Research Article GU J Sci, Part A, 11(3): 518-534 (2024) 10.54287/gujsa.1505905

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Neutral Benchmarking of Survival Models in Health Sciences: Comparative Study of Classical and Machine Learning Techniques

Sumaila ABUBAKARI[1*] Filiz KARAMAN[1]

[1] Yildiz Technical University, Institute of Applied Sciences, Department of Statistics, İstanbul, Türkiye

Keywords

Survival Modeling
Machine Learning
Benchmarking
Biomarkers
Oncology

Abstract

Survival analysis plays a central role in diverse research fields, especially in health sciences. As an analytical tool, it can be used to help improve patients' survival time, or at least, reduce the prospects of recurrence in cancer studies. However, approaches to the predictive performance of the current survival models mainly center on clinical data along with the classical survival methods. For censored "omics" data, the performance of survival models has not been thoroughly studied, either often due to their high dimensionality issues or reliance on binarizing the survival time for classification analysis. We aim to present a neural benchmark approach that analyzes and compares a broad range of classical and state-of-the-art machine learning survival models for "omics" and clinical datasets. All the methods considered in our study are evaluated using predictability as a performance measure. The study is systematically designed to make 36 comparisons (9 methods over 4 datasets, i.e., 2 clinical and 2 omics), and shows that, in practice, predictability of survival models does vary across real-world datasets, model choice, as well as the evaluation metric. From our results, we emphasize that performance criteria can play a key role in a balanced assessment of diverse survival models. Moreover, the Multitask Logistic Regression (MTLR) showed remarkable predictability for almost all the datasets. We believe this outstanding performance presents a unique opportunity for a wider use of MTLR for survival risk factors. For translational clinicians and scientists, we hope our findings provide practical guidance for benchmark studies of survival models, as well as highlight potential areas of research interest.

Cite

Abubakari, S., & Karaman, F. (2024). Neutral Benchmarking of Survival Models in Health Sciences: Comparative Study of Classical and Machine Learning Techniques. *GU J Sci, Part A, 11*(3), 518-534. doi:10.54287/gujsa.1505905

Author ID (ORCID Number)

0000-0003-4375-6273 Sumaila ABUBAKARI
0000-0002-8491-674X Filiz KARAMAN

Article Process

Submission Date 27.06.2024
Revision Date 05.07.2024
Accepted Date 11.07.2024
Published Date 30.09.2024

1. INTRODUCTION

Survival models are some of the most popular analytical techniques in the field of Statistics that are designed to handle censored observations. In the sense of application, they span a vast majority of fields; Medicine (Salerno & Li, 2023), Education (Arib, 2023), Gadget reliability (Karim & Islam, 2019), and Loan default (Thackham, 2022). Survival analysis is usually seen as a unique technique for its ability to deal with issues of censoring—a scenario where the exact survival time of a patient is not exactly known, either due to the event not being observed within the study time, or partial information of their survival time is known. For those subjects who are censored at the end of the study, we know that their survival time is, at least more than the stated time of the study. Censoring is often grouped into left-, right-, and interval censorship, with the most common being right censoring. The differences in these types lie in the range of the exact survival time we observe. For instance, in right-censoring, we observe the lower limit, the upper limit in left-censoring, and both in the interval-censoring. Comprehensive discussions on various forms of censoring are illustrated in (Klein & Moeschberger, 2003; Gijbels, 2010). In this article, we adopt the terminology of survival analysis in which the "status" variable is binarized, and the time to observe the event is referred to as the survival time.

*Corresponding Author, e-mail: abubakarisumaila3@gmail.com

Over the years, various survival models have been maintained, improved, or extended to achieve different research goals. These include both traditional and machine learning techniques—estimating survivor functions, comparing two or more survival curves, and/or the joint cumulative effect of complex risk factors on the survival time. Often, the inference in survival analysis is obtained from one or a hybrid of different modeling schemes. For instance, the Kaplan-Meier (KM) estimator (Kaplan & Meier, 1958) uses the nonparametric approach to estimate the survivor function while the log-rank test (Peto & Peto, 1972) is used to compare two or more survivor functions. Though the KM estimator is simple and easier to interpret, the parametric approach is preferred for instances in which the distribution of the survival times is pre-determined or assumed. When the target is to estimate the effect of risk factors on survival time, the most popular go-to technique is the CoxPH (Cox, 1972) since the baseline hazard is unspecified, while the effect of predictor(s) is specified parametrically.

In the last few decades, high-throughput techniques have enormously generated data at a faster rate and on large scale ("omics") from cellular processes. For example, following the rapid progression of technology in DNA microarrays, survival prospects of cancer patients and other forms of diseases have efficiently been improved due to such technologies presenting better paths to evaluate gene expression levels (microarray data extraction). Thus, one can run a survival genomic analysis, focusing on specific genotypes for clinical insights. For instance, non-small cell lung cancer (NSCLC) is considered a chief cause of lung cancer mortalities today. It is believed that the survival rate is influenced by differentially expressed genes (DEGS) between normal lung tissue and NSCLC. It was found that the overall survival rate was highly correlated with DEGs, and enriched in factors such as angiogenesis, DNA replication, and cell cycle (Liu et al., 2019). A challenging task from such microarray data, however, is the enormity of gene expression data used to discriminate between defective cells and normal cells, even for a unit gene. Simply put, we have to deal with the problem of multiple simultaneous hypothesis testing. In fairly low-dimensional data, the Bonferroni correction (Dunn, 1961) sufficiently deals with the problem of multiple testing. However, due to its conservative nature, when tonnes of genes are tested, a small proportion is detected. To overcome this problem, the proposed false discovery rate (FDR) (Benjamini & Hochberg, 1995) uses a method that adjusts for the conservativeness in the Bonferroni correction approach. Even so, genomic data suffers from the problem of the "curse of dimensionality" ($p \gg n$). The statistical techniques to deal with this sort of data go beyond the traditional methods, due to the high dimensional space of the risk factors, coupled with the high collinearity of some genes in the gene expression levels. Again, to overcome this challenge, many researchers have proposed efficient approaches; for example, penalized Cox regression, which trains, tests, and validates the high dimensional data (Dai & Breheny, 2019; Shih & Emura, 2021).

Several survival models have been proposed in the last few decades, from the traditional approaches to the contemporary machine learning models. Numerous investigations in the literature provide a great overview of survival models using right-censored datasets, with little or no focus on time-dependent covariates (see (Wang et al., 2019)). Nonetheless, a limited number of these studies provide comprehensive real-world dataset comparisons, and very few also approach the analysis from a practical point of view. Our motivation for this study stems from providing a fairly broad benchmark study that uses clinical and omics datasets in the health sciences. Our study seeks to improve knowledge and understanding of survival models, as well as to guide clinical decisions. The rest of this paper is organized as follows: in Section 2, we give a general overview of the classical CoxPH model, its proposed regularized extensions for dealing with high-dimensionality, and the modern ML survival models in health science. We also give a comprehensive literature review of some carefully selected articles in cancer research, discussing them in light of commonly used ML survival models for clinical and omics data, and thereafter revisit some benchmark studies in Section 3. In Section 4, we introduce the 4 datasets considered in this study, as well as the 9 models and procedures used for the comparative study. In Section 5, we give the results, and the discussion is provided in Section 6. Finally, we conclude our comparative study in Section 7 with some concluding remarks and implications for clinical researchers.

2. OVERVIEW OF CLASSICAL SURVIVAL AND MACHINE LEARNING MODELS

Generally, there are two arms of survival models. Classical models comprise parametric, semiparametric, and nonparametric models. On the other hand, contemporary ML models comprise state-of-the-art deep neutral

learning-based methods and ensemble-based methods. We briefly review both arms of modeling in the next sections.

2.1. Traditional Survival Models

The CoxPH model (Cox, 1972) is the most common traditional survival approach used to evaluate the dependency of survival time on risk factors (predictor variables). It is built on the validity of the PH assumption, mathematically stated by;

$$h(t|\mathrm{X}) = h_0(t)\exp\left(\sum_{j=1}^{p} x_{ij}\,\beta_j\right) \tag{1}$$

where $h_0(t) \geq 0$ is the baseline hazard function, $X = (x_1, x_2, \dots, x_p)$ is a vector of covariates in the model, and $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ is a vector of coefficients. The unspecified baseline hazard function in equation (1) implies that it assumes no functional form, i.e., a ratio of two hazard functions is free of the baseline hazard (cancels out).

Estimating the regression coefficients in CoxPH requires maximizing the likelihood function. However, due to the presence of censored observations, a *full* maximum likelihood estimation (MLE) is impracticable. To overcome this problem, the partial likelihood method is proposed, which takes into consideration censored and uncensored observations in the dataset (Cox, 1975). For right-censored data, this likelihood function is stated by;

$$L(\beta) = \prod_{j=1}^{n}\left\{\frac{\exp(\beta^T x_j)}{\sum_{k\in R(t_j)}\exp\,(\beta^T x_k)'}\right\}^{\delta_j} \tag{2}$$

where $R(t_j) = \{k: t_k \geq t_j\}$ represents the risk set at time t_j. Note that the risk set comprises both the censored and uncensored subjects before time t_j. The goal is to estimate a vector of regression coefficients, $\hat{\beta}$, by maximizing Equation (2).

2.1.1. Penalized Cox Survival Models

Generally, microarray data are known to have overwhelmingly few observations for too many variables ($p \gg n$). For example, gene microarrays have been used to identify significant disease-related genes in a single-wise gene analysis. However, this approach is suboptimal as it fails to identify and explain the complex associations between diseases, genes, and environments. One solution to this problem is the so-called *regularization* approach. Regularization is a technique used to improve and prevent the overfitting of a probabilistic model. In the context of gene microarrays (omics), one of the aims is to identify the most significant risk factors/features among hundreds of thousands of features linked to the outcome variable. Here, the features are selected by imploring different penalty functions on the assumption of sparsity—thus, of the tens of thousands of genes, few of the genes may have significance on the patient's survival time (Hastie et al., 2015). Penalized Cox models are extensions of the CoxPH model, proposed with varying penalized functions—notably among them; lasso-, ridge-, elastic-net, and OSCAR-Cox models (Ye & Liu, 2012; Shih & Emura, 2021). Subject to their respective penalty functions, the regression coefficients are obtained by a minimization (negative) of the partial log-likelihood, shown below;

$$
\begin{aligned}
\hat{\beta}_{lasso} &= \text{argmin}\left\{-\sum_{i=1}^{n}\delta_i\left(X_i\beta-\log\left(\sum_{i=1}^{n}\exp\left(X_i\beta\right)\right)\right)+\lambda\sum_{k=1}^{p}|\beta_k|\right\}\\
\hat{\beta}_{ridge} &= \text{argmin}\left\{-\sum_{i=1}^{n}\delta_i\left(X_i\beta-\log\left(\sum_{i=1}^{n}\exp\left(X_i\beta\right)\right)\right)+\lambda\sum_{k=1}^{p}|\beta_k^2|\right\}\\
\hat{\beta}_{EN} &= \text{argmin}\left\{-\sum_{i=1}^{n}\delta_i\left(X_i\beta-\log\left(\sum_{i=1}^{n}\exp\left(X_i\beta\right)\right)\right)+\lambda\sum_{k=1}^{p}|\beta_k^2|+(1-\alpha)\sum_{k=1}^{p}\beta_k^2\right\}\\
\hat{\beta}_{OSCAR} &= \text{argmin}\left\{-\sum_{i=1}^{n}\delta_i\left(X_i\beta-\log\left(\sum_{i=1}^{n}\exp\left(X_i\beta\right)\right)\right)+\lambda_1\|\beta\|_1+\lambda_2\|T\beta\|_1\right\}
\end{aligned}
\tag{3}
$$

where λ is the '*tuning parameter*' used to regulate the degree of regularization, δ_i is an indicator representing uncensored observation. In the case of the Octagonal Shrinkage and Clustering Algorithm for Regress (OSCAR), T is the sparse (symmetric) edge set matrix obtained by setting a graph structure where each considered feature is a node. For regularization, $\lambda = 0$ means no regularization is performed, while for $\lambda \to \infty$, the regression coefficients tend to be contained (i.e., regularized). Lasso-Cox uses L_1-norm regularizer while ridge-Cox uses L_2-norm regularizer, with the elastic-net using a combination of both penalties. It is important to point out, though, that the various forms of penalization in Equation (3) could also be incorporated into the cost functions of recent machine learning techniques, which we next introduce in the section below.

2.2. Modern Machine Learning (ML) Methods

The traditional survival models are rarely sufficient for capturing complex nonlinear dependencies between the survival time and the predictor variables. To this end, there has been a rising increase in the use of ML models in healthcare, especially for their remarkable performance. Also for their domain adaptability and ability to improve predictive accuracy, applications of ML models span diverse areas of research, for example, security (Liang et al., 2019).

The basic underlying idea of ML is to make a computer run powerful algorithms on complex input data so as to recognize hard-to-discover patterns. Machine learning systems are generally classified based on the task or a set of tasks to accomplish. This classification depends on whether the ML system learns through human supervision (supervised learning) or via other means such as; unsupervised, semi-supervised, reinforcement ML. Of these systems, supervised ML is the most interesting for survival data analysis, especially for classifying and predicting the target variable. We briefly give an overview of common ML methods.

2.2.1. Support Vector Machines (SVMs)

As a supervised ML approach, SVMs have been successful in dealing with regression and classification problems, in addition to the success of their adaptability to fitting survival data (Smola & Schölkopf, 2004). The general aim of SVMs lies in maximizing the distance between two classes while at the same time finding a separate hyperplane that minimizes wrong classification. This hyperplane also attempts to stay so far from close observations so that the individuals found on the edge of the separating hyperplane constitute the supporting vectors, which on the whole determine the classification.

Although using linear classifiers in SVMs is often efficient and enhances performance, in the case of high-dimensional datasets, linear SVM classifiers tend to be poor discriminants. Interestingly, the SVM classifier overcomes this problem by using a high-dimensional kernel function that handles both nonlinearity and high dimensionality. SVMs have also been extended to handle regression and survival data. For example, an SVM-based method, SurvivalSVM, was proposed by Van Belle et al. (2008) for survival modeling. With a modified

penalty term, this model is a variation of the penalized log-likelihood function. SurvivalSVM differs from other models by treating the prognosis problem as a ranking problem, rather than directly incorporating hazard estimation. A further extension known as support vector regression for censored data, SVRc (Khan & Zubek, 2008) was developed to factor in an asymmetric cost function for uncensored and censored data.

2.2.2. Random Survival Forests (RSF)

Random forests constitute another ensemble method purposefully developed for making predictions via tree-structured models (Breiman, 2001). The modeling framework is similar to bagging—thus, to grow the trees, it involves randomly bootstrapping from the training set. The central difference between the two is that, instead of using all the covariates or attributes when splitting a node, an RF uses a random subset of attributes to search for the best variables. Random survival forests have been shown to improve predictive performance due to randomization which reduces the degree of correlation among the trees. As an ensemble learner, RSFs are formed by averaging several base learners, similar to how regression problems are modeled. In the framework of survival ML, the base learner is a survival tree while the ensemble is the cumulative hazard function obtained by averaging the Nelson-Aalen's cumulative hazard function of individual trees (Ishwaran et al., 2011).

Ishwaran et al. (2008) first proposed Random Survival Forests (RSF) as a random forest variation for modeling survival data. Multiple models are generated from a large number of resamples. The result of the ensemble prediction is then averaged across the base learners or the outcome of a majority vote. In our application, the core features of RSF are that we assess the performance of the survival tree rather than using mean square error (mse) as in traditional regression analysis, or the confusion matrix as in classification problems. Additionally, we employ each node's log-rank estimation as the stopping rule.

2.2.3. Boosting-Based Methods (CoxBoosting)

Boosting is another popular ensemble method based on the combination of base learners into a strong learner which represents the final output (Freund, 1990). The principal concept of boosting is to iteratively update a set of predictors by repeatedly learning weak classifiers and adding them to a final strong classifier. Updating is done by minimizing a pre-assigned loss function. Note that after a weak learner is added, weights in the data are readjusted, so-called "re-weighting".

The Cox boosting model (De Bin, 2016) was proposed and designed based on the classical Cox model, where the boosting is applied in estimating the risk factors, i.e., regression coefficients, as in Equation (1). The β's are updated iteratively either by using the *mboost* method or by the partial log-likelihood function. CoxBoost is therefore a gradient-boosting algorithm in which the L_2-norm partial log-likelihood is used (our choice in this study). Note that there are 2 main factors to consider in a boosting procedure: the first one serves as a benchmark to control the weakness of the estimators, while the second parameter specifies the number of boosting iterations to be performed to meet the stopping criterion. The second parameter is necessary to avoid overfitting.

2.2.4. Neural Networks

In recent years, the ML methods discussed here, so far, have been classified as classical ML models, in contrast to the more emerging complex ML neural networks such as deep learning. Inspired by the complicated functionality of the human brain, artificial neural networks (ANN) are a collection of algorithms that are interconnected to process pieces of information in response to input data. Like boosting methods, many neural network methods such as *Cox-nnet* (Ching et al., 2018) and *DeepSurv* have been proposed as extensions to the popular classical Cox regression. The general idea is to have a collection of cost functions to estimate the survival probability or the hazard of patients, assisted by neurons in hidden layers of deep learning architecture. For example, gene microarrays can be represented with these hidden layers, with no stringent regard to the proportional hazard assumption.

While the Cox-nnet and DeepSurv perform favorably well on high-dimensional data, they are PH-based neural networks, hence their predictive power depends on the validity of the PH assumption. To overcome the PH

dependency, DNNSurv (Zhao & Feng, 2019) proposed a new deep neural network model that uses the pseudo-value approach to estimate the survival probability.

2.2.5. Multi-Task Learning (MTL)

Up till now, all the models discussed stress on optimizing one objective. Multi-task learning as an ML model focuses on training the model to perform multiple tasks (objectives) concurrently. Especially in survival models where the dataset is time-dependent, it is more informative to perform tasks concurrently and have the parameters estimated from a joint optimization of numerous likelihood functions—thus, each task corresponds to an objective. In MTL, the aim is to improve the generalized performance of the model by relying on the information shared across multiple tasks. This concept can be implemented in deep learning (DL). For example, Yu et al. (2011) proposed the multi-task logistic regression model (MTLR) as a survival model for multiple time points to use a logistic regression model to predict survival for each. In this case, parameters are jointly estimated by maximizing the joint likelihood function.

2.3 Performance Evaluation

When the time-to-event data are laced with censored observations, the predictive performance of survival models may not be adequately evaluated by the traditional ranking or classification metrics. The most common base metric for evaluating survival models is the *concordance index* (C-Index) (Harrell Jr et al., 1996). As a metric, the index is defined as the ratio of correctly ordered pairs (concordant) to the overall number of possible evaluation pairs. The index values of this metric lie in the range $[0,1]$. An index of 1 is interpreted as a *perfect* concordance between the event times and the risk. In the same vein, a value of 0 is interpreted as perfect discordance and a value of 0.5 means the model does no better prediction than the toss of a fair coin. Unfortunately, this metric is not unbiased when the amount of censoring in the data is high, thus far, Uno et al. (2011) have proposed an alternative estimator to deal with this situation.

The Brier score (Brier, 1950) is another popular metric for evaluating the predictive performance of survival models. It can be thought of as a cost function that measures the mean squared difference between the predicted probabilities and the true classes. Like the C-Index, it ranges within $[0,1]$. A metric score of 0 is interpreted as *perfect accuracy* and a metric score of 1 is interpreted as *perfect inaccuracy*. For a given survival time interval, a mathematical integration of multiple Brier scores can be computed as the overall average measure of the performance. This is referred to as the integrated Brier score (ibrier). Incorporating the censoring information into the Brier score method has also been extended (Graf et al., 1999). Other performance metrics for survival models include Royston's D index, Mean Absolute Error (MAE), and the time-dependent AUC.

3. REVIEW OF BENCHMARK STUDIES IN HEALTH RESEARCH

With a primary focus on omics and cancer datasets, we selected articles from popular scientific repositories (e.g., Scopus) by stressing ML techniques and the ubiquitous classical CoxPH and its extensions. We should point out that in doing this, the *exclusion criterion* was to drop articles not focused on survival analysis in health science. At the same, we searched for articles in this field with high citations in the last 5 years and with articles whose main foci are on "comparison", "benchmarking", and similar derivatives. After a careful review of the search results, we selected 8 of these papers to study the survival methods explored in them. Based on the techniques, we selected the most commonly used ML survival models in health science (Table 1).

3.1 Benchmark Studies in Health Science

In Moncada-Torres et al. (2021), the classical CoxPH is compared with SVMs, Random Forests, and XGBoost with decision trees as the base learner. To evaluate the performance of these methods, the authors used non-metastatic breast cancer data using the Concordance Index. The performance was identical for all the methods except for the XGBoost which outperformed the rest considered. In another comparative study of the Cox model against random survival forests (RSF) and support vector machine (SVM), Kim et al. (2022) found that the Cox performed slightly better than the RSF and the SVM, in terms of assessing the prognostic prospect of resected non-metastatic pancreatic ductal adenocarcinoma of patients.

In a study to evaluate recurrence patterns and the survivability of gastric cancer patients who underwent chemotherapy and radiation therapy, Akcay et al. (2020) explored ML techniques such as Random Forests, XGBoost, support vector classification, and the Naive Gaussian Bayes techniques. The study concluded the XGBoost and the Random Forest were the best predictors of overall survivability and peritoneal metastases.

Using the SEER database of Lung cancer patients to explore more predictive information, Lynch et al. (2017) explored, relative to the CoxPH model, the predictive performance of ML methods such as Decision Trees, SVMs, Gradient Boosting Machines (GBM), and a custom ensemble. Though the performance of these ML techniques was comparatively similar to the classical Cox proportional method, the Gradient Boosting approach proved to be the most efficient. In a similar retrospective study that focused on prognostic predictive modeling of Breast cancer patients, Xiao et al. (2022) compared the performance of three competing models, namely; Random survival Forests, penalized CoxPH, and SVMs. The study found the Cox model and the SVM to have marginally outperformed the RSF. In a cohort study of breast cancer, Aivaliotis et al. (2021) found the RSF to capture complex non-proportional hazard patterns, they also found out that the RSF overfits the data when compared to the classical CoxPH models—besides the fact that it is less interpretable. To explore the survival outcomes of bladder cancer patients, Bhambhvani et al. (2021) assessed the predictive performance of a multivariable CPM with Artificial Neural Networks (ANN). Intending to predict overall survival (OS) and 5-year specific survival (DSS), this study concluded that ANNs improve predictability in bladder cancer patients than in a multivariable CoxPH model, except for the complexities in the interpretation of ANNs. In a more comprehensive and large-benchmark study, Herrmann et al. (2021), under a multi-omic data setting, benchmarked some eleven methods built around random forests, boosting, and penalized regression. Using the KM estimation and the Cox model as reference methods, the block forest method—a variant of the Random Forest method was found to outperform the classical Cox method, however marginally. Further, the study pointed out that the performance of these methods, to a degree, varied on account of the multi-omic structure of the data.

Richter and Khoshgoftaar (2018) investigated the advances in statistical and ML techniques, the gaps in the literature, and several approaches for developing cancer risk models utilizing structured clinical patient data. The authors concluded that the most popular statistical technique in survival analysis is the CoxPH model while the most common ML approaches are neural networks, SVMs, and decision trees. In the same vein, other researchers have focused on recent advancements in cancer research, concluding that ensemble approaches, decision trees, and artificial neural networks are some of the most used ML methods for setting up survival models. More precisely, a recent comprehensive and methodological literature review by Deepa and Gunavathi (2022) summed up, to date, the majority of cancer research on machine- and deep-learning applications in survival analysis. The comprehensive review cited papers used in ML models—Support Vector Machines, Random Forests, and Support Vector Regression to predict the factors affecting survivability, as well as XGBoost to forecast recurrence and disease progression.

In light of this brief literature review, we deduced that the most widely used modeling techniques in cancer studies include CoxPH, boosting-based methods, Support Vector Machines (SVMs), deep learning-based models, and Random Survival Forests (RSF). Note that it is exceedingly difficult to include every technique in benchmark research. For this reason, our review of survival models considered commonly employed methods, both in the classical sense and the modern ML techniques in health sciences. In Table 1, we present a summary of the articles considered for the review of survival models—methods, performance metrics, and the cancer type.

4. MATERIALS AND METHODS

Clinical data sets—we considered 2 clinical datasets as summarized in Table 2.

- Lung dataset: This data contains 7 features of 228 patients diagnosed with advanced lung cancer from the North Central Cancer Treatment Group (NCCTG). This data can be obtained from the survival package in R and contains the survival information of the patients.

- Veteran dataset: Contains lung cancer data from a randomized trial of 2 treatment regimens with 6 features and 137 patients. This data is also readily available in the survival package in R.

Omics data sets—we considered 2 omics datasets as summarized in Table 2.

- Ovarian1 dataset: Ovarian cancer gene expression data from curatedOvarianData package with "GSE49997_eset" as the data ID. The study was on OVCAD Consortium conducting a study to validate the impact of a molecular subtype on ovarian cancer outcomes. The curation of this is given by Ganzfried et al. (2013).

- Ovarian2 dataset: Ovarian cancer gene expression data from curatedOvarianData package with "GSE30161_eset" as the data ID. The study was on multi-gene expression predictors of single drug responses to adjuvant treatment in ovarian cancer: predicting platinum resistance. The curation of this is given by Ganzfried et al. (2013).

Table 1. *Summary of selected benchmarking studies in oncology (cancer) focusing on survival models of classical and modern machine learning model techniques.*

Study	Technique	Type of cancer	Evaluation metric
Herrmann et al. (2021)	Boosting-based Penalized regression-based Random-forest-based	Liver, Blood, Lung, Skin Brain, Kidney, Stomach, Colon Ovarian, Pancreatic, Bladder, Breast Head-neck (SEER, TCGA)	Brier score Concordance Index
Moncada-Torres et al. (2021)	CoxPH Random forest SVMs XGBoosting	Breast cancer	Concordance Index
Xiao et al. (2022)	Random forest Penalized-based regression Support Vector Machines	Breast cancer	Brier score Concordance Index AUC D-index
Kim et al. (2022)	CoxPH Random survival forests SVMs	Pancreatic (SEER/KOTUS-BP)	AUC sensitivity specificity
Akcay et al. (2020)	XGBoosting Random forests SVMs Logistic regression Gaussian Naive Bayes (GNB) algo. Multi-layer perceptron	Gastric cancer	AUC sensitivity specificity
Bhambhvani et al. (2021)	Multivariable CoxPH Artificial Neural Networks	Bladder cancer (SEER)	AUC
Lynch et al. (2017)	Decision Trees Multivariable CoxPH Support Vector Machines Gradient Boosting Linear regression	Lung cancer (SEER)	Root Mean Square Error
Aivaliotis et al. (2021)	CoxPH Random forests	Breast cancer (UK)	Brier score Concordance Index

Table 2. *Summary of datasets considered in the comparative study in this paper. We round the rate of censorship to three decimal places*

Dataset source	No. of observations	No. of features	Data class	Censoring rate
Veteran (Kalbfleisch & Prentice, 2011)	137	8	Clinical	0.066
NCCTG Lung (Loprinzi et al., 1994)	228	9	Clinical	0.276
Ovarian 1 (Ganzfried et al., 2013)	194	16050	Omics	0.706
Ovarian 2 (Ganzfried et al., 2013)	58	19818	Omics	0.379

4.1. Benchmarking Procedure and Methods

Our study is aimed at a neutral comparison of classical survival models to the state-of-the-art ML models. Generally, the performance of probabilistic models is often influenced by the design and choice of datasets, and for this reason, we chose two data types—clinical and omics datasets. For each of the 9 models considered in this study, we evaluated the models' performances in terms of their predictability. Here, predictability is in reference to 4 popular evaluation metrics. Note that when the amount of censoring in the test data is high, Harrell's c-index is known to be a biased estimator, and for this reason, we have included an alternative estimate, Uno's c-index which uses the inverse probability of censoring weighting (ipcw). We also calculated the standard deviation (SD). All the analysis in our comparative study is conducted using **R 4.3.1**.

The models considered in the study are outlined in Table 3. For the comparison of the methods to be on neutral grounds, we use the default settings for the hyperparameter tuning (except for CoxPH). The evaluation of the methods are carried out on real-world data sets (2 clinical, 2 omics) as shown in Table 2. Using RStudio, we run the analysis repeatedly with 5-fold cross validation. In each run, we split the entire data into a *training* data set (80%) and the remaining 20% used as the *testing* set. Each model is trained using the training set while the evaluation metrics are computed using the testing set. In a case where there is a feature selection step, the 5-fold cross validation is still applied to the nested feature selection. We provide a supplementary data (*Supplementary-Table-1*) on the details of packages and parameters for the methods.

All the 4 datasets in the comparative study are benchmark datasets—the ovarian datasets have already being manually curated (clinical) and the expression data in them have also being uniformly processed.

5. RESULTS

To exhaustively assess the strengths and weaknesses of the various survival models, we settled on 9 typical methods from our extensive review of the literature. We then study their statistical performance to 4 diverse datasets. The basis of our assessment is on 4 metrics on predictability; Harell's C-Index, Uno's C-Index, Brier's score, and the time-dependent AUC. The primary focus is on extensive comparison of models from the traditional approaches to the state-of-the-art machine learning—*Classical models*: Cox, Cox_lasso, Cox_ridge, Cox_elasticNet; *Advanced ML*: RSF, SurvivalSVM, CoxBoosting; *Neural Network model*: Deepsurv.

5.1. Practical Consideration in Performance Assessment

Owing to the varying characteristics of data collection in domains like medical field (e.g., omics data), not all survival models meet the feasibility criteria for application to all data types. For instance, the classical CoxPH methods cannot handle high-dimensional data (i.e., $p \gg n$) where there are more features (p) than samples (n). For this reason, the 2 omics datasets, Ovarian_1, Ovarian_2, are not feasible for the classical Cox method, as seen in Table 4. Further, the results highlight, for example, that SurvivalSVM is readily applicable to all the 2 clinical datasets, whereas it is not readily applicable to Ovarian_2 dataset.

The multitask linear regression (MTLR) approach, conspicuously, outperformed the rest of the learners on all the evaluation metrics as presented in Table 4 (bolden mean values), across all the data sets, except for

Ovarian_2 where the method is not applicable. More so, looking at the average degree of performance between the clinical and omic data sets suggests that the learners (methods) depend on the type of data set.

For instance, Figure 1 shows the comparative performance of all the methods using Uno's C-index, which uses the inverse probability of censoring weighted, IPWC, (See supplementary material for results of other metrics). To appreciate the gain in model predictability relative to the Cox-based ML method, we compare the CoxBoosting method with the conventional Cox-based methods (i.e., Cox_EN, Cox_Lasso), which are typically employed as the standard of comparison in many medical investigations. Figure 2 presents our results which demonstrate similar performance across the datasets. For this comparison, we observe similar predictability for both the clinical and omic datasets when measured by both Harrell's C-Index and Uno's C-Index. This observation suggests that the performance of clinical and complex health data using state-of-the-art ML may not be as straightforward as is the case in some fields.

Table 3. *Summary table of traditional and state-of-the-art ML models considered in the benchmarking, R functions with their listed parameters.*

Model name	Function	R package	Default parameters
Cox	coxph	survival	none
Lasso Cox (cox_lasso)	penalized glmnet	penalized glmnet (omics)	lambda1 = 1, lambda2 = 0 alpha = 1.0, nfold = 5, type.measure =’C’
Ridge Cox (cox_ridge)	penalized glmnet	penalized glmnet (omics)	lambda1 = 0, lambda2 = 1 alpha = 0.0, nfold = 5L, type.measure =’C’
Elastic net (cox_en)	penalized glmnet	penalized glmnet (omics)	lambda1 = 1, lambda2 = 1 alpha = 0.50, nfold = 5L, type.measure =’C’
Random survival forest	rfsrc	randomForestSRC	ntree = 1000, mtry = 10
SurvivalSVM	survivalsvm	survivalsvm	margin = 0.050, bound = 10, eig.tol = 1e-05 sgf.sv = 5, sigf = 7, maxiter = 20 conv.tol = 1e-07, posd.tol =1e-08
Cox Boosting model	coxboost	Coxboost	stepnumber = 10, penalty number = 100
Deep survival (Deepsurv)	deepsurv	survivalmodels	frac = 0.3, activation = ‘relu’, dropout = 0.10, early_stopping = T, num_nodes = c(4L, 8L, 4L, 2L), epochs = 100, batch_size = 32
Multitask Logistic Regression	mtlr	MTLR	C1 = 1, normalize = T, train_biases = T

Table 4. *Summary assessment of the average performance of the methods. The third and sixth columns show the best-performing method/learners for each data and evaluation metric. We calculated the values by averaging over the cross-validation (CV) iterations. For the methods failing to yield the model coefficients, we represent them by '–'. Column 'ci' represents 95% confidence intervals; based on using quantiles of the t-distribution. It is possible that the number of risk factors might differ from the sum of risk factors in each category—as a result of rounding errors. Observe that the intervals, in this case, are only meant to offer an idea of stability of the average value (but are not, in reality, valid intervals due to the independent-observation assumption (Bengio & Grandvalet, 2003). Bold values refer to the most outperforming method.*

		Harrell's C-Index			**Uno's C-Index**			**Time-dependent AUC**		
Data source	Learner	Mean	sd	ci	Mean	sd	ci	Mean	sd	ci
	CoxPH	0.619	0.049	[0.609, 0.628]	0.601	0.047	[0.592, 0.611]	0.661	0.069	[0.648, 0.675]
	Cox_Lasso	0.614	0.047	[0.605, 0.624]	0.489	0.070	[0.475, 0.503]	0.654	0.062	[0.642, 0.667]
	Cox_Ridge	0.614	0.049	[0.604, 0.624]	0.482	0.071	[0.468, 0.496]	0.654	0.064	[0.641, 0.666]
	Cox_EN	0.509	0.025	[0.504, 0.514]	0.059	0.138	[0.032, 0.087]	0.513	0.035	[0.506, 0.520]
Lung cancer (NCCTG)	CoxBoost	0.629	0.047	[0.620, 0.639]	0.608	0.044	[0.599, 0.617]	0.674	0.067	[0.660, 0.687]
	RSF	0.603	0.044	[0.594, 0.611]	0.578	0.043	[0.569, 0.586]	0.635	0.062	[0.623, 0.647]
	survival	0.531	0.054	[0.520, 0.541]	0.521	0.048	[0.512, 0.531]	0.537	0.074	[0.522, 0.551]
	Deepsurv	0.516	0.058	[0.501, 0.531]	0.412	0.152	[0.373, 0.451]	0.527	0.080	[0.506, 0.547]
	MTLR	**0.864**	0.027	[0.858, 0.869]	**0.834**	0.030	[0.840, 0.828]	**0.935**	0.023	[0.930, 0.939]
	CoxPH	0.685	0.049	[0.676, 0.695]	0.670	0.051	[0.659, 0.680]	0.785	0.068	[0.771, 0.798]
	Cox_Lasso	0.706	0.048	[0.697, 0.716]	0.623	0.054	[0.613, 0.634]	0.809	0.063	[0.797, 0.822]
	Cox_Ridge	0.706	0.048	[0.697, 0.716]	0.623	0.054	[0.613, 0.634]	0.809	0.063	[0.797, 0.822]
	Cox_EN	0.703	0.051	[0.693, 0.713]	0.616	0.082	[0.600, 0.632]	0.805	0.069	[0.792, 0.819]
Veteran lung cancer	CoxBoost	0.694	0.048	[0.685, 0.704]	0.674	0.049	[0.664, 0.684]	0.801	0.082	[0.785, 0.818]
	RSF	0.712	0.058	[0.701, 0.724]	0.700	0.056	[0.688, 0.711]	0.798	0.065	[0.785, 0.811]
	survival	0.692	0.061	[0.680, 0.704]	0.677	0.062	[0.664, 0.689]	0.792	0.079	[0.776, 0.808]
	Deepsurv	0.637	0.072	[0.623, 0.651]	0.556	0.141	[0.528, 0.584]	0.707	0.109	[0.685, 0.728]
	MTLR	**0.869**	0.034	[0.863, 0.876]	**0.863**	0.033	[0.857, 0.870]	**0.938**	0.033	[0.931, 0.944]
	CoxPH	–	–	–	–	–	–	–	–	–
	Cox_Lasso	0.527	0.084	[0.511, 0.544]	0.501	0.262	[0.449, 0.553]	0.515	0.106	[0.494, 0.536]
	Cox_Ridge	0.500	0.000	–	–	–	–	0.500	0.000	–
	Cox_EN	0.521	0.082	[0.504, 0.537]	0.528	0.237	[0.480, 0.575]	0.506	0.104	[0.485, 0.526]
Ovarian 1 dataset	CoxBoost	0.468	0.101	[0.448, 0.488]	0.485	0.249	[0.436, 0.535]	0.465	0.111	[0.443, 0.487]
	RSF	0.544	0.092	[0.526, 0.562]	0.430	0.263	[0.378, 0.483]	0.542	0.104	[0.521, 0.562]
	survival	0.546	0.082	[0.530, 0.563]	0.645	0.238	[0.598, 0.693]	0.533	0.112	[0.511, 0.555]
	Deepsurv	0.499	0.074	[0.484, 0.514]	0.367	0.310	[0.305, 0.428]	0.500	0.100	[0.480, 0.520]
	MTLR	**0.691**	0.087	[0.674, 0.708]	**0.599**	0.320	[0.535, 0.662]	**0.713**	0.106	[0.692, 0.734]
	CoxPH	–	–	–	–	–	–	–	–	–
	Cox_Lasso	–	–	–	–	–	–	–	–	–
	Cox_Ridge	–	–	–	–	–	–	–	–	–
	Cox_EN	0.521	0.082	[0.504, 0.537]	0.528	0.237	[0.480, 0.575]	0.506	0.104	[0.485, 0.526]
Ovarian 2 dataset	CoxBoost	0.468	0.101	[0.448, 0.488]	0.485	0.249	[0.436, 0.535]	0.465	0.111	[0.443, 0.487]
	RSF	0.544	0.092	[0.526, 0.562]	0.430	0.263	[0.378, 0.483]	**0.542**	0.104	[0.521, 0.562]
	survival	**0.546**	0.082	[0.530, 0.563]	**0.645**	0.238	[0.598, 0.693]	0.533	0.112	[0.511, 0.555]
	Deepsurv	–	–	–	–	–	–	–	–	–
	MTLR	–	–	–	–	–	–	–	–	–

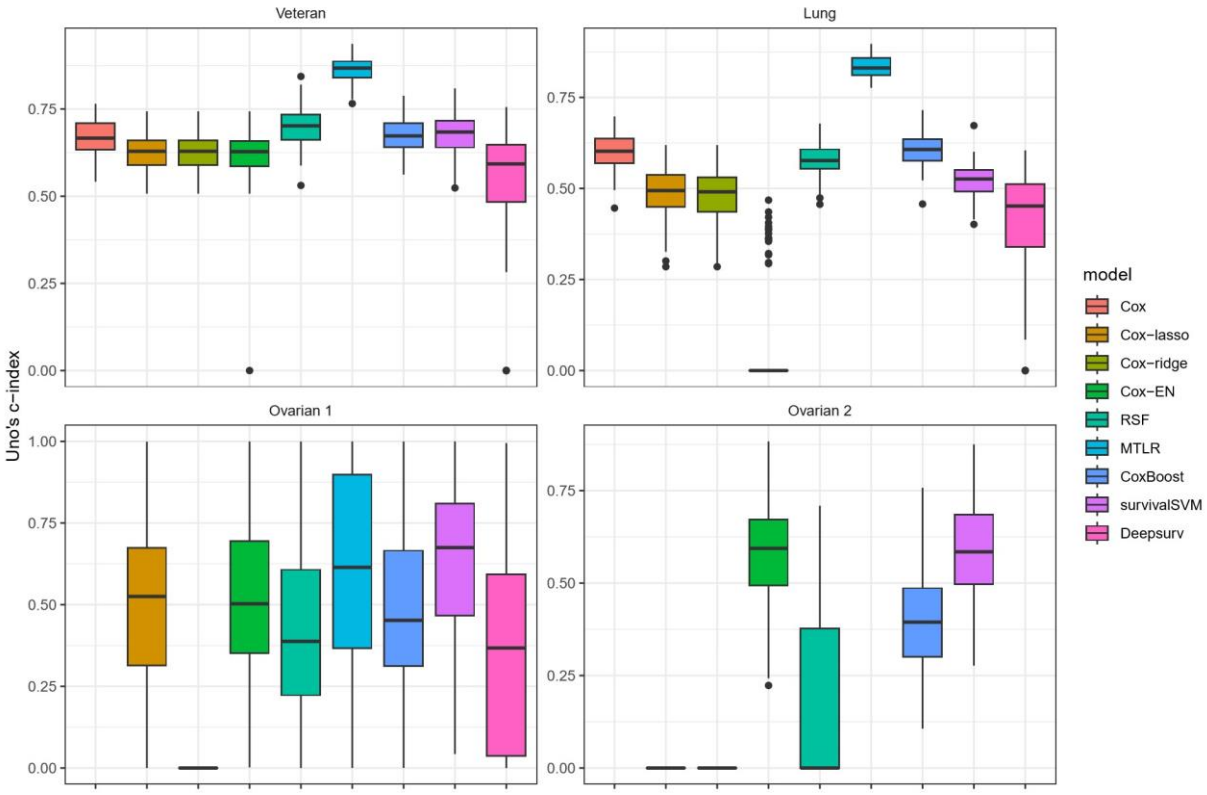

Figure 1. *Visualization of Uno's concordance index for each of the 9 learners considered on all the 4*

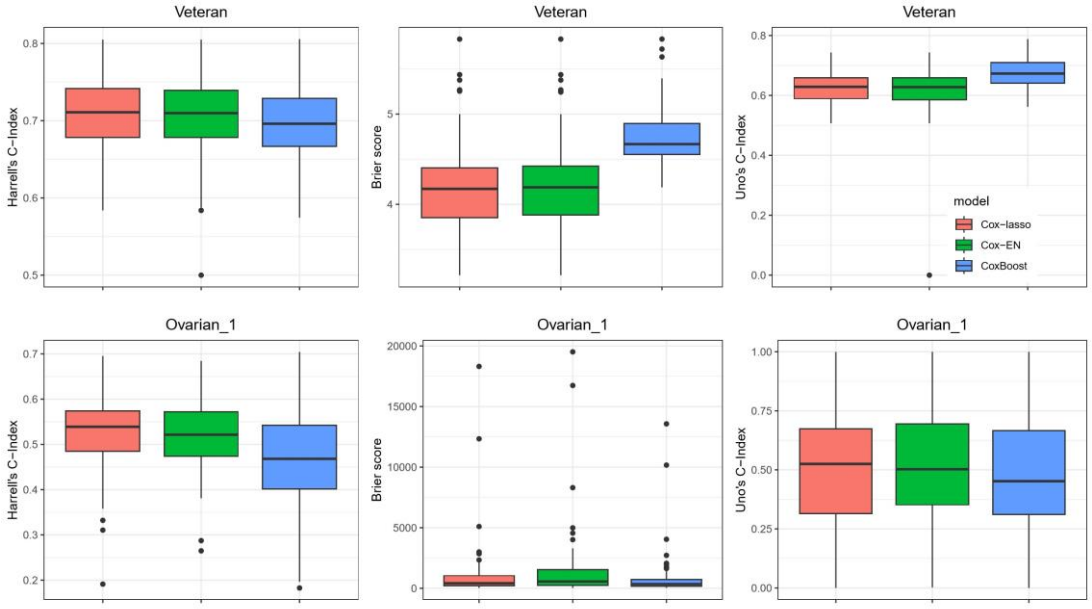

Figure 2. *Comparison of model predictability between ML Cox-based method and the traditional penalized Cox-related learners. Top-left: Harrell's metric on Veteran dataset. Top-middle: Brier's metric on Veteran dataset. Top right: Uno's metric on Veteran dataset. Bottom-left: Harrell's metric on Ovarian 1 dataset. Bottom-middle: Brier's metric on Ovarian 1 dataset. Bottom-right: Uno's metric on Ovarian dataset.*

6. DISCUSSION

More principally, caution should be exercised when interpreting the results of our comparative study. This is so because of the assessment methods used to measure the performance of the competing models. Stated differently, the relative standing of the models tends to change significantly from one scale of assessment measure to the other, in addition to the default settings of the packages considered in the benchmarking. For example, when "omics" data is considered, the *CoxBoosting* method under C-Index performs poorly as opposed to Uno's index, integrated Brier score (ibrier), and the time-dependent measures. This observation is identical to the study by Zhang et al. (2022). But if we considered the low-dimensional datasets (i.e., NCCTG Lung cancer and Veteran Lung cancer datasets), the C-Index outperforms the Brier score. What this observation shows is that the predictive performance can significantly vary with the assessment scale employed in the modeling. Of the three measures (i.e., C-index, ibrier score, and time-dependent AUC), the AUC-scale measure in nearly all cases outperformed the other two measures. It is important however, to mention at this point that the C-Index as a discriminant measure should not be given preference over the ibrier score since according to Herrmann et al. (2021), the C-Index cannot be considered an appropriate scoring rule. Thus, for prognostic calibration, choose the Brier score over the Harrell's index while for ease of interpretability, the C-Index might be preferred over the Brier score.

For predictive power, the MTLR model outperformed all the other models in nearly all the datasets we considered. A couple of reasons could be cited for this observation of performance and this is explained in detail in Yu et al. (2011). One reason worth mentioning is the fact that the MTLR model is robustly designed to allow concurrent fitting of multiple logistic regression models while directly taking into consideration the survivor function. Despite the great performance of the MTLR, not many researchers have employed its application in the modeling of omics survival data Zhang et al. (2022). This presents an opportunity for more translational medical researchers to consider this approach in their analytical toolkit.

Ideally, the best-performing model in a comparative study of competing methods is unlikely to exist in all likely scenarios because a high-performing model under one consideration could be a low-performing model under a different criterion. This should necessarily imply that the methods in our study that did not perform well might be great competitors under different criteria–interpretability, predictability, or problems with the "curse of dimensionality".

7. CONCLUSION

In our benchmark comparative study, we extensively evaluated the importance of survival methods in practice, where the selected models are based on our comprehensive review of ML benchmark studies, and applying these models to diverse datasets in biomedical studies. Using model predictability as the basis of performance, we explored a wide range of survival models from the traditional CoxPH methods to the state-of-the-art ML models, where the evaluation metric was assessed on 4 scales—Harrell-, Uno- c-index, Brier score, and time-dependent AUC. In our study, we did not attempt a comprehensive survey of tuning procedures for the 9 models considered. The core reason, in practice, is that the default set of hyperparameters is often used, hence the decision to rely on the default parameters in our study. Nevertheless, we point out that different targeted penalized techniques for a given data might cause varying performances for the learners covered in our study. Our research's conclusions will provide some degree of guidance for clinicians and translation researchers, while at the same time pointing to areas of potential study in the scope of benchmarking methodology and survival approaches.

There has been, in recent years, a distinct change in course in how time-to-event data are modeled, from the classical approach of directly modeling the hazard estimator to building several models using the survivor function. In theory, though, modeling the hazard function paves a great way to identify key biomarkers/risk factors related to the survival prospects of patients. However, if the primary goal is to accurately predict the patient's survival, then modeling directly on the survival probability greatly improves predictability. To this end, modeling methods such as the survivalSVM, and MTLR, which employed direct modeling of the survivor, demonstrated better performance based on predictability. This finding is consistent with the analysis by Yu et al. (2011) about the effective performance of their MTLR method.

In our benchmark study, it is interesting that MTLR comparatively showed exceptionally high model predictability. This high performance may be explained on several fronts. As Yu et al. (2011) extensively discussed, the performance of the MTLR is attributable to 3 central reasons; dynamic modeling, direct modeling of the survivor function, and concurrent construction of multiple logistic regression learners. To account for nonlinearity in datasets, many researchers built on the proposed MTLR by incorporating neural networks (see Fotso (2018)). Surprisingly, very few studies have used MTLR, either by using clinical data or high-dimensional omics data. Thus far, we think there is potential to employ MTLR more extensively for predicting risk factors in survival modeling, given its remarkably high model predictability.

One of the most important measures for evaluating survival studies is model predictability, with evaluation indices such as Harrell's c-index being the most widely used. Though Harrell's c-index is a ranked-based metric capable of evaluating predicted outcomes with censored data, a couple of versions of it are also available, for example, the Uno's c-index uses the IPCW technique to evaluate predicted outcomes. Other than the concordance indices, other metrics for evaluation include the Brier score and the time-dependent AUC, where the time interval is divided into several time points for evaluation, similar to the general idea of AUC in binary classification. Since model predictability can be evaluated on several metrics, we suggest that a combination of different metrics should be employed to help in comprehensively assessing the fitted model.

Though many survival models have algorithms capable of fitting both omics and clinical data, several recently developed methods are uniquely designed for high-dimensionality in omics data (e.g., genomics, transcriptomics). For example, CoxBoost as an ML learner is tailored to handle the curse of dimensionality ($p \gg n$) in omics data. The goal of developing these data-specific techniques and/or learners is to efficiently capture the distinctive features of the omics or clinical data. Apart from the fact that the performance of a model's predictability (real-world datasets) is influenced by the data type (clinical or omics), another aspect of the data that can affect predictability is *data modality*, which we did not cover in this paper.

AUTHOR CONTRIBUTIONS

Conceptualization, S.A; methodology, S.A and F.K, Software; S.A and F.K, Validation, F.K, formal analysis; S.A and F.K, data curation, S. A, manuscript-original draft, S.A; visualization, S.A and F.K, supervision, F.K. All authors have read and legally accepted the final version of the article published in the journal.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Aivaliotis, G., Palczewski, J., Atkinson, R., Cade, J. E., & Morris, M. A. (2021). A comparison of time to event analysis methods, using weight status and breast cancer as a case study. *Scientific Reports*, *11*(1), 14058. https://doi.org/10.1038/s41598-021-92944-z

Akcay, M., Etiz, D., & Celik, O. (2020). Prediction of survival and recurrence patterns by machine learning in gastric cancer cases undergoing radiation therapy and chemotherapy. *Advances in Radiation Oncology*, *5*(6), 1179-1187. https://doi.org/10.1016/j.adro.2020.07.007

Arib, M. A. A. (2023). Survival analysis of students not graduated on time using cox proportional hazard regression method and random survival forest method. *Journal of Statistics and Data Science*, 13-21. https://doi.org/10.33369/jsds.v2i1.24312

Bengio, Y., & Grandvalet, Y. (2003). No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, *16*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bhambhvani, H. P., Zamora, A., Shkolyar, E., Prado, K., Greenberg, D. R., Kasman, A. M., Liao, J., Shah, S., Srinivas, S., Skinner, E. C., & Shah, J. B. (2021). Development of robust artificial neural networks for

prediction of 5-year survival in bladder cancer. *Urologic Oncology: Seminars and Original Investigations*, *39*(3), 193.e7-193.e12. https://doi.org/10.1016/j.urolonc.2020.05.009

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1-3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, *14*(4), e1006076. https://doi.org/10.1371/journal.pcbi.1006076

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269-276. https://doi.org/10.1093/biomet/62.2.269

Dai, B., & Breheny, P. (2019). Cross-validation approaches for penalized Cox regression. *Statistical Methods in Medical Research*, *33*(4), 702-715. https://doi.org/10.1177/09622802241233770

De Bin, R. (2016). Boosting in cox regression: A comparison between the likelihood-based and the model-based approaches with focus on the r-packages CoxBoost and mboost. *Computational Statistics*, *31*, 513-531. https://doi.org/10.1007/s00180-015-0642-2

Deepa, P., & Gunavathi, C. (2022). A systematic review on machine learning and deep learning techniques in cancer survival prediction. *Progress in Biophysics and Molecular Biology*. https://doi.org/10.1016/j.pbiomolbio.2022.07.004

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52-64. https://doi.org/10.1080/01621459.1961.10482090

Fotso, S. (2018). Deep neural networks for survival analysis based on a multi-task framework. *arXiv Preprint arXiv:1801.05512*. https://doi.org/10.48550/arXiv.1801.05512

Freund, Y. (1990). Boosting a weak learning algorithm by majority. In M. FULK & J. CASE (Eds.), *Colt proceedings 1990* (pp. 202-216). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-146-8.50019-9

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G., Huttenhower, C., & Waldron, L. (2013). curatedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. *Database*, *2013*. https://doi.org/10.1093/database/bat013

Gijbels, I. (2010). Censored data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(2), 178-188. https://doi.org/10.1002/wics.80

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*(17-18), 2529-2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5

Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361-387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4%3C361::AID-SIM168%3E3.0.CO;2-4

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC press. https://doi.org/10.1201/b18401

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., & Boulesteix, A. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, *22*(3), bbaa167. https://doi.org/10.1093/bib/bbaa167

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). *Random survival forests*. https://doi.org/10.1214/08-AOAS169

Ishwaran, H., Kogalur, U. B., Chen, X., & Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *4*(1), 115-132. https://doi.org/10.1002/sam.10103

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data.* (2nd Ed.). John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457-481. https://doi.org/10.1080/01621459.1958.10501452

Karim, Md. R., & Islam, M. A. (2019). *Reliability and Survival Analysis*. Springer Singapore. https://doi.org/10.1007/978-981-13-9776-9

Khan, F. M., & Zubek, V. B. (2008). Support vector regression for censored data (SVRc): A novel tool for survival analysis. *2008 Eighth IEEE International Conference on Data Mining*, 863-868. https://doi.org/10.1109/ICDM.2008.50

Kim, H., Park, T., Jang, J., & Lee, S. (2022). Comparison of survival prediction models for pancreatic cancer: Cox model versus machine learning models. *Genomics & Informatics*, *20*(2). https://doi.org/10.5808/gi.22036

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data* (2nd ed). Springer.

Liang, F., Hatcher, W. G., Liao, W., Gao, W., & Yu, W. (2019). Machine learning for security and the internet of things: The good, the bad, and the ugly. *IEEE Access*, *7*, 158126-158147. https://doi.org/10.1109/ACCESS.2019.2948912

Liu, C., Chen, Y., Deng, Y., Dong, Y., Jiang, J., Chen, S., Kang, W., Deng, J., & Sun, H. (2019). Survival-based bioinformatics analysis to identify hub genes and key pathways in non-small cell lung cancer. *Translational Cancer Research*, *8*(4). https://tcr.amegroups.org/article/view/30209

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., & Klatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North central cancer treatment group. *Journal of Clinical Oncology*, *12*(3), 601-607. https://doi.org/10.1200/JCO.1994.12.3.601

Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics*, *108*, 1-8. https://doi.org/10.1016/j.ijmedinf.2017.09.013

Moncada-Torres, A., Maaren, M. C. van, Hendriks, M. P., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, *11*(1), 6968. https://doi.org/10.1038/s41598-021-86327-7

Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, *135*(2), 185-198. https://doi.org/10.2307/2344317

Richter, A. N., & Khoshgoftaar, T. M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine*, *90*, 1-14. https://doi.org/10.1016/j.artmed.2018.06.002

Salerno, S., & Li, Y. (2023). High-dimensional survival analysis: Methods and applications. *Annual Review of Statistics and Its Application*, *10*(1), 25-49. https://doi.org/10.1146/annurev-statistics-032921-022127

Shih, J., & Emura, T. (2021). Penalized cox regression with a five-parameter spline model. *Communications in Statistics-Theory and Methods*, *50*(16), 3749-3768. https://doi.org/10.1080/03610926.2020.1772305

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Thackham, M. (2022). *Survival analysis: Applications to credit risk default modelling* [PhD thesis, Macquarie University]. https://doi.org/10.25949/19436723.v1

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L.-J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, *30*(10), 1105-1117. https://doi.org/10.1002/sim.4154

Van Belle, V., Pelckmans, K., Suykens, J. A., & Van Huffel, S. (2008). *Survival SVM: A practical scalable algorithm*. In: Proceedings of the 16th European Symposium on Artificial Neural Networks (pp. 89-94).

Wang, P., Li, Y., & Reddy, C. K. (2019). *Machine learning for survival analysis: A survey*. *51*(6). https://doi.org/10.1145/3214306

Xiao, J., Mo, M., Wang, Z., Zhou, C., Shen, J., Yuan, J., He, Y., & Zheng, Y. (2022). The application and comparison of machine learning models for the prediction of breast cancer prognosis: Retrospective cohort study. *JMIR Medical Informatics*, *10*(2), e33440. https://doi.org/10.2196/33440

Ye, J., & Liu, J. (2012). Sparse methods for biomedical data. *ACM Sigkdd Explorations Newsletter*, *14*(1), 4-15. https://doi.org/10.1145/2408736.2408739

Yu, C., Greiner, R., Lin, H., & Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*, *24*.

Zhang, Y., Wong, G., Mann, G., Muller, S., & Yang, J. Y. (2022). SurvBenchmark: Comprehensive benchmarking study of survival analysis methods using both omics data and clinical data. *GigaScience*, *11*, giac071. https://doi.org/10.1093/gigascience/giac071

Zhao, L., & Feng, D. (2019). Dnnsurv: Deep neural networks for survival analysis using pseudo values. IEEE *Journal of Biomedical and Health Informatics*, *24*(11), 3308-3314. https://doi.org/10.1109/JBHI.2020.2980204