Towards evidence-based guidance on variable selection methods for multivariable regression models

Georg Heinze

Institute of Clinical Biometrics

Acknowledgment

Joint work
with
Theresa Ullmann, Daniela Dunkler

and
Topic Group 2
‚Selection of variables and functional forms for multivariable regression models‘
of the
STRATOS initiative

Topic Group 2:
Chairs: Georg Heinze, Aris Perperoglou, Willi Sauerbrei

Members: Michal Abrahamowicz, Harald Binder, Daniela Dunkler, Frank Harrell, Marc Henrion, Michael Kammer, Mariana Nold, Matthias Schmid, Theresa Ullmann

Agenda

- The STRATOS initiative: TG2 and others
- The role of data-driven variable selection in empirical research
- Principled data screening to avoid random data explorations
- Variable selection in practice – a systematic review
- Performance of variable selection methods – a simulation study
  - Expectations about different methods
  - Where do those methods meet our expectations?
- Including functional form selection in variable selection – current activities

The STRATOS initiative and TG2

STRATOS INITIATIVE

STRATOS Objectives

- Provide accessible and evidence-based guidance for key topics in the design and analysis of observational studies
- Guidance is intended for applied statisticians and other data analysts with varying levels of statistical education, experience and interests

STRATOS is led by an **Executive Committee** and a **Steering Group**. It has the following topic groups and cross-cutting panels:

	Topic Groups (TGs)
1	Missing data
2	Selection of variables and functional forms in multivariable analysis
3	Initial data analysis
4	Measurement error and misclassification
5	Study design
6	Evaluating diagnostic tests and prediction models
7	Causal Inference
8	Survival analysis
9	High-dimensional data

	Panels
MP	Membership
PP	Publications
GP	Glossary
WP	Website
RP	Literature Review
BP	Bibliography
SP	Simulation Studies
DP	Data Sets
TP	Knowledge Translation
CP	Contact Organizations
VP	Visualisation

STRATOS INITIATIVE

Publications

Topic Group 2 (TG2) Publications

These publications were written on behalf of STRATOS-TG2. They center around the topic group's core topic, the selection of variables and functional forms in multivariable models.

A Systematic Categorization of Performance Measures for Estimated Non-Linear Associations Between an Outcome and Continuous Predictors

Ullmann T, Heinze G, Abrahamowicz M, Perperoglou A, Sauerbrei W, Schmid M, Dunkler D, for TG2 of the STRATOS initiative, 2025. A Systematic Categorization of Performance Measures for Estimated Non-Linear Associations Between an Outcome and Continuous Predictors. WIREs Computational Statistics 17(3), e70042. https://doi.org/10.1002/wics.70042

▶ Abstract

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

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, e0308543. https://doi.org/10.1371/journal.pone.0308543

▶ Abstract

Regression without regrets –initial data analysis is a prerequisite for multivariable regression

Heinze, G., Baillie, M., Lusa, L., Sauerbrei, W., Schmidt, C.O., Harrell, F.E., Huebner, M., on behalf of TG2 and TG3 of the STRATOS initiative, 2024. Regression without regrets –initial data analysis is a prerequisite for multivariable regression. BMC Med Res Methodol 24, 178. https://doi.org/10.1186/s12874-024-02294-3

▶ Abstract

On this page

What are the aims of modeling?

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Why statistical
explanatory modeling
differs from
predictive modeling

Shmueli (2010), *Statistical Science*

Galit Shmueli discusses the distinction between explaining and predicting (Preview)

- … or to describe?

1.3 Descriptive Modeling

Although not the focus of this article, a third type of modeling, which is the most commonly used and developed by statisticians, is descriptive modeling. This type of modeling is aimed at summarizing or representing the data structure in a compact manner. Unlike explanatory modeling, in descriptive modeling the reliance on an underlying causal theory is absent or incorporated in a less formal way. Also, the focus is at the measurable level rather than at the construct level. Unlike predictive modeling, descriptive modeling is not

Statistics in Medicine

WILEY
Statistics
in Medicine

COMMENTARY

To Explain, to Predict, or to Describe: Figuring out the Study Goal [Commentary on "On the Uses and Abuses of Regression Models" by Carlin and Moreno-Betancur]

Galit Shmueli

Statistics in Medicine

WILEY

Statistics
in Medicine 2025

FEATURED ARTICLE OPEN ACCESS

On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching

John B. Carlin[1,2,3] | Margarita Moreno-Betancur[1,2]

- "It can be argued that each research data analysis task has either
  - (i) a **descriptive purpose**—characterizing the distribution of a feature or health outcome in a population,
  - (ii) a **predictive purpose**—producing a model or algorithm for predicting future values of an outcome given individual characteristics, or
  - (iii) a **causal purpose**—investigating the extent to which health outcomes in a population would be different if a particular intervention were made."

To Describe, to Predict or to Explain?

- **Descriptive models**
  - In order to describe a feature/outcome in a population, one uses models to connect expected values of 'independent' variables
  - Capture the data structure parsimoniously
    - Which variables is the outcome associated with, and how? ← VARIABLE SELECTION
    - Smoothing/functional forms: efficient estimation of expected values ← FUNCTIONAL FORM ESTIMATION
- **Prediction models**
  - Interest in accurate predictions for future application
- **Explanatory (causal) models**
  - Interest in effect of an intervention on an individual's outcome

Often several modeling goals simultaneously:

- Transparent prediction models (D + P)
- Counterfactual prediction models (E + P)

(Shmueli, 2010)

Prespecification and selection of variables and functional forms depends on the modeling aim

Modeling aim	Prespecification of variables based on expertise	What data-driven selection may add	What functional form estimation may add
Description	What are the variables I want to consider?	Remove irrelevant predictors	Remove local biases resulting from incorrect specification
Prediction	Availability, chronology, costs, assumed associations with Y	Reduce model size, reduce prediction error	Optimize calibration, reduce prediction error
Explanation	Confounders that need to be adjusted for (DAGs)	Reduce MSE of effect estimate	Remove residual confounding resulting from too simplistic assumptions

Heinze *et al.*
BMC Medical Research Methodology (2024) 24:178
https://doi.org/10.1186/s12874-024-02294-3

BMC Medical Research Methodology

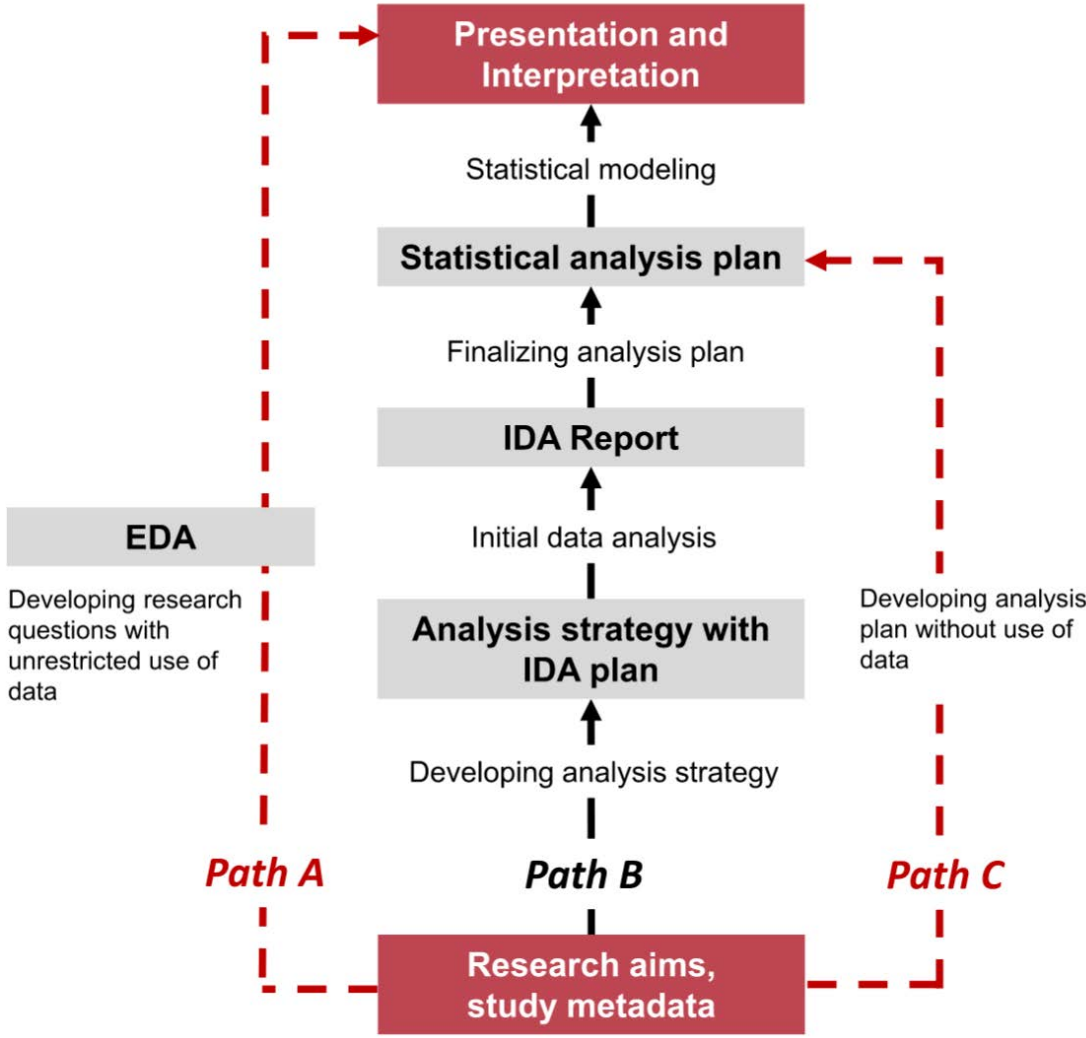
RESEARCH Open Access

Regression without regrets –initial data analysis is a prerequisite for multivariable regression

Georg Heinze[1*], Mark Baillie[2], Lara Lusa[3,4], Willi Sauerbrei[5], Carsten Oliver Schmidt[6], Frank E. Harrell[7], Marianne Huebner[8] on behalf of TG2 and TG3 of the STRATOS initiative

Prespecification vs. data driven approach vs. IDA

- Path A: EDA (exploratory data analysis) leaves door open for HARKing and other unaccounted sources of bias and variation
- Path C: full prespecification often unrealistic
- Path B:
  Prespecify analysis strategy in SAP v1
  - Specify „Initial Data Analysis (IDA)“ plan:
    - IDA≠EDA: no outcome-predictor snooping!
  - Perform IDA: provide foundation for
    - Later interpretation of results
    - Later presentation of results
    - Amendments of SAP
  - Finalize SAP v2
  - Perform analysis according to SAP v2



Heinze et al, Regression without regrets‘, BMC Med Res Meth 2024

Domains of IDA (data screening)

- Prerequisites – metadata
  - IDA makes data fit for purpose
  - needs to build on analysis strategy
  - Define roles of variables of interest
- Missing values
- Univariate distributions (predictors+outcome)
- Multivariate distribution (predictors)

Heinze et al, Regression without regrets', BMC Med Res Meth 2024

Table 2 Check list for an initial data analysis (IDA) plan

Topic	Item	Features
Prerequisites		
Research aim	PRE1	Define the research aim (descriptive, predictive, or causal) and phase of research (exploratory or confirmatory)
Analysis strategy	PRE2	Check specification of models and roles of variables in the models
Data dictionary	PRE3	For variables identified in PRE2, and any additional structural variables, check variable labels, definitions, values, units of measurement, data type, etc
Domain expertise	PRE4	When discussing analysis strategy with a domain expert, address: key predictors, structural variables for IDA, predictor grouping, expected missing values proportion, and predictor distributions/correlations
IDA screening domain: Missing values (predictor and outcome variables)		
Participant (unit) missingness	M1	Describe: number of potentially eligible but not assessed, assessed but not recruited, and recruited but didn't contribute data
Variable (item) missingness	M2	Provide missing value count and proportion for each predictor and the outcome variable. Distinguish by reason, if applicable
Complete cases	M3	Describe complete observations for outcome and predictors in any model described in PRE2
Patterns	M4	Investigate missing value patterns across all variables, structured by structural variables. Display as tables or appropriate visualizations
Missing values – Optional extensions		
Predictors	ME1	Investigate predictors of missingness (complete vs incomplete cases)
IDA screening domain: Univariate descriptions (structural variables, predictors and outcome)		
Categorical variables	U1	Summarize category frequencies and proportions, with appropriate plots. Summarize frequencies of collapsed categories as well
Continuous variables	U2	Inspect distributions with high-resolution histogram, summary of key quantiles (e.g. 1st, 5th, 25th, 50th, 75th, 90th, 99th) extreme values (5 highest and 5 lowest), measures of central tendency (mean) and dispersion (Gini mean difference, standard deviation, interquartile range). Include number of distinct values. Describe mode of the data and its frequency. Similarly, inspect distributions of transformed variables, if applicable
Univariate analyses – Optional extensions		
Sparsity	UE1	Create distributional plots to identify observations with extreme values
IDA screening domain: Multivariate descriptions (structural variables and predictors)		
Association	V1	Visualize and summarize the association of each predictor with the structural variables
Correlation	V2	Quantify association (e.g., pairwise correlations) between all key predictors in a matrix or heatmap
Interactions, if applicable	V3	Evaluate bivariate distributions of the predictors specified in interactions, incorporating appropriate graphical displays
Multivariate analyses – Optional extensions		
Correlation	VE1	Compare results from different association metrics
Clustering	VE2	Visualize clustering of predictors using a dendrogram to show closely associated predictors
Redundancy	VE3	Compute Variance Inflation Factors or fit parametric additive models to assess the predictability of each predictor from the remaining predictors

Potential impact of (outcome-agnostic) IDA

- Remove predictors
  - because of excessive missing values
  - because of redundancy
- Transform predictors
  - reparametrize a set of correlated predictors to increase interpretability:
    see Gregorich et al, 2021
  - to remove impact of high (low) data points
  - See Heinze et al, 2024
- Decide on functional form
  - Sparse distributions may call for simpler functional forms

Prediction modeling in the times of a pandemic

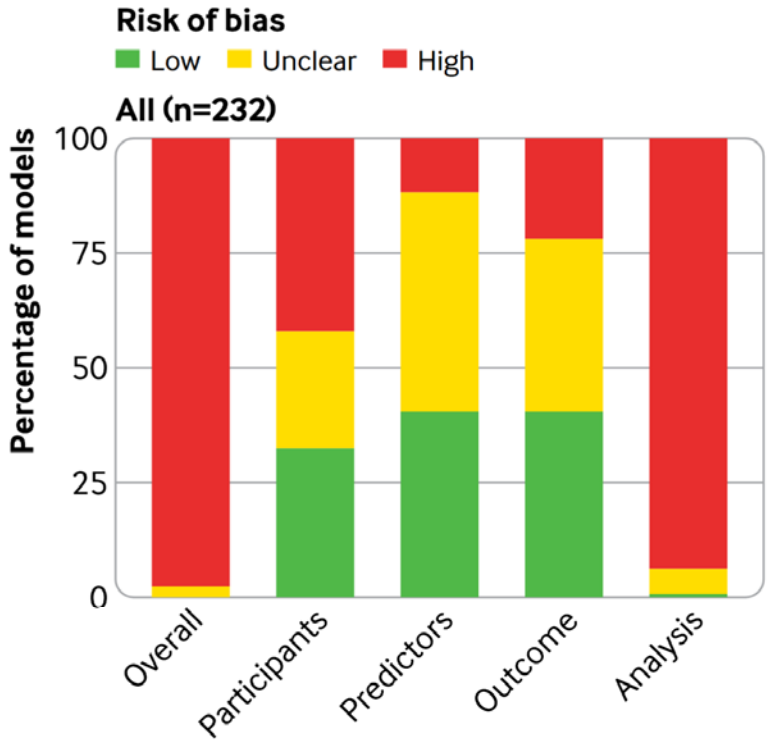
- Wynants et al reviewed 232 models prediction models for Covid-19

RESEARCH

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

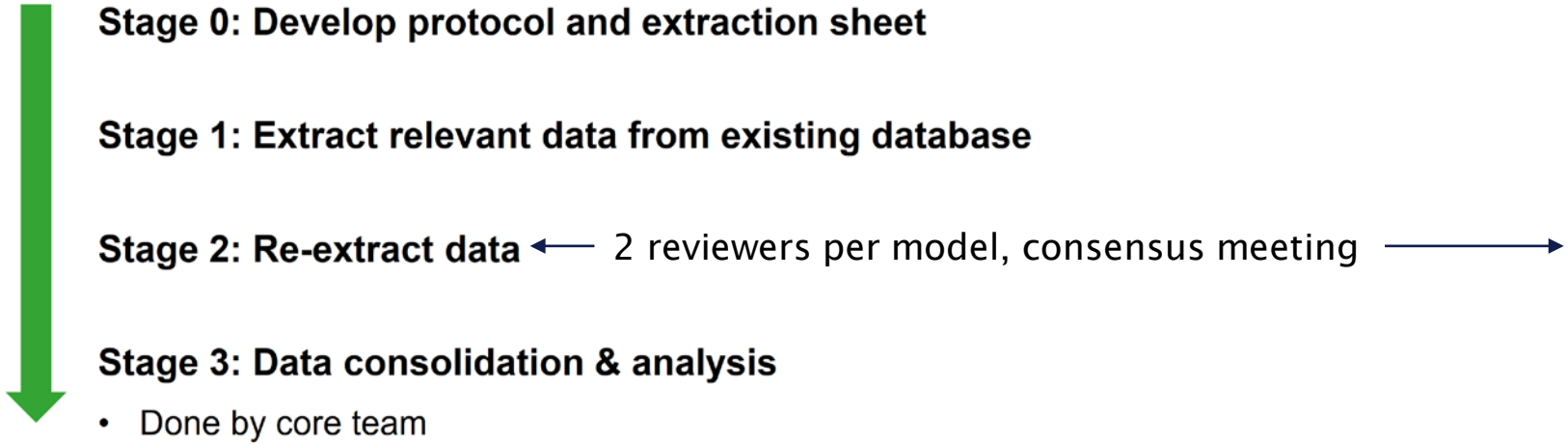
Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] Johanna A Damen,[8,9] Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,13] Paula Dhiman,[4,5] Maria C Haller,[7,14] Michael O Harhay,[15,16] Liesbet Henckaerts,[17,18] Pauline Heus,[8,9] Michael Kammer,[7,19] Nina Kreuzberger,[20] Anna Lohmann,[21] Kim Luijken,[21] Jie Ma,[5] Glen P Martin,[22] David J McLernon,[23] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[24,25] Chunhu Shi,[26] Nicole Skoetz,[19] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[27] René Spijker,[8,9,28] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[29,30] Sander M J van Kuijk,[31] Bas C T van Bussel,[1,32] Iwan C C van der Horst,[32] Florien S van Royen,[8] Jan Y Verbakel,[33,34] Christine Wallisch,[7,35,36] Jack Wilkinson,[22] Robert Wolff,[37] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

BMJ, 2020



A snapshot of the practice of model building

- We reanalyzed these studies for aspects of model building
- 181 models were selected based on our inclusion criteria



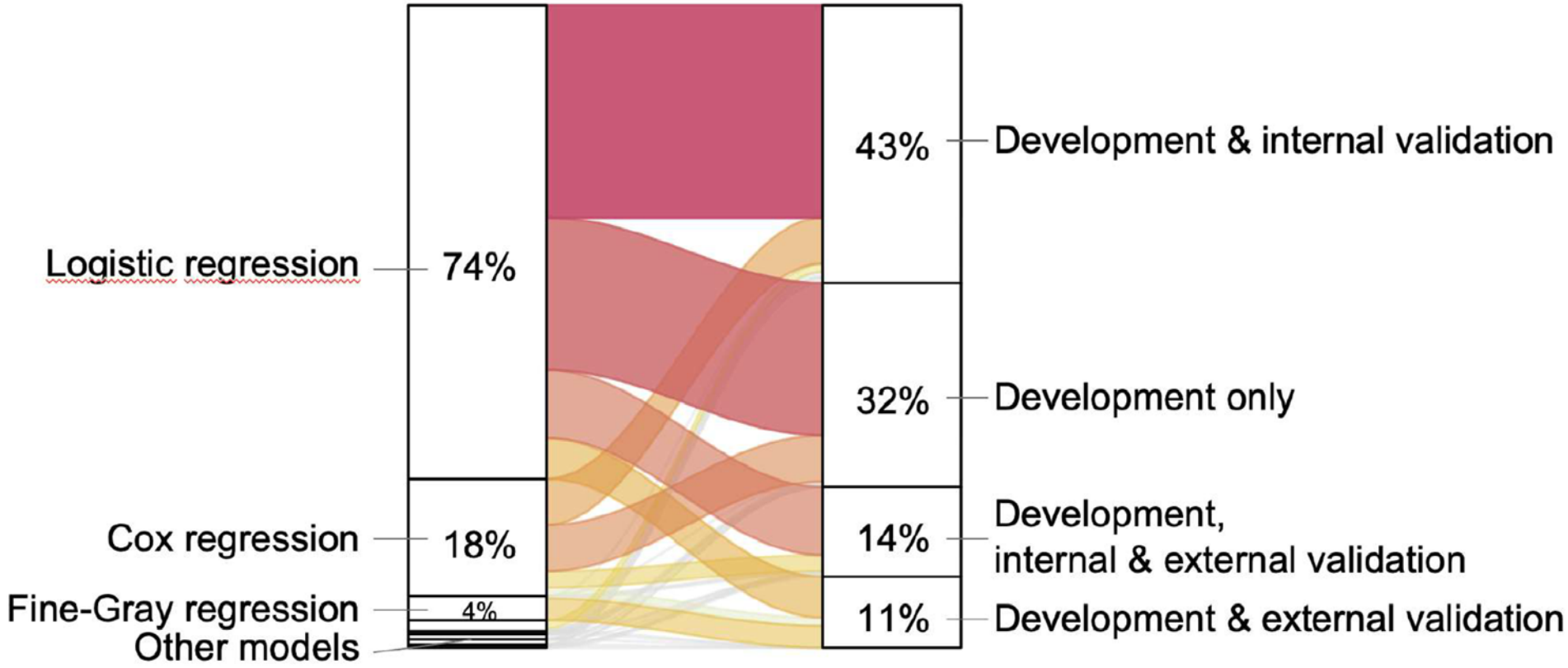
A big thank you to all our reviewers and supporters

Alexander Gieswinkel (Mainz)	James Chirombo (Blantyre)	Mariana Nold (Jena)
Alice Schneider (Berlin)	Johannes Vey (Heidelberg)	Moritz Pamminger (Vienna)
Andreas Klinger (Vienna)	Laure Wynants (Maastricht)	Theresa Ullmann (Vienna)
Daniel Schulze (Berlin)	Linard Hoessly (Basel)	Ulrike Grittner (Berlin)
Daniela Dunkler (Vienna)	Lorena Hafermann (Berlin)	Willi Sauerbrei (Freiburg)
David McLernon (Aberdeen)	Manuel Feißt (Berlin)	

- Core team of TG2:
  Michael Kammer (Vienna), Gregor Buch (Mainz), Marc Henrion (Malawi), G.H.

A snapshot of the practice of model building

Data extraction of 181 models completed February 2025



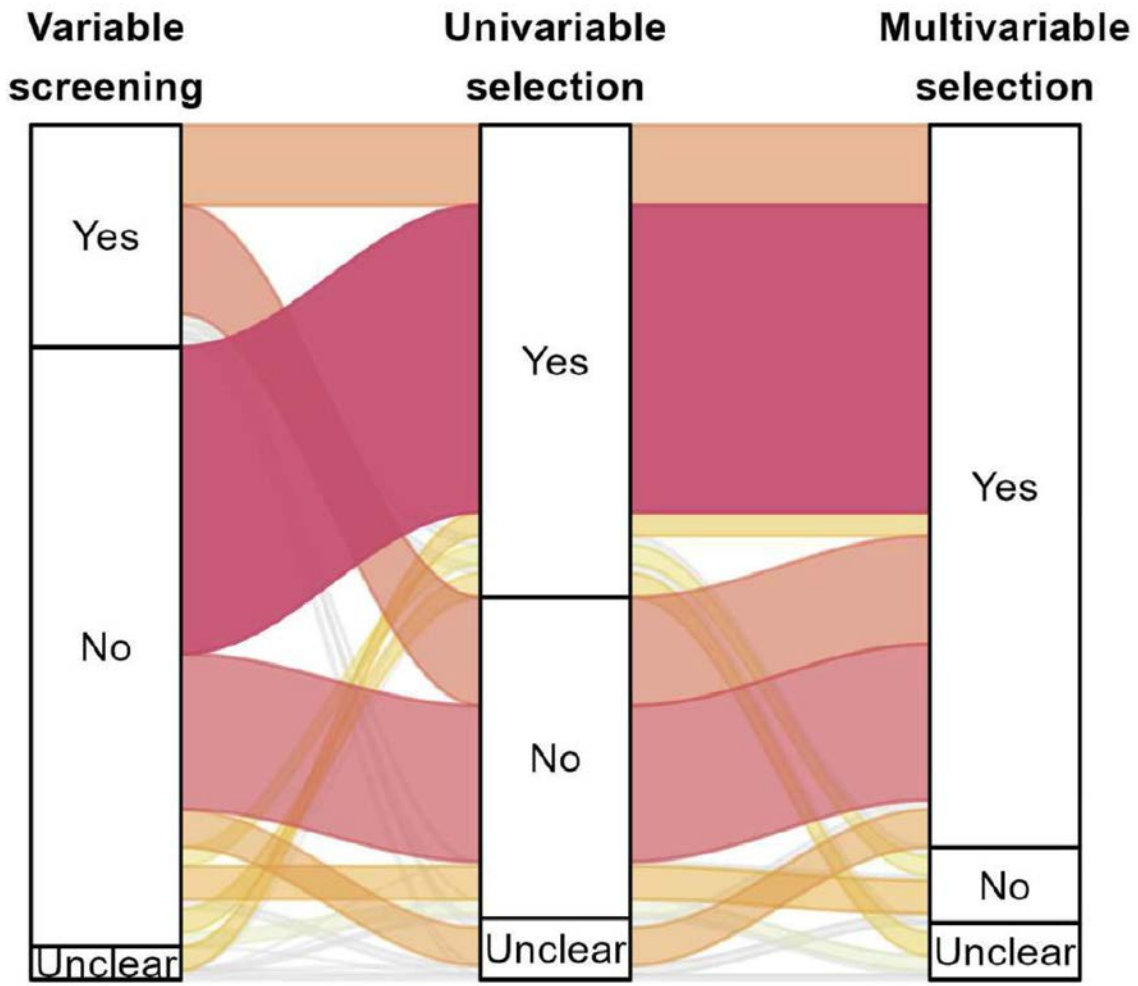
Median sample size 344 (IQR 156 - 982) with median 68 events (IQR 35 - 169)

16

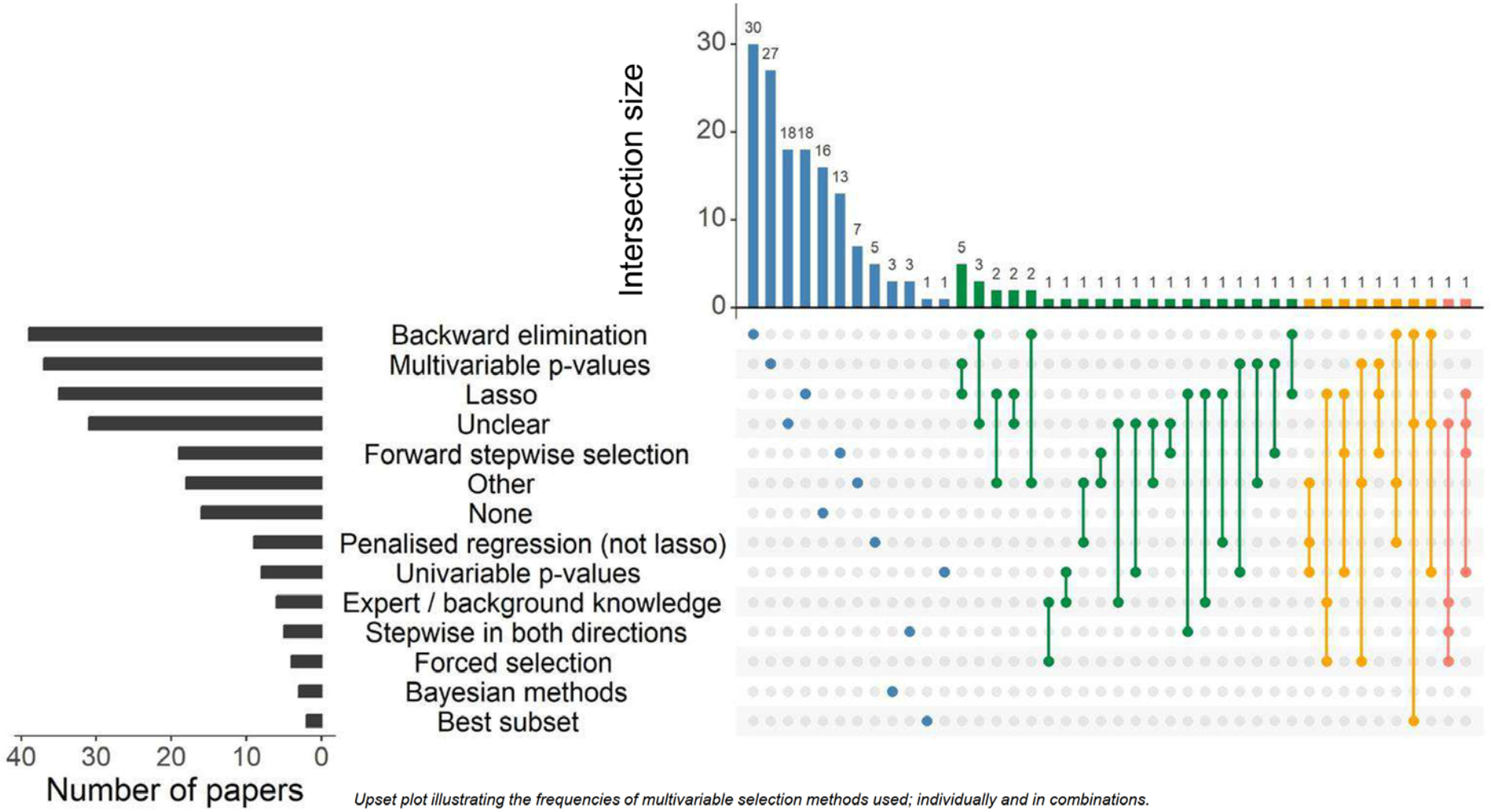
Results of model building review: variable selection

- Main path:
  - No variable screening
  - Univariable selection
  - and multivariable selection

- That is quite the opposite of what we preach

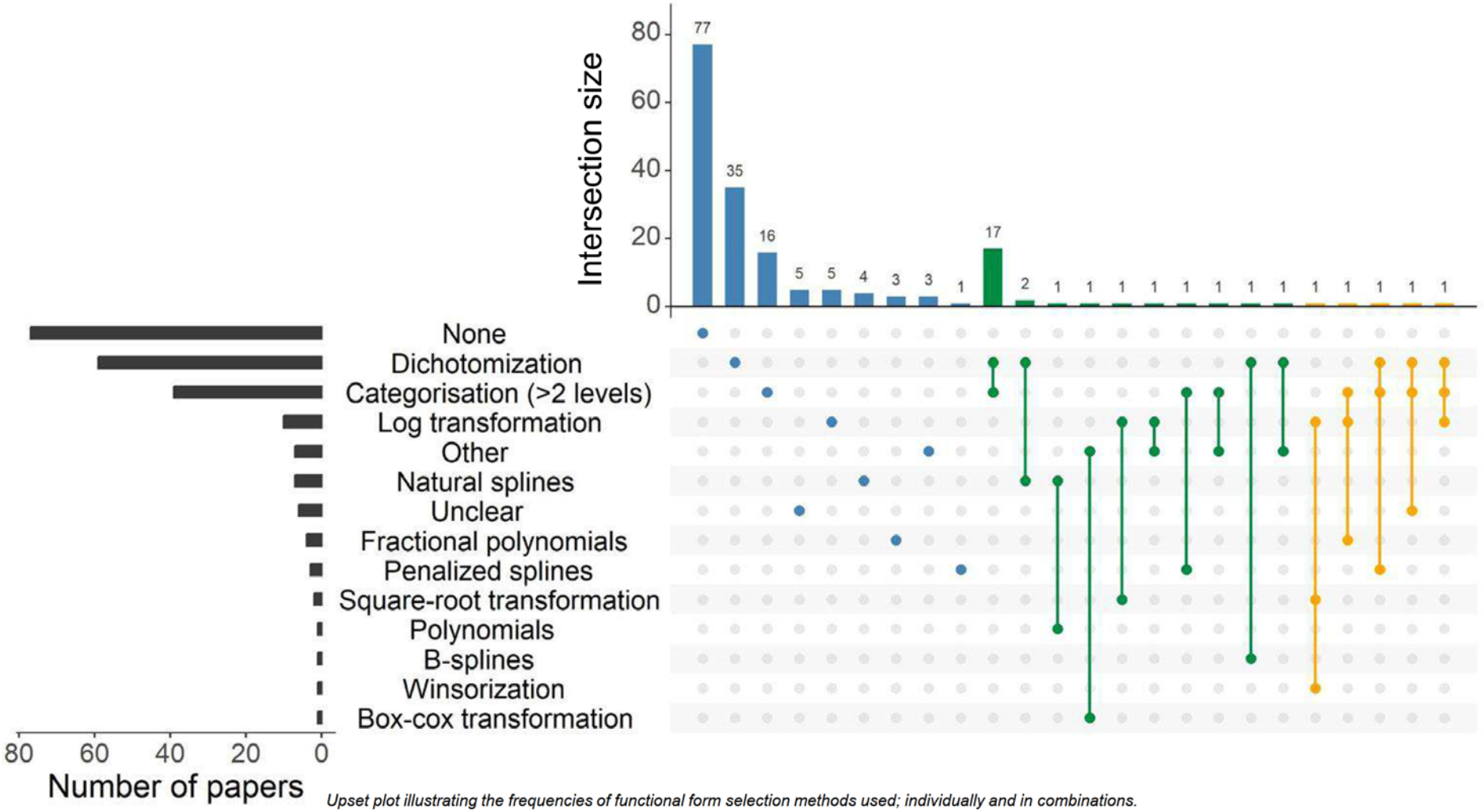


Results of model building review: variable selection methods



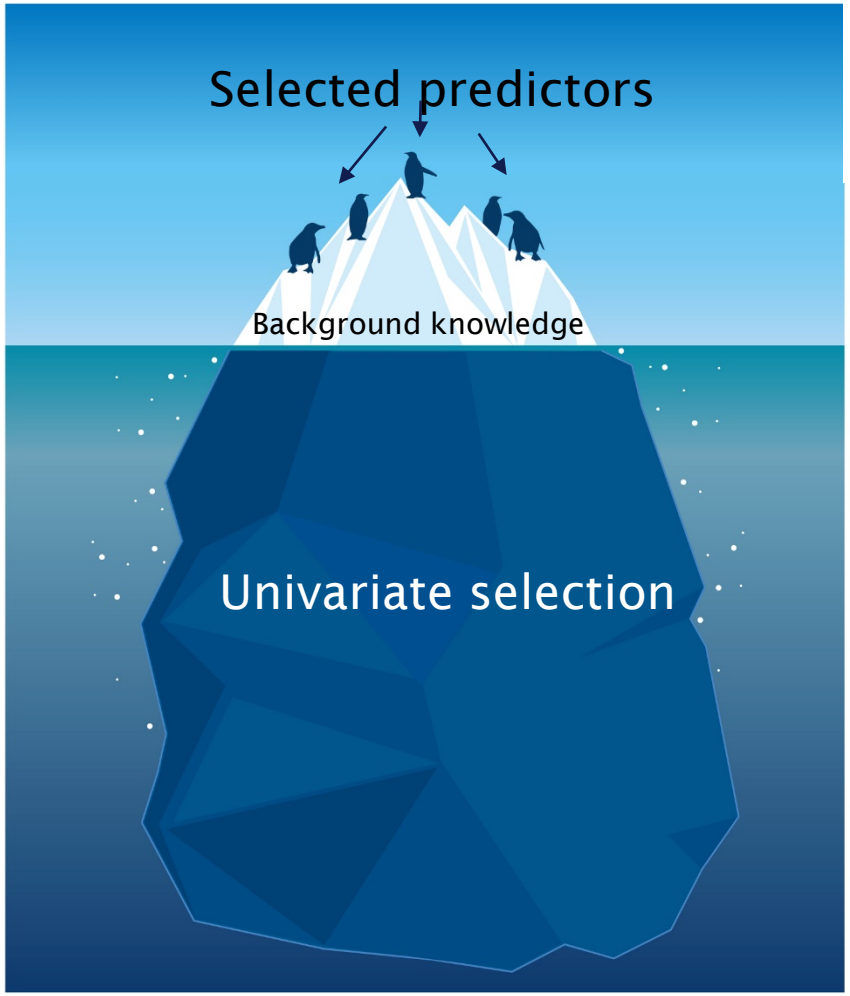
Upset plot illustrating the frequencies of multivariable selection methods used; individually and in combinations.

Results of model building review: functional forms



Upset plot illustrating the frequencies of functional form selection methods used; individually and in combinations.

Variable selection: current practice



Inappropriate Use of Bivariable Analysis to Screen Risk Factors for Use in Multivariable Analysis

Guo-Wen Sun, Thomas L. Shook, and Gregory L. Kay*

J Clin Epidemiol Vol. 49, No. 8, pp. 907–916, 1996

- Many authors have advised against using univariable selection (e.g., Sun et al. 1996, Harrell 2015, Heinze and Dunkler, 2017)
- Our review of model building strategies and many others identified that univariate selection ("bivariable analysis") is still in wide use
  - (and its actual, silent use may be even much more widespread)
- Outcome-agnostic data screening is often not performed, or not mentioned
  - Was it done before locking the statistical analysis methods?

Comparing methods - a review and new simulations

REVIEW ARTICLE

Biometrical Journal

Variable selection – A review and recommendations for the practicing statistician

Georg Heinze | Christine Wallisch | Daniela Dunkler

PLOS ONE

STUDY PROTOCOL

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

Theresa Ullmann[1], Georg Heinze[1], Lorena Hafermann[2], Christine Schilhart-Wallisch[1,3], Daniela Dunkler[1]*, for TG2 of the STRATOS initiative[¶]

Our recommendations:

- Generate an initial working set of variables
- Decide whether VS should be applied, by which method and to which variables
  - Only variables where inclusion is questionable
  - BE preferred
  - Only with EPV>25
  - 10<EPV<25: with shrinkage methods
  - EPV<10: no
- Perform sensitivity analyses and stability investigations
- Solve the problem of selective inference

Challenge these recommendations

Main aims of this simulation study

- Descriptive modeling:
  - Correct/false inclusion/exclusion rates
  - Bias/variance/MSE of regression coefficients (also descriptive)
  - Validity of confidence intervals
- Predictive modeling:
  - Prediction error
  - Calibration
  - Variable importance ranking (also descriptive)

Data generating mechanisms

- 20 potential predictors, among them 10 ‚true' and 10 ‚non-' predictors
- Distributions:
  - Resembling real data (NHANES, mixed types) or multivariate normal
- Correlation structures:
  - Resembling real correlation structure (NHANES), or independence
- True effects:
  - Standardized coefficients distributed from 1.5 to 0.25
  - Continuous and binary outcome

	linear regression	logistic regression	
		event rate 0.3	event rate 0.05
main scenario...			
...with linear effects	setting 1 $R^2 = 0.45$	setting 4 $R^2_{CS} = 0.40$	setting 6 $R^2_{CS} = 0.16$
...with nonlinear effects	setting 1b $R^2 = 0.45$	setting 4b $R^2_{CS} = 0.43$	setting 6b $R^2_{CS} = 0.20$
low R^2 scenario...			
...with linear effects	setting 2 $R^2 = 0.15$	setting 5 $R^2_{CS} = 0.13$	setting 7 $R^2_{CS} = 0.05$
...with nonlinear effects	setting 2b $R^2 = 0.15$	setting 5b $R^2_{CS} = 0.14$	setting 7b $R^2_{CS} = 0.07$
high R^2 scenario...			
...with linear effects	setting 3 $R^2 = 0.7$		
...with nonlinear effects	setting 3b $R^2 = 0.7$		

Sample sizes and EPV values:

linear regression	n	100	200	400	*500*	800	1600	3200	6400
	EPV	5	10	20	*25*	40	80	160	320
logistic regression, event rate 0.3	n	183	365	730	*1667*	1461	2922	5844	11,687
	EPV	2.75	5.48	10.95	*25*	21.92	43.83	87.66	175.31
logistic regression, event rate 0.05	n	2000	4000	8000	*10,000*	–	–	–	–
	EPV	5	10	20	*25*	–	–	–	–

Ullmann et al, 2026

Methods

- We considered three groups of methods(criteria), with different target characteristics:
  - ‚Mild selectors' (noise removers): target TPR>0.95, FPR<0.75
    - Lasso(CV), BE(0.5)
    - Aiming at removing the noise (screening)
  - ‚Intermediate selectors' (predictor selectors): target TPR>0.8, FPR<0.2
    - BE(AIC), AdaLasso(CV), Rlasso(CV)
    - Aiming at prediction performance
  - ‚Strict selectors' (model identifiers): target TPR>0.5, FPR<0.1
    - BE(0.05), BE(BIC), Rlasso(BIC)
    - Aiming at identifying the data generating model
- Comparisons within these groups:
  - At which sample size/R^2 can these methods be safely used?

Estimands and targets

- Regression coefficients and their 95% confidence intervals (when available)
- Selection behaviour
- Predictions in validation set

Selection:
Mild (noise removers)
Intermediate (predictor selectors)
Strict (model identifiers)

Performance measures

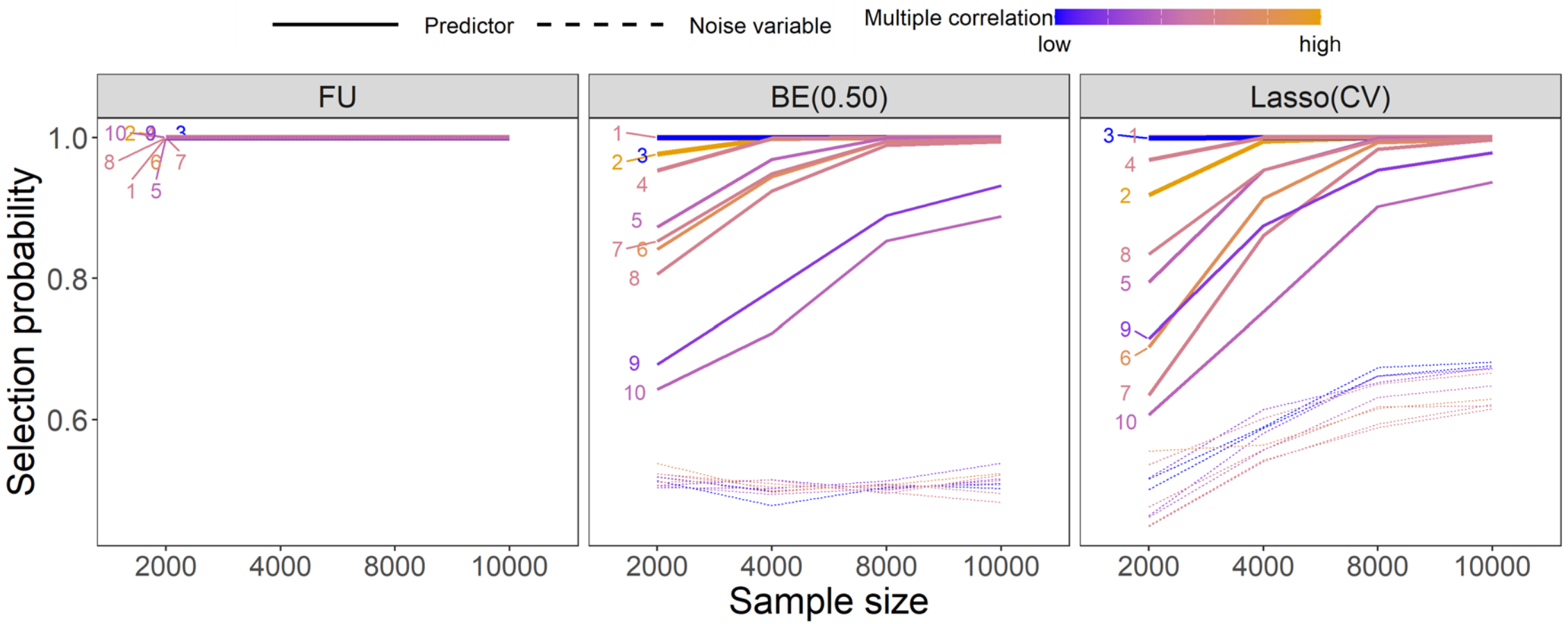
- Targeting regression coefficients: bias, MSE
- Targeting importance of variables: Kendall τ_B (true vs. estimated β_{STD})
- Targeting CI: coverage, width, probability to exclude 0
- Targeting selection:
  TPR, FPR, Model identification rate, over- and underselection, model size
- Targeting predictions: local/global bias, RMSPE, MAE, AUC, calibration

Some results

- Some results of the simulation study based on our interactive Shiny app
  - Focus on scenario with:
    binary outcome, realistic data, R2=0.05 (weak signal), event rate 0.05
  - Comparison with event rate 0.3, and with R2=0.13 (strong signal)
- Performance measures are shown per sample size:
  - For noise removers:
    selection probabilities, coverage probabilities
  - For prediction model selectors:
    RMSPE, model size
  - For descriptive model identifiers:
    model selection rates, variable importance ranking

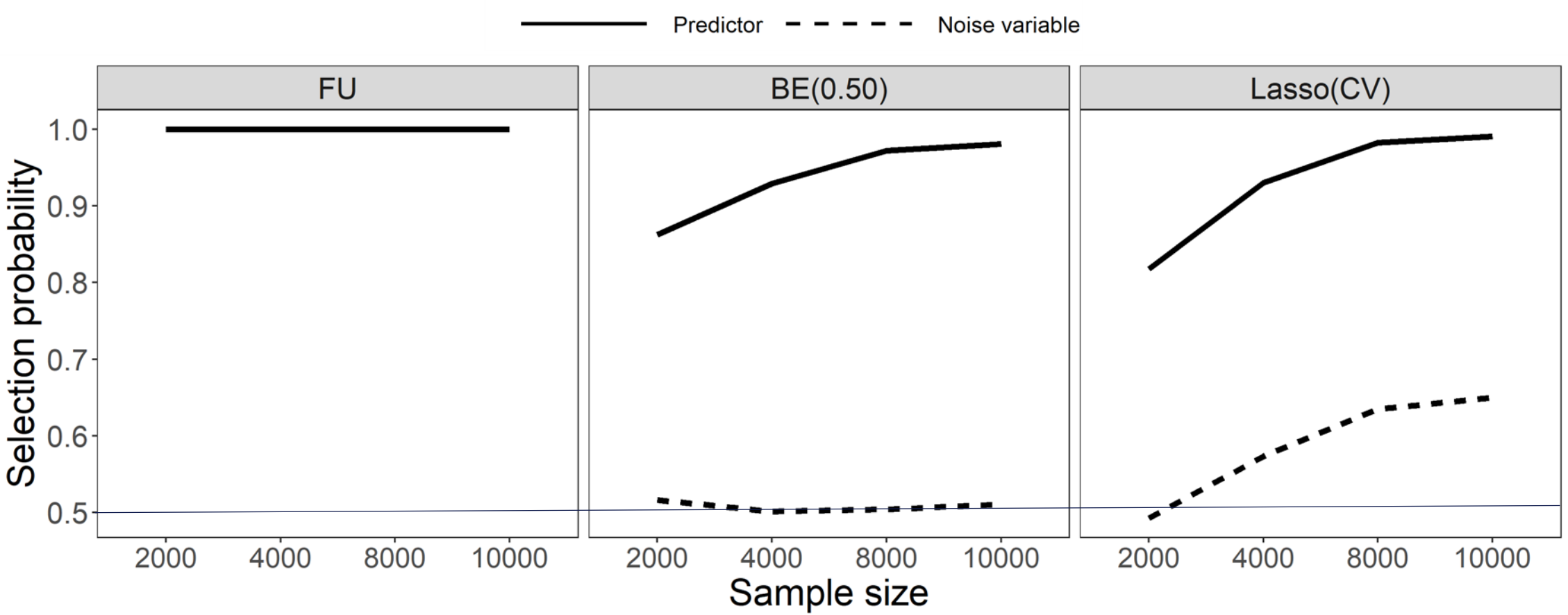
Noise removers: selection probabilities

- Logistic regression, event rate 0.05, R2=0.05, realistic predictor distributions



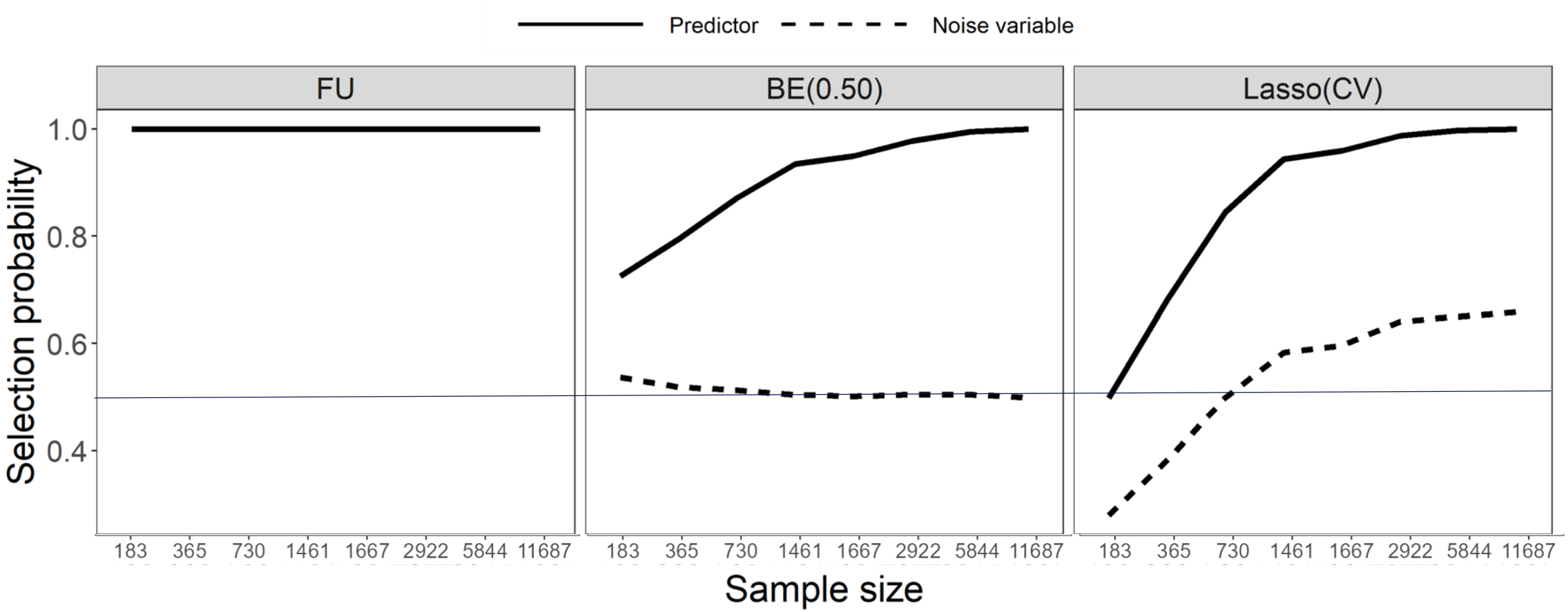
Noise removers: selection probabilities (averaged)

- Logistic regression, event rate 0.05, R2=0.05, realistic predictor distributions



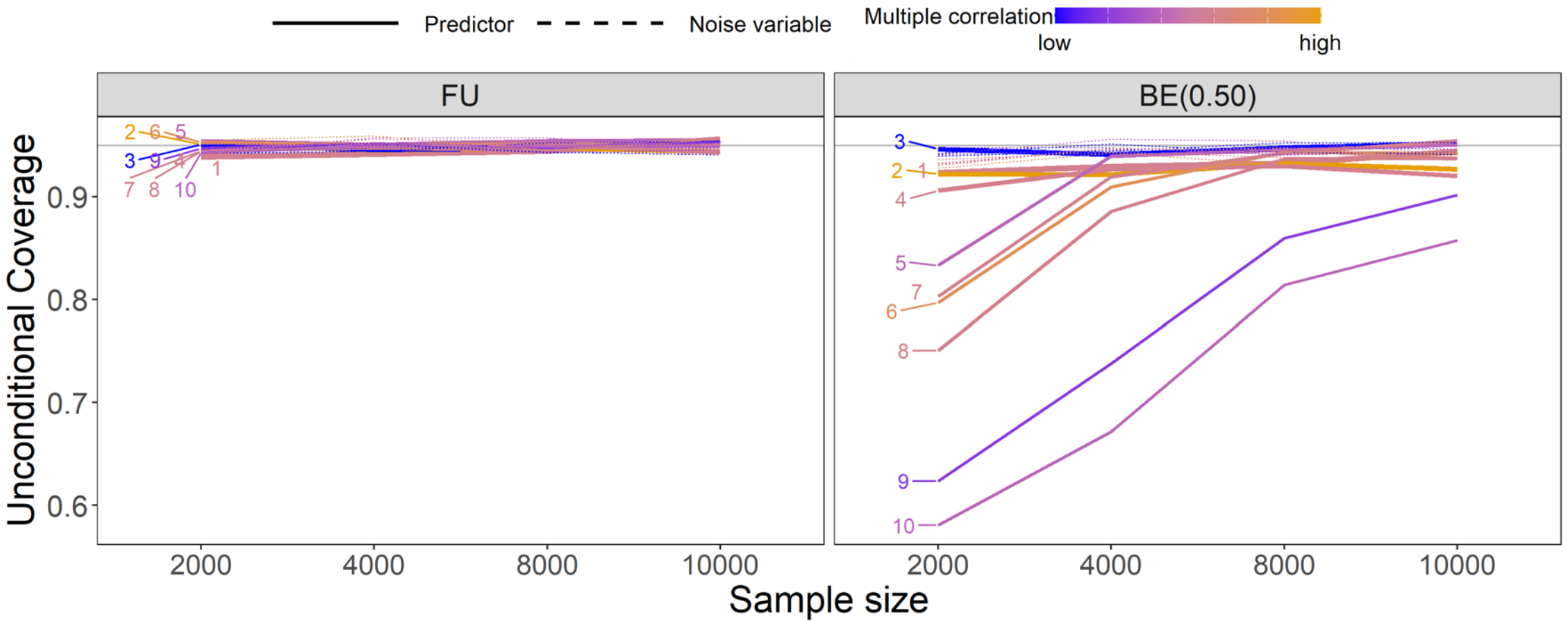
Noise removers: selection probabilities (averaged)

- Logistic regression, event rate 0.3, R2=0.05, realistic predictor distributions



Noise removers: coverage of 95%CI

- Logistic regression, event rate 0.05, R2=0.05, realistic predictor distributions

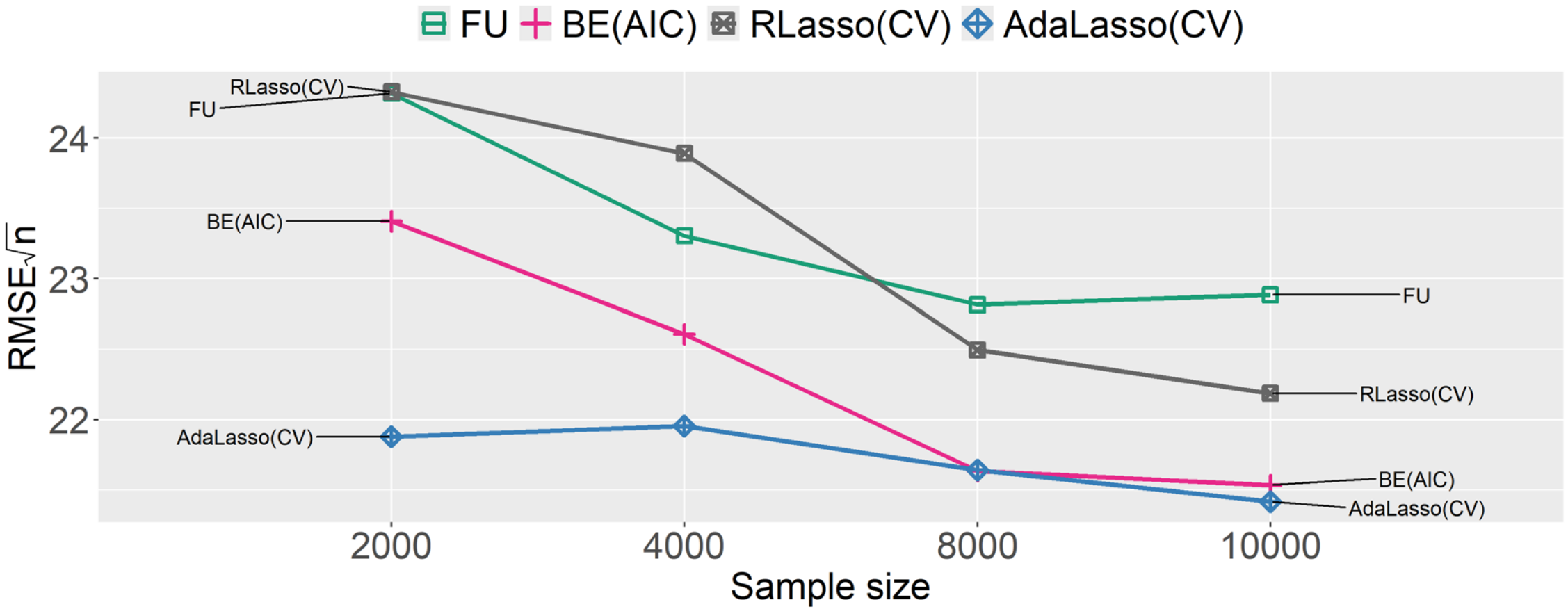


Noise removers: summary

- Inconsistent behaviour of Lasso
  - With higher sample sizes, more noise variables are selected
- Weak signal: needs large sample sizes to meet targets
- Strong signal, BE(0.5):
  - LogReg: ≥20 EPV
  - LinReg: >25 EPV
- Coverage of confidence intervals largely violated, even for very mild selection (BE(0.5)) and high sample sizes
  - Need for addressing postselection inference!

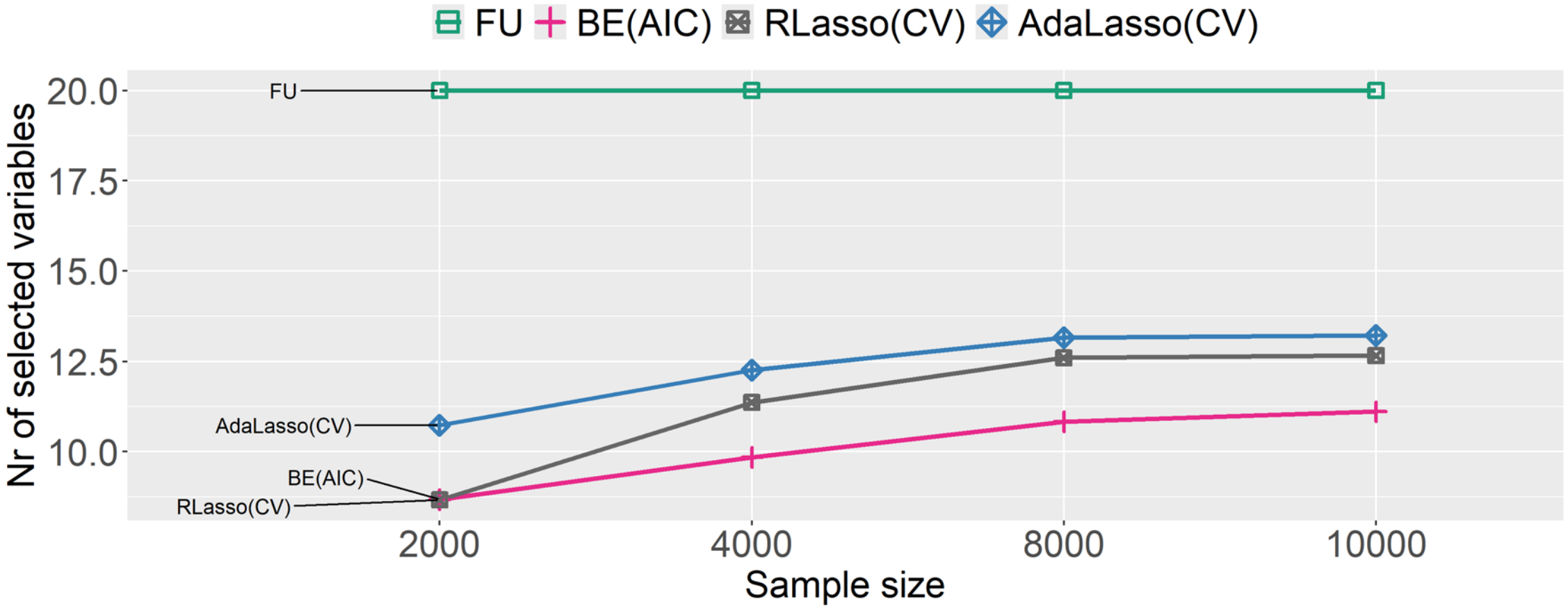
Predictor selectors: RMSPE (root mean squared prediction error)

- Logistic regression, event rate 0.05, R2=0.05, realistic predictor distributions



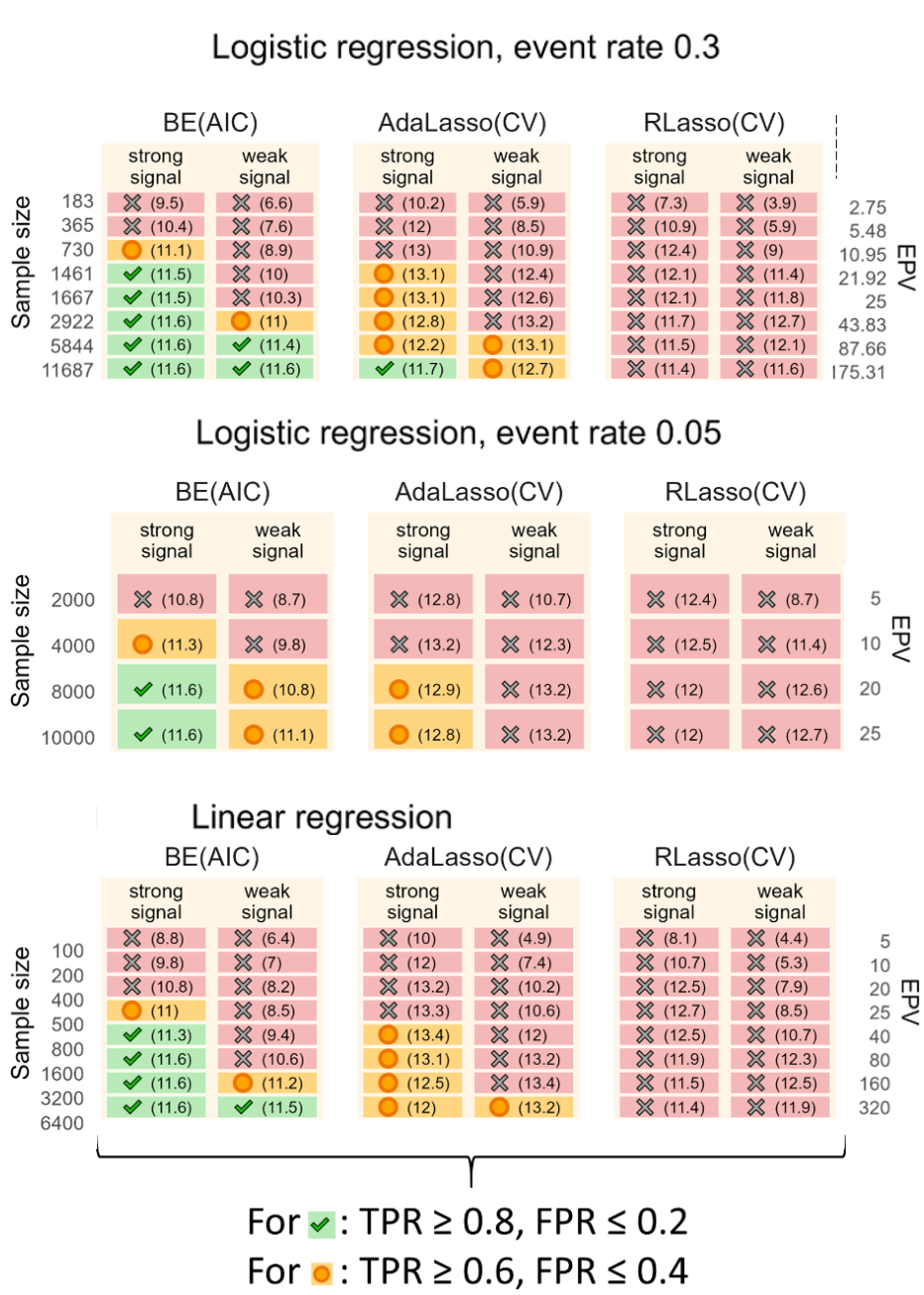
Predictor selectors: model size

- Logistic regression, event rate 0.05, R2=0.05, realistic predictor distributions



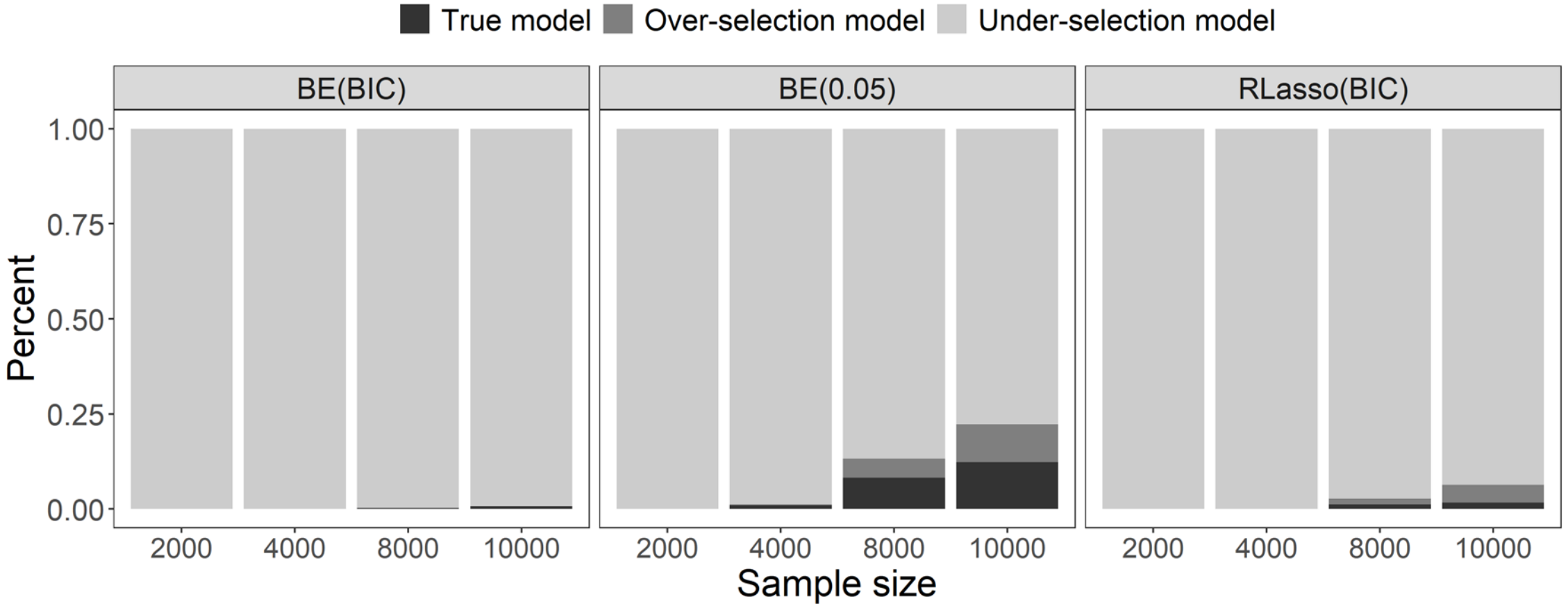
Predictor selectors: summary

- While AdaLasso achieves better prediction performance (AUC, RMSPE, ICI), it achieves a higher model size than RLasso or BE(AIC)
  - It always includes noise predictors
- RLasso has problems with a noise predictor highly correlated to a real predictor
  - This improved with independent predictors



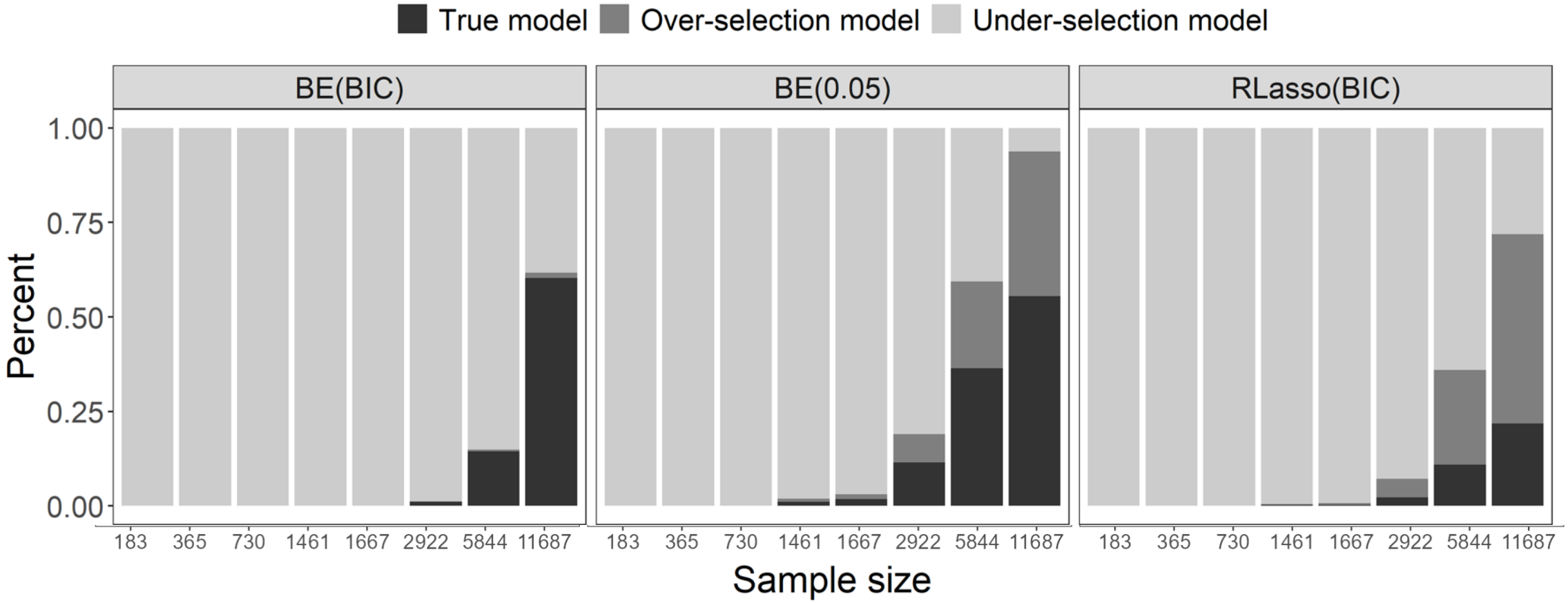
Model identifiers: model selection rates

- Logistic regression, event rate 0.05, R2=0.05, realistic predictor distributions



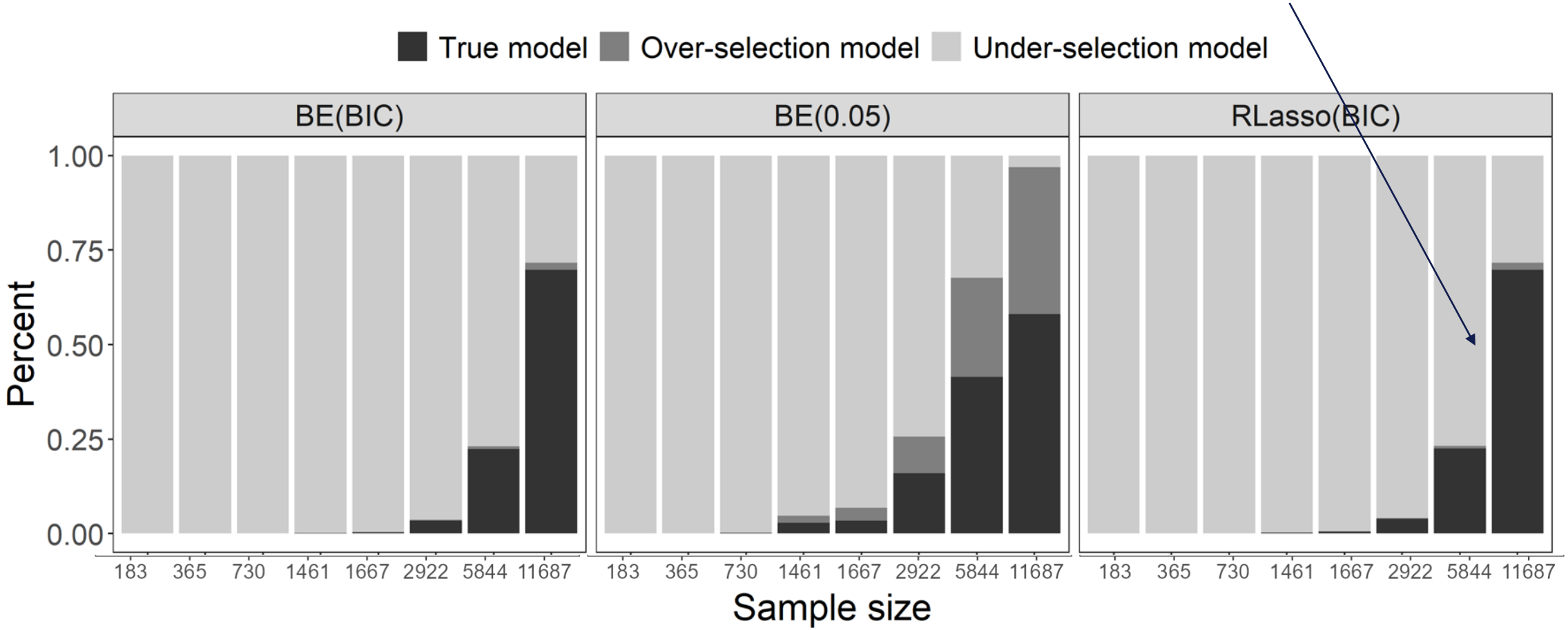
Model identifiers: model selection rates

- Logistic regression, event rate 0.3, R2=0.05, realistic predictor distributions



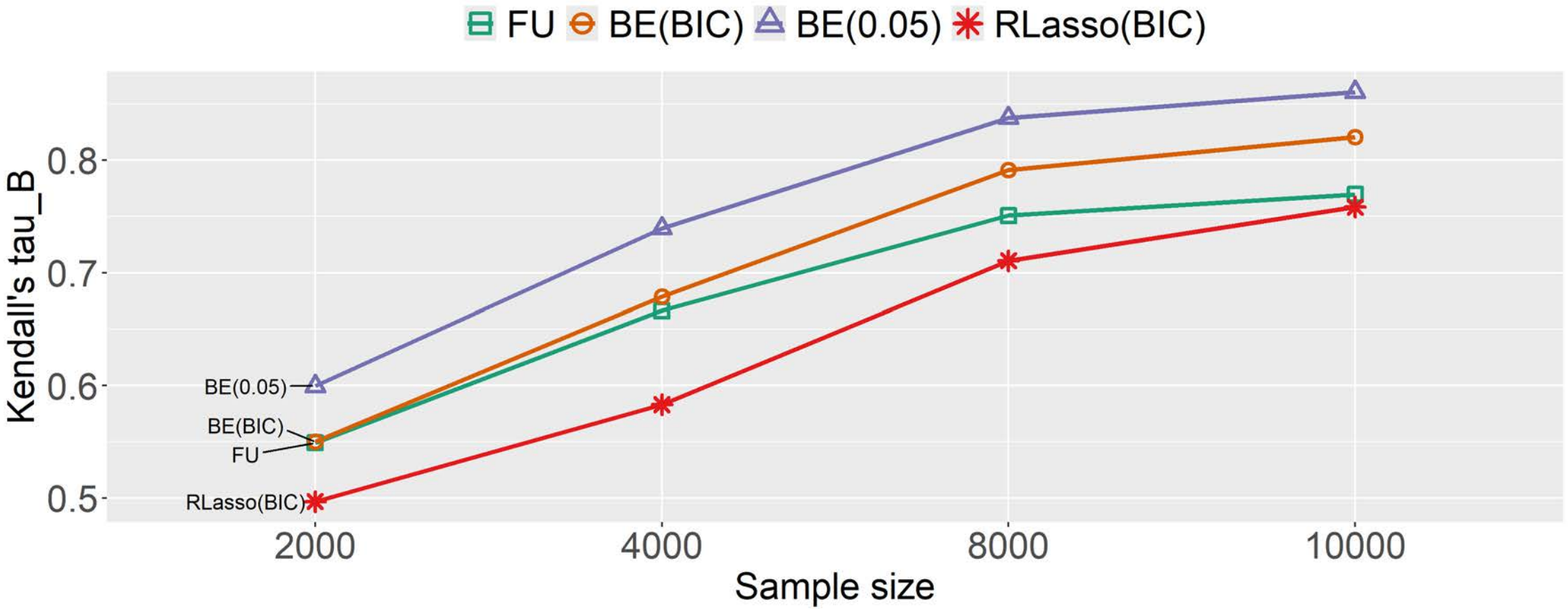
Model identifiers: model selection rates

- Logistic regression, event rate 0.3, R2=0.05, no correlation between predictors



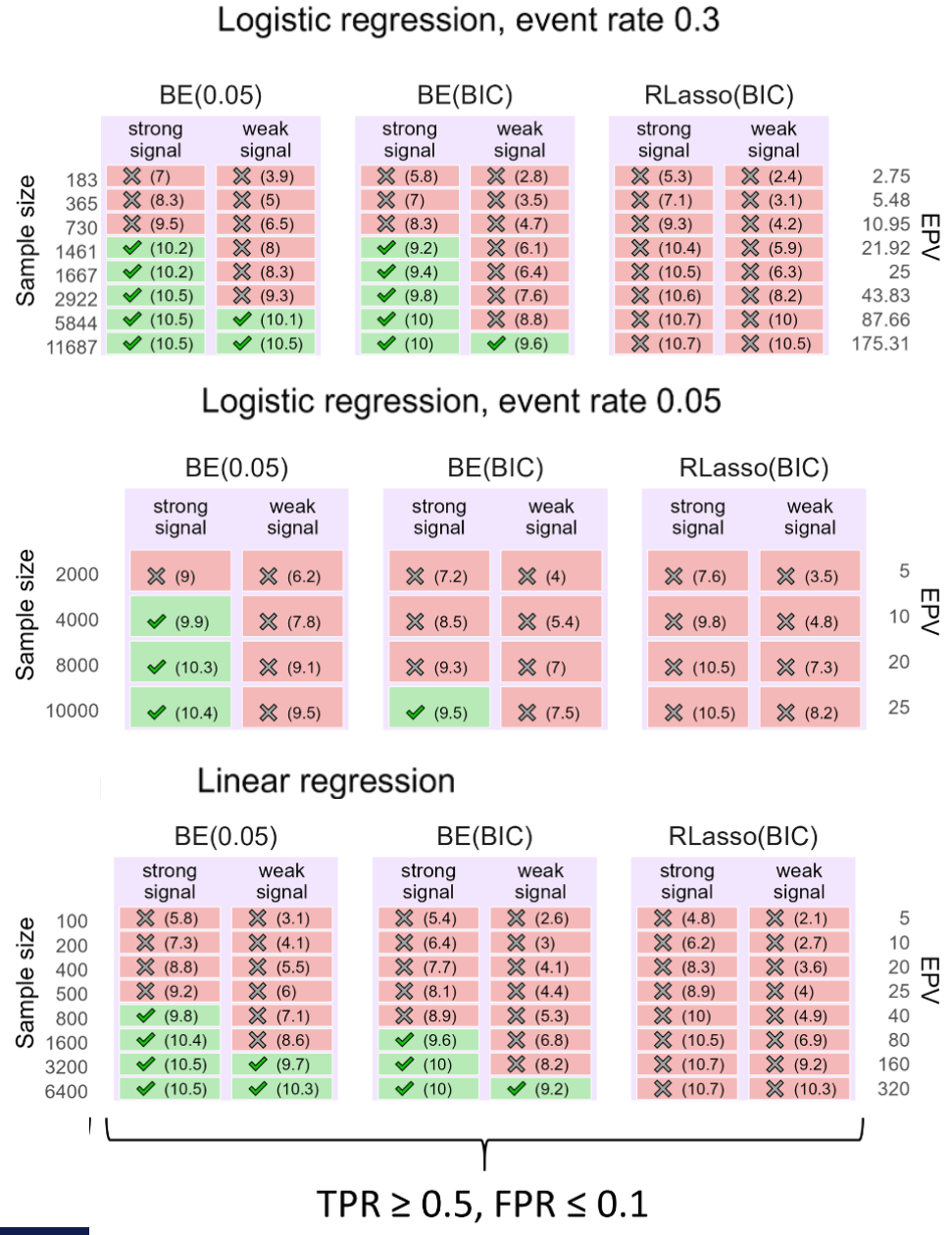
Model identifiers: importance ranking (of predictors)

- Logistic regression, event rate 0.05, R2=0.05, realistic predictor distributions

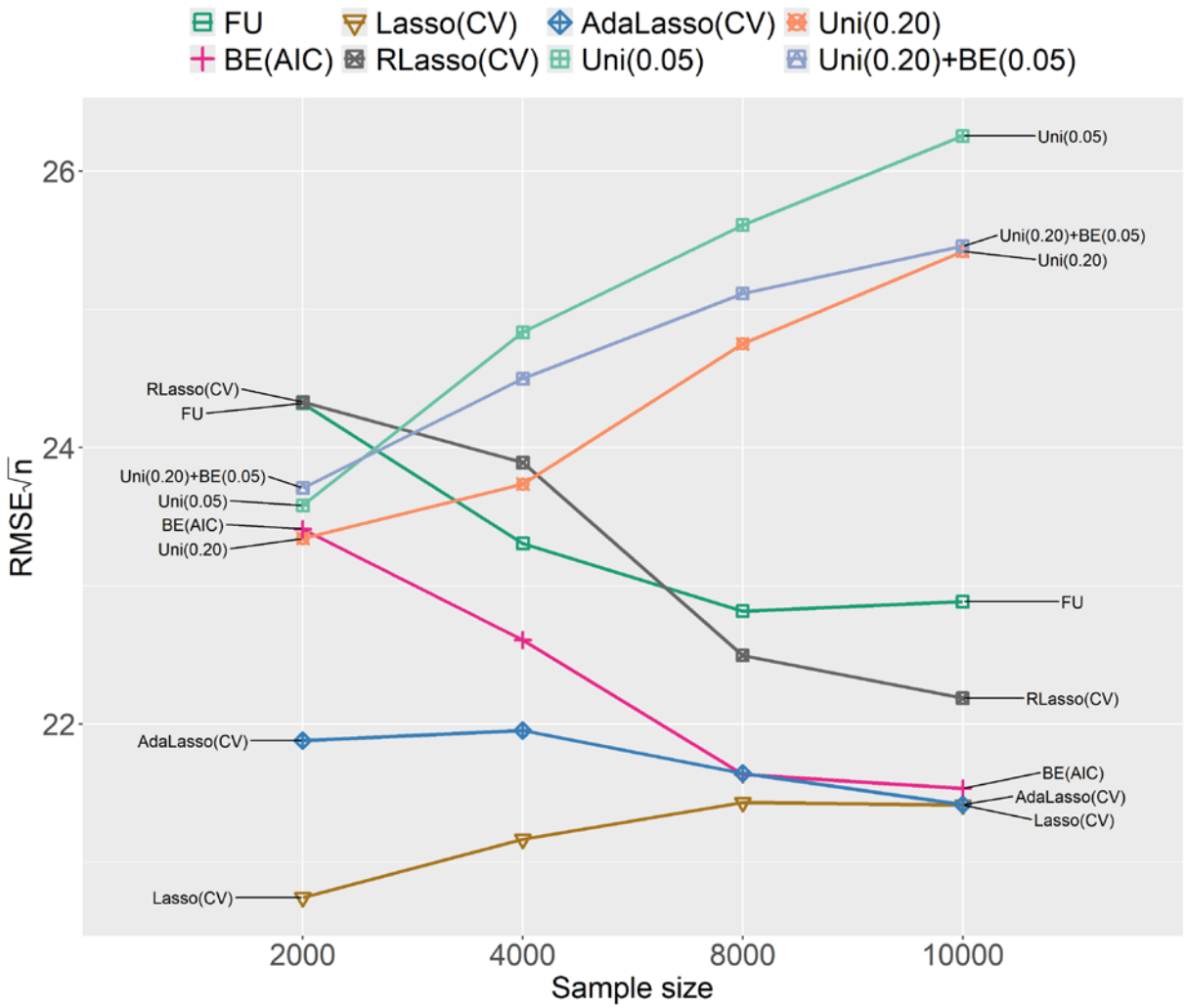


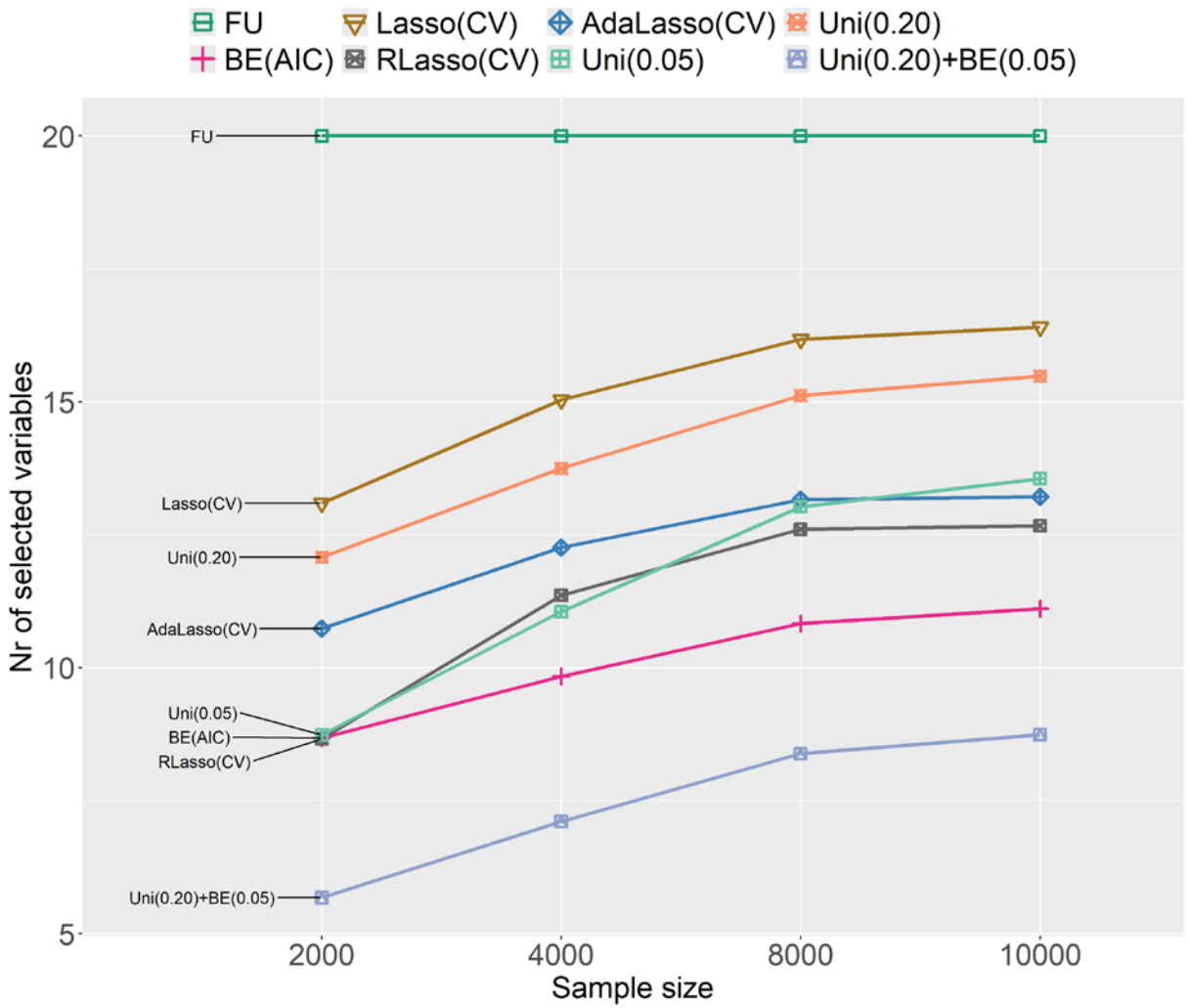
Model identifiers: summary

- The ‚true model' is rarely selected, even at high sample sizes or strong signals
- Mostly predictors are missed
- Rlasso overselected (correlation noise-true predictor)
- Our mild criteria (TPR>0.5, FPR<0.1) are met by BE(0.05) with strong signals and EPV>=20/40, but BE(BIC) needed higher EPV
- BE is better able than RLasso to deal with correlation patterns



Now the fun part: what about univariate selection?





Some summary of simulation study

- An interactive Shiny app allows us to perform targeted comparisons
- Scenarios were well aligned such that ‚patterns' were similar between LinearReg, LogReg(0.05), and LogReg(0.3)
- Sample size and R^2 are the key decisive quantities
- Consistent VS methods: those for which performance increase with sample size:
  - Lasso worked fine as ‚prediction machine'
  - Any further restrictions (AdaLasso, Rlasso) to reduce model size decreased its performance
  - BE(AIC) better than its reputation
- Inconsistent: performance may decrease with sample size
  - Any type of univariate selection, whether or not combined with BE
- At low sample sizes: there is no way around regularization
  - At the cost of selection properties
  - Performance of RLASSO improved when we removed the strong correlation of a noise variable with a strong predictor
  - Others have found superior performance of Relaxed Lasso (with tuned criterion) (Hastie, Tibshirani, Tibshirani, Statistical Science 2020)
  - Indicates necessity of several different studies to compare methods
- Don't trust the confidence intervals, even not at BE(0.50)

Towards recommendations

- What are your expectations?
  - Mild, intermediate, strict selection?
  - Is the signal/sample size big enough?
  - Choose accordingly but avoid univariate selection
  - Only variables where inclusion is unclear should be subjected to selection
- Stability analyses: use the bootstrap/subsampling to evaluate model stability

Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling

Christine Wallisch[1] | Daniela Dunkler[1] | Geraldine Rauch[2,3] | Riccardo de Bin[4] | Georg Heinze[1]

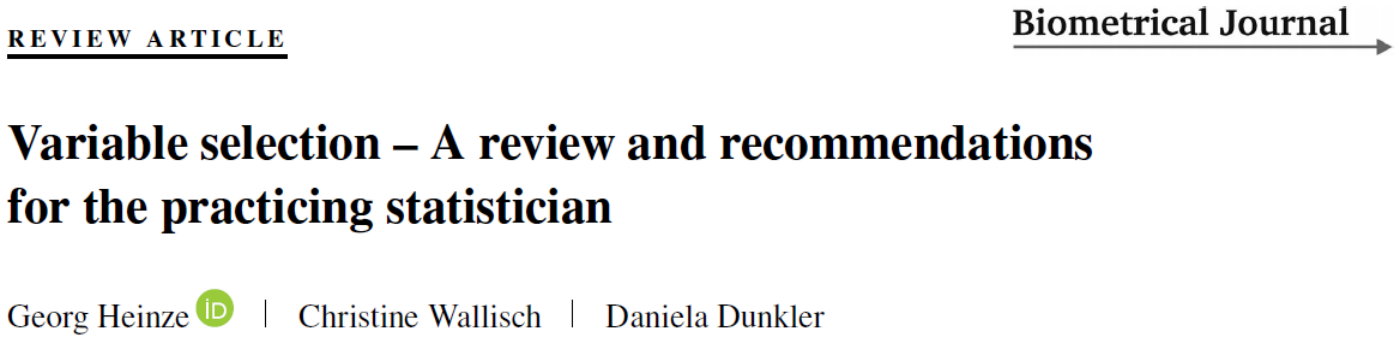
Statistics in Medicine

Statistics in Medicine. 2021;40:369–381.

- Report not only final model, but also how you got there

Recommendations on inference

- In our review we gave some pragmatic recommendations on inference



- Main roadmap: inference is OK in starting (‚global') model
  - Reduced model can be seen as projection
- Bootstrap may help in assessing uncertainties when selection was performed
  - See also prediction stability plots (Riley and Collins, 2023)
  - Probably not fully correct (see work of Leeb and Pötscher, 2005, 2008)
  - Conservative (Akbari et al, 2026; Wallisch et al, 2021)

Combining variable with functional form selection: first the DON'T

Journal of Clinical Epidemiology 98 (2018) 133–143

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Poor performance of clinical prediction models: the harm of commonly applied methods

Ewout W. Steyerberg[a,b,*], Hajime Uno[c], John P.A. Ioannidis[d,e,f,g], Ben van Calster[a,h], Collaborators

- Training cohort:
  48 events out of 870
- 37 candidate predictors
  - All continuous predictors dichotomized
  - Univariable selection ($p<0.05$), then backward selection ($p<0.05$)
- Simulation of this strategy using small subsets of a large contemporary cohort
  (2,051 events among N=19,66
  - Poor performance: median AUROC = 0.74
- Better performance if using external knowledge and continuous predictors (AUROC =0.836)

Combining variable with functional form selection:
Better grounded approaches

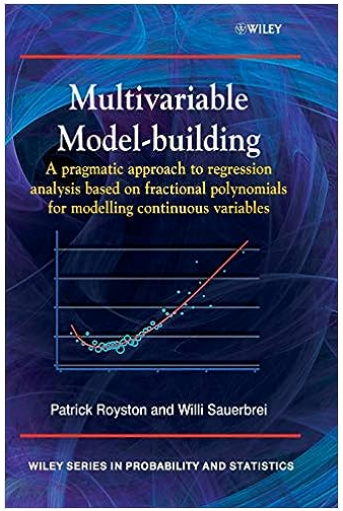
- Multivariable fractional polynomials (MFP, Sauerbrei, Royston and Binder, StatMed 2007)
- Multivariable regression splines (MRVS, Royston and Sauerbrei, STATA J 2007)
- Penalty-based approaches (see Kovacs, CompStat 2025):

Table 1 Summary of the feature selection algorithms investigated in this paper

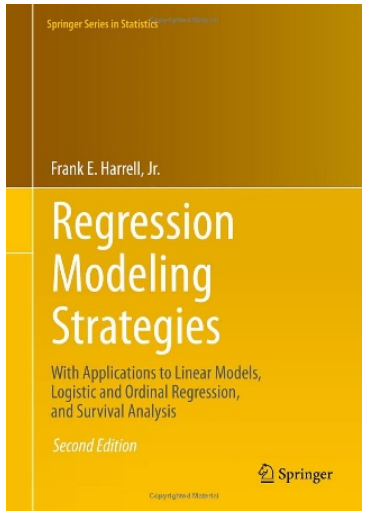
Cluster	Algorithm	Original paper	Software implementation
Stepwise	Forward selection	Efroymson (1960)	R package *gam* (Hastie 2018)
	Backward elimination		
	Bidirectional elimination		
Boosting	GAMBoost	Binder and Tutz (2008) and Schmid and Hothorn (2008)	R package *GAMBoost* (Binder and Tutz 2008)
	Modified backfitting	Belitz and Lang (2008)	R package a *R2BayesX* (Umaluf et al. 2015)
Regularization methods	Double penalty approach	Marra and Wood (2011)	R package *mgcv* (Wood 2017)
	Shrinkage approach		
	Cosso method	Lin and Zhang (2006)	R package *cosso* (Zhang and Lin 2013)
	Non-negative garrote component selection	Cantoni et al.(2011)	Own R script bases on the *ncvreg* package (Breheny and Huang 2011)
HSIC based methods	mRMR	De Jay et al. (2013)	R package *mRMRe* (De Jay et al. 2013)
	HSIC-Lasso	Climente-González et al. (2019)	Python package *pyHSICLasso* (Climente-González et al. 2019)

Combining variable and functional form selection: Textbook philosophies

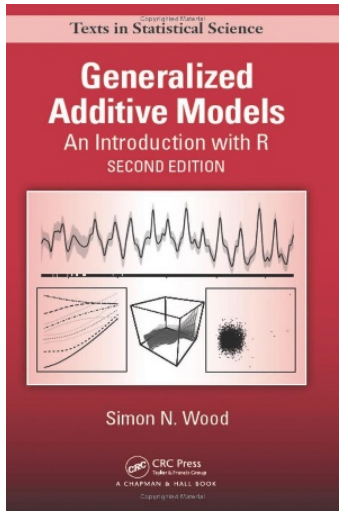
	Multivariable fractional polynomials (mfp2)	Restricted cubic splines (rms)	Penalized/thin plate splines (mgcv)
Selection	Significance-based	No	Penalty-based
Smoothing	Global: x^{p_1}, x^{p_2}	Local: spline based	Local: spline based
Basis functions (4df)	2 per variable (FP2)	4 per variable	‚many' per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



Royston and Sauerbrei, 2008



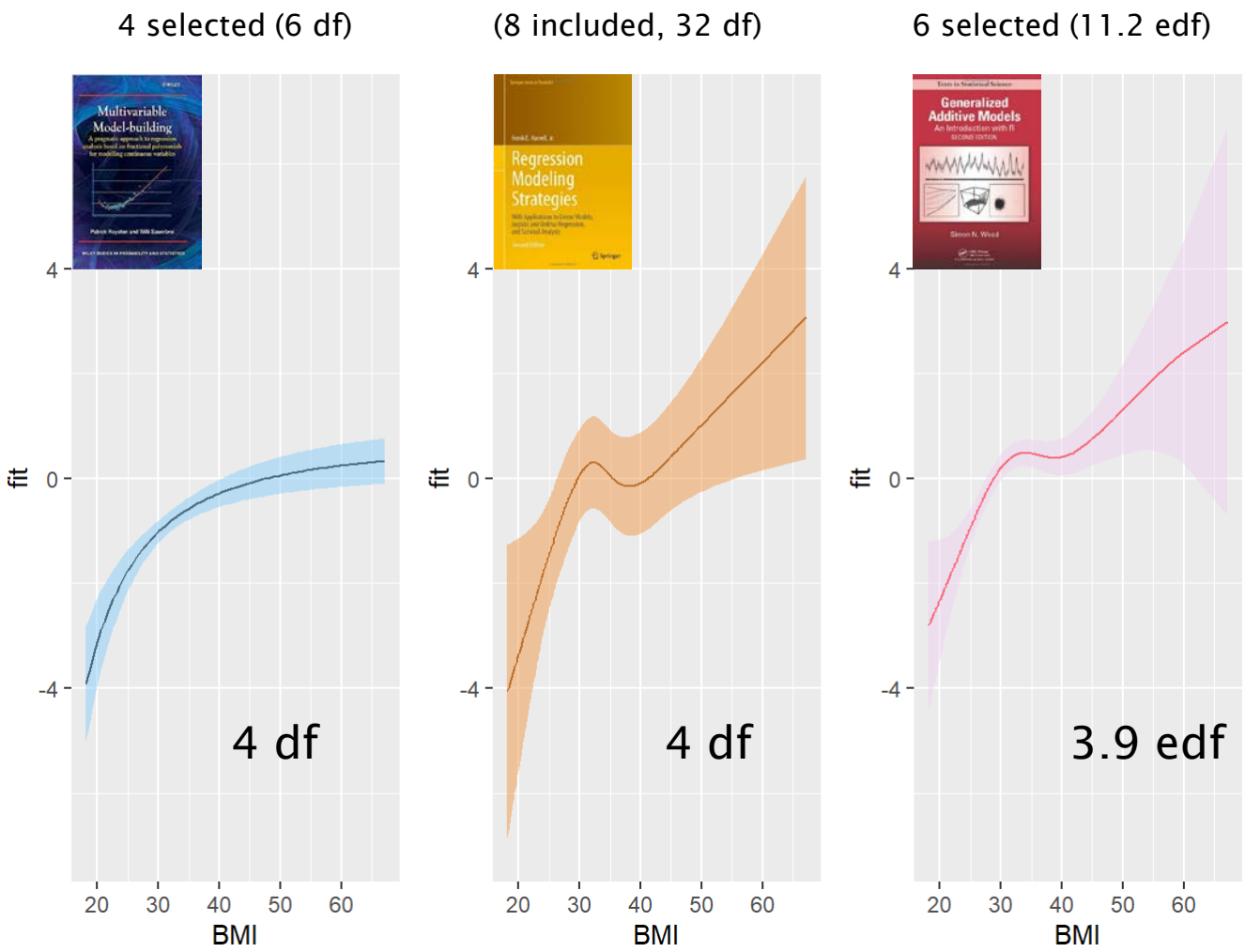
Harrell, 2015



Wood, 2017

Comparison in Pima data set:

- Predicting diabetes onset (yes/no) in 768 members of Pima nation
- 8 cont. predictors
- Partial linear predictors for BMI:



Some ongoing TG2 activities

- Performance measures for evaluating nonlinear effects:

A Systematic Categorization of Performance Measures for Estimated Non-Linear Associations Between an Outcome and Continuous Predictors

Theresa Ullmann[1] | Georg Heinze[1] | Michal Abrahamowicz[2] | Aris Perperoglou[3] | Willi Sauerbrei[4] | Matthias Schmid[5] | Daniela Dunkler[1] | TG2 of the STRATOS Initiative

Wiley Interdisciplinary Reviews: Computational Statistics, 2025; 17:e70042
https://doi.org/10.1002/wics.70042

- Simulation studies:
  - Prince T, Kappenberg F, Dunkler D, Sauerbrei W, Heinze G, Schmid M:
    A comparison of variable selection approaches for spline regression (multivariable)
    Lead: Group of Matthias Schmid, Bonn, D
  - Ullmann T, Dunkler D, Heinze G (Vienna):
    A comparison of spline estimators for nonlinear associations (univariable)

References

- Akbari N, Grittner U, Heinze G, Dunkler D. (2026) “A robust zero-corrected variance estimator to improve full-model inference after variable selection”. Submitted.
- Carlin, J.B. and Moreno-Betancur, M. (2025) “On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching,” *Statistics in Medicine*, 44(13–14), p. e10244. Available at: https://doi.org/10.1002/sim.10244.
- Gregorich, M. *et al.* (2021) “Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution,” *International Journal of Environmental Research and Public Health*, 18(8), p. 4259. Available at: https://doi.org/10.3390/ijerph18084259.
- Harrell , F.E. (2015) *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Cham: Springer International Publishing (Springer Series in Statistics). Available at: https://doi.org/10.1007/978-3-319-19425-7.
- Hastie, T., Tibshirani, Robert and Tibshirani, Ryan (2020) “Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons,” *Statistical Science*, 35(4). Available at: https://doi.org/10.1214/19-STS733.
- Heinze, G. *et al.* (2024) “Regression without regrets –initial data analysis is a prerequisite for multivariable regression,” *BMC Medical Research Methodology*, 24(1), p. 178. Available at: https://doi.org/10.1186/s12874-024-02294-3.
- Heinze, G. and Dunkler, D. (2017) “Five myths about variable selection,” *Transplant International*, 30(1), pp. 6–10. Available at: https://doi.org/10.1111/tri.12895.
- Heinze, G., Wallisch, C. and Dunkler, D. (2018) “Variable selection – A review and recommendations for the practicing statistician,” *Biometrical Journal*, 60(3), pp. 431–449. Available at: https://doi.org/10.1002/bimj.201700067.
- Kovács, L. (2024) “Feature selection algorithms in generalized additive models under concurvity,” *Computational Statistics*, 39(2), pp. 461–493. Available at: https://doi.org/10.1007/s00180-022-01292-7.
- Leeb, H. and Pötscher, B.M. (2005) “Model Selection and Inference: Facts and Fiction,” *Econometric Theory*, 21(1), pp. 21–59. Available at: https://www.jstor.org/stable/3533623 (Accessed: December 11, 2025).
- Leeb, H. and Pötscher, B.M. (2008) “Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?,” *Econometric Theory*, 24(2), pp. 338–376. Available at: https://www.jstor.org/stable/20142496 (Accessed: December 11, 2025).
- Riley, R.D. and Collins, G.S. (2023) “Stability of clinical prediction models developed using statistical or machine learning methods,” *Biometrical Journal*, 65(8), p. 2200302. Available at: https://doi.org/10.1002/bimj.202200302.
- Royston, P. and Sauerbrei, W. (2007) “Multivariable Modeling with Cubic Regression Splines: A Principled Approach,” *The Stata Journal: Promoting communications on statistics and Stata*, 7(1), pp. 45–70. Available at: https://doi.org/10.1177/1536867X0700700103.

References (cont'd)

- Royston, P. and Sauerbrei, W. (2008) *Multivariable Model-Building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. 1st ed. Wiley (Wiley Series in Probability and Statistics). Available at: https://doi.org/10.1002/9780470770771.
- Sauerbrei, W., Royston, P. and Binder, H. (2007) "Selection of important variables and determination of functional form for continuous predictors in multivariable model building," *Statistics in Medicine*, 26(30), pp. 5512–5528. Available at: https://doi.org/10.1002/sim.3148.
- Shmueli, G. (2010) "To Explain or to Predict?," *Statistical Science*, 25(3). Available at: https://doi.org/10.1214/10-STS330.
- Shmueli, G. (2025) "To Explain, to Predict, or to Describe: Figuring out the Study Goal [Commentary on 'On the Uses and Abuses of Regression Models' by Carlin and Moreno-Betancur]," *Statistics in Medicine*, 44(13–14), p. e10307. Available at: https://doi.org/10.1002/sim.10307.
- Steyerberg, E.W. *et al.* (2018) "Poor performance of clinical prediction models: the harm of commonly applied methods," *Journal of Clinical Epidemiology*, 98, pp. 133–143. Available at: https://doi.org/10.1016/j.jclinepi.2017.11.013.
- Sun, G.-W., Shook, T.L. and Kay, G.L. (1996) "Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis," *Journal of Clinical Epidemiology*, 49(8), pp. 907–916. Available at: https://doi.org/10.1016/0895-4356(96)00025-X.
- Ullmann, T. *et al.* (2024) "Evaluating variable selection methods for multivariable regression models: A simulation study protocol," *PLOS ONE*. Edited by S. Tian, 19(8), p. e0308543. Available at: https://doi.org/10.1371/journal.pone.0308543.
- Ullmann, T. *et al.* (2025) "A Systematic Categorization of Performance Measures for Estimated Non-Linear Associations Between an Outcome and Continuous Predictors," *WIREs Computational Statistics*, 17(3), p. e70042. Available at: https://doi.org/10.1002/wics.70042.
- Ullmann, T., *et al.* (2026) "Towards evidence-based guidance on variable selection methods for multivariable regression models". Submitted.
- Wallisch, C. *et al.* (2021) "Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling," *Statistics in Medicine*, 40(2), pp. 369–381. Available at: https://doi.org/10.1002/sim.8779.
- Wood, S.N. (2017) *Generalized additive models: an introduction with R*. Second edition. Boca Raton, FL: CRC Press, Taylor & Francis Group (Chapman & Hall/CRC texts in statistical science series). Available at: https://doi.org/10.1201/9781315370279.
- Wynants, L. *et al.* (2020) "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, p. m1328. Available at: https://doi.org/10.1136/bmj.m1328.

https://stratostg2.github.io

https://www.stratos-initiative.org

Towards evidence-based guidance on variable selection methods for multivariable regression models

Theresa Ullmann [1], Georg Heinze [1], Michael Kammer [1,2], Daniela Dunkler [*1], for TG2 of the STRATOS initiative

[1]Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria

[2]Division of Nephrology and Dialysis, Department of Medicine III, Medical University of Vienna, Vienna, Austria

Submitted, 2026