \leq \cdots \leq p_{n}^{R}(n) \),

\[
r^R(u) = \sum_{k=1}^{n-u+1} p_{n-u+1+\alpha}(k) \times \left[ 1 - F_k \left( -\log \left( \frac{C^R_k(i)}{\alpha/2} \right) \right) \right],
\]

where \( C^R_k(u) = -2 \sum_{j=\alpha}^{n} \log \left( \frac{p_j^R}{p_j^R|p_j^R \leq \alpha} \right) \).

The above steps can be straightforwardly adjusted in order to compute the p-value for \( H^{u/n}(L) \), denoted by \( r^L(u) \).

Based on our own investigations we suggest using thresholds that are at least twice \( \alpha \), that is, \( t = 0.1 \) or \( t = 0.2 \), see Figure 1 for details. In the analysis therein we use \( t = 0.2 \).

**Remark 2.1.** There are many one-sided composite tests that combine test statistics from multiple sources, for example, one-sided sum test (Pocock, Geller, and Tsiatis 1987; Frick 1994), approximate likelihood ratio test (Tang, Gnecco, and Geller 1989; Pollmann 1996), Max test (Tarone 1981). For each of these tests, as well as our selected test, there exists a data generation for which the test is optimal. We favor Zaykin’s combining method, since it handles efficiently p-values that are stochastically larger than uniform, a setting which may arise if the study effects have mixed signs. However, the r-value and confidence lower bounds described below can be applied using any valid one-sided composite test, so researchers can choose their favorite intersection test instead.

**2.3. The r-Value**

The p-value of the test with the minimal replicability requirement, that is, with \( u = 2 \), is simply referred to as the r-value. The null hypothesis \( H^{2/n} \) is true if at most one study has an effect in either direction. An evidence is replicable if the r-value is below the nominal level for the Type I error, since then the conclusion is that at least two studies have an effect in the same direction.

Formally, the p-value for the composite null hypothesis \( H^{u/n} : H^{u/n}(R) \cap H^{u/n}(L) \) is

\[
r(u) = \min\{r^R(u), r^L(u)\}.
\]

Testing the composite null \( H^{u/n} \) for a given \( u \) allows us to make inference that is robust to \( u - 1 \) possibly outlier studies (in the sense that at least one of the remaining studies has a true effect), without presupposing them.

**Remark 2.2.** With \( u = 1 \), this test reduces to Pearson’s test described in Owen (2009), which is useful for powerful identification of effects that are consistently decreasing or consistently increasing across the \( n \) studies. This test has greater power than a test based on Fisher’s combining method using two-sided p-values when direction of the treatment effect is consistent across studies, while not requiring us to know the common direction.

**2.4. Confidence Lower Bounds for Replicability of Effect Direction**

In order to establish, with \( 1 - \alpha \) confidence, lower bounds on the number of studies with decreased effects and the number with increased effects, we test in order \( H^{u/n}(L) \) and \( H^{u/n}(R) \) for increasing values of \( u \) (Heller 2011). Let \( u_{\text{max}}^{L} \) be the maximal value of \( u \) for which \( H^{u/n}(L) \) was rejected at significance level \( \alpha/2 \). Therefore, if \( r^L(1) > \alpha/2 \) then \( u_{\text{max}}^{L} = 0 \) (no evidence of a decreased effect), and if \( r^L(1) \leq \alpha/2 \):

\[
u_{\text{max}}^{L} = \arg\max_{u \in \{1, \ldots, n\}} \{ r^L(k) \leq \alpha/2, k = 1, \ldots, u \}.
\]

Then we can conclude with \( 1 - \alpha/2 \) confidence, that there are at least \( u_{\text{max}}^{L} \) studies with decreased effect. Similarly, we compute \( u_{\text{max}}^{R} \) (with \( R \) instead of \( L \) in the above expression). Therefore, with \( 1 - \alpha \) confidence, there are at least \( u_{\text{max}}^{R} \) studies with decreased effects and \( u_{\text{max}}^{R} \) studies with increased effects. In other words, these are lower bounds to the number of studies with decreased and increased effects, respectively.

**2.5. Enhancing the Meta-analysis Report with Replicability Analysis Findings**

By adding the r-value we provide an objective measure of the confidence that the finding is not driven by a single study. A result that r-value \( \leq 0.05 \) is useful for strengthening the scientific finding, by concluding that at least two studies have a treatment effect in the same direction. However, a result that
3. Simulations

Simulation studies are carried out to examine the power of the aforementioned, assumption-free, tests for establishing replicability of effect in various settings of interest.

3.1. Simulation Settings

For study \( i \in \{1, \ldots, n\} \), the estimated effect size \( \hat{\theta}_i \) is sampled from the normal distribution with mean \( \bar{\theta} \) and standard error \( SE_\theta = \sqrt{1/n_{CI} + 1/(n_{T})} \) with control and treatment group sizes \( n_{CI} \) and \( n_{T} \), respectively. We examined a wide range of values for \( \theta \), \( n \) and \( (n_{CI}, n_{T}) \).

Since the qualitative conclusions are similar for the various values of \( n, n_{CI}, n_{T} \), we display in this section results for \( n = 8 \), with equal group sizes \( n_{CI} = n_{T} = 25 \). Simulations for \( n = 4, 20 \) and other group sizes are shown in Appendix B, supplementary materials. For the effects vector \( \theta \), we considered three different settings: (a) the fixed effects setting, in which the effects are fixed to the same value for each data generation, namely, a common effects value for all the nonnull studies; (b) the random effects setting, in which the effects distribution is the one assumed in the RE model, so for each data generation, iid random samples \( \theta \) are drawn from \( N(\mu, \tau^2) \) where \( \tau^2 = 1 \) unless specified otherwise and (c) the mixture setting, where the effects are a mixture of zero and nonzero signals either consistent(homogeneous)
heterogeneous) or inconsistent. Meta-analyses were performed via the R package meta where inverse variance weighting is used for pooling. Heterogeneity (\( \tau^2 \)) was estimated via the DerSimonian–Laird method. The meta-analysis confidence intervals assume normality of the estimated effect (for the FE meta-analysis) and in addition normality of the random effects (for the RE meta-analysis). The number of iterations was set at \( 4 \times 10^4 \) where the simulation error is less than 0.0025.
3.2. Simulation Results

We find that the power of the replicability analysis can be much higher than that of the meta-analysis in many realistic settings. Figure 1 shows four settings of mixed signals, for eight studies: an increased treatment effect only in a single study (furthest left column), the same effect in only two studies, three studies with effect where only one has a decreased effect, and three consistent studies with increased effects but heterogeneous. Truncation threshold $t$ can be any value between 0 < $t$ ≤ 1, but in systematic reviews, we see that values below the nominal level (i.e., $t \leq 0.05$) or $t = 1$ (i.e., no truncation) should not be considered, since they lead to loss in power. With $t = 1$, studies with too weak signal (or none at all) will overshadow the rest, and with $t = 0.05$ the power of the pooled result will then rely on the power of the statistically significant studies. Therefore we see it is important to limit the truncation threshold in a way such that studies that have signal but low power (i.e., $p$-value over but close to 0.05) be detected.

We examined three truncation values for the test statistic, and found that for all values of $u$, as well as for detecting inconsistency, the best truncation value is at 0.2. Its advantage over $t = 0.1$, $t = 0.5$ is especially large in the setting with mixed and heterogeneous signs: 0.1 consistently has less or almost equal power, but the Type I error of $t = 0.5$ for $u = 3$ is exaggerated when $H^{3/n}$ is false (columns 1–3, row 3). Therefore, from henceforth we only consider truncation at $t = 0.2$.

As expected, the power to reject $H^{2/n}$ when it is false is lower than that of rejecting $H^{1/n}$, but it increases to one as the signal strengthens in two studies in the same direction. The power to detect inconsistency increases to one as well when the signals are mixed. Arguably, the meta-analysis that will be carried out for these FE data generations is a RE meta-analysis because of the nonnegligible heterogeneity. The RE meta-analysis rejects the null hypothesis of no overall treatment effect at most 5% of the time when the treatment effect is present in only one study, thus, providing better protection against the danger of concluding there is a treatment effect based on a single study than the FE meta-analysis. The power to detect an overall signal with the RE model is also very low in the other two settings: at most 15% when a treatment effect in the same direction is present in two studies; at most 5% when the treatment effect is inconsistent. However, in these two settings (middle and right columns) the replicability can be established with power increasing to one as the absolute value of the treatment effects increases.

Figure 2 shows FE settings where the non-null treatment effect is common, and the number of studies with this common effect increases from zero to eight. Studies without this common effect are set at zero. The rejection rate of the replicability null hypothesis $H^{2/n}$ (i.e., with minimal replicability requirement) is far greater than that of the RE meta-analysis null hypothesis when the number of non-null studies is at least two, showing an increased power in the replicability test even with a low portion of studies with nonnull signal. This happens while maintaining robustness in face of a single “outlier” study. As the number of studies with the null effect decreases, the performance of RE meta-analysis becomes almost as good as of replicability analysis, probably due to decreased heterogeneity.

Figure 3 shows RE settings with high and moderate heterogeneity. For this data generation, the effect sizes are nonzero with probability one. Therefore, the minimal replicability null hypothesis is never true. The power for discovering minimum replicability, that is, the test of $H^{2/8}$, is greater in all settings than the power to discover the overall effect by the RE meta-analysis. As expected, the power decreases as $u$, the minimum number of studies with effect in the same direction we want to discover, increases.

When the data is generated according to the RE model, the probability of inconsistency, that is, of having at least one positive and one negative treatment effect, increases as the overall mean approaches zero and as the heterogeneity increases. For $\theta_i \sim N(\mu, \tau^2)$, the probability to have an inconsistent configuration is $1 - \Phi(\mu/\tau)^2 - (1 - \Phi(\mu/\tau))^8$, which is $1 - \left(\frac{1}{2}\right)^{8-1} \approx 1$ when $\mu = 0$. We declare (and thus detect) inconsistency if the lower bound for both the decreasing effect and the increasing effect is at least one. The probability of detecting that the effects are inconsistent in the setting considered in Figure 3 reached $\approx 60\%$ in the setting with $\mu = 0$ and high heterogeneity, and deteriorated quickly as $\mu$ increased. Potentially more powerful tests for detecting inconsistency are available (Gail and Simon 1985; Piantadosi and Gail 1993), but these tests do not provide lower bounds on the number of studies with effect in each direction.

4. Case Studies from the Cochrane Library

We provide examples of meta-analyses in the breast cancer domain for which we can, and cannot, claim replicability. For each example, we report the $r$-value (as described in Section 2.3 with $\alpha = 0.05$) and the 95% confidence lower bounds on the number of studies with effect in each direction (as described in Section 2.4). Moreover, we provide recommendations on how to
incorporate these new analyses in the abstract and forest plots of Cochrane reviews.

The first example is based on a RE meta-analysis in review CD005211 (Figure 4). The primary objective was to review randomized controlled trials evaluating the effectiveness of exercise interventions in preventing, minimizing, or improving upper-limb dysfunction due to breast cancer treatment. The outcome in 4 is shoulder abduction ROM in degrees, early versus delayed post operative exercise. In this example, all of the studies report estimates with compatible overall effect: an increase in shoulder abduction ROM when postoperative exercise is early. Still, two of the three studies do not find statistical significance, neither does the meta-analysis. The authors’ main result is that: “Although the pooled data showed no statistically significant difference between early and delayed exercise groups (MD: 14.47 degrees, 95% CI: −2.28 to 31.21); the analysis showed considerable heterogeneity I² = 93%.” As was previously suggested in Sections 1 and 2.4, we suggest adding the r-value and lower confidence bounds on the number of studies, as follows: “The evidence toward shoulder abduction ROM was replicable,
Figure 4. The effect of exercise interventions on shoulder abduction ROM in degrees (early vs. delayed postoperative exercise, review CD005211). The evidence toward replicability is strong: the $2/3$ $r$-value = 0.0476; the 95% lower bound on the number of studies with increased effect is two, although the meta-analysis fails to declare a significant finding and there is only one study with a significant difference. Truncation is done with $t = 0.2$.

Figure 5. The effects of Combination chemotherapy on progression-free survival (review CD008792). The evidence is consistent: the $2/8$ $r$-value $< 10^{-5}$. The 95% lower bound on the number of studies with increased effect is seven studies out of eight, and no studies with decreased effect, whereas there are six studies with statistically increased effect. Truncation is done with $t = 0.2$.

with $r$-value = 0.0476. Moreover, with 95% confidence, we can conclude that at least two out of three studies had an increased effect.”

The second example is based on a FE meta-analysis in review CD008792 (Figure 5), where their main objective is to assess the effect of combination chemotherapy compared to the same drugs given sequentially in women with metastatic breast cancer. The authors report “Combination chemotherapy had a higher risk of progression than sequential single agent chemotherapy with an overall HR of 1.16 (95% CI 1.03–1.31; $p = 0.01$) and there was no statistically significant heterogeneity ($I^2 = 26\%; p = 0.22$).” We conquer with this result ($r$-value $< 10^{-5}$) where we report that replicability of increased effect is assured with 95% in at least seven out of eight studies, where in fact only six have $p$-value $< 0.05$.

The third example is based on a RE meta-analysis in review CD003366 (Figure 6). The authors compare chemotherapy regimens on overall effect in Leukopaenia. Pooling 28 studies, the RE meta-analysis fails to declare any significant difference between regimens, due to the highly significant yet contradicting results. The authors write: “Overall, there was no difference in the risk of Leukopaenia (RR 1.07; 95% CI 0.97–1.17; $p = 0.16$; participants = 6564; Analysis 5.2) with significant heterogeneity across the studies ($I^2 = 90\%; p < 0.0001$).” We suggest adding: “There is inconsistent evidence for the direction of effect: an increased risk of Leukopaenia in at least 16 studies and a decreased risk in at least 6 studies (with 95% confidence).” This is an example where the contribution of our quantification manifests, as opposed to the approach of mere counting. Among the 28 studies, there are only nine studies with significant increased effect, and five studies with significant decreased effect. We can identify clear inconsistency in the findings across the studies in this meta-analysis which requires caution when selecting studies for the analysis.

5. Replicability Assessment in the Breast Cancer Domain

We took all the updated Cochrane Collaboration systematic reviews in the breast cancer domain. Our eligibility criteria were as follows: (a) the review included forest plots; (b) at least one FE primary outcome was reported as significant at the 0.05 level, which is the default significant level used in Cochrane Reviews; (c) the meta-analysis of at least one of the primary outcomes was
based on at least three studies (d) there was no reporting in the review of unreliable/biased primary outcomes or poor quality of available evidence, and (e) the data is available for download. We consider as primary outcomes the outcomes that were defined as primary by the review authors. If none were defined we selected the most important findings from the review summaries and treated the outcomes for these findings as primary. In the breast cancer domain 62 updated (up to February 2018) reviews were published by the Cochrane Breast Cancer Group in the Cochrane library, out of which we analyzed 23 reviews that met our eligibility criteria (16, 12, 5, 2, and 4 reviews was excluded due reasons a, b, c, d, and e, respectively). Out of the 23 eligible reviews, in 8 reviews we had enough evidence to establish replicability (i.e., an \( r \)-value at most 0.05) for all the primary outcomes with meta-analysis \( p \)-values at most 0.05.

We analyzed a total of 248 primary outcomes contributed by the eligible systematic reviews of which 108 were FE meta-analyses, as reported by the authors. Out of the 71 outcomes with a statistically significant FE \( p \)-value, 64 were replicable (\( r \)-value \( \leq \) 0.05). For the 64 replicable findings, we rule out the danger that the discovery is entirely driven by one study. Thus, the evidence on the treatment effect is more trustworthy.

For the 248 primary outcomes, Table 1 summarizes the consistency evidence. As expected, among the nonsignificant outcomes the fraction of studies supporting consistency is smaller than among the significant outcomes. Ten inconsistent outcomes were detected and all of them were analyzed via RE model by the authors, warranting further research into why the effects are inconsistent across studies rather than simply turning to RE model. It is worth noting that about 16.7% of the significant meta-analyses are in fact not supported by the replicability analysis where the \( r \)-value > 0.05, indicating that the significance of these meta-analyses was most likely in light of a single study with an extreme result.

### 6. The Special Case of a Nonnull Common Effect

In this section we develop a version of the method introduced in Section 2 under the common effect assumption of FE meta-analysis. Assuming that the nonnull studies have a common effect, a powerful test statistic for the intersection hypothesis is the one used by the FE model. Specifically, \( \theta_i \in \{0, \theta\} \) for \( i = 1, \ldots, n \). The FE model alternative is that all studies have a common effect, that is, \( \theta_i = \theta \) for all \( i \).
The evidence toward consistency and inconsistency in the 245 meta-analyses, for significant (at the 0.2 level) meta-analysis outcomes (column 2) and nonsignificant meta-analysis outcomes (column 1).

| P-value | Nonsignificant meta-analysis | Significant meta-analysis |
|---------|-----------------------------|---------------------------|
|         | Consistent                  | Inconsistent               | Not enough evidence |
|         | 13                          | 9                         | 94                   |
|         | 109                         | 1                         | 22                   |

For a subset of studies \( i = \{i_1, \ldots, i_{n-u+1}\} \), the estimated common effect is

\[
\hat{\theta}_i = \frac{\sum_{k=1}^{n-u+1} \frac{\hat{\theta}_k}{SE_k}}{\sum_{k=1}^{n-u+1} \frac{1}{SE_k}},
\]

and the standard error is

\[
SE_i = \frac{1}{\sqrt{\sum_{k=1}^{n-u+1} \frac{1}{SE_k^2}}}.\]

The \( p \)-value for the intersection hypothesis \( H^R_i \) is

\[
P^R_i = 1 - \Phi\left(\frac{\hat{\theta}_i}{SE_i}\right).
\]

The \( p \)-value for \( H^L_i \) is \( P^L_i = 1 - P^R_i \). The \( p \)-value for \( H^{u/n} \), \( X \in \{L, R\} \) is

\[
r_{FE}(u) = \max_{i \in \Pi(\alpha)} P^X_i.
\]

The \( p \)-value for \( H^{u/n} : H^{u/n}(R) \cap H^{u/n}(L) \) is

\[
r_{FE}(u) = 2 \min(r^R_{FE}(u), r^L_{FE}(u)).
\]

Intuitively, the \( r \)-values should be larger than the meta-analysis \( p \)-value since a stronger scientific claim is made by rejecting \( H^{u/n} \) than by rejecting the FE meta-analysis null hypothesis. We formalize this in the following proposition.

**Proposition 6.1.** Let \( p = 2 \min(p^L_{1,\ldots,n}, p^R_{1,\ldots,n}) \) be the FE meta-analysis \( p \)-value. Then, if \( \theta_i \in [0, \theta] \) for \( i = 1, \ldots, n \), for \( u \in \{2, \ldots, n\} \); \( p < r_{FE}(u) \) if \( \hat{\theta}_{i(1,\ldots,n)} < 0 \), \( p^L_{1,\ldots,n} < r^L_{FE}(u) \); if \( \hat{\theta}_{(1,\ldots,n)} > 0 \), \( p^R_{1,\ldots,n} < r^R_{FE}(u) \).

See, Appendix A, supplementary materials for proof; for further simulations see Figures B1 and B3 in Appendix B, supplementary materials. Figure 7 shows the sensitivity of the FE model to the setting with exactly one non-null study.

When the number of studies is small it is difficult to observe whether the assumption of common effect is reasonable. Take for example scenarios where the common effect assumption is true but there is a mass of studies at effect zero (see Figure 1 columns 1 and 2), in such a case neither the FE nor the RE meta-analyses are appropriate models. Therefore, unless the common effect is clearly valid, we recommend using the general approach that we propose in Section 2.2 based on the one-sided \( p \)-values of each study, where studies can have unequal parameters and follow any distribution. We recommend using the assumption-free version in Section 2.2 since it remains valid regardless of the validity of model assumptions, unless the common effect assumption can be justified apriori.

### 7. Alternative Approaches

Other methods that offer to assess the replicability of results not relying on the pooled effect, confidence and/or its \( p \)-value, are of three types. The first are methods that rely on mere vote counting approaches that still lead to conflicting conclusions despite similar point estimates and intervals (Kirsch et al. 2008; Turner et al. 2008). Another parametric metric is estimating the proportion of replicating studies, which was proposed by Mathur and VanderWeele (2019) under the assumptions of the RE meta-analysis model.

In summary, the methods of Mathur and VanderWeele (2019, 2020a, 2020b) provide estimates for the proportions of: \( \theta_i \) that agree in direction with an original study, \( \theta_i \) that pass a prespecified threshold \( q \) (that lines up with the direction of the pooled overall effect). Most part of these (a) require a large enough number of studies such that the estimated proportion follows the normal approximation, and (b) rely on the estimation of \( \tau^2 \) which can be quite inflated when the studies experience an outlier, or when pooling estimates that are not identically distributed. The metric is useful for comparing across different yet related meta-analyses, but it is also sensitive to large or outlying studies. Despite their advantages, these methods neither offer statistically supported conclusions, nor applicable when a meta-analysis contains less than 10 studies. Their uncertainty relies on the mean of dichotomous values, which in turn yields the estimated heterogeneity \( \tau^2 \), and therefore their uncertainty increases as the number of studies decreases (Veroniki et al. 2016).

A different tool to monitor the consistency of an effect is sensitivity analysis: the meta-analysis is repeated, each time omitting one of the studies (Anzures-Cabrera and Higgins 2010).
The output is a plot of the results of these meta-analyses, called an "exclusion sensitivity plot" (Bax et al. 2006) or leave-one-out analysis (Gaudino et al. 2020). Gaudino et al. (2020)'s second figure is a forest plot for the analysis of 17 studies (Figure 8, left panel) where the inference was based on the FE model due to the low heterogeneity. They further justified their choice by an exclusion sensitivity plot (Figure 8, right panel) where they report that "sensitivity analyses confirmed the solidity of the primary analysis." However, this plot shows that the significance of the pooled estimated effect hinges solely on the results of SYNTAX: the single analysis omitting it has a corresponding overall estimate that is statistically insignificant, whereas each of the remaining 16 analyses (that include the study SYNTAX) unanimously declare significant findings. The calculation of the r-value in this example gives a clear and interpretable quantitative indication: $r(2) = 0.23$. The minimal replicability requirement, that is, $2/16$ replicability, is not met.

In meta-analyses with a larger number of studies, a result can be driven by a small fraction of studies, not just by one study. In such a case, a leave-one-out plot can not help identifying that. See, for example first forest plot in Hughes et al. (2020), Figure 9. It studies the association between blood pressure lowering with antihypertensive therapy and the incidence of dementia or cognitive impairment. The first outcome is the effect of blood pressure lowering treatment on cognitive decline, where they combine 13 studies (Figure 9, left panel). The inference was based on the RE model because of low heterogeneity, the conclusion was significant cognitive reduction of subjects treated with blood pressure lowering treatments with antihypertensive, compared with control. Out of the 13 studies, one had $p$-value $< 0.05$ but following the replicability analysis under the common effect assumption detailed in Section 6 we declare that with 95% confidence, no increased or decreased effects can be inferred. This is an example of a scenario where the significance of a single study (or multiple studies) can disappear when examined in a pool of studies, where studies with insignificant effects can contribute to the overall significance of the pooled effect.

8. Discussion

We provided examples, mainly from the Cochrane library, to demonstrate the benefit from complementing the meta-analysis with a report of the r-value and lower bounds on the number of studies with increasing and decreasing effect; although, it should be clear that the methods we offer can be used in any meta-analysis. We recommend adding the quantified information to the forest plot of the meta-analysis as a convention. We specifically encourage adding it to the two-page abstract of the Cochrane systematic review, which is a standalone document (published in MEDLINE) that briefly reports the main results and author’s conclusions.

The importance of detecting replicability for trusting the evidence in favor (or harm) of a treatment is manifested, for example, in the FDA requirement for at least two studies finding an effect (MDI, Metered Dose Inhaler and Drug, Dry Powder Inhaler DPI 1998). Possible implication of this work is for the development of drugs for less common diseases. Small sample size available in each study may fail to reach the "at least two studies" benchmark. But a handful of such studies with a proven effect in at least two studies may offer such support. In this case it is important to use the combining method in Equation (1) with a threshold $t \geq 0.1$.

According to the Cochrane Handbook for Systematic Reviews of Interventions, an outlier study is defined as the one with a markedly different intervention effect estimate for a given treatment comparison, where an influential one influences aspects of the model such as parameter estimates, heterogeneity and inconsistency. An influential study is not necessarily outlier, and vice versa. Although, some tools of detecting outlier studies are incapable of separating influential apart from outliers, where an influential study is likely to be labeled as an outlier unless a convincing argument was made based on a funnel/forest plot.

The key difference of our approach from methods of outlier detection is that we evaluate the tenacity of the pooled result despite deleting any number of studies, where we report the
biggest number of studies deleted maintaining the same result. This quantifies the replicability of the pooled effect with protection from outlier studies. It can be viewed as a quantification of the number of influential studies to the combined conclusion.

Outlier detection tools (e.g., Zhang, Fu, and Carlin 2015) aim to detect the studies with the effect estimate that most differ from the pooled estimated effect of the remaining studies. Such tools are useful for detecting departure in effect estimates but without considering single-study variance which affects the study weight in the estimated pooled effect and heterogeneity and thus, the significance of the combined conclusion.

Zhang, Fu, and Carlin (2015) seem to overcome this complication, with outlier forward search algorithm. It starts by fitting the hypothesized data generating model to a subset of the data which is gradually incremented by adding the remaining studies according to their closeness to the postulated model while monitoring changes in the estimated parameters. This tool can be incorporated with our quantification in order to detect which are the \( u^L_{\text{max}} \) or \( u^R_{\text{max}} \) inconsistent studies (whichever is smaller) or an outlier (when the bound is 1).

It may be thought that if the number of studies is large, the meta-analysis cannot be driven by one outlying study. However, we found four fairly large FE analyses, with 17, 11, 9, and 7 studies, for which the meta-analysis \( p \)-value was significant, with the 95% confidence interval not covering 1, yet the \( r \)-value > 0.05.

Establishing replicability for two studies out of four is a stronger statement than out of 20. This is where introducing \( u_{\text{max}} \) bounds offers a flexible view on replicability. The appropriate size of \( u_{\text{max}} \) relative to \( n \) is a question to be explored in each scientific discipline. Nevertheless, the \( r \)-value representing a minimal requirement, is valuable when \( n \) is small or large: a significant \( r \)-value when pooling a small number of studies reflects strong evidence toward replicability of effects; a nonsignificant \( r \)-value for numerous studies salvages from unfounded results. Table 1 shows that among the 132 meta-analyses with a significant \( p \)-value, 16.7% in fact rely on a
single study at most with a nonreplicable result, that is, \( r \)-value > 0.05.

High heterogeneity can be a result of numerous reasons, whether inevitable or not. Regardless of its source, heterogeneity may lead to studies having opposite signs for estimated effect sizes which makes pooling the estimated effects meaningless. An unperceiving mean effect does not disclose the contradictions between the different studies, but rather diminish the valuable underlying findings. When pooling a big number of studies it would be tolerable to observe both negative and positive effect estimates, regardless of the true magnitudes of effect or between study variance. Assuming the true effect is positive, it would be expected that a small portion of the studies report negative estimates. A statistically founded calculation of the bounds \( u_{\text{max}}^R \) and \( u_{\text{max}}^L \) reflects the strength of each registered signal.

The suggested complementary replicability analysis gives insight into the consistency of the effects. For example, we see from Table 1 that among the outcomes with a nonsignificant RE meta-analysis \( p \)-value, we have evidence supporting consistency in 11 and inconsistency in 9 of the 79 meta-analyses with a nonsignificant RE meta-analysis \( p \)-value.

High heterogeneity can also appear when the dispersion of the estimated effects contributes most of the overall variance, as argued by Borenstein et al. (2017), leading to nonsignificant RE meta-analysis \( p \)-value. If, in spite of that, we observe that the effects have consistent signals, it can still lead to a nonsignificant \( p \)-value. Using alternative estimation methods such as Best Linear Unbiased Prediction (BLUP) or (Bayesian) shrinkage estimates may offer improvement to the meta-analysis estimates, but as such they cannot be incorporated into our suggested method, since point estimates do not convey information regarding the uncertainty and hence about replicability or consistency (in our framework). However, assessing replicability and consistency may very well be a relevant goal for Bayesian treatment. They are likely to involve mixture priors, such as suggested method, since point estimates do not convey information regarding the uncertainty and hence about replicability or consistency (in our framework). However, assessing replicability and consistency may very well be a relevant goal for Bayesian treatment. They are likely to involve mixture priors, such as

\[
\text{The only caveat is that meta-analysis is prone to publication bias, where only significant results (at } p \text{-value } \leq 0.05 \text{) are published. The Cochrane reviews are known to be careful during their search for eligible studies, avoiding as much as possible this problem. In other areas, where this may not be feasible, using conditional } p \text{-values rather than the raw ones in the procedures may circumvent the problem (with unfortunate loss of some power.) More concretely, we can use the result of (Zhao, Small, and Su 2019) that showed that for one-sided tests in a one-dimensional exponential family, the conditional distribution of the } p \text{-value divided by } \alpha, \text{ given that the } p \text{-value was at most } \alpha, \text{ is stochastically larger than the uniform (0,1) distribution for any } \alpha. \text{ Thus, limiting our inference only to the set of hypotheses with } p \text{-values at most } 0.05 \text{(for example), and using a valid combining function on these } p \text{-values divided by } 0.05, \text{ provides a valid inference immune at least to some forms of publication bias (specifically, that studies are published only if they are below a certain threshold that is at least } 0.05\).

### Supplementary Materials

An R package implementing the methods proposed in this article is now available for download at CRAN, under the name “metarep” (https://cran.r-project.org/web/packages/metarep/index.html). R-codes for generating the reported simulations and reproducing the examples are available on GitHub (https://github.com/IJaljuli/r-value). Supplemental file includes (a) proof for Proposition 6.1, and (b) results of simulations like in Section 3 for \( n = 4, 20 \) and both equal and unequal group sizes.

### Funding

Iman Jaljuli gratefully acknowledges The Baroness Ariane de Rothschild Women’s Doctoral grant. Yoav Benjamini gratefully acknowledges Joint US-NSF and US-Israel BSF [grant no. 2016746]. Orestis A. Panagiotou gratefully acknowledges U.S. Agency for Healthcare Research and Quality [grant no. IR03HS025840] and reports personal fees from International Consulting Associates Inc. outside the submitted work. Ruth Heller gratefully acknowledges Israeli Science Foundation [grant 1049/16].

### References

- Anagnostou, O. A., and Trikalinos, T. A. (2015), “On Effect Measures, Heterogeneity, and ‘The Laws of Nature,” Epidemiology, 26, 710–713. [373]
- Anzures-Cabrera, J., and Higgins, J. P. T. (2010), "Graphical Displays for Meta-analysis: An Overview with Suggestions for Practice," Research Synthesis Methods, 1, 66–80. [381]
- Bax, L., Yu, L. M., Ikeda, N., Tsuruta, H., and Moons, K. M. (2006), “Development and Validation of MIX: Comprehensive Free Software for Meta-analysis of Causal Research Data,” BMC Medical Research Methodology, 6, 50. [382]
