Confirmatory studies in methodological statistical research

Concept and illustration

F. Julian D. Lange[1,2] Anne-Laure Boulesteix[1,2]
julian.lange@ibe.med.uni-muenchen.de

[1]Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich
[2]Munich Center for Machine Learning (MCML)

Conference of the Central European Network, September 2023

Outline

1. Confirmatory methodological studies and preregistration: Why and how?

2. Illustration

3. Discussion, conclusion and takeaways

Confirmatory methodological studies and preregistration: Why and how?

Replication crisis in applied research

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

(Ioannidis, 2005)

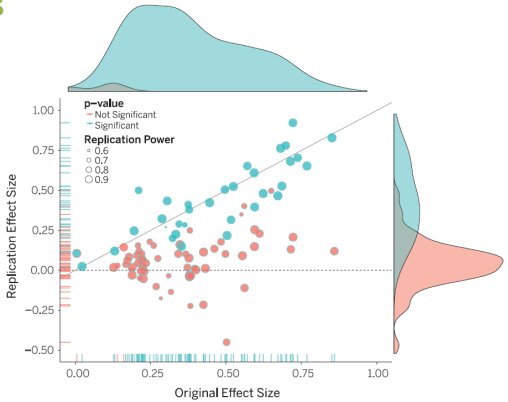
Estimating the reproducibility of psychological science

(Open Science Collaboration, 2015)

NEWS | 09 December 2021

Half of top cancer studies fail high-profile reproducibility effort

(Mullard, 2021)



(Open Science Collaboration, 2015)

Replication crisis in applied research

Some of the reasons for the crisis:

- Multiplicity of analysis strategies and researcher degrees of freedom (RDFs)
- Practices like selective reporting and *p*-hacking
- Publication bias
- Failing to distinguish between exploratory and confirmatory research and hypothesizing after results are known (HARKing)

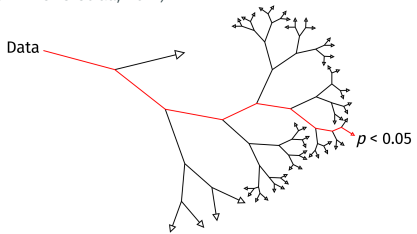
But those issues only exist in *applied* research fields like psychology, right? Well...

General Article

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn

(Simmons et al., 2011)



Special Section on Replicability in Psychological Science: A Crisis of Confidence?

An Agenda for Purely Confirmatory Research

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit

(Wagenmakers et al., 2012)

Replication crisis in methodological research?

Some of the reasons for the crisis:

- Multiplicity of analysis strategies and researcher degrees of freedom (RDFs)
- Practices like selective reporting and *p*-hacking
- Publication bias
- Failing to distinguish between exploratory and confirmatory research and hypothesizing after results are known (HARKing)

But those issues only exist in *applied* research fields like psychology, right? Well...

...in methodological statistical/ML research:

- Many study design and analysis choices with large impact on results: datasets, performance measures, competing methods, etc. (Nießl et al., 2022)
- Newly proposed methods almost always seem to outperform existing methods (Norel et al., 2011; Boulesteix et al., 2013; Buchka et al., 2021)
- Distinction between exploratory and confirmatory research rarely discussed or considered

Distinguishing between exploratory and confirmatory research

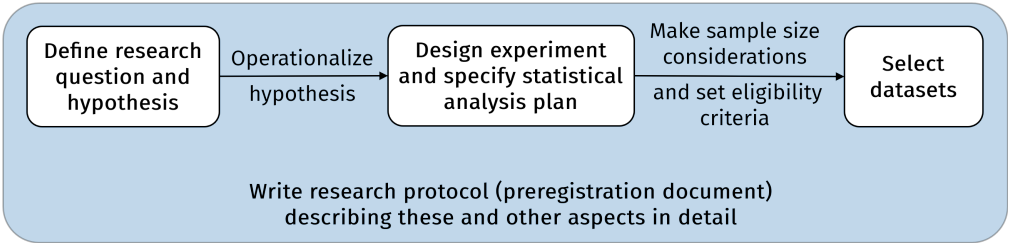
Distinction between exploratory (hypothesis-generating) and confirmatory (hypothesis-testing) research:

- Why important? Only for purely confirmatory analyses null hypothesis testing and *p*-values retain their diagnostic value, clear separation reduces risk of bias and false-positive results

- In practice? Often no clear distinction

- How to ensure confirmatory research? In applied research → **Preregistration!**
  → public, time-stamped registration of study plans (hypotheses, study design, methods and analysis plan) prior to collecting or accessing data
  → Why not try that for real-data methodological research?

Workflow of a real-data confirmatory methodological study via preregistration I

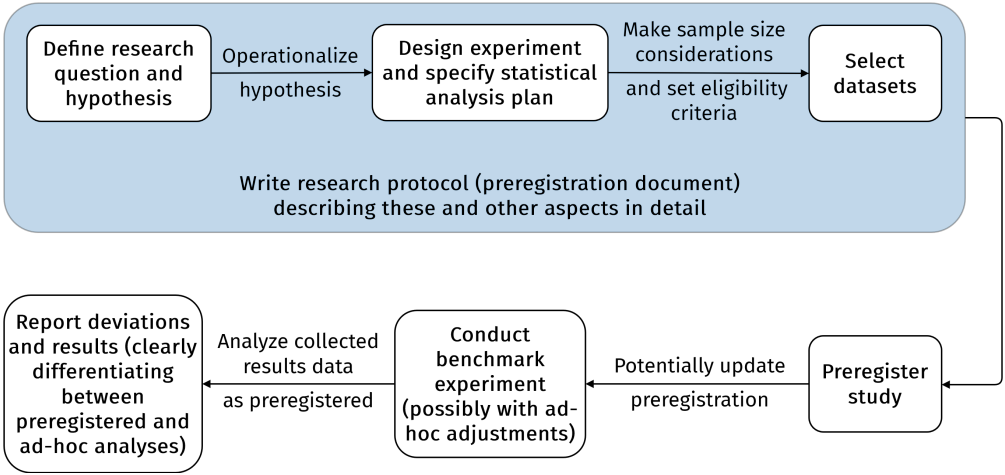


- Effectiveness of preregistration depends on comprehensiveness of the document
- For applied research: Several templates available to aid researchers
- For methodological research: We developed the following checklist for this document for real-data methodological studies

A study protocol checklist for real-data methodological research

Administrative information	Basic study information, registration details, protocol version and history, abstract
Introduction	Study rationale, background, research questions and hypotheses
Datasets	Population, sources, sample size considerations, eligibility criteria, selection process and its results
Benchmark experiment plan	Study design, studied methods and measures, preprocessing procedure, methods configuration
Analysis plan	Confirmatory analyses, contingencies, sensitivity analyses
Software, hardware and reproducibility	Software packages, methods implementations, statement on reproducibility efforts
Prior knowledge, neutrality and dissemination	Knowledge about the selected datasets and prior analyses of them, neutrality statement regarding methods, dissemination plan

Workflow of a real-data confirmatory methodological study via preregistration II



Illustration

Illustration: Study protocol

- Hypothesis: “Random forests need more events per variable (EPV) than logistic regression to achieve a stable/good predictive performance (AUC)”
  $\rightarrow$ Result from van der Ploeg et al. (2014) based on simulated data, but does it hold up when analyzing many real datasets?

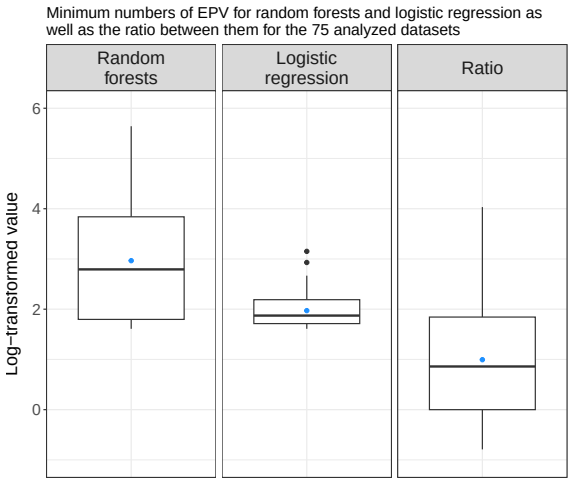
- Datasets: 75 datasets from OpenML

- Design: For each cross-validation iteration for every dataset $\rightarrow$ split training data into 24 subsets (fixed numbers of EPV between 5-500) $\rightarrow$ train RF and LR models on the complete training data and its 24 subsets $\rightarrow$ evaluate models on the test data

- Operationalization and confirmatory analysis: ”good/stable performance” $\stackrel{\text{def}}{=}$ surpassing 95% of ”max. possible” performance (AUC of model using complete training data) $\rightarrow$ identify minimum number of EPV where that criteria is fulfilled $\rightarrow$ take ratio of minimum number for RF and LR $\rightarrow$ test null hypothesis that ratio $= 1$

Results of confirmatory analysis

- Average number of EPV needed to reach their predictive performance potential: 7.18 for LR, 19.41 for RF $\rightarrow$ RF require 2.7 times more EPV
- p-value of Wilcoxon signed-rank test < 0.001

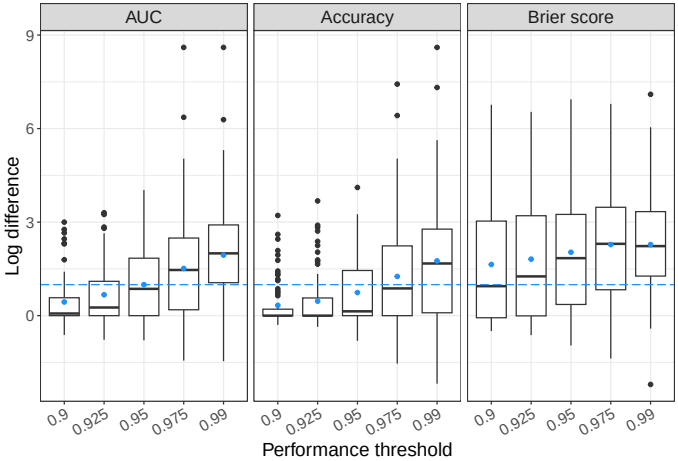


Sensitivity analysis

Investigation of results' sensitivity to design and analysis choices by repeating the confirmatory analysis pipeline for each combination of the following considered options:

Design or analysis choice	Considered options
Dataset selection	All 75 datasets (conf. analysis), eight subgroups ($<$ or $\geq$ median of n, p, EPV^{tot}, pct^{evt})
Performance measure	AUC (conf. analysis), accuracy, Brier score
Performance threshold in evaluation metric (stability threshold)	90%, 92.5%, 95% (conf. analysis), 97.5%, 99%
Handling of missing values in evaluation metric	20%-threshold rule (conf. analysis), worst, mean, weighted
Aggregation of evaluation metric values across CV iterations ("within" a dataset)	Geometric mean (conf. analysis), median
Aggregation of evaluation metric values across datasets	Geometric mean (conf. analysis), median

Impact of the choice of performance measure and threshold on study result



- Vastly different results across the 15 considered analysis strategies
- Reported factor by which random forests require more EPV than logistic regression could have been between 1.39 and 9.78
- Shows the importance and value of prespecification and preregistration

Discussion, conclusion and takeaways

Discussion and conclusion

- Preregistration and research protocols can prevent over-optimistic results and help researchers avoid human biases
- Issues for real-data methodological studies $\rightarrow$ mainly related to datasets, e.g., definition of population of interest, loss of blinding/access prior to analysis, time-consuming selection process, limited number of available datasets, ...
- In general: Preregistration is not a silver bullet for the replication crisis

Takeaways

- Most importantly: Consider the exploratory-confirmatory distinction when conducting methodological research and interpreting your results
- If you do come across a methodological research question suitable for a strictly confirmatory study with real datasets → Why not give preregistration (aided by our study protocol checklist) a try?
- Even if it's not a real-data study but a confirmatory simulation study → Checklist can largely be used simulation studies
- And even if you don't plan on doing strictly confirmatory research → Checklist and preregistration could make you to think about often overlooked aspects that impact research findings, reducing the risk of human biases and over-optimistic reporting

"You must not fool yourself, and you are the easiest person to fool." — Richard Feynman

Thank you!

Want more details on the illustration (incl. full protocol) and/or the complete protocol checklist?
→ See my MA thesis and/or the preliminary version of complete checklist with all items, both of which as well as these slides are available on my LMU website:
www.ibe.med.uni-muenchen.de/mitarbeiter/mitarbeiter/julian-lange/index.html

References

Boulesteix, A.-L., Lauer, S., and Eugster, M. J. (2013). A plea for neutral comparison studies in computational sciences. *PloS ONE*, 8:e61562.

Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R., and Boulesteix, A.-L. (2021). On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology*, 22:152.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8):e124.

Mullard, A. (2021). Half of top cancer studies fail high-profile reproducibility effort. *Nature*, 600(7889):368–369.

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., and Boulesteix, A.-L. (2022). Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *WIREs Data Mining and Knowledge Discovery*, 12(2):e1441.

Norel, R., Rice, J. J., and Stolovitzky, G. (2011). The self-assessment trap: can we all be better than average? *Molecular Systems Biology*, 7:537.

References

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

van der Ploeg, T., Austin, P. C., and Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14(1):137.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6):632–638.