

A Large-Scale Neutral Comparison Study of Survival Models on Low-Dimensional Data

Burk, L.^{1,2,3,4} Zobolas, J.⁵ Bischl, B.^{2,4} Bender, A.^{2,4} Wright, M. N.^{1,3} Sonabend, R.^{6,7}

¹Leibniz Institute for Prevention Research and Epidemiology – BIPS

²LMU Munich ³University of Bremen

⁴Munich Center for Machine Learning (MCML)

⁵Institute for Cancer Research, Oslo

⁶OSPO Now ⁷Imperial College, London

July 29rd, 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
- 8 evaluation measures

Quick Summary



2

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

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

Quick Summary



2

- 32 tasks
- 18 learners
- 2 tuning measures
- 8 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	3	5	7	25
Observed Events	101	194	336	1034	5616
Cens. %	6	32	48	74	95

Learners



5

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

¹Lang et al. (2019)

Learners



5

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

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

¹Lang et al. (2019)

Learners



5

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

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

¹Lang et al. (2019)

Learners



5

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

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

¹Lang et al. (2019)

Learners



5

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

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

¹Lang et al. (2019)

Learners



5

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

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

¹Lang et al. (2019)

List of Learners (Baseline, Classical)



6

Name	Abbreviation	Package
Kaplan-Meier	KM	survival
Nelson-Aalen	NA	survival
Akritas	AK	survivalmodels
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
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** of two criteria 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** of two criteria 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** of two criteria 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)

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** of two criteria 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)
- **Fallback:** Impute result with KM

“Well, technically...”



9

Exceptions to the previously stated rules:

“Well, technically...”



9

Exceptions to the previously stated rules:

- Some learners (RRT, Par) have small, finite search spaces \Rightarrow exhaustive grid search

“Well, technically...”



9

Exceptions to the previously stated rules:

- Some learners (RRT, Par) have small, finite search spaces \Rightarrow exhaustive grid search
- Task **veteran** has so few observations \Rightarrow 4 outer resampling folds, ensuring min. 30 observed events per outer fold

“Well, technically...”



9

Exceptions to the previously stated rules:

- Some learners (**RRT**, **Par**) have small, finite search spaces \Rightarrow exhaustive grid search
- Task **veteran** has so few observations \Rightarrow 4 outer resampling folds, ensuring min. 30 observed events per outer fold
- **CoxBoost** learner tunes itself with internal CV \Rightarrow set to use 3 folds as well

“Well, technically...”



9

Exceptions to the previously stated rules:

- Some learners (**RRT**, **Par**) have small, finite search spaces \Rightarrow exhaustive grid search
- Task **veteran** has so few observations \Rightarrow 4 outer resampling folds, ensuring min. 30 observed events per outer fold
- **CoxBoost** learner tunes itself with internal CV \Rightarrow set to use 3 folds as well
- We tune **cv.glmnet** for **alpha**, while it tunes itself for **lambda**

Evaluation



10

- Main Results:

Evaluation



10

- Main Results:
 - Friedman rank sum tests

Evaluation



10

- Main Results:
 - Friedman rank sum tests
 - Critical difference plots² based on Bonferroni-Dunn tests

²Demšar (2006)

Evaluation



10

- Main Results:
 - Friedman rank sum tests
 - Critical difference plots² based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, [Scoring Rules](#)

²Demšar (2006)

Evaluation



10

- Main Results:
 - Friedman rank sum tests
 - Critical difference plots² based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, [Scoring Rules](#)
- Tuned on 2 different measures

²Demšar (2006)

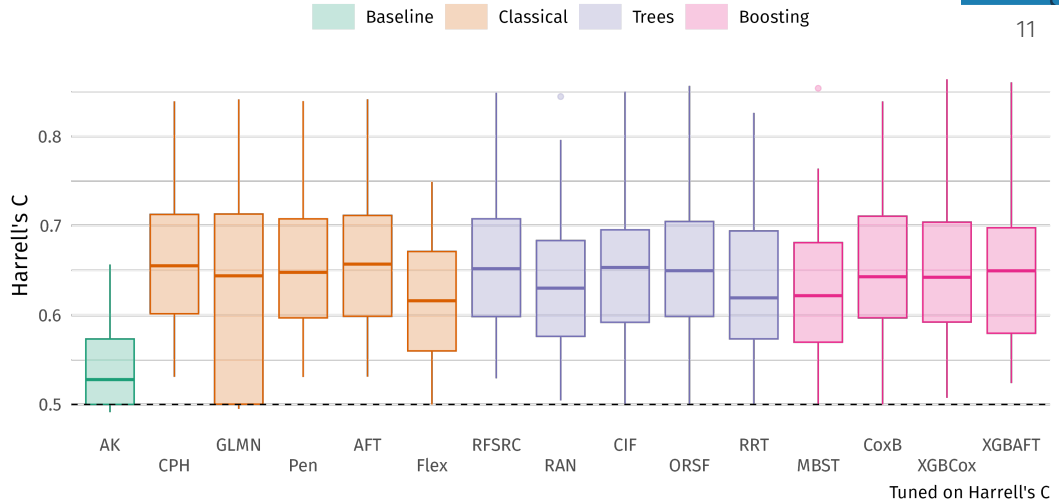
- Main Results:
 - Friedman rank sum tests
 - Critical difference plots² based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, [Scoring Rules](#)
- Tuned on 2 different measures
 - Harrell's C (Discrimination)

²Demšar (2006)

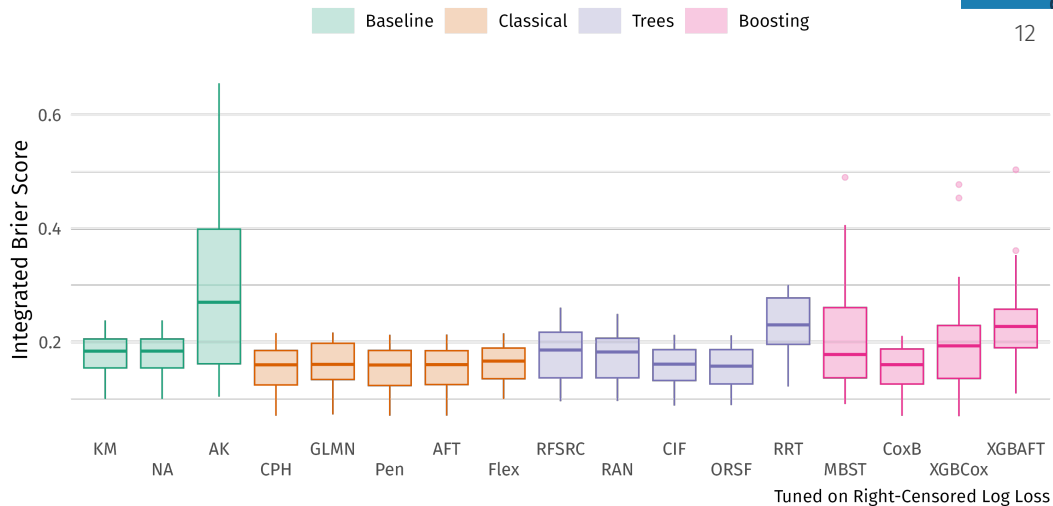
- Main Results:
 - Friedman rank sum tests
 - Critical difference plots² based on Bonferroni-Dunn tests
- 3 types of metrics: Discrimination, Calibration, [Scoring Rules](#)
- Tuned on 2 different measures
 - Harrell's C (Discrimination)
 - Right-Censored Log Loss (Scoring Rule)

²Demšar (2006)

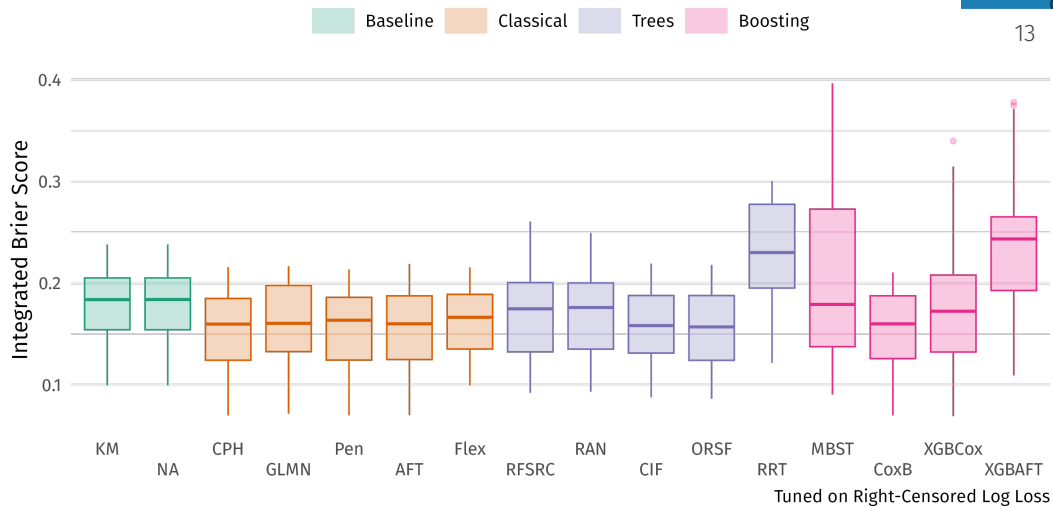
Boxplot (Harrel's C, higher is better)



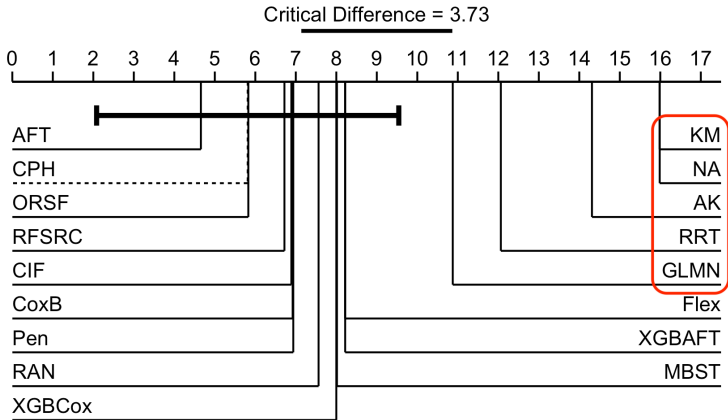
Boxplot (IBS, lower is better)



Boxplot (IBS, truncated)

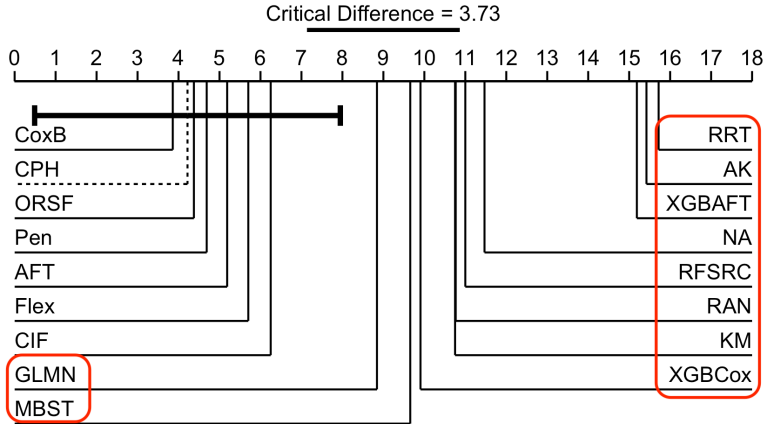


Critical Difference: Bonferroni-Dunn (Harrell's C)



Evaluation measure: Harrell's C
Tuning measure: Harrell's C

Critical Difference: Bonferroni-Dunn (IBS/RCLL)



Evaluation measure: Integrated Survival Brier Score (ISBS)
Tuning measure: Right-Censored Log Loss (RCLL)

Closing Remarks



16

- Only computationally feasible due to resources of ARCC³

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

Closing Remarks



16

- Only computationally feasible due to resources of ARCC³
 - Sequential runtime \approx 18 years

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

Closing Remarks



16

- Only computationally feasible due to resources of ARCC³
 - Sequential runtime ≈ 18 years
 - Effective runtime (incl reruns) ≈ 32 days

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

Closing Remarks



16

- Only computationally feasible due to resources of ARCC³
 - Sequential runtime ≈ 18 years
 - Effective runtime (incl reruns) ≈ 32 days
- Experimental design is not perfect, but it was possible to conduct

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

Closing Remarks



16

- Only computationally feasible due to resources of ARCC³
 - Sequential runtime \approx 18 years
 - Effective runtime (incl reruns) \approx 32 days
- Experimental design is not perfect, but it was **possible** to conduct
- **Conclusion:** Cox regression — hard to beat since 1972!

³Advanced 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



18

-  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](#).