Survival Skills for Surviving al Data Analysis

Antje Jahn

University of Applied Sciences Darmstadt

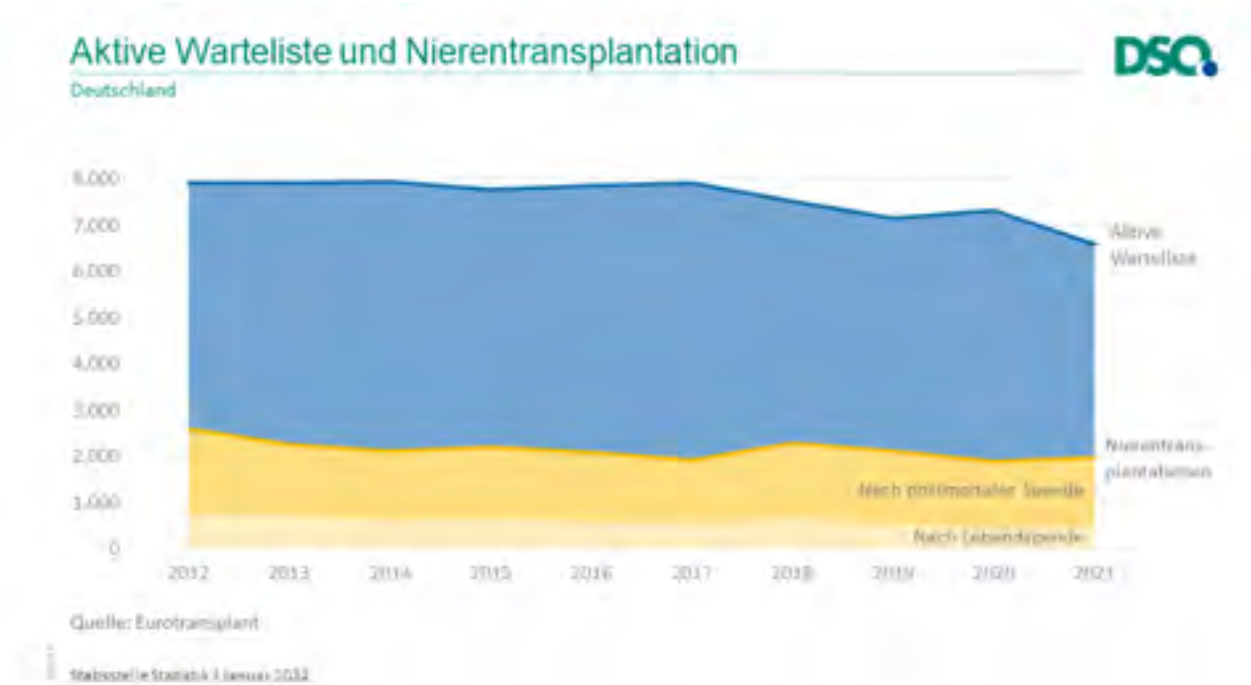
KSFE 2023

Outline

- Motivating Example: Prediction of post-transplant survival
- Censoring and its Implications
- Differences between Regression & Machine Learning for Survival Prediction
- Methods comparison on data of ≈ 50.000 donor kidney recipients

Motivating Example

- A kidney transplant is often the treatment of choice for patients with end-stage kidney disease
- There is a gap between supply and demand for kidney transplantation



- About 15-20% of procured kidneys are finally not used, mostly the so-called marginal kidneys [1].

Motivating Example

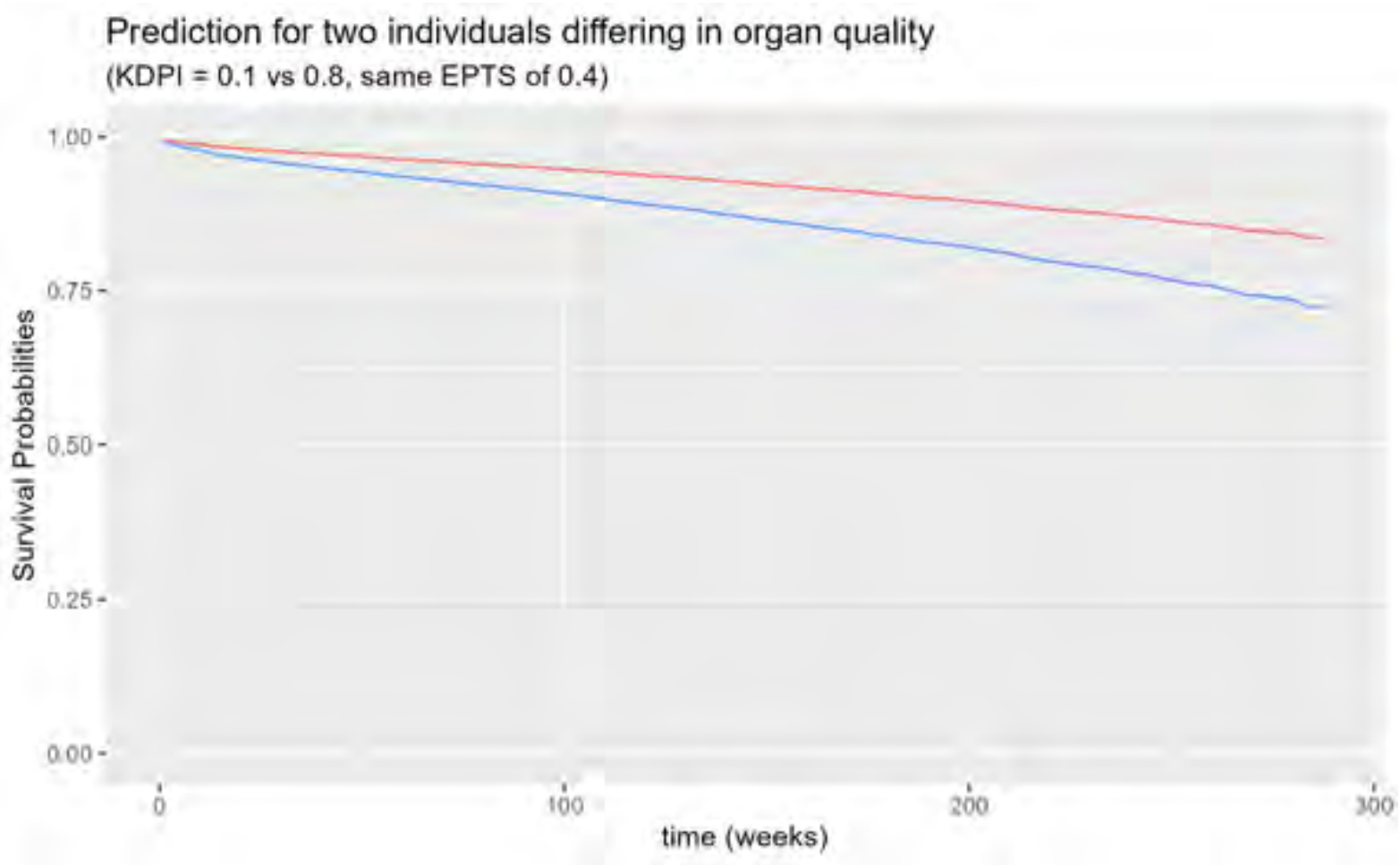
Reliable statistical methods for predicting a patients' individual risk for survival after kidney transplantation are of great importance to support

- the individual risk-benefit assessment before transplantation
- defining individual centers' and patients' kidney acceptance criteria
- guiding organ allocation policies
- making informed use of marginal kidneys

Examples: Eurotransplant Senior Program (ESP), Eurotransplant Rescue Allocation Program, US Kidney Accelerated Placement project (KAP)

Motivating Example

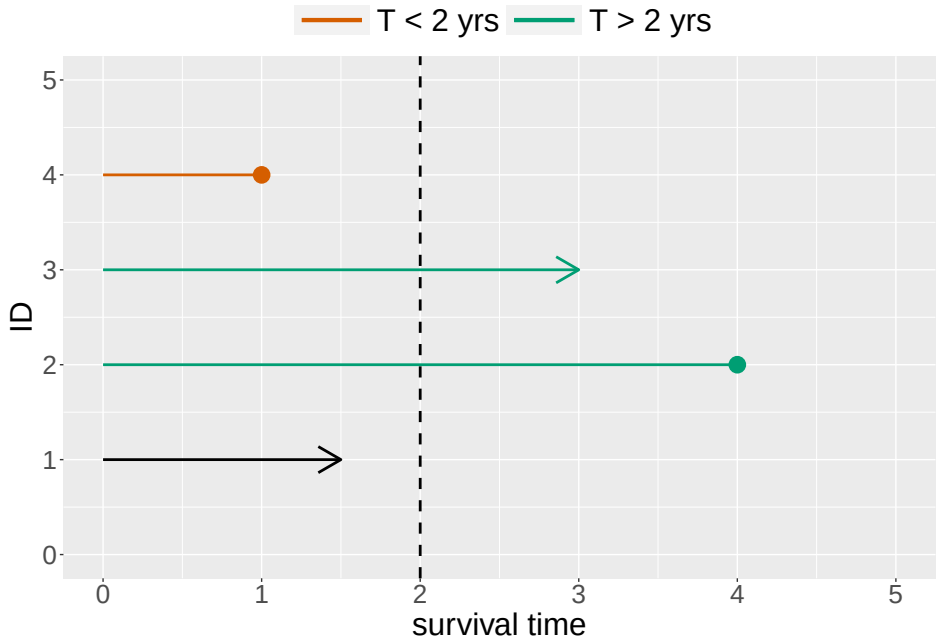
Exemplarily predictions derived from data of $n \approx 50.000$ donor kidney receivers



Outline

- Motivating Example: Prediction of post-transplant survival
- Censoring and its Implications
- Differences between Regression & Machine Learning for Survival Prediction
- Methods comparison on data of ≈ 50.000 donor kidney recipients

Censoring and its implications

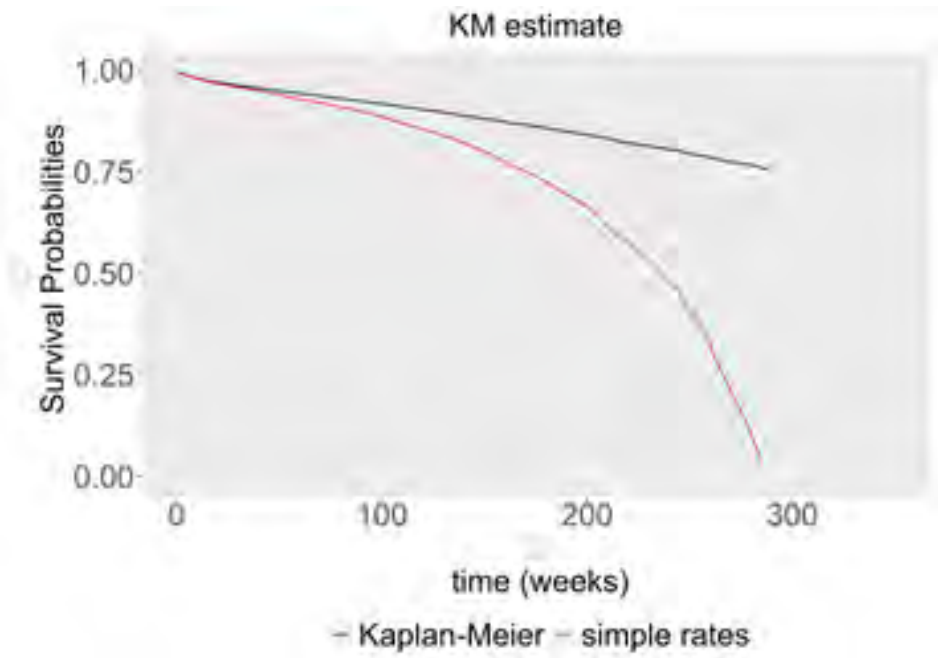


- Censored obs are missing for classification at some time t
- Censored obs hold partial information on survival
- Longer survival implies a higher risk for censoring

Censoring is not just missing data, but requires survival analysis

Censoring and its implications (1)

- Discarding censored observations from analyses causes systematic bias
- Example

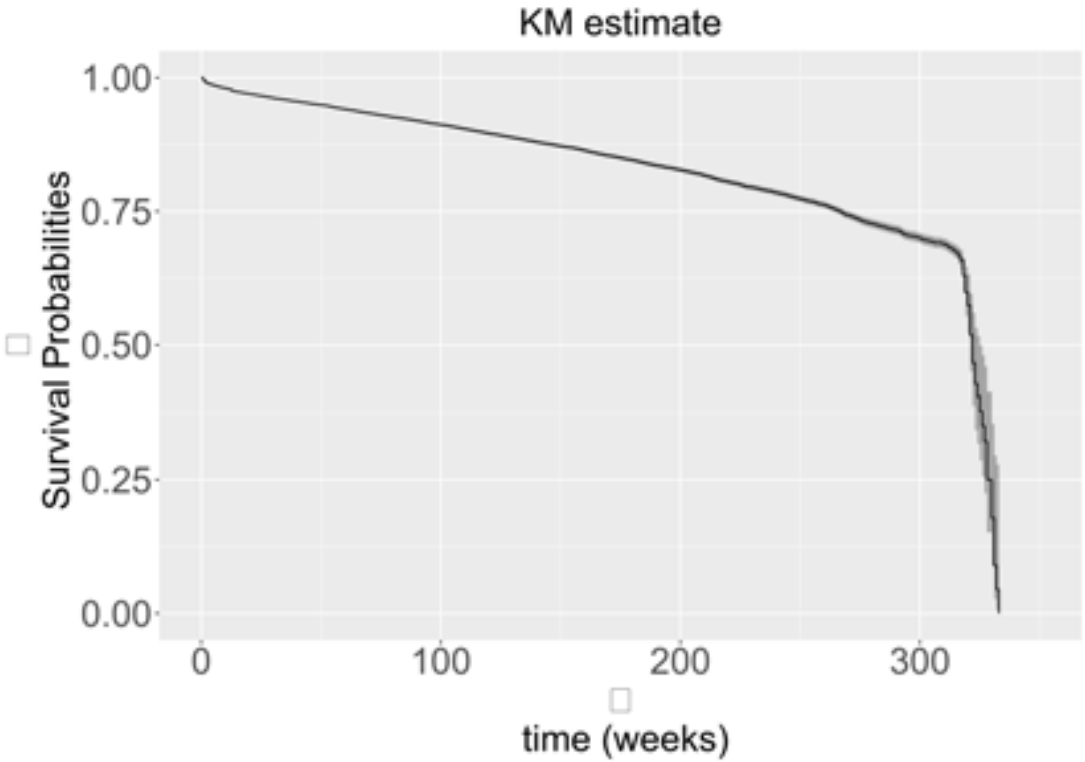


Simple rates:

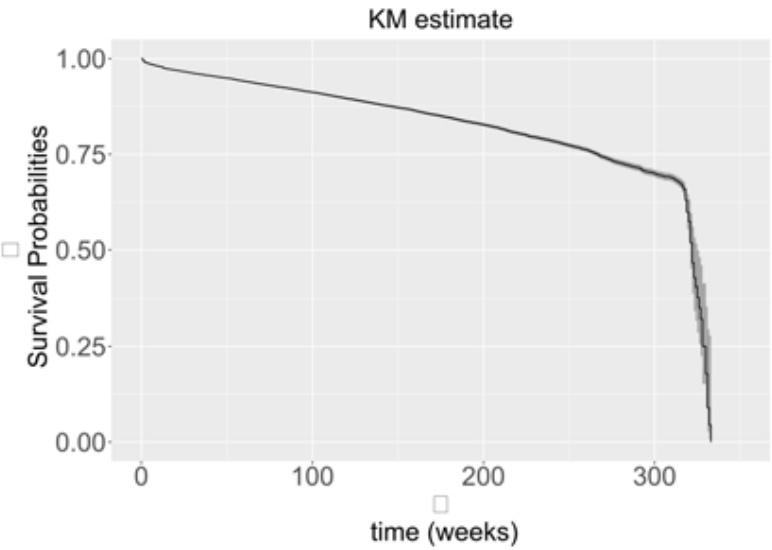
$$\frac{|\{\text{i; i survives t}\}|}{|\{\text{i; i not censored before t}\}|}$$

Censoring and its implications (2)

- Even Kaplan-Meier-Estimates (that consider censoring) can give implausible results as seen with UNOS data

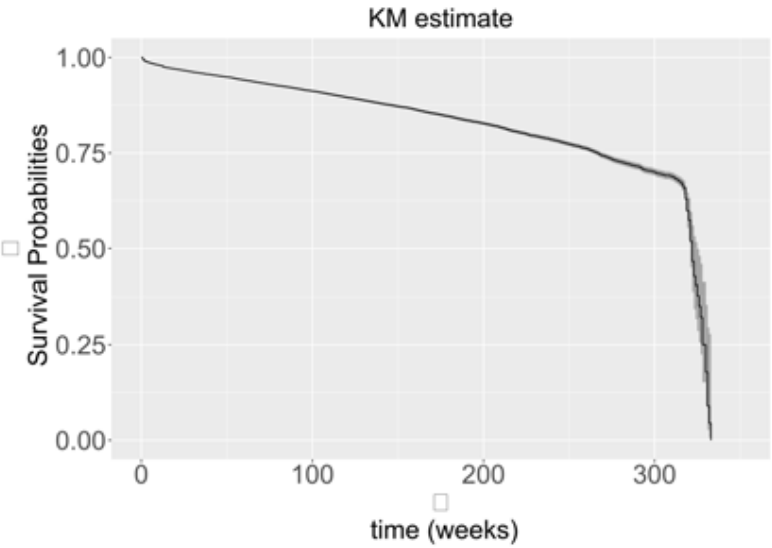


Censoring and its implications (2)



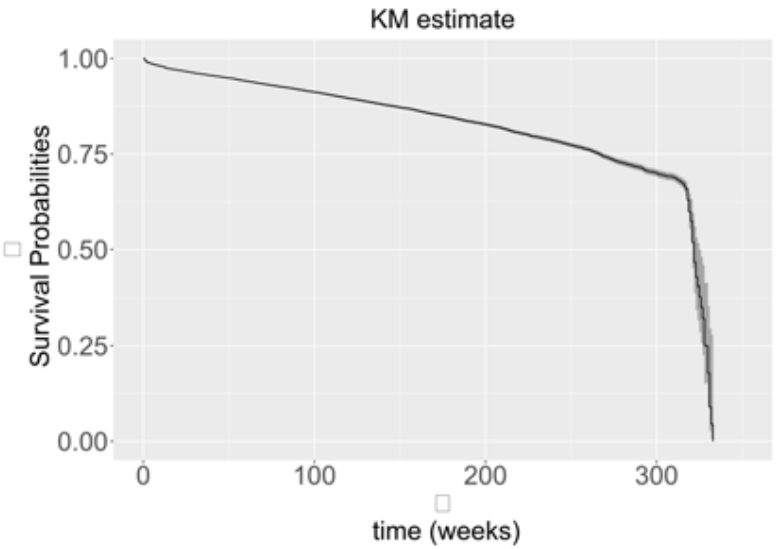
- 1st patient in: Jan 2014, DB closure: Sep 2020 (330 wks)
- FU visits are scheduled yearly $\Rightarrow$ After wk 278 the risk sets reduce to patients with events

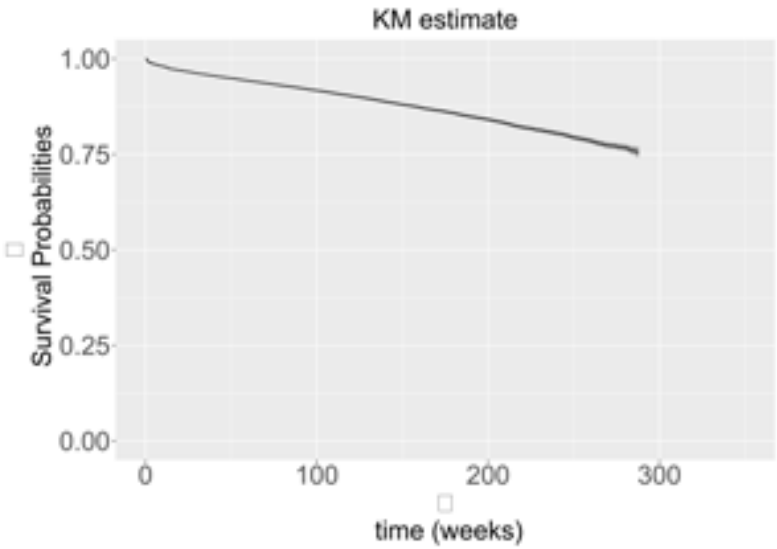
Censoring and its implications



- 1st patient in: Jan 2014, DB closure: Sep 2020 (330 wks)
- FU visits are scheduled yearly $\Rightarrow$ After wk 278 the risk sets reduce to patients with events
- Administrative censoring has been added at Jul 2019

Censoring and its implications





- 1st patient in: Jan 2014, DB closure: Sep 2020 (330 wks)
- FU visits are scheduled yearly $\Rightarrow$ After wk 278 the risk sets reduce to patients with events
- Administrative censoring has been added at Jul 2019

When to censor is not always easy to answer [2]

Outline

- Motivating Example: Prediction of post-transplant survival
- Censoring and its Implications
- Differences between Regression & Machine Learning for Survival Prediction
- Methods comparison on data of ≈ 50.000 donor kidney recipients

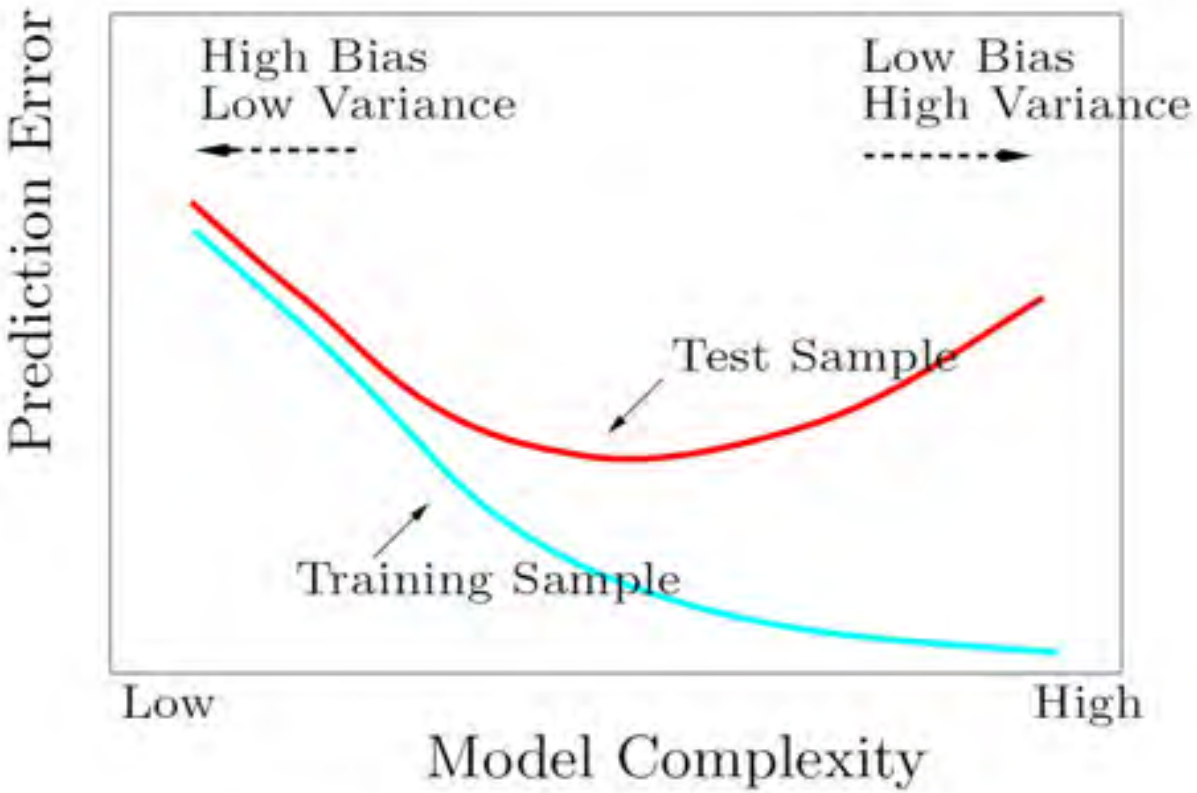
Regression & Machine Learning for Survival Prediction

- Risk of bias in prediction models for living kidney transplantation [3]
- The models were mainly derived by regression

  - Median number of 3.8 events per candidate predictors
  - Correcting for optimism infrequently provided - despite of guidelines
  - 11/29 models were neither internally nor externally validated
  - Weak strategies for model-building (e.g. screening of predictors)

Regression & Machine Learning for Survival Prediction

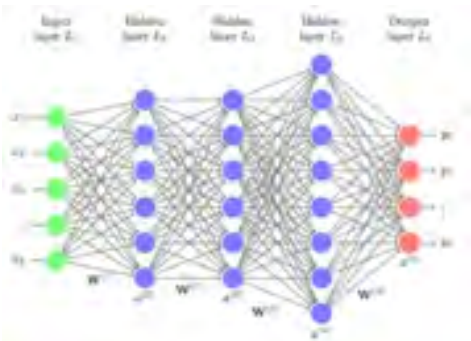
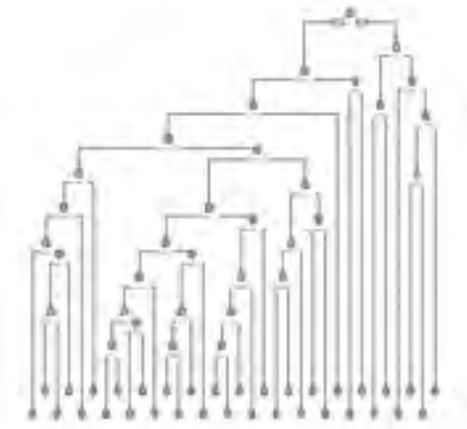
Overfitting:



Hastie et. al: The Elements of Statistical Learning. Springer

Regression & Machine Learning for Survival Prediction

- In Machine Learning the risk of overfitting is well-known and rarely ignored
- Example for UNOS data on n=36346 with k=3909 observed events

Cox model	Neural Network	Trees & Forests
Linear predictor $b^T X$		
95 parameters	3889 parameters	Flexible number of parameters

Regression & Machine Learning for Survival Prediction

- Correcting for overfitting and optimism is often integrated within the machine learning algorithm, e.g. by
  - Penalizing loss functions
  - Tuning hyperparameters by cross-validation
  - Bagging with weak learners
  - Early stopping of the fitting algorithm
  - Estimating accuracy from out-of-bag data
- However: Machine Learning has its own risk of bias by improper handling of censoring

Regression & Machine Learning for Survival Prediction

- Discarding censored observations was observed in 9 out of 12 machine learning applications for predicting kidney graft failure [4]
- In regression modeling this is observed infrequently
- Potential reason: Software comfort zone?
  - Decade-long history of survival packages in R
  - Integration of survival methods into Python scikit-learn since 2020

Outline

- Motivating Example: Prediction of post-transplant survival
- Censoring and its Implications
- Differences between Regression & Machine Learning for Survival Prediction
- Methods comparison on data of ≈ 50.000 donor kidney recipients

Methods comparison

Objective

Comparison of prediction performance for post-transplant survival when statistical and machine learning methods both consider censoring and overfitting

- Compared methods
  - Cox regression model with backward variable selection
  - Cox regression model with Lasso penalization and cross-validated shrinkage parameter
  - Random Survival Forest with default tuning parameters
  - Feed-forward neural network (DeepSurv) with hyperparameter tuning (inner layers, nodes per layer, drop-out rate)
- All methods are trained on a 70% random sample and evaluated on a 30% test sample

A sketch on statistical learning for event time data

- Let T^* and C be the time to event and censoring, respectively
- T^* is (for some subjects) unobserved, only T and D is observed:

$$T = min(T^*, C) \qquad D = \mathbf{I}(T^* \leq C)$$

- Fit a model using T and D only, that minimizes some loss function for censored data

The Cox regression model

Parametric model of the hazard function depending on covariate vector x

$$\begin{aligned} h(t|x) &= \lim_{\Delta \downarrow 0} \frac{P(t \leq T^* < t + \Delta \mid T^* \geq t, X = x)}{\Delta} \\ &= h_0(t) \exp(b^T x) \end{aligned}$$

Log partial likelihood loss function:

$$LL(b) = \sum_{i=1}^{n} d_i [b^T x_i - \log \Big(\sum_{j \in R_i} \exp(b^T x_j) \Big)]$$

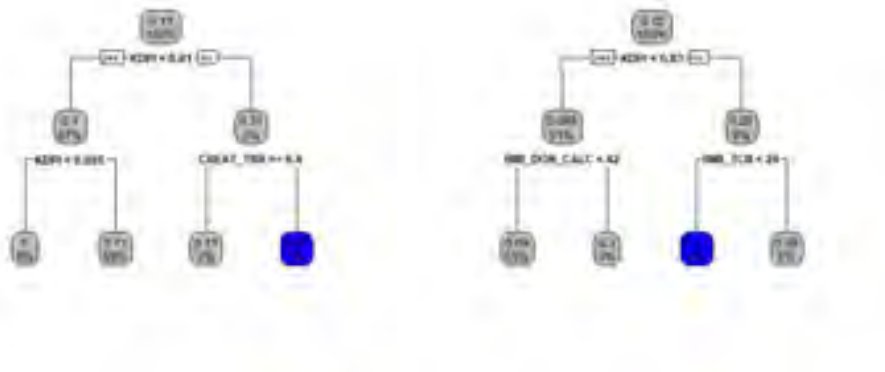
Loss function with Lasso-regularization:

$$Loss(b) = -LL(b) + \lambda \sum_{j=1}^{p} |b_j|$$

Methods comparison: Random Survival Forest

General idea: Average the predictions of several weak learners (the trees)

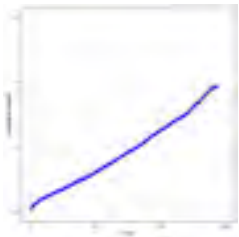
(a) Grow trees from bootstrap samples and random variable candidates for splitting



(b) Calculate Nelson-Aalen estimate of CHF per terminal node K

$$\hat{H}_K^b(t) = \sum_{i \in K, t_i \leq t} \frac{d_i}{n_i(K)}$$

(c) Predict the CHF per subject by its average over the trees



Methods comparison: Feed forward neural network - DeepSurv

- Proportional hazards assumption relaxing linearity and additivity:

$$h(t|x) = h_0(t)\exp(\phi_w(x))$$

- Predicting $\phi_w(x)$ by a neural net with loss-function

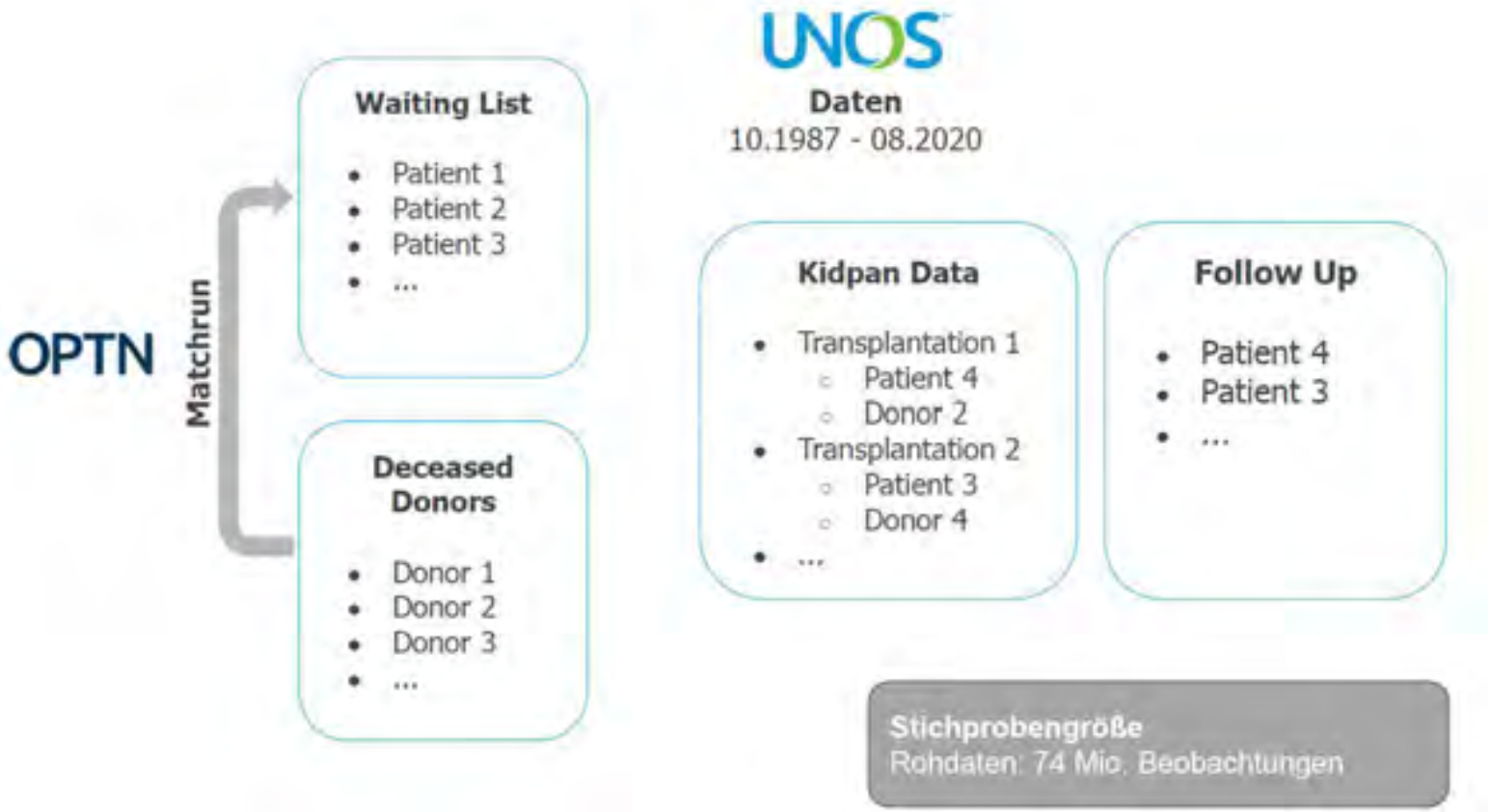
$$Loss_\lambda(w) = -\sum_{i=1}^{n} d_i\left[\phi_w(x_i) - \log\left(\sum_{j\in R_i}\exp(\phi_w(x_j))\right)\right] + \lambda||w||_2^2$$

- For prediction use $\hat{H}(t|x) := \hat{H}_0(t)\exp(\hat{\phi}_w(x))$ with Breslow estimator $\hat{H}_0(t)$

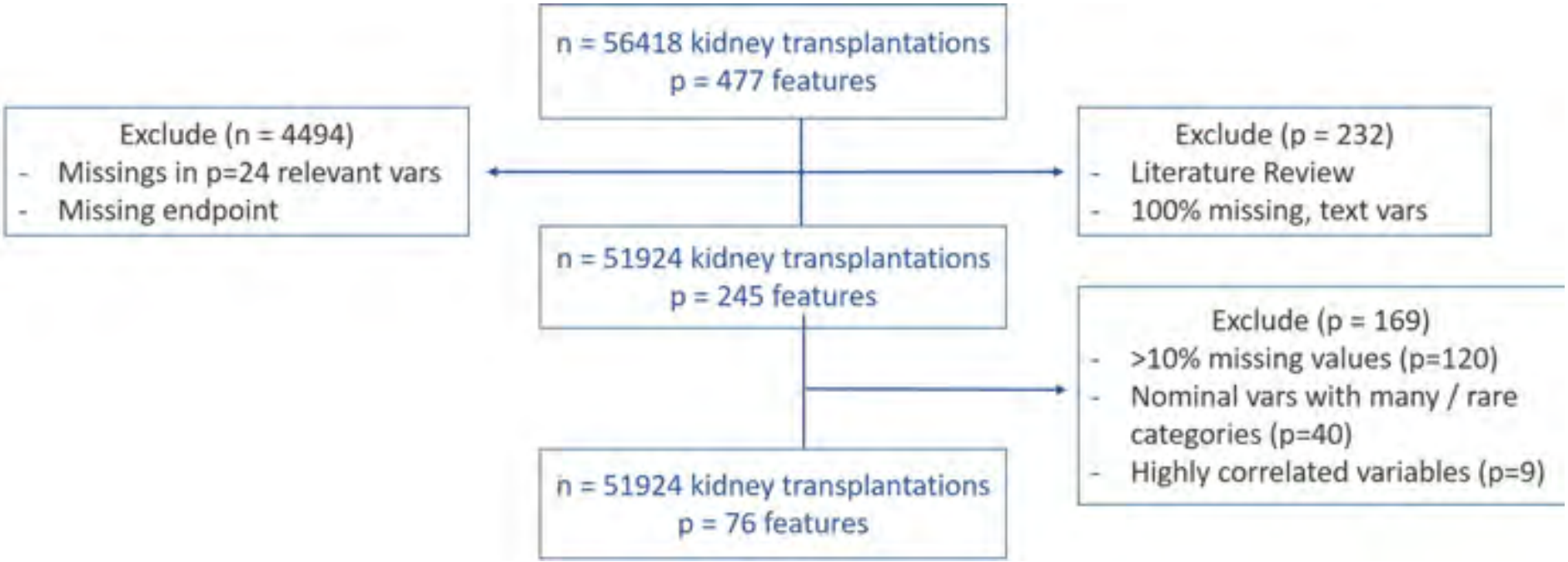
Methods comparison

	Non-prop. hazards	Non-linear predictors	Effect estimation
Cox Backward	-	-	✓
Cox Lasso	-	-	✓
Random Survival Forest	✓	✓	-
Neural Network	-	✓	-

Methods comparison: Data



Methods comparison: Data



Methods comparison: Results - Discrimination

- Consider prediction for patient A is worse than for patient B
  - A dies before B: concordant pair
  - B dies before A: discordant pair
  - censoring masks the comparison: uninformative pair
- C-Index: Proportion of concordant among all informative pairs of patients

Methods comparison: Results - Discrimination

- Consider prediction for patient A is worse than for patient B
  - A dies before B: concordant pair
  - B dies before A: discordant pair
  - censoring masks the comparison: uninformative pair
- C-Index: Proportion of concordant among all informative pairs of patients

Cox Backward	Cox Lasso	Random Forest[(1)]	Neural Network
0.644	0.643	0.619	0.640

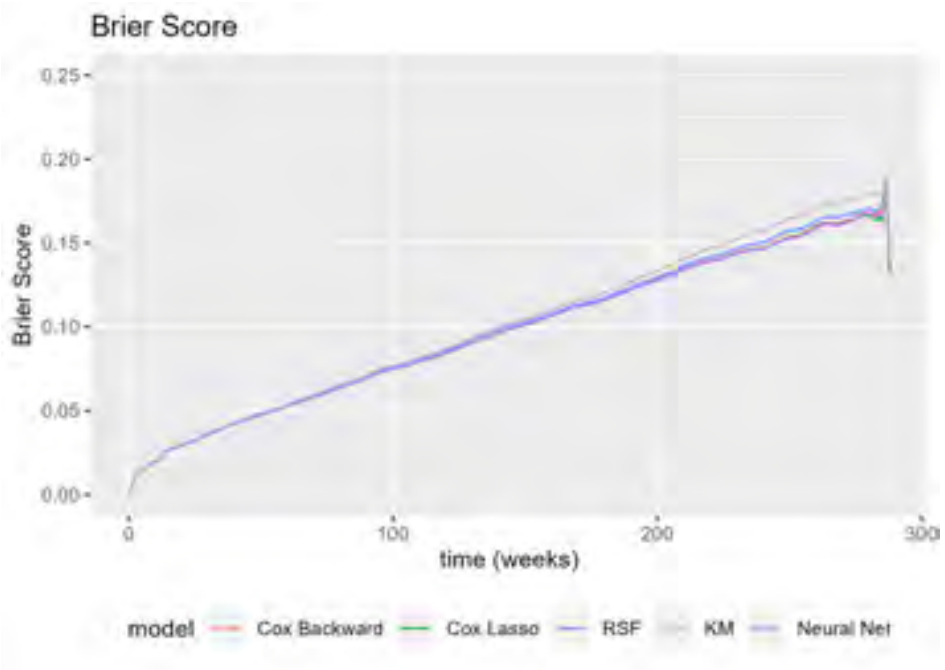
[(1)] based on OOB data and with score ϕ defined as mean CHF over observed event times

Methods comparison: Results - Accuracy

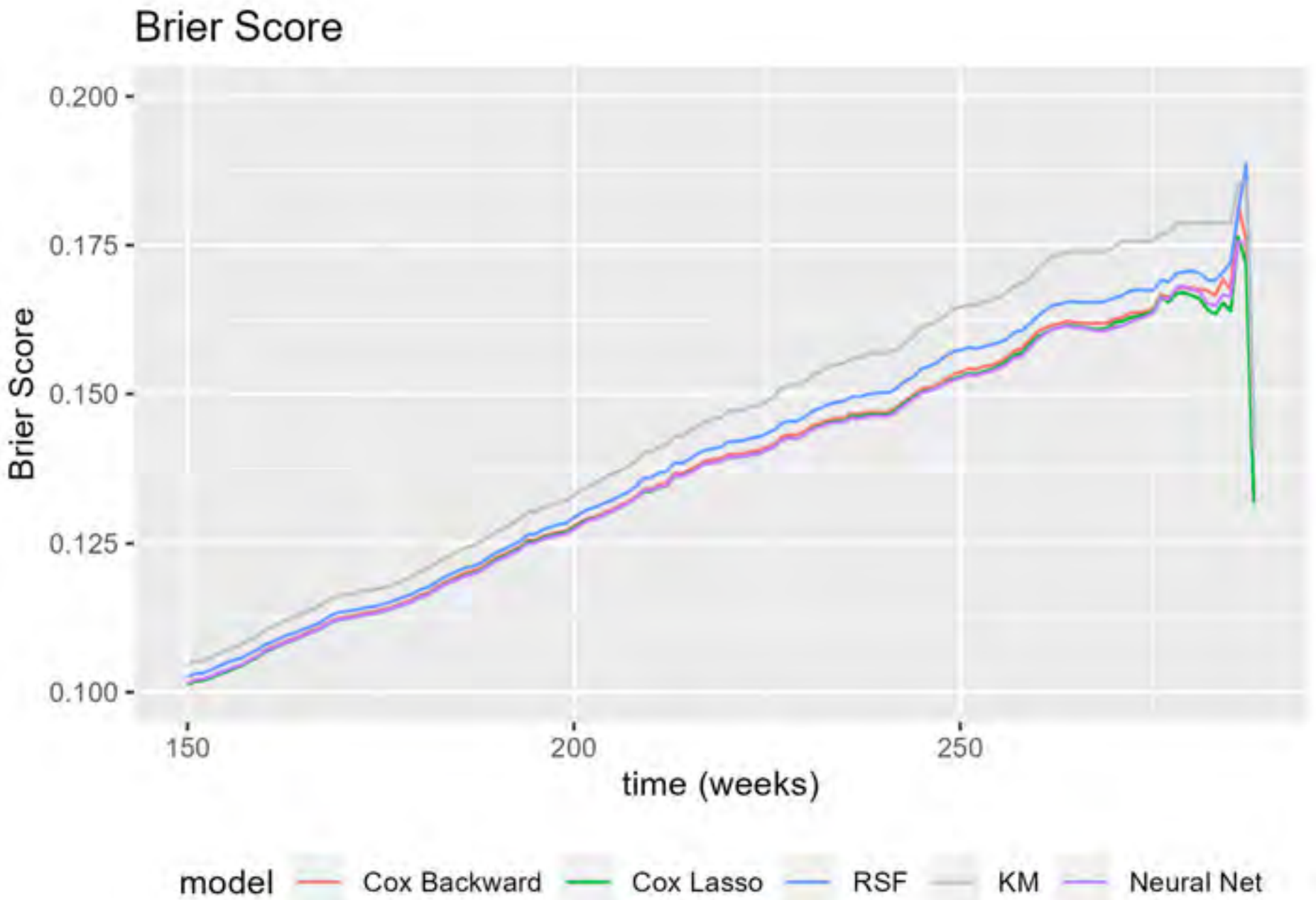
- Squared error of predicted probability to survive time t, $\hat{S}(t)$
  - Patient survives t: $(1 - \hat{S}(t))^2$
  - Patient dies before t: $(0 - \hat{S}(t))^2$
  - Censoring masks information: Represented by other patients

Methods comparison: Results - Accuracy

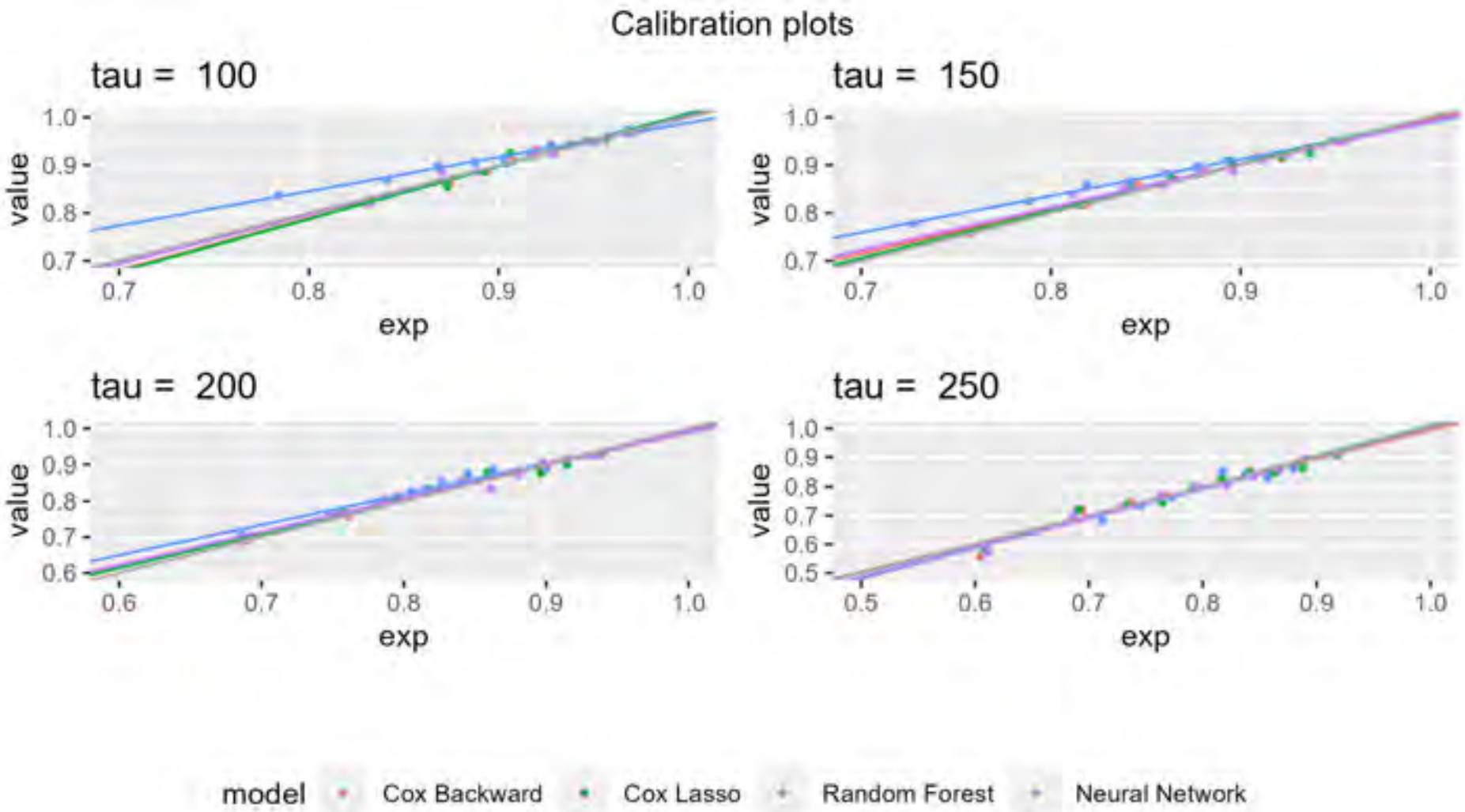
- Squared error of predicted probability to survive time t, $\hat{S}(t)$
  - Patient survives t: $(1 - \hat{S}(t))^2$
  - Patient dies before t: $(0 - \hat{S}(t))^2$
  - Censoring masks information: Represented by other patients
- Mean Squared Error (Brier Score) over time:



Methods comparison: Results - Accuracy



Methods comparison: Results - Calibration



Methods comparison: Summary

- Random Survival Forest showed poorest prediction performance - potentially by skipping hyperparameter tuning
- Only slight differences between Cox Regression and Neural Network
- Potential reasons:
  - Machine Learning needs more data
  - Data needs larger models
  - Negligible non-linear and non-additive effects

Summary

- Overfitting is not a problem of machine learning only
- Censoring has a long tradition in classical statistics, but is sometimes less carefully addressed in machine learning
- Systematic methods' comparisons are required to inspect the "promise of machine learning" [5]

Acknowledgements

- This was joint work with Gunter Grieser and Lukas Klein
- We thank the Organ Procurement and Transplantation Network (OPTN) for sharing the data on kidney transplantations

The data reported here have been supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and Transplantation Network.

The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S.

Government. Based on OPTN data as of June 20, 2020

UNOS

References

(1) Bae S et al.: *Changes in Discard Rate After the Introduction of the Kidney Donor Profile Index (KDPI)*. American Journal of Transplantation 2016. doi:10.1111/ajt.13769

(2) Lesko C et al.: *When to censor?*. American Journal of Epidemiology 2018. doi: 10.1093/aje/kwx281

(3) Haller M et al.: *Prediction models for living organ transplantation are poorly developed, reported, and validated: a systematic review* . Journal of Clinical Epidemiology 2022, doi: 10.1016/j.jclinepi.2022.01.025

(4) Senanayake S et al.: *Machine learning in predicting graft failure following kidney transplantation: A systematic review of published predictive models*. International Journal of Medical Informatics 2019, doi: 10.1016/j.ijmedinf.2019.103957

(5) Gotlieb N et al.: *The promise of machine learning applications in solid organ transplantation*. npj Digital Medicine 2022, doi: 10.1038/s41746-022-00637-2

antje.jahn@h-da.de

Back-up-Folien :