On the Design of Simulation Studies

Anne-Laure Boulesteix, Tim Morris, Michal Abrahamowicz
Inst. for Medical Information Processing, Biometry, and Epidemiology
LMU Munich, Germany

for the Simulation Panel of the STRATOS initiative

September 10th 2021, ROeS Conference, Salzburg

The simulation panel

- Main goal of STRATOS: providing evidence-based guidance for the choice of statistical methods
- What is “evidence” in the methodological context?

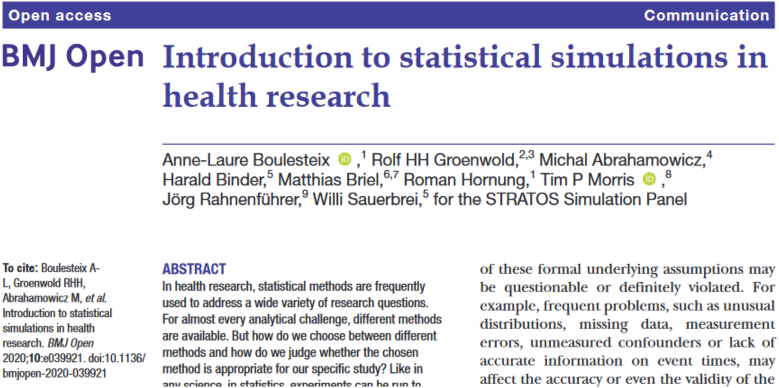
→ Key role of simulations!

- Main goal of simulation panel: deriving guidance to design, perform and report simulation studies

Chairs: Michal Abrahamowicz, Anne-Laure Boulesteix

Members: Harald Binder, Rolf Groenwold, Victor Kipnis, Jessica Myers Franklin, Tim Morris, Willi Sauerbrei, Pamela Shaw, Ewout Steyerberg, Ingeborg Waernbaum

An introduction for level 1 audience



Open access | Communication

BMJ Open

Introduction to statistical simulations in health research

Anne-Laure Boulesteix ,[1] Rolf HH Groenwold,[2,3] Michal Abrahamowicz,[4] Harald Binder,[5] Matthias Briel,[6,7] Roman Hornung,[1] Tim P Morris ,[8] Jörg Rahnenführer,[9] Willi Sauerbrei,[5] for the STRATOS Simulation Panel

To cite: Boulesteix A-L, Groenwold RHH, Abrahamowicz M, *et al*. Introduction to statistical simulations in health research. *BMJ Open* 2020;**10**:e039921. doi:10.1136/bmjopen-2020-039921

ABSTRACT

In health research, statistical methods are frequently used to address a wide variety of research questions. For almost every analytical challenge, different methods are available. But how do we choose between different methods and how do we judge whether the chosen method is appropriate for our specific study? Like in any science, in statistics, experiments can be run to

of these formal underlying assumptions may be questionable or definitely violated. For example, frequent problems, such as unusual distributions, missing data, measurement errors, unmeasured confounders or lack of accurate information on event times, may affect the accuracy or even the validity of the

Ordinal endpoints in randomized clinical trials

Example: Mantle Cell Lymphoma (MCL) elderly trial
(Kluin-Nelemans, Hoster et al., NJEM 2012)

Treatment response	R-FC	R-CHOP
	$n = 246$	$n = 239$
Early death	8 (3%)	9 (3%)
Progressive disease	35 (14%)	13 (5%)
Stable disease	11 (4%)	11 (5%)
Partial response	62 (25%)	89 (37%)
Complete remission, unconfirmed	32 (13%)	36 (15%)
Complete remission, confirmed	98 (40%)	81 (34%)

Available methods

1. **Wilcoxon test**
2. Cochran-Armitage trend test
3. Proportional odds logistic regression
4. **Dichotomization**, then chi-square/Fisher for 2 × 2 table
5. **Chi-square/Fisher's exact test** for $2 \times K$ table
6. Tests based on maximally selected chi-square statistics
   - exact (Boulesteix, Biometrical Journal 2006)
   - asymptotic (Boulesteix et al., SAGMB 2007)

Non-neutrality disclosure

A statistician
'WE DEMONSTRATE WITH A [DODGY] SIMULATION STUDY THAT OUR PROPOSED METHOD X HAS EXCELLENT PERFORMANCE'

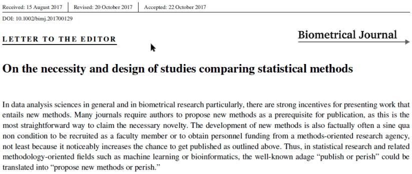
(picture by T. Morris)

- I developed the tests based on maximally selected chi-square statistics
- But I am not under pressure to make them look good (no grant, no PhD student on this project)

→ Not neutral, but hopefully not strongly biased

A plea for “neutral comparison studies”

- not introducing any new method
- neutral authors (unbiased, equally familiar with methods)



Received: 15 August 2017 | Revised: 20 October 2017 | Accepted: 22 October 2017

DOI: 10.1002/bimj.201700129

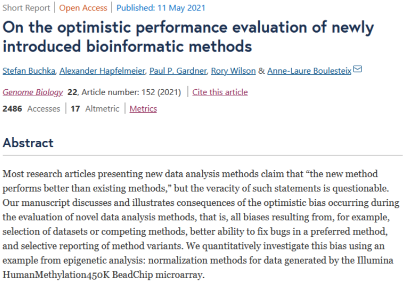
LETTER TO THE EDITOR

Biometrical Journal

On the necessity and design of studies comparing statistical methods

In data analysis sciences in general and in biometrical research particularly, there are strong incentives for presenting work that entails new methods. Many journals require authors to propose new methods as a prerequisite for publication, as this is the most straightforward way to claim the necessary novelty. The development of new methods is also factually often a sine qua non condition to be recruited as a faculty member or to obtain personnel funding from a methods-oriented research agency, not least because it noticeably increases the chance to get published as outlined above. Thus, in statistical research and related methodology-oriented fields such as machine learning or bioinformatics, the well-known adage “publish or perish” could be translated into “propose new methods or perish.”

Letter by the simulation panel (ALB, Abrahamowicz, Binder & Sauerbrei, 2018)



Short Report | Open Access | Published: 11 May 2021

On the optimistic performance evaluation of newly introduced bioinformatic methods

Stefan Buchka, Alexander Hapfelmeier, Paul P. Gardner, Rory Wilson & Anne-Laure Boulesteix ✉

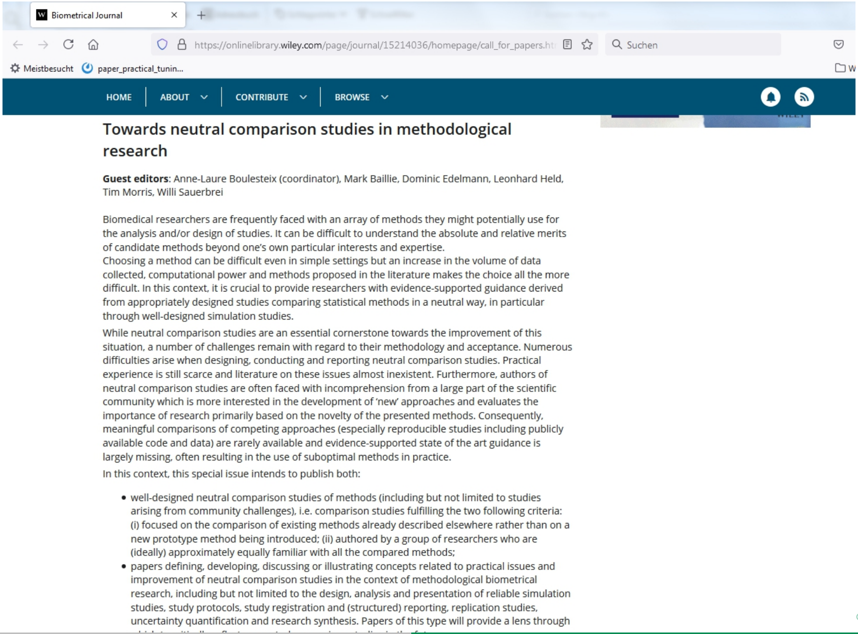
Genome Biology **22**, Article number: 152 (2021) | Cite this article

2486 Accesses | **17** Altmetric | Metrics

Abstract

Most research articles presenting new data analysis methods claim that “the new method performs better than existing methods,” but the veracity of such statements is questionable. Our manuscript discusses and illustrates consequences of the optimistic bias occurring during the evaluation of novel data analysis methods, that is, all biases resulting from, for example, selection of datasets or competing methods, better ability to fix bugs in a preferred method, and selective reporting of method variants. We quantitatively investigate this bias using an example from epigenetic analysis: normalization methods for data generated by the Illumina HumanMethylation450K BeadChip microarray.

Anecdotal evidence of the “new method bias”



Towards neutral comparison studies in methodological research

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

Biomedical researchers are frequently faced with an array of methods they might potentially use for the analysis and/or design of studies. It can be difficult to understand the absolute and relative merits of candidate methods beyond one's own particular interests and expertise.
Choosing a method can be difficult even in simple settings but an increase in the volume of data collected, computational power and methods proposed in the literature makes the choice all the more difficult. In this context, it is crucial to provide researchers with evidence-supported guidance derived from appropriately designed studies comparing statistical methods in a neutral way, in particular through well-designed simulation studies.

While neutral comparison studies are an essential cornerstone towards the improvement of this situation, a number of challenges remain with regard to their methodology and acceptance. Numerous difficulties arise when designing, conducting and reporting neutral comparison studies. Practical experience is still scarce and literature on these issues almost inexistent. Furthermore, authors of neutral comparison studies are often faced with incomprehension from a large part of the scientific community which is more interested in the development of 'new' approaches and evaluates the importance of research primarily based on the novelty of the presented methods. Consequently, meaningful comparisons of competing approaches (especially reproducible studies including publicly available code and data) are rarely available and evidence-supported state of the art guidance is largely missing, often resulting in the use of suboptimal methods in practice.

In this context, this special issue intends to publish both:

- well-designed neutral comparison studies of methods (including but not limited to studies arising from community challenges), i.e. comparison studies fulfilling the two following criteria: (i) focused on the comparison of existing methods already described elsewhere rather than on a new prototype method being introduced; (ii) authored by a group of researchers who are (ideally) approximately equally familiar with all the compared methods;
- papers defining, developing, discussing or illustrating concepts related to practical issues and improvement of neutral comparison studies in the context of methodological biometrical research, including but not limited to the design, analysis and presentation of reliable simulation studies, study protocols, study registration and (structured) reporting, replication studies, uncertainty quantification and research synthesis. Papers of this type will provide a lens through

Simulation study

Y: the treatment group ($Y = \{0, 1\}$)
In our simulation: $P(Y = 0) = P(Y = 1) = 0.5$

X: the ordinal endpoint ($X \in \{1, \ldots, K\}$)

$\pi_{i,k} = P(X = k|Y = i)$, for $i = 0, 1$ and $k = 1, \ldots, K$

$H_0 : \forall k \; \pi_{0,k} = \pi_{1,k}$
$H_1 : \exists k \; \pi_{0,k} \neq \pi_{1,k}$

Simulation settings are characterised by n and $\pi_{i,k}$.

Which $\pi_{i,k}$ should we consider?

A matter of perspective:

- **Methodological statistician**: interested in general trends and in settings that allow for observation and understanding of the differences between methods

→ calls for special settings that are not necessarily realistic

Example: $\boldsymbol{\pi}_0 = (1/6, 2/6, 3/6)^\top$, $\boldsymbol{\pi}_1 = (1/6, 3/6, 2/6)^\top$

- **Applied statistician**: interested in their own setting (observed—at the analysis stage, or assumed—at the planning stage)

→ calls for "representative" realistic settings

Realistic setting: example

	R-FC $n = 246$	R-CHOP $n = 239$
Early death	8	9
Progressive disease	35	13
Stable disease	11	11
Partial response	62	89
Complete remisision, unconfirmed	32	36
Complete remission, confirmed	98	81

This yields the "realistic setting":

- $\pi_0 := \hat{p}_{RFC} = (0.03, 0.14, 0.04, 0.25, 0.13, 0.40)^\top$
- $\pi_1 := \hat{p}_{RCHOP} = (0.03, 0.05, 0.05, 0.37, 0.15, 0.34)^\top$

This is only one example. One could use *many* such trials sampled from the *population* of trials with ordinal endpoint to derive realistic, representative π_0's and π_1's.

Summary of results

- no universal best method in terms of power
- own interpretation for each test: in terms of medians, odds, cutpoint, etc.
- maxselE and (to a lesser extent) lrm have increased type 1 error for small n
- Fisher and chi-square perform suboptimally for large K and for trends (as opposed to cutpoints)
- trend test and Wilcoxon perform well overall, but fail in case of non-monotonous pattern and are outperformed for $K = 3$.
- price of maxselA's interpretability is **(sometimes/often)** a (slightly) reduced power

Methodological challenges (choice of simulation settings)

- What does **sometimes/often** mean? It implicitly refers to a population of scenarios, but how is this population defined?
- Infinitely many potentially relevant scenarios
  → making simulation script available?
  → starting replicating simulation studies? (Lohmann et al., 2021)
- Non-neutrality of simulation design
  → crowd-sourcing the design of simulations?
  → starting replicating simulation studies? (Lohmann et al., 2021)

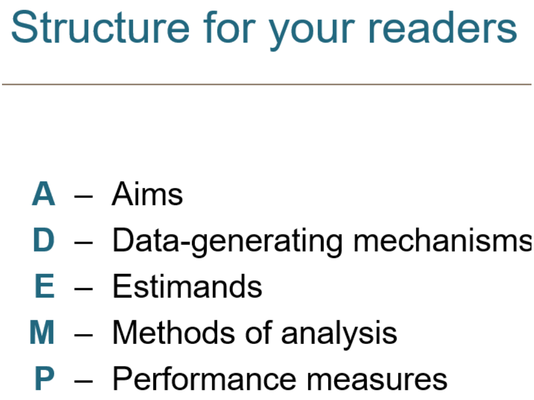
Further issues (from Tim Morris' slides)

- Regarding the number of repetitions:

  “We need to quantify uncertainty due to using a finite number of repetitions (Monte Carlo error).”
- Regarding the study's aim:

  “Think of different phases: proof-of-concept (like pre-clinical work), trying to hone a method (like dose-finding), comparison of competing methods in non-ideal situations (phase III), understanding when a method breaks (phase IV)”

Reporting simulation studies



Structure for your readers

A – Aims

D – Data-generating mechanisms

E – Estimands

M – Methods of analysis

P – Performance measures

(Morris et al., SIM 2019; slide by T. Morris)

Towards structured reporting of simulation studies

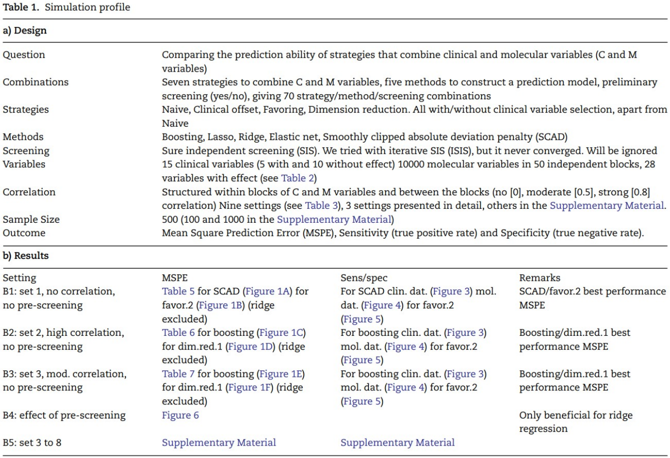


Table 1. Simulation profile

a) Design	
Question	Comparing the prediction ability of strategies that combine clinical and molecular variables (C and M variables)
Combinations	Seven strategies to combine C and M variables, five methods to construct a prediction model, preliminary screening (yes/no), giving 70 strategy/method/screening combinations
Strategies	Naive, Clinical offset, Favoring, Dimension reduction. All with/without clinical variable selection, apart from Naive
Methods	Boosting, Lasso, Ridge, Elastic net, Smoothly clipped absolute deviation penalty (SCAD)
Screening	Sure independent screening (SIS). We tried with iterative SIS (ISIS), but it never converged. Will be ignored
Variables	15 clinical variables (5 with and 10 without effect) 10000 molecular variables in 50 independent blocks, 28 variables with effect (see Table 2)
Correlation	Structured within blocks of C and M variables and between the blocks (no [0], moderate [0.5], strong [0.8] correlation) Nine settings (see Table 3), 3 settings presented in detail, others in the Supplementary Material.
Sample Size	500 (100 and 1000 in the Supplementary Material)
Outcome	Mean Square Prediction Error (MSPE), Sensitivity (true positive rate) and Specificity (true negative rate).

b) Results

Setting	MSPE	Sens/spec	Remarks
B1: set 1, no correlation, no pre-screening	Table 5 for SCAD (Figure 1A) for favor.2 (Figure 1B) (ridge excluded)	For SCAD clin. dat. (Figure 3) mol. dat. (Figure 4) for favor.2 (Figure 5)	SCAD/favor.2 best performance MSPE
B2: set 2, high correlation, no pre-screening	Table 6 for boosting (Figure 1C) for dim.red.1 (Figure 1D) (ridge excluded)	For boosting clin. dat. (Figure 3) mol. dat. (Figure 4) for favor.2 (Figure 5)	Boosting/dim.red.1 best performance MSPE
B3: set 3, mod. correlation, no pre-screening	Table 7 for boosting (Figure 1E) for dim.red.1 (Figure 1F) (ridge excluded)	For boosting clin. dat. (Figure 3) mol. dat. (Figure 4) for favor.2 (Figure 5)	Boosting/dim.red.1 best performance MSPE
B4: effect of pre-screening	Figure 6		Only beneficial for ridge regression
B5: set 3 to 8	Supplementary Material	Supplementary Material	

De Bin et al. (Briefings in Bioinformatics, 2020)

Translational simulation studies — Remember...

A matter of perspective:

- **Applied statistician**: interested in their own setting (observed—at the analysis stage, or assumed—at the planning stage)

But: impossible to cover all relevant settings in a single study

Solution: sharing simulation code in user-friendly way?

Thank you for your attention!

Thanks to all STRATOS colleagues, in particular Tim Morris, Michal Abrahamowicz and Willi Sauerbrei, and to the DFG for funding!