



University of  
Zurich UZH



Philipps  
Universität  
Marburg



Swiss National  
Science Foundation

# Simulation Studies for Methodological Research: State of the Art, Issues, and Potential Solutions

---

**Samuel Pawel<sup>1</sup>    Björn Siepe<sup>2</sup>**

April 17, 2025 – Berlin Methods Colloquium

<sup>1</sup>Department of Biostatistics, Center for Reproducible Science, University of Zurich

<sup>2</sup>Department of Psychology, University of Marburg

# Agenda

Introduction

Questionable research practices in simulation studies

Simulation studies in Psychology

Potential improvements

Discussion

# Introduction

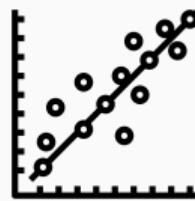
---

# Quantitative methodological research

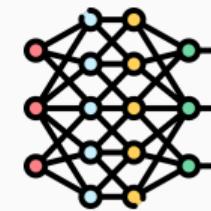
- **Diverse fields:** Statistics, psychometrics, bioinformatics, ecology, econometrics, machine learning, ...

# Quantitative methodological research

- **Diverse fields:** Statistics, psychometrics, bioinformatics, ecology, econometrics, machine learning, ...
- Common question: **Which data analysis methods work well when?**

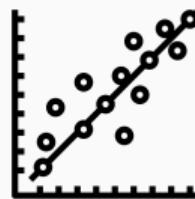


vs.



# Quantitative methodological research

- **Diverse fields:** Statistics, psychometrics, bioinformatics, ecology, econometrics, machine learning, ...
- Common question: **Which data analysis methods work well when?**

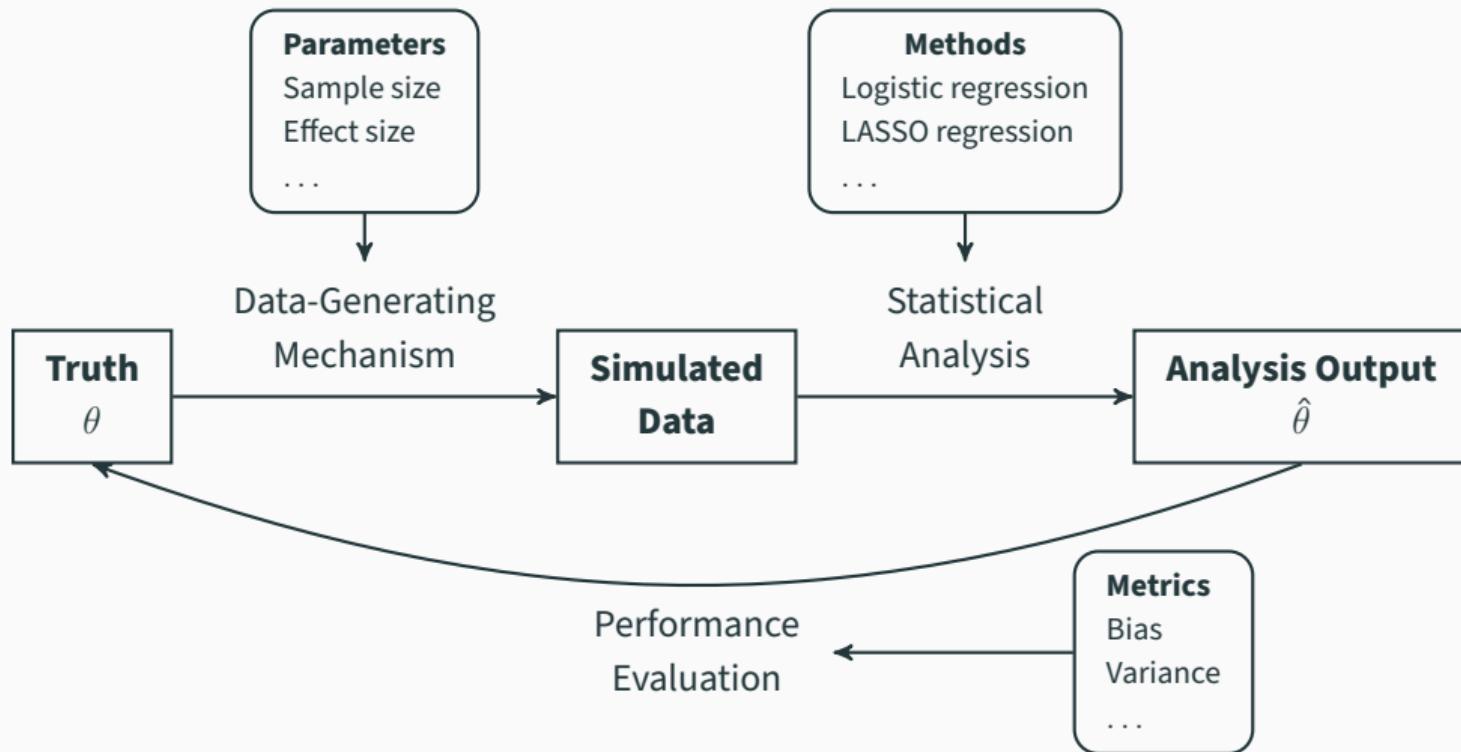


vs.



- Tools:
  - Formal analysis and **mathematical proofs** → theory
  - Application to **real data sets** → case studies
  - **Simulation studies** → controlled experiments

# Simulation studies



# Simulation studies are commonly used

Journal	Article contains simulation study
Journal of the American Statistical Association	186/200 = <b>93%</b>
Statistics in Medicine	104/115 = <b>90%</b>
Psychological Methods	98/179 = <b>55%</b>
Research Synthesis Methods	94/306 = <b>31%</b>

Literature review from Pawel et al. (2024a)

# Simulation studies can be influential

Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives

L Hu, PM Bentler - Structural equation modeling: a ..., 1999 - Taylor & Francis

This article examines the adequacy of the "rules of thumb" conventional cutoff criteria and several new alternatives for various fit indexes used to evaluate model fit in practice. Using a 2...

☆ Save ⚡ Cite Cited by 116305 Related articles All 9 versions

A **simulation study** of the number of events per variable in logistic regression analysis

P Peduzzi, J Concato, E Kemper, TR Holford... - Journal of clinical ..., 1996 - Elsevier

... In a **simulation study** of forward stepwise multiple linear regression, Freedman and Pee [3] demonstrated that the ... In **simulation studies** of the effect of EPV on proportional ... Peter Peduzzi. ...

☆ Save ⚡ Cite Cited by 8827 Related articles All 9 versions

# Simulation studies impact implementation of research

Post-anaesthesia pulmonary complications after use of muscle relaxants (POPULAR): a multicentre, prospective observational study

E Kirmeier, LI Eriksson, H Lewald... - *The Lancet* ..., 2019 - [thelancet.com](https://www.thelancet.com)

Background Results from retrospective studies suggest that use of neuromuscular blocking agents during general anaesthesia might be linked to postoperative pulmonary ...

☆ Save 99 Cite Cited by 303 Related articles All 37 versions

## Statistical analysis

Sample size was estimated using the rule of ten.<sup>19</sup>

Sample size =

$$\frac{10 \times \text{number of factors and cofactors}}{\text{Incidence of postoperative pulmonary complications}}$$

19 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49**: 1373–79.

# There can be problems with simulation studies

van Smeden et al. BMC Medical Research Methodology (2016) 16:163  
DOI 10.1186/s12874-016-0267-3

BMC Medical Research  
Methodology

RESEARCH ARTICLE

Open Access



CrossMark

## No rationale for 1 variable per 10 events criterion for binary logistic regression analysis

Maarten van Smeden<sup>1\*</sup> Joris A. H. de Groot<sup>1</sup>, Karel G. M. Moons<sup>1</sup>, Gary S. Collins<sup>2</sup>,  
Douglas G. Altman<sup>2</sup>, Marinus J. C. Eijkemans<sup>1</sup> and Johannes B. Reitsma<sup>1</sup>

“The current **evidence supporting [the rule of ten] is weak** [...] there is an urgent need for new research to provide guidance for supporting sample size considerations for binary logistic regression” van Smeden et al. (2016)

## Related new arXiv preprint

### **Handling Missingness, Failures, and Non-Convergence in Simulation Studies: A Review of Current Practices and Recommendations**

Samuel Pawel  <sup>1</sup>, František Bartoš  <sup>2,\*</sup>, Björn S. Siepe  <sup>3,\*</sup>, Anna Lohmann  <sup>4,5,\*</sup>

## Related new arXiv preprint

### Handling Missingness, Failures, and Non-Convergence in Simulation Studies: A Review of Current Practices and Recommendations

Samuel Pawel <sup>1</sup>, František Bartoš <sup>2,\*</sup>, Björn S. Siepe <sup>3,\*</sup>, Anna Lohmann <sup>4,5,\*</sup>

- Review of 482 simulation studies published in JASA, SiM, PM, RSM:
  - **23.0%** mention missingness / failures / non-convergence

### Handling Missingness, Failures, and Non-Convergence in Simulation Studies: A Review of Current Practices and Recommendations

Samuel Pawel <sup>1</sup>, František Bartoš <sup>2,\*</sup>, Björn S. Siepe <sup>3,\*</sup>, Anna Lohmann <sup>4,5,\*</sup>

- Review of 482 simulation studies published in JASA, SiM, PM, RSM:
  - **23.0%** mention missingness / failures / non-convergence
  - **19.1%** report frequency
  - **13.9%** report handling
  - **46.7%** share code

### Handling Missingness, Failures, and Non-Convergence in Simulation Studies: A Review of Current Practices and Recommendations

Samuel Pawel  <sup>1</sup>, František Bartoš  <sup>2,\*</sup>, Björn S. Siepe  <sup>3,\*</sup>, Anna Lohmann  <sup>4,5,\*</sup>

- Review of 482 simulation studies published in JASA, SiM, PM, RSM:
  - **23.0%** mention missingness / failures / non-convergence
  - **19.1%** report frequency
  - **13.9%** report handling
  - **46.7%** share code
- Missingness classification, handling approaches, case-study

## Issues in simulation studies

*“...extensive simulation studies show that the proposed method performs  
**better than existing methods ...”***

# Issues in simulation studies

*“...extensive simulation studies show that the proposed method performs better than existing methods ...”*

- Over-Optimism (e.g., Ullmann et al., 2022)

# Issues in simulation studies

*“...extensive simulation studies show that the proposed method performs better than existing methods ...”*

- Over-Optimism (e.g., Ullmann et al., 2022)
- Issues similar to other empirical research  
(Boulesteix et al., 2020)

# Issues in simulation studies

*“...extensive simulation studies show that the proposed method performs better than existing methods ...”*

- Over-Optimism (e.g., Ullmann et al., 2022)
- Issues similar to other empirical research  
(Boulesteix et al., 2020)
- Insufficient reporting standards (e.g., Hoaglin and Andrews, 1975)

# Issues in simulation studies

*“...extensive simulation studies show that the proposed method performs better than existing methods ...”*

- Over-Optimism (e.g., Ullmann et al., 2022)
- Issues similar to other empirical research  
(Boulesteix et al., 2020)
- Insufficient reporting standards (e.g., Hoaglin and Andrews, 1975)
- Reproducibility? (e.g., Luijken et al., 2023)

# Issues in simulation studies

*“...extensive simulation studies show that the proposed method performs better than existing methods ...”*

- Over-Optimism (e.g., Ullmann et al., 2022)
- Issues similar to other empirical research  
(Boulesteix et al., 2020)
- Insufficient reporting standards (e.g., Hoaglin and Andrews, 1975)
- Reproducibility? (e.g., Luijken et al., 2023)



xkcd.com (CC-BY-NC)

# Meta-research on simulation studies

STATISTICS IN MEDICINE

*Statist. Med.* 2006; **25**:4279–4292

Published online 31 August 2006 in Wiley InterScience  
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

## The design of simulation studies in medical statistics

Andrea Burton<sup>1, 2, \*, †</sup>, Douglas G. Altman<sup>1</sup>, Patrick Royston<sup>1, 3</sup> and Roger L. Holder<sup>4</sup>

## On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses

Elizabeth KOEHLER, Elizabeth BROWN, and Sébastien J.-P. A. HANEUSE

DOI: 10.1002/bimj.202200104

DISCUSSION

Biometrical Journal →

## Against the “one method fits all data sets” philosophy for comparison studies in methodological research

Carolin Strobl<sup>1</sup> | Friedrich Leisch<sup>2</sup>

*Multivariate Behavioral Research*, 35 (2), 137-167

Copyright © 2000, Lawrence Erlbaum Associates, Inc.

## Design and Analysis of Monte Carlo Experiments: Attacking the Conventional Wisdom

Anders Skrondal

## Some Thoughts on Simulation Studies to Compare Clustering Methods

Christian Hennig

DOI: 10.1002/bimj.202200222

RESEARCH ARTICLE

Biometrical Journal →

## Phases of methodological research in biostatistics—Building the evidence base for new methods

Georg Heinze<sup>1</sup> | Anne-Laure Boulesteix<sup>2</sup> | Michael Kammer<sup>1,3</sup> | Tim P. Morris<sup>4</sup> |  
Ian R. White<sup>4</sup> | on behalf of the Simulation Panel of the STRATOS initiative

• • •

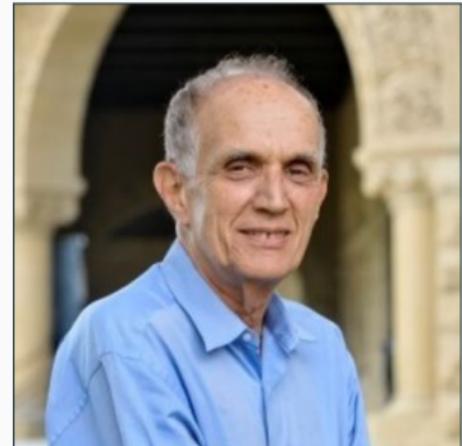
# **Questionable research practices in simulation studies**

---

# Neutrality in simulation studies

*“In fact it is **very difficult to run an honest simulation** comparison, and **easy to inadvertently cheat** by choosing favorable examples, or by not putting as much effort into optimizing the dull old standard as the exciting new challenger.”*

Brad Efron (2001)



<https://statistics.stanford.edu/people/bradley-efron>

# Our study

Received: 25 March 2022 | Revised: 5 January 2023 | Accepted: 9 January 2023  
DOI: 10.1002/bimj.202200091

**RESEARCH ARTICLE**

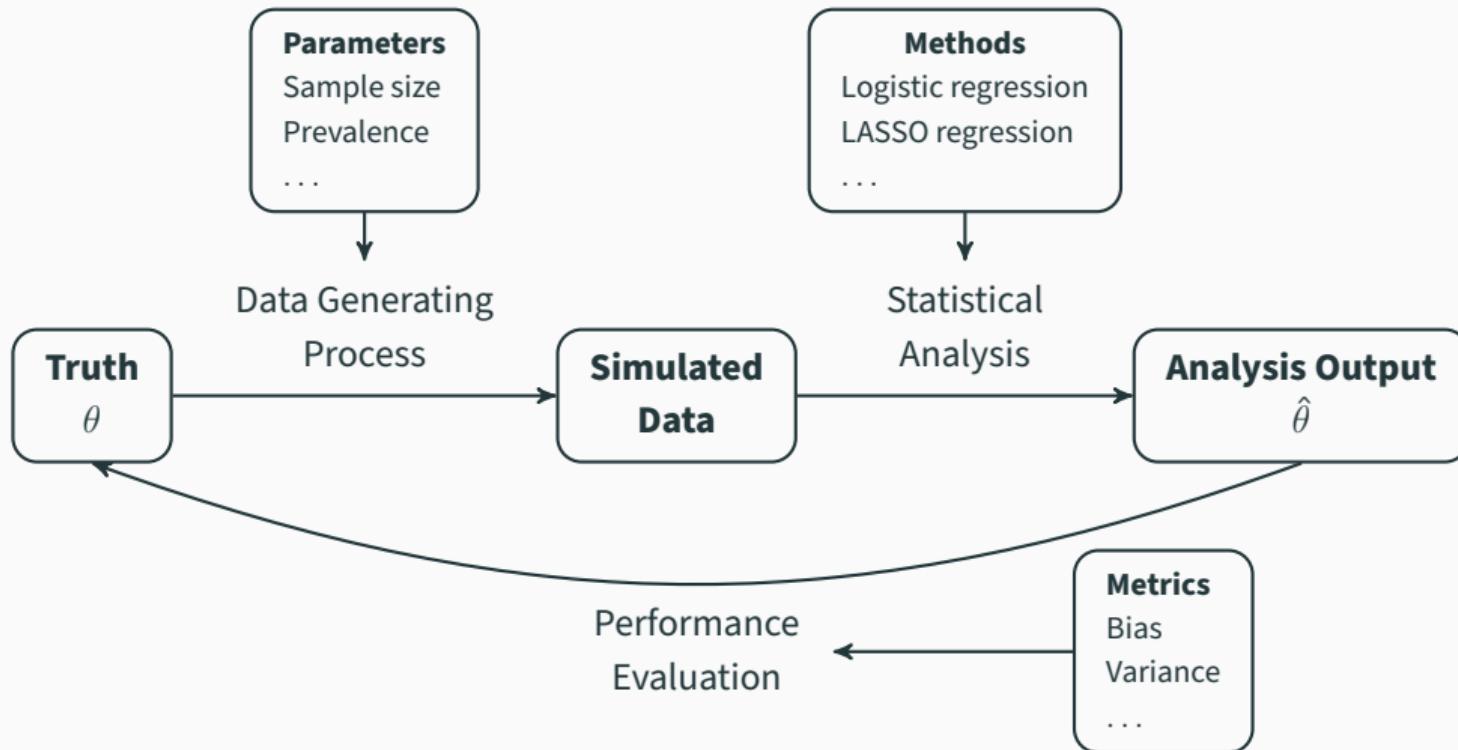
**Biometrical Journal** →

## Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method

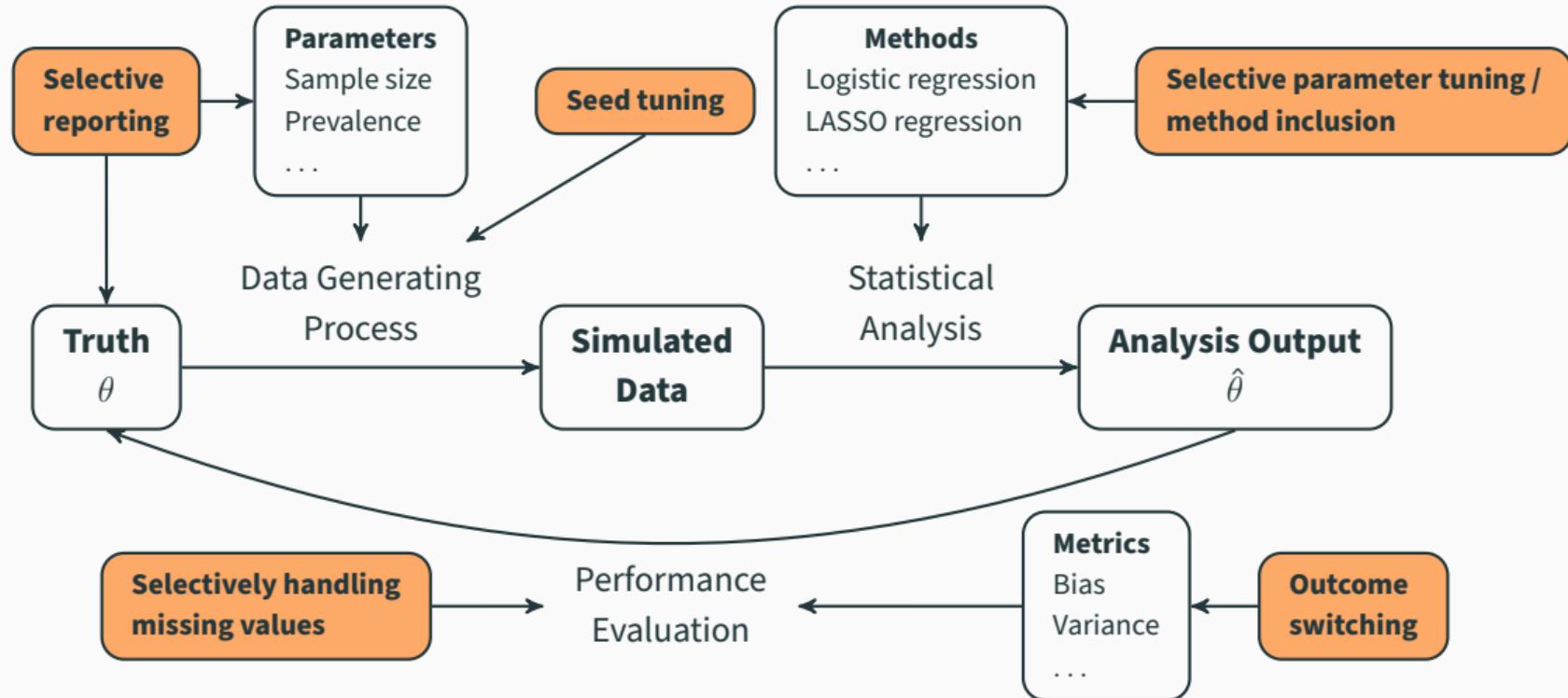
Samuel Pawel  | Lucas Kook  | Kelly Reeve 

- Which **questionable research practices** (QRPs) exist in simulation studies?
- How can QRPs **impact the conclusions** of a study?
- How can QRPs be **addressed**?

# Questionable research practices in simulation studies



## Questionable research practices in simulation studies

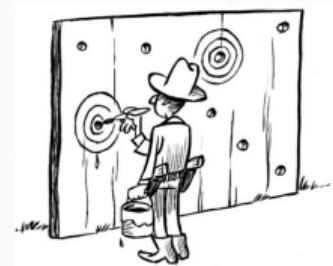


See Table 1 in doi:10.1002/bimj.202200091 for more ORPs.

# Questionable research practices in simulation studies

## Root causes

- **Pressure to publish** novel and positive results
- **Low requirements** from journals
- **Cognitive biases** (e.g., confirmation or hindsight bias)
- **Low awareness** in scientific community

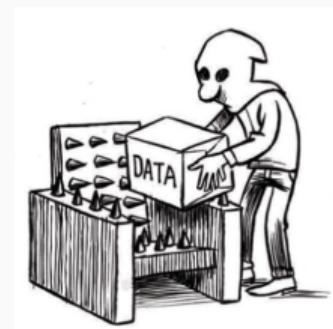


Dirk-Jan Hoek (CC-BY)

# Questionable research practices in simulation studies

## Root causes

- **Pressure to publish** novel and positive results
- **Low requirements** from journals
- **Cognitive biases** (e.g., confirmation or hindsight bias)
- **Low awareness** in scientific community



## Potential consequences

- **Overoptimistic conclusions**
- **Publication bias**
- **Misinformed decisions**



Dirk-Jan Hoek (CC-BY)

# QRP Illustration

Received: 25 March 2022

Revised: 5 January 2023

Accepted: 9 January 2023

DOI: 10.1002/bimj.202200091

RESEARCH ARTICLE

Biometrical Journal

## Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method

Samuel Pawel  | Lucas Kook  | Kelly Reeve 

“By **deliberately using several QRPs**, we were able to **present a method with no expected benefits** [...] **as an improvement** over [...] well-established competitors.”

# **Simulation studies in Psychology**

---

# Literature Review

*“Statisticians ... often pay too little attention to their own principles of design”*(Hoaglin & Andrews, 1975)

**Statistical Computing**

This Department will carry articles of high quality on all aspects of computation in statistics. Papers describing new algorithms, programs, or statistical packages will not receive coverage if the program, although completely documented, program must be available free of charge and without a licensing test of the program by the referee. The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program. The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

**The Reporting of Computation-Based Results in Statistics**

DAVID C. HOAGLIN\* and DAVID F. ANDREWS\*\*

STATISTICS IN MEDICINE  
*Statist. Med.* 2006; 25:4279–4292  
Published online 31 August 2006 in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton<sup>1, 2, \*, †</sup>, Douglas G. Altman<sup>1</sup>, Patrick Royston<sup>1, 3</sup> and Roger L. Holder<sup>4</sup>

TUTORIAL IN BIOSTATISTICS      WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris<sup>1</sup> | Ian R. White<sup>1</sup> | Michael J. Crowther<sup>2</sup>

# Literature Review

*“Statisticians ... often pay too little attention to their own principles of design”*(Hoaglin & Andrews, 1975)

**Statistical Computing**

This Department will carry articles of high quality on all aspects of computation in statistics. Papers describing new algorithms, programs, or statistical packages will not receive coverage if the program, although completely documented, program must be available free of charge and without a licensing test of the program by the referee. The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program. The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

**The Reporting of Computation-Based Results in Statistics**

DAVID C. HOAGLIN\* and DAVID F. ANDREWS\*\*

STATISTICS IN MEDICINE  
*Statist. Med.* 2006; 25:4279–4292  
Published online 31 August 2006 in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton<sup>1, 2, \*, †</sup>, Douglas G. Altman<sup>1</sup>, Patrick Royston<sup>1, 3</sup> and Roger L. Holder<sup>4</sup>

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris<sup>1</sup> | Ian R. White<sup>1</sup> | Michael J. Crowther<sup>2</sup>

**This project:**

# Literature Review

*“Statisticians ... often pay too little attention to their own principles of design”*(Hoaglin & Andrews, 1975)

The screenshot shows a journal article from the journal "Statistical Computing". The title is "Statistical Computing". Below the title, there is a short description of the department's focus on computation in statistics. The main article title is "The Reporting of Computation-Based Results in Statistics" by David C. Hoaglin\* and David F. Andrews\*\*. The article is from the journal "STATISTICS IN MEDICINE" (Statist. Med. 2006; 25:4279-4292). It was published online on 31 August 2006 in Wiley InterScience (www.interscience.wiley.com) with DOI: 10.1002/sim.2673. The abstract discusses the design of simulation studies in medical statistics, featuring authors Andrea Burton, Douglas G. Altman, Patrick Royston, and Roger L. Holder. The article is part of a special issue titled "TUTORIAL IN BIOSTATISTICS" and is associated with the journal "WILEY Statistics in Medicine". The footer includes the names Tim P. Morris, Ian R. White, and Michael J. Crowther.

**Statistical Computing**

This Department will carry articles of high quality on all aspects of computation in statistics. Papers describing new algorithms, programs, or statistical packages will not receive coverage if the program, although completely documented, must be available free of charge. Papers describing a working test of the program by the referee.

The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.

The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

**The Reporting of Computation-Based Results in Statistics**

DAVID C. HOAGLIN\* and DAVID F. ANDREWS\*\*

STATISTICS IN MEDICINE  
Statist. Med. 2006; 25:4279-4292  
Published online 31 August 2006 in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton<sup>1, 2, \*, †</sup>, Douglas G. Altman<sup>1</sup>, Patrick Royston<sup>1, 3</sup> and Roger L. Holder<sup>4</sup>

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris<sup>1</sup> | Ian R. White<sup>1</sup> | Michael J. Crowther<sup>2</sup>

## This project:

- Review of **100 recent simulation studies** in psychology

# Literature Review

*“Statisticians ... often pay too little attention to their own principles of design”*(Hoaglin & Andrews, 1975)

The screenshot shows a section of a journal page titled "Statistical Computing". The title is underlined and bolded. Below it is a descriptive text about the department's focus on computation in statistics. Underneath this is a section titled "The Reporting of Computation-Based Results in Statistics" by David C. HOAGLIN\* and David F. ANDREWS\*\*. The text below the authors' names is as follows:

STATISTICS IN MEDICINE  
Statist. Med. 2006; 25:4279–4292  
Published online 31 August 2006 in Wiley InterScience  
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton<sup>1, 2, \*, †</sup>, Douglas G. Altman<sup>1</sup>, Patrick Royston<sup>1, 3</sup> and Roger L. Holder<sup>4</sup>

TUTORIAL IN BIOSTATISTICS      WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris<sup>1</sup> | Ian R. White<sup>1</sup> | Michael J. Crowther<sup>2</sup>

## This project:

- Review of **100 recent simulation studies** in psychology
- Psychological Methods, Behavior Research Methods, Multivariate Behavioral Research

# Literature Review

*“Statisticians ... often pay too little attention to their own principles of design”*(Hoaglin & Andrews, 1975)

The screenshot shows a section of a journal page titled "Statistical Computing". The title is underlined and bolded. Below it is a descriptive text about the department's focus on computation in statistics. Underneath this is a section titled "The Reporting of Computation-Based Results in Statistics" by David C. HOAGLIN\* and David F. ANDREWS\*\*. Below the authors' names is a note about statistics in medicine, mentioning the journal "Statist. Med." from 2006, volume 25, pages 4279-4292. It also notes the publication online on 31 August 2006 in Wiley InterScience. The abstract begins with "The design of simulation studies in medical statistics" by Andrea Burton<sup>1, 2, \*, †</sup>, Douglas G. Altman<sup>1</sup>, Patrick Royston<sup>1, 3</sup> and Roger L. Holder<sup>4</sup>. The footer includes "TUTORIAL IN BIOSTATISTICS" and the "WILEY Statistics in Medicine" logo. At the bottom, it says "Using simulation studies to evaluate statistical methods" and lists authors: Tim P. Morris<sup>1</sup> | Ian R. White<sup>1</sup> | Michael J. Crowther<sup>2</sup>.

## This project:

- Review of **100 recent simulation studies** in psychology
- Psychological Methods, Behavior Research Methods, Multivariate Behavioral Research
- Coding of various aspects of reporting

# Overview Paper



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

© 2024 American Psychological Association  
ISSN: 1082-989X

Psychological Methods

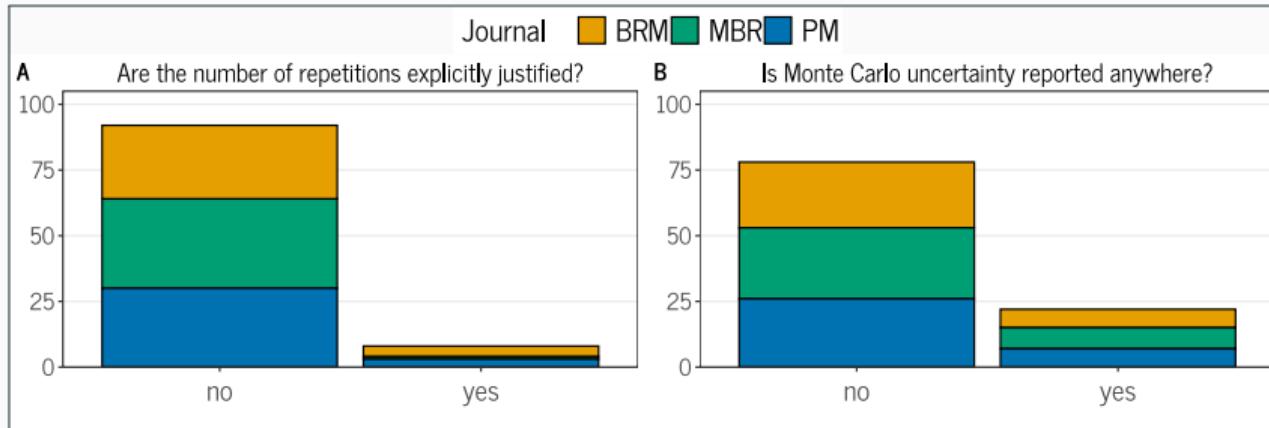
<https://doi.org/10.1037/met0000695>

## Simulation Studies for Methodological Research in Psychology: A Standardized Template for Planning, Preregistration, and Reporting

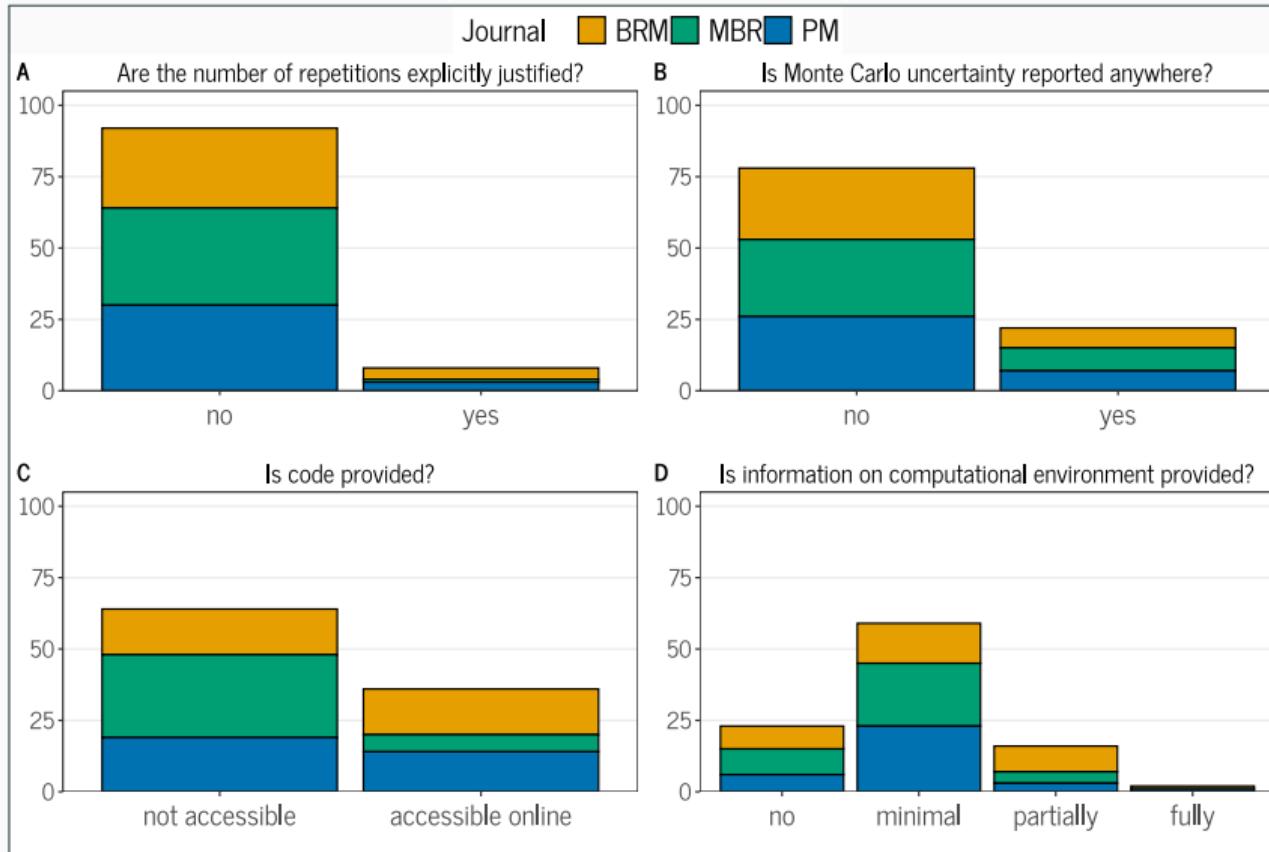
Björn S. Siepe<sup>1</sup>, František Bartoš<sup>2</sup>, Tim P. Morris<sup>3</sup>, Anne-Laure Boulesteix<sup>4, 5</sup>,  
Daniel W. Heck<sup>1</sup>, and Samuel Pawel<sup>6, 7</sup>

# Main Results

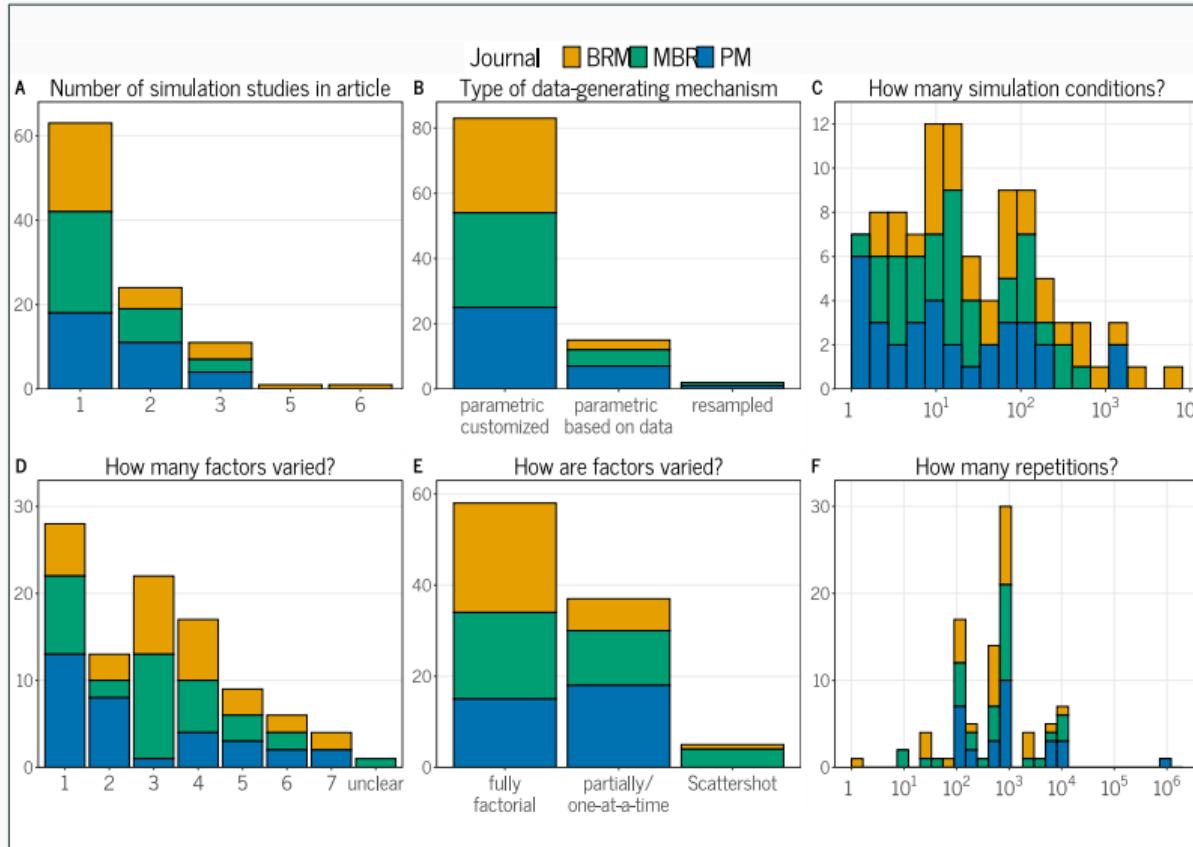
# Main Results



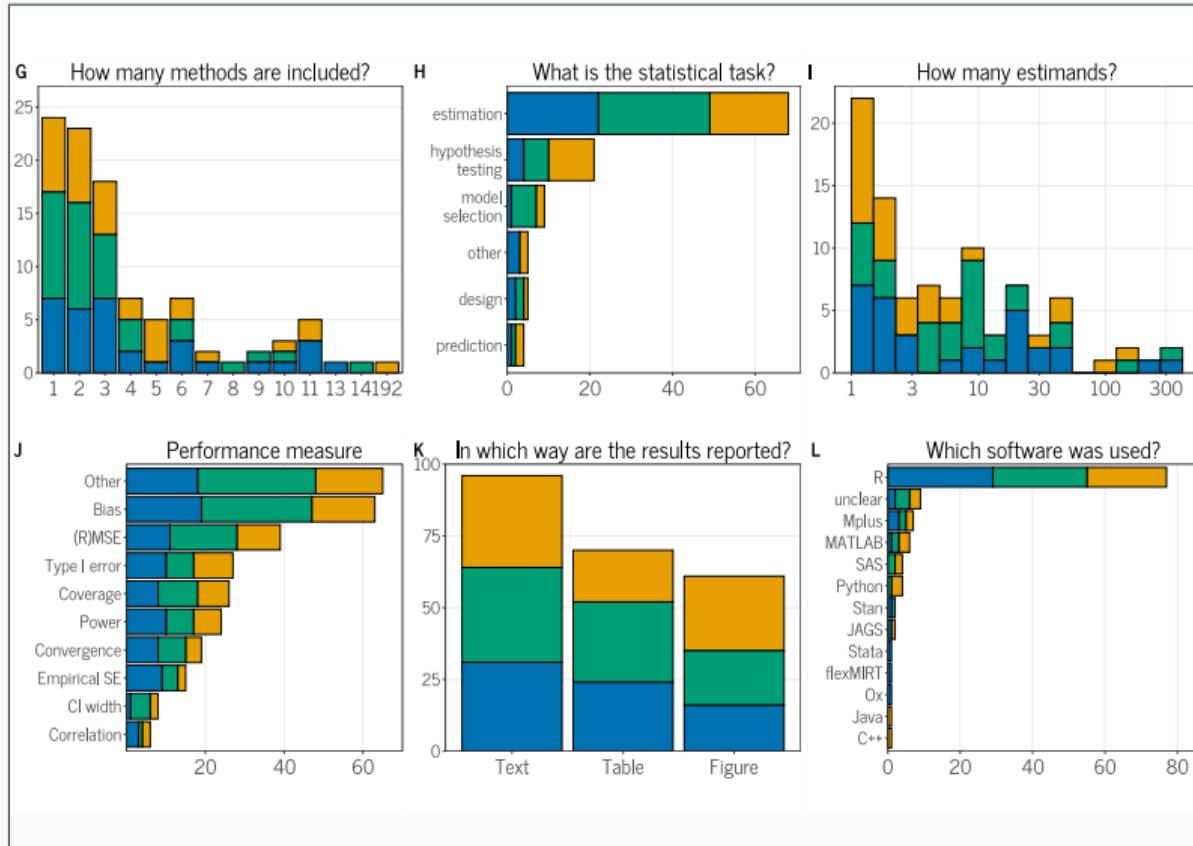
# Main Results



# Additional Results



# Additional Results



# Reporting Suggestions

# Reporting Suggestions

**Table 3**

*Definitions of Common Performance Measures, their Estimates, Monte Carlo Standard Errors (MCSE), and Number of Simulation Repetitions  $n_{\text{sim}}$  to Achieve a Desired MCSE<sub>\*</sub>.*

Performance measure	Definition	Estimate	MCSE	$n_{\text{sim}}$
Bias	$E(\hat{\theta}) - \theta$	$(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta$	$\sqrt{S_{\hat{\theta}}^2 / n_{\text{sim}}}$	$S_{\hat{\theta}}^2 / \text{MCSE}_*^2$
Relative bias	$\{E(\hat{\theta}) - \theta\} / \theta$	$\{(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta\} / \theta$	$\sqrt{S_{\hat{\theta}}^2 / (\theta^2 n_{\text{sim}})}$	$S_{\hat{\theta}}^2 / (\text{MCSE}_*^2 \theta^2)$
Mean square error (MSE)	$E\{(\hat{\theta} - \theta)^2\}$	$\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}$	$\sqrt{S_{(\hat{\theta}-\theta)^2}^2 / n_{\text{sim}}}$	$S_{(\hat{\theta}-\theta)^2}^2 / \text{MCSE}_*^2$
Root mean square error (RMSE)	$\sqrt{E\{(\hat{\theta} - \theta)^2\}}$	$\sqrt{\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}}$	$\sqrt{S_{(\hat{\theta}-\theta)^2}^2 / (4n_{\text{sim}} \widehat{\text{MSE}})}$	$S_{(\hat{\theta}-\theta)^2}^2 / (4\widehat{\text{MSE}} \text{MCSE}_*^2)$
Empirical variance	$\text{Var}(\hat{\theta})$	$S_{\hat{\theta}}^2$	$S_{\hat{\theta}}^2 \sqrt{2 / (n_{\text{sim}} - 1)}$	$1 + 2(S_{\hat{\theta}}^2)^2 / \text{MCSE}_*^2$
Empirical standard error	$\sqrt{\text{Var}(\hat{\theta})}$	$\sqrt{S_{\hat{\theta}}^2}$	$\sqrt{S_{\hat{\theta}}^2 / \{2(n_{\text{sim}} - 1)\}}$	$1 + S_{\hat{\theta}}^2 / (2\text{MCSE}_*^2)$
Coverage	$\Pr(\text{CI includes } \theta)$	$\sum_{i=1}^{n_{\text{sim}}} \mathbb{1}(\text{CI}_i \text{ includes } \theta) / n_{\text{sim}}$	$\sqrt{\widehat{\text{Cov}}(1 - \widehat{\text{Cov}}) / n_{\text{sim}}}$	$\widehat{\text{Cov}}(1 - \widehat{\text{Cov}}) / \text{MCSE}_*^2$
Power (or Type I error rate)	$\Pr(\text{Test rejects } H_0)$	$\sum_{i=1}^{n_{\text{sim}}} \mathbb{1}(\text{Test}_i \text{ rejects } H_0) / n_{\text{sim}}$	$\sqrt{\widehat{\text{Pow}}(1 - \widehat{\text{Pow}}) / n_{\text{sim}}}$	$\widehat{\text{Pow}}(1 - \widehat{\text{Pow}}) / \text{MCSE}_*^2$
Mean CI width	$E(\text{CI}_{\text{upper}} - \text{CI}_{\text{lower}})$	$\sum_{i=1}^{n_{\text{sim}}} (\text{CI}_{i,\text{upper}} - \text{CI}_{i,\text{lower}}) / n_{\text{sim}}$	$\sqrt{S_W^2 / n_{\text{sim}}}$	$S_W^2 / \text{MCSE}_*^2$
Mean of generic statistic $G$	$E(G)$	$\sum_{i=1}^{n_{\text{sim}}} G_i / n_{\text{sim}}$	$\sqrt{S_G^2 / n_{\text{sim}}}$	$S_G^2 / \text{MCSE}_*^2$

*Note.* Table adapted from Table 6 in Morris et al. (2019)

## Potential improvements

---

# How to address questionable research practices?

## Researchers

- Preregistered simulation protocols
- Adversarial collaboration
- Blinding of analysis
- Transparent reporting (e.g., disclose non-neutrality)



# How to address questionable research practices?

## Researchers

- Preregistered simulation protocols
- Adversarial collaboration
- Blinding of analysis
- Transparent reporting (e.g., disclose non-neutrality)



## Reviewers, journals, funders

- Encourage simulation protocols
- Incentivize neutrality and transparency in simulation studies
- Deincentivize outperforming state-of-the-art methods

# Simulation study protocols

STATISTICS IN MEDICINE

*Statist. Med.* 2006; **25**:4279–4292

Published online 31 August 2006 in Wiley InterScience

(www.interscience.wiley.com) DOI: 10.1002/sim.2673

## The design of simulation studies in medical statistics

Andrea Burton<sup>1,2,\*†</sup>, Douglas G. Altman<sup>1</sup>, Patrick Royston<sup>1,3</sup> and Roger L. Holder<sup>4</sup>

**“When planning a simulation study, it is recommended that a detailed protocol be produced, giving full details of how the study will be performed, analysed and reported.”**

Burton et al. (2006)

# Simulation study protocols

## Advantages

- + Planning and reporting
- + Transparency and replicability
- + Can be preregistered
- ? Less/more work

→ **How to structure protocol?**

0. Detailed protocol of all aspects of the simulation study
  - a. Justifications for all the decisions made
1. Clearly defined aims and objectives
2. Simulation procedures
  - a. Level of dependence between simulated datasets
  - b. Allowance for failures
  - c. Software to perform simulations
  - d. Random number generator to use
  - e. Specification of the starting seeds
3. Methods for generating the datasets
4. Scenarios to be investigated
5. Statistical methods to be evaluated
6. Estimates to be stored for each simulation and summary measures to be calculated over all simulations
7. Number of simulations to be performed
8. Criteria to evaluate the performance of statistical methods for different scenarios
  - a. Assessment of bias
  - b. Assessment of accuracy
  - c. Assessment of coverage
9. Presentation of the simulation results

Proposal from Burton et al. (2006)

# The ADEMP-PreReg template

## ADEMP-PreReg Template for Simulation Studies

March 20, 2025

Version: 1.1  
Last updated: 2024-11-18

Protocol template based on:

- **ADEMP structure** (Morris et al., 2019)
- **Open science** aspects
- **Reproducibility** aspects

# The ADEMP-PreReg template – Different versions

The screenshot shows the Overleaf project page for the "ADEMP-PreReg Simulation Study Template". The page includes a preview of the document, project details like authorship and last update, and download links for Open as Template, View Source, and View PDF.

**ADEMP-PreReg Simulation Study Template**

ADEMP-PreReg  
Template for Simulation Studies

November 1, 2023

**Author:** Björn S. Siepe, František Bartoš, Tim P. Morris, Anne-Laure Boulesteix, Daniel W. Heck, Samuel Pawel

**Last Updated:** 7 months ago

**License:** Creative Commons CC BY 4.0

**Abstract:** ADEMP-PreReg is a step-by-step template that researchers can use for the design, potential preregistration, and reporting of their simulation studies.

**Tags:** Project / Lab Report

[Find More Templates](#)

The screenshot shows a Microsoft Word document titled "ADEMP-PreReg Template for Simulation Studies". The document header includes the title and version information. The main content consists of two sections: "1 Instructions" and "1.1 General Information". Both sections provide a brief overview of the template's purpose and how it can be used.

**ADEMP-PreReg**  
Template for Simulation Studies

Version: 0.1.0  
Last updated:  
Preregistration template designed by  
Björn S. Siepe, František Bartoš, Tim P. Morris, Anne-Laure Boulesteix, Daniel W. Heck, and Samuel Pawel

## 1 Instructions

### 1.1 General Information

This template can be used to plan and/or preregister Monte Carlo simulation studies according to the ADEMP framework (Morris, White, and Crowther 2019). The preprint associated with this template is (Siepe et al. 2023). Alternative Google Docs and Word versions of this template are available at (<https://github.com/siepe/ADEMP-PreReg>). To time-stamp your protocol, we recommend uploading it to the Open Science Framework (<https://osf.io/>) or Zenodo (<https://zenodo.org/>). When using this template, please cite the associated preprint (Siepe et al. 2023). If you have any questions or suggestions for improving the template, please contact us via the ways described at (<https://github.com/siepe/ADEMP-PreReg>).

LaTeX, Overleaf

MS/Libre office, Google docs

# The ADEMP-PreReg template - A living document

The screenshot shows the GitHub repository page for "ADEMP-PreReg". The repository is public and has 25 commits. The README file contains the ADEMP Preregistration template. The repository has 1 watch, 2 forks, and 3 stars. It was created on Oct 31, 2023.

**Code** | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights

**ADEMP-PreReg** Public

Go to file + <> Code

**main** 25 Commits

**About**  
ADEMP preregistration protocol for simulation studies

Readme  
Activity  
3 stars  
1 watching  
2 forks  
Report repository

**Releases** 1

Initial Template Version (Latest) on Oct 31, 2023

**Packages**  
No packages published Publish your first package

**Contributors** 2

SamCH93  
bsiepe Björn Siepe

**README**

## ADEMP Preregistration

DOI: 10.5281/zenodo.10057884

This repository contains the ADEMP preregistration (ADEMP-PreReg) template introduced in Siepe et al. (2023, <https://doi.org/10.31234/osf.io/ufov6>), which is based on the ADEMP reporting structure from Morris et al.

<https://github.com/bsiepe/ADEMP-PreReg>

# The ADEMP-PreReg template - Overview

1. Instructions
2. General information
3. Aims
4. Data-generating mechanism
5. Estimands and targets
6. Methods
7. Performance Measures
8. Computational details

## 7 Performance Measures

### 7.1 Which performance measures will be used?

*Explanation:* Please provide details on why they were chosen and on how these measures will be calculated. Ideally, provide formulas for the performance measures to avoid ambiguity. Some models in psychology, such as item response theory or time series models, often contain multiple parameters of interest, and their number may vary across conditions. With a large number of estimated parameters, their performance measures are often combined. If multiple estimates are aggregated, specify how this aggregation will be performed. For example, if there are multiple parameters in a particular condition, the mean of the individual biases of these parameters or the bias of each individual parameter may be reported.

#### Example

Our primary performance measures are the type I error rate (in conditions where the true effect is zero) and the power (in conditions where the true effect is non-zero) to reject the null hypothesis of no difference between the control and treatment condition. The null hypothesis is rejected if the  $p$ -value for the null hypothesis of no effect is less than or equal to the conventional threshold of 0.05. The rejection rate (the type I error rate or the power, depending on the data generating mechanism) is estimated by

$$\widehat{\text{RRate}} = \frac{\sum_{i=1}^{n_{\text{sim}}} 1(p_i \leq 0.05)}{n_{\text{sim}}}$$

where  $1(p_i \leq 0.05)$  is the indicator of whether the  $p$ -value in simulation  $i$  is equal to or less than 0.05. We use the following formula to compute the MCSE of the rejection rate

$$\text{MCSE}_{\widehat{\text{RRate}}} = \sqrt{\frac{\widehat{\text{RRate}}(1 - \widehat{\text{RRate}})}{n_{\text{sim}}}}.$$

# The ADEMP-PreReg template

## Purposes

# The ADEMP-PreReg template

## Purposes

- Blueprint for **planning, reporting & reviewing** of simulation studies

# The ADEMP-PreReg template

## Purposes

- Blueprint for **planning, reporting & reviewing** of simulation studies
- **Preregistration** brings multiple benefits similar to other empirical research
  - Avoid QRPs
  - Increase transparency
  - Improve informativeness

## Limitations

# The ADEMP-PreReg template

## Purposes

- Blueprint for **planning, reporting & reviewing** of simulation studies
- **Preregistration** brings multiple benefits similar to other empirical research
  - Avoid QRPs
  - Increase transparency
  - Improve informativeness

## Limitations

- Preregistration could be **faked**

# The ADEMP-PreReg template

## Purposes

- Blueprint for **planning, reporting & reviewing** of simulation studies
- **Preregistration** brings multiple benefits similar to other empirical research
  - Avoid QRPs
  - Increase transparency
  - Improve informativeness



## Limitations

- Preregistration could be **faked**
- May **slow down** exploratory research

doi:10.5281/zenodo.7994221

## Current trends

# Current trends



## Special Collection: “Neutral Comparison Studies in Methodological Research”

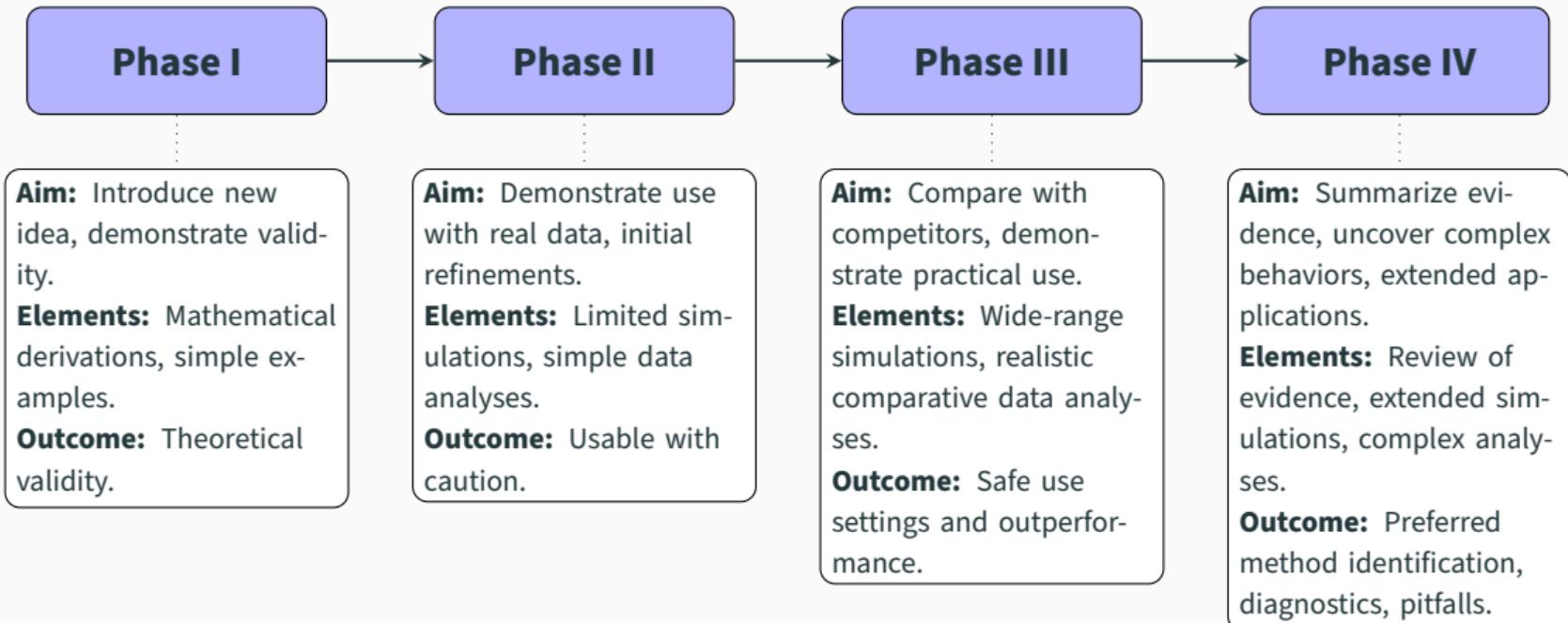
[Virtual Issues](#) | First published: 14 December 2023 | Last updated: 19 February 2024

Biometriicians are frequently faced with a multitude of methods they might use for the analysis and/or design of studies. Choosing an appropriate method is a challenge, and neutral comparison studies are an essential step towards providing practical guidance. This Special Collection contains both papers defining, developing, discussing or illustrating concepts related to the design and interpretation of neutral comparison studies, and reports of neutral comparison studies of methods that address specific biostatistical problems.

**Guest editors:** Anne-Laure Boulesteix, Mark Baillie, Dominic Edelmann, Leonhard Held, Tim Morris, Willi Sauerbrei

- Focus on “**neutral comparison studies**” (Boulesteix et al., 2013)
- Some journals adopt **reproducibility checks** (Wrobel et al., 2024)
- **Various fields** discuss how to improve methodological research (e.g., Robinson and Vitek, 2019; Van Mechelen et al., 2023; Herrmann et al., 2024)
- **Meta-research** on simulation/benchmarking studies continues

# Phases of methodological research (Heinze et al., 2024)



Based on Heinze et al. (2024)

## Against “one method fits all [data sets]” (Strobl and Leisch, 2024)

# Against “one method fits all [data sets]” (Strobl and Leisch, 2024)

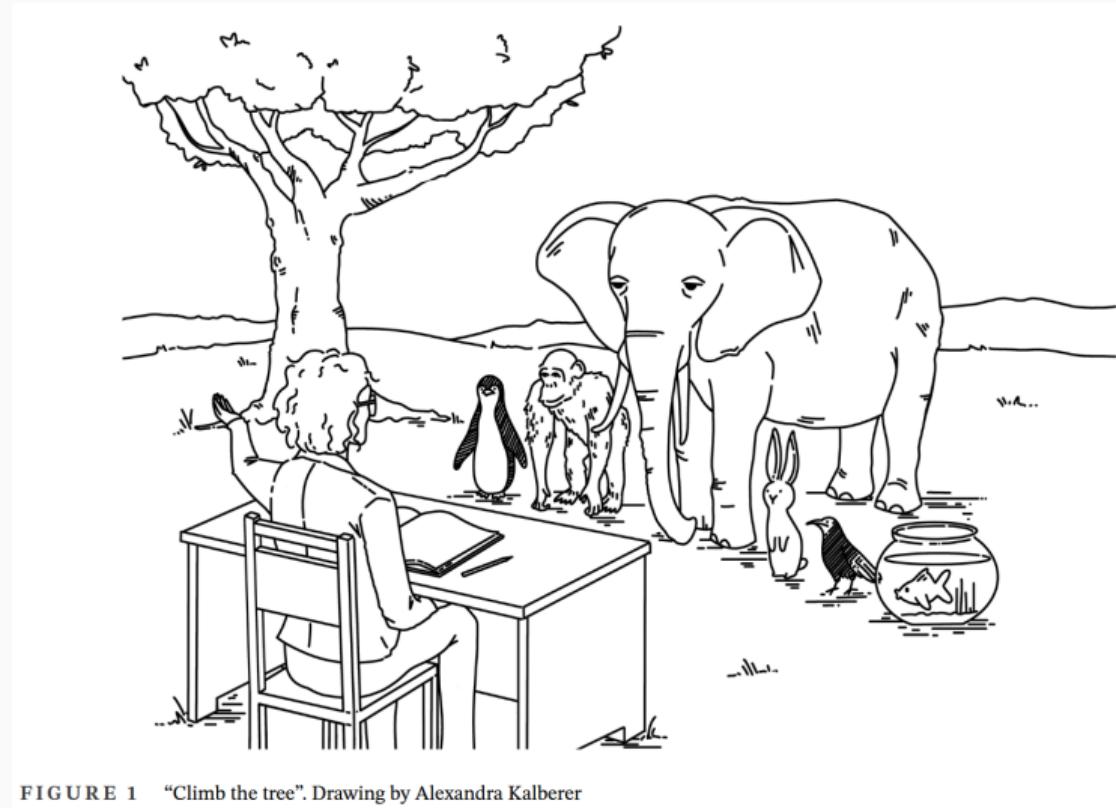


FIGURE 1 “Climb the tree”. Drawing by Alexandra Kalberer

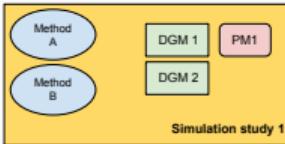
# WIP: Synthetic benchmarking

## Separate Studies (Status Quo)

Paper 1  
(new method)

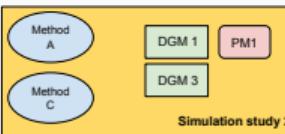


Paper 2  
(new method &  
simulation)



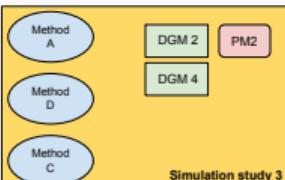
Comparison not possible

Paper 3  
(new method & DGM)



Comparison not possible

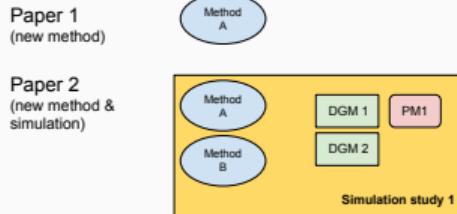
Paper 4  
(new method & DGM  
& PM)



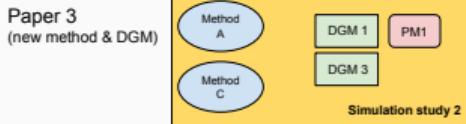
DGM: Data-Generating Mechanism  
PM: Performance Measure

# WIP: Synthetic benchmarking

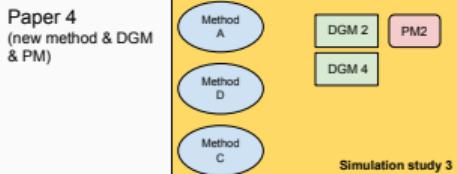
Separate Studies (Status Quo)



Comparison not possible



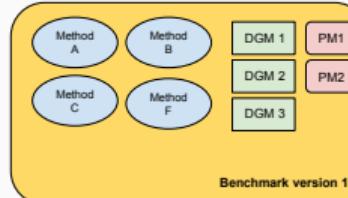
Comparison not possible



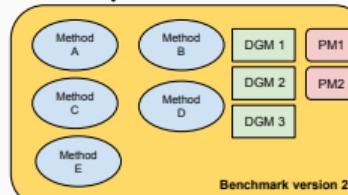
DGM: Data-Generating Mechanism  
PM: Performance Measure

Continuous Synthetic Benchmarking (Proposal)

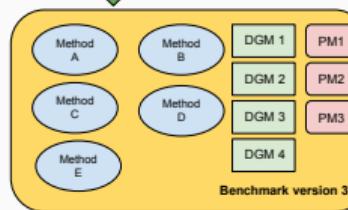
Paper 5  
(collects methods, DGMs, PMs)



Extends (new method)



Extends (new DGM & PM)



## **Discussion**

---

# Conclusions

Received: 25 March 2022 | Revised: 5 January 2023 | Accepted: 9 January 2023

DOI: 10.1082/bmjj.202200091

RESEARCH ARTICLE

Biometrical Journal

**Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method** 

Samuel Pawel  | Lucas Kook  | Kelly Reeve 



© 2024 American Psychological Association  
ISSN: 1082-989X

Psychological Methods

<https://doi.org/10.1037/met0000695>

Simulation Studies for Methodological Research in Psychology:  
A Standardized Template for Planning, Preregistration, and Reporting

Björn S. Siepe<sup>1</sup>, František Bartoš<sup>2</sup>, Tim P. Morris<sup>3</sup>, Anne-Laure Boulesteix<sup>4, 5</sup>,  
Daniel W. Heck<sup>1</sup>, and Samuel Pawel<sup>6, 7</sup>

- **Simulation studies** are ubiquitous in methodological research
- Simulation studies can be impacted by **questionable research practices** and misaligned **incentives**
- **Protocols** have potential to improve simulation studies
- Meta-research, discussions, and reforms needed to **increase awareness** and **improve standards**

# Open questions

- Which simulation studies require which **degree of rigour**?
- How to avoid **cheating in preregistration**?
- How can journals/researchers/reviewers/communities promote **good practices**?
- Other ways to **improve** simulation studies?



xkcd.com (CC-BY-NC)

# A multidisciplinary collaboration



František Bartoš



Lucas Kook



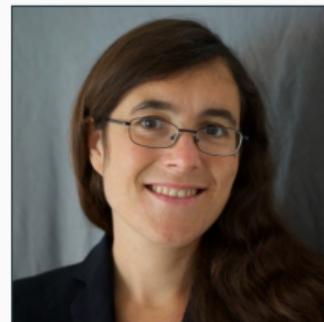
Kelly Reeve



Daniel W. Heck



Tim P. Morris



Anne-Laure Boulesteix



Anna Lohmann

## TODO Get In Touch + QR CODE

# References i

- Boulesteix, A.-L., Hoffmann, S., Charlton, A., and Seibold, H. (2020). A replication crisis in methodological research? *Significance*, 17(5):18–21.  
doi:10.1111/1740-9713.01444.
- Boulesteix, A.-L., Lauer, S., and Eugster, M. J. A. (2013). A plea for neutral comparison studies in computational sciences. *PLoS ONE*, 8(4):e61562.  
doi:10.1371/journal.pone.0061562.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.  
doi:10.1214/ss/1009213726.
- Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.  
doi:10.1002/sim.2673.
- Heinze, G., Boulesteix, A.-L., Kammer, M., Morris, T. P., and and, I. R. W. (2024). Phases of methodological research in biostatistics—building the evidence base for new methods. *Biometrical Journal*. doi:10.1002/bimj.202200222.
- Hennig, C. (2018). Some thoughts on simulation studies to compare clustering methods. *Archives of Data Science, Series A*, 5(1). doi:10.5445/KSP/1000087327/24.
- Herrmann, M., Lange, F. J. D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., Rügamer, D., Hüllermeier, E., Boulesteix, A.-L., and Bischl, B. (2024). Position: Why we must rethink empirical research in machine learning. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18228–18247. PMLR.
- Hoaglin, D. C. and Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, 29(3):122–126.  
doi:10.1080/00031305.1975.10477393.
- Hu, L.-t. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55. doi:10.1080/10705519909540118.

# References ii

- Kirmeier, E., Eriksson, L. I., Lewald, H., Jonsson Fagerlund, M., et al. (2019). Post-anaesthesia pulmonary complications after use of muscle relaxants (popular): a multicentre, prospective observational study. *The Lancet Respiratory Medicine*, 7(2):129–140. doi:10.1016/s2213-2600(18)30294-7.
- Koehler, E., Brown, E., and Haneuse, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162. doi:10.1198/tast.2009.0030.
- Luijken, K., Lohmann, A., Alter, U., Gonzalez, J. C., Clouth, F. J., Fossum, J. L., Hesen, L., Huizing, A. H. J., Ketelaar, J., Montoya, A. K., Nab, L., Nijman, R. C. C., de Vries, B. B. L. P., Tibbe, T. D., Wang, Y. A., and Groenwold, R. H. H. (2023). Replicability of simulation studies for the investigation of statistical methods: The replisims project. doi:10.48550/ARXIV.2307.02052. arXiv preprint.
- Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102. doi:10.1002/sim.8086.
- Pawel, S., Bartoš, F., Siepe, B. S., and Lohmann, A. (2024a). Handling missingness, failures, and non-convergence in simulation studies: A review of current practices and recommendations. doi:10.48550/arXiv.2409.18527. URL <https://arxiv.org/abs/2409.18527>. arXiv preprint.
- Pawel, S., Kook, L., and Reeve, K. (2024b). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, (e2200091):1–19. doi:10.1002/bimj.202200091.
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379. doi:10.1016/s0895-4356(96)00236-3.
- Robinson, M. D. and Vitek, O. (2019). Benchmarking comes of age. *Genome Biology*, 20(1):205. doi:10.1186/s13059-019-1846-5.
- Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D. W., and Pawel, S. (2024). Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting. *Psychological Methods*. doi:10.31234/osf.io/ufgy6. URL <https://doi.org/10.1037/met0000695>. to appear.

# References iii

- Skrondal, A. (2000). Design and analysis of monte carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, 35(2):137–167.  
doi:10.1207/s15327906mbr3502\_1.
- Strobl, C. and Leisch, F. (2024). Against the “one method fits all data sets” philosophy for comparison studies in methodological research. *Biometrical Journal*.  
doi:10.1002/bimj.202200104.
- Ullmann, T., Beer, A., Hünemörder, M., Seidl, T., and Boulesteix, A.-L. (2022). Over-optimistic evaluation and reporting of novel cluster algorithms: An illustrative study. *Advances in Data Analysis and Classification*. doi:10.1007/s11634-022-00496-5. Advance online publication.
- Van Mechelen, I., Boulesteix, A., Dangl, R., Dean, N., Hennig, C., Leisch, F., Steinley, D., and Warrens, M. J. (2023). A white paper on good research practices in benchmarking: The case of cluster analysis. *WIREs Data Mining and Knowledge Discovery*, 13(6). doi:10.1002/widm.1511.
- van Smeden, M., de Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., and Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1). doi:10.1186/s12874-016-0267-3.
- Wrobel, J., Hector, E. C., Crawford, L., McGowan, L. D., da Silva, N., Goldsmith, J., Hicks, S., Kane, M., Lee, Y., Mayrink, V., Paciorek, C. J., Usher, T., and Wolfson, J. (2024). Partnering with authors to enhance reproducibility at JASA. *Journal of the American Statistical Association*, 119(546):795–797.  
doi:10.1080/01621459.2024.2340557.