Multiplicity of design and analysis options in benchmark studies based on real data: An illustration

Christina Sauer (née Nießl), Moritz Herrmann[2], Chiara Wiedemann[1], Giuseppe Casalicchio[2], Anne-Laure Boulesteix[1]

Workshop "Open Replicable Research", Munich, October 5, 2023

[1]Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, LMU Munich

[2]Department of Statistics, LMU Munich)

Data sharing in methodological statistical research

- Different types of research
  - Applied research (e.g. clinical research): Use statistical methods to address a research question

Data sharing in methodological statistical research

- Different types of research
  - Applied research (e.g. clinical research): Use statistical methods to address a research question
  - Methodological statistical research: Development and evaluation of statistical methods

Data sharing in methodological statistical research

- Different types of research
  - Applied research (e.g. clinical research): Use statistical methods to address a research question
  - Methodological statistical research: Development and evaluation of statistical methods
- Data sharing in methodological statistical research

Data sharing in methodological statistical research

- Different types of research
  - Applied research (e.g. clinical research): Use statistical methods to address a research question
  - Methodological statistical research: Development and evaluation of statistical methods
- Data sharing in methodological statistical research
  - Sharing of data sets used to develop and evaluate statistical methods: Often not as problematic as for applied research

Data sharing in methodological statistical research

- Different types of research
  - Applied research (e.g. clinical research): Use statistical methods to address a research question
  - Methodological statistical research: Development and evaluation of statistical methods

- Data sharing in methodological statistical research
  - Sharing of data sets used to develop and evaluate statistical methods: Often not as problematic as for applied research
  - Sharing/reporting of other types of "data" (e.g., code, how the multiplicity of design/analysis options in benchmark studies was addressed, variability of results): Still not common practice

Data sharing in methodological statistical research

- Different types of research
  - Applied research (e.g. clinical research): Use statistical methods to address a research question
  - Methodological statistical research: Development and evaluation of statistical methods

- Data sharing in methodological statistical research
  - Sharing of data sets used to develop and evaluate statistical methods: Often not as problematic as for applied research
  - Sharing/reporting of other types of "data" (e.g., code, **how the multiplicity of design/analysis options in benchmark studies was addressed, variability of results**): Still not common practice

Illustrating the multiplicity of design and analysis options in benchmark studies

Example benchmark study: Survival prediction using multi-omics data

- Benchmark studies comparing the performance of different statistical methods based on real/simulated data are an essential part of methodological statistical research

Example benchmark study: Survival prediction using multi-omics data

- Benchmark studies comparing the performance of different statistical methods based on real/simulated data are an essential part of methodological statistical research

- Example study: Herrmann et al. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3), bbaa167.
  $\rightarrow$ Comparison of 13 survival prediction methods (incl. Lasso, CoxBoost, blockForest, Cox regression)

Example benchmark study: Survival prediction using multi-omics data

- Benchmark studies comparing the performance of different statistical methods based on real/simulated data are an essential part of methodological statistical research

- Example study: Herrmann et al. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3), bbaa167.
  $\rightarrow$ Comparison of 13 survival prediction methods (incl. Lasso, CoxBoost, blockForest, Cox regression)

Choice	Herrmann et al. (2021)	Alternative options
Data sets	18 real data sets with 5 multi-omics groups, $n \geq 100$, $\geq 5\%$ eff. cases	
Primary performance measure	Integrated Brier score	
Missing performance values	Ad-hoc 20%-threshold rule	
Aggregation across data sets	Mean	

Example benchmark study: Survival prediction using multi-omics data

- Benchmark studies comparing the performance of different statistical methods based on real/simulated data are an essential part of methodological statistical research

- Example study: Herrmann et al. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3), bbaa167.
  $\rightarrow$ Comparison of 13 survival prediction methods (incl. Lasso, CoxBoost, blockForest, Cox regression)

Choice	Herrmann et al. (2021)	Alternative options
Data sets	18 real data sets with 5 multi-omics groups, $n \geq 100$, $\geq 5\%$ eff. cases	$n, n_e, p, clin <$ or $\geq$ than median of orig. 18 data sets
Primary performance measure	Integrated Brier score	
Missing performance values	Ad-hoc 20%-threshold rule	
Aggregation across data sets	Mean	

Example benchmark study: Survival prediction using multi-omics data

- Benchmark studies comparing the performance of different statistical methods based on real/simulated data are an essential part of methodological statistical research

- Example study: Herrmann et al. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3), bbaa167.
  $\rightarrow$ Comparison of 13 survival prediction methods (incl. Lasso, CoxBoost, blockForest, Cox regression)

Choice	Herrmann et al. (2021)	Alternative options
Data sets	18 real data sets with 5 multi-omics groups, $n \geq 100$, $\geq 5\%$ eff. cases	$n, n_e, p, clin <$ or $\geq$ than median of orig. 18 data sets
Primary performance measure	Integrated Brier score	Uno's C-index
Missing performance values	Ad-hoc 20%-threshold rule	
Aggregation across data sets	Mean	

Example benchmark study: Survival prediction using multi-omics data

- Benchmark studies comparing the performance of different statistical methods based on real/simulated data are an essential part of methodological statistical research

- Example study: Herrmann et al. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3), bbaa167.
  $\rightarrow$ Comparison of 13 survival prediction methods (incl. Lasso, CoxBoost, blockForest, Cox regression)

Choice	Herrmann et al. (2021)	Alternative options
Data sets	18 real data sets with 5 multi-omics groups, $n \geq 100$, $\geq 5\%$ eff. cases	$n, n_e, p, clin <$ or $\geq$ than median of orig. 18 data sets
Primary performance measure	Integrated Brier score	Uno's C-index
Missing performance values	Ad-hoc 20%-threshold rule	Weighted, random, mean
Aggregation across data sets	Mean	

Example benchmark study: Survival prediction using multi-omics data

- Benchmark studies comparing the performance of different statistical methods based on real/simulated data are an essential part of methodological statistical research

- Example study: Herrmann et al. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3), bbaa167.
  $\rightarrow$ Comparison of 13 survival prediction methods (incl. Lasso, CoxBoost, blockForest, Cox regression)

Choice	Herrmann et al. (2021)	Alternative options
Data sets	18 real data sets with 5 multi-omics groups, $n \geq 100$, $\geq 5\%$ eff. cases	$n, n_e, p, clin <$ or $\geq$ than median of orig. 18 data sets
Primary performance measure	Integrated Brier score	Uno's C-index
Missing performance values	Ad-hoc 20%-threshold rule	Weighted, random, mean
Aggregation across data sets	Mean	Median, rank, best0.05

Multiplicity of design and analysis options in benchmark studies

- Multiplicity of different options when designing and analysing a benchmark study (data sets, DGPs, evaluation criteria, etc.)

Multiplicity of design and analysis options in benchmark studies

- Multiplicity of different options when designing and analysing a benchmark study (data sets, DGPs, evaluation criteria, etc.)
- Possible consequences caused by the multiplicity of options
  - Researchers might be concerned about how their choices affect the results
  - Researchers might (subconsciously) modify the benchmark study until it yields a favourable/reasonable result $\rightarrow$ risk of optimistic bias

Exploiting the multiplicity of different design and analysis options

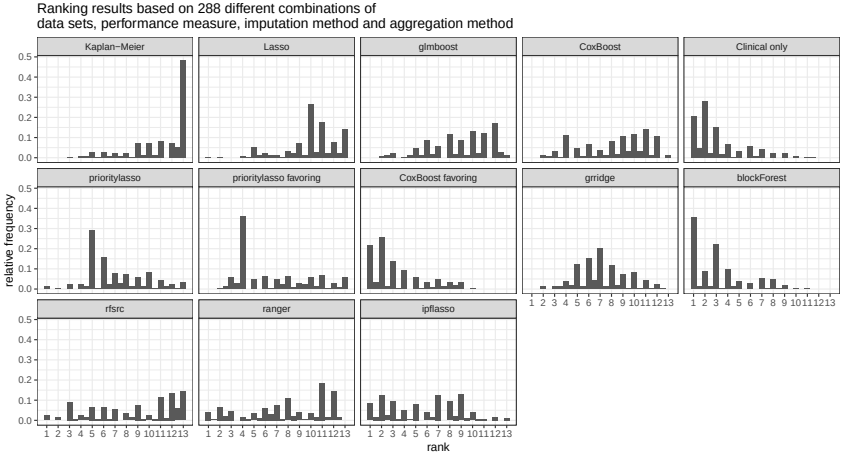
Choice	Herrmann et al. (2021)	Alternative options	No. of options
Data sets	18 real data sets with 5 multi-omics groups, $n \geq 100$, $\geq 5\%$ eff. cases	$n, n_e, p, clin <$ or $\geq$ than median of orig. 18 data sets	9
Primary performance measure	Integrated Brier score	Uno's C-index	2
Missing performance values	Ad-hoc 20%-threshold rule	Weighted, random, mean	4
Aggregation across data sets	Mean	Median, rank, best0.05	4
			$= 288$

Exploiting the multiplicity of different design and analysis options

Choice	Herrmann et al. (2021)	Alternative options	No. of options
Data sets	18 real data sets with 5 multi-omics groups, $n \geq 100$, $\geq 5\%$ eff. cases	$n, n_e, p, clin <$ or $\geq$ than median of orig. 18 data sets	9
Primary performance measure	Integrated Brier score	Uno's C-index	2
Missing performance values	Ad-hoc 20%-threshold rule	Weighted, random, mean	4
Aggregation across data sets	Mean	Median, rank, best0.05	4
			$= 288$

- In total: 288 combinations of design and analysis options
- Compare the resulting 288 rankings of the 13 survival prediction methods

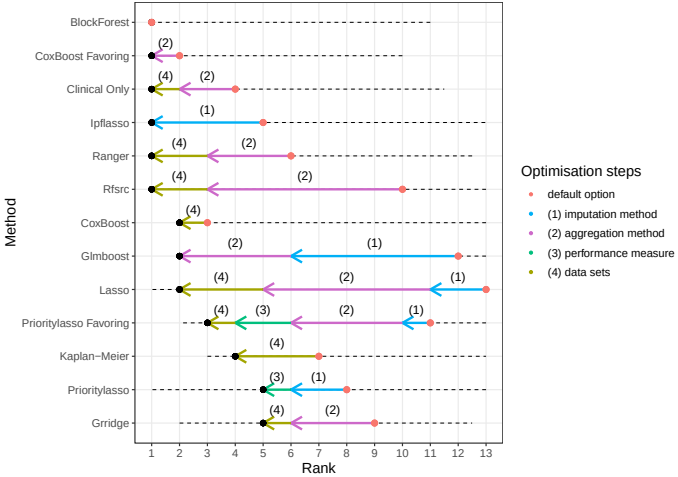
Overall variability of method rankings



→ Any method can achieve almost any rank

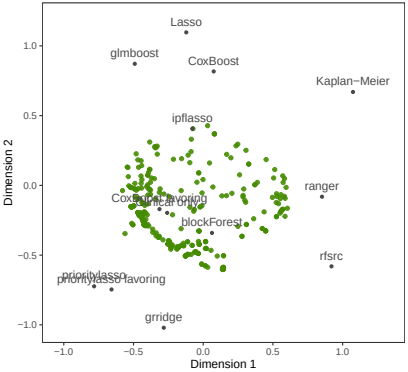
Stepwise optimisation

"No researcher would try all possibilities." → Stepwise optimisation



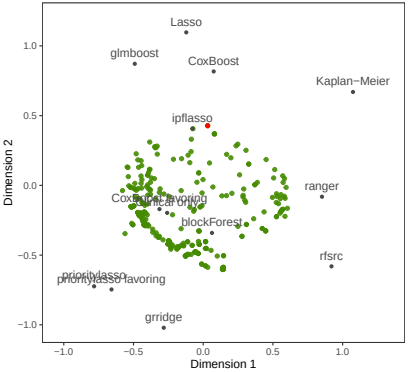
Assessing the impact of individual design and analysis choices

Impact of individual design and analysis choices



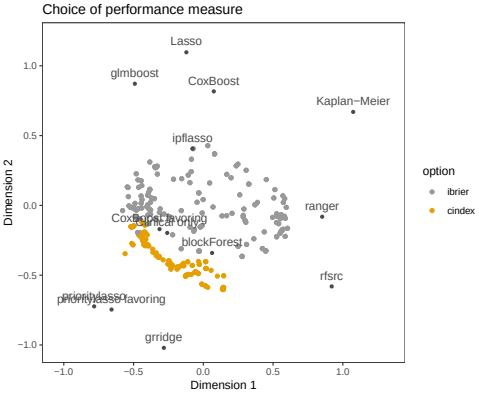
- Locate methods and option combinations such that distances between them correspond to ranks using multidimensional unfolding

Impact of individual design and analysis choices



- Locate methods and option combinations such that distances between them correspond to ranks using multidimensional unfolding

Impact of individual design and analysis choices

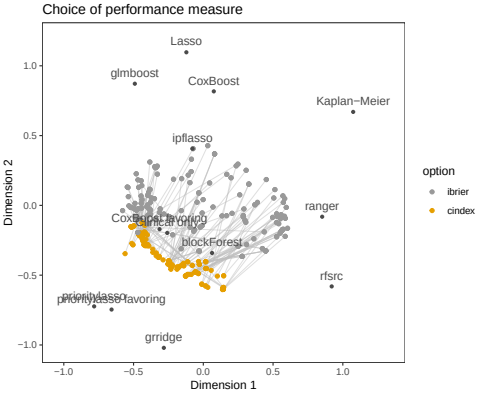


- Locate methods and option combinations such that distances between them correspond to ranks using multidimensional unfolding

For each design or analysis choice:

- Colour each point according to the option that was used in the respective combination

Impact of individual design and analysis choices



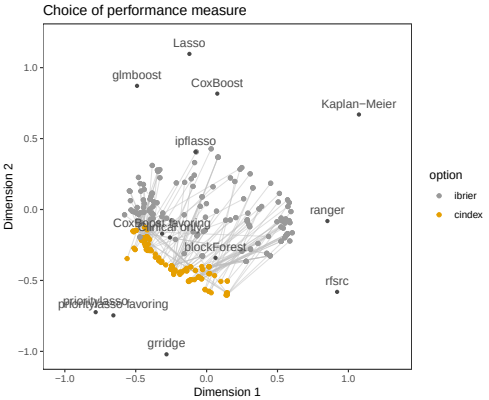
- Locate methods and option combinations such that distances between them correspond to ranks using multidimensional unfolding

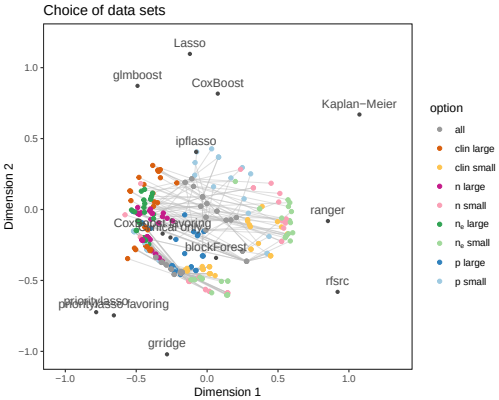
For each design or analysis choice:

- Colour each point according to the option that was used in the respective combination

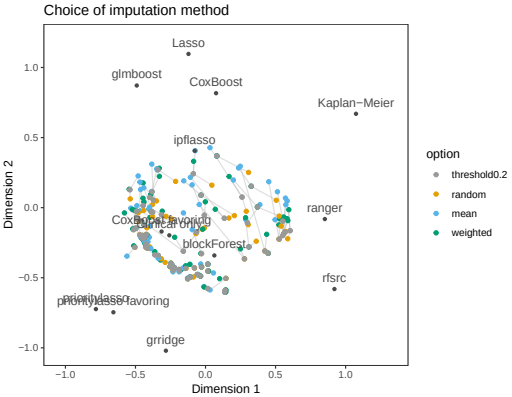
- Connect each point representing the option used in Herrmann et al. (2021) to points representing alternative options given that the other three choices remain the same

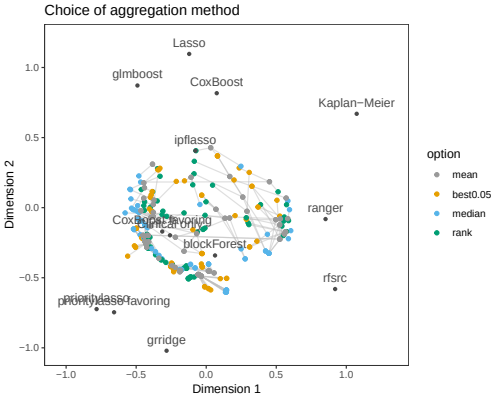
Impact of individual design and analysis choices





Impact of individual design and analysis choices





Conclusion

Conclusion

- Results of benchmark studies can be highly variable with respect to design and analysis choices $\rightarrow$ risk of introducing an optimistic bias

Conclusion

- Results of benchmark studies can be highly variable with respect to design and analysis choices → risk of introducing an optimistic bias
- Reporting of variability using the multidimensional unfolding approach → graphical assessment of results with respect to a large number of different combinations of design and analysis options
  - Intuitive overview of the variability of results
  - Identification of critical choices that substantially affect the results and should be investigated in more detail

References

C. Nießl, M. Herrmann, C. Wiedemann, G. Casalicchio, and A.-L. Boulesteix.
Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results.
WIREs Data Mining and Knowledge Discovery, 12:e1441, 2022.

M. Herrmann, P. Probst, R. Hornung, V. Jurinovic, and A.-L. Boulesteix.
Large-scale benchmark study of survival prediction methods using multi-omics data.
Briefings in Bioinformatics, 22, 2021.
bbaa167.

S. Hoffmann, F. Sch, R. Elsas, R. Wilson, U. Strasser, and A.-l. Boulesteix.
The multiplicity of analysis strategies jeopardizes replicability : lessons learned across disciplines.
Royal Society Open Science, 8:201925, 2021.

S. Pawel, L. Kook, and K. Reeve.
Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method.
Biometrical Journal, page 2200091.

Thank you!