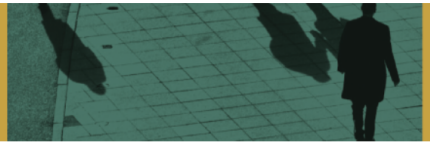
Replicability in Science: Day 2, Paper 4: How to measure replication

giovanni_parmigiani@dfci.harvard.edu

Padova, July 9, 2024

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS



Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 3, *pp.* 1145–1166

New statistical metrics for multisite replication projects

Maya B. Mathur

Stanford University, USA

and Tyler J. VanderWeele

Harvard University, Boston, USA

[Received April 2018. Final revision March 2020]

one-lab versus many-labs replication

”Many Labs projects select multiple original studies and subject each to a multisite replication”

prelude

meta-analysis and replicability

Ganzfried etal 2007

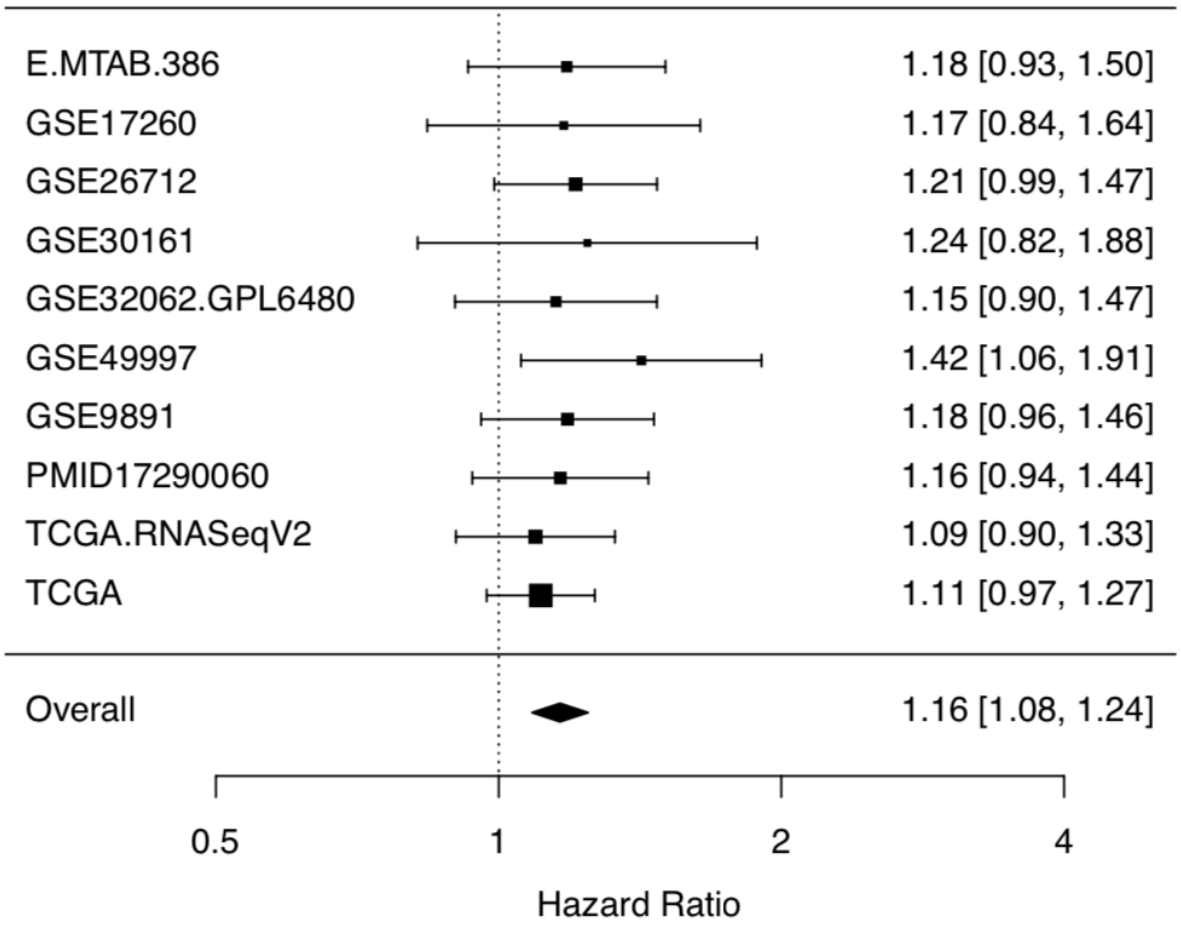
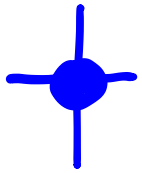


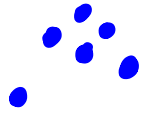
Figure 2: Validation of CXCL12 as an independent predictor of survival
This figure shows a forest plot as in Figure 1, but the CXCL12 expression levels were adjusted for debulking status (optimal versus suboptimal) and tumor stage. The p-value for the overall HR, found in res$pval, is 1.8e-05.

aggregate replicability

Zhong etal 2007

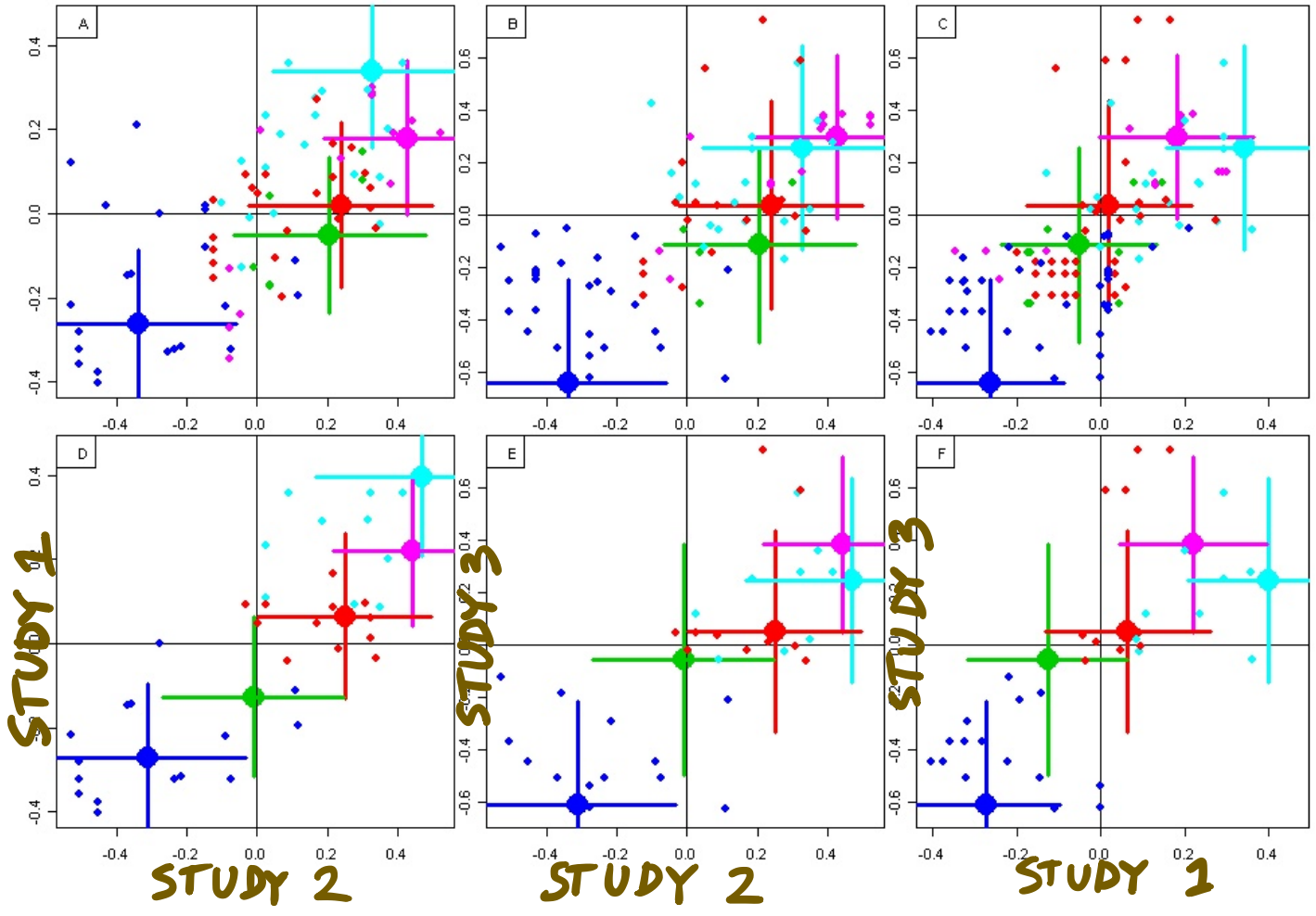


SIGNATURE



GENES WITHIN SIGNATURE

FILTERED UNFILTERED



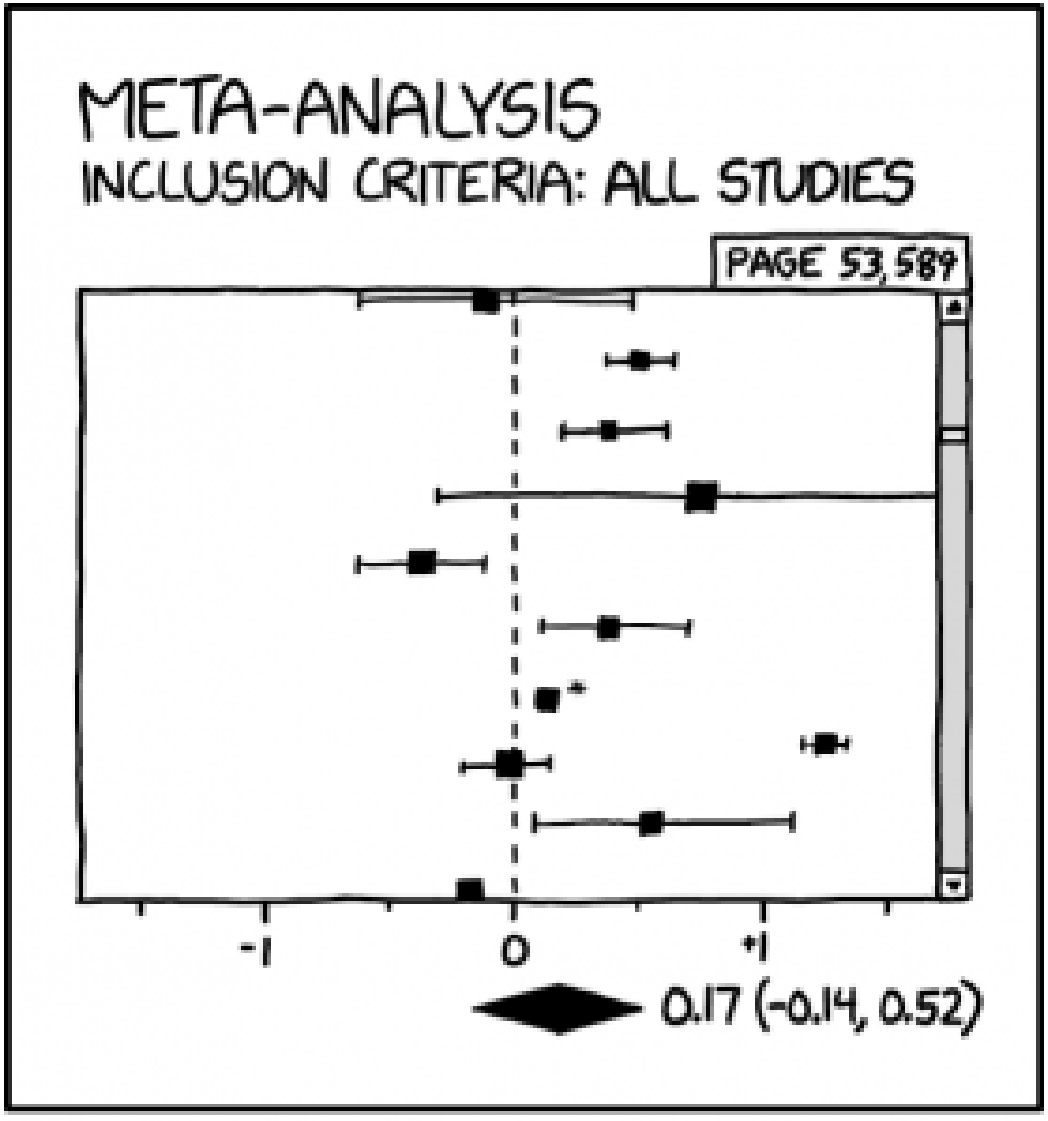
SCALE: PREDICTION STRENGTH (COX EFFECT SIZE)

metrics

pooled effect size

”First, nearly all many-to-one designs report a pooled estimate of the effect size in the replications.

The pooled estimate is usually estimated by meta-analysing effect sizes from the replications or by fitting a mixed model to individual subject data.”



BAD NEWS: THEY FINALLY DID A META-ANALYSIS OF ALL OF SCIENCE, AND IT TURNS OUT IT'S NOT SIGNIFICANT.

significance agreement

”whether the replication study obtains a statistically significant p-value and an effect estimate in the same direction as in the original study (assuming that the original study itself obtained a significant p-value).

significance agreement is challenging to interpret because it is a function not only of the nominal α-level (e.g. 0.05), but also of power in both the original and the replication study. ”

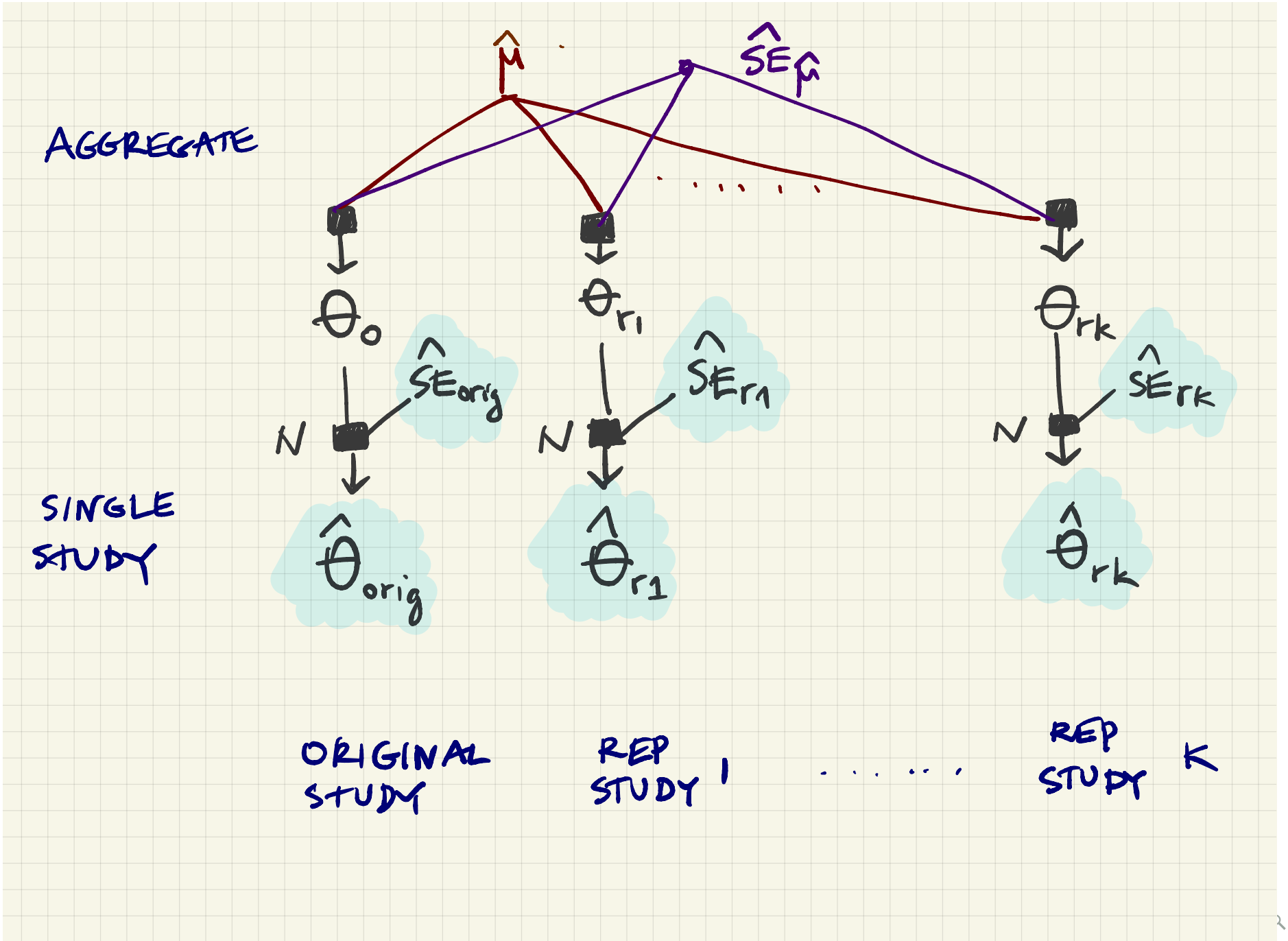
prediction interval

Patil 2016

”use the original study to construct a prediction interval representing a plausible range for the effect estimate in the replication study, assuming that the replication and the original study are generated from the same distribution.

If indeed the two studies are generated from the same distribution, then regardless of power in either study there is, by construction, a 95% probability that the replication effect estimate will fall inside the prediction interval. ”

conceptual framework behind Mathur (I think)



consistency of original with replications

”the probability that, if indeed the original is consistent with the replications in this sense, its estimate would be as extreme or more extreme than it actually was.

$$P_{\text{orig}} = 2\left[1 - \Phi\left\{\frac{|\hat{\theta}_{\text{orig}} - \hat{\mu}|}{\sqrt{(\hat{\tau}^2 + \widehat{\text{SE}}^2_{\text{orig}} + \widehat{\text{SE}}^2_{\hat{\mu}})}}\right\}\right]. \tag{4.1}$$

proportion of population effects agreeing in direction

$$\tilde{\theta}_{\text{rep}j} = \hat{\mu} + (\hat{\theta}_{\text{rep}j} - \hat{\mu})\sqrt{\left(\frac{\hat{\tau}^2}{\hat{\tau}^2 + \widehat{\text{SE}}^2_{\text{rep}j}}\right)}.$$

$$\hat{P}_{>0} = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}(\tilde{\theta}_{\text{rep}j} > 0)$$

proportion of meaningfully strong population effects

proportion of effects that are stronger than a non-null threshold, q

coda

decision theoretic replication

with Sajia Darwish

imagine each study is to gather evidence to support a binary choice between a_1 and a_2.

a measure of replication is then the proportion of times in which studies support the decision initially supported by the first study.

this will depend on the decision maker(s) in ways that go beyond the data, and includes the consequences of the actions and generally prior information