The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines

A.-L. Boulesteix and S. Hoffmann

MetaArXiv Preprints

Submit a Preprint Search Donate Sign Up Sign In

The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines

AUTHORS

Sabine Hoffmann, Felix Schönbrodt, Ralf Elsas, Rory Wilson, Ulrich Strasser, Anne-Laure Boulesteix

Seite: 1 von 39 − + 70%

The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines

Sabine Hoffmann,[1,2] Felix Schönbrodt,[1,3] Ralf Elsas,[1,4] Rory Wilson,[5] Ulrich Strasser,[6] Anne-Laure Boulesteix[1,2,7*]

[1]Open Science Center of the Ludwig Maximilian University (LMU) Munich, Germany

[2]Institute for Medical Information Processing, Biometry, and Epidemiology, Medical School, LMU Munich, Germany

[3]Department of Psychology, Psychological Methods and Assessment, LMU Munich, Germany

[4]Institute for Finance & Banking, Munich School of Management, LMU Munich, Germany

[5]Research Unit of Molecular Epidemiology, Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[6]Department of Geography, University of Innsbruck, Austria

[7]Department of Statistics, Faculty of Mathematics, Computer Science and Statistics, LMU Munich, Germany

Abstract

For a given research question, there are usually a large variety of possible analysis strategies acceptable according to the scientific standards of the field, and there are concerns that this multiplicity of analysis strategies plays an important role in the non-replicability of research findings. Here, we define a general framework on common ...

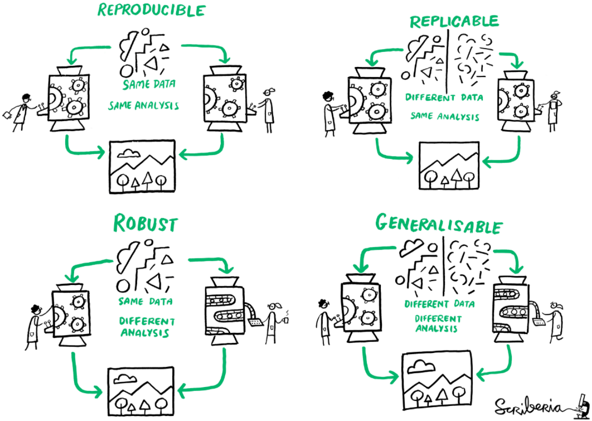
See more

Preprint DOI

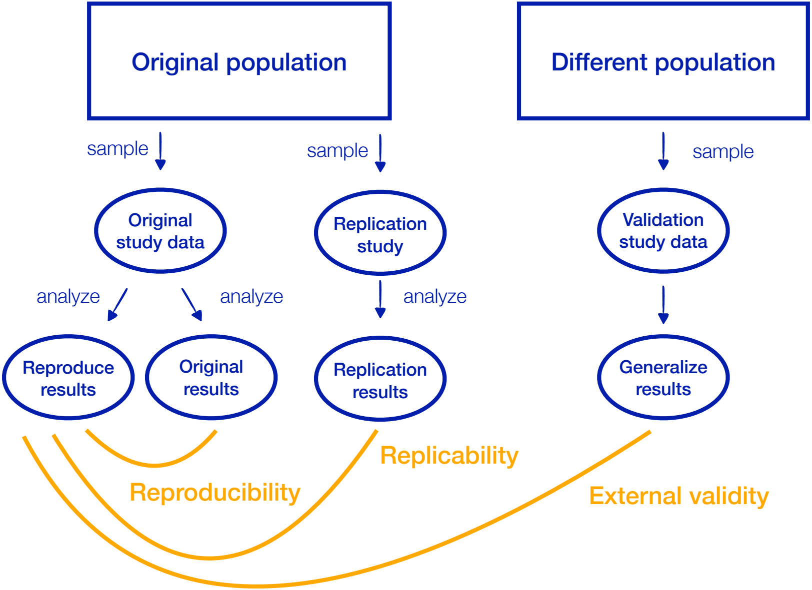
10.31222/osf.io/afb9p

License

Remember Boris Hejblum's talk...



The Turing Way Community and Scriberia. Illustrations from the turing way book dashes, 2020. URL https://doi.org/10.5281/zenodo.3332808. https://doi.org/10.5281/zenodo.3695300. [Hejblum et al., 2020]


Original population
Different population
sample
sample
sample
Original study data
Replication study
Validation study data
analyze
analyze
analyze
Reproduce results
Original results
Replication results
Generalize results
Replicability
Reproducibility
External validity

The replication crisis in science

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

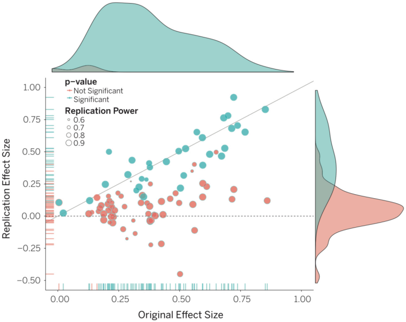
Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability

Corollary 4: The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true. Flexibility increases the potential for transforming what would be "negative" results into "positive" results,

The replication crisis in science

Psychology
[Open Science Collaboration, 2015]



Preclinical research
[Freedman et al., 2015]

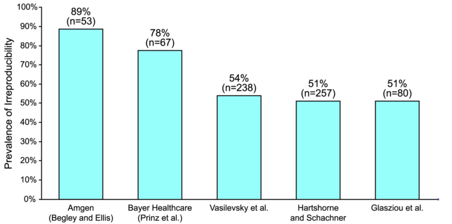


Fig 1. Studies reporting the prevalence of irreproducibility. Source: Begley and Ellis [6], Prinz et al. [7], Vasilevsky [8], Hartshorne and Schachner [5], and Glasziou et al. [9].

Reasons for the non-replicability of research findings

- Fraud and scientific misconduct
  [Ince, 2011, Chandler et al., 2012, Anaya et al., 2017]

Reasons for the non-replicability of research findings

- Fraud and scientific misconduct
  [Ince, 2011, Chandler et al., 2012, Anaya et al., 2017]
- Publication bias
  [Sterling, 1959, Easterbrook et al., 1991, Begg and Mazumdar, 1994]

Reasons for the non-replicability of research findings

- Fraud and scientific misconduct
  [Ince, 2011, Chandler et al., 2012, Anaya et al., 2017]
- Publication bias
  [Sterling, 1959, Easterbrook et al., 1991, Begg and Mazumdar, 1994]
- Combining the multiplicity of possible analysis strategies with selective reporting
  [Ioannidis, 2005b, Gelman and Loken, 2014, Goodman et al., 2016]

The multiplicity of analysis strategies in empirical research

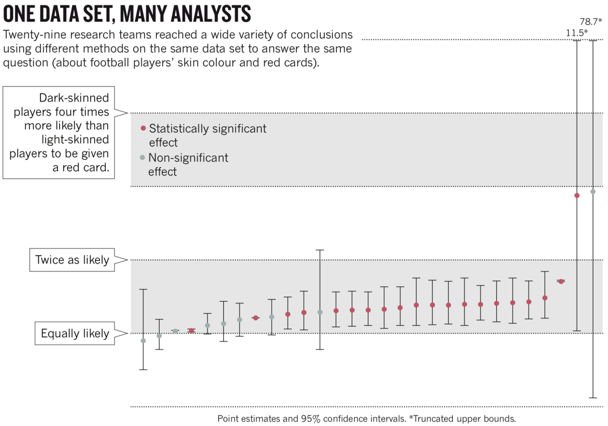
Are football referees more likely to give red cards to players with dark skin than to players with light skin? [Silberzahn and Uhlmann, 2015]

MICHAEL REGAN/GETTY

Mario Balotelli, playing for Manchester City, is shown a red card during a match against Arsenal.

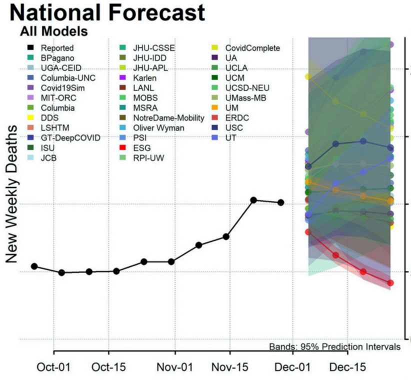
The multiplicity of analysis strategies in empirical research

Are football referees more likely to give red cards to players with dark skin than to players with light skin? [Silberzahn and Uhlmann, 2015]

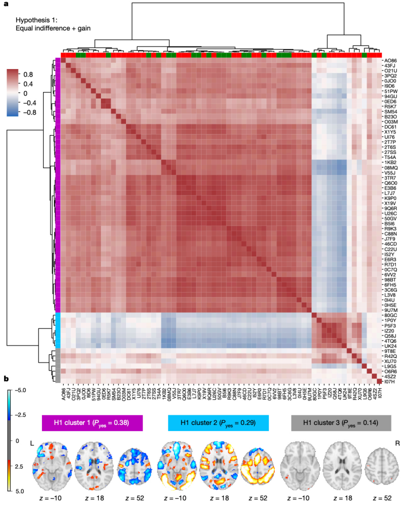


The multiplicity of analysis strategies in empirical research

... beyond hypothesis testing:



COVID-19 modelling



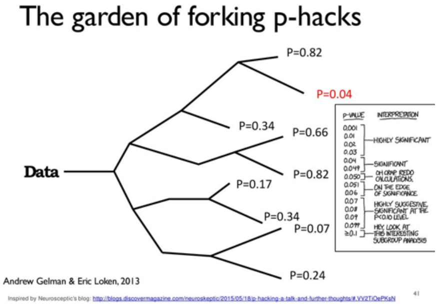
[Botvinik-Nezer et al., 2020]

The multiplicity of analysis strategies and selective reporting

Well-investigated in the context of hypothesis testing

Known as:

- "fishing for significance"
- "p-hacking"



t-test or Mann-Whitney
with equal variance or not
keeping outliers
or excluding them
or winsorizing them
excluding a subgroup
excluding missing values
or imputing them

Fishing for significance/p-hacking

[Ioannidis, 2005a]:

"Give me information on a single gene and 200 patients, half of them dead, please. I bet that I can show that this gene affects survival ($p<0.05$) even if it does not. One can do analyses: counting or ignoring exact follow-up, censoring at different timepoints, excluding specific causes of death, exploiting subgroup analyses, using dozens of different cut-offs to decide what constitutes inappropriate gene expression, and so forth. Without highly specified a priori hypotheses, there are hundreds of ways to analyse the dullest dataset. Thus, no matter what my discovery eventually is, it should not be taken seriously, unless it can be shown that the same exact mode of analysis gets similar results in a different dataset. Validation becomes even more important when datasets become complex and analytical options increase exponentially."

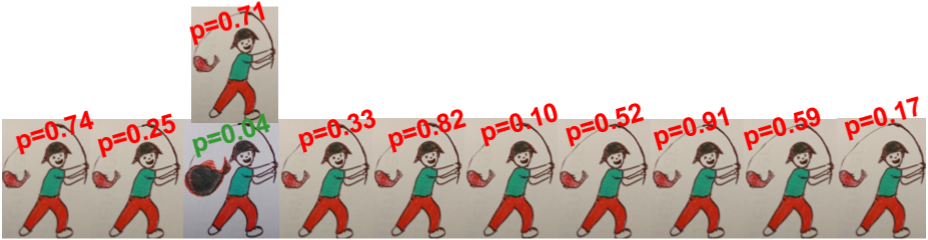
Fishing for significance/p-hacking... in simple words

If we test enough times, we finally get something significant—even if there is actually nothing.

If we fish many fishes, it is likely that one of them will be big even if fishes are usually small in this lake.

But this result will most likely not be confirmed in replication studies!

if I try to fish again in the same place—for validation purposes—and only try once, the fish this time will probably not be big again...



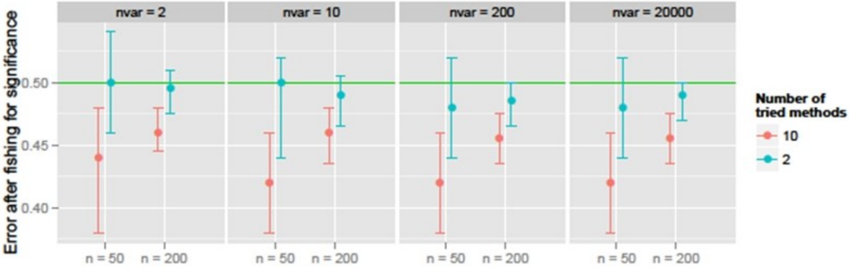
The multiplicity of analysis strategies and selective reporting

- cherry-picking
- data dredging
- data snooping
- ...

The multiplicity of analysis strategies and selective reporting

... beyond hypothesis testing:

- $K = 2$ or $K = 10$ supervised learning algorithms
- sample size $n = 50$ or $n = 200$
- $nvar = 2, 10, 200, 20000$ variables



[Boulesteix et al., 2017]

Also widespread in methodological research...

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 16 2010, pages 1990–1998
doi:10.1093/bioinformatics/btq323

Gene expression

Advance Access publication June 26, 2010

Over-optimism in bioinformatics: an illustration

Monika Jelizarow[1], Vincent Guillemot[1,2], Arthur Tenenhaus[2], Korbinian Strimmer[3] and Anne-Laure Boulesteix[1,*]

[1]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany, [2]SUPELEC Sciences des Systèmes (E3S)-Department of Signal Processing and Electronics Systems - 3, rue Joliot Curie, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France and [3]Department of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany

Associate Editor: John Quackenbush

ABSTRACT

Motivation: In statistical bioinformatics research, different optimization mechanisms potentially lead to 'over-optimism' in published papers. So far, however, a systematic critical study concerning the various sources underlying this over-optimism is lacking.

Results: We present an empirical study on over-optimism using high-dimensional classification as example. Specifically, we consider a 'promising' new classification algorithm, namely linear discriminant analysis incorporating prior knowledge on gene functional groups

it would be wrong to report only favorable datasets without mentioning and/or discussing the other results. This strategy induces an optimistic bias. This aspect of over-optimism is quantitatively investigated in the study by Yousefi *et al.* (2010) and termed as 'optimization of the dataset' in this article.

The second source of over-optimism, which is related to the optimal choice of the dataset mentioned above, is the optimal choice of a particular setting in which the superiority of the new algorithm is more pronounced. For example, researchers could report the results obtained after a particular feature filtering which favors the

In this talk

Interdisciplinary perspective on the multiplicity of analysis strategies and lessons learned across disciplines:

- general framework to describe sources of uncertainty arising in empirical research
- impact on the replicability of research findings
- potential solutions proposed across disciplines

Sources of uncertainty in empirical research

The multiplicity of analysis strategies in empirical research

Does meat intake increase the risk of colorectal cancer?

Prediction of the future water mass in seasonal snowpack

The multiplicity of analysis strategies in empirical research

Does meat intake increase the risk of colorectal cancer?

Prediction of the future water mass in seasonal snowpack

$\downarrow$ $\downarrow$

θ $\mathscr{Y}$

The multiplicity of analysis strategies in empirical research

Does meat intake increase the risk of colorectal cancer?

- Define input and outcome variables
- Handle outliers and missing values

	X_1		X_p	Y
1				
n				

θ

Prediction of the future water mass in seasonal snowpack

- Collect and process input data
- Handle outliers and missing values

	X_1			X_p
1				
t				

$\mathcal{Y}$

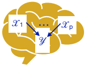
The multiplicity of analysis strategies in empirical research

Does meat intake increase the risk of colorectal cancer?

- Define input and outcome variables
- Handle outliers and missing values

	X_1		X_p	Y
1				
n				

- Choose variables to include in model
- Choose functional form



θ

Prediction of the future water mass in seasonal snowpack

	X_1			X_p
1				
t				

- Collect and process input data
- Handle outliers and missing values

θ

- Specify model structure
- Choose values for model parameters

$\mathcal{Y}$

The multiplicity of analysis strategies in empirical research

Does meat intake increase the risk of colorectal cancer?

Prediction of the future water mass in seasonal snowpack

- Define input and outcome variables
- Handle outliers and missing values

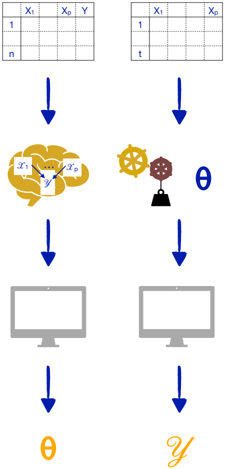
- Collect and process input data
- Handle outliers and missing values

- Choose variables to include in model
- Choose functional form

- Specify model structure
- Choose values for model parameters

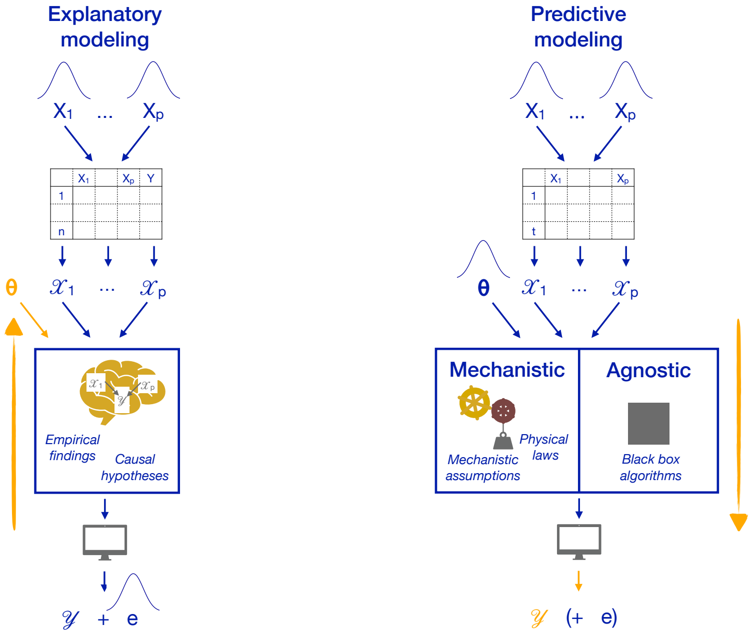
- Choose method and method settings to estimate parameter vector

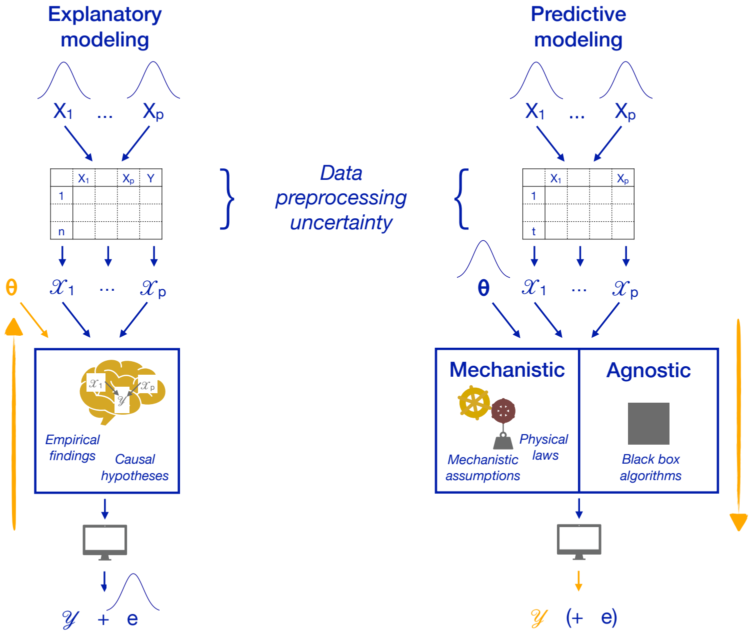
- Choose methods to run and to analyze simulations

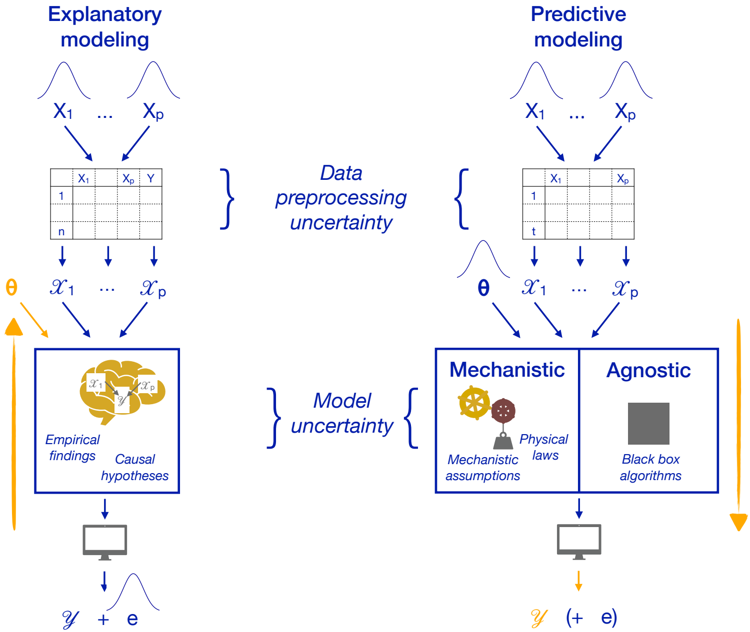


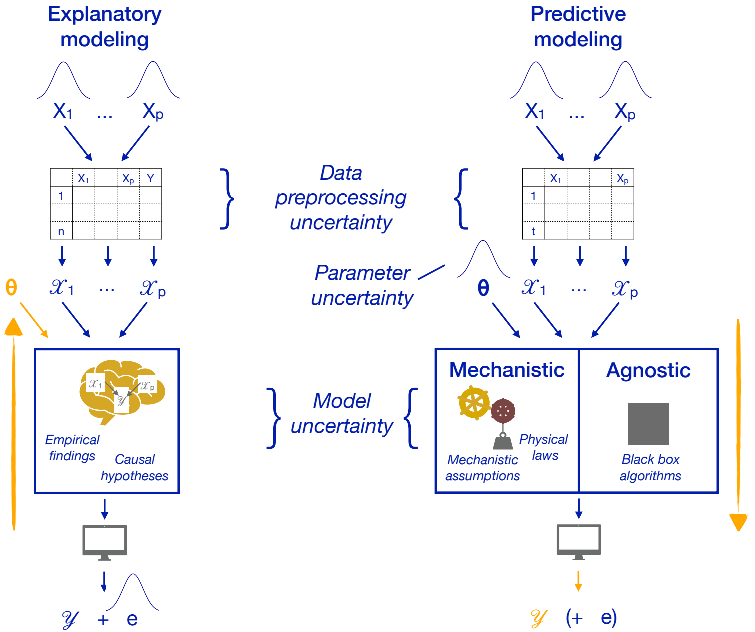
θ θ θ θ θ

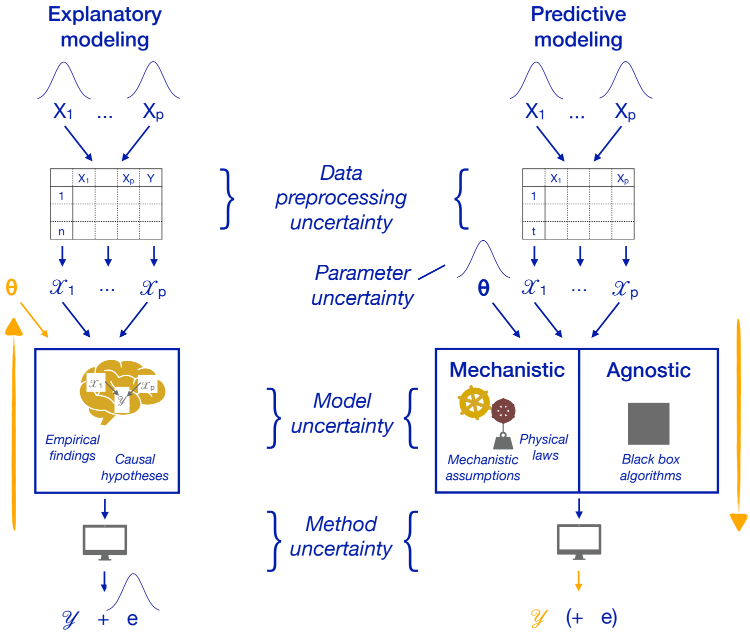
y y y y y

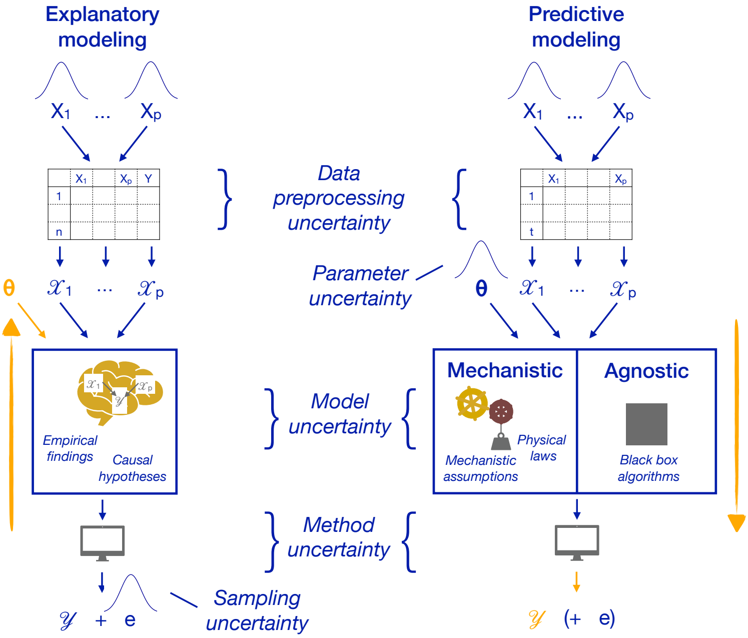

Explanatory modeling
X_1 … X_p
X_1 X_p Y
1
n
θ
$\mathscr{X}_1$ … $\mathscr{X}_p$
Empirical findings
Causal hypotheses
$\mathscr{Y}$ + e
Predictive modeling
X_1 … X_p
X_1 X_p
1
t
θ
$\mathscr{X}_1$ … $\mathscr{X}_p$
Mechanistic
Agnostic
Physical laws
Mechanistic assumptions
Black box algorithms
$\mathscr{Y}$ (+ e)

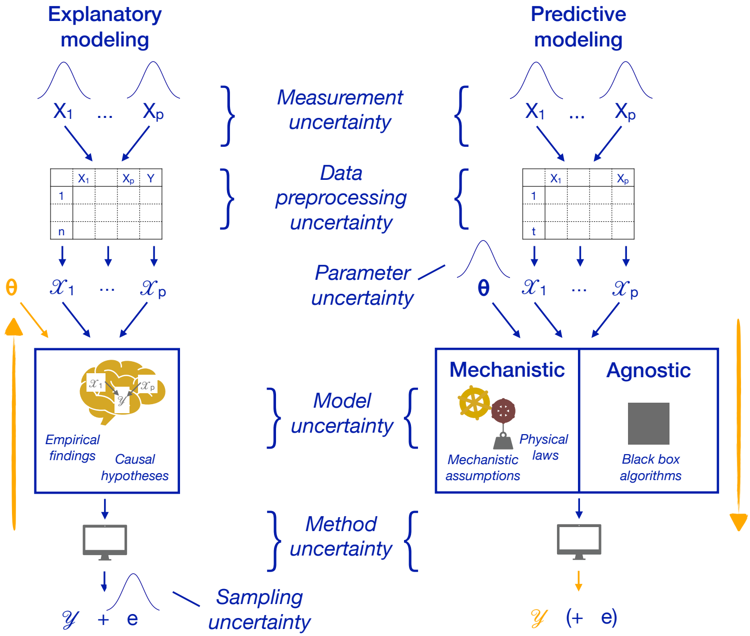

Explanatory modeling
Predictive modeling
X1 ... Xp
X1 Xp Y
1
n
X1 ... Xp
X1 Xp
1
t
Data preprocessing uncertainty
θ
𝒳1 ... 𝒳p
θ
𝒳1 ... 𝒳p
Empirical findings
Causal hypotheses
Mechanistic
Agnostic
Physical laws
Mechanistic assumptions
Black box algorithms
𝒴 + e
𝒴 (+ e)


Explanatory modeling
Predictive modeling
X1 ... Xp
X1 Xp Y
1
n
X1 Xp
1
t
Data preprocessing uncertainty
θ
𝒳1 ... 𝒳p
Empirical findings
Causal hypotheses
Model uncertainty
Mechanistic
Agnostic
Physical laws
Mechanistic assumptions
Black box algorithms
𝒴 + e
𝒴 (+ e)

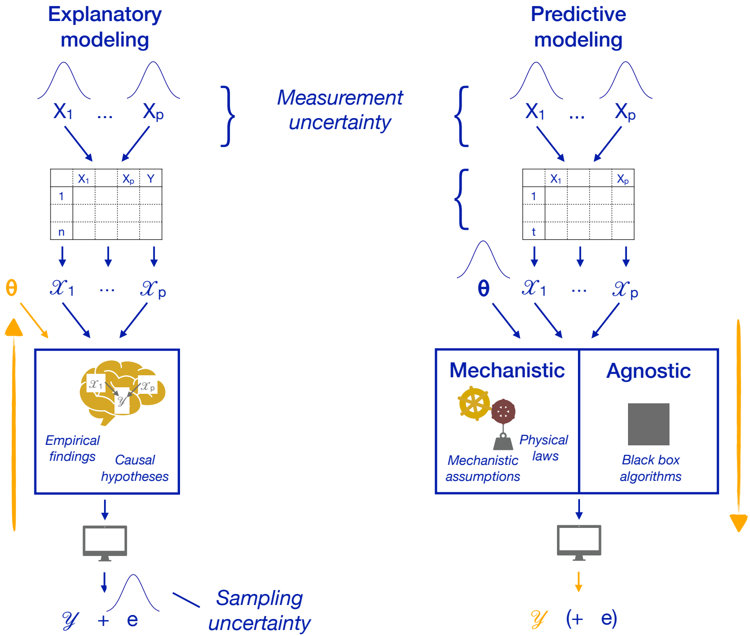

Explanatory modeling
Predictive modeling
X1 ... Xp
X1 Xp Y
1
n
Data preprocessing uncertainty
X1 Xp
1
t
Parameter uncertainty
θ
𝒳1 ... 𝒳p
Model uncertainty
Empirical findings
Causal hypotheses
Mechanistic
Agnostic
Physical laws
Mechanistic assumptions
Black box algorithms
𝒴 + e
𝒴 (+ e)

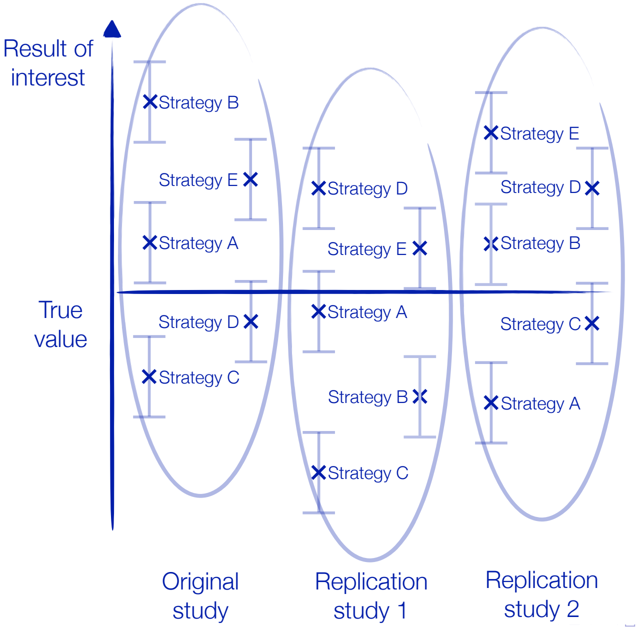

Explanatory modeling
Predictive modeling
X1 ... Xp
X1 Xp Y
1
n
X1 Xp
1
t
Data preprocessing uncertainty
Parameter uncertainty
θ
𝒳1 ... 𝒳p
Model uncertainty
Empirical findings
Causal hypotheses
Mechanistic
Agnostic
Physical laws
Mechanistic assumptions
Black box algorithms
Method uncertainty
𝒴 + e
𝒴 (+ e)

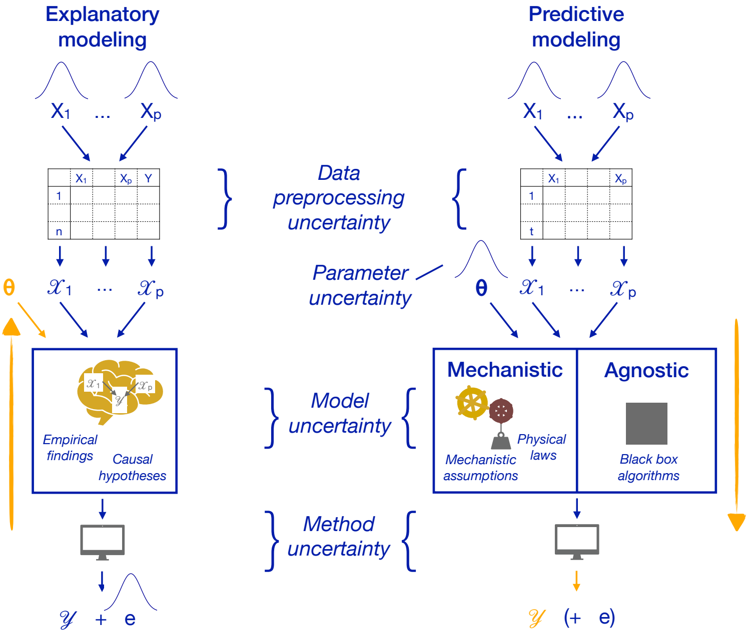

Explanatory modeling
Predictive modeling
X1 ... Xp
X1 Xp Y
1
n
X1 Xp
1
t
Data preprocessing uncertainty
Parameter uncertainty
θ
𝒳1 ... 𝒳p
Empirical findings
Causal hypotheses
Mechanistic
Physical laws
Mechanistic assumptions
Agnostic
Black box algorithms
Model uncertainty
Method uncertainty
𝒴 + e
Sampling uncertainty
𝒴 (+ e)


Explanatory modeling
Predictive modeling
X1 ... Xp
Measurement uncertainty
X1 Xp Y
1
n
t
Data preprocessing uncertainty
θ
Parameter uncertainty
Empirical findings
Causal hypotheses
Mechanistic
Agnostic
Physical laws
Mechanistic assumptions
Black box algorithms
Model uncertainty
Method uncertainty
Sampling uncertainty
+ e
(+ e)

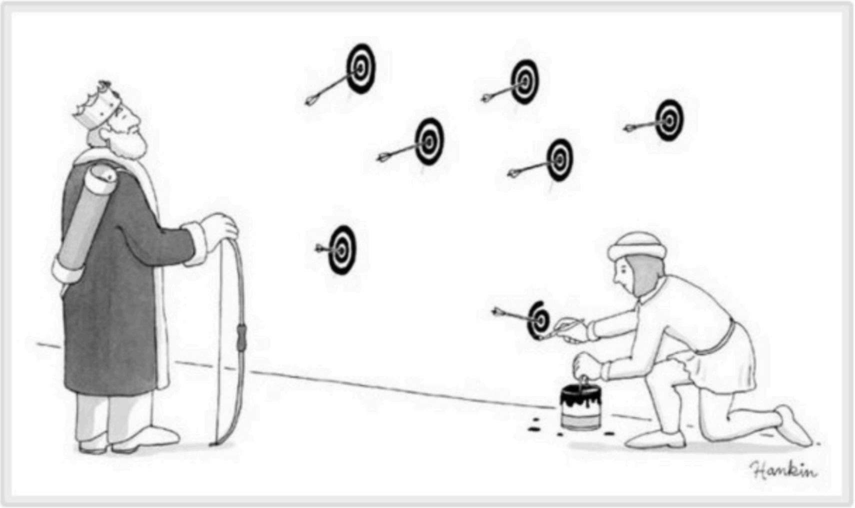
Impact on the replicability of research findings


Explanatory modeling
Predictive modeling
X1 ... Xp
Measurement uncertainty
X1 Xp Y
1
n
X1 Xp
t
θ
𝒳1 ... 𝒳p
Empirical findings
Causal hypotheses
Mechanistic
Agnostic
Physical laws
Mechanistic assumptions
Black box algorithms
𝒴 + e
Sampling uncertainty
𝒴 (+ e)

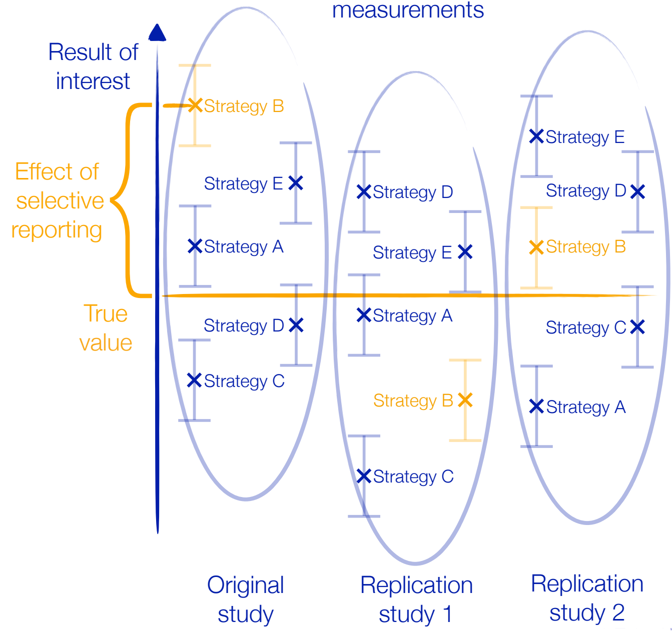

Result of interest
True value
Strategy B
Strategy E
Strategy A
Strategy D
Strategy C
Strategy D
Strategy E
Strategy A
Strategy B
Strategy C
Strategy E
Strategy D
Strategy B
Strategy C
Strategy A
Original study
Replication study 1
Replication study 2


Explanatory modeling
Predictive modeling
X_1 ... X_p
X_1 ... X_p
X_1 X_p Y
1
n
X_1 X_p
1
t
Data preprocessing uncertainty
Parameter uncertainty
θ $\mathcal{X}_1$ ... $\mathcal{X}_p$
θ $\mathcal{X}_1$ ... $\mathcal{X}_p$
Empirical findings
Causal hypotheses
Mechanistic
Agnostic
Physical laws
Mechanistic assumptions
Black box algorithms
Model uncertainty
Method uncertainty
$\mathcal{Y}$ + e
$\mathcal{Y}$ (+ e)

Selective reporting of analyses strategies



Small sample size and imprecise measurements

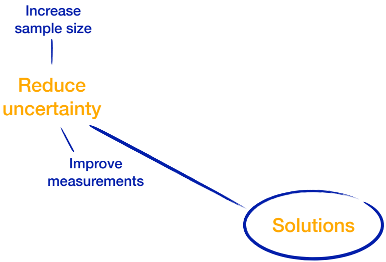


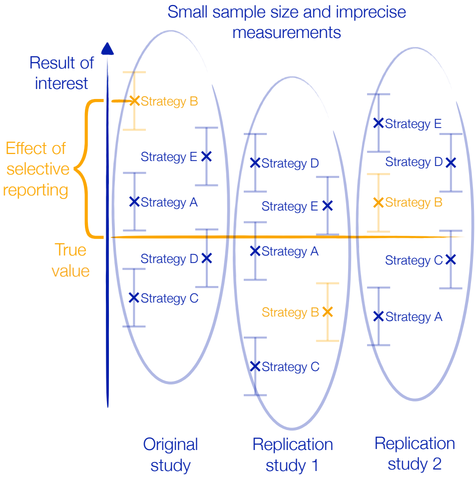
Lessons learned across discisplines

Solutions

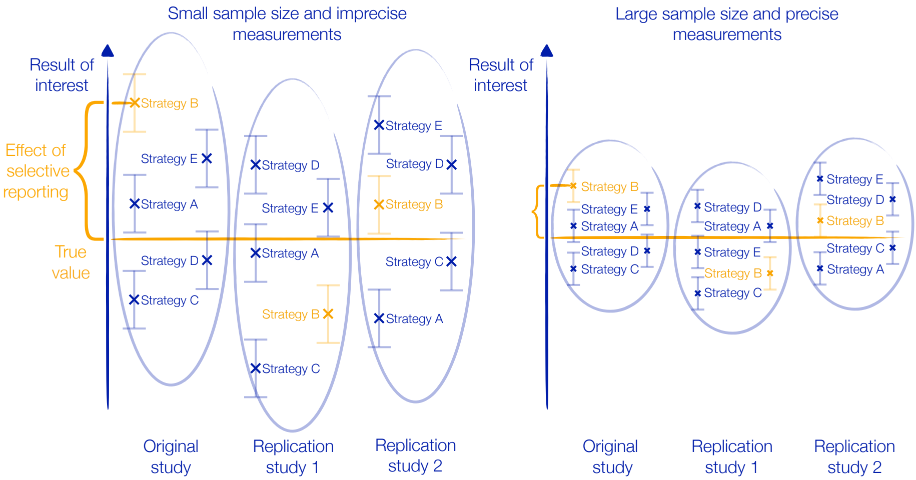
Reduce uncertainty

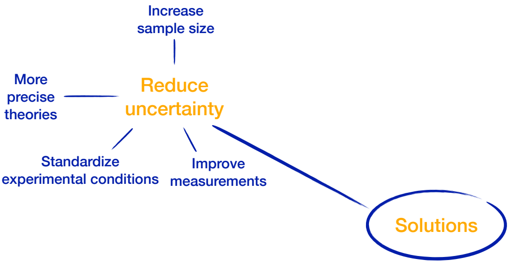
Solutions


Increase
sample size
Reduce
uncertainty
Improve
measurements
Solutions



Small sample size and imprecise measurements
Result of interest
Effect of selective reporting
True value
Strategy B
Strategy E
Strategy A
Strategy D
Strategy C
Strategy D
Strategy E
Strategy A
Strategy B
Strategy C
Strategy E
Strategy D
Strategy B
Strategy C
Strategy A
Original study
Replication study 1
Replication study 2


Small sample size and imprecise measurements
Result of interest
Effect of selective reporting
True value
Strategy B
Strategy E
Strategy A
Strategy D
Strategy C
Original study
Replication study 1
Replication study 2
Large sample size and precise measurements
Result of interest

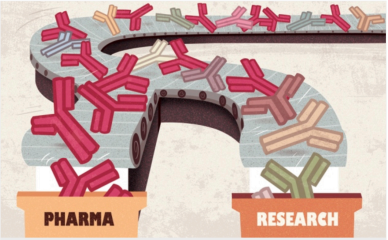

Increase
sample size
More
precise
theories
Reduce
uncertainty
Standardize
experimental conditions
Improve
measurements
Solutions

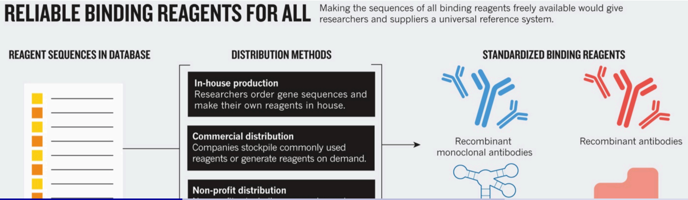
[Bradbury and Plückthun, 2015]: Standardize experimental conditions

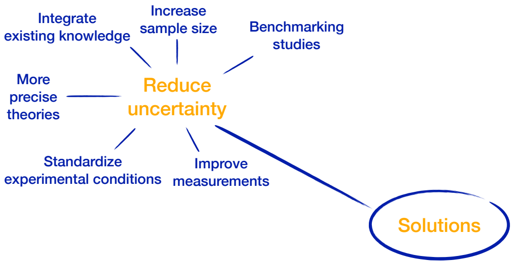

BLAME IT ON THE ANTIBODIES

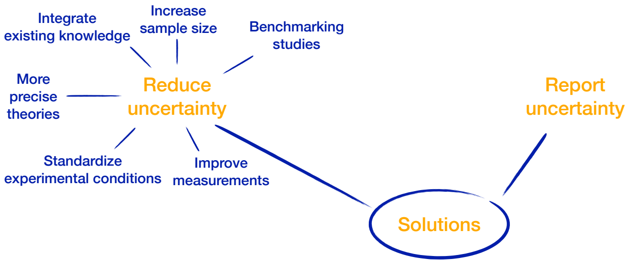
Antibodies are the workhorses of biological experiments, but they are littering the field with false findings. A few evangelists are pushing for change.

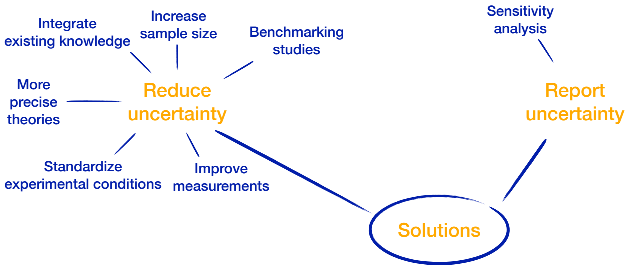
BY MONYA BAKER

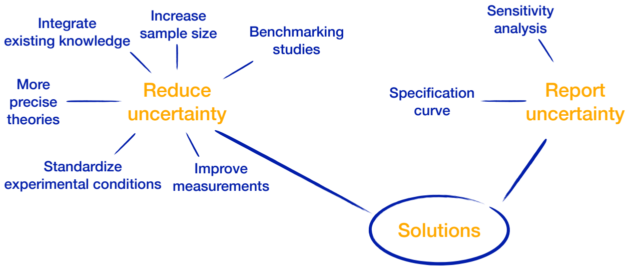




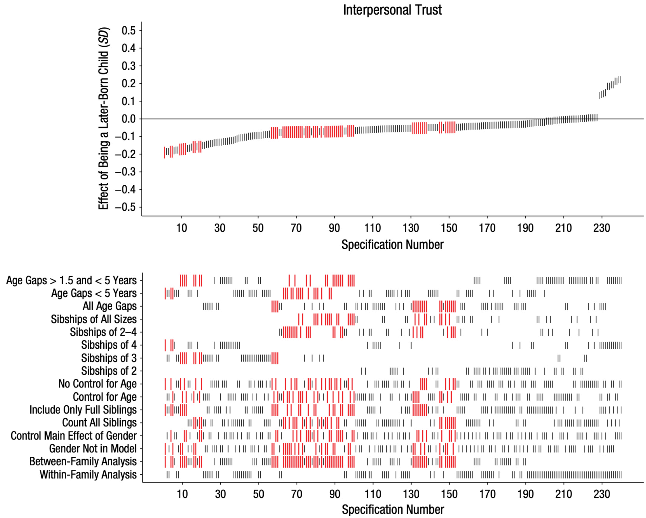

Integrate
existing knowledge
Increase
sample size
Benchmarking
studies
More
precise
theories
Reduce
uncertainty
Standardize
experimental conditions
Improve
measurements
Solutions


Integrate
existing knowledge
Increase
sample size
Benchmarking
studies
More
precise
theories
Reduce
uncertainty
Report
uncertainty
Standardize
experimental conditions
Improve
measurements
Solutions

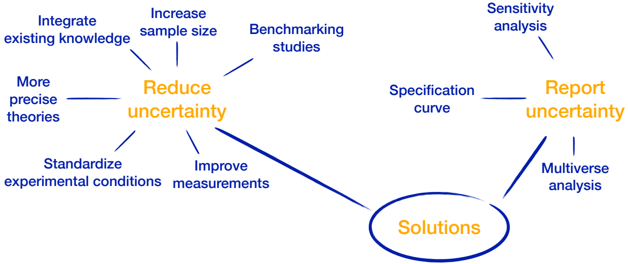

Integrate
existing knowledge
Increase
sample size
Benchmarking
studies
More
precise
theories
Reduce
uncertainty
Standardize
experimental conditions
Improve
measurements
Sensitivity
analysis
Report
uncertainty
Solutions


Solutions
Reduce uncertainty
Integrate existing knowledge
Increase sample size
Benchmarking studies
More precise theories
Standardize experimental conditions
Improve measurements
Report uncertainty
Sensitivity analysis
Specification curve

Specification curve analysis [Simonsohn et al., 2020]



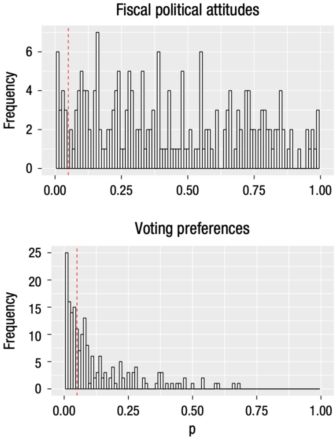
[Rohrer et al., 2017]


Integrate existing knowledge
Increase sample size
Benchmarking studies
More precise theories
Reduce uncertainty
Standardize experimental conditions
Improve measurements
Sensitivity analysis
Specification curve
Report uncertainty
Multiverse analysis
Solutions

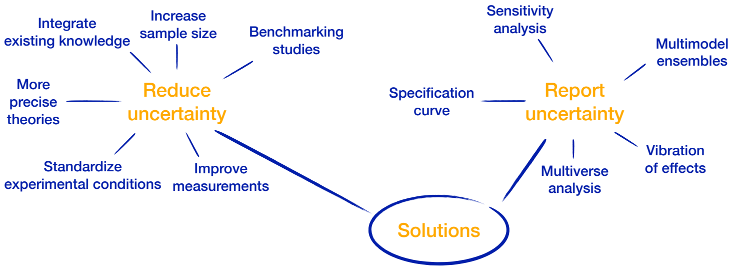
Reporting data pre-processing uncertainty

Table 1. Processing choices

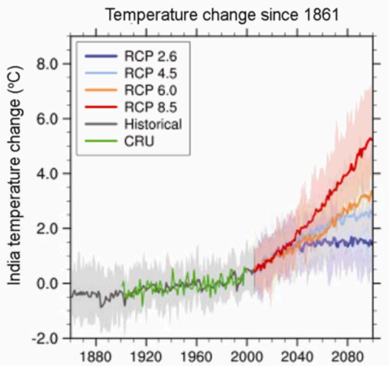
1. Assessment of fertility (F)—high vs low.
   (a) F1: high = cycle days 7–14; low = cycle days 17–25
   (b) F2: high = cycle days 6–14; low = cycle days 17–27
   (c) F3: high = cycle days 9–17; low = cycle days 18–25
   (d) F4: high = cycle days 8–14; low = cycle days 1–7 and 15–28
   (e) F5: high = cycle days 9–17; low = cycle days 1–8 and 18–28
2. Next menstrual onset (NMO)
   (a) NMO1: reported start date previous menstrual onset + computed cycle length
   (b) NMO2: reported start date previous menstrual onset + reported cycle length
   (c) NMO3: reported estimate of next menstrual onset
3. Assessment of relationship status (R) (single vs relationship)
   (a) R1: single = response options 1 and 2; relationship = response options 3 and 4
   (b) R2: single = response option 1; relationship = response options 2, 3, and 4
   (c) R3: single = response option 1; relationship = response options 3 and 4
4. Exclusion of women based on cycle length (ECL)
   (a) ECL1: no exclusion based on cycle length
   (b) ECL2: exclusion of participants with computed cycle length greater than 25 or less than 35 days
   (c) ECL3: exclusion of participants with reported cycle length greater than 25 or less than 35 days
5. Exclusion of women based on certainty ratings of start dates of two previous menstrual periods (EC)
   (a) EC1: no exclusion based on certainty ratings
   (b) EC2: exclusion of participants who are not certain about at least one start date (i.e., sure less than 6)



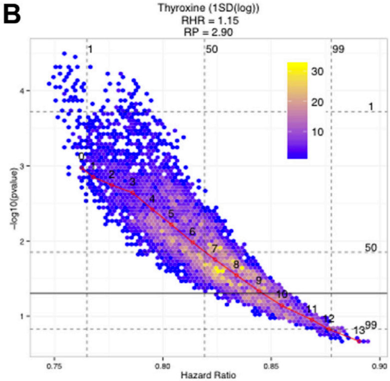
Multiverse analysis [Steegen et al., 2016]


Integrate existing knowledge
Increase sample size
Benchmarking studies
More precise theories
Reduce uncertainty
Standardize experimental conditions
Improve measurements
Sensitivity analysis
Multimodel ensembles
Specification curve
Report uncertainty
Multiverse analysis
Vibration of effects
Solutions

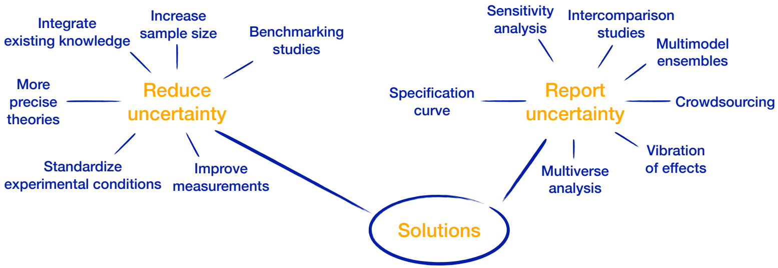
Reporting model uncertainty



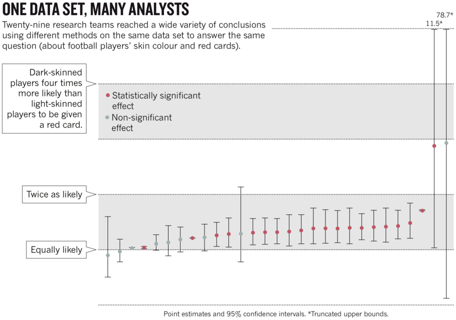
Multi-model projections
[Chaturvedi et al., 2012]

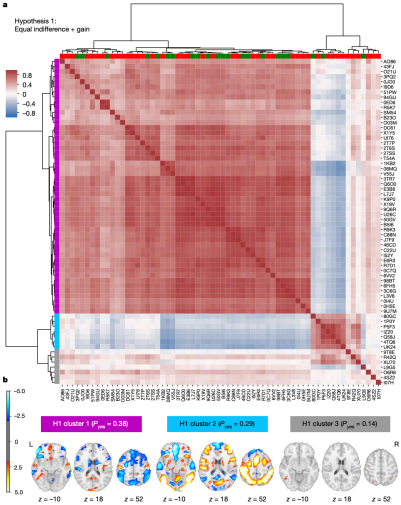


Vibration of effects
[Patel et al., 2015]

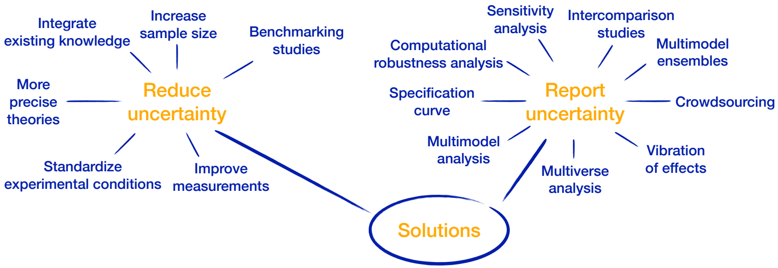

Integrate
existing knowledge
Increase
sample size
Benchmarking
studies
More
precise
theories
Reduce
uncertainty
Standardize
experimental conditions
Improve
measurements
Sensitivity
analysis
Intercomparison
studies
Multimodel
ensembles
Specification
curve
Report
uncertainty
Crowdsourcing
Multiverse
analysis
Vibration
of effects
Solutions

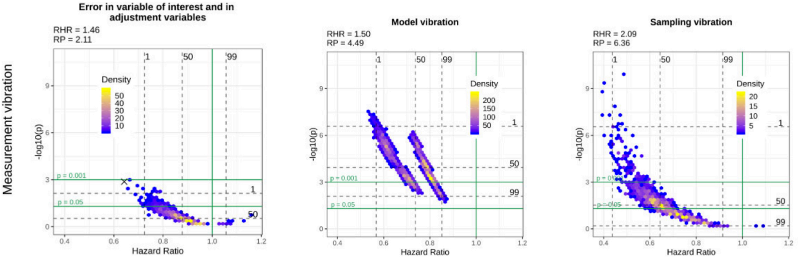
Crowdsouring





[Silberzahn and Uhlmann, 2015] [Botvinik-Nezer et al., 2020]


Integrate
existing knowledge
Increase
sample size
Benchmarking
studies
More
precise
theories
Reduce
uncertainty
Standardize
experimental conditions
Improve
measurements
Sensitivity
analysis
Intercomparison
studies
Computational
robustness analysis
Multimodel
ensembles
Specification
curve
Report
uncertainty
Crowdsourcing
Multimodel
analysis
Multiverse
analysis
Vibration
of effects
Solutions

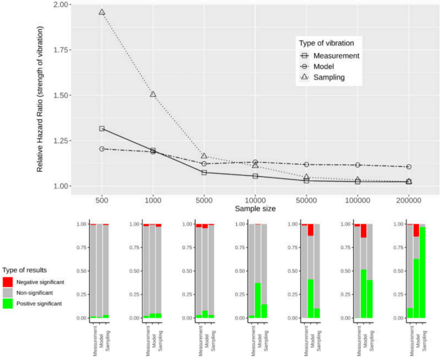


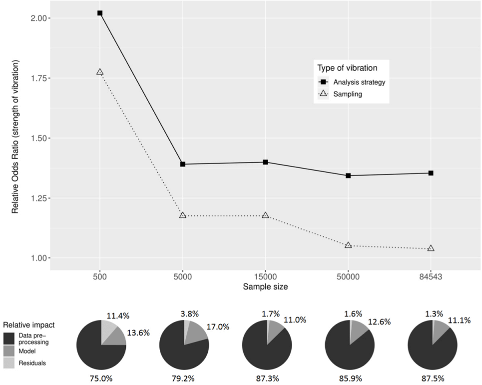
International Journal of Epidemiology, 2020, 1–13
doi: 10.1093/ije/dyaa164
Original Article

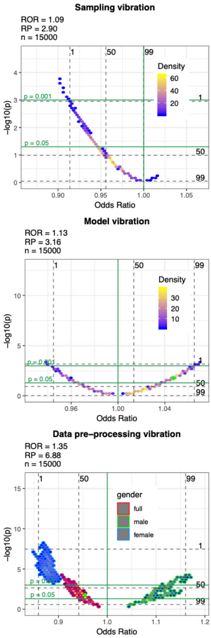
Original Article

Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework

Simon Klau[1,2,*,†] **Sabine Hoffmann,**[1,3†] **Chirag J Patel,**[4]
John P A Ioannidis,[5,6,7,8,9] **and Anne-Laure Boulesteix**[1,3]



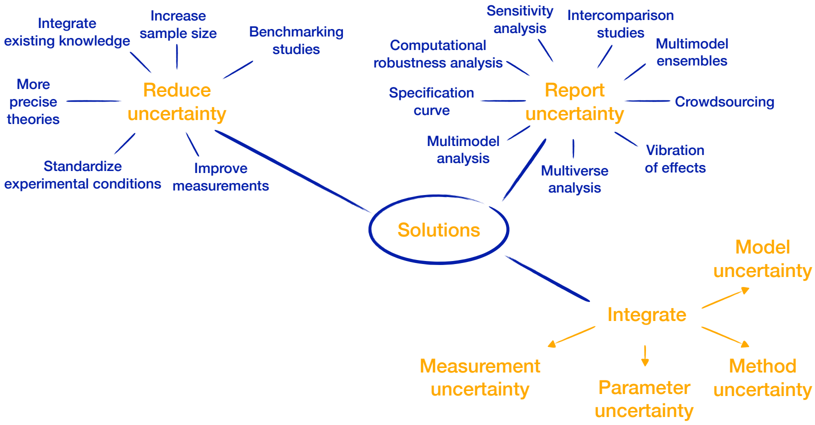


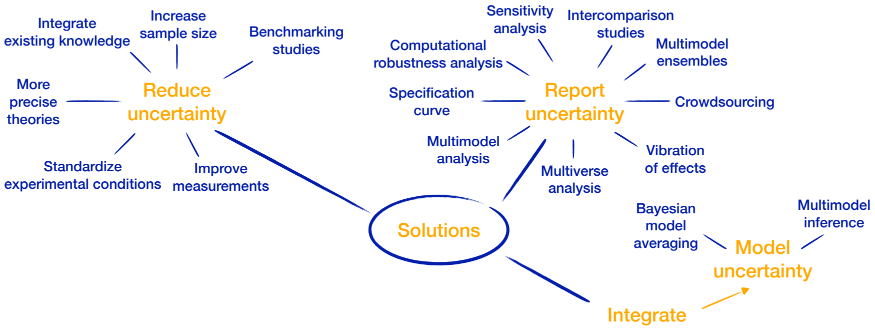


LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN INSTITUT FÜR STATISTIK

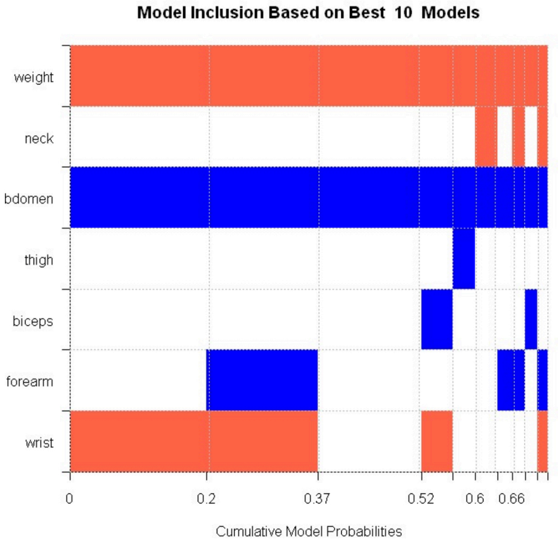
Simon Klau, Felix Schönbrodt, Chirag Patel, John Ioannidis, Anne-Laure Boulesteix, Sabine Hoffmann

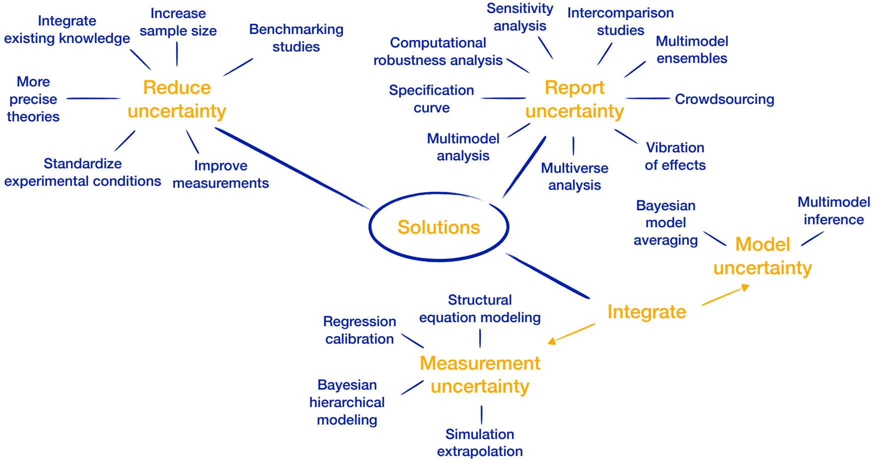
Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology


Solutions
Reduce uncertainty
Integrate existing knowledge
Increase sample size
Benchmarking studies
More precise theories
Standardize experimental conditions
Improve measurements
Report uncertainty
Sensitivity analysis
Intercomparison studies
Computational robustness analysis
Multimodel ensembles
Specification curve
Crowdsourcing
Multimodel analysis
Multiverse analysis
Vibration of effects
Integrate
Model uncertainty
Measurement uncertainty
Parameter uncertainty
Method uncertainty

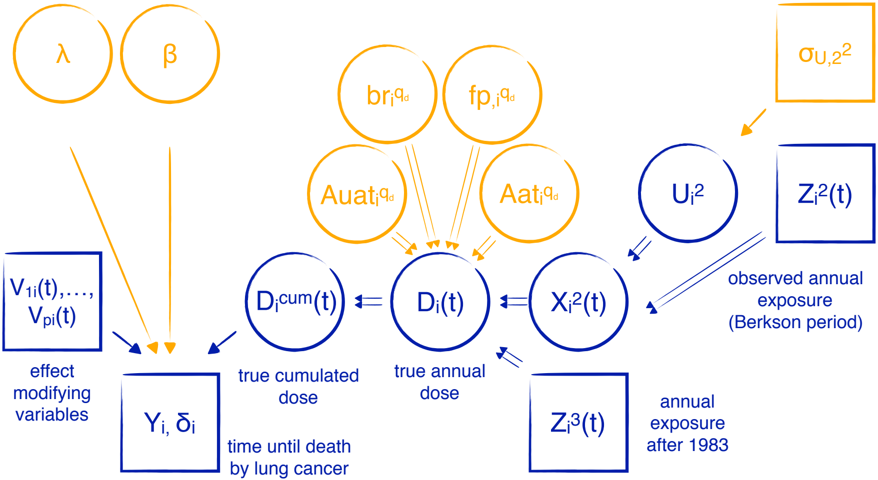

Solutions
Reduce uncertainty
Integrate existing knowledge
Increase sample size
Benchmarking studies
More precise theories
Standardize experimental conditions
Improve measurements
Report uncertainty
Sensitivity analysis
Intercomparison studies
Computational robustness analysis
Multimodel ensembles
Specification curve
Crowdsourcing
Multimodel analysis
Multiverse analysis
Vibration of effects
Integrate
Model uncertainty
Bayesian model averaging
Multimodel inference

Bayesian model averaging

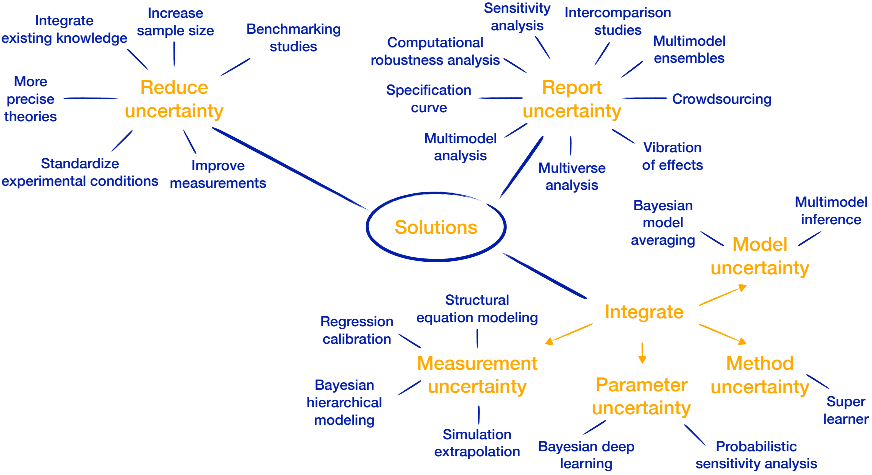


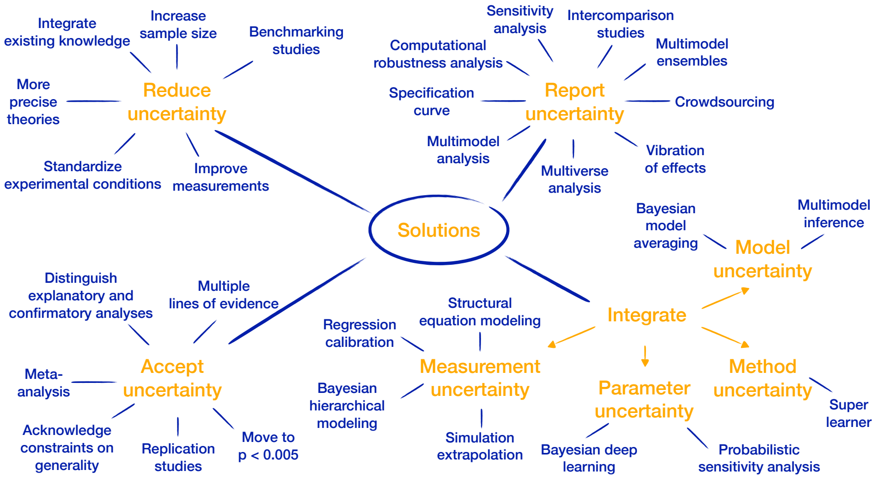

Integrate existing knowledge
Increase sample size
Benchmarking studies
More precise theories
Reduce uncertainty
Standardize experimental conditions
Improve measurements
Sensitivity analysis
Intercomparison studies
Computational robustness analysis
Multimodel ensembles
Specification curve
Report uncertainty
Crowdsourcing
Multimodel analysis
Multiverse analysis
Vibration of effects
Solutions
Bayesian model averaging
Multimodel inference
Model uncertainty
Integrate
Structural equation modeling
Regression calibration
Measurement uncertainty
Bayesian hierarchical modeling
Simulation extrapolation

Accounting for complex patterns of measurement error



Joint work with Raphael Rehms and Nicole Ellenbach


Solutions
Reduce uncertainty
Integrate existing knowledge
Increase sample size
Benchmarking studies
More precise theories
Standardize experimental conditions
Improve measurements
Report uncertainty
Sensitivity analysis
Intercomparison studies
Computational robustness analysis
Multimodel ensembles
Specification curve
Crowdsourcing
Multimodel analysis
Multiverse analysis
Vibration of effects
Integrate
Model uncertainty
Bayesian model averaging
Multimodel inference
Measurement uncertainty
Structural equation modeling
Regression calibration
Bayesian hierarchical modeling
Simulation extrapolation
Parameter uncertainty
Bayesian deep learning
Probabilistic sensitivity analysis
Method uncertainty
Super learner


Solutions
Reduce uncertainty
Integrate existing knowledge
Increase sample size
Benchmarking studies
More precise theories
Standardize experimental conditions
Improve measurements
Report uncertainty
Sensitivity analysis
Intercomparison studies
Computational robustness analysis
Multimodel ensembles
Specification curve
Crowdsourcing
Multimodel analysis
Multiverse analysis
Vibration of effects
Accept uncertainty
Distinguish explanatory and confirmatory analyses
Multiple lines of evidence
Meta-analysis
Acknowledge constraints on generality
Replication studies
Move to p < 0.005
Integrate
Model uncertainty
Bayesian model averaging
Multimodel inference
Measurement uncertainty
Regression calibration
Structural equation modeling
Bayesian hierarchical modeling
Simulation extrapolation
Parameter uncertainty
Bayesian deep learning
Probabilistic sensitivity analysis
Method uncertainty
Super learner

Recommendations

Step 1:
Be aware of the multiplicity of possible analysis strategies

Recommendations

Step 1:
Be aware of the multiplicity of possible analysis strategies

Step 2:
If possible, reduce sources of uncertainty before the analysis

Recommendations

Step 1:
Be aware of the multiplicity of possible analysis strategies

Step 2:
If possible, reduce sources of uncertainty before the analysis

Step 3a:
If possible, integrate remaining sources of uncertainty into the analysis

Recommendations

Step 1:
Be aware of the multiplicity of possible analysis strategies

Step 2:
If possible, reduce sources of uncertainty before the analysis

Step 3a:
If possible, integrate remaining sources of uncertainty into the analysis

Step 3b:
Report the results of alternative analysis strategies to assess the robustness of results

Recommendations

Step 1:
Be aware of the multiplicity of possible analysis strategies

Step 2:
If possible, reduce sources of uncertainty before the analysis

Step 3a:
If possible, integrate remaining sources of uncertainty into the analysis

Step 3b:
Report the results of alternative analysis strategies to assess the robustness of results

Step 4:
Acknowledge the inherent uncertainty in your findings

Recommendations

Step 1:
Be aware of the multiplicity of possible analysis strategies

Step 2:
If possible, reduce sources of uncertainty before the analysis

Step 3a:
If possible, integrate remaining sources of uncertainty into the analysis

Step 3b:
Report the results of alternative analysis strategies to assess the robustness of results

Step 4:
Acknowledge the inherent uncertainty in your findings

Step 5:
Publish all research code, data and material

Conclusion

- Multidisciplinary efforts are essential to avoid reinventing the wheel in every discipline and to generating enough momentum to bring about change

Conclusion

- Multidisciplinary efforts are essential to avoid reinventing the wheel in every discipline and to generating enough momentum to bring about change

  "The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines" by S. Hoffmann, F. Schönbrodt, R. Elsas, R. Wilson, U. Strasser and A. Boulesteix, available on Meta-Arxiv, preprint DOI: `10.31222/osf.io/afb9p`

Conclusion

- Multidisciplinary efforts are essential to avoid reinventing the wheel in every discipline and to generating enough momentum to bring about change
  "The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines" by S. Hoffmann, F. Schönbrodt, R. Elsas, R. Wilson, U. Strasser and A. Boulesteix, available on Meta-Arxiv, preprint DOI: `10.31222/osf.io/afb9p`
- Increasing amounts of data that are not recorded for research in many disciplines

Conclusion

- Multidisciplinary efforts are essential to avoid reinventing the wheel in every discipline and to generating enough momentum to bring about change
  "The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines" by S. Hoffmann, F. Schönbrodt, R. Elsas, R. Wilson, U. Strasser and A. Boulesteix, available on Meta-Arxiv, preprint DOI: `10.31222/osf.io/afb9p`
- Increasing amounts of data that are not recorded for research in many disciplines
- Reproducibility and transparency as first steps to increase the replicability and credibility of research findings

Thank you for your attention!

Anaya, J., van der Zee, T., and Brown, N. (2017).
Statistical infarction: A postmortem of the cornell food and brand lab pizza publications.
Technical report, PeerJ Preprints.

Begg, C. B. and Mazumdar, M. (1994).
Operating characteristics of a rank correlation test for publication bias.
Biometrics, pages 1088–1101.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., et al. (2020).
Variability in the analysis of a single neuroimaging dataset by many teams.
Nature, pages 1–7.

Boulesteix, A.-L., Hornung, R., and Sauerbrei, W. (2017).
On fishing for significance and statistician‚Äôs degree of freedom in the era of big molecular data.

In *Berechenbarkeit der Welt?*, pages 155–170. Springer.

Bradbury, A. and Plückthun, A. (2015).
Reproducibility: Standardize antibodies used in research.
Nature News, 518(7537):27–29.

Chandler, R. E., Thorne, P., Lawrimore, J., and Willett, K. (2012).
Building trust in climate science: data products for the 21st century.
Environmetrics, 23(5):373–381.

Chaturvedi, R. K., Joshi, J., Jayaraman, M., Bala, G., and Ravindranath, N. (2012).
Multi-model climate change projections for india under representative concentration pathways.
Current Science, 103(7):791–802.

Easterbrook, P. J., Gopalan, R., Berlin, J., and Matthews, D. R. (1991).
Publication bias in clinical research.
The Lancet, 337(8746):867–872.

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015).
The economics of reproducibility in preclinical research.
PLoS Biol, 13(6):e1002165.

Gelman, A. and Loken, E. (2014).
The statistical crisis in science data-dependent analysis—a “garden of forking paths”—explains why many statistically significant comparisons don’t hold up.
American Scientist, 102(6):460–465.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016).
What does research reproducibility mean?
Science translational medicine, 8(341):341ps12–341ps12.

Hejblum, B. P., Kunzmann, K., Lavagnini, E., Hutchinson, A., Robertson, D. S., Jones, S. C., and Eckes-Shephard, A. H. (2020).
Realistic and robust reproducible research for biostatistics.

Ince, D. (2011).
The duke university scandal—what can be done?

Significance, 8(3):113–115.

Ioannidis, J. P. (2005a).
Microarrays and molecular research: noise discovery?
Lancet (London, England), 365(9458):454–455.

Ioannidis, J. P. A. (2005b).
Why most published research findings are false.
PLoS Med, 2(8):e124.

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

Patel, C. J., Burford, B., and Ioannidis, J. P. A. (2015).
Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations.
J Clin Epidemiol, 68(9):1046–58.

Rohrer, J. M., Egloff, B., and Schmukle, S. C. (2017).

Probing birth-order effects on narrow traits using specification-curve analysis.
Psychol Sci, 28(12):1821–1832.

Silberzahn, R. and Uhlmann, E. L. (2015).
Crowdsourced research: Many hands make tight work.
Nature, 526(7572):189–91.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020).
Specification curve analysis.
Nature Human Behaviour, pages 1–7.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016).
Increasing transparency through a multiverse analysis.
Perspect Psychol Sci, 11(5):702–712.

Sterling, T. D. (1959).
Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa.
Journal of the American Statistical Association, 54(285):30–34.