Ongoing research towards state-of-the-art in variable and functional form selection for statistical models

Georg Heinze, Aris Perperoglou, Willi Sauerbrei

for TG2 of the STRATOS initiative

Sauerbrei *et al. Diagnostic and Prognostic Research* (2020) 4:3
https://doi.org/10.1186/s41512-020-00074-3

Diagnostic and Prognostic Research

COMMENTARY Open Access

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

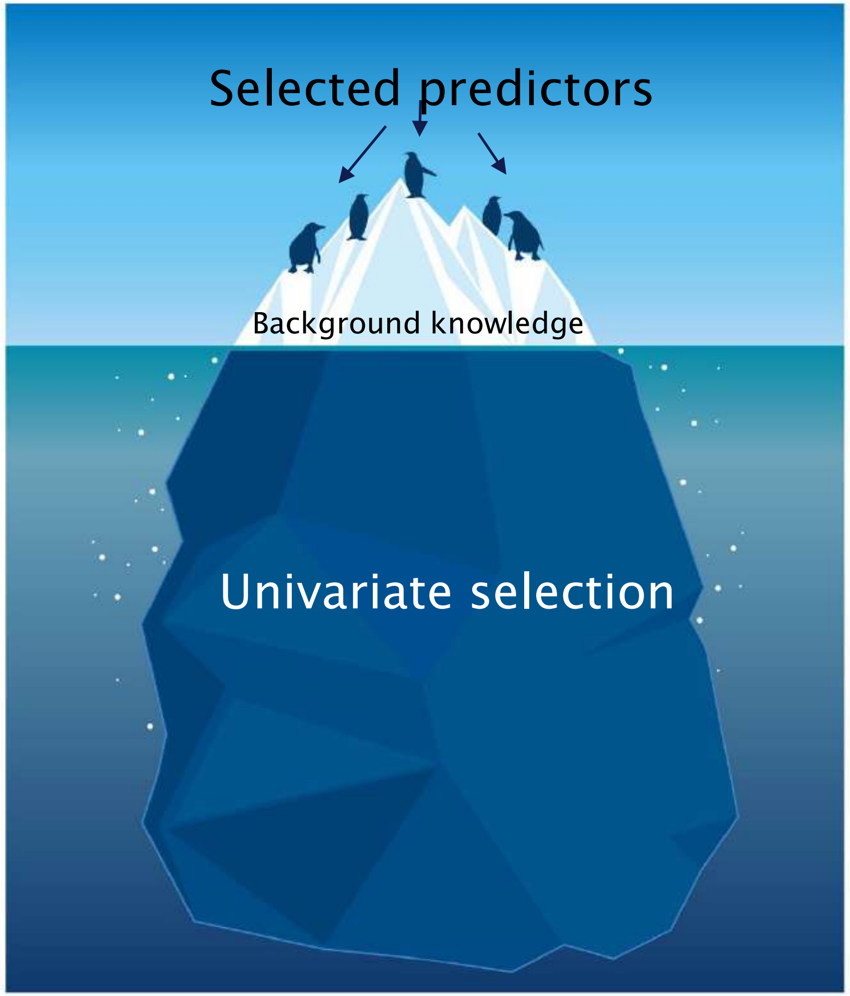
Willi Sauerbrei[1*], Aris Perperoglou[2], Matthias Schmid[3], Michal Abrahamowicz[4], Heiko Becher[5], Harald Binder[1], Daniela Dunkler[6], Frank E. Harrell Jr[7], Patrick Royston[8], Georg Heinze[6] and for TG2 of the STRATOS initiative

Slide from ISCB 2019

Towards recommendations – research required!

1. Investigation and comparison of the properties of **variable selection strategies**
2. **Comparison of spline procedures** in univariable and multivariable contexts
3. How to model one or more variables with a ‚**spike-at-zero**'?
4. Comparison of **multivariable procedures for model and function selection**
5. **Role of shrinkage** to correct for bias introduced by data-dependent modelling
6. Evaluation of new approaches for **post-selection inference**
7. Adaptation of procedures for **very large sample sizes** needed?

Variable selection: current practice



- Various reviews of model building strategies identified univariate selection still in wide use
- (and its actual, silent use may be even much more widespread)
- TG2 is conducting a review of model building strategies in COVID-19 prediction models

Variable selection – poor guidance?

RESEARCH ARTICLE

Review of guidance papers on regression modeling in statistical series of medical journals

Christine Wallisch[1,2]*, Paul Bach[1,3], Lorena Hafermann[1], Nadja Klein[3], Willi Sauerbrei[4], Ewout W. Steyerberg[5], Georg Heinze[2], Geraldine Rauch[1]*, on behalf of topic group 2 of the STRATOS initiative[¶]

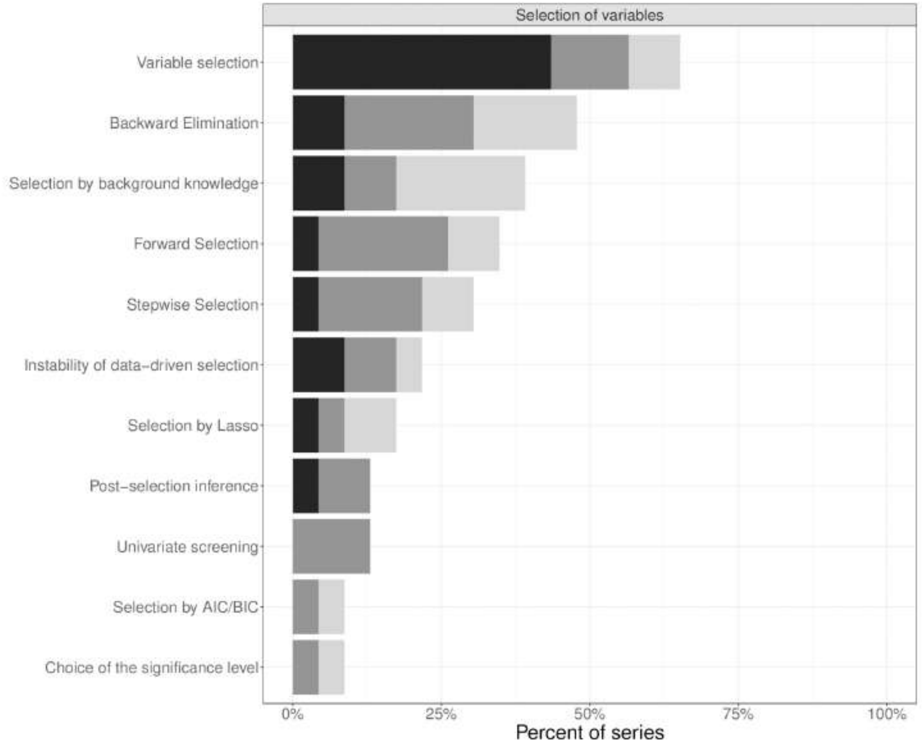


Fig 5. Extent of explanation of aspects of selection of variables in statistical series: One sentence only (light grey), more than one sentence to one paragraph (grey) and more than one paragraph (black).

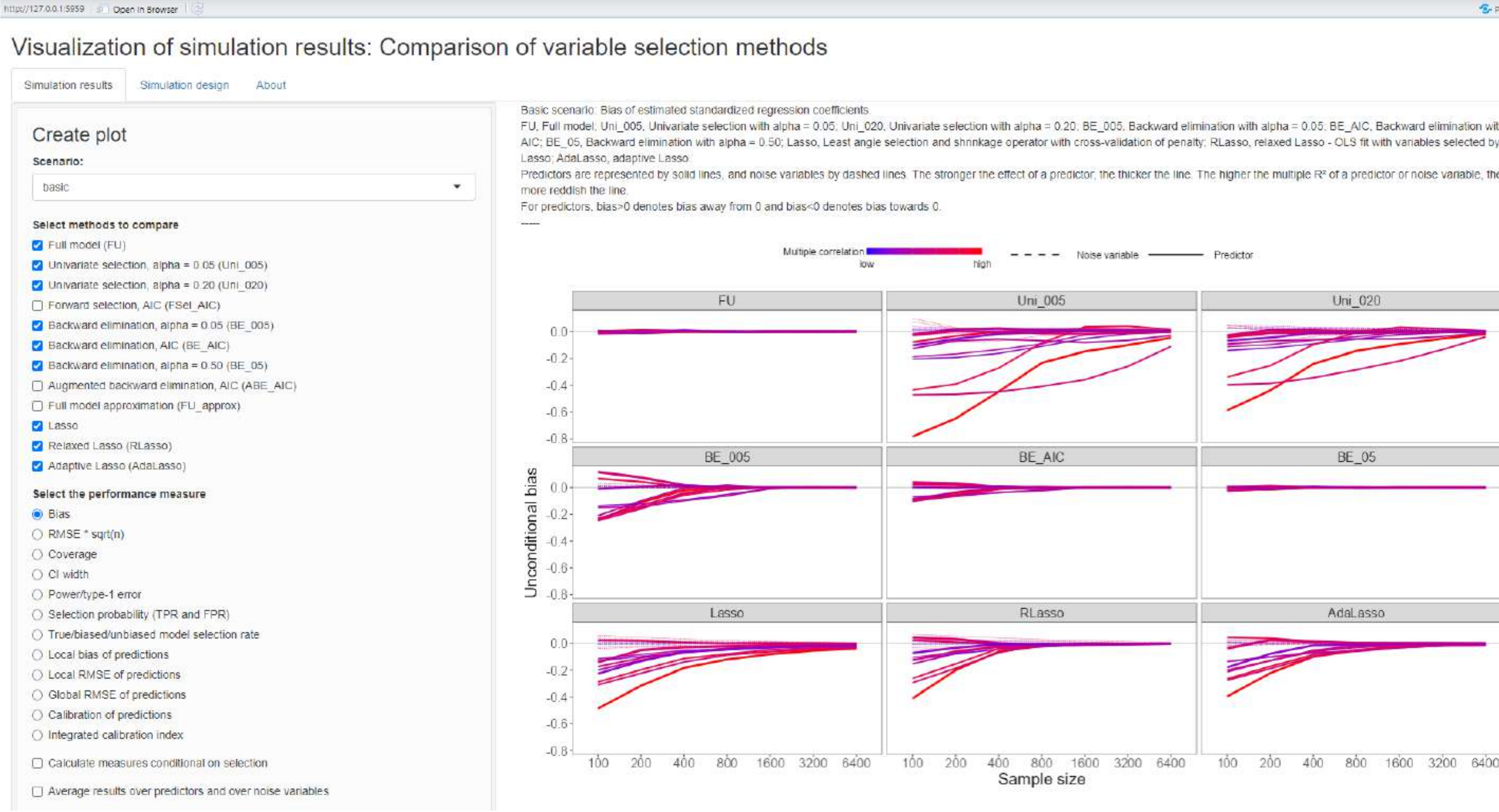
https://doi.org/10.1371/journal.pone.0262918.g005

Wallisch et al, 2022:

Selection of variables was mentioned in 15 series (65%) and described extensively in ten series (43%) (Fig 5). However, specific variable selection methods were rarely described in detail. *Backward elimination, selection based on background knowledge, forward selection*, and *stepwise selection* were the most frequently described selection methods in seven to eleven series (30–48%). *Univariate screening*, which is still popular in medical research, was only described in three series (13%) in up to one paragraph. Other aspects of variable selection were hardly ever mentioned. *Selection based on AIC/BIC*, relating to best subset selection or stepwise selection based on these information criteria, and the *choice of the significance level* were found in 2 series only (9%). Relative frequencies of aspects mentioned in articles are detailed in Figs 1–3 in S5 File.

Ongoing work

- Simulation studies:
  - Ullmann (Vienna)
  - Kipruto (Freiburg)
- Education:
  TG2 workshops
  at ROeS 2021,
  Maastricht 2023
- Lectures and workshops
  by Willi Sauerbrei,
  Frank Harrell,
  Georg Heinze
  & Daniela Dunkler
  and others



Key messages about variable selection

- Purpose of model?   Descriptive, explanatory, or predictive?
  - Large effects on coefficient estimates (simulation study by Th. Ullmann)
  - Smaller effects on predictions (see also Riley & Collins, BIomJ 2023: https://doi.org/10.1002/bimj.202200302)
- Relative performance of methods depends on sample size
- The true ‚model' is rarely identified
- Inference is wrong – where does it matter?
- *Accept model uncertainty as another source of variation!*

The role of background knowledge
Some STRATOS-triggered cooperations (project SAMBA)

- Good examples for background knowledge:
  Nottingham Prognostic Index (Breast Cancer)
  Framingham risk score (Cardiovascular Med.)

- But: Poorly conducted studies generate background „knowledge“ that is of little use
- *Critically appraise the source of the background knowledge!*

Hafermann et al. BMC Medical Research Methodology (2021) 21:196
https://doi.org/10.1186/s12874-021-01373-z
BMC Medical Research Methodology

RESEARCH Open Access

Statistical model building: Background “knowledge” based on inappropriate preselection causes misspecification

Check for updates

Lorena Hafermann[1*], Heiko Becher[2], Carolin Herrmann[1], Nadja Klein[3], Georg Heinze[4] and Geraldine Rauch[1]

Functional form selection: Spline procedures

REVIEW **Open Access**

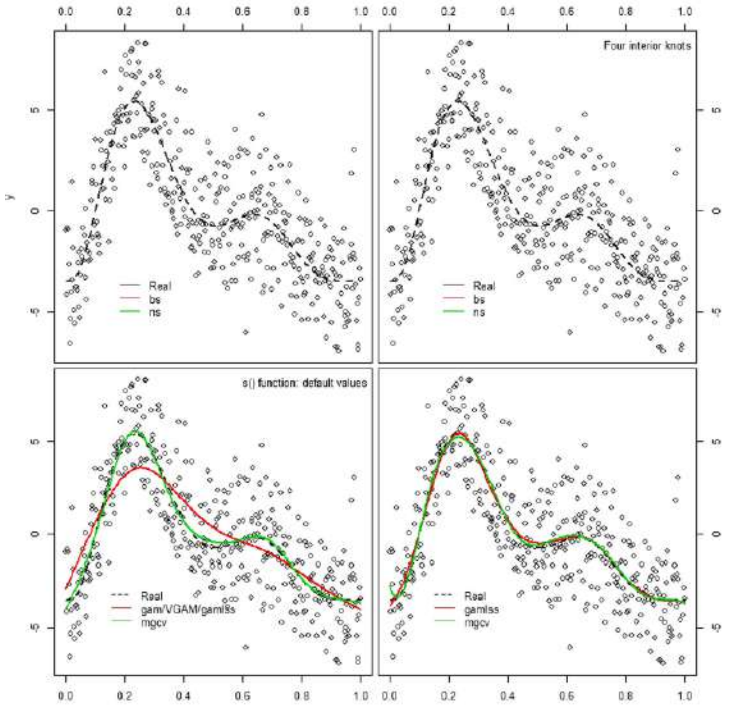
A review of spline function procedures in R

Check for updates

Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of TG2 of the STRATOS initiative

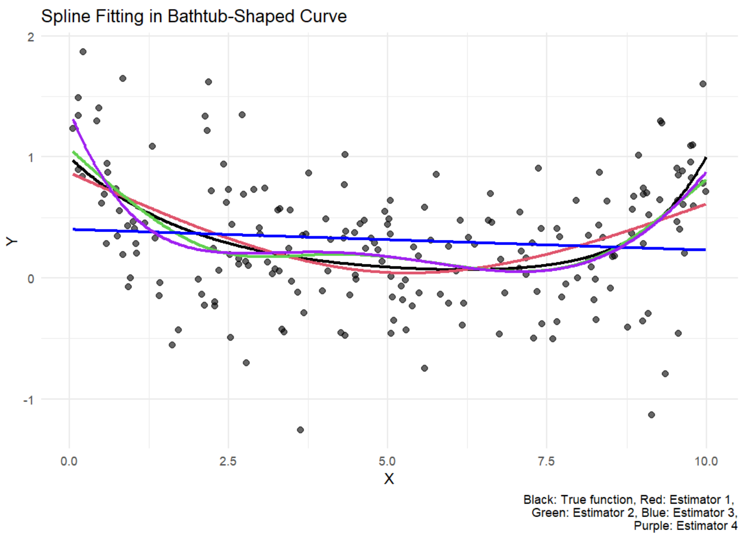
Perperoglou *et al. BMC Medical Research Methodology* (2019) 19:46
https://doi.org/10.1186/s12874-019-0666-3

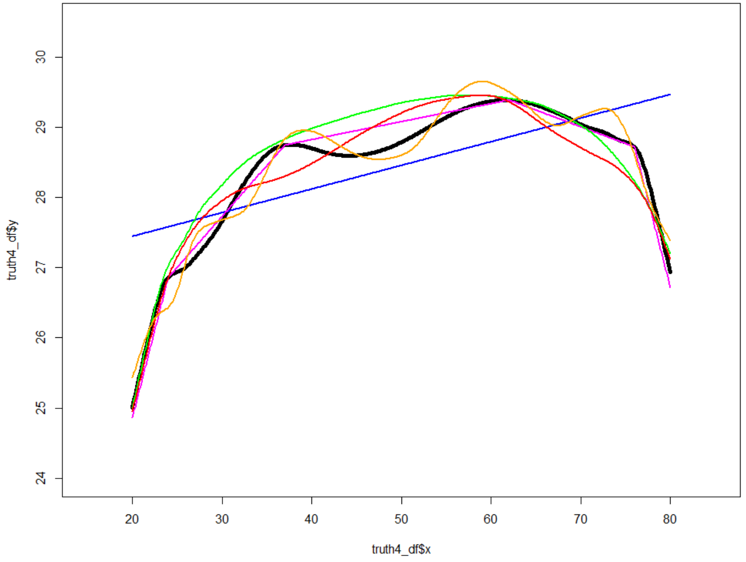
BMC Medical Research Methodology



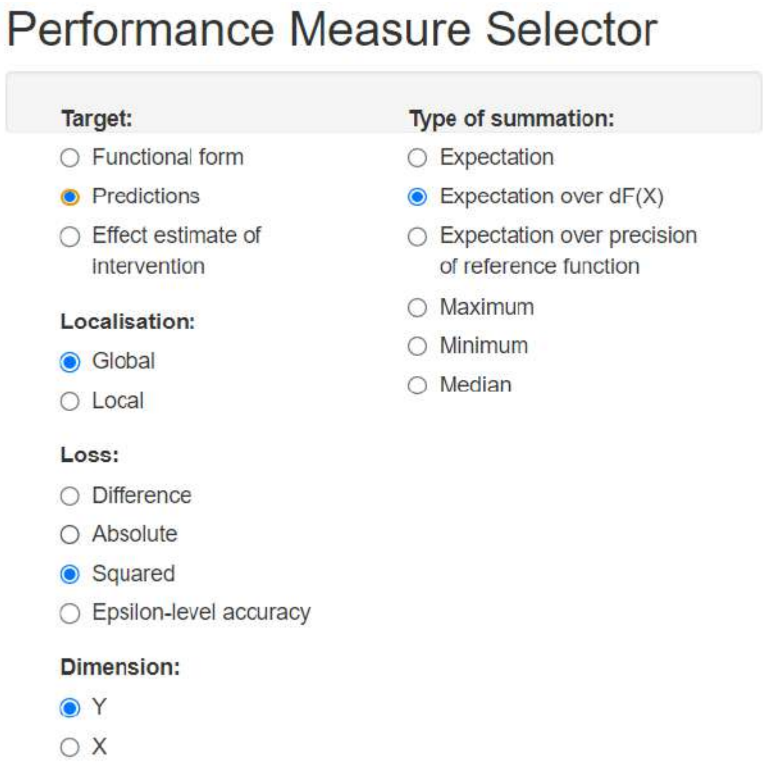
Spline procedures: current questions

- Which estimator is the ‚best‘?





True
Estimator 1
Estimator 2
Estimator 3
Estimator 4



→Many different performance measures can be ‚designed‘ by combining different aspects

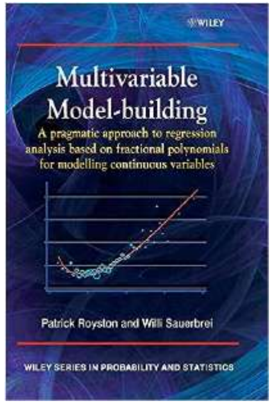
→Which of them are suitable?

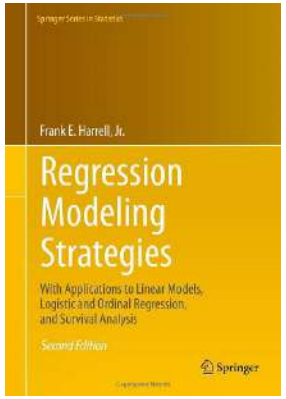
Currently: evaluation of performance measures
Next step: comparison of splines

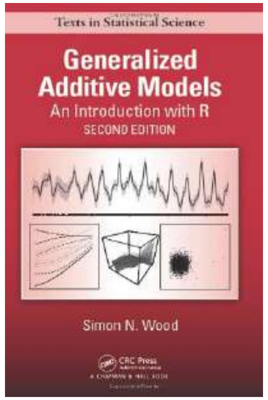
Combining variable and functional form selection

- Several‚philosophies‘:

	Multivariable fractional polynomials (mfp2)	Restricted cubic splines (rms)	Penalized/thin plate splines (mgcv)
Selection	Significance-based	No	Penalty-based
Smoothing	Global: x^{p_1}, x^{p_2}	Local: spline based	Local: spline based
Basis functions (4df)	2 per variable (FP2)	4 per variable	‚many‘ per variable



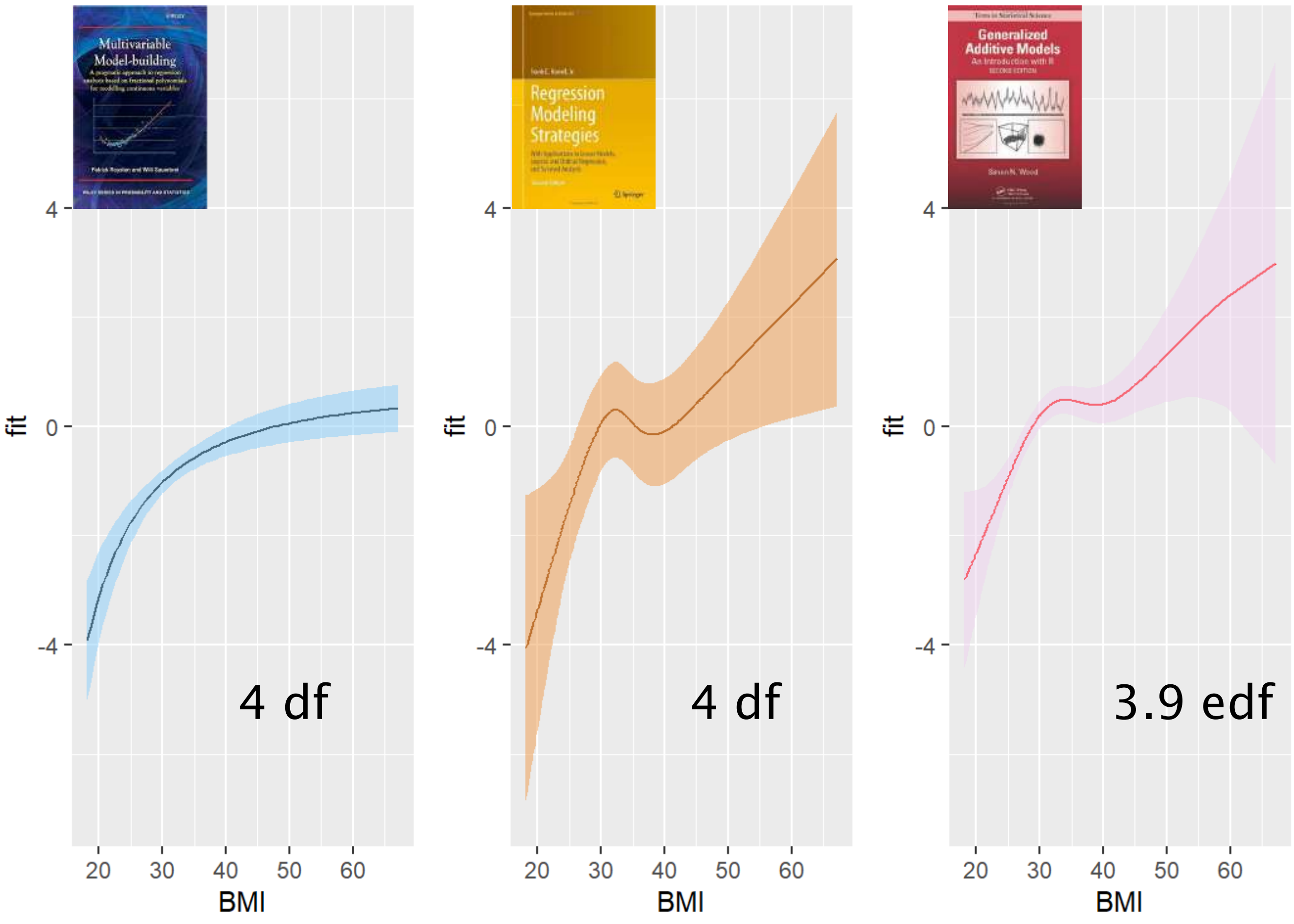




Comparison in Pima data set:

- Predicting diabetes onset (yes/no) in 768 members of Pima nation
- 8 cont. predictors

4 selected (6 df) | (8 included, 32 df) | 6 selected (11.2 edf)



- Partial linear predictors for BMI:

Role of shrinkage

STUDY PROTOCOL

Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low-dimensional data

Edwin Kipruto *, Willi Sauerbrei

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany

* Edwin.Kipruto@imbi.uni-freiburg.de

A. Variable selection methods

Method	Tuning parameters	Initial estimates
Lasso	10-fold CV, AIC & BIC	N/A
Garrote	10-fold CV, AIC & BIC	OLS, ridge and lasso
Alasso*	10-fold CV, AIC & BIC	OLS, ridge and lasso
Rlasso*	10-fold CV, AIC & BIC	N/A
Best subset	10-fold CV, AIC & BIC	N/A
BE*	10-fold CV, AIC & BIC	N/A

B. Post-estimation shrinkage methods:
(i) Global [10], (ii) parameterwise [9] and (iii) Breiman's method [5]
Estimation method: (i) leave-one-out CV and (ii) 10-fold CV

PlosOne 2022

Talk by Edwin Kipruto @CEN 2023

Evaluation of selective inference
STRATOS-triggered cooperation

- Phase II-III study
  (Neutral,
  rather new methods,
  limited range of scenarios)
- Software still in its infancy
- Sel.Inf.-CI were proposed for fixed-λ LASSO
- They empirically work also with tuned λ, but are sometimes conservative
- They don‘t work well with Adaptive Lasso

Kammer *et al.*
BMC Medical Research Methodology (2022) 22:206
https://doi.org/10.1186/s12874-022-01681-y
BMC Medical Research Methodology

RESEARCH Open Access

Evaluating methods for Lasso selective inference in biomedical research: a comparative simulation study

Michael Kammer[1,2], Daniela Dunkler[1], Stefan Michiels[3] and Georg Heinze[1*]

Adaptation for big data sets

- Big data sets (many observations) make any p-values ridiculously small
- How to separate relevant from irrelevant effects?
- Model size depends on purpose of the model
  - Should the model be communicable or a ‚black-box'?
  - Can the model be applied electronically (e.g. on EHRs)?
  - Model approximation/projection?
- Ongoing project in the context of MFP
  (Willi Sauerbrei, Patrick Royston, Aris Perperoglou)

STRATOS cooperations between TGs

- TG2-TG4: effects of measurement error on functional form estimation
  Exciting design of an adversarial simulation study with distributed roles
  data generator – data analysts – performance evaluator

- TG2-TG3: ‚Regression without regrets'
  - Initial data analysis before regression analysis
  Paper to be submitted soon; previous talks at ISCB 2020, IBC 2022

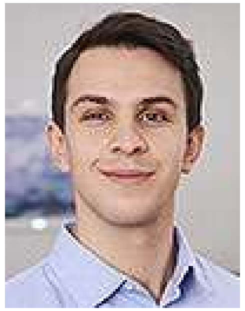
Simulation studies –
key instruments to compare approaches

Boulesteix et al, *Significance* 2020

There is a clear need for more neutral comparisons and replications of methodological statistical research, but how should such studies be performed? Surprisingly, the design of comparison studies of statistical methods has hardly been addressed

Pawel et al, *BiomJ* 2023:

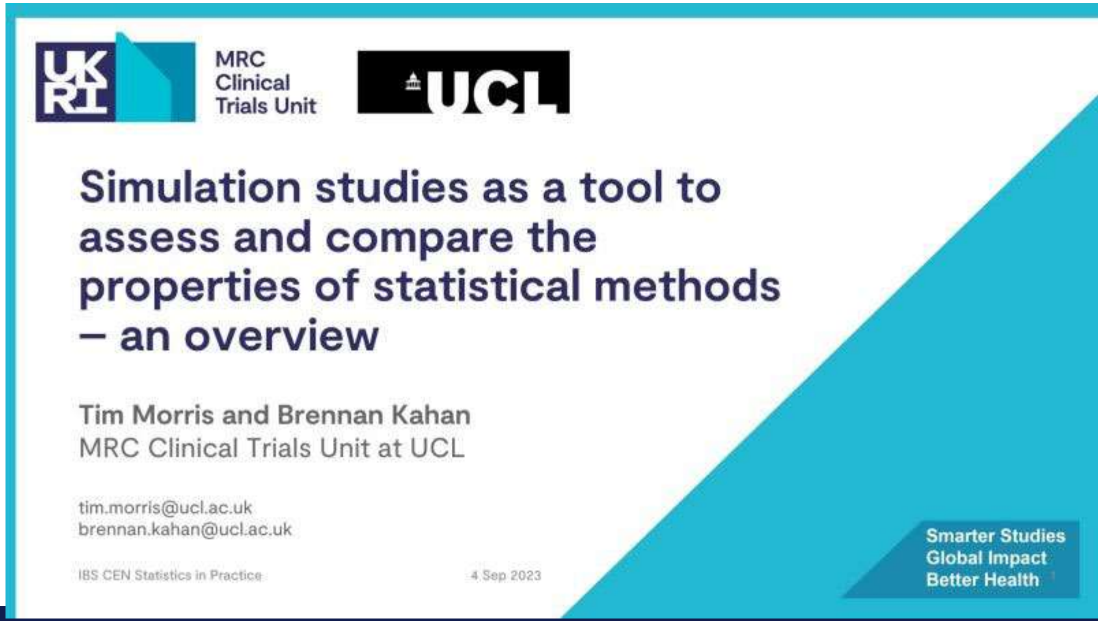
We **show how easy it is to make the method appear superior** over well-established competitor methods if no protocol is in place and various questionable research practices are employed



Heinze, Boulesteix, Kammer, Morris, White (STRATOS Simulation Panel), *BiomJ* 2023:
Biostatistical methods are typically developed and evaluated in **four phases**;
only after **Phase IV** we know when a method **is or is not the preferred method**
Each phase needs different type of simulation study

- Special issue in Biometrical Journal devoted to 'Neutral Comparison Studies' (to be released very soon)

Statistics in Practice, CEN2023:



Conclusion

- In many areas, we have enough methods
  but we don‘t know yet which one to recommend/discourage from
- We need evidence generated in neutral comparison studies of Phases III and IV:
  - Simulation studies
  - Comparative studies based on real data sets with real scientific questions
- Which methods are ready to use:
  - by level-1 data analysts?
  - by level-2 statisticians?

References

- Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M (STRATOS). A review of spline function procedures in R. BMC Med Res Meth 2019
- Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, Dunkler D, Harrell FE, Royston P, Heinze G (STRATOS). State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. DiagnPrognRes 2020
- Wallisch C, Bach P, Hafermann L, Klein N, Sauerbrei W, Steyerberg EW, Heinze G, Rauch G (STRATOS). Review of guidance papers on regression modeling in statistical series of medical journals. Plos One 2022
- Heinze G, Boulesteix A-L, Kammer M, Morris TP, White IR (STRATOS). Phases of methodological research in biostatistics – Building the evidence base for new methods. Biometrical Journal 2023
- Boulesteix A-L, Hofmann S, Charlton A, Seibold H. A replication crisis in methodological research? Significance 2020
- Kammer M, Dunkler D, Michiels S, Heinze G. Evaluating methods for Lasso selective inference in biomedical research: a comparative simulation study. BMC Med Res Meth 2022
- Hafermann L, Becher H, Herrmann C, Klein N, Heinze G, Rauch G. Statistical model building: Background “knowledge” based on inappropriate preselection causes misspecification. BMC Med Res Meth 2021
- Harrell FE. Regression Modeling Strategies, 2nd edition. Springer, 2015
- Kipruto E, Sauerbrei W. Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression – protocol of a simulation study. PlosOne 2022
- Pawel S, Kook L, Reeve K. Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. Biometrical Journal 2023
- Riley R, Collins G. Stability of clinical prediction models developed using statistical or machine learning methods. Biometrical Journal 2023
- Royston P, Sauerbrei W. Multivariable model-building. Wiley 2008
- Wood SN. Generalized Additive Models, 2nd Edition. Taylor and Francis, 2017