

# A Large-Scale Neutral Comparison Study of Survival Models

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

<sup>1</sup>Leibniz Institute for Prevention Research and Epidemiology – BIPS

<sup>2</sup>LMU Munich   <sup>3</sup>University of Bremen

<sup>4</sup>Munich Center for Machine Learning (MCML)

<sup>5</sup>University of Oslo   <sup>6</sup>TU Dortmund

<sup>7</sup>OSPO Now   <sup>8</sup>Imperial College, London

Biometric Colloquium — March 1st, 2024

# Introduction

---



1

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

# Introduction

---



1

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

# Introduction

---



1

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

# Introduction

---



1

- 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

---



1

- 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

---



1

- 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

---



1

- 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

---



1

- 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

---



2

- 32 tasks
- 18 learners
- 2 tuning measures
- 9 evaluation measures
  
- **Large-scale**  $\Rightarrow$  Generalizability
- **Neutral**  $\Rightarrow$  Fair comparison

## Quick Summary

---



2

- 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

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

---



4

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

---



5

18 learners implemented in R and available via the `mlr3`<sup>1</sup> framework

---

<sup>1</sup>Lang et al. (2019)

# Learners

---



5

18 learners implemented in R and available via the `mlr3`<sup>1</sup> framework

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

---

<sup>1</sup>Lang et al. (2019)



# Learners

---



5

18 learners implemented in R and available via the `mlr3`<sup>1</sup> framework

- **Baseline:** Kaplan-Meier & Nelson-Aalen
- **Classical:** Cox, penalized (L1,L2), parametric (AFT)

---

<sup>1</sup>Lang et al. (2019)

# Learners

---



5

18 learners implemented in R and available via the `mlr3`<sup>1</sup> framework

- **Baseline:** Kaplan-Meier & Nelson-Aalen
- **Classical:** Cox, penalized (L1,L2), parametric (AFT)
- **Trees:** Individuals and ensembles

---

<sup>1</sup>Lang et al. (2019)

# Learners

---



5

18 learners implemented in R and available via the `mlr3`<sup>1</sup> framework

- **Baseline:** Kaplan-Meier & Nelson-Aalen
- **Classical:** Cox, penalized (L1,L2), parametric (AFT)
- **Trees:** Individuals and ensembles
- **Boosting:** Gradient- and likelihood-based

---

<sup>1</sup>Lang et al. (2019)

# Learners

---



5

18 learners implemented in R and available via the `mlr3`<sup>1</sup> framework

- **Baseline:** Kaplan-Meier & Nelson-Aalen
- **Classical:** Cox, penalized (L1,L2), parametric (AFT)
- **Trees:** Individuals and ensembles
- **Boosting:** Gradient- and likelihood-based
- **Other:** Akritas, SVM

---

<sup>1</sup>Lang et al. (2019)

## List of Learners (Baseline, Classical)



6

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
Akritis	AK	survivalmodels
Survival SVM	SSVM	survivalsvm

# List of Learners (Trees, Boosting)



7

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)	XGB Cox	xgboost
Gradient Boosting (AFT objective)	XGB AFT	xgboost

# Tuning

---



8

- Tuning spaces discussed with learner authors

# Tuning

---



8

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



# Tuning

---



8

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

# Tuning

---



8

- 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

---



8

- 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

---



8

- 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

---



9

- Main Results:

# Evaluation

---



9

- Main Results:
  - Friedman rank sum tests

# Evaluation

---



9

- Main Results:
  - Friedman rank sum tests
  - Critical difference plots<sup>2</sup> based on Bonferroni-Dunn tests

---

<sup>2</sup>Demšar (2006)

# Evaluation

---



9

- Main Results:
  - Friedman rank sum tests
  - Critical difference plots<sup>2</sup> based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, [Scoring Rules](#)

---

<sup>2</sup>Demšar (2006)



# Evaluation

---



9

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

---

<sup>2</sup>Demšar (2006)

# Evaluation

---



9

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

---

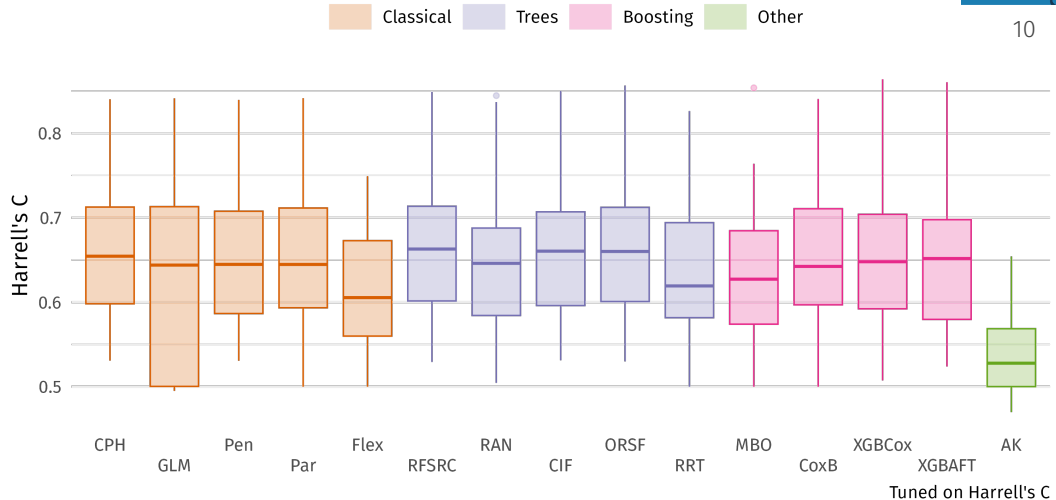
<sup>2</sup>Demšar (2006)

- Main Results:
  - Friedman rank sum tests
  - Critical difference plots<sup>2</sup> 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)

---

<sup>2</sup>Demšar (2006)

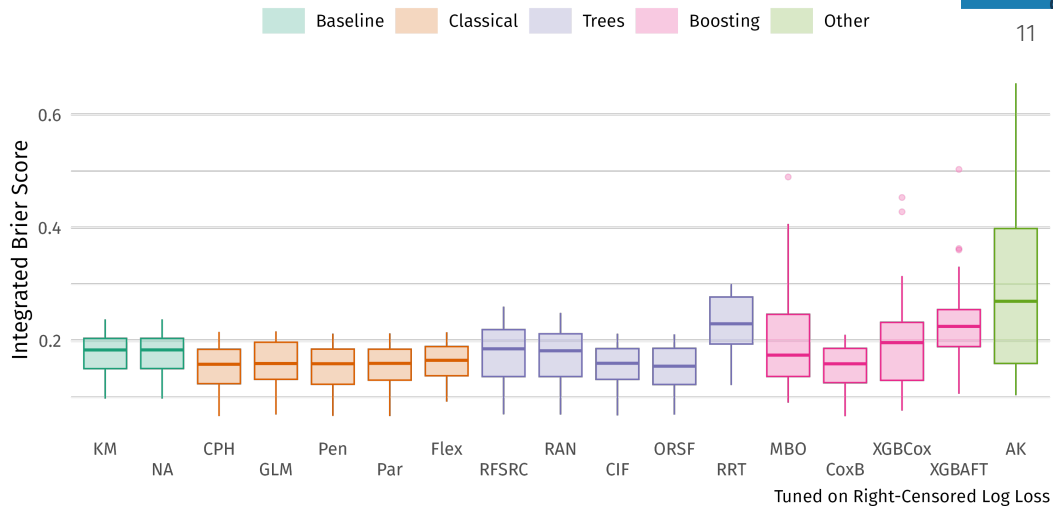
# Boxplot (Harrel's C, higher is better)



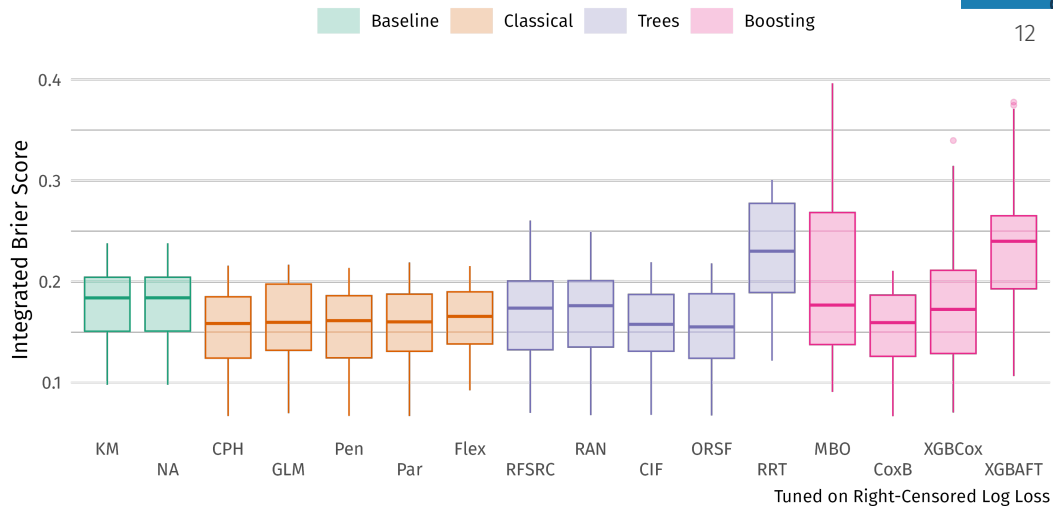
## Boxplot (IBS, lower is better)



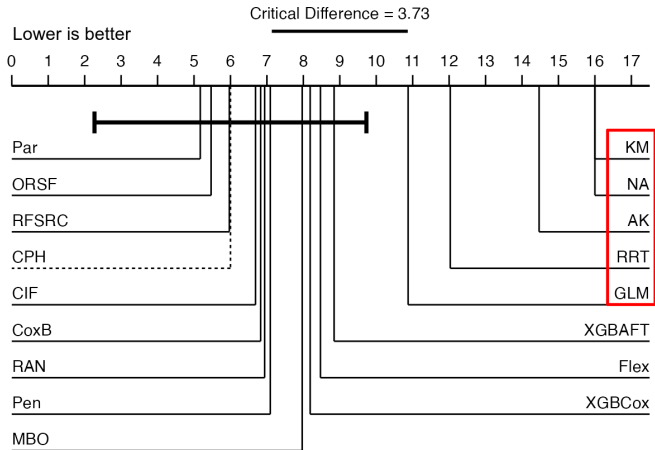
11



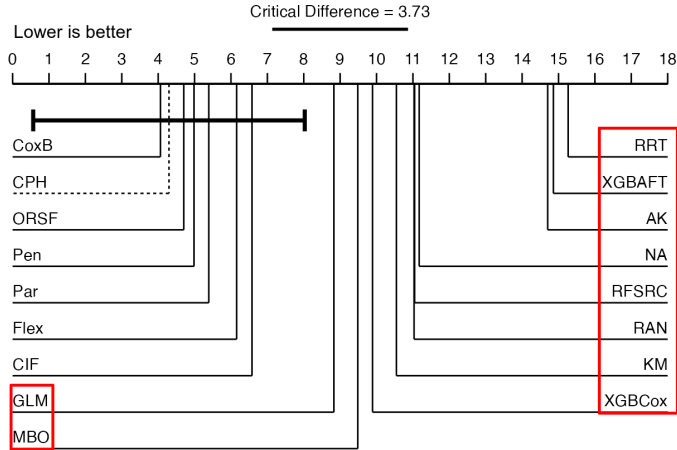
# Boxplot (IBS, truncated)



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



# Critical Difference: Bonferroni-Dunn (IBS/RCLL)





## Closing Remarks

---



15

- Only computationally feasible due to resources of ARCC<sup>3</sup>

---

<sup>3</sup>Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

## Closing Remarks

---



15

- Only computationally feasible due to resources of ARCC<sup>3</sup>
  - Sequential runtime  $\approx 18$  years

---

<sup>3</sup>Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

## Closing Remarks

---



15

- Only computationally feasible due to resources of ARCC<sup>3</sup>
  - Sequential runtime  $\approx 18$  years
  - Effective runtime  $\approx 32$  days

---

<sup>3</sup>Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

## Closing Remarks

---



15

- Only computationally feasible due to resources of ARCC<sup>3</sup>
  - Sequential runtime  $\approx 18$  years
  - Effective runtime  $\approx 32$  days
- Experimental design is not perfect, but it was possible to conduct

---

<sup>3</sup>Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

## Closing Remarks

---



15

- Only computationally feasible due to resources of ARCC<sup>3</sup>
  - Sequential runtime  $\approx 18$  years
  - Effective runtime  $\approx 32$  days
- Experimental design is not perfect, but it was possible to conduct
- Results still need processing, checking, ...

---

<sup>3</sup>Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

## Closing Remarks

---



15

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

---

<sup>3</sup>Advanced Research Computing Center, Beartooth Computing Environment, University of Wyoming.

Thank you for your attention!



[www.leibniz-bips.de/en](http://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](mailto:burk@leibniz-bips.de)





## References I

---



17

-  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](#).