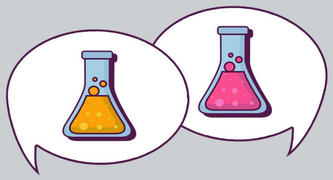
How much reproducibility do we want? An app and model for understanding replication

Kleber Neves, Pedro Tan & Olavo Amaral

Institute of Medical Biochemistry Leopoldo De Meis, Federal University of Rio de Janeiro - Brazil

INTRODUCTION

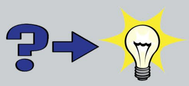
Science can never be 100% reproducible and one can waste resources both by having too little reproducibility or by taking excessive measures to have too much of it. With the many efforts in the last few years to produce estimates of the reproducibility rate of different fields - including our own -, this question becomes a central one.



To address this issue, we developed a computational model and an associated R Shiny App where one can explore reproducibility rates in different scenarios by varying sample sizes, bias and publication incentives.

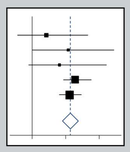
This works both as (1) a model to help us interpret empirical estimates of reproducibility and (2) a web app that can be used for teaching and learning about reproducibility.

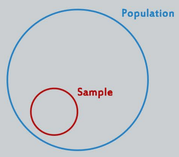
THE MODEL



Scientists choose an effect to investigate. These can be dichotomous or vary in their sizes and signals. They can be either replications or original effects.

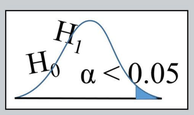
Effect sizes vary between labs. They will be subject to both sampling error and interlab variation, and will thus vary each time they are tested.





To test an effect, one draws a sample of a preset size, from normal populations where the difference between means is equal to the effect size being investigated.

Measurement error is added to the sample values. After this, a t-test is performed, comparing the two samples. A p-value is obtained.



Depending on publication incentives for replications and non-significant results, the findings are either published or not.

METHODS

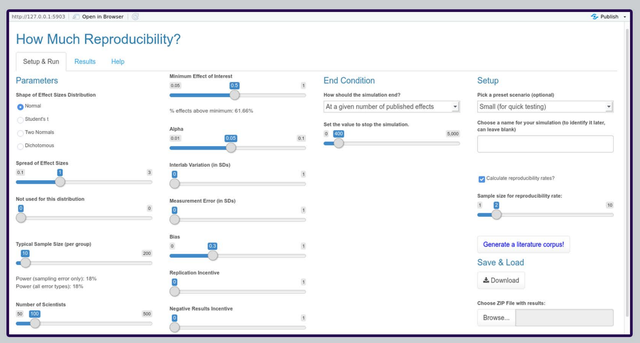
After many iterations, various measures can be obtained from the scientific literature generated with the model (e.g. positive predictive value, rate of false positives, etc).

Replication initiatives can also be simulated using the same model. Replication experiments are performed to compare different measures of replication success and the reproducibility rates obtained with each of them.

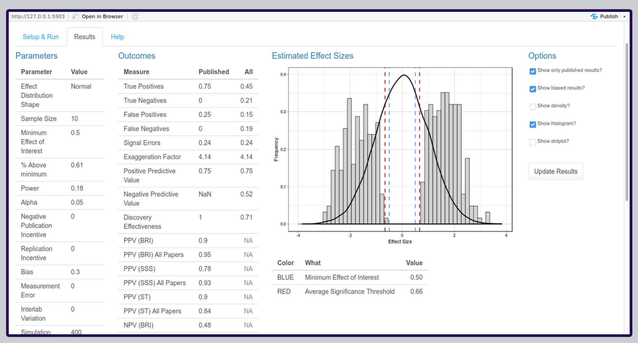
The model was implemented using the R language and the Shiny Apps environment and packages in RStudio.

SHINY APP

The setup tab is where users can set the parameters for the simulation and for how long it should run. After a simulation is finished, from the same interface, the user can save and load results for later analysis.



The results tab, where users can see the characteristics of the literature generated by the model, including the distribution of effect sizes, as histograms, dotplots or density curves. The user can choose to see published and unpublished findings, comparing them to the underlying distribution from where effect sizes are derived, detecting biases by visual inspection.



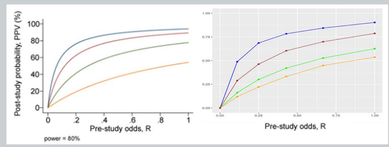
Parameter	Value
Effect Distribution Shape	Normal
Sample Size	10
Minimum Effect of Interest	0.5
% Above minimum	0.61
Power	0.18
Alpha	0.05
Negative Publication Incentive	0
Replication Incentive	0
Bias	0.3
Measurement Error	0
Interlab Variation	0
Simulation	400

Measure	Published	All
True Positives	0.75	0.45
True Negatives	0	0.21
False Positives	0.25	0.15
False Negatives	0	0.19
Signal Errors	0.24	0.24
Exaggeration Factor	4.14	4.14
Positive Predictive Value	0.75	0.75
Negative Predictive Value	NaN	0.52
Discovery Effectiveness	1	0.71
PPV (BRI)	0.9	NA
PPV (BRI) All Papers	0.95	NA
PPV (SSS)	0.78	NA
PPV (SSS) All Papers	0.93	NA
PPV (ST)	0.9	NA
PPV (ST) All Papers	0.84	NA
NPV (BRI)	0.48	NA

Color	What	Value
BLUE	Minimum Effect of Interest	0.50
RED	Average Significance Threshold	0.66

VALIDATION

Scenario description	Power (%)	R	Bias (%)	PPV (%)	Dichotomous Sim PPV (%)*
Adequately powered RCT with little bias and 1:1 pre-study odds	80	1:1	10	85	84.9
Confirmatory meta-analysis of good quality RCTs	95	2:1	30	85	85.2
Meta-analysis of small inconclusive studies	80	1:3	40	41	40.8
Underpowered, but well-performed phase I/II RCT	20	1:5	20	23	23.9
Underpowered, poorly performed phase I/II RCT	20	1:5	80	17	17.5
Adequately powered exploratory epidemiological study	80	1:10	30	20	20.6
Underpowered exploratory epidemiological study	20	1:10	30	12	11.9
Discovery-oriented exploratory research with massive testing	20	1:1000	80	0,1	0.12
Discovery-oriented exploratory research with massive testing, but with more limited bias (more standardized)	20	1:1000	20	0,15	0.14

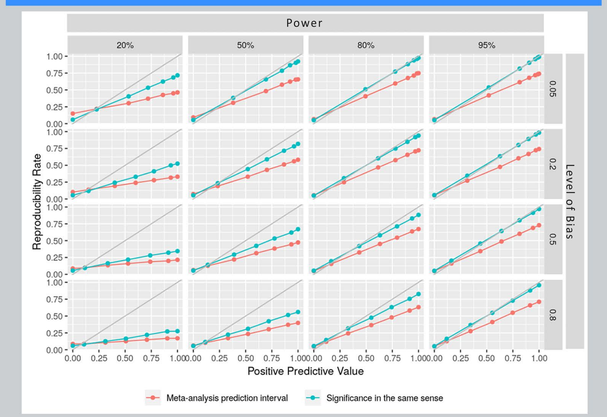


A comparison of the analytical results from Ioannidis (2005) and our model. A dichotomous population of effects with a fixed effect size (which inform sample size calculations) is simulated with the model. Graphs show how the PPV change with increasing pre-study odds.

Reference: Ioannidis (2005). PLoS medicine, 2(8), e124.

* average of 30 simulations

MEASURING REPRODUCIBILITY

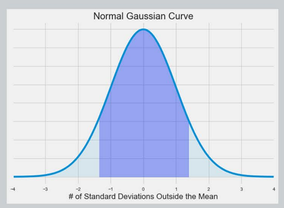


Reproducibility rate, measured in two ways, compared to the positive predictive value of the literature. Three replication attempts are made for each published result. Replication success is defined as the original effect falling within the prediction interval of the meta-analysis of the replications (red) or the meta-analysis being significantly different from 0 and in the same sense (blue). The gray line is a reference where they are equal (x = y).

CONCLUSIONS

We plan to release the app to the public as an interactive online tool, as well as to expand our analysis of the model, in particular, to use continuous effect sizes instead of dichotomous ones.

	Positive	Negative
Positive	True Positive (A)	False Negative (C)
Negative	False Positive (B)	True Negative (D)



We also want to evaluate the sensitivity and specificity of different measures of replication success, as well as to test the impact of various incentives and study parameters on the reliability of the scientific literature, as would be measured by replication efforts.

kleber.na@gmail.com