A Large-Scale Neutral Comparison Study of Survival Models

Burk, L.[1,2,3,4] Zobolas, J.[5] Bischl, B.[2,4] Bender, A.[2,4] Wright, M. N.[1,3] Lang, M.[6] Sonabend, R.[7, 8]

[1]Leibniz Institute for Prevention Research and Epidemiology – BIPS
[2]LMU Munich [3]University of Bremen
[4]Munich Center for Machine Learning (MCML)
[5]University of Oslo [6] TU Dortmund
[7]OSPO Now [8]Imperial College, London

Biometric Colloquium — March 1st, 2024

Introduction

- There are many survival learners ("models") to choose from

Introduction

- There are many survival learners (“models”) to choose from
- Advantages and Disadvantages often unclear, specific to setting

Introduction

- There are many survival learners ("models") to choose from
- Advantages and Disadvantages often unclear, specific to setting
- Various comparisons exist in literature

Introduction

- There are many survival learners (“models”) to choose from
- Advantages and Disadvantages often unclear, specific to setting
- Various comparisons exist in literature
- Limited scope (learners, tasks, evaluation measures)

Introduction

- There are many survival learners ("models") to choose from
- Advantages and Disadvantages often unclear, specific to setting
- Various comparisons exist in literature
- Limited scope (learners, tasks, evaluation measures)
- Focus on individual / new method $\Rightarrow$ no neutral comparison

Introduction

- There are many survival learners ("models") to choose from
- Advantages and Disadvantages often unclear, specific to setting
- Various comparisons exist in literature
- Limited scope (learners, tasks, evaluation measures)
- Focus on individual / new method $\Rightarrow$ no neutral comparison
- No (or limited) quantitative comparison

Introduction

- There are many survival learners (“models”) to choose from
- Advantages and Disadvantages often unclear, specific to setting
- Various comparisons exist in literature
- Limited scope (learners, tasks, evaluation measures)
- Focus on individual / new method $\Rightarrow$ no neutral comparison
- No (or limited) quantitative comparison

Introduction

- There are many survival learners (“models”) to choose from
- Advantages and Disadvantages often unclear, specific to setting
- Various comparisons exist in literature
- Limited scope (learners, tasks, evaluation measures)
- Focus on individual / new method $\Rightarrow$ no neutral comparison
- No (or limited) quantitative comparison

$\Rightarrow$ Needs comprehensive comparison!

Quick Summary

- 32 tasks
- 18 learners
- 2 tuning measures
- 9 evaluation measures

Quick Summary

- **32** tasks
- **18** learners
- **2** tuning measures
- **9** evaluation measures

- **Large-scale** $\Rightarrow$ Generalizability
- **Neutral** $\Rightarrow$ Fair comparison

Quick Summary

- **32** tasks
- **18** learners
- **2** tuning measures
- **9** evaluation measures

- **Large-scale** $\Rightarrow$ Generalizability
- **Neutral** $\Rightarrow$ Fair comparison

$\Rightarrow$ The largest survival benchmark to date as far as we know

Scope

The “Standard Setting”:

- Single-event outcome: $\delta_i \in \{0, 1\}$
- Low-dimensional: $2 \leq p < n$
- No time-varying covariates
- Right-censoring only
- At least 100 observed events

Tasks

32 tasks collected from R packages on CRAN

	Minimum	q25%	Median	q75%	Maximum
N	137	446	820	2378	52410
p	2	4	5	7	25
Observed Events	101	194	323	699	5616
Cens. %	6	32	48	74	95

Learners

18 learners implemented in R and available via the `mlr3` [1] framework

- **Baseline**: Kaplan-Meier & Nelson-Aalen

[1] Lang et al. (2019)

Learners

18 learners implemented in R and available via the `mlr3` [1] framework

- **Baseline**: Kaplan-Meier & Nelson-Aalen
- **Classical**: Cox, penalized, parametric

[1] Lang et al. (2019)

Learners

18 learners implemented in R and available via the `mlr3` [1] framework

- **Baseline**: Kaplan-Meier & Nelson-Aalen
- **Classical**: Cox, penalized, parametric
- **Trees**: Individual and ensembles thereof

[1] Lang et al. (2019)

Learners

18 learners implemented in R and available via the `mlr3` [1] framework

- **Baseline**: Kaplan-Meier & Nelson-Aalen
- **Classical**: Cox, penalized, parametric
- **Trees**: Individual and ensembles thereof
- **Boosting**: Gradient- and likelihood-based

[1] Lang et al. (2019)

Learners

18 learners implemented in R and available via the `mlr3` [1] framework

- **Baseline**: Kaplan-Meier & Nelson-Aalen
- **Classical**: Cox, penalized, parametric
- **Trees**: Individual and ensembles thereof
- **Boosting**: Gradient- and likelihood-based
- **Other**: Akritas, SVM

[1] Lang et al. (2019)

List of Learners (Baseline, Classical)

Name	Abbreviation	Package
Kaplan-Meier	KM	`survival`
Nelson-Aalen	NA	`survival`
Cox Regression	CPH	`survival`
Penalized Cox Regression (L1, L2)	GLM	`glmnet`
Penalized Cox Regression (L1, L2)	Pen	`penalized`
Parametric (AFT)	Par	`survival`
Flexible Parametric Splines	Flex	`flexsurv`
Akritas	AK	`survivalmodels`
Survival SVM	SSVM	`survivalsvm`

List of Learners (Trees, Boosting)

Name	Abbreviation	Package
Decison Tree	RRT	`rpart`
Random Survival Forest	RFSRC	`randomForestSRC`
Random Survival Forest	RAN	`ranger`
Conditional Inference Forest	CIF	`partykit`
Oblique RSF	ORSF	`aorsf`
Model-Based Boosting	MBO	`mboost`
Likelihood-Based Boosting	CoxB	`CoxBoost`
Gradient Boosting (Cox objective)	XGBCox	`xgboost`
Gradient Boosting (AFT objective)	XGBAFT	`xgboost`

Tuning

- Tuning spaces discussed with learner authors

Tuning

- Tuning spaces discussed with learner authors
- **Resampling**: Nested cross-validation (5-fold outer, 3-fold inner)

Tuning

- Tuning spaces discussed with learner authors
- **Resampling**: Nested cross-validation (5-fold outer, 3-fold inner)
- **Strategy**: Random Search

Tuning

- Tuning spaces discussed with learner authors
- **Resampling**: Nested cross-validation (5-fold outer, 3-fold inner)
- **Strategy**: Random Search
- **Budget**: Tuning stopped if either is reached

Tuning

- Tuning spaces discussed with learner authors
- **Resampling**: Nested cross-validation (5-fold outer, 3-fold inner)
- **Strategy**: Random Search
- **Budget**: Tuning stopped if either is reached
    1. Number of evaluations: $n_{\text{evals}} = n_{\text{parameters}} \times 50$

Tuning

- Tuning spaces discussed with learner authors
- **Resampling**: Nested cross-validation (5-fold outer, 3-fold inner)
- **Strategy**: Random Search
- **Budget**: Tuning stopped if either is reached
    1. Number of evaluations: $n_{\text{evals}} = n_{\text{parameters}} \times 50$
    2. Tuning time of 150 hours ($6\frac{1}{4}$ days)

Evaluation

- Main Results:

Evaluation

- Main Results:
    - Friedman rank sum tests

Evaluation

- Main Results:
  - Friedman rank sum tests
  - Critical difference plots[2] based on Bonferroni-Dunn tests

[2] Demšar (2006)

Evaluation

- Main Results:
  - Friedman rank sum tests
  - Critical difference plots[2] based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, Scoring Rules

[2] Demšar (2006)

Evaluation

- Main Results:
  - Friedman rank sum tests
  - Critical difference plots[2] based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, Scoring Rules
- Tuned on 2 different measures

[2] Demšar (2006)

Evaluation

- Main Results:
  - Friedman rank sum tests
  - Critical difference plots[2] based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, Scoring Rules
- Tuned on 2 different measures
  - Harrell's C (Discrimination)

[2] Demšar (2006)

Evaluation

- Main Results:
  - Friedman rank sum tests
  - Critical difference plots[2] based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, Scoring Rules
- Tuned on 2 different measures
  - Harrell’s C (Discrimination)
  - Right-Censored Log Loss (Scoring Rule)

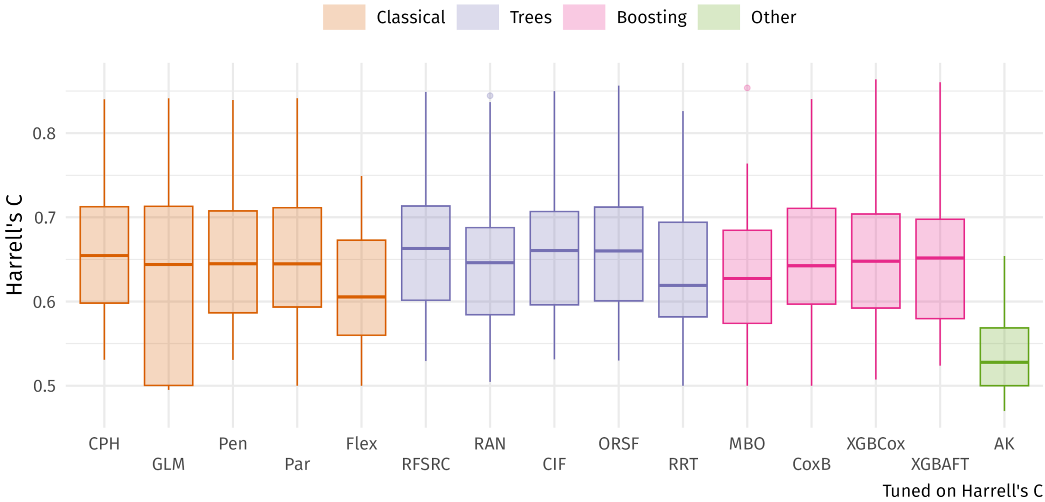
[2] Demšar (2006)

Evaluation

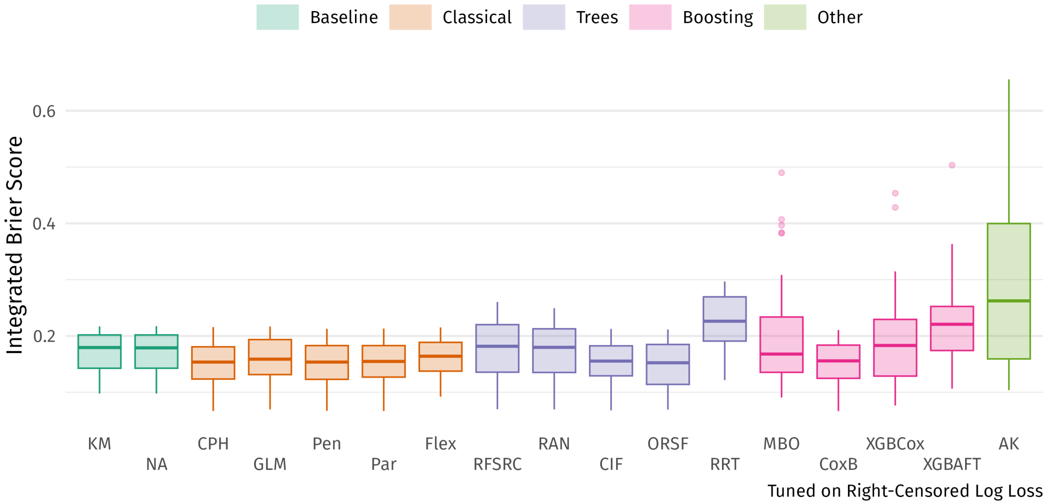
- Main Results:
  - Friedman rank sum tests
  - Critical difference plots[2] based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, Scoring Rules
- Tuned on 2 different measures
  - Harrell's C (Discrimination)
  - Right-Censored Log Loss (Scoring Rule)
- Evaluation spans all 3 types

[2] Demšar (2006)

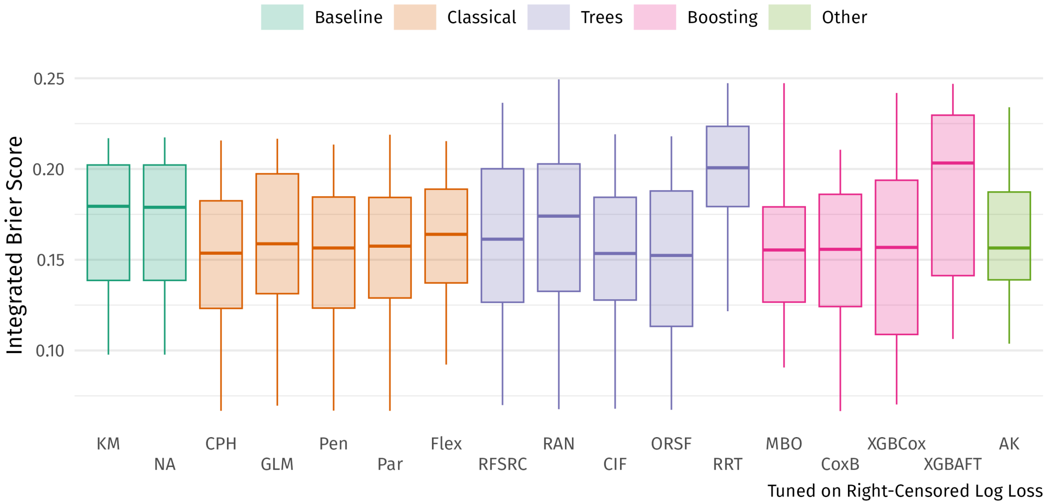
Boxplot (Harrel’s C, higher is better)



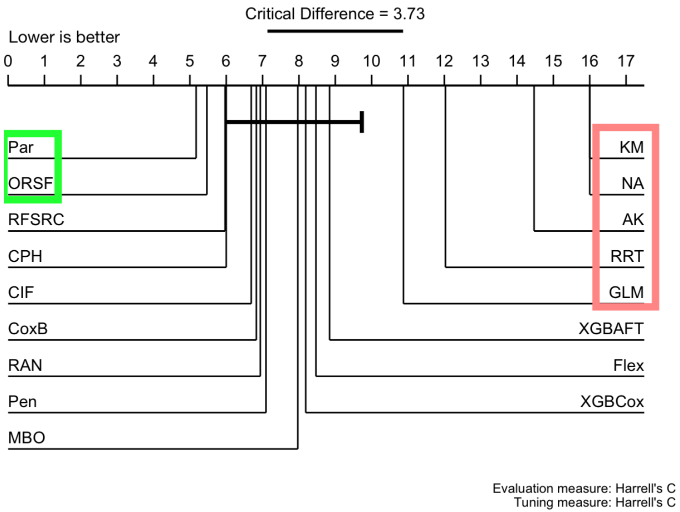
Boxplot (IBS, lower is better)



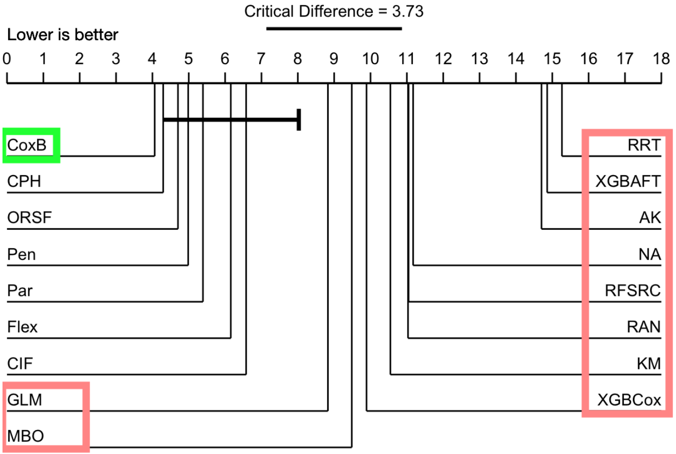
Boxplot (IBS, truncated)



Critical Difference: Bonferroni-Dunn (Harrell's C)



Critical Difference: Bonferroni-Dunn (IBS/RCLL)



Evaluation measure: Integrated Brier Score (Improper)
Tuning measure: Right-Censored Log Loss

Closing Remarks

- Experimental design is not perfect, but it was possible to conduct

Closing Remarks

- Experimental design is not perfect, but it was possible to conduct
- Only computationally feasible due to resources of ARCC [3]

[3] Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

Closing Remarks

- Experimental design is not perfect, but it was possible to conduct
- Only computationally feasible due to resources of ARCC [3]
  - Sequential runtime: $\approx$ 18 years

[3] Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

Closing Remarks

- Experimental design is not perfect, but it was possible to conduct
- Only computationally feasible due to resources of ARCC [3]
  - Sequential runtime: $\approx$ 18 years
  - Effective runtime: 32 days ($\approx$ 200x decrease)

[3] Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

Closing Remarks

- Experimental design is not perfect, but it was possible to conduct
- Only computationally feasible due to resources of ARCC [3]
  - Sequential runtime: $\approx$ 18 years
  - Effective runtime: 32 days ($\approx$ 200x decrease)
- Results still need processing, checking, ...

[3] Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

Closing Remarks

- Experimental design is not perfect, but it was possible to conduct
- Only computationally feasible due to resources of ARCC [3]
  - Sequential runtime: $\approx$ 18 years
  - Effective runtime: 32 days ($\approx$ 200x decrease)
- Results still need processing, checking, ...
- **Preliminary conclusion**: Cox regression — hard to beat since 1972!

[3] Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

Closing Remarks

- Experimental design is not perfect, but it was possible to conduct
- Only computationally feasible due to resources of ARCC [3]
  - Sequential runtime: $\approx$ 18 years
  - Effective runtime: 32 days ($\approx$ 200x decrease)
- Results still need processing, checking, ...
- **Preliminary conclusion**: Cox regression — hard to beat since 1972!
- The "standard setting" $\approx$ the "do you need ML?"-setting

[3] Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

Thank you for your attention!

www.leibniz-bips.de/en

Contact
Lukas Burk
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Achterstraße 30
D-28359 Bremen
burk@leibniz-bips.de

References I

Demšar, Janez (2006). "Statistical comparisons of classifiers over multiple data sets". In: Journal of Machine learning research 7.1, pp. 1–30.

Lang, Michel et al. (2019). "mlr3: A modern object-oriented machine learning framework in R". In: Journal of Open Source Software 4.44, p. 1903. DOI: 10.21105/joss.01903.