

Neutral comparisons of statistical methods: opportunities, challenges, and common issues

Theresa Ullmann¹, Georg Heinze¹, Anne-Laure-Boulesteix², Daniela Dunkler¹

¹Medical University of Vienna, ²LMU Munich

September 17th, 2025

Statistical methods: from innovation... to implementation?

Biometrics

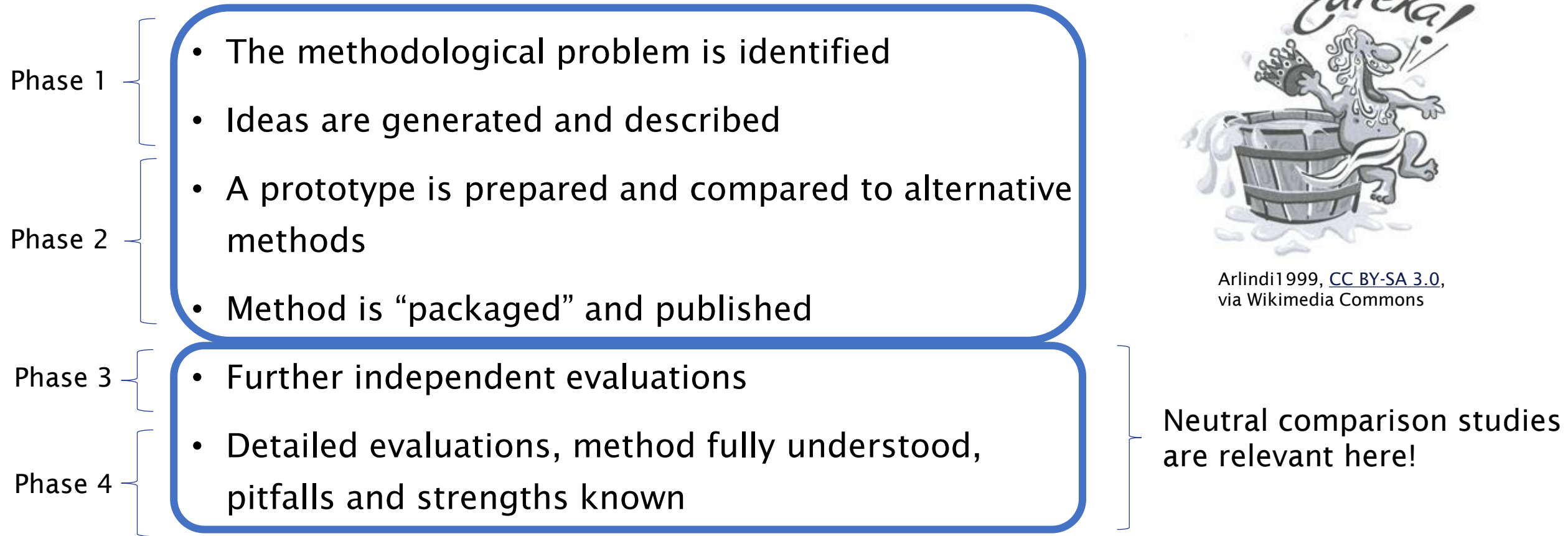


BMC Medical Research Methodology



- Every issue of these journals is full of newly developed methods
- Lots of *innovation* – but what about *implementation*?
- How many of these new methods find their way into routine applications?

From ideas to trustworthy applications: a journey



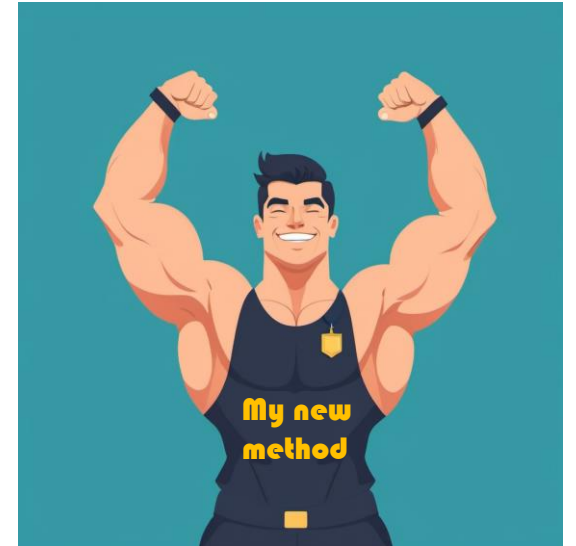
Arlindi1999, [CC BY-SA 3.0](#),
via Wikimedia Commons

Heinze, Boulesteix et al., *Biometrical Journal* (2024):
Phases of methodological research in biostatistics

Why are neutral comparison studies important?

Articles which introduce *new methods*...

- ... usually only include limited comparisons, and
 - ... are likely to be optimistically biased towards the new methods
- Typical claim: “This new method outperforms competitors!”



<https://openart.ai>

Neutral comparison studies can provide a more comprehensive and less biased picture!

What are neutral comparison studies?

- The design of a **comparison study** includes
 - the *methods* under comparison,
 - the *data sets* (real or simulated) on which the methods are compared,
 - the *performance criteria* used to assess the methods.
- A **neutral** comparison study is characterized by ...
 - the *comparison of existing* methods (not introducing a new method), and
 - *neutrality* of the authors, i.e.,
 - the authors do not have a vested interest in a particular method, and
 - are as a group approximately equally familiar with all considered methods.

Neutral comparison studies: providing guidance

- Neutral comparison studies can provide *evidence-supported guidance*:

“Which statistical method should I use in which situation?”

- Traditionally, it was rather difficult to publish and/or get funding for neutral comparison studies – “We want new methods!”
- In recent years, some steps forward:
 - STRATOS initiative
 - Special issue in Biometrical Journal

Steps forward



STRengthening Analytical Thinking for Observational Studies

Accessible and Accurate Guidance in the Design and Analysis of Observational Studies

<https://stratos-initiative.org/>

STRATOS advocates for
neutral comparison studies!

23 articles



Special Collection: “Neutral Comparison Studies in Methodological Research”

Virtual Issues | First published: 14 December 2023 | Last updated: 19 February 2024

Biometricians 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 Edelman, Leonhard Held, Tim Morris, Willi Sauerbrei

Challenges in neutral comparison studies

- Suppose you want to perform a neutral comparison study
- You will likely encounter some issues and challenges
- I will illustrate some of these using an example study

PLOS ONE

Evaluating variable selection methods for multivariable regression models: A simulation study protocol

Theresa Ullmann, Georg Heinze, Lorena Hafermann, Christine Schilhart-Wallisch, Daniela Dunkler ,
for TG2 of the STRATOS initiative 

Published: August 9, 2024 • <https://doi.org/10.1371/journal.pone.0308543>

An example: comparison of variable selection methods

- Comparison of frequently used data-driven variable selection methods for multivariable regression (linear or logistic)
- Methods: e.g. backward elimination, Lasso,...
- Simulated data (“true model” known)
- Performance criteria: e.g. selection rates of variables (true predictors and noise variables), bias of regression coefficients,...

Perfect neutrality: always realistic?

- Openly disclose possible conflicts of interest

As mentioned above, neutrality is an important goal when conducting systematic comparison studies. “Perfect” neutrality may be the ultimate goal, but this ideal can be difficult to achieve in practice. While we aim to be as neutral as possible, we disclose (for the purpose of full transparency) that one of the methods for variable selection included in our comparison, namely augmented backwards elimination, was originally proposed by two authors of the present study protocol [15]. Our goal was to not let this fact influence our choice of study design, though unconscious biases can never be fully excluded. Striving for as much neutrality as pos-

Ullmann et al. 2024, PLOS ONE

- But: using a disclaimer does not mean “anything goes”!

Perfect neutrality: always realistic?

Strategies to improve neutrality and “equal familiarity with methods”:

- Ask other experts to contribute (in our case: feedback from fellow STRATOS members)
- Pre-registration: publish a peer-reviewed study protocol
 - You get feedback from reviewers
 - Prevents you from changing the study design based on the results (at least not without good reason!)
- “Blinding” (not used in our study, but interesting concept): different persons/teams perform...
 - ...data simulation (“dataset A”, “dataset B”,...)
 - ...method implementation (“method X”, “method Y”,...)
 - ...results evaluation (“on dataset A, method Y performs better than method X”,...)

Treat all methods fairly

- The data generating mechanisms should not favour any specific method
 - Should be okay in our study
 - Counter-example: when linear regression is compared with machine learning methods, and data is only generated according to a linear model with only linear effects
- The scope of parameter tuning should be comparable across all methods
 - In our study: methods are relatively straightforward, tuning of hyperparameters of Lasso variants not very complicated
 - In other cases, tuning might require expert knowledge for complex methods

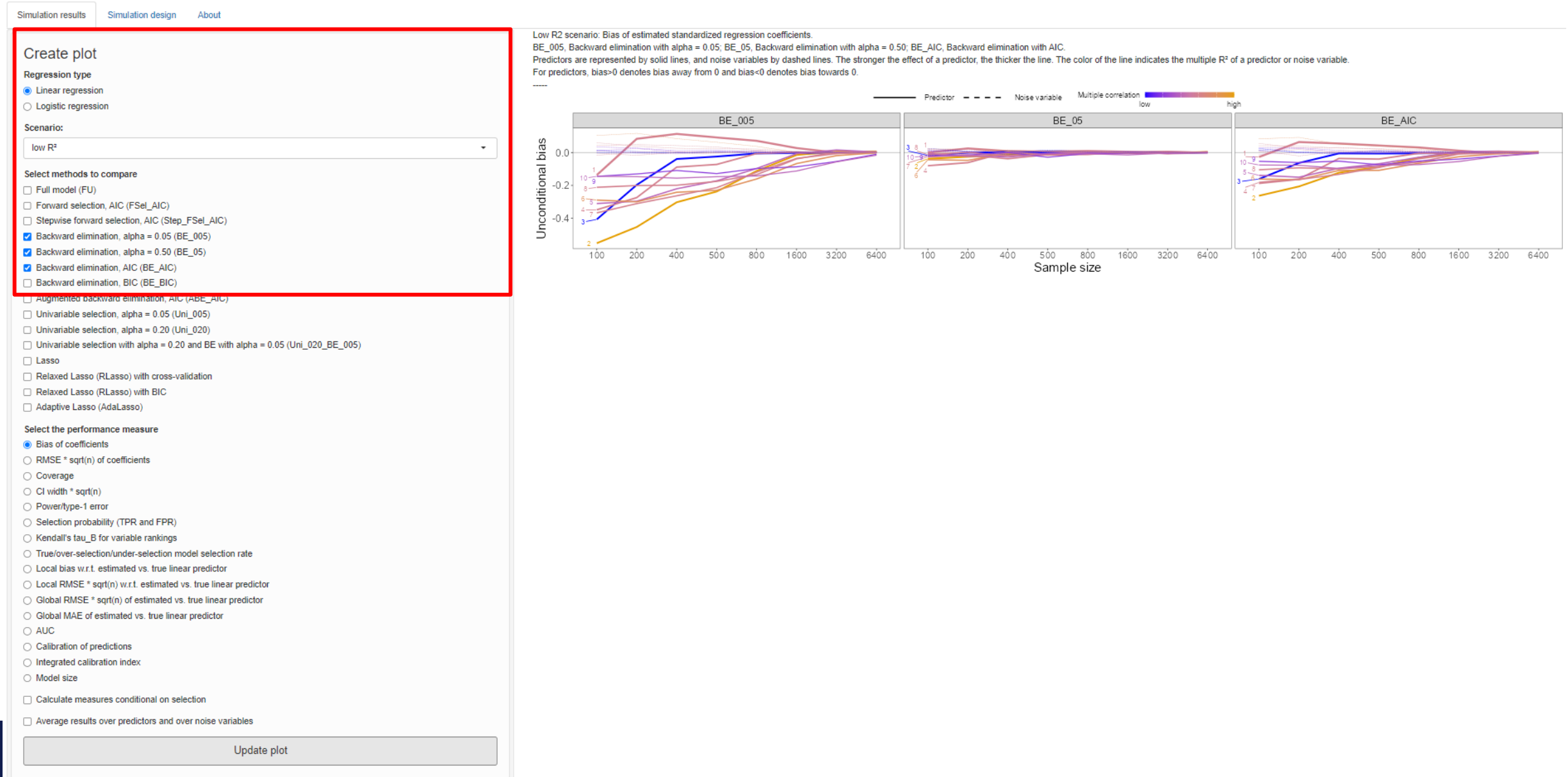
Reporting the results of a neutral comparison study

- The results of neutral comparison studies can be very extensive
 - Data simulation: many data-generating mechanisms/simulation settings
 - Many methods
 - Many performance measures
 - ...
- Challenge: how to report all these results?
 - Multiple papers?
 - Only report a subset of results? But then issue with selective reporting?
- We developed a Shiny app to make all results available



<https://openart.ai>

Interactive reporting of results in a Shiny app



Interactive reporting of results in a Shiny app

Create plot

Regression type

☒ Linear regression

☐ Logistic regression

Scenario:

low R^2 ▼

Select methods to compare

☐ Full model (FU)

☐ Forward selection, AIC (FSel_AIC)

☐ Stepwise forward selection, AIC (Step_FSel_AIC)

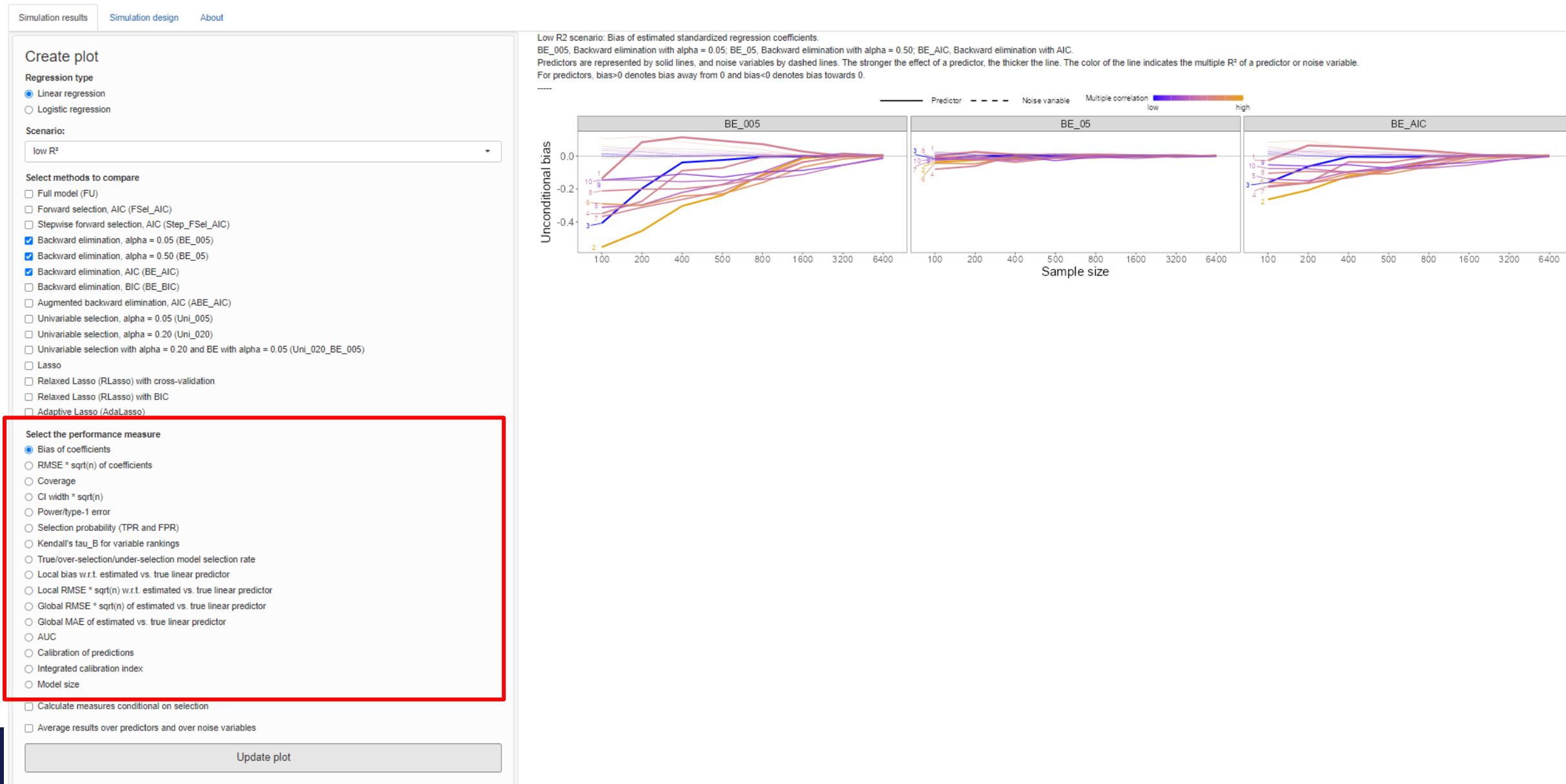
☒ Backward elimination, alpha = 0.05 (BE_005)

☒ Backward elimination, alpha = 0.50 (BE_05)

☒ Backward elimination, AIC (BE_AIC)

☐ Backward elimination, BIC (BE_BIC)

Interactive reporting of results in a Shiny app

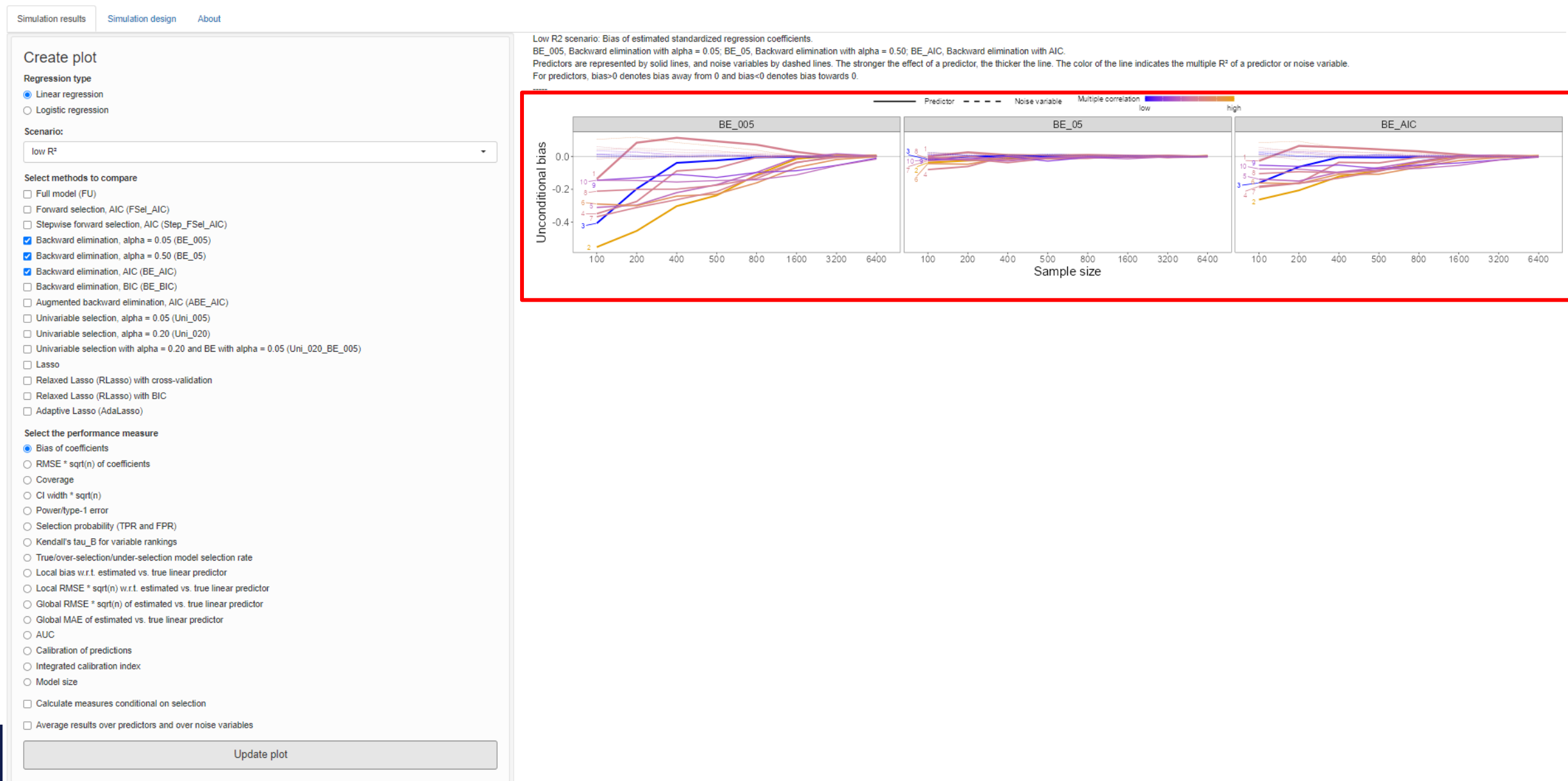


Interactive reporting of results in a Shiny app

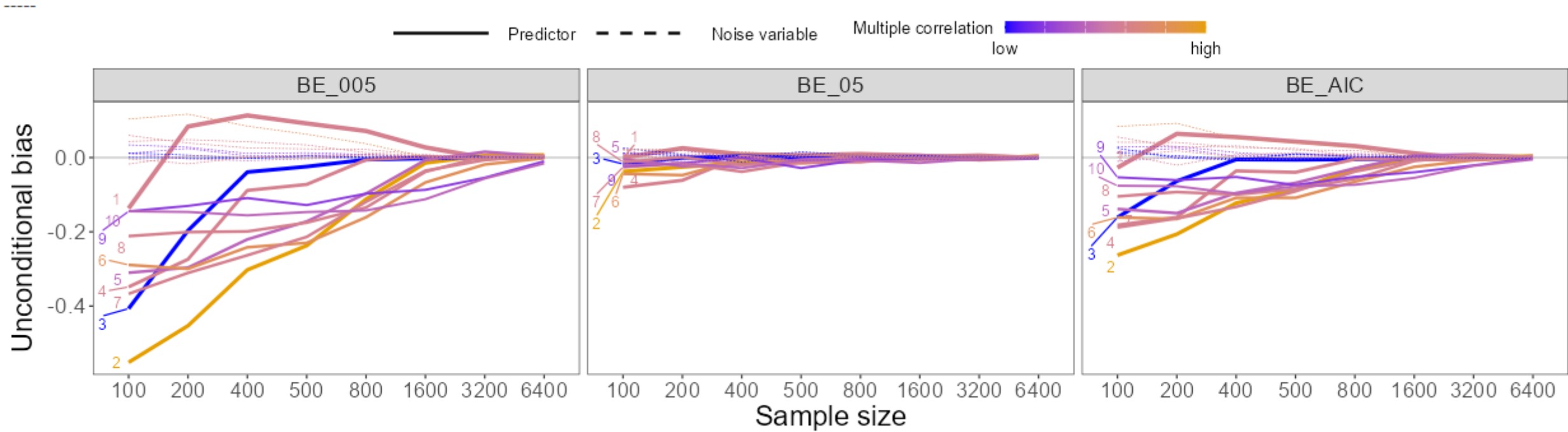
Select the performance measure

- ☒ Bias of coefficients
- ☐ RMSE * sqrt(n) of coefficients
- ☐ Coverage
- ☐ CI width * sqrt(n)
- ☐ Power/type-1 error
- ☐ Selection probability (TPR and FPR)
- ☐ Kendall's tau_B for variable rankings
- ☐ True/over-selection/under-selection model selection rate
- ☐ Local bias w.r.t. estimated vs. true linear predictor
- ☐ Local RMSE * sqrt(n) w.r.t. estimated vs. true linear predictor
- ☐ Global RMSE * sqrt(n) of estimated vs. true linear predictor
- ☐ Global MAE of estimated vs. true linear predictor
- ☐ AUC
- ☐ Calibration of predictions
- ☐ Integrated calibration index
- ☐ Model size

Interactive reporting of results in a Shiny app



Interactive reporting of results in a Shiny app



Reporting the results of a neutral comparison study

- How to draw meaningful conclusions from the results?
- Typically, we are not that interested in the performance of a method in very specific, somewhat “artificial” simulation settings...
- ... but we would like to *understand* the behavior of the method!
- Analogy: “understanding a method = learning how to drive a car”
- “Executive summary”?



Wikimedia Commons

Making data and code openly available

- Data and code for a comparison study should be openly available (upload e.g., on Zenodo, Github,...)
- Preparing the files can take some time...
- ... but is vital for transparency and reproducibility!
- This also allows to extend the study later on, e.g., adding new methods
- We will make our code available in a Github repository
- But: some knowledge about the code required to add further methods

Neutral comparison studies are not perfect, but vital

- To sum up, some challenges exist in neutral comparison studies
- No neutral comparison study will be “perfect” and provide a “final answer”
- Moreover, results of such studies rely on the specific choices for the study design
 - E.g., in our study: correlation structure of simulated data
 - Would results change when choosing different design options?

But:

- This does not mean that neutral comparison studies are useless
- Should serve as motivation to conduct *more* neutral comparison studies, and to design them carefully!

References

- Heinze, G., Boulesteix, A. L., Kammer, M., Morris, T. P., White, I. R., for the Simulation Panel of the STRATOS Initiative (2024). Phases of methodological research in biostatistics—building the evidence base for new methods. *Biometrical Journal*, 66(1), 2200222.
- Ullmann, T., Heinze, G., Hafermann, L., Schilhart-Wallisch, C., Dunkler, D., for TG2 of the STRATOS initiative (2024). Evaluating variable selection methods for multivariable regression models: A simulation study protocol. *Plos One*, 19(8), e0308543.