Prognosis Research in Medicine – Pitfalls, Progress and Pathways to Excellence

Georg Heinze

Institute of Clinical Biometrics

Outline

- Prognosis research – aims and major challenges
- Pitfalls, and how to avoid them
- Some contributions
- Pathways to excellence

Purpose of models: To Explain or to Predict?

- **Descriptive models**
  - Interest in describing the data structure parsimoniously.
  - “Describe how outcome varies with predictors.“
- **Predictive models**
  - Interest in predicting outcome for future application.
  - “Predict how outcomes will be, given the predictors.”
- **Explanatory models**
  - Interest in inferring causal effects of interventions on outcome.
  - “Explain why outcomes differ depending on the intervention.”

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

(Shmueli, 2010)

- Similar considerations by Hernan et al, 2019; and Carlin and Moreno-Betancur, 2023

Prognosis research

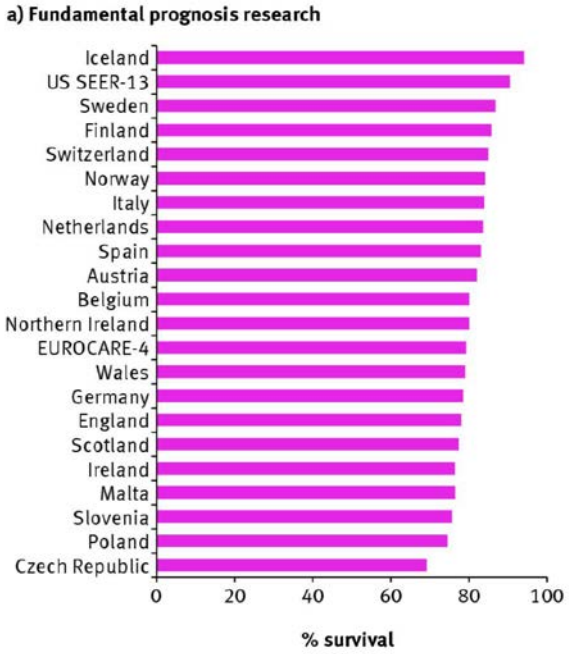
- In their PROGRESS series, Hemingway et al (2013) defined prognosis research as

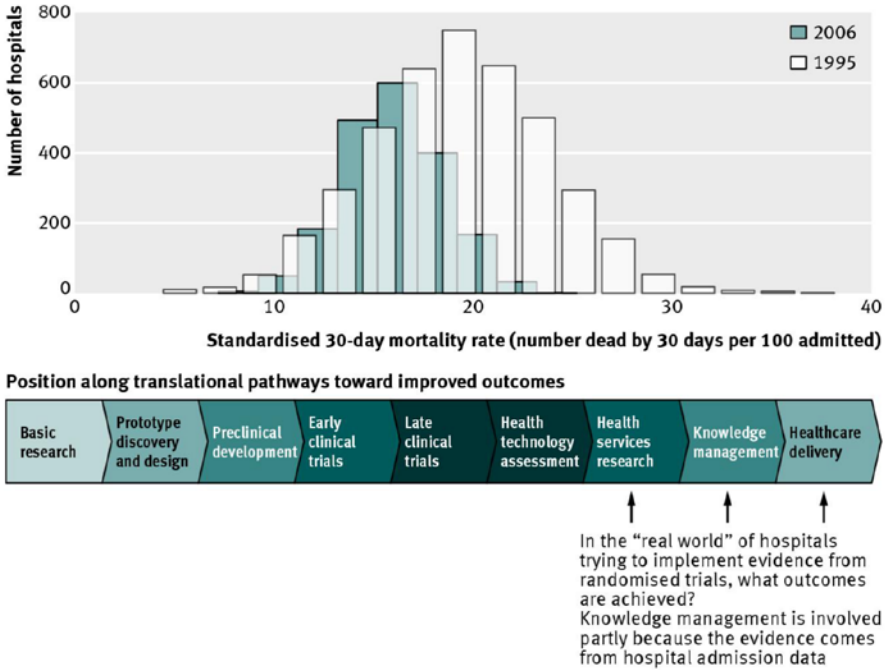
  „*… the investigation of the relations between future outcomes (endpoints) among people with a given baseline health state (startpoint) in order to improve health*"

- They distinguish the four interrelated research themes:
  - Fundamental (*descriptive*) prognosis research
  - Prognostic factor research
  - Prognostic model research
  - Stratified medicine research

Fundamental prognosis research

- According to Hemingway et al (2013), fundamental prognosis research refers to **describing outcomes and investigating variation in outcomes** across different groups → compare Shmueli (2010)'s notion of ‚descriptive models'





Prognostic factor research

Breast Cancer Research and Treatment **52**: 289-303, 1998.

Methodological challenges in the evaluation of prognostic factors in breast cancer

Douglas G. Altman[1] and Gary H. Lyman[2,3]
[1]*Imperial Cancer Research Fund Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK;* [2]*Medical Statistics Unit, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK;* [3]*H Lee Moffitt Cancer Center and Research Institute at the University of South Florida, Tampa, Florida, USA*

- Phases of prognostic factor research (Altman & Lyman, 1998):
  - Phase I: exploratory studies (hypothesis generating)
  - Phase II: exploratory studies that use a prognostic marker to
    - Discriminate between patients at high or low risk
    - Indicate which subsets likely benefit from therapy
  - Phase III: confirmatory studies of a-priori hypotheses to proof which markers…
    - Discriminate …
    - Indicate …
  - Develop a prognostic model combining many prognostic variables
    - Maximize the ability to predict outcomes for groups or individuals

REMARK guidelines!

Prognostic model research

- Key steps in model development:
- Literature research
  - Systematic reviews using PROBAST (upcoming: PROBAST+AI) tool
  - Identification of existing models with low risk of bias
  - Review of prognostic factor studies
  - Which prognostic factors have been used/not used?
- Validation of existing models
  - Assessment of discrimination in target population
  - Assessment of calibration (in-the-large, slope, local) in target population
- Updating of existing models (if necessary)
  - Recalibration
  - Reestimation
  - Adding predictors, dropping predictors
- Development of a totally new model (if necessary)

Prognostic factor research

TRIPOD+AI guidelines!

Systematic reviews

ELSEVIER Journal of Clinical Epidemiology 145 (2022) 126–135

REVIEW

Prediction models for living organ transplantation are poorly developed, reported, and validated: a systematic review

Maria C. Haller[a,b], Constantin Aschauer[c], Christine Wallisch[a], Karen Leffondré[d], Maarten van Smeden[e], Rainer Oberbauer[c], Georg Heinze[a,*]

[a] Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Section for Clinical Biometrics, Medical University of Vienna, Vienna, Austria

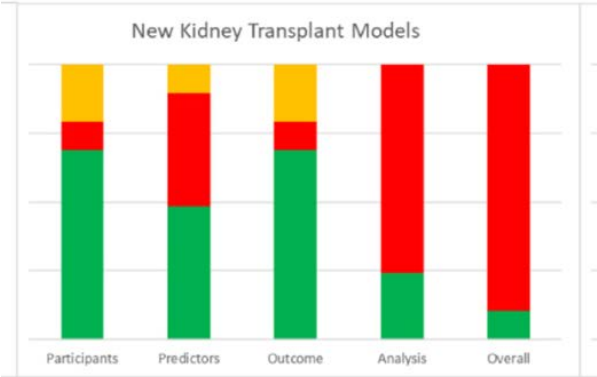
[b] Department for Internal Medicine III, Nephrology and Hypertension Diseases, Transplantation Medicine and Rheumatology, Ordensklinikum Linz, Linz, Austria

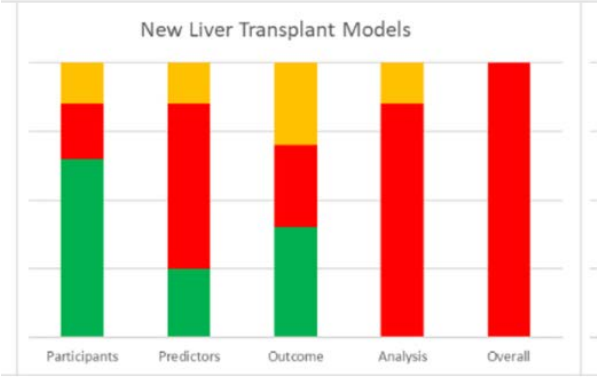
[c] Division of Nephrology and Dialysis, Department of Medicine III, Medical University of Vienna, Vienna, Austria

[d] University of Bordeaux, INSERM, Bordeaux Population Health Research Center, UMR1219, Bordeaux, France

[e] Julius Center for Health Science and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Accepted 31 January 2022; Available online 4 February 2022





Risk of bias: unclear high low

- Most frequent problems:
  - Participants: subjective eligibility criteria, posttransplant information,
  - Predictors: from the future,
  - Outcome: arbitrary definitions, too short horizon
  - Analysis: small sample size, mishandling of missing data, weak strategies for model building, inappropriate model performance evaluation

How to avoid pitfalls: consider PROBAST+AI

- 2015:

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies

Robert F. Wolff, MD*; Karel G.M. Moons, PhD*; Richard D. Riley, PhD; Penny F. Whiting, PhD; Marie Westwood, PhD; Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Jos Kleijnen, MD, PhD; and Sue Mallett, DPhil; for the PROBAST Group†

- 2021:

BMJ Open

Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence

Gary S Collins [1,2] Paula Dhiman [1,2] Constanza L Andaur Navarro [3] Jie Ma [1] Lotty Hooft [3,4] Johannes B Reitsma [3] Patricia Logullo [1,2] Andrew L Beam [5,6] Lily Peng [7] Ben Van Calster [8,9,10] Maarten van Smeden [3] Richard D Riley [11] Karel GM Moons [3,4]

2024: **PROBAST+AI: An updated quality, risk of bias and applicability assessment tool for prediction models using regression or artificial intelligence methods**

Karel G.M. Moons (0000-0003-2118-004X) [1*],
Johanna A.A. Damen (0000-0001-7401-4593)[1],
Tabea Kaul (0000-0002-4402-5379)[1],
Lotty Hooft (0000-0002-7950-2980)[1],
Constanza Andaur Navarro (0000-0002-7745-2887)[1],
Paula Dhiman (000-0002-0989-0623)[2],
Andrew L. Beam (0000-0002-6657-2787)[3],
Ben Van Calster (0000-0003-1613-7450)[4],
Leo Anthony Celi (0000-0001-6712-6626)[5],
Spiros Denaxas (0000-0001-9612-7791)[6],
Alastair K. Denniston (0000-0001-7849-0087)[7],
Marzyeh Ghassemi (0000-0001-6349-7251)[8],
Georg Heinze (0000-0003-1147-8491)[9],
André Pascal Kengne (0000-0002-5183-131X)[10],
Lena Maier-Hein (0000-0003-4910-9368)[11],
Xiaoxuan Liu (0000-0002-1286-0038)[7,12,19,20],
Patricia Logullo (0000-0001-8708-7003)[2],
Melissa D. McCradden (0000-0002-6476-2165)[13],
Nan Liu (0000-0003-3610-4883)[14],
Lauren Oakden-Rayner (0000-0001-5471-5202)[15],
Karandeep Singh (0000-0001-8980-2330)[16],
Daniel S. Ting (0000-0003-2264-7174)[14,17],
Laure Wynants (0000-0002-3037-122X)[18],
Bada Yang (0000-0002-9317-4995)[1],
Johannes B. Reitsma (0000-0003-4026-4345)[1],
Richard D. Riley (0000-0001-8699-0735)[19,20],
Gary S. Collins (0000-0002-2772-2316)[2],
Maarten van Smeden (0000-0002-5529-1541)[1]

Provisionally accepted

PROBAST+AI signalling questions

- **Participants** and **data sources**:

- Were appropriate data sources used?
  - How was data collected? How were measurements done? Fairness?
- Was an appropriate study design used?
  - Longitudinal cohort studies?
  - Selective sampling (case-control) with appropiate adjustments (calibration)?
  - Data quality?
- Did the in- and exclusions of study participants result in a representative data set?
  - Representative for target application?
  - No exclusion of ‚difficult' patients?
  - Handling of marginalized subgroups?

PROBAST+AI signalling questions

- **Predictors** domain:

- Were predictors defined in the same way for all participants?
- Was any pre-processing of predictors similar for all participants?
- Were predictor assessments made without knowledge of outcome data?
- Were the predictors included in the model available at the time the model was intended to be used?

PROBAST+AI signalling questions

- **Outcome** domain:

- Were outcomes defined and assessed appropriately?
- Were outcomes defined and assessed in a similar way for all participants?
- Were outcome assessments made without use or knowledge of predictor data?
- Was the time interval between predictor assessment and outcome assessment appropriate?

PROBAST+AI signalling questions

- **Analysis** domain:

- Was there evidence that the sample size was reasonable?
- Were continuous and categorical predictors handled appropriately?
- Were participants with missing or censored data handled appropriately in the analysis?
- If methods to address class imbalance were used, was the model or the model predictions recalibrated?
- Were methods used to address potential model overfitting?

PROBAST+AI signalling questions

- Additional questions for **performance evaluation**:

- Was model evaluation based on **only apparent performance avoided**?
- Were participants with **missing or censored data** handled appropriately in the analysis?
- If methods to address **class imbalance** were used, was the evaluation done in a dataset without imbalance correction?
- If data splitting was done to create **training and test datasets**, was there evidence that data leakage was avoided?
- If resampling methods were used to evaluate model performance, were **all model development steps replicated in the resampling** process?
- Was the predictive performance of the model evaluated appropriately, e.g., **calibration, discrimination, and net benefit**?

Prognostic model research: new model development

- Prognostic factor/model research: evidence available?
  - Which predictors to consider?
- Data set(s) available?
  - Sample size for development
  - Multicenter collaboration: cross-validation?
  - Quality of data? Prospectively collected/retrospective?
- Research protocol and Statistical Analysis Plan
  - Participants – Predictors – Outcome - Analysis
  - Data cleaning and data screening (IDA)
  - Predictor specification
  - Outcome specification
  - Model specification and model selection
  - Model diagnostics and model performance
  - Describing the model

TRIPOD+AI guidelines!

Reporting of prediction models: TRIPOD+AI

RESEARCH METHODS AND REPORTING

TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods

Gary S Collins,[1] Karel G M Moons,[2] Paula Dhiman,[1] Richard D Riley,[3,4] Andrew L Beam,[5] Ben Van Calster,[6,7] Marzyeh Ghassemi,[8] Xiaoxuan Liu,[9,10] Johannes B Reitsma,[2] Maarten van Smeden,[2] Anne-Laure Boulesteix,[11] Jennifer Catherine Camaradou,[12,13] Leo Anthony Celi,[14,15,16] Spiros Denaxas,[17,18] Alastair K Denniston,[4,9] Ben Glocker,[19] Robert M Golub,[20] Hugh Harvey,[21] Georg Heinze,[22] Michael M Hoffman,[23,24,25,26] André Pascal Kengne,[27] Emily Lam,[12] Naomi Lee,[28] Elizabeth W Loder,[29,30] Lena Maier-Hein,[31] Bilal A Mateen,[17,32,33] Melissa D McCradden,[34,35] Lauren Oakden-Rayner,[36] Johan Ordish,[37] Richard Parnell,[12] Sherri Rose,[36] Karandeep Singh,[38] Laure Wynants,[40] Patricia Logullo[1]

For numbered affiliations see end of the article

Correspondence to: G S Collins gary.collins@csm.ox.ac.uk (or @GSCollins on Twitter; ORCID 0000-0002-2772-2316)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2024;385:e078378 http://dx.doi.org/10.1136/bmj-2023-078378

Accepted: 17 January 2024

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement was published in 2015 to provide the minimum reporting recommendations for studies developing or evaluating the performance of a prediction model. Methodological advances in the field of prediction have since included the widespread use of artificial intelligence (AI) powered by machine learning methods to develop prediction models. An update to the TRIPOD statement is thus needed. TRIPOD+AI provides harmonised guidance for reporting prediction model studies, irrespective of whether regression modelling or machine learning methods have been used. The new checklist supersedes the TRIPOD 2015 checklist, which should no longer be used. This article describes the development of TRIPOD+AI and presents the expanded 27 item checklist with more detailed explanation of each reporting recommendation, and the TRIPOD+AI for Abstracts checklist. TRIPOD+AI aims to promote the complete, accurate, and transparent reporting of studies that develop a prediction model or evaluate its performance. Complete reporting will facilitate study appraisal, model evaluation, and model implementation.

Some of our own contributions

- Initial data analysis (Heinze et al, 2024)
- Correlated predictors (Gregorich et al, 2021)
- Prespecification of predictors by background knowledge (Hafermann et al, 2021, 2022)
- Data-driven selection (Heinze et al, 2018; Ullmann et al, 2024)
- Non-linear functional forms (Sauerbrei et al, 2020)
- Missing data imputation (Deforth et al, 2024)
- Regularization: to tune or not to tune (Sinkovec et al, 2021)
- Model ~~explanation~~ description (Wallisch et al, 2021)
- Putting research into context: Phases of methodological research (Heinze et al, 2024)

Initial data analysis

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

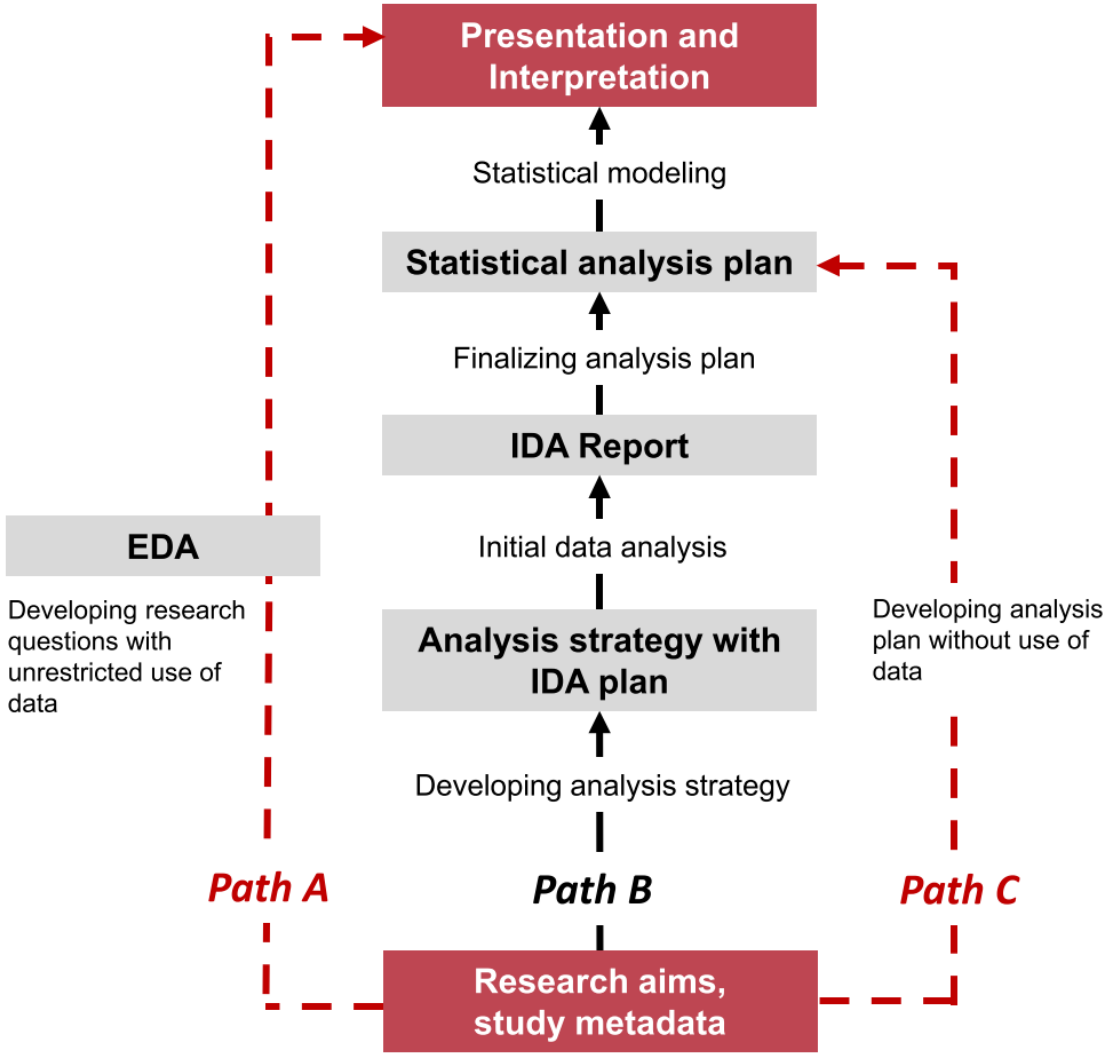
BMC Medical Research Methodology

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

- Provided a checklist of items to be addressed at initial data analysis for prediction or descriptive modeling task
- Main domains: missing data, univariate distributions, multivariate analyses (without outcome!)
- Golden rule of IDA :
  „Do not assess predictor-outcome association!“ (similar to blinding in RCTs)



Follow-up project: SAPI

- SAPI – statistical analysis plan with initial data analysis (IDA) plan
- Lead: Marianne Huebner, Carsten Oliver Schmidt, Lara Lusa, Georg Heinze, Willi Sauerbrei, Gary Collins

- Step 1: Write SAPI version 1
  Written without detailed knowledge of data, includes specification of IDA
- Step 2: Perform Initial data analysis according to SAPI v1, evaluate IDA results and:
- Step 3: Write SAPI version 2
  Update/refine SAPI v1 because of IDA results

- Golden rule of IDA:
  „Do not assess predictor-outcome association!“ (similar to blinding in RCTs)

Correlated predictors

International Journal of
Environmental Research and Public Health

2021

Article

Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution

Mariella Gregorich [1], Susanne Strohmaier [1,2], Daniela Dunkler [1] and Georg Heinze [1,*]

The symptoms:

- Highly variable regression coefficients
- Large standard errors
- Numerical instability

- 56 citations to date

Table 3. Some options to deal with collinearity by research aim. With 'symptoms', we mean typical consequences of collinearity such as inflated standard errors and unstable parameter estimates.

Method	Explanation	Remark
Descriptive research aim		
Variable omission	Omit one of the variables involved in the collinearity	Removes the symptoms, but leads to different interpretation of the model
Summary score	Combine several nearly collinear variables into a summary score and include only the summary score in the regression model	Removes the symptoms, retains most of the predictive value of the model, but leads to different interpretation of the model
Predictive research aim		
Use information criteria	Information criteria such as Akaike's can be used to guide model building	Information criteria guide the analyst in a search for the most predictive model
Explanatory research aim		
Use causal reasoning	Specification of variables (exposure of interest, confounders) is necessitated by causal reasoning	Neither exposure nor confounders should be omitted as this violates assumptions needed to identify the causal estimand of interest

Predictor selection: where does all the background knowledge come from?

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

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

entropy

Article

Using Background Knowledge from Preceding Studies for Building a Random Forest Prediction Model: A Plasmode Simulation Study

Lorena Hafermann [1], Nadja Klein [2,*], Geraldine Rauch [1], Michael Kammer [3] and Georg Heinze [3,*]

- „Background knowledge" may result from inappropriate methods
- How relevant is background knowledge
  - Depending on sample size
  - Depending on predictability

Variable selection

DOI: 10.1002/bimj.201700067

REVIEW ARTICLE

Biometrical Journal

Variable selection – A review and recommendations for the practicing statistician

Georg Heinze | Christine Wallisch | Daniela Dunkler

- 956 citations to date ☺

PLOS ONE

STUDY PROTOCOL

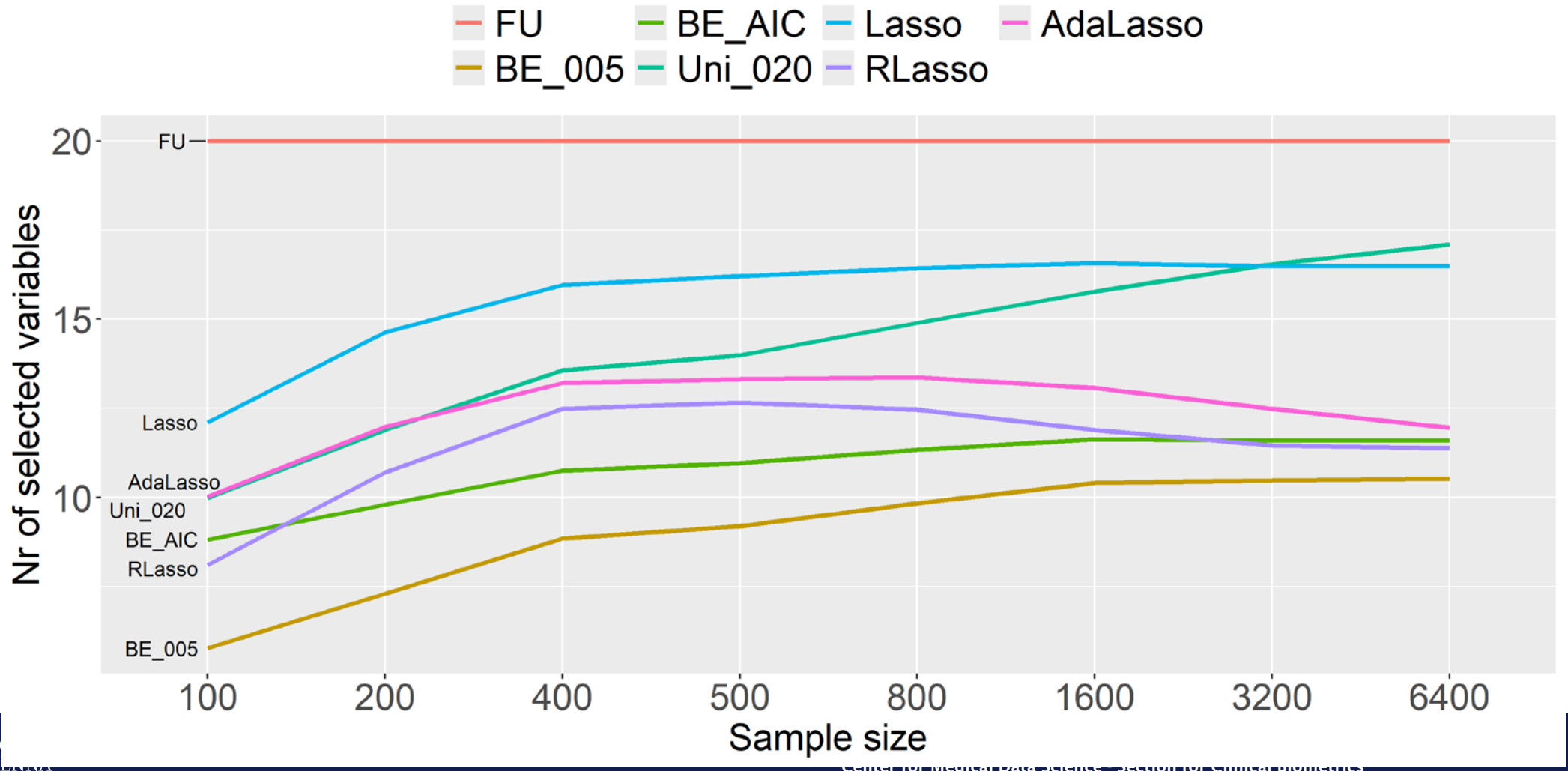
Evaluating variable selection methods for multivariable regression models: A simulation study protocol

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

- Protocol for a simulation study
  Results were recently presented at IBC, Atlanta

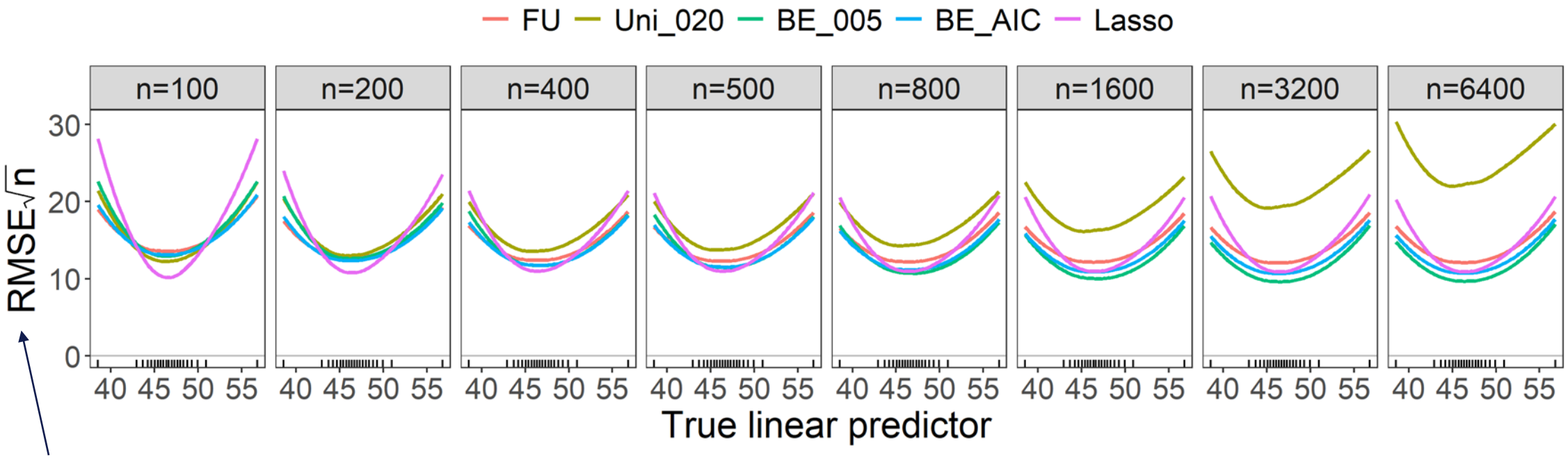
Results (1): main scenario, model size

Main scenario: Model size (nr of selected variables).
FU, Full model; BE_005, Backward elimination with alpha = 0.05; BE_AIC, Backward elimination with AIC; Uni_020, Univariable selection with alpha = 0.20; Lasso, Least angle selection and shrinkage operator with cross-validation of penalty; RLasso, relaxed Lasso - OLS fit with variables selected by Lasso, lambda tuned with cross-validation; AdaLasso, adaptive Lasso.



Results (2): main scenario, local prediction error

Main scenario: Local root mean squared error w.r.t. estimated vs. true linear predictor, multiplied with square root of sample size, averaged over simulations and smoothed with a LOESS smoother. FU, Full model; BE_005, Backward elimination with alpha = 0.05; BE_AIC, Backward elimination with AIC; Uni_020, Univariable selection with alpha = 0.20; Lasso, Least angle selection and shrinkage operator with cross-validation of penalty.



$$\sqrt{\frac{n}{n_{sim}}\sum_{i=1}^{n_{sim}}\left(x\hat{\beta}^{(i)} - x\beta\right)^2},$$

with x = observation vector in test set

- Lasso: larger prediction errors towards the boundaries
- Starting from $n = 1600$, BE_005 dominates the other methods.

Predictor selection - overall conclusions

- Performance of variable selection methods depends on sample size and R^2:
  worse performance for smaller sample sizes and lower R^2
- No ‘one-size-fits-all’ method:
  ranking of methods depends on performance measure
- Do not use univariable selection, neither on its own nor in combination with backward elimination
- A ‘true’ data generating mechanism is hardly ever identified
  (exception: large sample size and high R^2)
  - ➢ We should not ‘believe’ in a model that was found by variable selection
  - ➢ The selected model is just an ‘example model’ out of many

Continuous predictors

- How to include continuous predictors?

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

Procedures for simultaneous variable and functional form selection (1)

- MFP (Multivariable fractional polynomials) is an algorithm that combines variable selection with functional form selection.
- It uses stepwise (backward/forward) selection and at each steps reevaluates functional form selection.
- Parameters:
  - Selection criterion (AIC/BIC/significance level)
  - Significance level for functional form selection
  - Complexity of FP (1, 2, 3, …)
  - Variables ‘safe’ to be included (no matter which p-value)
- Described in Royston and Sauerbrei, 2008
- Implementation: R package mfp2 (available on CRAN)

Procedures for simultaneous variable and functional form selection (2)

- Although in principle possible, there is no widely accepted other algorithm for simultaneous VS&FF selection
- MFP principle can be used with splines: multivariable regression splines (MVRS) procedure (Royston and Sauerbrei, 2007)

- rms package: fit restricted cubic splines for continuous variables (default: 4df)
  - Remove only ‘very insignificant’ variables (Harrell, 2015)

Example: CRASH-2

Research

Predicting early death in patients with traumatic bleeding: development and validation of prognostic model

BMJ 2012 ; 345 doi: https://doi.org/10.1136/bmj.e5166 (Published 15 August 2012)
Cite this as: *BMJ* 2012;345:e5166

Article | Related content | Metrics | Responses | Peer review

Pablo Perel, senior clinical lecturer [1], David Prieto-Merino, lecturer, medical statistics [2], Haleema Shakur, senior lecturer [1], Tim Clayton, senior lecturer, medical [2], Fiona Lecky, clinical professor [3], honorary professor [4], honorary consultant [5], Omar Bouamra, medical statistician [6], Rob Russell, senior lecturer [7], Mark Faulkner, paramedic advisor [8], Ewout W Steyerberg, professor [9], Ian Roberts, professor [1]

https://biostat.org/data

CRASH-2					
crash2.html	crash2.rda	crash2.dta	NA	NA	Ccrash2.html

Training: N=15,000

Validation: N=4,127

Predictors:

- Age
- Sex
- Glasgow coma scale (1-15)
- Systolic blood pressure
- Heart rate
- Respiratory rate
- Capillary refill time
- Type of injury (3 types)
- Time since injury

Example: CRASH-2

MFP

Selection criterion: AIC

Complexity: max. 4 DF (FP2)

mfp2::mfp2()

```
i Initial degrees of freedom:
   age gcs sbp sexmale hr cc rr injurytime injurytype1 injurytype2
df   4   4   4       1  4  4  4          4           1           1

i Visiting order: gcs, age, rr, sbp, cc, injurytime, hr, injurytype1, injurytype2, sexmale

--------------------
i Running MFP Cycle 1
--------------------

Variable: gcs (keep = FALSE)
                  Powers    DF    AIC
null              NA        10    11735.7
linear            1         11    9636.1
FP1               0.5       12    9622.7
FP2               -2, -0.5  14    9615.0
Selected: FP2

Variable: age (keep = FALSE)
                  Powers    DF    AIC
null              NA        11    9785.8
linear            1         12    9611.0
FP1               3         13    9599.5
FP2               -2, 3     15    9599.0
Selected: FP2

Variable: rr (keep = FALSE)
                  Powers    DF    AIC
null              NA        12    9727.9
linear            1         13    9595.0
FP1               2         14    9598.5
FP2               0, 0.5    16    9578.5
Selected: FP2
```

RMS

Selection criterion: p>0.5

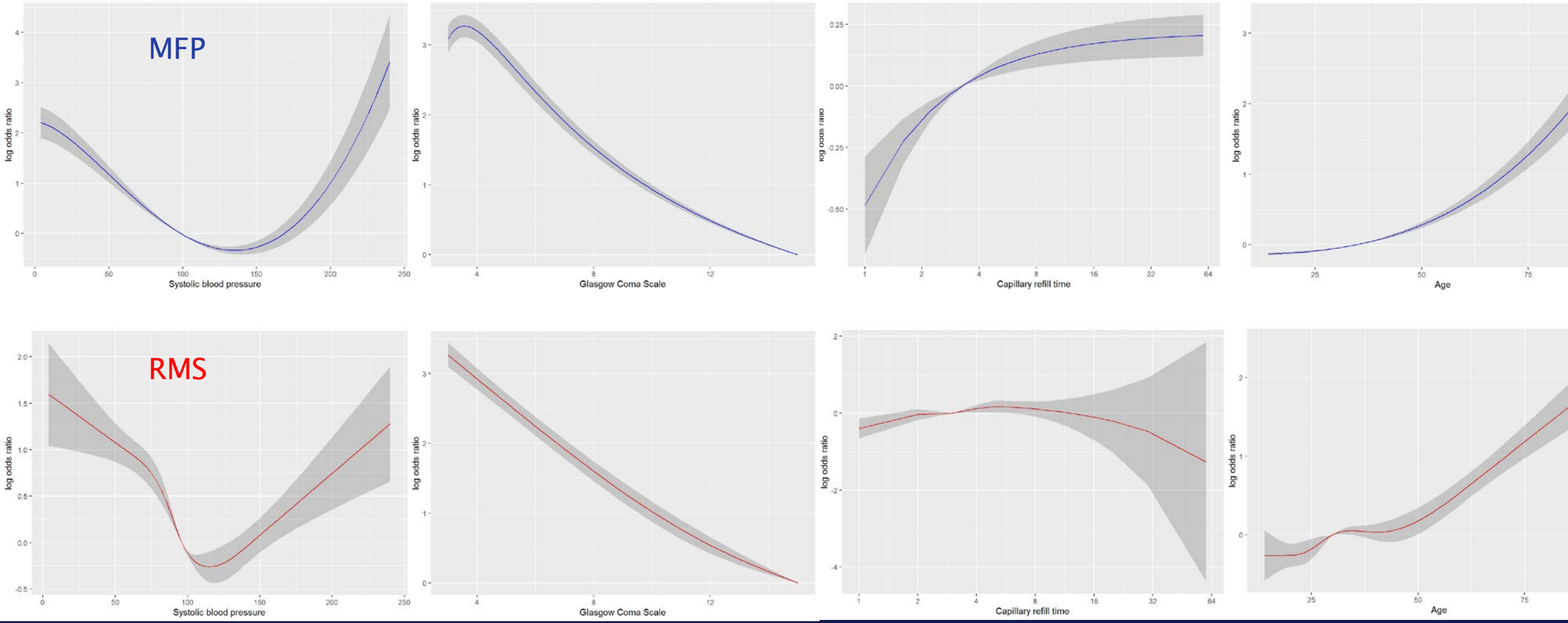
Complexity: RCS with 4DF

rms::lrm()

```
                Wald Statistics          Response: earlydeath

Factor          Chi-Square d.f. P
age              163.18     4   <.0001
 Nonlinear        18.34     3   0.0004
gcs             1444.75     2   <.0001
 Nonlinear        19.98     1   <.0001
sbp              202.92     4   <.0001
 Nonlinear       123.64     3   <.0001
sex                0.00     1   0.9986
hr                18.46     4   0.0010
 Nonlinear        15.92     3   0.0012
cc                20.20     4   0.0005
 Nonlinear        12.92     3   0.0048
rr               146.97     4   <.0001
 Nonlinear        25.03     3   <.0001
injurytime        21.20     4   0.0003
 Nonlinear         9.83     3   0.0201
injurytype        12.12     2   0.0023
TOTAL NONLINEAR  245.02    19   <.0001
TOTAL           2346.82    29   <.0001
```

CRASH-2: (Selected) modeling results



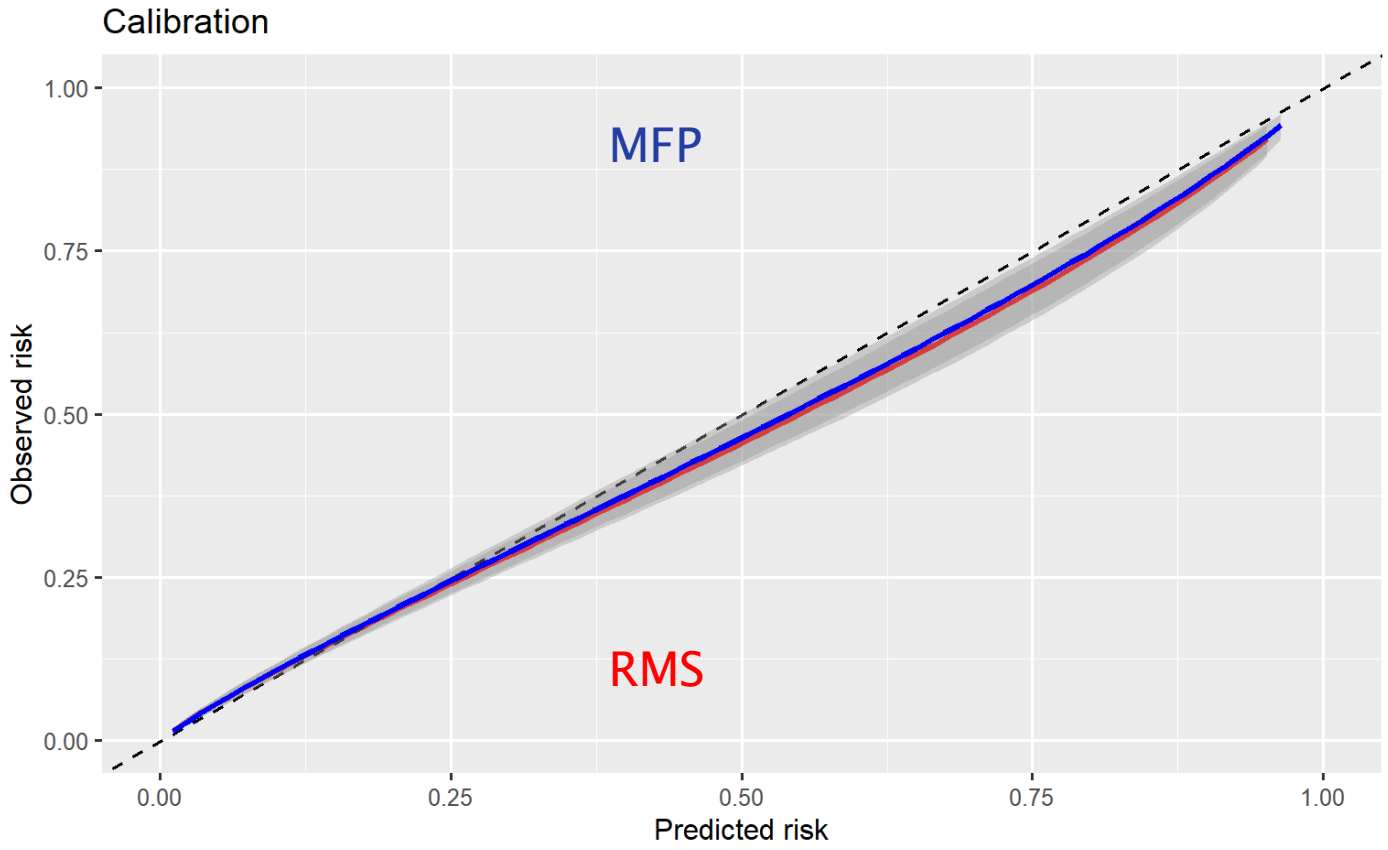
CRASH-2: Results of validation

MFP

Measure	Value
AUROC	0.8191
Brier	**0.0971**
ICI	**0.0112**

RMS

Measure	Value
AUROC	**0.8235**
Brier	0.0973
ICI	0.0123



Missing data imputation

ELSEVIER

Check for updates

Journal of Clinical Epidemiology 176 (2024) 111539

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

The performance of prognostic models depended on the choice of missing value imputation algorithm: a simulation study

Manja Deforth[a], Georg Heinze[b], Ulrike Held[a,*]

[a]*Department of Biostatistics at the Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland*

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

Accepted 16 September 2024; Published online 24 September 2024

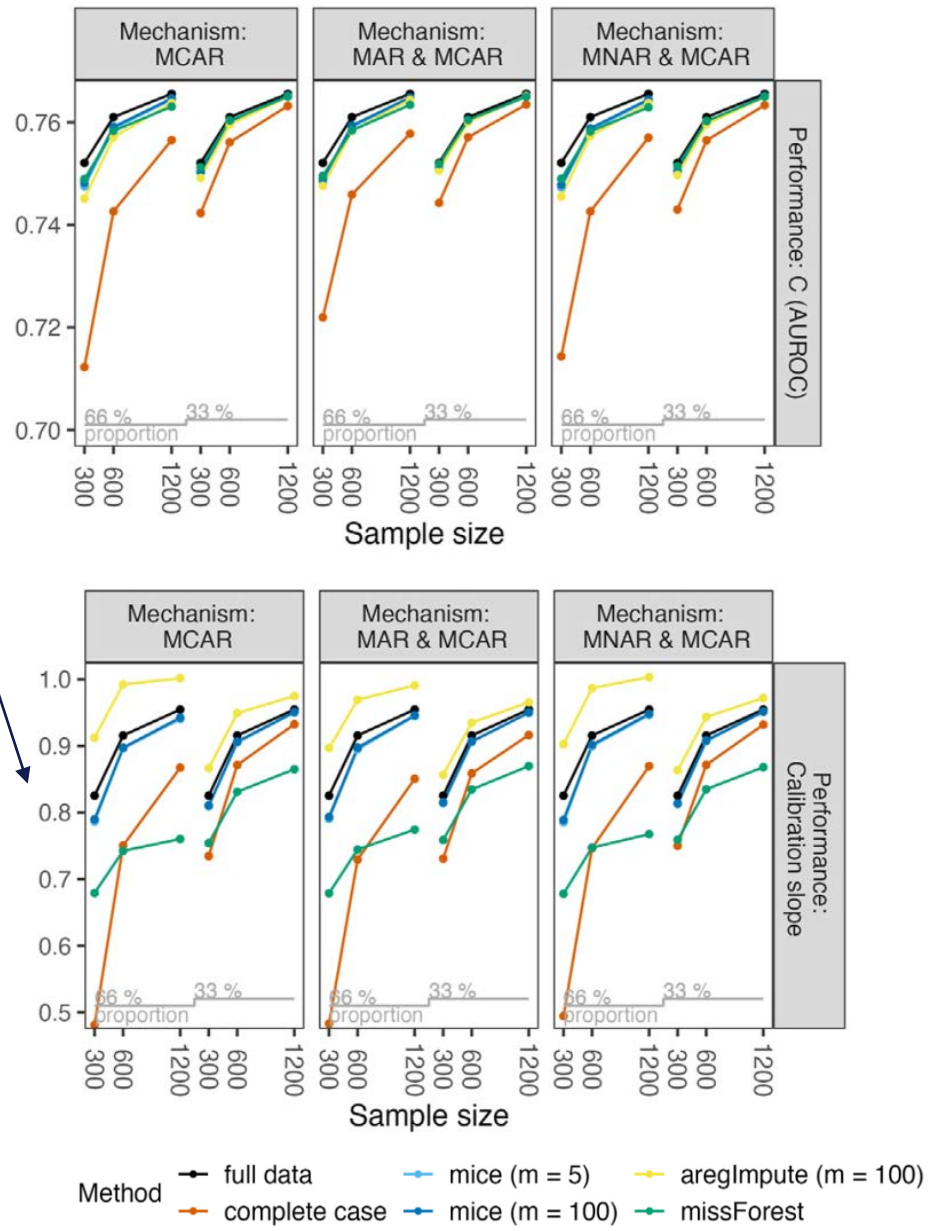
- Investigated three different imputation methods in model development
- Nonlinear associations between variables and nonlinear functional forms in outcome model (resembling real long-Covid study)

Table 3. Comparison of the three imputation methods

Imputation method	mice	aregImpute	missForest
Implementation (R package::function)	mice::futuremice	Hmisc::aregImpute	missForest::missForest
Imputation models	Linear-additive models	Flexible additive models with restricted cubic spline transformations (4 knots)	Random forest with max. 100 trees, splits based on three randomly selected variables
Number of burn-in iterations per imputation chain	4	3	max. 9
Total length of imputation chains	5	103	max. 10
Number of chains	5, 100	1	1
Data basis of imputation models	Original sample with iterated imputations	Bootstrap resamples from original sample with iterated imputations	Bootstrap resamples from original sample with iterated imputations
Number of imputations *m*	5, 100	100	1
Imputed values	Predictive mean matching	Predictive mean matching based on a bootstrap approximation of the full Bayesian predictive distribution	Predictions from random forest

Missing data imputation: results

- Overall, among the imputation methods
  - missForest was slightly superior for AUROC
  - aregImpute performed best in terms of calibration
- Surprisingly, calibration of models after aregImpute were superior even to full data analysis (before amputing data)
- This could be explained by the combination of:
  - Correctly specified imputation models (nonlinearities!) → lead to unbiased imputations
  - Only random noise in the imputations → amputation/imputation acts just like shrinkage factor
  - The shrinkage improves the calibration slopes



Talking about shrinkage: To tune or not to tune?

Šinkovec *et al. BMC Medical Research Methodology* (2021) 21:199
https://doi.org/10.1186/s12874-021-01374-y

BMC Medical Research Methodology

RESEARCH Open Access

To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets

Hana Šinkovec[1], Georg Heinze[1], Rok Blagus[2] and Angelika Geroldinger[1*]

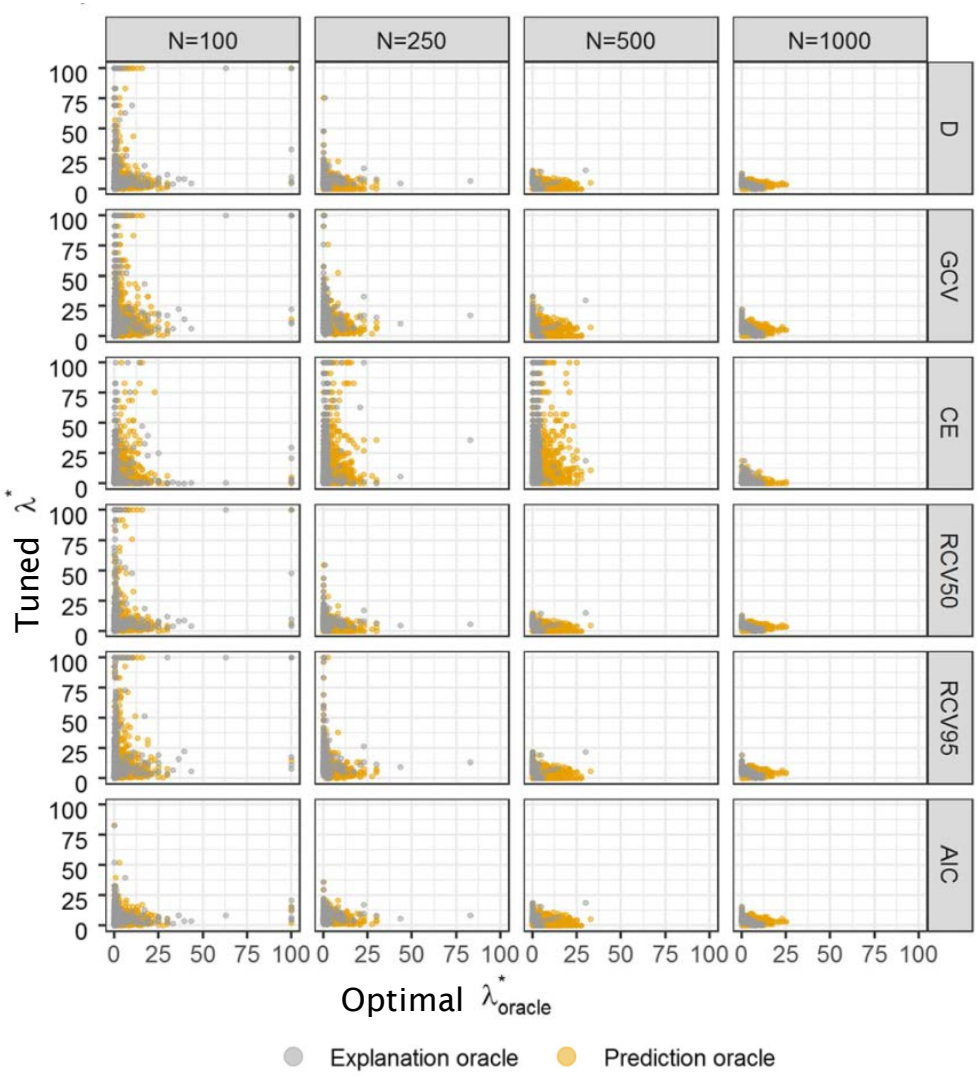
- We investigated logistic ridge regression with tuned and fixed penalty
- Tuned penalty: different methods
- Fixed penalty according to width of prior interval for regression coefficients („weak“, „strong“)

Results: to tune or not to tune?

- The tuned and optimal penalty strength were negatively correlated:

- Need strong penalty but tuned penalty is weak
- Need weak penalty but tuned penalty is strong

- → The costs of tuning hyperparameters is often neglected, but can be significant!



Model description

Wallisch *et al.*
BMC Medical Research Methodology (2021) 21:284
https://doi.org/10.1186/s12874-021-01487-4

BMC Medical Research Methodology

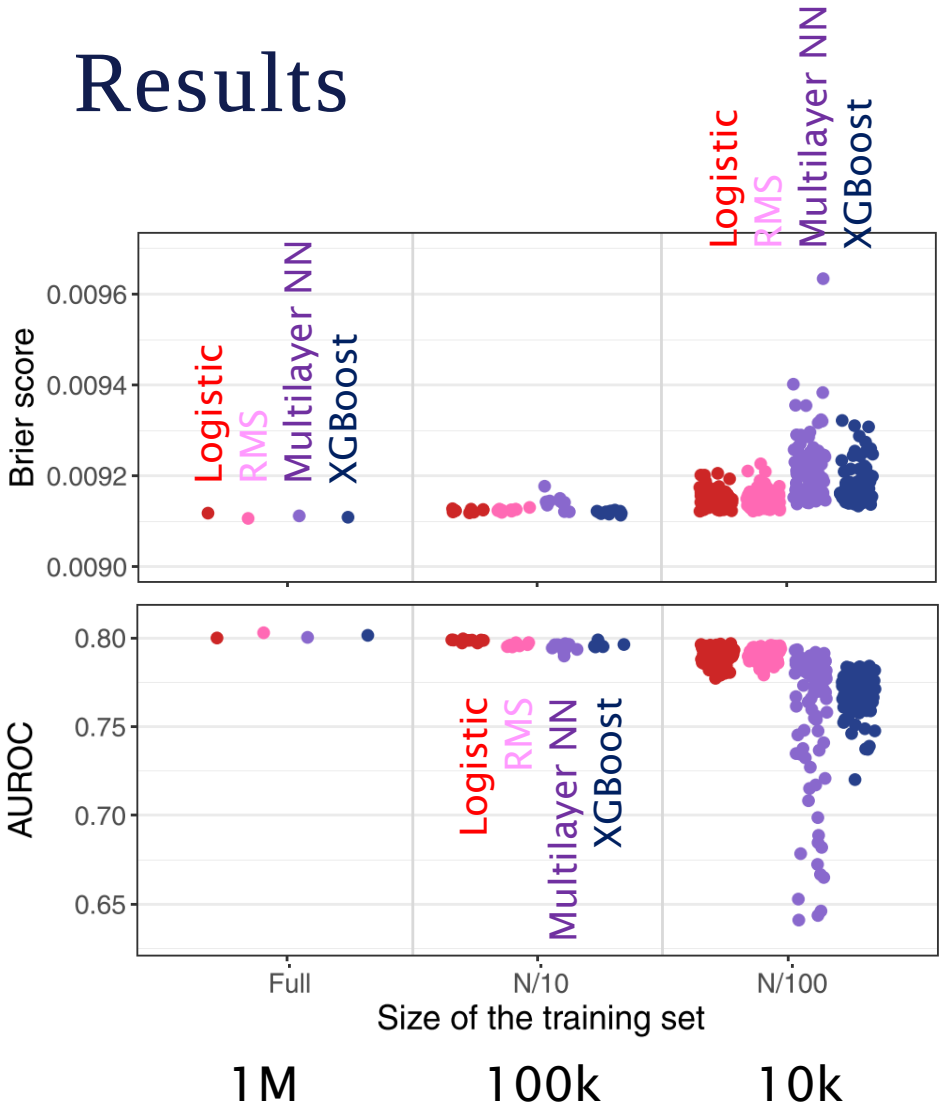
RESEARCH Open Access

The roles of predictors in cardiovascular risk models - a question of modeling culture?

Christine Wallisch[1], Asan Agibetov[2], Daniela Dunkler[1], Maria Haller[1,3], Matthias Samwald[2], Georg Dorffner[2] and Georg Heinze[1*]

- „Model explanation“ (description): How do predictions vary with the values of a predictor?
- We compared partial dependence plots and individual conditional expectation (ICE) plots obtained
  - In cardiovascular risk prediction
  - In a large development data set (1M), validation set = 500k, event rate = 1%, ~20 predictors
  - Comparing
    Logistic regression (linear-additive), RMS strategy with splines and pre-specified penalties for higher terms, Multilayer Neural Network, XGBoost

Results



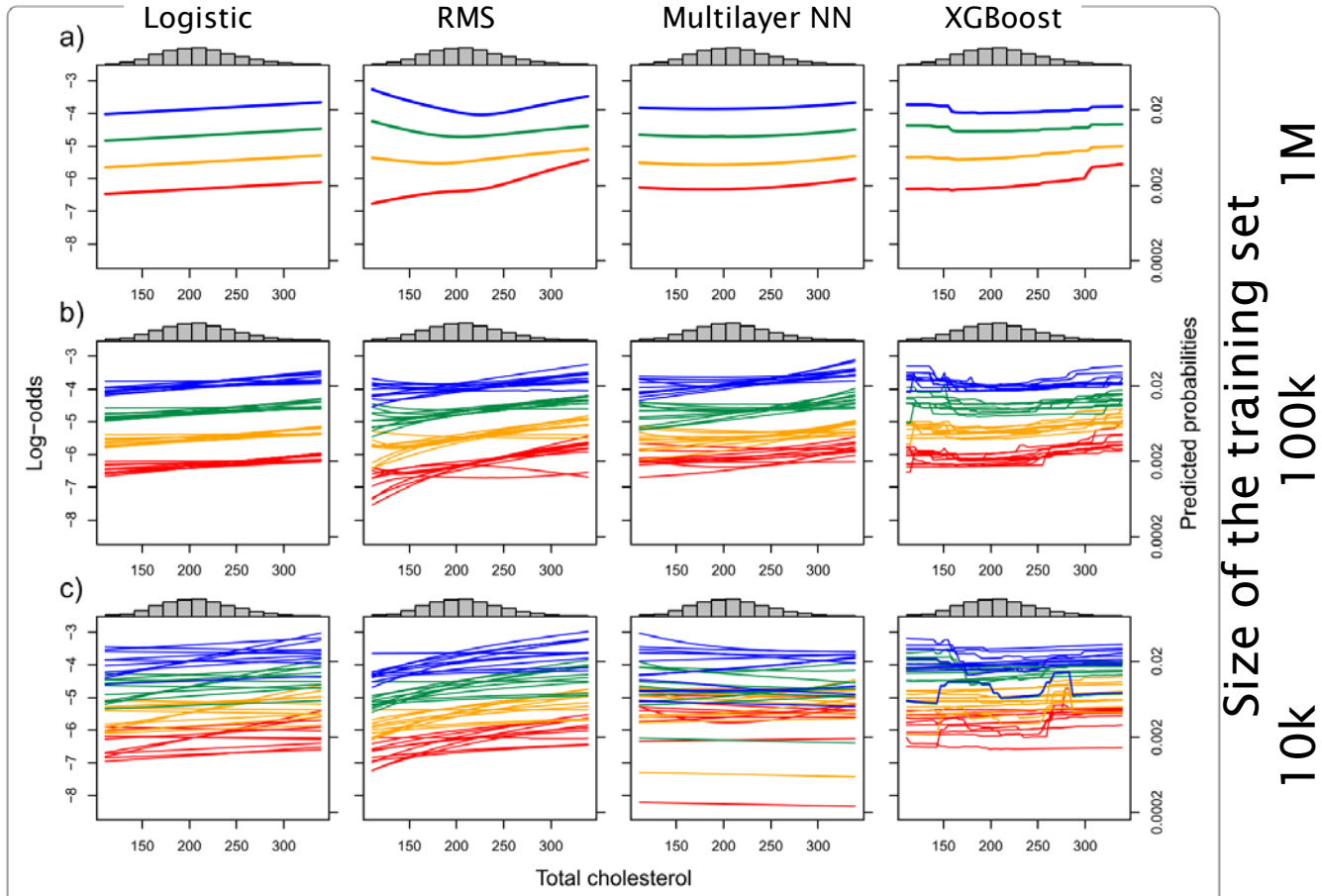


Fig. 5 Partial dependence of estimated risk on total cholesterol, showing how average predictions vary with total cholesterol while keeping all other predictors fixed. Red: age fixed at 40 years and sex set to female; yellow: 50 years, female; green: 60 years, female; blue: 70 years, female. The models (SLNN-LR, GAM, MLNN, and XGBoost) were fitted at **a** full data availability **b** data availability of 1/10 and **c** data availability of 1/100. In **c** 10 out of 100 models were randomly selected

Stratified medicine research

- Hemingway et al (2013): ‚*The use of prognostic information to tailor treatment decisions to an individual or a group of individuals with similar characteristics*‘

- Example:

JAMA Network | Open

Original Investigation | Nephrology

Survival Benefit of First Single-Organ Deceased Donor Kidney Transplantation Compared With Long-term Dialysis Across Ages in Transplant-Eligible Patients With Kidney Failure

Susanne Strohmaier, PhD; Christine Wallisch, PhD; Michael Kammer, PhD; Angelika Geroldinger, PhD; Georg Heinze, PhD; Rainer Oberbauer, MD, MSc; Maria C. Haller, MD, MSc

- Target trial emulation: each trial compared transplanted to those still-on-waiting list

Survival benefit example

Figure 2. Restricted Mean Survival Times for All-Cause Mortality and Differences Thereof

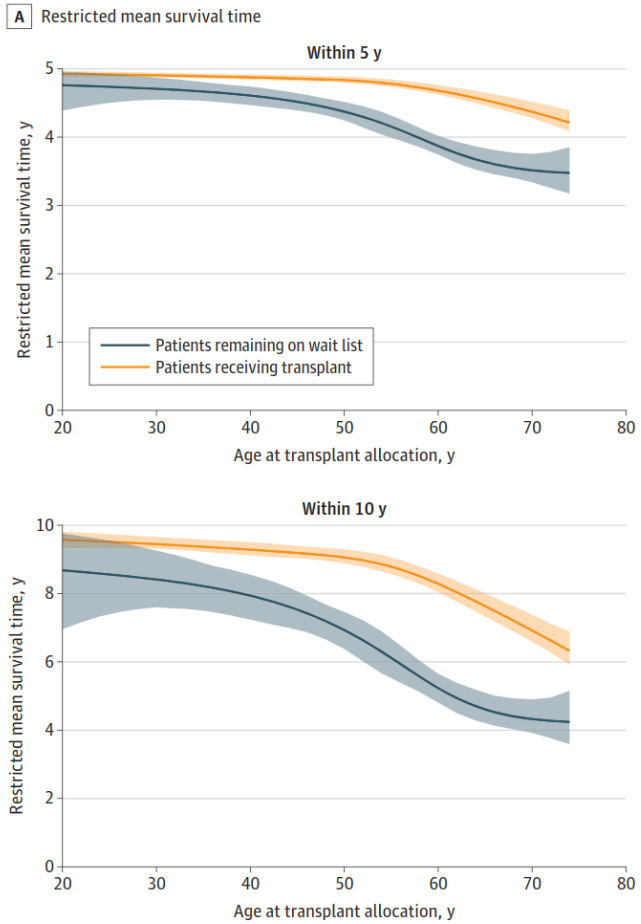
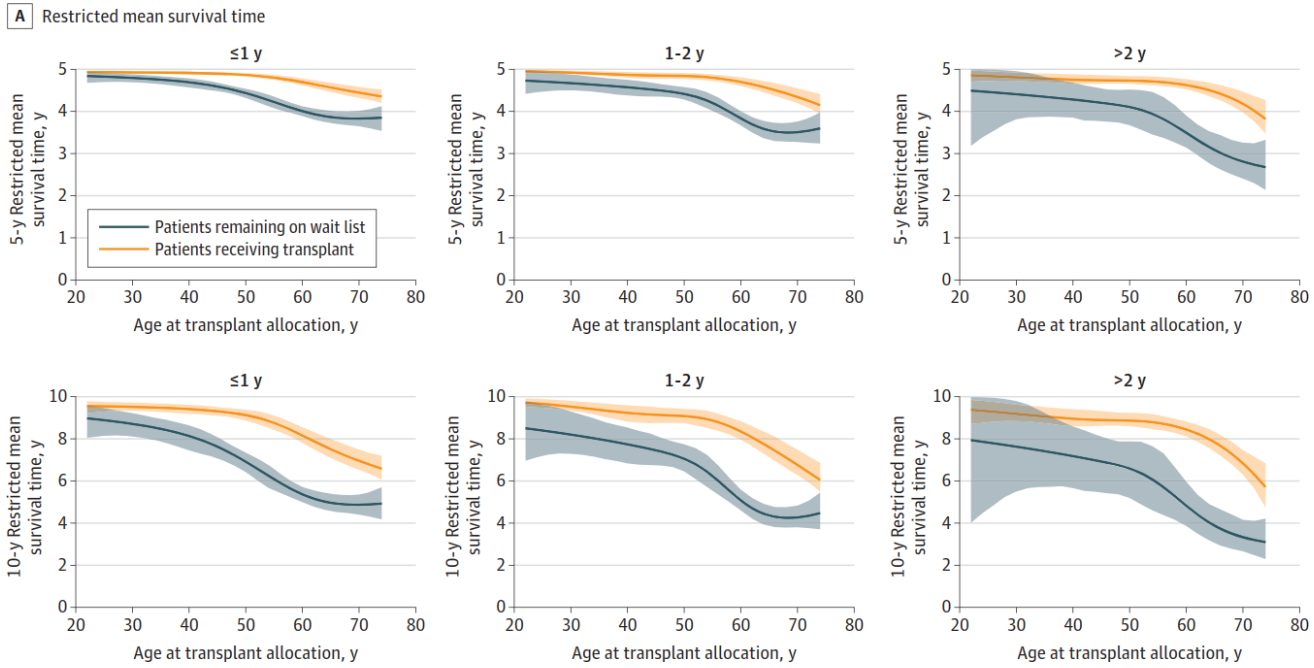


Figure 3. Restricted Mean Survival Times Conditional on Wait-listing Duration



A, Five-year and 10-year restricted mean survival times for all-cause mortality. B, Five-year and 10-year restricted mean survival times for all-cause mortality differences. Shaded areas indicate 95% CIs.

Pathways to excellence (1)

- Clearly distinguish between descriptive, predictive and causal research questions:
  - Fundamental (*descriptive*) prognosis research … descriptive
  - Prognostic factor research … predictive
  - Prognostic model research … predictive
  - Stratified medicine research … causal

- Carlin and Moreno-Betancur (2023):

  ‚… *it should be emphasised that most areas of health and medicine advance by examining questions of all three types.*‘

  ‚*Unfortunately, this fundamental taxonomy of research questions has barely penetrated the teaching and practice of biostatistics, especially with respect to regression models.*‘

Pathways to excellence (2)

- Descriptive research is about summarizing outcomes in a population or about quantifying differences in outcomes between different subjects
- Predictive research is about (improving) accuracy of predictions
- Causal research is about effects of alternative interventions within the same subjects
  - This excludes research questions like ‚effect of sex‘, ‚effect of age‘, …!

Pathways to excellence (3)

- In all domains, estimates are preferred over tests
  - ‚*We would like to quantify the difference*‘ > ‚*We would like to infer if there is a difference*‘
  - (Ir)relevance of null hypotheses in descriptive research?
  - (Ir)relevance of p-values and confidence intervals in multivariable models?
  - Quantify the uncertainty, but with no cut offs

Pathways to excellence (4)

- The tedious homework of statisticians:

- Prespecification of analysis plans: SAPI

- Conducting analysis in reproducible way: same data, same code, same results!

- Transparent reporting of what was done
  → EQUATOR network
  https://www.equator-network.org/

Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions
Observational studies	STROBE	Extensions
Systematic reviews	PRISMA	Extensions
Study protocols	SPIRIT	PRISMA-P
Diagnostic/prognostic studies	STARD	TRIPOD
Case reports	CARE	Extensions
Clinical practice guidelines	AGREE	RIGHT
Qualitative research	SRQR	COREQ
Animal pre-clinical studies	ARRIVE	
Quality improvement studies	SQUIRE	Extensions
Economic evaluations	CHEERS	Extensions

See all 655 reporting guidelines

References (1)

- Altman, D.G., Lyman, G.H., 1998. Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 52, 289–303. https://doi.org/10.1023/A:1006193704132
- Carlin, J.B., Moreno-Betancur, M., 2023. On the uses and abuses of regression models: a call for reform of statistical practice and teaching. https://doi.org/10.48550/ARXIV.2309.06668
- Collins, G.S., Dhiman, P., Andaur Navarro, C.L., Ma, J., Hooft, L., Reitsma, J.B., Logullo, P., Beam, A.L., Peng, L., Van Calster, B., Van Smeden, M., Riley, R.D., Moons, K.G., 2021. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 11, e048008. https://doi.org/10.1136/bmjopen-2020-048008
- Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J.B., Van Smeden, M., Boulesteix, A.-L., Camaradou, J.C., Celi, L.A., Denaxas, S., Denniston, A.K., Glocker, B., Golub, R.M., Harvey, H., Heinze, G., Hoffman, M.M., Kengne, A.P., Lam, E., Lee, N., Loder, E.W., Maier-Hein, L., Mateen, B.A., McCradden, M.D., Oakden-Rayner, L., Ordish, J., Parnell, R., Rose, S., Singh, K., Wynants, L., Logullo, P., 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ e078378. https://doi.org/10.1136/bmj-2023-078378
- Deforth, M., Heinze, G., Held, U., 2024. The performance of prognostic models depended on the choice of missing value imputation algorithm: a simulation study. Journal of Clinical Epidemiology 176, 111539. https://doi.org/10.1016/j.jclinepi.2024.111539
- Gregorich, M., Strohmaier, S., Dunkler, D., Heinze, G., 2021. Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. IJERPH 18, 4259. https://doi.org/10.3390/ijerph18084259
- Hafermann, L., Becher, H., Herrmann, C., Klein, N., Heinze, G., Rauch, G., 2021. Statistical model building: Background “knowledge” based on inappropriate preselection causes misspecification. BMC Med Res Methodol 21, 196. https://doi.org/10.1186/s12874-021-01373-z

References (2)

- Hafermann, L., Klein, N., Rauch, G., Kammer, M., Heinze, G., 2022. Using Background Knowledge from Preceding Studies for Building a Random Forest Prediction Model: A Plasmode Simulation Study. Entropy 24, 847. https://doi.org/10.3390/e24060847
- Harrell , F.E., 2015. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Springer Series in Statistics. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-19425-7
- 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
- Heinze, G., Wallisch, C., Dunkler, D., 2018. Variable selection – A review and recommendations for the practicing statistician. Biometrical J 60, 431–449. https://doi.org/10.1002/bimj.201700067
- Hemingway, H., Croft, P., Perel, P., Hayden, J.A., Abrams, K., Timmis, A., Briggs, A., Udumyan, R., Moons, K.G.M., Steyerberg, E.W., Roberts, I., Schroter, S., Altman, D.G., Riley, R.D., for the PROGRESS Group, 2013. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ 346, e5595–e5595. https://doi.org/10.1136/bmj.e5595
- Hernán, M.A., Hsu, J., Healy, B., 2019. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. CHANCE 32, 42–49. https://doi.org/10.1080/09332480.2019.1579578
- Moons, K.G.M., Wolff, R.F., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S., 2019. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 170, W1. https://doi.org/10.7326/M18-1377
- Perel, P., Prieto-Merino, D., Shakur, H., Clayton, T., Lecky, F., Bouamra, O., Russell, R., Faulkner, M., Steyerberg, E.W., Roberts, I., 2012. Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ 345, e5166–e5166. https://doi.org/10.1136/bmj.e5166

References (3)

- Royston, P., Sauerbrei, W., 2008. Multivariable Model-Building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables, 1st ed, Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/9780470770771
- Royston, P., Sauerbrei, W., 2007. Multivariable Modeling with Cubic Regression Splines: A Principled Approach. The Stata Journal 7, 45–70. https://doi.org/10.1177/1536867X0700700103
- Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H., Dunkler, D., Harrell, F.E., Royston, P., Heinze, G., for TG2 of the STRATOS initiative 2020. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagn Progn Res 4, 3, s41512-020-00074–3. https://doi.org/10.1186/s41512-020-00074-3
- Shmueli, G., 2010. To Explain or to Predict? Statist. Sci. 25. https://doi.org/10.1214/10-STS330
- Šinkovec, H., Heinze, G., Blagus, R., Geroldinger, A., 2021. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. BMC Med Res Methodol 21, 199. https://doi.org/10.1186/s12874-021-01374-y
- Strohmaier, S., Wallisch, C., Kammer, M., Geroldinger, A., Heinze, G., Oberbauer, R., Haller, M.C., 2022. Survival Benefit of First Single-Organ Deceased Donor Kidney Transplantation Compared With Long-term Dialysis Across Ages in Transplant-Eligible Patients With Kidney Failure. JAMA Netw Open 5, e2234971. https://doi.org/10.1001/jamanetworkopen.2022.34971
- 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
- Wallisch, C., Agibetov, A., Dunkler, D., Haller, M., Samwald, M., Dorffner, G., Heinze, G., 2021. The roles of predictors in cardiovascular risk models - a question of modeling culture? BMC Med Res Methodol 21, 284. https://doi.org/10.1186/s12874-021-01487-4
- Wolff, R.F., Moons, K.G.M., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S., for the PROBAST Group†, 2019. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 170, 51. https://doi.org/10.7326/M18-1376