

Beyond Cox Models: Assessing the Performance of Machine-Learning Methods in Non-Proportional Hazards and Non-Linear Survival Analysis

Ivan Rossi¹, Flavio Sartori¹, Cesare Rollo¹, Giovanni Birolo^{1,*†},
Piero Fariselli^{1,*}, and Tiziana Sanavia^{1,*}

¹Department of Medical Sciences, University of Turin, Turin, Italy

^{*}These authors jointly supervised this work.

[†]Corresponding author: giovanni.birolo@unito.it

Apr 2025

1 Abstract

Survival analysis often relies on Cox models, assuming both linearity and proportional hazards (PH). This study evaluates machine and deep learning methods that relax these constraints, comparing their performance with penalized Cox models on a benchmark of three synthetic and three real datasets. In total, eight different models were tested, including six non-linear models of which four were also non-PH. Although Cox regression often yielded satisfactory performance, we showed the conditions under which machine and deep learning models can perform better. Indeed, the performance of these methods has often been underestimated due to the improper use of Harrell's concordance index (C-index) instead of more appropriate scores such as Antolini's concordance index, which generalizes C-index in cases where the PH assumption does not hold. In addition, since occasionally high C-index models happen to be badly calibrated, combining Antolini's C-index with Brier's score is useful to assess the overall performance of a survival method. Results on our benchmark data showed that survival prediction should be approached by testing different methods to select the most appropriate one according to sample size, non-linearity and non-PH conditions. To allow an easy reproducibility of these tests on our benchmark data, code and documentation are freely available at <https://github.com/compbio-med-unito/survhive>.

2 Introduction

Survival analysis represents a collection of statistical and machine learning procedures where the outcome variable of interest is a time-to-event, representing not only whether or not an event occurred, but also when that event occurred, which is common in several contexts, ranging from component failure in engineering to adverse events in medicine. The prediction of these time-to-event cases makes survival analysis much more challenging than other analyses like regression or classification. Another important aspect of survival analysis is the presence of censored cases, i.e. when the observed event does not occur during the study’s timeframe (for example, death or disease relapse after the end of a clinical study).

The Cox proportional hazards (CoxPH) model [5] is by far the most well known and widely used statistical approach in survival analysis, which extends linear regression to survival outcomes and assumes that the hazard functions of all individuals are proportional (PH), greatly simplifying the fit of the model by turning it into a regression problem. Furthermore, CoxPH often achieves high predictive performance, showing that it is reasonably robust to minor deviations from its underlying assumptions. Moreover, being also directly interpretable, it is not surprising that it has remained the most popular model in survival analysis for the last fifty years since its introduction.

However, datasets that are strongly non-linear and/or non-PH do exist and this motivated the development of new methods based on more recent techniques. Thus, successful machine learning methods were adapted to survival analysis, such as CoxNet [22], a version of the CoxPH regression with Elastic-Net regularization, Random Survival Forests (RSF) [11] and Gradient Boosting Survival Analysis (GrBSA) [10] models, based on Random Forests and Gradient Boosting algorithms, respectively. More recently, many deep learning survival methods have been developed to address this task. The first has been DeepSurv [12], a straightforward deep learning extension of CoxPH allowing non-linearity while preserving the PH assumption and the Cox loss. Other deep versions of the CoxPH model have been introduced after DeepSurv, for instance FastCPH [26], an extension of the LassoNet[15] method to survival outcomes. DeepHit [14] takes a more radical departure from CoxPH, predicting a discrete probability distribution on binned event times, avoiding the PH assumption and other parametric distributions. Deep Survival Machines [16] aim to model the underlying event outcome distribution as a mixture of some fixed parametric (Weibull or Log-normal) distributions, whose parameters, as well as the mixing weights, are modeled using deep neural networks. Another survival model based on deep learning, SurvTRACE [25], employs a transformer that incorporates an attention mechanism instead of the more traditional network architectures used by the previous methods.

Machine learning and especially deep learning survival predictors advertise themselves as being more powerful than CoxPH, claiming to provide more accurate predictions on real datasets, when linearity and PH assumptions are not guaranteed. However, they also present their own disadvantages, ranging from

their substantially larger computational costs to the need for more data and the requirement for complex training strategies, often coupled with a reduced interpretability of the results.

Another critical aspect affecting survival analysis is the choice of the evaluation metrics for predictive models. The Concordance Index (C-index) is one of the most used metrics, that quantifies the rank correlation between the actual survival times and the predictions of a model, accounting for censoring. Multiple estimators have been proposed, but the popularity of the PH assumption led to the adoption of Harrell’s C-index [9] as the de-facto standard metric. It measures the ability of a survival model to order subjects by estimating the proportion of correctly ordered pairs of individuals among all comparable pairs in the dataset. Individuals with higher risk scores have a higher probability of an event at every time value. This means that the ranking is fixed between individuals and is the same at all times. However, in non-PH models, the ranking of individuals is not necessarily fixed and can change depending on the time, i.e. the survival curves predicted for two individuals can cross. In these cases, a generalization of Harrell’s C-index must be used to account for the rank dependence on the time. To this aim, Antolini’s C-index [1] has been introduced as an appropriate evaluation metric for non-PH models. Nevertheless, Harrell’s C-index has been used, incorrectly, to evaluate non-PH models, often trying to partially account for the time dependence of ranking by averaging the C-index scores computed from the survival probability at several fixed times.

Therefore, we decided to better investigate advantages and drawbacks of the most recent survival analysis methods by performing their systematic comparison on six different datasets, including three synthetic datasets designed to violate some (or all) of CoxPH’s base assumptions and three real clinical datasets. We also explored how the choice of the evaluation metric affects the results. Our study provides insights into the conditions under which machine learning approaches can outperform CoxPH, highlighting the importance of the type of data in determining the most suitable survival analysis methodology. To allow an easy reproducibility of our analyses on different datasets, we organized the code used in this work into the SurvHive package [2], a Python-based framework developed to provide a unified and consistent API for executing several survival analysis methods, which often differ in input/output formats and capabilities. Compatible with both *scikit-learn* [18, 4] and *scikit-survival* [19] libraries, it standardizes data handling, supports hyperparameter tuning, time-dependent risk evaluation, and cross-validation for censored data. By using adapters instead of re-implementing methods, SurvHive minimizes bias and preserves the integrity of original models. Code and documentation are freely available at <https://github.com/comphiomed-unito/survhive>.

3 Methods

3.1 Survival methods

We tested eight different survival methods. Four of them are from the *scikit-survival* library: the classical CoxPH [24] and the machine learning-based CoxNet [22] (which adds the ElasticNet regularization to CoxPH), GrBSA [10] (GradientBoostingSurvivalAnalysis using the CoxPH loss and thus also assuming PH) and RSF [11] (RandomSurvivalForest, a non-PH method).

The other four deep learning methods have different sources:

- DeepHit [14], implemented by the DeepHitSingle class from *PyCox* [13] package;
- Deep Survival Machines (DSM) [16], from *Auton Survival* [17] package;
- FastCPH [26], from *LassoNet* [15] package;
- SurvTRACE, implemented by the SurvTRACESingle class from *SurvTRACE* [25] package.

3.2 Datasets

In order to compare the methods, we used six different datasets. Three of them are synthetic datasets specifically designed to test the limits of the Cox-based methods, while the others are real data coming from clinical studies.

3.2.1 Synthetic datasets

We generated the three synthetic datasets following different assumptions:

LinPH : proportional hazards from a linear combination of features,

NonLinPH : proportional hazards from a non-linear function of the features,

NonPH : non proportional hazards.

The second and third datasets violate one or both CoxPH assumptions and are meant to stress-test the methods, in order to better characterize their limitations.

All three datasets have 20 features x_1, \dots, x_{20} , whose value is randomly sampled from a normal distribution. Events are randomly sampled from a $[0, 10]$ time interval according to the survival function estimated from the features.

Both LinPH and NonLinPH generative models have a baseline hazard that is generated from random noise in a bounded time interval. Specifically, LinPH model computes the risk from a linear combination of the first 8 features with random coefficients, while NonLinPH model uses a fixed function of the first 4 features: $x_1 + x_2 x_3 + \cos(6x_4)$. Finally, NonPH model splits the time interval into 16 same-sized intervals and yields survival functions that are piecewise constant

on the intervals. The probability p_i of an event for each interval is computed using the soft-max function on the first 16 features (one for each time interval) multiplied by 16: $p_i = \frac{\exp(16x_i)}{\sum_{j=1}^{16} \exp(16x_j)}$.

The synthetic datasets were generated with the goal of strongly violating the linearity and PH assumptions while also being highly predictable, meaning that the event time is mostly determined by the features and not overly affected by the random sampling. In this way, a good model will be able to achieve high predictive performance, allowing us to discriminate better between the different models.

3.2.2 Clinical datasets

In addition to the synthetic datasets described above, we also considered three publicly available survival datasets, commonly used to assess the performance of predictive models:

FLChain : the Assay of Serum Free Light Chain [23], including 6,524 samples and 8 features;

GBSG2 : the Rotterdam tumour bank and German Breast Cancer Study Group [7, 21], including 3,668 samples and 8 features;

METABRIC : the Molecular Taxonomy of Breast Cancer International Consortium [6], including 1,903 samples and 9 features.

FLChain and METABRIC datasets were preprocessed according to Kvamme et al. [13], while for GBSG2 the same preprocessing schema of Royston and Altman [20] was applied, with the only difference that the data from the two studies were merged to create a single dataset.

3.3 Metrics

Model performance was evaluated using all the most commonly used metrics in survival analysis:

1. Harrell’s C-index [9], calculated at three time quartiles (25%, 50%, 75%);
2. Antolini’s Concordance Index [1], the extension of Harrell’s C-index for models with time-dependent risk ranking;
3. Brier score [3], calculated at multiple time quantiles;
4. Area under the receiver operator curve (AUROC) at multiple time quantiles.

Note that Harrell’s C-index is not appropriate for models where the risk ranking between individuals is not fixed (that is, when survival curves can cross each other). All PH models have fixed ranking, but for non-PH models this is generally not the case. However, we computed it for all models to show the impact of using Harrell’s C-index instead of Antolini’s when the comparison includes non-PH models.

3.4 Testing

We used a nested cross-validation approach for the hyper-parameter optimization and the model evaluation. For each dataset, model performance was estimated using an outer 3-fold cross-validation. For each fold, independent hyper-parameter optimization was performed on the training set with a full grid search using an inner 5-fold cross-validation, which was repeated twice with reshuffling, for a total of 10 validation sets. The hyper-parameters’ combination yielding the highest average Antolini’s C-index was then selected to refit the model on the whole training set. Finally, the fitted model was evaluated on a hold-out set for each of the three test sets of the outer cross-validation.

All computations were performed on a virtual machine with 12 Intel Xeon Skylake processors and 1 Nvidia Tesla T4 GPU. The parameter-optimization runs were performed using 12 concurrent processes for the non-GPU methods and a single process for the GPU-accelerated ones.

4 Results

Table 1 reports the tested survival methods, highlighting whether they follow one or both fundamental CoxPH assumptions, i.e. the hazard functions of all individuals are proportional (PH assumption) and the proportionality is computed through a linear combination of the features (linearity assumption).

Method	Linearity	PH
CoxPH/CoxNet	Yes	Yes
GrBSA	No	Yes
RSF	No	No
FastCPH	No	Yes
DeepHit	No	No
DeepSurvivalMachines (DSM)	No	No
SurvTRACE	No	No

Table 1: **Tested methods and their characterization.**

The three synthetic datasets satisfy either both assumptions (LinPH), or only PH (NonLinPH), or none of them (NonPH), providing a “controlled” environment to evaluate the methods. On the other hand, checking the validity of these assumptions on the real datasets is not straightforward, as we do not directly observe the complete survival distribution of each patient but only the final event time. Therefore, we used the synthetic datasets to guide the interpretation of the results observed in the clinical datasets FLChain, GBSG2, and Metabric.

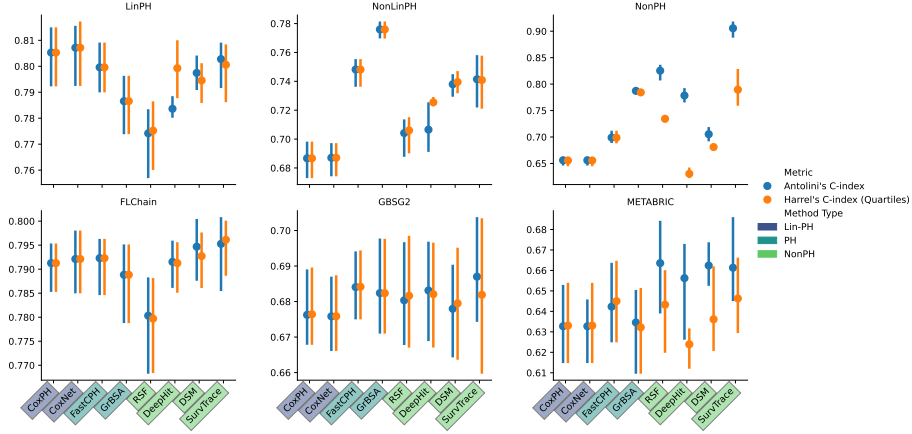


Figure 1: **C-index performance across all methods and datasets.** Error bars are displayed for both Harrell’s (orange bars) and Antolini’s (blue bars) C-indices. Specifically, each dot represents the average C-index, while the bars show the C-index range resulting from the 3-fold cross-validation. Methods’ assumptions are highlighted with different colors in the horizontal labels.

4.1 Concordance index

For each dataset, the models’ performance has been evaluated three times with a 3-fold cross-validation. As discussed above, Harrell’s C-index was developed in a PH context, assuming a fixed risk ranking between individuals. Unfortunately, it has often been incorrectly used for evaluating non-PH models, often taking its average at multiple times in an attempt to account for the rank dependence upon time. However, this is not enough: for comparisons including non-PH models, Antolini’s C-index should be used, since it generalizes Harrell’s C-index to models without fixed risk ranking. To show the bias introduced by this misuse, we computed both Antolini’s and Harrell’s C-indices (with Harrell’s computed as an average on the first, second and thirds quartiles of event times). The results for all datasets and methods are plotted in Figure 1 and full data is reported in Supp. Table 1.

On the synthetic LinPH dataset, CoxPH and CoxNet methods are the top performers as expected. In principle, other more powerful models should be able to reach similar scores, but only FastCPH and SurvTRACE show close performance to the Cox-based methods. Both Harrell’s and Antolini’s C-indices are quite similar for all methods in this dataset, except for DeepHit, whose predictions seem to deviate significantly from PH.

In the NonLinPH dataset, non-linear methods achieved much higher performance than the linear CoxPH and CoxNet. Here, the best performing methods are FastCPH and GrBSA, which are PH methods with non-linear risks, with GrBSA showing the highest performance. FastCPH performed only slightly better than DSM and SurvTRACE. DeepHit and RSF do better than the linear

models too.

The gap in performance among the methods becomes very large on the synthetic NonPH dataset. Both CoxPH and CoxNet achieved an average C-index of 0.66, with no significant differences between Harrell’s and Antolini’s metrics. Despite the assumptions of these two methods are strongly violated in this dataset, they are still able to provide some prediction power. On the other hand, SurvTRACE, which does not make any assumption, was able to achieve an Antolini’s C-index of 0.9. It is also worth highlighting that GrBSA, which assumes PH, almost reached an Antolini’s C-index of 0.8, showing that some PH methods can still perform well even when PH does not hold. In this dataset, we were also able to observe large differences between the C-indices for the non-PH methods, showing a Harrell’s C-index much lower than Antolini’s metric. According to our expectations, Harrell’s C-index strongly penalizes non-PH methods, making the performance of SurvTRACE seem comparable to those of GrBSA, while Antolini’s shows a 0.1 of advantage for SurvTRACE.

Turning our attention to the real clinical datasets, almost all tested methods achieved high C-indices on the FLChain dataset, without significant differences between Harrell’s and Antolini’s metrics, suggesting a behavior similar to that in the LinPH synthetic dataset, except for RSF, which performed visibly worse, albeit by a small margin. Even if both SurvTRACE and DSM appear to be on average slightly better than the others, the good performance of both CoxPH and CoxNet suggests that this dataset does not heavily deviate from linearity and PH assumptions.

Survival methods applied to GBSG2 clinical dataset performed almost all equally well, with a slight advantage for non-linear methods. Low differences between Harrell’s and Antolini’s C-index observed in non PH methods, i.e. FastCPH and GrBSA, suggest that the PH assumption does not seem restrictive in this dataset.

Conversely, in METABRIC the non-PH methods are the best performing, even though this advantage is appreciable only when using Antolini’s C-index, while Harrell’s C-index values are comparable to those observed for the other methods. PH methods performed noticeably worse, with a slight advantage for non-linear PH methods (FastCPH). We can conclude that PH is a limiting assumption in METABRIC and perhaps linearity too, at least to a small extent.

4.2 Calibration

The C-index, as a rank correlation, only evaluates the ability of a model to predict the order of the events. However, a high C-index does not guarantee that the model can also reliably predict the probability of observing the occurrence of an event before a given time. Indeed, arbitrarily rescaling the times (preserving the order) would not change the C-index at all. To evaluate the probability estimation, a model calibration is necessary. The most popular evaluation metric for calibration is the Brier score [8].

Figure 2 displays the Brier scores for all methods across the synthetic and real datasets, compared with the Antolini’s C-indices. Full data is available in

Supp. Table 1. Note that we rescaled the Brier score to allow for an easier comparison with the C-index (see figure legend for details).

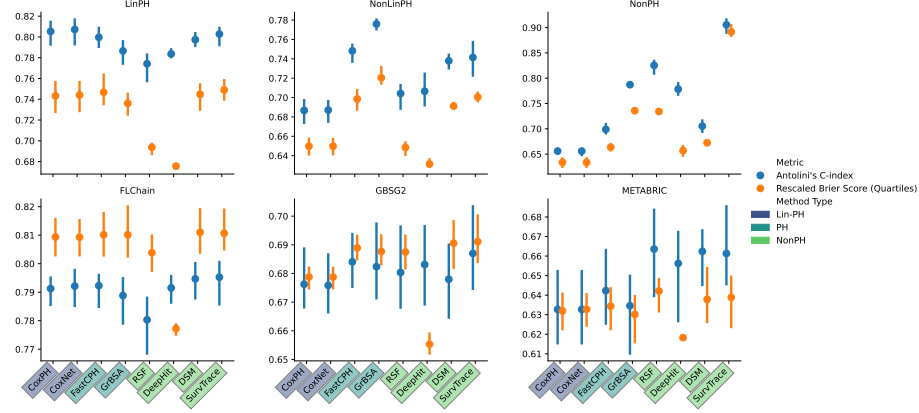


Figure 2: **Brier score performance performance across all methods and datasets.** Error bars are displayed for the rescaled Brier score (orange bars) and Antolini’s C-index (blue bars). Specifically, each dot represents the average score value, while the bars show the range of each score resulting from the 3-fold cross-validation. The original Brier score BS has been rescaled as $1 - 2BS$ to have the same range and direction of the C-index, allowing an easier visual comparison. Methods’ assumptions are highlighted with different colors in the horizontal labels.

Generally, we observed that models showing a good Antolini’s C-index tend to be more calibrated, with a Pearson Correlation of 0.87 between the two scores. The only systematic exception is DeepHit, that achieved consistently poor Brier scores across all the datasets, ranking last in all but one dataset. Indeed, if we exclude DeepHit, the Pearson correlation between Antolini’s C-index and Brier score increases to 0.92.

In most datasets method ranking is not overly affected by the choice of metrics, including the synthetic datasets, FLChain and GBSG2. However, there are some discrepancies between the two. In Metabric, the Brier score is flatter than the C-index, providing little discrimination between methods.

4.3 Time dependent ROC-AUC

Another commonly used metric is the time-dependent ROC-AUC, which is again based on ranking and is closely related to Antolini’s C-index. Indeed, the correlation between the two is very high (0.97) and thus using the ROC-AUC does not provide much additional information with respect to the C-index. Full data is available in Supp. Table 1.

4.4 Ablation study on the training sample size

In practical applications, sample size has also a crucial role in determining which method will yield the best survival model for a specific dataset. Therefore, we performed an ablation study calculating, for each method, the Antolini’s C-index at different subsamples of the synthetic datasets, considering a training sample size ranging between 300 and 2400 samples. In order to factor out the effect of event censoring, for each synthetic data set, the subsamples have been selected from the same pool of 10.000 points in such a way that each subsample contains the same fraction (30%) of censored events. Results for the three synthetic datasets are displayed in Figure 3 and reported in Supp. Table 2.

In general, it is reasonable to expect that the performance for all methods will increase at larger sample sizes, since the models are able to learn more complex relations between input features and outcomes. However, saturation effects might occur.

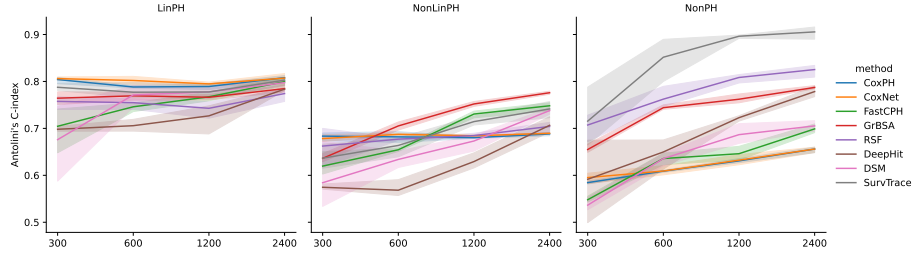


Figure 3: **Performance of methods at different subsamples of the synthetic datasets.** Y-axis displays the Antolini’s C-index, while the x-axis shows the sample size.

For the LinPH dataset, the best models are CoxPH and CoxNet and they achieved performance saturation at the lowest sample size of 300, showing that, if their assumptions hold, these models do not require many samples to achieve their best result. Other models like SurvTRACE and GrBSA also showed similar trends, even though at slightly lower C-index values. However, models like FastCPH, DSM and DeepHit performed significantly worse at 300 samples and improved their C-index at higher numbers, but never exceeding the performance obtained by CoxPH and CoxNet.

In the NonLinPH dataset, at 300 samples CoxPH and CoxNet were the best performers, while the other methods, theoretically better suited to the dataset, failed to learn the complex non-linear interactions between features. This changed at higher sample size, where the non-linear models performance continued to grow, overtaking the linear models that already had reached saturation. GrBSA only needed 600 samples to surpass CoxPH/CoxNet and become the best performing method, while FastCPH and SurvTRACE overcame the linear models at 1200 samples. DSM, RSF and DeepHit required at least 2400 samples to perform better than the linear approaches. Overall, it seems that the

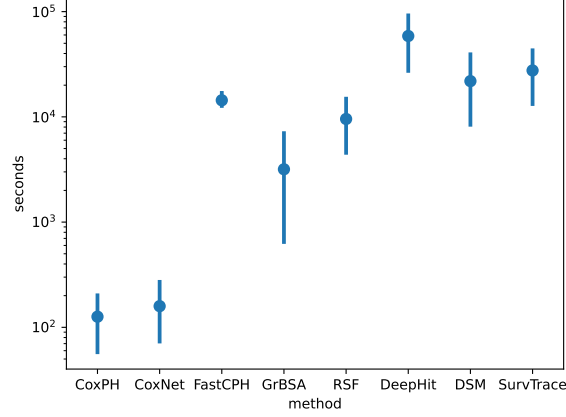


Figure 4: **Time elapsed for hyper-parameter optimization and training of each method.** The error bars represent the time elapsed in the three real clinical datasets. The vertical axis scale is logarithmic.

non-linear models did not achieve performance saturation even at the largest sample size of 2400.

In the NonPH dataset, SurvTRACE outperformed the other methods at all sample sizes. RSF achieved almost the same performance as SurvTRACE at 300 samples, but failed to keep the pace starting from 600 samples, remaining a distant second. Except for SurvTRACE, which appears to achieve saturation at 1200 samples, no other methods showed saturation in this dataset. Interestingly, even both CoxPH and CoxNet methods, whose assumptions are violated here, were able to improve their C-indices with increasing number of samples, differently from the other datasets. It is also notable that these linear models were able to outperform some non-linear and non-PH methods (FastCPH and DSM) at the 300-sample size.

4.5 Running time

Another factor to consider when applying machine-learning-based methods is the computational cost for training the models, in particular when hyper-parameter optimization and the consequent cross-validation are taken into account. Depending on the available computational resources, a long computational time for training can severely limit the optimization process, reducing the number of hyper-parameter combinations that can be tested. Figure 4 shows the total time required for hyper-parameter optimization and training of each method. As could be expected, CoxPH and CoxNet, as linear methods, are the fastest methods, requiring only a few minutes. The GrBSA and RSF are the fastest non-linear methods, taking less than one hour and up to three hours on FLChain

(the largest real dataset), respectively. The other deep learning methods are even slower, with DeepHit reaching 19 hours on FLChain. For all methods, the larger the dataset, the longer the computational time is, with the exception of FastCPH, which took roughly the same time on all three datasets.

5 Conclusions

Linearity and PH assumptions play a critical role in determining the best method for a given survival dataset. When they both hold, there seems to be no reason to go beyond a CoxPH regression, which provides the best performance, while also having the benefits of fast computational training times and a clear, statistically-sound interpretation. In addition, even when linearity and PH assumptions are violated, it can still provide good predictions proving to be remarkably robust, in particular at small sample sizes, where methods allowing for non-linearity and non-PH may not have enough data to converge to a better solution than CoxPH. Furthermore, CoxNet shows almost no differences in performance with respect to the CoxPH model, probably due to the high sample-size over feature-number ratio that makes regularization unnecessary even at lower sample sizes.

When enough samples are available, more complex methods may provide more accurate models when the Cox assumptions are not verified. Since there is no direct way to predict when this occurs, in practical terms we can only try to fit the models and check whether they perform better than Cox. When doing this, it is very important to use appropriate metrics for non-PH methods, such as Antolini’s C-index and Brier’s score to check the model calibration. Using the original Harrell’s metric to assess the C-index might be deleterious, since this metric implicitly assumes a constant risk ranking, a condition not necessarily satisfied by non-PH models. Even though in our study Antolini’s C-index and Brier’s score often showed concordance in ranking models, we recommend using both, as they evaluate different aspects and occasionally high C-index models happen to be badly calibrated.

Although no method emerged as the best performer in all datasets, we can make some general considerations. SurvTRACE consistently showed good- or high- performance across all datasets in both Antolini’s C-index and Brier’s score, even at lower sample sizes, never failing too badly. Considering its performance in the synthetic datasets, this method showed its capability to adapt to rather extreme conditions. DSM performed similarly to SurvTRACE in most tests at large dataset sizes, however it performed slightly worse at lower sample sizes. On the other hand, DeepHit never overcame the other methods in Antolini’s C-index, performing quite badly at lower sample sizes and in terms of Brier’s score, thus providing no advantages with respect to other methods. When considering the whole optimization time, DeepHit also resulted as the slowest method. Finally, non-linear PH methods FastCPH and GrBSA performed similarly, with GrBSA showing better C-indices in some datasets and requiring much less computational resources than FastCPH.

Since the best method in machine learning is always dependent on the dataset, testing different methods is the only way to know which one to use. From our study, we suggest to start testing with CoxNet, GrBSA and SurvTRACE, since other methods do not appear to provide compelling benefits over those three. To address this, we therefore included the benchmark data into the SurvHive package [2], providing support for hyperparameter tuning, time-dependent risk evaluation, and cross-validation for censored data.

6 Acknowledgements

The authors thank the Italian Ministry for Education, University and Research under the programme “Ricerca Locale ex-60%” and PNRR M4C2 HPC -1.4 “CENTRI NAZIONALI - Spoke 8 for fellowship support. In addition, the authors thank the European Union’s Horizon 2020 projects Brainteaser (Grant Agreement ID: 101017598) and GenoMed4All (Grant Agreement ID: 101017549). Dr. Tiziana Sanavia thanks PRIN project “Investigating the role of NF-YA isoform/lncRNA axis in mesoderm specification” of the Italian Ministry for Education, University and Research (Grant ID: 20224TWKNJ).

References

- [1] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, November 2005. ISSN 1097-0258. doi: 10.1002/sim.2427.
- [2] Giovanni Birolo, Ivan Rossi, Flavio Sartori, Cesare Rollo, Tiziana Sanavia, and Piero Fariselli. Survhive: a package to consistently access multiple survival-analysis packages. 2025. doi: 10.48550/ARXIV.2502.02223. URL <https://arxiv.org/abs/2502.02223>.
- [3] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. ISSN 0027-0644. doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.
- [4] Gilles Buitinck, Mathieu Louppe, Fabian Blondel, Andreas Pedregosa, Olivier Mueller, Vlad Grisel, Peter Niculae, Alexandre Prettenhofer, Jaques Gramfort, Robert Grobler, Jake Layton, Arnaud Vanderplas, Brian Joly, Holt, and Gael Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013.
- [5] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246. URL <http://www.jstor.org/stable/2985181>.
- [6] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith

- Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 04 2012. ISSN 1476-4687. doi: 10.1038/nature10983.
- [7] John A. Foekens, Harry A. Peters, Maxime P. Look, Henk Portengen, Manfred Schmitt, Michael D. Kramer, Nils Brünner, Fritz Jänicke, Marion E. Meijer-van Gelder, Sonja C. Henzen-Logmans, Wim L. J. van Putten, and Jan G. M. Klijn. The Urokinase System of Plasminogen Activation and Prognosis in 2780 Breast Cancer Patients1. *Cancer Research*, 60(3):636–643, 2000. ISSN 0008-5472.
- [8] Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, December 2006. ISSN 1521-4036. doi: 10.1002/bimj.200610301. URL <http://dx.doi.org/10.1002/bimj.200610301>.
- [9] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–87, 1996.
- [10] T. Hothorn. Survival ensembles. *Biostatistics*, 7(3):355–373, December 2005. ISSN 1468-4357. doi: 10.1093/biostatistics/kxj011.
- [11] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), September 2008. ISSN 1932-6157. doi: 10.1214/08-aos169.
- [12] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1.
- [13] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. URL <http://jmlr.org/papers/v20/18-424.html>.
- [14] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11842.

- [15] Ismael Lemhadri, Feng Ruan, and Robert Tibshirani. LassoNet: Neural networks with feature sparsity. *Proceedings of machine learning research*, 130:10–18, 2021. URL <https://api.semanticscholar.org/CorpusID:213619646>.
- [16] Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021.
- [17] Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. 2022. doi: 10.48550/arXiv.2204.07276.
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.
- [19] Sebastian Pölsterl. Scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL <http://jmlr.org/papers/v21/20-729.html>.
- [20] Patrick Royston and Douglas G. Altman. External validation of a cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13(1):33, 2013. ISSN 1471-2288. doi: 10.1186/1471-2288-13-33. URL <https://doi.org/10.1186/1471-2288-13-33>.
- [21] M Schumacher, G Bastert, H Bojar, K Hübner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, RL Neumann, and HF Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, October 1994. ISSN 1527-7755. doi: 10.1200/JCO.1994.12.10.2086.
- [22] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 2011. ISSN 1548-7660. doi: 10.18637/jss.v039.i05.
- [23] Terry Therneau et al. A package for survival analysis in s. *R package version*, 2(7):2014, 2015.
- [24] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, February 1997. ISSN 1097-0258. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.

- [25] Zifeng Wang and Jimeng Sun. Survtrace: transformers for survival analysis with competing events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB 2022. ACM, August 2022. doi: 10.1145/3535508.3545521.
- [26] Xuelin Yang, Louis Abraham, Sejin Kim, Petr Smirnov, Feng Ruan, Benjamin Haibe-Kains, and Robert Tibshirani. Fastcph: Efficient survival analysis for neural networks. August 2022. doi: 10.48550/ARXIV.2208.09793.