SurvBenchmark: comprehensive benchmarking study of survival analysis methods using both omics data and clinical data

Yunwei Zhang[1,2,] Germaine Wong[3,4,5], Graham Mann[6], Samuel Muller[1,7^,] Jean Y.H. Yang[1,2*^]

[1] School of Mathematics and Statistics, The University of Sydney, Sydney, Australia
[2] Charles Perkins Centre, The University of Sydney, Sydney, Australia
[3] Sydney School of Public Health, The University of Sydney, NSW, Sydney, Australia.
[4] Centre for Kidney Research, Kids Research Institute, The Children's Hospital at Westmead, NSW, Sydney, Australia.
[5] Centre for Transplant and Renal Research, Westmead Hospital, NSW, Sydney, Australia.
[6] John Curtin School of Medical Research, Australian National University, Canberra, Australia
[7] Department of Mathematics and Statistics, Macquarie University, Sydney, Australia

^ Equal contribution

*To whom correspondence should be addressed.

Abstract

Survival analysis is a branch of statistics that deals with both, the tracking of time and of the survival status simultaneously as the dependent response. Current comparisons of the performance of survival models mostly focus on classical clinical data with traditional statistical survival models, with prediction accuracy being often the only measurement of model performance. Moreover, survival analysis approaches for censored omics data have not been fully studied. The typical solution is to truncate survival time, to define a new status variable, and to then perform a binary classification analysis.

Here, we develop a benchmarking framework that compares survival models for both clinical datasets and omics datasets, and that not only focuses on classical statistical survival models but also incorporates state-of-art machine learning survival models with multiple performance evaluation measurements including model predictability, stability, flexibility and computational issues. Our comprehensive comparison framework shows that optimality is dataset and analysis method dependent. The key result is that there is no one size fits all solution for any of the criteria and any of the methods. Some methods with a high C-index suffer from computational exhaustion and instability. The implications of our framework give researchers an insight on how different survival model implementations vary over real world datasets. We highlight that care is needed when selecting methods and recommend specifically not to consider the C-index as the only performance evaluation metric as alternative metrics measure other performance aspects.

Code availability: https://github.com/SydneyBioX/SurvBenchmark

Contact: jean.yang@sydney.edu.au

1. Introduction

Survival models are statistical models designed for data that have censored observations, that is time-to-event data, which are commonly found in health, biological organisms and mechanical systems. We will follow the terminology in survival analysis and capture the event of interest through a 'status' variable, which is typically a binary class outcome; we call the waiting time to this status event the 'survival' time. This class of models has wide applicability as it can be applied to study the duration of marriage in sociology, the duration of employment in labor economics and much more in addition to the survival type in the datasets considered in this article. Therefore, survival models target both outcomes, whereas neither regression analysis on time nor classification analysis on status explain both simultaneously (Schober and Vetter, 2018b).

Multiple survival models have been developed over the last decades and there are many good reviews and comparisons of those models in the literature. However, there is a lack of comparison studies from a practical viewpoint, that is a lack of extensive real world datasets comparisons, particularly in the biomedical field. This motivates us to develop a benchmarking framework for that viewpoint to provide a better understanding of survival models and to inform on how to guide decision making. To better understand what has been done, we perform an exhaustive search for various types of available survival analysis methods together with performance evaluations for different types of datasets.

For comparison studies that include real world datasets in health, we found that they typically have specific focus such as on a certain disease (e.g. colon or pancreatic cancer), or on a certain data platform (e.g. omics or clinical). For example, Schober and Vetter (Schober and Vetter, 2018a) and Ahmed et al. (Ahmed et al., 2007) conduct reviews on traditional survival models such as the Kaplan-Meier (KM) method and the Cox Proportional Hazards (CoxPH) model with a focus on clinical data with an induce anesthesia state and a specific colon cancer type, respectively. Lee and Lim (Lee and Lim, 2019) apply the penalized Cox model, survival support vector machine, random survival forest (RSF), and Cox boosting models on large genomic data. To date, no systematic review encompasses datasets obtained from multiple disease types and includes both, clinical categorical and clinical continuous data, as well as large omics data.

We also conclude from these recent comparison studies that they are either within traditional models (KM method, CoxPH model) or within modern machine learning (ML) methods. With the emergence of different modelling approaches from various disciplines addressing challenges such as different model assumptions, a recent comprehensive survey article by Wang et al. (Wang et al., 2019) summarizes three categories of statistical survival models and ML methods with a focus on theoretical mathematical detail. However, their study does not discuss practical implications of the various methods, in particular there is no comparison of performance using real world datasets. There is a need for better guidance on what data analysis strategy to use.

To this end, we develop a benchmarking framework: SurvBenchmark that considers multiple aspects with several evaluation metrics on a large collection of real world health datasets to guide method selection and new method development.

2. Survival models and their evaluation

Survival models can deal with data that explain censored observations with a bivariate outcome variable, consisting of 'time' (the minimum of 'time-to-event' and 'censoring time') and 'event' (often categorized as "class 1" meaning survival event did occur, and "class 0" otherwise). There are two key features of such censored survival objects. One, the class label "0" (referring to a "no-event") does not mean "class 0"; instead, it represents that the event-outcome is "unknown", or, using the technical term, "censored". Two, an additional tracking time measurement is included as part of the response. In this section, we briefly introduce survival models together with their evaluation metrics.

There are two main branches of survival models: classical statistical survival models, which include parametric models, nonparametric models and semi-parametric models; and modern ML survival models, which include ensemble based methods and state-of-the art deep learning based approaches. Both sets of models are briefly reviewed in the following two sections.

2.1 Classical survival models

The Cox Proportional Hazards (CoxPH) model (Cox, 1972) is the most widely used classical survival model. CoxPH works on the hazard function which is the multiplication of a baseline hazard function and the exponential of a linear combination of covariates in the data and the estimated parameters.
The hazard function is given by

$$h(t,x) = h_0(t)\, e^{(\sum_{j=1}^{p} \beta_j x_j)}, \qquad (1)$$

where $x = (x_1, x_2, \ldots, x_p)$ is the covariate vector and $h_0(t)$ is the baseline hazard function.

This is a semi-parametric model as the baseline hazard function is canceled out when taking the ratio of two hazard functions.

The penalized Cox model is another extension of the CoxPH model that helps to prevent overfitting. The L1 regularized CoxPH model adds a scaled sum of absolute values of the magnitude of model coefficients, that is $\lambda_1 \sum_{j=1}^{p} |\beta_j|$ as the regularization term to the partial log-likelihood. Other regularizers can be used such as L2 regularization, that is $\lambda_2 \sum_{j=1}^{p} (\beta_j)^2$, or other scaled sums of non-negative penalties of the $\beta's$, such as in the following general penalized partial log-likelihood:

$$\log(L(\beta)) - \lambda \sum_{j=1}^{p} \pi(\beta_j), \qquad (2)$$

where $L(\beta)$ is the partial likelihood as for example given in Tibshirani ((Tibshirani, 1997), Equation 2) and then optimization takes place (Van Houwelingen, 2004; Do et al., 2013; Huang and Liang, 2018). Using the L1 penalty in Equation (2) gives the Lasso Cox estimation and using the L2 penalty gives the Ridge Cox solution, respectively. If instead of a single regularization term we consider a weighted average of the L1 and L2 penalty, we obtain the ElasticNet Cox model. One remarkable characteristic of the Lasso Cox model and the Elastic Net Cox model is that they can simultaneously perform feature selection and prediction, because some of the beta parameters can be penalized all the way to 0 when maximizing Equation (2).

2.2 Modern machine learning models

Recent years have seen a strong surge in the use of modern ML methods as a result of their exceptional performance, in health and in many other areas, such as in finance (Gogas and Papadimitriou, 2021), the environment (Chen et al., 2017) and internet of things (Lakshmanaprabu et al., 2019). Notable examples in health include the application of Random Survival Forest (RSF) on complex metabolomics data (Dietrich et al., 2016) and the application of SurvivalSVM to the survival of prostate cancer patients (Van et al., 2008). Both approaches are survival analysis extensions to two widely used ML algorithms (Random Forest and Support Vector Machine) for binary classification.

SurvivalSVM was developed by Van Belle and colleagues (Van Belle et al., 2011) for time-to-event data. It is a variant of the regularized partial log-likelihood function (2) above but has a different penalty term. In contrast to using $\lambda \sum_{j=1}^{p} \pi(\beta_j)$, SurvivalSVM uses penalized splines and then applies both, ranking constraints and regression constraints to the corresponding partial log-likelihood function. SVM with those constraints enables models for high-dimensional omics data to have more flexible structure, e.g. additive (non)-linear models. One distinct feature of SurvivalSVM is that it treats the prognostic problem as a ranking problem and therefore, the estimation of the hazards is not directly incorporated in the model.

RSF was first proposed by Ishwaran (Ishwaran et al., 2008) as an extension of Random Forest to censored survival data. Random Forest (Breiman, L., 2001) is a non-parametric bagging based ensemble learning method that adds variation in the training datasets by bootstrapping the data. Multiple models are generated based on various bootstrapped datasets, and the ensemble prediction result is an average of these multiple models or the result of a majority vote. The key components in our application of RSF are that we use Harrell's C-index to evaluate the survival tree instead of the mean square error for regression problems or confusion matrix for classification problems, and that we use the log-rank score in each node as the stopping rule.

Another ensemble based approach is the boosting method, which contains multiple learners and sequentially gives more weight to weak learners to enhance predictability. For example, the Cox boosting model (Bin and De Bin, 2016; Binder et al., 2013, 2009; Binder and Schumacher, 2009) is developed based on Cox models with boosting being applied to the estimation of the regression beta parameters in Equation (1). There are two popular approaches to update β: the first is the model-based approach that leads to the mboost method, the second is the likelihood based approach that leads to the CoxBoost method (benchmarked in this study).

These models so far only focus on optimizing a single objective. Because survival data is time dependent, it is natural to have multiple tasks related to one or more time points of interest. This naturally leads to multi-task learning, a method that deals with the need to predict for more than a single response variable, based on joint optimization of multiple likelihood functions corresponding to each task. The multi-task logistic regression model (MTLR) by Yu et al. (Yu et al., 2011) is a survival model for multiple time points, where for each, the task is to predict survival using a logistic regression model and the parameters from each model are estimated simultaneously in the maximization of the joint likelihood function.

More recently, the ML and artificial intelligence (AI) communities refer to the methods above as classical ML methods due to the emergence of deep learning (DL), a conceptual advancement based on neural networks (NN). In survival analysis, recently a number of DL survival models were developed such as Cox-nnet (Ching et al., 2018a), DeepSurv (Katzman et al., 2018) and DeepHit (Ryu et al., 2020). The key concept in DL survival models is having different loss functions that particularly target either the hazard or the survival probability for those neurons in hidden layers when building the DL architecture. High dimensional complex biological information can be better represented with the application of those hidden layers (Ching et al., 2018b) and through relaxing the proportional hazard assumption.

2.3 Feature selection methods applied to survival models

Fundamental to any ML model is feature selection. A critical component for the models is the input features. In parallel to having a good prediction model, it is important to understand the drivers (features) that are behind the good model performance. Wrapper and filter (Bagherzadeh-Khiabani et al., 2016) are two types of widely used feature selection methods applied together in different types of survival models. We briefly review traditional and advanced methods used in this benchmark work here.

Traditional statistical sequential feature selection methods such as stepwise selection methods, including backward or forward selection with different criteria such as AIC or BIC, are wrapper methods used with the Cox model. With the emergence of omics data in the 1990s, the statistics community embraced the development of differential expression (DE) analysis that is a filter type feature selection method to select promising genes. Expression levels are usually measured by log-fold-change that can vary for omics data and due to the large p small n framework, "parallel univariate strategies" have been used (such as Wilcoxon-Mann-Whitney or t-test) to identify DE genes when comparing two groups.

Further, there are some advanced modern filter feature selection methods for omics data. The genetic algorithm (GA) is one filtering method that is inspired by Charles Darwin's natural evolution theory; the selection of features here is actually an optimization problem which is a binary classification problem that uses linear discrimination analysis in the design (Saeys et al., 2007; Coombes, Kevin R, 2017). GA was developed by John Holland (Holland et al., 1992) and two major steps are crossover, which exchanges the genes of parents and mutation, which adds diversity in the next generation.

2.4 Classical performance evaluation metric for survival data

Classically, survival analysis is evaluated in three broad settings: the concordance index, the Brier score and the time-dependent AUC. Similar to evaluating classification and regression models, metrics for calibration and discrimination frameworks are developed with incorporating censoring by applying rank based methods or error based methods together with a weighting scheme.

2.4.1 C-index and its extension in survival analysis

C-indices in survival analysis are concordance based methods where 'concordance' measures how close a prediction is to the truth. The first C-index for survival analysis was introduced by Frank. E. Harrell (Harrell, 1982), as a performance measure that does not depend on time. A C-index can range from 0 to 1, where 1 means perfect performance and 0 means worst possible performance. If a model would not take into account any information from the data, that is a random prediction is made, then the corresponding C-index would be around 0.5. For most clinical datasets, a C-index around or larger than 0.6 is considered an acceptable prediction. Harrell's C-index (Newson, 2006) defines concordance by looking at ranks of pairs of subjects in the data (there are n choose 2 pairs for data with n subjects). Harrell's C index further depends on the censoring distribution of the data, is motivated by

Kendall's tau statistic and is closely related to Somers' D. When ranking the subjects, censored subjects are excluded; and pairs included in the formula are only those comparable, non-censored pairs. There are different versions of the C-index, where the differences come from the different ways that censored subjects are ranked.

We will use the following three C-indices. First, Begg's C-index (Marchevsky and Wick, 2011) uses KM estimators to incorporate both censored and uncensored subjects by assigning different weights to them. When we have a high percentage of censoring, there might be a high bias for this evaluation metric. Second, Uno et al (Uno et al., 2011; Harrell et al., 1984) develop a new way to calculate the rank with the help of inverse probability of censoring weight (IPCW). Third, the GH C-index (Gönen and Heller, 2005) changes the concordance function into a probability function based on the Cox model estimation and then approximates its distribution which is robust to censoring.

2.4.2 Brier score

The Brier score (Gerds and Schumacher, 2006) (Schmid et al., 2011) uses IPCW to handle censored subjects when measuring discrepancy between the estimated values and the actual values. This score can be considered as a similar measure to the mean squared error (MSE) in regression models to some extent. Like the MSE, the Brier score takes a value greater than 0 that depends on the data and the smaller the Brier score the better. However, to have better interpretability, the integrated Brier score (IBS) is introduced which also takes values between 0 and 1 - it averages the loss over time in situations where there is no interest in a particular time point but performance is with regards to all time points as a whole.

2.4.3 Time-dependent AUC

The time-dependent AUC is also inspired from binary classification model evaluations. The receiver operating characteristic (ROC) curve is the curve that plots the sensitivity versus the false positive rate. The area under the ROC curve is called AUC (area under the curve). In survival analysis, as we have multiple time points, the traditional AUC cannot be applied directly. Chambless and Diao (Chambless and Diao, 2006) were the first to propose a time-dependent AUC for survival analysis. They define the AUC(t) as the probability that a person with disease onset by time t has a higher score than the person with no event by time t. By specifying different time points, AUC values can be visualized through the time-dependent AUC plot.

3. Material and Methods

3.1 Datasets: six clinical and ten omics data sets

Clinical data - Six clinical datasets with different sample sizes and disease types are selected. Of these, three datasets are publicly available and the other three (ANZDATA, UNOS_Kidney data and Melanoma data as described below) can be requested (see references in Table 1).

- Veteran data (Kalbfleisch and Prentice, 2002) is a survival dataset from the randomised trial of two treatment regimens for lung cancer and the data is obtained from the R "survival" package via the function "data(veteran)". There are 6 measured features in this data.
- PBC data (Fleming and Harrington, 2005) from the Mayo Clinic trial in primary biliary cirrhosis (Pbc) of the liver conducted between 1974 and 1984; it can be obtained from the R "RandomForestSRC" package via function "data(pbc)". This data has 5 clinical features with a total of 312 patients.
- Lung data (Loprinzi et al., 1994) contains patient survival information with advanced lung cancer from the North Central Cancer Treatment Group and is available from the R "survival" package via the function "data(lung)". This data has 7 features with a total of 228 patients.
- ANZ data (ANZDATA), this is the Australia & New Zealand Dialysis and Transplant Registry data containing graft survival information and electronic clinical records for kidney transplantation recipients in Australia and New Zealand from 30th June 2006 to 13th November 2017. This data contains records for both living and deceased donors and also multi-organs transplants. We processed the raw data, restricting the transplant date to be after 2008-09-18 and retained deceased donor kidney transplants only. Missing records are excluded and this leaves us with 3323 patients and 38 features containing patient, donor and donor-recipient human leukocyte antigen (HLA) compatibility.
- UNOS_Kidney data is the organ transplant data based on the Organ Procurement and Transplantation Network (OPTN)-United Network for Organ Sharing (UNOS) in the US (based on OPTN data as of March, 2020). We selected a random sample of 3000 records (setting the seed function in R to 202001) associated with deceased donor kidney transplantation only with 99 features containing recipients, donors and donor-recipient HLA compatibility. Missing values are imputed using the R package "MICE".
- Melanoma_clinical data, extracted from melanoma data (Mactier et al., 2014; Mann et al., 2013) which is a melanoma disease in-house dataset collected as part of multi-omics study (Wang et al., 2020), is the part that contains clinical information for patients. After deleting all missing values, we have 88 patients with stage three melanoma disease measured by 14 clinical features.

Omics data - We consider eight published data and two in-house melanoma cancer omics datasets.

- Two ovarian cancer gene expression datasets are downloaded from the R package "curatedOvarianData" (Ganzfried et al., 2013). The curation of this data collection closely follows from Waldron et al. (2014) and the analysis pipeline from Yoshihara et al. (2012). Ovarian1 is obtained using the R function "data("GSE49997_eset")" with 194 patients and 16047 genes and Ovarian2 is the "GSE30161_eset" data with 58 patients and 19816 genes.
- Another six gene expression datasets are available online from Li and colleagues' paper. (Li et al., 2016; Sorlie et al., 2003; Beer et al., 2002; Bullinger et al., 2004; van de Vijver et al., 2002; van 't Veer et al., 2002; van Houwelingen et al., 2006). We named these data as GE_1, GE_2, GE_3, GE_4, GE_5 and GE_6. For GE_3, log2 transformation is applied, followed by a KNN imputation with 10 nearest points. For GE_6, median normalisation is applied. For others, no further pre-processing is performed by us.
- Melamona_itraq and Melanoma_nano are two melanoma omics datasets, the first is a protein expression dataset from the iTRAQplatform and the second is a Nanostring dataset from the above melanoma study and pre-processing steps are described in the respective papers. The itraq protein expression data has 41 patients with 640 proteins. The nanostring data has 45 patients with 204 nanostrings.

A summary of the size and censoring rate of all datasets can be found in Table 1.

3.2 Benchmarking framework/procedure

Benchmarking methods: All methods evaluated are described in detail in the Supplementary Excel file.

Evaluation metrics: We examine model performance metrics that can be broadly grouped into 4 categories and assessing performance in terms of each methods' flexibility, predictability, stability and computational efficiency (detailed in Table 2).

We measure model flexibility by looking at whether a given method can handle different data modality, different level of sparsity, represents multiple ways including the type of data required (clinical, omics), type of input required (categorical, numerical), sparsity of the data allowed (yes, no) and prediction ability evaluation metrics allowed.

We measure model predictability using three different metrics: C-index, time-dependent AUC and Brier score. We apply 4 different modified versions of C-index: Harrell's C-index, Begg's C-index, Uno's C-index and GH C-index. For identification of different time points, we equally divided the survival time ranging from the 1st quartile to the 3rd quartile into 15 time points for each dataset, and therefore, we obtained 15 AUC values corresponding to each time point. As for the Brier score, we calculated the raw Brier score and the IBS.

We measure model computational efficiency using both computational time and memory. Computational time is calculated using the "Sys.time" function in R. Memory is calculated using the "Rprof" function in R and the total memory used is summarized for each experiment.

We measure model stability using model reproducibility and the standard deviation (SD) of model predictability metrics. Model reproducibility is defined as the proportion of successful runs among all the runs attempted. For each model predictability metric, we calculated its SD. We then ranked the values for all the methods for each dataset from the most stable (smallest SD) to the least stable (largest SD).

All compared methods (Supp Table 1) and evaluation metrics (Table 2) are applied and evaluated on those listed real world datasets (Section 3.1). We apply 20 times (runs) repeated 5-fold cross validation using RStudio server (based on R software) with 15 cores in parallel. For each run, the whole data is split into a training dataset (80%) and a testing dataset (20%) with each method trained using the training dataset and values of evaluation metrics calculated using the testing dataset. Detail about the packages and parameters can be found in (Supp Table 1) and functions used to evaluate the methods are shown in (Table 2).

4. Results

4.1 SurvBenchmark framework for comprehensive benchmarking

To comprehensively evaluate the strength and weaknesses of the various survival analysis approaches, we select 20 methods from the extensive literature and examine their performance when applied to 16 datasets, including both clinical and omics data. The performance of each method is measured against 11 metrics representing multiple aspects including model feasibility, model predictability, model stability and computational efficiency. There are three key aspects of our comparison framework SurvBenchmark as depicted in Figure 1: (i) Practical oriented framework which is applied to a broad range of datasets and which includes a taxonomic methods system to evaluate multiple aspects; (ii) Extensive comparison of methods from classical to state of the art machine learning (ML) approaches (terminologies used in this study includes classical Cox-based methods: Cox, Cox_bw_AIC, Cox_bw_BIC, Cox_bw_p, Lasso_Cox, Ridge_Cox, EN_Cox; Cox-based modern ML methods: CoxBoost, CoxBoost (GA), CoxBoost(DE); Cox-based methods: both classical Cox-based and modern Cox-based; and MTLR-based methods: MTLR, MTLR(GA), MTLR(DE)); (iii) Comprehensive evaluation of the model performance with the utilisation of a customizable weighting framework.

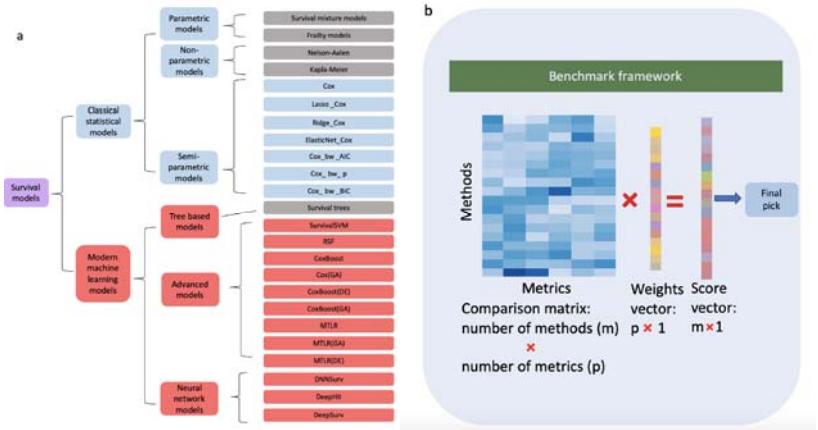

Figure 1: SurvBenchmark--Schematic view of our benchmark framework.
(a) An overview of survival models. We broadly classify current models into two categories; traditional statistical models (top group) and modern machine learning models (bottom group). Each of these categories can be further subdivided as presented in the hierarchical chart. All models in blue and red colored boxes are implemented in this current benchmark study.
(b) A graphical representation of the SurvBenchmark framework. The methods and evaluation metrics are summarized in a matrix with a flexible user defined weights vector.

4.2 Practical consideration in assessing model performance

Many comparison studies define method performance solely in terms of method predictability, with only a few studies taking into account computational time. Often the feasibility of the method is not properly considered or discussed. Practically, it is paramount that a method can be applied to the data at hand, based on both the flexibility (data modality, sparsity) and computational requirement.

Given the diverse collection of data characteristics that is now available in the biomedical field, not all survival approaches are feasible to be applied to all data types. For example, some classical Cox models (Figure 2a, top left from column 1 to 10, row 1 to 4; a blue box indicates 'not feasibility' of the method) cannot handle large p (features) small n (samples) datasets (such as GE-1) which is a distinct feature of any molecular (omics) study. Advanced feature selection methods together with ML survival models such as CoxBoost(DE) can only take numerical data as the input (purple box for input type, where model characteristics are coded using 0, 1 and 2 with questions defined as below. Is input type numeric only? Yes: numerical only. No: both numerical and categorical are ok. Is output type survival risk? Yes: survival risk. No: survival probability. Can the model handle n<p situation? Yes: it can. No: it cannot. The other case: output is the rank of survival risk.).

Next, we look at the computational aspect, and we notice that deep learning based methods are computationally inefficient. The star icon, which reflects that the method takes a long time and requires a lot of memory, indicates this (Figure 2a). From the many stars, we observe that RSF(default) (5 stars) and MTLR (4 stars) are not as computationally efficient as Cox-based approaches such as Lasso_Cox (1 star) and CoxBoost (no stars).

Lastly, a summary tabulating the feasibility associated with each of the evaluation metrics for prediction is provided in Figure 2b. The results highlight that Begg's C-index and GH C-index are applicable only for Cox methods (red boxes indicate feasible), that the integrated Brier score can be calculated for Cox model and RSF (red), and that the Brier score cannot be calculated for SurvivalSVM (blue).

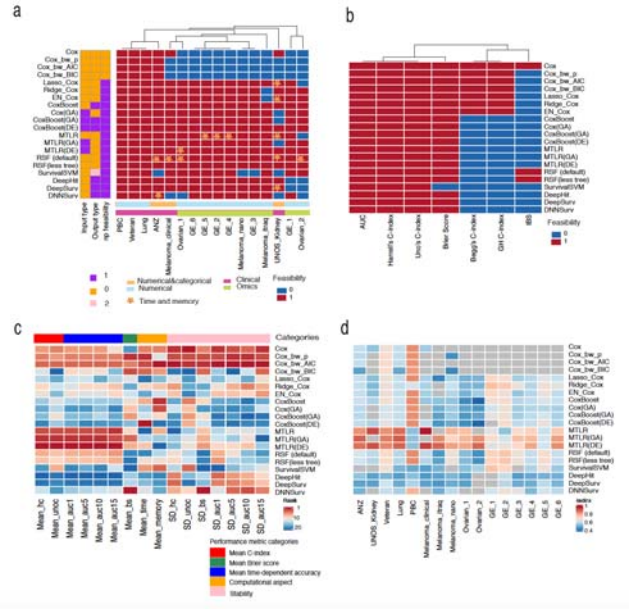

Figure 2: Summary heatmaps
(a) Summary for method flexibility and computational efficiency. (row: methods, column: datasets, feasible: red, not feasible: blue, feasible, star: memory and time consuming, model characteristics: yes, 1, no, 0, the other case, 2) (b) Prediction ability evaluation metric flexibility. (row: methods, column: prediction ability evaluation metrics, feasible: red, not feasible: blue) (c) Rank heatmap for method overall performance (row: methods, column: performance metrics, red to blue: top to bottom rank) (d) Harrell's C-index heatmap (row: datasets, column: methods)

4.3 Performance evaluation from multiple perspectives: no 'one size fits all'

To achieve a comprehensive overview of different survival approaches, we assess method performance from multiple perspectives across a large collection of datasets. Here, we color the methods according to their performances for all three broad categories: model predictability, model stability and computational efficiency (Figure 2c shows ranks of those methods where red means the best and blue the worst; similarly, Figure 2d shows Harrell's C-index values with red referring to high values and blue to small values). We find that no method performs optimally across all three categories and there are various trade-offs among the categories.

For model predictability, we use seven different measures based on C-index, Brier score and time-dependent AUC. Here, MTLR-based approaches perform significantly better than others, which is most apparent by looking at the performance results using C-index and time-dependent AUC. In order to further examine whether MTLR-based approaches have similar performance across all datasets, we show our examination on one specific criteria (the most popular Harrell's C-index; Mean_hc). In Figure 2d we demonstrate that MTLR has optimal performance for all but one of the six clinical datasets with PBC having optimal performance for one of the clinical datasets. Variants of MTLR (MTLR(GA) and MTLR(DE)) outperformed MTLR when applied to any of the ten omics datasets suggesting the performance of the approaches depend on the type of dataset.

For computational efficiency as measured by computational time and memory usage, the best performing methods are classical Cox-based models and CoxBoost. In particular, Cox, Cox_bw_AIC and Cox_bw_BIC are the top three performing methods for computational time (Figure 2c). For model stability, we have seven criteria and they are based on calculating the standard deviation (SD) of predictability metrics described above. Similar to the the computational efficiency performance, when using SD-criteria, Cox, Cox_bw_AIC and Cox_bw_BIC are also the top 3 performing methods in all but one criteria, the exception is the standard deviation of Brier score (SD_bs), where DNNSurv ranks first suggesting its ability to discriminate survival probabilities for different observations.

Despite having the best performance in both computational efficiency and stability for Cox-based methods, their predictability falls behind MTLR-based methods. On the other hand, MTLR-based methods clearly have the best predictability but they are not efficient and stable enough (He and Sun, 2015).

In conclusion, these observations demonstrate that no method performs optimally for all those categories.

4.4 Cox-based modern ML methods have similar prediction performance compared to classical Cox-based methods

To understand the gain in model predictability from Cox-based modern ML methods (CoxBoost, Coxboost (GA)), we compare those models with traditional Cox-based methods (Lasso_Cox, EN_Cox) which are used as a gold standard method in many studies. Our results indicate that they have similar performance (Figure 3) across a large collection of datasets. For example, in the ANZ data, which is a representative clinical dataset, we observe similar model predictability measured by both Harrell's C-index and Brier score. For GE_5, a representative dataset of omics with large p small n data characteristics, the same conclusion is drawn. This suggests the performance of modern ML methods in complex health and clinical data is not as clear cut as in some other domains.

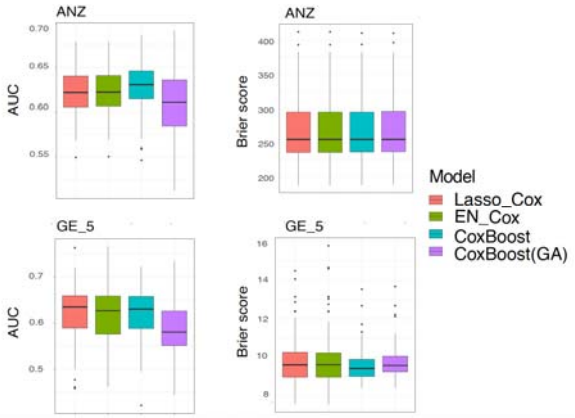

Figure 3: Prediction ability for cox-based methods.
Top left: Harrell's C-index on ANZ data. Top right: Brier score on ANZ data. Bottom left: Harrell's C-index on GE_5. Bottom right: Brier score on GE_5.

4.5 Data dependent model performance related to short and long time prediction

To discover the model performance over time, we visualize the model performance using the time-dependent AUC curves for all methods. Here we observe among two representative clinical datasets (PBC, UNOS_Kidney) and two omics datasets (GE_2, GE_4) in Figure 4, not all curves are parallel to each other, indicating that the behaviour of model predictability for different time points is data dependent (see supplementary plot 1 for further results). For UNOS_Kidney data and GE_2 data, there is consistency of short-time, medium-time and long-time prediction as the curves are approximately horizontal. In contrast, for PBC data and GE_4 data, methods such as RSF, CoxBoost(GA) possess different predictability along those time points.

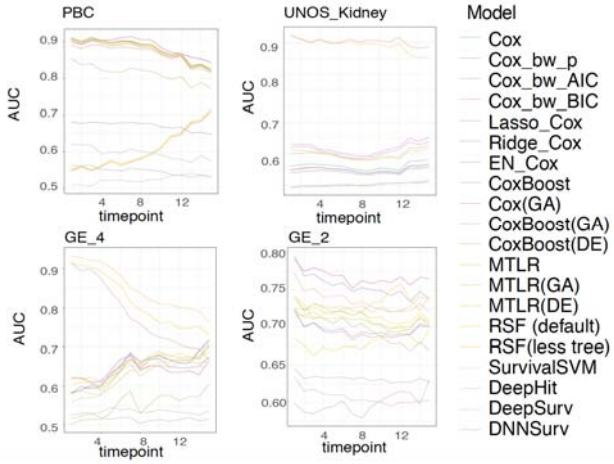

Figure 4: Time-dependent AUC curves.
Top left: PBC data Top right: UNOS_US data Bottom left: GE_4 data. Bottom right: GE_2 data.

5. Discussion

This benchmark study comprehensively evaluated the relevance and usefulness of survival models in practice, where emphasis was placed on the performance in diverse real life datasets. In our review we assessed a broad variety of survival methods from classical Cox-PH models to modern ML models including DNNSurv. The findings of our systematic assessment will provide specific guidance for translational scientists and clinicians, as well as define areas of potential study in both survival methodology and benchmarking strategies.

5.1 Shift from hazard function to survival function

In recent years, there is a clear shift in how survival data is analysed, from modelling directly the hazard function to building models directly on survival functions. Conceptually, modelling hazard functions is a good way to identify key risk factors related to various patients' risk levels. On the other hand, if our key criterion is to produce an accurate survival probability prediction (predictability), modelling survival probability instead of hazard function shows the improvement in predictability. Methods including MTLR, DNNSurv and SurvivalSVM which directly model the survival function showed better performance in terms of model predictability, this is consistent with what Li and colleagues (Yu et al., 2011) have commented on when discussing the performance of their proposed MTLR method.

5.2 Discussion of MTLR

It is striking that MTLR shows remarkable model predictability in our benchmark study and we now highlight technical advantages, disadvantages as well as its applications. Numerous reasons could contribute to the better model prediction performance observed in our study with MTLR-based approaches. These include the three main reasons as discussed by Li and colleagues (Yu et al., 2011): the direct modeling of the survival function, simultaneous building of multiple logistic regression models and the dynamic modeling. Interestingly, the majority of extended MTLR models since 2011 are based on neural networks as researchers extend the concept to account for nonlinearity in datasets (Fotso, 2018). To date, only a limited number of studies have applied MTLR in health using clinical data in HIV patients (Bisaso et al., 2018) or on large omics datasets to predict patient survival in Breast Invasive Carcinoma and Kidney Renal Clear Cell cancers (Wang et al., IEEE 2017). Given its outstanding model predictability observed for most of the datasets in our study, we believe there may be an opportunity to use MTLR more widely for survival risk modelling in Health contexts.

5.3 Suggestion to look at multiple indices

Model predictability is one of the key metrics to assess survival studies with Harrell's C-index being currently the most popular model assessment metric. As this kind of ranking based concordance measurement is suitable to evaluate predicted outcomes with censored data, various concordance indices are developed using different methods to handle censoring such as Uno's C-index using IPCW. Besides concordance indices, other predictability metrics such as the time-dependent AUC, which applies a similar idea as the AUC in binary classification problem but needs to divide the whole time interval into multiple time points, are also adopted in some survival studies (Li et al., 2020). Given that the model predictability could actually be measured by multiple types of indices, we suggest that hybrid evaluation metrics should be applied in practice to provide relatively comprehensive assessments for the fitted model.

5.4 Models targeting clinical data vs omics data

Though many survival approaches are applicable to both clinical and omics data, there are a number of recently developed approaches that are specifically tailored for high dimensional omics data, such as CoxBoost. The rationale behind developing data-specific methods is to better capture the distinct data characteristics in either the clinical or omics studies. Clinical data such as electronic health records (EHR) usually include mixed modality variables, large sample sizes but has a relatively small number of features (i.e. large n and small p). In contrast, omics data naturally comes with large p (large collection of molecular features) and small sample size (small n) but their data type is homogenous. When it comes to various real life datasets, performances are also affected by many other aspects besides data type (clinical, omics) such as data modality and therefore, it is challenging to directly examine whether those tailored methods indeed improve the performance.

5.5 DNNSurv reproducibility

Unlike all other approaches, there is some concern over reproducibility of the results for the method DNNSurv in our empirical work. We found that the DNNSurv failed to run for some samples of the data under our repeated cross-validation setting. More specifically, among all 100 runs, we observe that DNNSurv has a 100% completion rate for 5 out of the 12 applicable datasets (supplementary plot 2) only. For the remaining 7 datasets, the completion rate is around 80% and completion rate was as low as 63% for the Melanoma_itraq data. The reason behind such instability is likely due to parameter tuning requirements for most of the DNN based methods for data with small sample sizes (Shaikhina and Khovanova, 2017).

Acknowledgements

The authors thank all their colleagues, particularly at The University of Sydney, Sydney Precision Bioinformatics Alliance and Charles Perkins Centre for their support and intellectual engagement.

Funding

The following sources of funding for each author, and for the manuscript preparation, are gratefully acknowledged: Australian Research Council Discovery Project grant (DP170100654) to JYHY and SM. Research Training Program Tuition Fee Offset and Stipend Scholarship and the Dean's International Postgraduate Research Scholarship (DIPRS) to YZ. The funding source had no role in the study design; in the collection, analysis, and interpretation of data, in the writing of the manuscript, and in the decision to submit the manuscript for publication.

Conflict of Interest: none declared.

References

Ahmad,R. and Bath,P.A. (2004) The use of Cox regression and genetic algorithm (CoRGA) for identifying risk factors for mortality in older people. Health Informatics Journal, 10, 221–236.

Ahmed,F.E. et al. (2007) Modeling survival in colon cancer: a methodological review. Mol. Cancer, 6, 15.

Bagherzadeh-Khiabani,F. et al. (2016) A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J. Clin. Epidemiol., 71, 76–85.

Beer,D.G. et al. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med., 8, 816–824.

Binder,H. et al. (2009) Boosting for high-dimensional time-to-event data with competing risks. Bioinformatics, 25, 890–896.

Binder,H. et al. (2013) Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures. Statistics in Medicine, 32, 1778–1791.

Binder,H. and Schumacher,M. (2009) Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. BMC Bioinformatics, 10.

Bin,R.D. and 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.

Bisaso,K.R. et al. (2018) A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. BMC Med. Inform. Decis. Mak., 18, 77.

Breiman,L. (2001) Random Forest. Machine Learning, 45, 5–32.

Bullinger,L. et al. (2004) Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. New England Journal of Medicine, 350, 1605–1616.

Chambless,L.E. and Diao,G. (2006) Estimation of time-dependent area under the ROC curve for long-term risk prediction. Statistics in Medicine, 25, 3474–3486.

Chen,W. et al. (2017) A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. CATENA, 151, 147–160.

Ching,T., Zhu,X., et al. (2018) Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. PLoS Comput. Biol., 14, e1006076.

Ching,T., Himmelstein,D.S., et al. (2018) Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface, 15.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. (2011) Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11)., 1845–1853.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11). Curran Associates Inc., Red Hook, NY, USA, 1845–1853.

Coombes, K. R. (2017) Genetic Algorithms for Feature Selection.

Cox,D.R. (1972) Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological), 34, 187–202.

Dietrich,S. et al. (2016) Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. Int. J. Epidemiol., 45, 1406–1420.

Do,K.-A. et al. (2013) Advances in Statistical Bioinformatics: Models and Integrative Inference for High-Throughput Data Cambridge University Press.

Fleming,T.R. and Harrington,D.P. (2005) Counting Processes and Survival Analysis. Wiley Series in Probability and Statistics.

Fotso, S. (2018) Deep neural networks for survival analysis based on a multi-task framework. arXiv preprint arXiv:1801.05512.

Ganzfried,B.F. et al. (2013) curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database, 2013.

Gerds,T.A. and Schumacher,M. (2006) Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. Biometrical Journal, 48, 1029–1040.

Gogas,P. and Papadimitriou,T. (2021) Machine Learning in Economics and Finance. Computational Economics, 57, 1–4.

Gönen,M. and Heller,G. (2005) Concordance probability and discriminatory power in proportional hazards regression. Biometrika, 92, 965–970.

Harrell,F.E. (1982) Evaluating the yield of medical tests. JAMA: The Journal of the American Medical Association, 247, 2543–2546.

Harrell,F.E. et al. (1984) Regression modelling strategies for improved prognostic prediction. Statistics in Medicine, 3, 143–152.

He,K. and Sun,J. (2015) Convolutional neural networks at constrained time cost. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Holland,J.H. et al. (1992) Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence MIT Press.

van Houwelingen,H.C. et al. (2006) Cross-validated Cox regression on microarray gene expression data. Stat. Med., 25, 3201–3216.

Huang,H.-H. and Liang,Y. (2018) Hybrid L1/2 2 method for gene selection in the Cox proportional hazards model. Computer Methods and Programs in Biomedicine, 164, 65–73.

Ishwaran,H. et al. (2008) Random survival forests. The Annals of Applied Statistics, 2, 841–860.

Kalbfleisch,J.D. and Prentice,R.L. (2002) The Statistical Analysis of Failure Time Data. Wiley Series in Probability and Statistics.

Katzman,J.L. et al. (2018) DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med. Res. Methodol., 18, 24.

Lakshmanaprabu,S.K. et al. (2019) Random forest for big data classification in the internet of things using optimal features. International Journal of Machine Learning and Cybernetics, 10, 2609–2618.

Lee,S. and Lim,H. (2019) Review of statistical methods for survival analysis using genomic data. Genomics Inform., 17, e41.

Li,G. et al. (2020) Development and validation of novel nomograms for predicting the survival of patients after surgical resection of pancreatic ductal adenocarcinoma. Cancer Med., 9, 3353–3370.

Li,Y. et al. (2016) A Multi-Task Learning Formulation for Survival Analysis. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Loprinzi,C.L. et al. (1994) Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. J. Clin. Oncol., 12, 601–607.

Mactier,S. et al. (2014) Protein signatures correspond to survival outcomes of AJCC stage III melanoma patients. Pigment Cell Melanoma Res., 27, 1106–1116.

Mann,G.J. et al. (2013) BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma. J. Invest. Dermatol., 133, 509–517.

Newson,R. (2006) Confidence Intervals for Rank Statistics: Somers' D and Extensions. The Stata Journal: Promoting communications on statistics and Stata, 6, 309–334.

Ryu,J.Y. et al. (2020) DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. Bioinformatics, 36, 3049–3055.

Saeys,Y. et al. (2007) A review of feature selection techniques in bioinformatics. Bioinformatics, 23, 2507–2517.

Sampson,J.R. (1976) Adaptation in Natural and Artificial Systems (John H. Holland). SIAM Review, 18, 529–530.

Schmid,M. et al. (2011) A Robust Alternative to the Schemper-Henderson Estimator of Prediction Error. Biometrics, 67, 524–535.

Schober,P. and Vetter,T.R. (2018a) Survival Analysis and Interpretation of Time-to-Event Data. Anesthesia & Analgesia, 127, 792–798.

Schober,P. and Vetter,T.R. (2018b) Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesth. Analg., 127, 792–798.

Shaikhina,T. and Khovanova,N.A. (2017) Handling limited datasets with neural networks in medical applications: A small-data approach. Artif. Intell. Med., 75, 51–63.

Sorlie,T. et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. U. S. A., 100, 8418–8423.

Tibshirani,R. (1997) THE LASSO METHOD FOR VARIABLE SELECTION IN THE COX MODEL. Statistics in Medicine, 16, 385–395.

Uno,H. et al. (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine, 30, 1105–1117.

Van Belle, V., Pelckmans, K., Suykens, J. A., & Van Huffel, S. (2008). Survival SVM: a practical scalable algorithm. ESANN, 89–94.

Van Belle, V., Pelckmans, K., Suykens, J. A., & Van Huffel, S. (2008, April). Survival SVM: a practical scalable algorithm. In ESANN (pp. 89-94).

Van Belle,V. et al. (2011) Improved performance on high-dimensional survival data by application of Survival-SVM. Bioinformatics, 27, 87–94.

Van Houwelingen,H.C. (2004) The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Trevor Hastie, Robert Tibshirani and Jerome Friedman, Springer, New York, 2001. No. of pages: xvi 533. ISBN 0-387-95284-5. Statistics in Medicine, 23, 528–529.

van 't Veer,L.J. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415, 530–536.

van de Vijver,M.J. et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med., 347, 1999–2009.

Wang, L., Li, Y., Zhou, J., Zhu, D., & Ye, J. (2017). Multi-task survival analysis. In 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 485–494.

Wang, L., Li, Y., Zhou, J., Zhu, D., & Ye, J. (2017, November). Multi-task survival analysis. In 2017 IEEE International Conference on Data Mining (ICDM) (pp. 485-494). IEEE.

Wang,K.Y.X. et al. Cross-Platform Omics Prediction procedure: a game changer for implementing precision medicine in patients with stage-III melanoma.

Wang,P. et al. (2019) Machine Learning for Survival Analysis. ACM Computing Surveys, 51, 1–36.

Zhao,L. and Feng,D. (2020) Deep Neural Networks for Survival Analysis Using Pseudo Values. IEEE Journal of Biomedical and Health Informatics, 24, 3308–3314.

Table1. Data table showing the names of datasets used in this paper in the first column. Datasets are ordered by the number of observations (second column, from smallest to largest). Censoring rate is rounded to 4 decimal places.

Datasets summary					
Dataset (name used in this paper)	No. of observations (ordered, smallest to largest)	No. of variables	Type of data	Censoring rate (rounded to 4 decimal places)	Reference
Melanoma_itraq	41	642	Omics	0.4146	(Wang *et al.*)
Melanoma_nano	45	206	Omics	0.4222	(Wang *et al.*)
Ovarian_2	58	19818	Omics	0.3793	(Ganzfried *et al.*, 2013)
GE_5	78	4753	Omics	0.5641	(van 't Veer *et al.*, 2002)
GE_3	86	6288	Omics	0.7209	(Bullinger *et al.*, 2004)
Melanoma_clinical	88	16	Clinical	0.3939	(Wang *et al.*)
GE_1	115	551	Omics	0.6670	(Sorlie *et al.*, 2003)
GE-_4	116	4753	Omics	0.5641	(van de Vijver *et al.*, 2002)
Veteran	137	8	Clinical	0.0657	(Kalbfleisch and Prentice, 2002))
Ovarian_1	194	16050	Omics	0.7062	(Ganzfried *et al.*, 2013)
Lung	228	9	Clinical	0.2763	(Loprinzi *et al.*, 1994)
GE_6	240	7401	Omics	0.4250	(van Houwelingen *et al.*, 2006)
GE_2	295	4921	Omics	0.7322	(Beer *et al.*, 2002)
PBC	312	7	Clinical	0.5994	(Fleming and Harrington, 2005)
UNOS_Kidney	3000	101	Clinical	0.7350	(OPTN data)
ANZ	3323	40	Clinical	0.8739	(ANZDATA)

Table 2. Model performance evaluation criteria

Evaluation criteria		
Class	Name	Description
Model flexibility	Type of data required	Describes the source of the data obtained, either clinical data or omics data.
	Data input	Describes the different types of data modality such as categorical, numerical or mixture.
	Data sparsity	Binary yes or no measure whether the model can handle a high level of sparsity in the data.
	Prediction ability evaluation metrics allowed	A certain evaluation metric can only be applied to a specific type of survival model. This is a 2-dim model cross predictability evaluation metric summary and whether a model can be assessed by a metric is recorded with yes or no values.
Model predicta-bility	Harrell's C-index	Harrell's method to calculate the C-index. This is applied to all methods using the R function "rcorr.cens" in package "Hmisc". This value ranges from 0 to 1 with 0.5 representing random guesses and the higher the value the better the concordance, i.e. model predictability.
	Begg's C-index	Begg's method to calculate the C-index. This is applied to applicable methods using the R function "BeggC" in package "survAUC".
	Uno's C-index	Uno's method to calculate the C-index. This is applied to all methods using the R function "UnoC" in package "survAUC".
	GH C-index	Gonen and Heller's method to calculate the C-index. This is applied to applicable methods using the R function "GHCI" in package "survAUC".
	Time-dependent AUC for time t	This is the Chambless and Diao estimator of cumulative/dynamic AUC for right-censored time-to-event data for time t. This is applied to all methods using the R function "AUC.cd" in package "sur-vAUC". The interpretation of this AUC value for a specific time t is the same as AUC in classification models.
	Brier score	This calculates the Brier score for all methods using the R function "pec" in package "pec". The smaller the value, the better the prediction.
	Integrated Brier score	This gives the scaled version of the Brier score. This value is between 0 and 1 (the smaller the value, the better the prediction) and for the constant prediction probability 0.5, the value is 0.25. This is ap-plied to feasible methods using the R function"crps" in package "pec".
Computational efficiency	Computational time	Represents how long it takes the method to run for each dataset. This is calculated using the "Sys.time" function in R.
	Total Memory	Represents how much memory is required to run the method for each dataset. This is calculated using the "Rprof" function in R which returns the memory consumed for each sub function and then we calculate the total memory which is the sum of all of those.
Model stability	Reproducibility	Measures the proportions of successful runs among all those 100 runs attempted. Some methods are not fully successful for all datasets, such as DNNSurv, Cox_bw_AIC and the corresponding successful proportions are plotted. (supplementary plot 3)
	SD of model predict-ability metrics	Measures the standard deviation (SD) of each model predictability metric. Values are ranked from 1 (smallest SD) to 20 (largest SD) for all those 20 methods within each dataset.