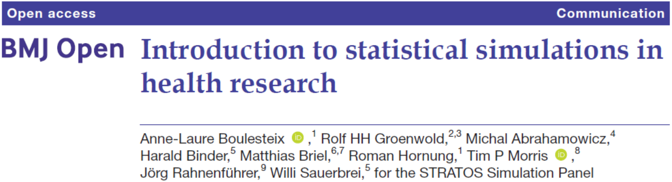
“Real-data-based” simulation studies

Anne-Laure Boulesteix, Christina Sauer, F. J. D. Lange
Inst. for Medical Information Processing, Biometry, and Epidemiology
LMU Munich and Munich Center for Machine Learning, Germany

June 3rd, 2025, Workshop “Designing Synthetic Benchmarks for Real-World Cohort Data”, Freiburg

Simulation studies: in which context?

- in **applications**, i.e. with focus on a **particular study** that has to be planned or analysed by the team conducting the simulation
- in **methodological research**, i.e. with focus on the general behaviour of methods in a whole application domain



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

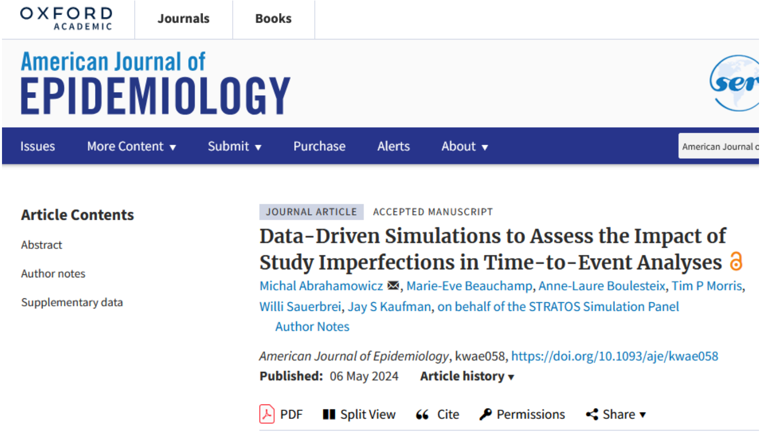
Introduction

Real-data-based simulations: applications

Real-data-based simulations: methodological

Miscellaneous

Applications: Abrahamowicz et al.



American Journal of EPIDEMIOLOGY

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Data-Driven Simulations to Assess the Impact of Study Imperfections in Time-to-Event Analyses

Michal Abrahamowicz, Marie-Eve Beauchamp, Anne-Laure Boulesteix, Tim P Morris, Willi Sauerbrei, Jay S Kaufman, on behalf of the STRATOS Simulation Panel

Author Notes

American Journal of Epidemiology, kwae058, https://doi.org/10.1093/aje/kwae058

Published: 06 May 2024 **Article history** ▾

Applications: Other examples

- Simulations to investigate the impact of measurement error (Brakenhoff et al., 2018; Boulesteix et al. for STRATOS Simulation Panel, 2020)
- Simulations for sample size calculations when there is no closed formula
- ...

Introduction

Real-data-based simulations: applications

Real-data-based simulations: methodological

Miscellaneous

Simulations in methodological research should...

- ... help us understand methods, when they work and when they don't, when they break down, etc.

- ... help us define standard methods and distinguish "niche methods" from methods that are safe to be used in common settings.

- ... be neutral.

Simulations in methodological research should...

- ... help us understand methods, when they work and when they don't, when they break down, etc.

→ Extreme (perhaps unrealistic) scenarios are needed.

- ... help us define standard methods and distinguish "niche methods" from methods that are safe to be used in common settings.

- ... be neutral.

Simulations in methodological research should...

- ... help us understand methods, when they work and when they don't, when they break down, etc.

→ Extreme (perhaps unrealistic) scenarios are needed.

- ... help us define standard methods and distinguish "niche methods" from methods that are safe to be used in common settings.

→ Realistic scenarios reflecting real data are needed.

- ... be neutral.

Simulations in methodological research should...

- ... help us understand methods, when they work and when they don't, when they break down, etc.

→ Extreme (perhaps unrealistic) scenarios are needed.

- ... help us define standard methods and distinguish "niche methods" from methods that are safe to be used in common settings.

→ Realistic scenarios reflecting real data are needed.

- ... be neutral.

→ The choice of scenarios should be "representative" for reality and not focus specifically on settings that favor one or the other method.

Real-data-based simulations

→ Extreme (perhaps unrealistic) scenarios are needed.

→ Realistic scenarios reflecting real data are needed.

→ The choice of scenarios should be “representative” for reality and not focus specifically on settings that favor one or the other method.

It calls for simulation studies that:

1. use real data sets as a basis
2. use a representative set of real data sets as a basis

Real-data-based simulations: What is that?

- **parametric**

→ parameters/characteristics extracted from a real data set are used in parametric data-generating mechanisms

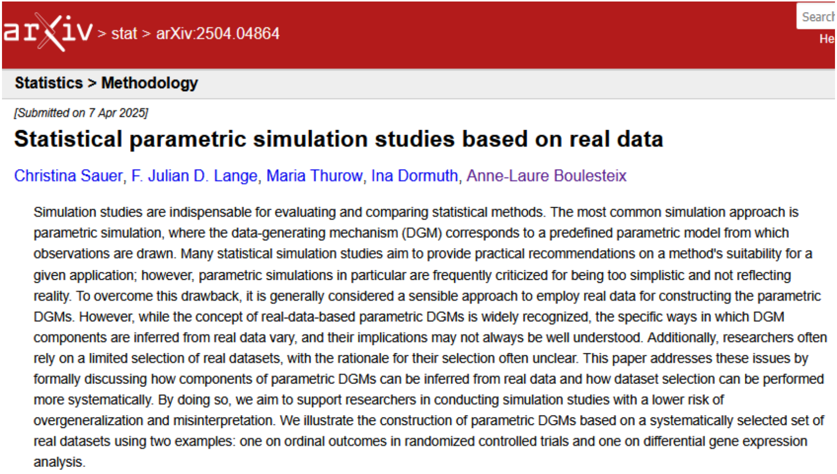
- **non-parametric**

→ e.g., sampling with/without replacement from a real data set

- **semi-parametric**

→ plasmode simulations as an important special case (Schreck et al., Stat Med 2024)

Sauer, Lange, et al., 2025



arXiv > stat > arXiv:2504.04864

Statistics > Methodology

[Submitted on 7 Apr 2025]

Statistical parametric simulation studies based on real data

Christina Sauer, F. Julian D. Lange, Maria Thurow, Ina Dormuth, Anne-Laure Boulesteix

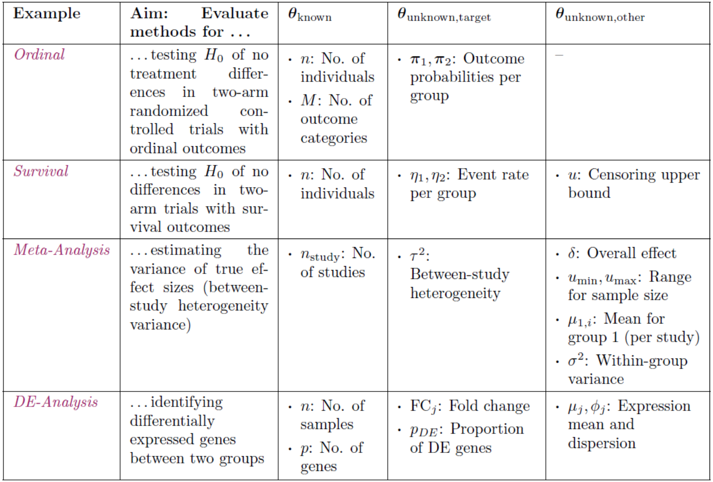
Simulation studies are indispensable for evaluating and comparing statistical methods. The most common simulation approach is parametric simulation, where the data-generating mechanism (DGM) corresponds to a predefined parametric model from which observations are drawn. Many statistical simulation studies aim to provide practical recommendations on a method's suitability for a given application; however, parametric simulations in particular are frequently criticized for being too simplistic and not reflecting reality. To overcome this drawback, it is generally considered a sensible approach to employ real data for constructing the parametric DGMs. However, while the concept of real-data-based parametric DGMs is widely recognized, the specific ways in which DGM components are inferred from real data vary, and their implications may not always be well understood. Additionally, researchers often rely on a limited selection of real datasets, with the rationale for their selection often unclear. This paper addresses these issues by formally discussing how components of parametric DGMs can be inferred from real data and how dataset selection can be performed more systematically. By doing so, we aim to support researchers in conducting simulation studies with a lower risk of overgeneralization and misinterpretation. We illustrate the construction of parametric DGMs based on a systematically selected set of real datasets using two examples: one on ordinal outcomes in randomized controlled trials and one on differential gene expression analysis.

Sauer, Lange, et al., 2025

Tasks in real-data-based parametric simulations:

- selection of the real data sets
  - in principle similar to benchmarking studies
  - should be well-justified
- extraction (from real data sets) of relevant information informing the data generating mechanism:
  - known parameters
  - unknown parameters (to be estimated):
    - related to the method's target
    - ... or not

Sauer, Lange, et al., 2025—Examples



Example	**Aim: Evaluate methods for . . .**	$\boldsymbol{\theta}_{\text{known}}$	$\boldsymbol{\theta}_{\text{unknown,target}}$	$\boldsymbol{\theta}_{\text{unknown,other}}$
Ordinal	. . . testing H_0 of no treatment differences in two-arm randomized controlled trials with ordinal outcomes	• n: No. of individuals • M: No. of outcome categories	• π_1, π_2: Outcome probabilities per group	–
Survival	. . . testing H_0 of no differences in two-arm trials with survival outcomes	• n: No. of individuals	• η_1, η_2: Event rate per group	• u: Censoring upper bound
Meta-Analysis	. . . estimating the variance of true effect sizes (between-study heterogeneity variance)	• n_{study}: No. of studies	• τ^2: Between-study heterogeneity	• δ: Overall effect • $u_{\min}, u_{\max}$: Range for sample size • $\mu_{1,i}$: Mean for group 1 (per study) • σ^2: Within-group variance
DE-Analysis	. . . identifying differentially expressed genes between two groups	• n: No. of samples • p: No. of genes	• FC$_j$: Fold change • p_{DE}: Proportion of DE genes	• μ_j, ϕ_j: Expression mean and dispersion

Sauer, Lange, et al., 2025—Results

- Results of real-data-based parametric simulations may substantially deviate from those obtained using simple fully researcher-defined scenarios.
- Results of real-data-based parametric simulations strongly depend on the considered real data set(s).
  - → Results obtained with *one* real data set are but a point in the space of all possible results.

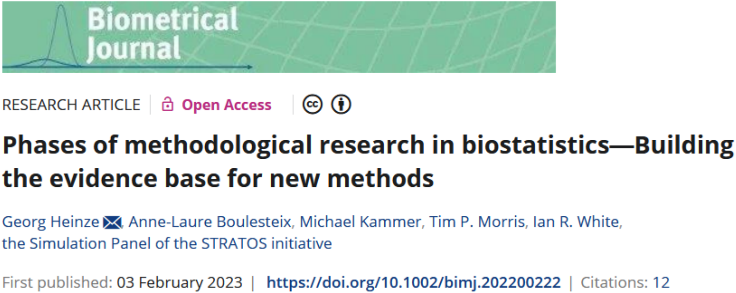
Introduction

Real-data-based simulations: applications

Real-data-based simulations: methodological

Miscellaneous

The phases of methodological research



Biometrical Journal

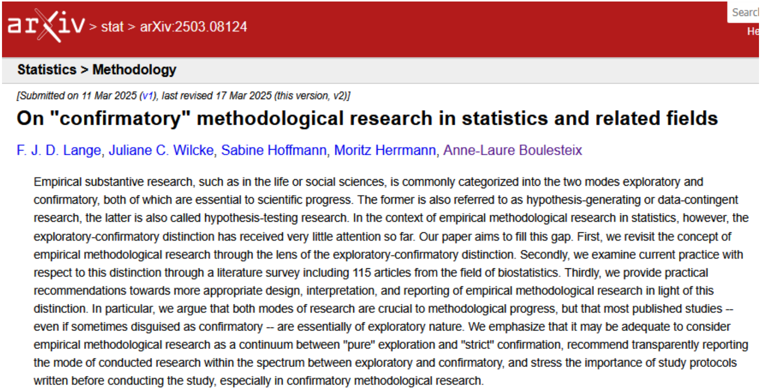
RESEARCH ARTICLE | Open Access

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

Georg Heinze, Anne-Laure Boulesteix, Michael Kammer, Tim P. Morris, Ian R. White, the Simulation Panel of the STRATOS initiative

First published: 03 February 2023 | https://doi.org/10.1002/bimj.202200222 | Citations: 12

Confirmatory methodological research



arXiv > stat > arXiv:2503.08124

Statistics > Methodology

[Submitted on 11 Mar 2025 (v1), last revised 17 Mar 2025 (this version, v2)]

On "confirmatory" methodological research in statistics and related fields

F. J. D. Lange, Juliane C. Wilcke, Sabine Hoffmann, Moritz Herrmann, Anne-Laure Boulesteix

Empirical substantive research, such as in the life or social sciences, is commonly categorized into the two modes exploratory and confirmatory, both of which are essential to scientific progress. The former is also referred to as hypothesis-generating or data-contingent research, the latter is also called hypothesis-testing research. In the context of empirical methodological research in statistics, however, the exploratory-confirmatory distinction has received very little attention so far. Our paper aims to fill this gap. First, we revisit the concept of empirical methodological research through the lens of the exploratory-confirmatory distinction. Secondly, we examine current practice with respect to this distinction through a literature survey including 115 articles from the field of biostatistics. Thirdly, we provide practical recommendations towards more appropriate design, interpretation, and reporting of empirical methodological research in light of this distinction. In particular, we argue that both modes of research are crucial to methodological progress, but that most published studies -- even if sometimes disguised as confirmatory -- are essentially of exploratory nature. We emphasize that it may be adequate to consider empirical methodological research as a continuum between "pure" exploration and "strict" confirmation, recommend transparently reporting the mode of conducted research within the spectrum between exploratory and confirmatory, and stress the importance of study protocols written before conducting the study, especially in confirmatory methodological research.

Confirmatory methodological research

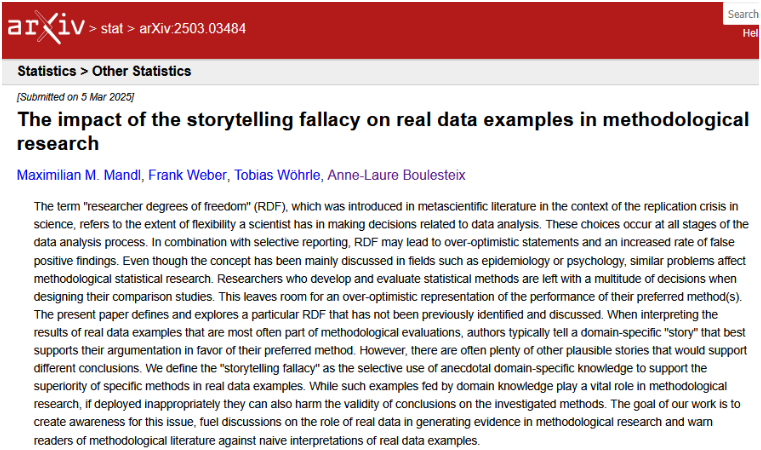
Our argument:

- The well-known distinction between confirmatory and exploratory research should be given attention in methodological research as well.
- Most currently published methodological studies are of exploratory nature.
- Both types of studies are crucial to scientific progress.

→ We need more confirmatory research.

Lange et al., 2025. https://arxiv.org/pdf/2503.08124v2

The storytelling fallacy



Statistics > Other Statistics

[Submitted on 5 Mar 2025]

The impact of the storytelling fallacy on real data examples in methodological research

Maximilian M. Mandl, Frank Weber, Tobias Wöhrle, Anne-Laure Boulesteix

The term "researcher degrees of freedom" (RDF), which was introduced in metascientific literature in the context of the replication crisis in science, refers to the extent of flexibility a scientist has in making decisions related to data analysis. These choices occur at all stages of the data analysis process. In combination with selective reporting, RDF may lead to over-optimistic statements and an increased rate of false positive findings. Even though the concept has been mainly discussed in fields such as epidemiology or psychology, similar problems affect methodological statistical research. Researchers who develop and evaluate statistical methods are left with a multitude of decisions when designing their comparison studies. This leaves room for an over-optimistic representation of the performance of their preferred method(s). The present paper defines and explores a particular RDF that has not been previously identified and discussed. When interpreting the results of real data examples that are most often part of methodological evaluations, authors typically tell a domain-specific "story" that best supports their argumentation in favor of their preferred method. However, there are often plenty of other plausible stories that would support different conclusions. We define the "storytelling fallacy" as the selective use of anecdotal domain-specific knowledge to support the superiority of specific methods in real data examples. While such examples fed by domain knowledge play a vital role in methodological research, if deployed inappropriately they can also harm the validity of conclusions on the investigated methods. The goal of our work is to create awareness for this issue, fuel discussions on the role of real data in generating evidence in methodological research and warn readers of methodological literature against naive interpretations of real data examples.

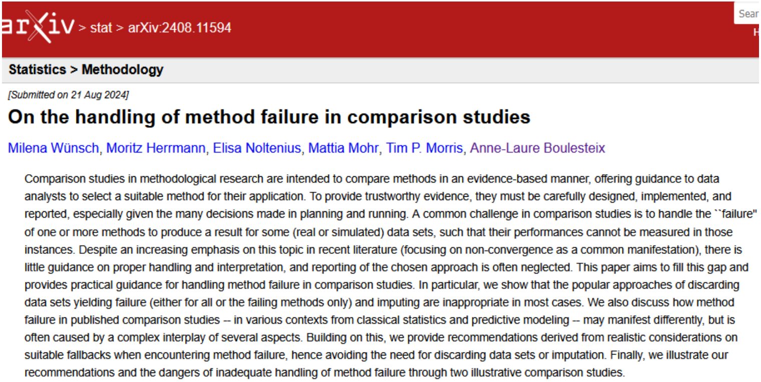
The storytelling fallacy

- In methodological articles, domain-based knowledge is sometimes used to argue in favor of the new method's superiority in the context of real-data applications.

  Example: From a clinical point of view, the variables selected by method A make more sense than those selected by method B, so method A is better.
- Our argument: such statements are misleading/over-optimistic because:
  - They are often based on $n = 1$ data set.
  - More importantly, there are often plenty of other plausible stories that would support different conclusions.

Mandl et al., 2025. https://arxiv.org/abs/2503.03484

What if a simulation repetition does not yield any result?



Statistics > Methodology

[Submitted on 21 Aug 2024]

On the handling of method failure in comparison studies

Milena Wünsch, Moritz Herrmann, Elisa Noltenius, Mattia Mohr, Tim P. Morris, Anne-Laure Boulesteix

Comparison studies in methodological research are intended to compare methods in an evidence-based manner, offering guidance to data analysts to select a suitable method for their application. To provide trustworthy evidence, they must be carefully designed, implemented, and reported, especially given the many decisions made in planning and running. A common challenge in comparison studies is to handle the ``failure" of one or more methods to produce a result for some (real or simulated) data sets, such that their performances cannot be measured in those instances. Despite an increasing emphasis on this topic in recent literature (focusing on non-convergence as a common manifestation), there is little guidance on proper handling and interpretation, and reporting of the chosen approach is often neglected. This paper aims to fill this gap and provides practical guidance for handling method failure in comparison studies. In particular, we show that the popular approaches of discarding data sets yielding failure (either for all or the failing methods only) and imputing are inappropriate in most cases. We also discuss how method failure in published comparison studies -- in various contexts from classical statistics and predictive modeling -- may manifest differently, but is often caused by a complex interplay of several aspects. Building on this, we provide recommendations derived from realistic considerations on suitable fallbacks when encountering method failure, hence avoiding the need for discarding data sets or imputation. Finally, we illustrate our recommendations and the dangers of inadequate handling of method failure through two illustrative comparison studies.

Wünsch et al., 2025. Statistics in Medicine (cond. accepted).

Thanks!

M. Herrmann
F. J. D. Lange
M. Mandl
C. Sauer (née Nießl)
M. Wünsch
M. Abrahamowicz (Montreal, CA)
G. Heinze (Vienna, AT)
T. Morris (UCL, UK)
W. Sauerbrei (Freiburg, DE)
M. Thurow, I. Dormuth (Dortmund, DE)

Sauer, Lange, et al., 2025:

arXiv:2504.04864

Thanks to the DFG (BO3139/4-3, 7-1, 7-2, 9-1) and the STRATOS Initiative!