Universität Marburg

Simulation Studies for Methodological Research: Status Quo, Problems, and Potential Solutions

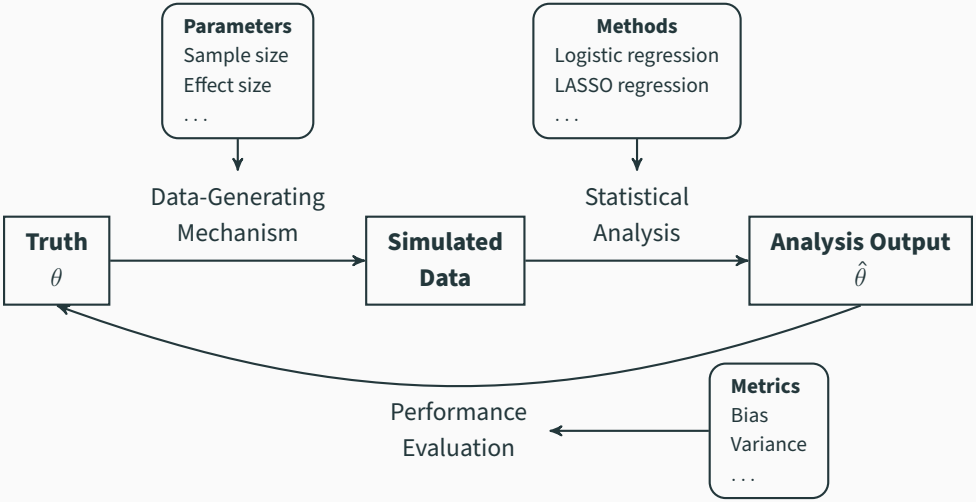
Björn Siepe[1]

October 1st, 2025 – Conference of the DGPs Methods and Evaluation Section 2025

[1]Psychological Methods Lab, Department of Psychology, Philipps-Universität Marburg

Simulation studies

Simulation studies



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*** *…"*

- 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*** *…"*

- Issues similar to other empirical research (Boulesteix et al., 2020)
- 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*** *…"*

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

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

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

- Issues similar to other empirical research (Boulesteix et al., 2020)
- Over-Optimism (e.g., Ullmann et al., 2022)
- Insufficient reporting standards (e.g., Hoaglin and Andrews, 1975)
- Reproducibility? (e.g., Luijken et al., 2023)
- Inadequate handling of missingness (Pawel et al., 2024a)

Issues in simulation studies

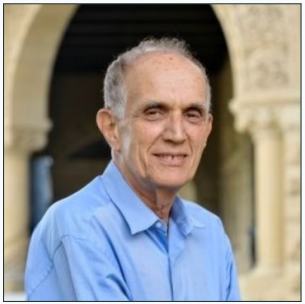
> *"…extensive simulation studies show that the proposed method performs* ***better than existing methods*** *…"*

- Issues similar to other empirical research (Boulesteix et al., 2020)
- Over-Optimism (e.g., Ullmann et al., 2022)
- Insufficient reporting standards (e.g., Hoaglin and Andrews, 1975)
- Reproducibility? (e.g., Luijken et al., 2023)
- Inadequate handling of missingness (Pawel et al., 2024a)
- Ignoring uncertainty (Koehler et al., 2009)

Neutrality in simulation studies

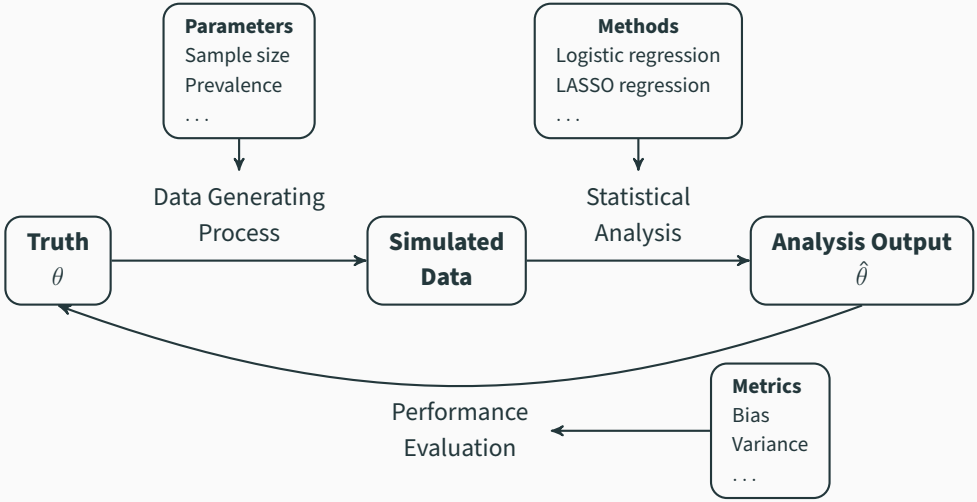
"*In fact it is* ***very difficult to run an honest simulation*** *comparison, and* ***easy to inadvertently cheat*** *by choosing favorable examples, or by not putting as much effort into optimizing the dull old standard as the exciting new challenger.*"

Brad Efron (2001)

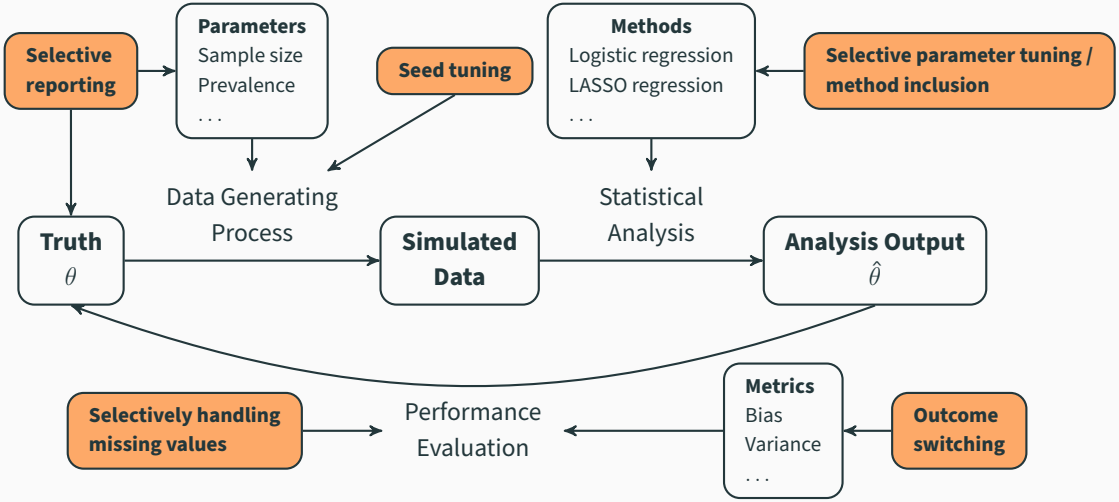


https://statistics.stanford.edu/people/bradley-efron

Questionable research practices in simulation studies



Questionable research practices in simulation studies

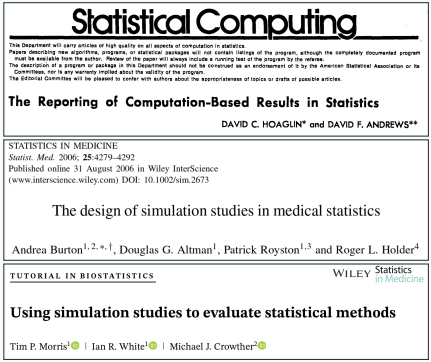


See Table 1 in doi:10.1002/bimj.202200091 for more QRPs

Literature Review

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 contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running 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[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1,3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

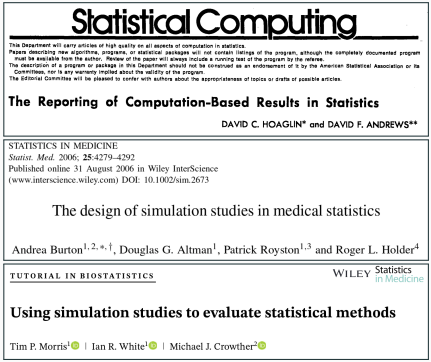
WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

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 contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running 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[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

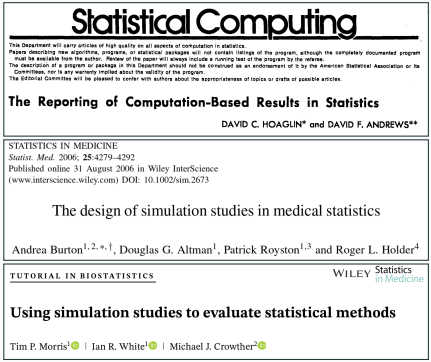
Using simulation studies to evaluate statistical methods

Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

This project:

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 contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running 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[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

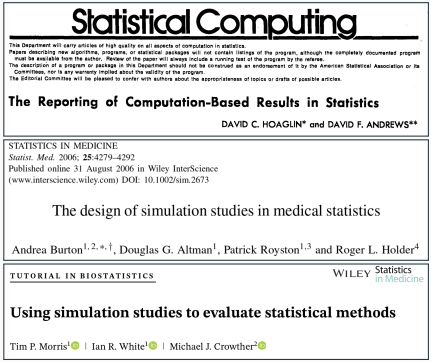
Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

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)



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 contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running 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[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

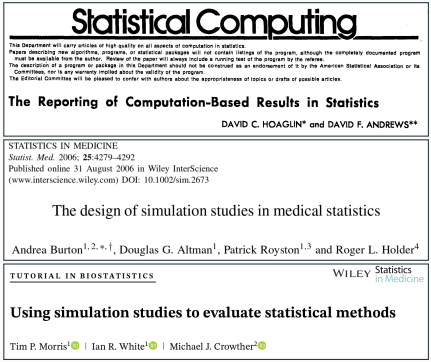
Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

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)



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 contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running 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[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

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

Psychological Methods

ISSN: 1082-989X

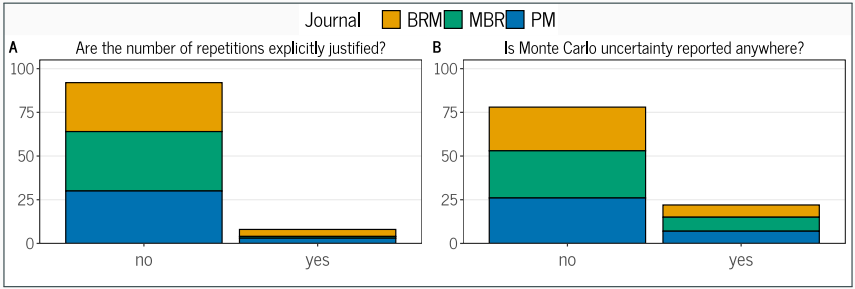
https://doi.org/10.1037/met0000695

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

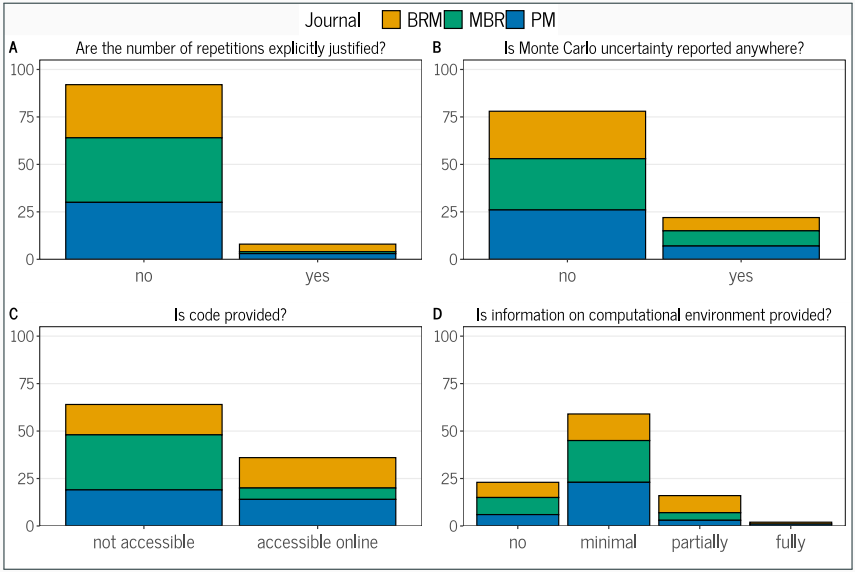
Björn S. Siepe[1], František Bartoš[2], Tim P. Morris[3], Anne-Laure Boulesteix[4, 5], Daniel W. Heck[1], and Samuel Pawel[6, 7]

Main Results

Main Results



Main Results



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 (Strobl and Leisch, 2024)

Simulation study protocols

Simulation study protocols

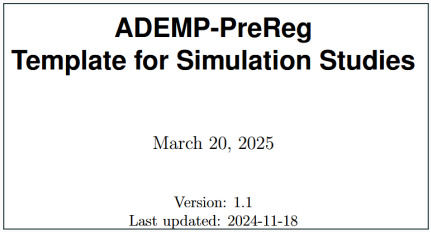
STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1,3] and Roger L. Holder[4]

"***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)

The ADEMP-PreReg template


ADEMP-PreReg
Template for Simulation Studies

March 20, 2025

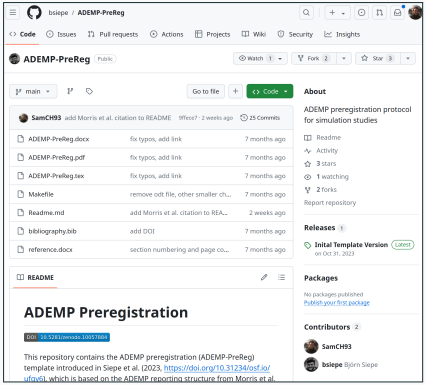
Version: 1.1
Last updated: 2024-11-18

Protocol template based on:

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

ADEMP: doi.org/10.1214/ss/1009213726, ADEMP-PreReg: doi.org/10.31234/osf.io/ufgy6

The ADEMP-PreReg template – A living document



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

LaTeX, Overleaf

MS/Libre office, Google docs

Open Science Framework

The ADEMP-PreReg template – Overview

1. Instructions
2. General information
3. **A**ims
4. **D**ata-generating mechanism
5. **E**stimands and targets
6. **M**ethods
7. **P**erformance 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

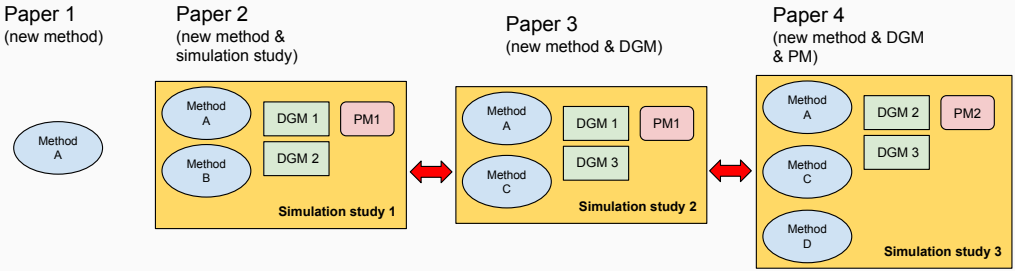
doi:10.5281/zenodo.7994221

Limitations

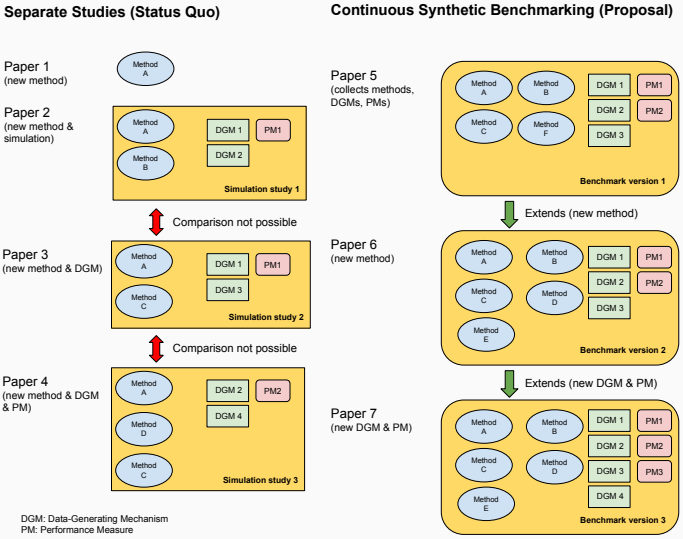
- Preregistration could be **faked**
- May **slow down** exploratory research

Incomparable Simulations

Incomparable Simulations



WIP: Synthetic benchmarking



Conclusions

- **Simulation studies** are ubiquitous in methodological research
- Simulation studies can be impacted by **questionable research practices** and misaligned **incentives**
- Adopting strategies from other fields has the potential to improve simulation studies
- Meta-research, discussions, and reforms needed to **increase awareness** and **improve standards**

A multidisciplinary collaboration

František Bartoš

Daniel W. Heck

Tim P. Morris

A.-L. Boulesteix

Anna Lohmann

Samuel Pawel

Get In Touch

- bjoern.siepe@uni-marburg.de
- https://bsiepe.github.io/

Paper & Slides

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.

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.

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.

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.

References ii

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.

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.