SurvMetrics: An R package for Predictive Evaluation Metrics in Survival Analysis

by Hanpu Zhou, Hong Wang, Sizheng Wang and Yi Zou

Abstract Recently, survival models have found vast applications in biostatistics, bioinformatics, reliability engineering, finance and related fields. But there are few R packages focusing on evaluating the predictive power of survival models. This lack of handy software on evaluating survival predictions hinders further applications of survival analysis for practitioners. In this research, we want to fill this gap by providing an "all-in-one" R package which implements most predictive evaluation metrics in survival analysis. In the proposed **SurvMetrics** R package, we implement concordance index for both untied and tied survival data; we give a new calculation process of Brier score and integrated Brier score; we also extend the applicability of integrated absolute error and integrated square error for real data. For models that can output survival time predictions, a simplified metric called mean absolute error is also implemented. In addition, we test the effectiveness of all these metrics on simulated and real survival data sets. The newly developed **SurvMetrics** R package is available on CRAN at https://CRAN.R-project.org/package=SurvMetrics and GitHub at https://github.com/skyee1/SurvMetrics.

1 Introduction

Survival analysis is an important part of biostatistics. It is frequently used to define prognostic indices for mortality or recurrence of a disease, and to study the outcome of treatment (Wang et al., 2019; Wijethilake et al., 2021; Fan et al., 2016; Wiens et al., 2016). Recent decades have witnessed many other applications of survival analysis in various disciplines such as finance, engineering and social sciences (Steyerberg et al., 2010; Bender et al., 2021; Wang et al., 2019). Evaluating the ability to predict future data is one of the most important considerations in the development of these survival models (Wang et al., 2019).

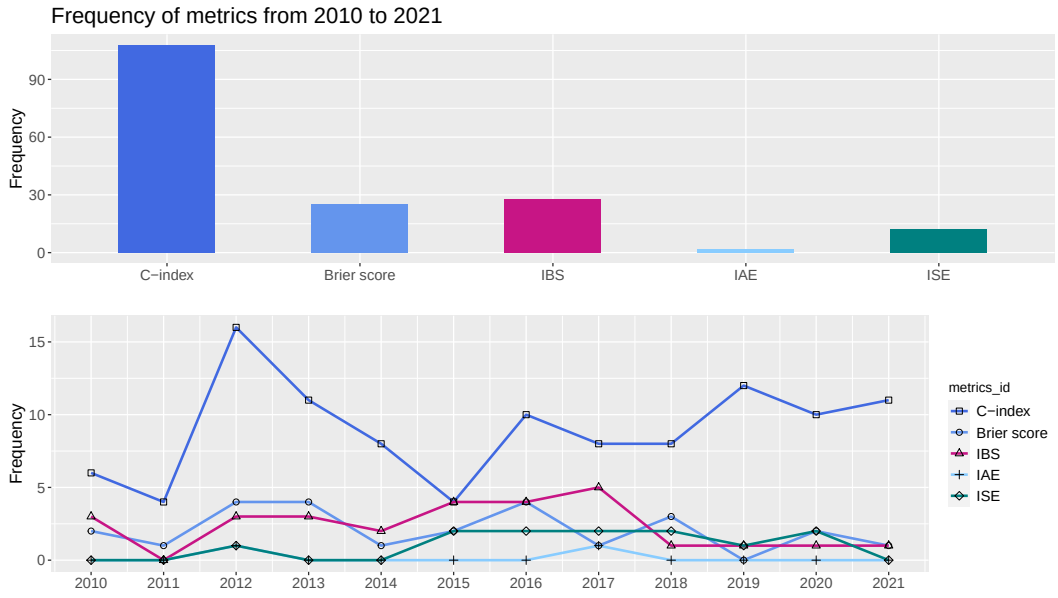

Figure 1: Frequency of survival model evaluation metrics appearing in journals such as "*Annals of Statistics*", "*Biometrika*", "*Journal of the American Statistical Association*", "*Journal of the Royal Statistical Society, Series B*", "*Statistics in Medicine*", "*Artificial Intelligence in Medicine*", "*Lifetime Data Analysis*" from 2010 to 2021.

When evaluating a prediction model, the predictive performance of the survival model is commonly addressed by appropriate measures which quantify the 'distance' or the agreement between the observed and predicted outcomes (Uno et al., 2011; Bylinskii et al., 2018). By investigating the predictive measures frequently used in the statistical literature (top or specialized statistical journals)

in the past decade (from 2010 to 2021) in Figure 1 and Figure 2, we have found that among these 136 research papers, the popular predictive metrics for survival models mainly include: concordance index (C-index), Brier score (BS), integrated Brier score (IBS), integrated absolute error (IAE), integrated square error (ISE) and mean absolute error (MAE) (Harrell et al., 1982; Brier, 1950; Graf et al., 1999; HooraMoradian et al., 2017; Schemper, 1992).

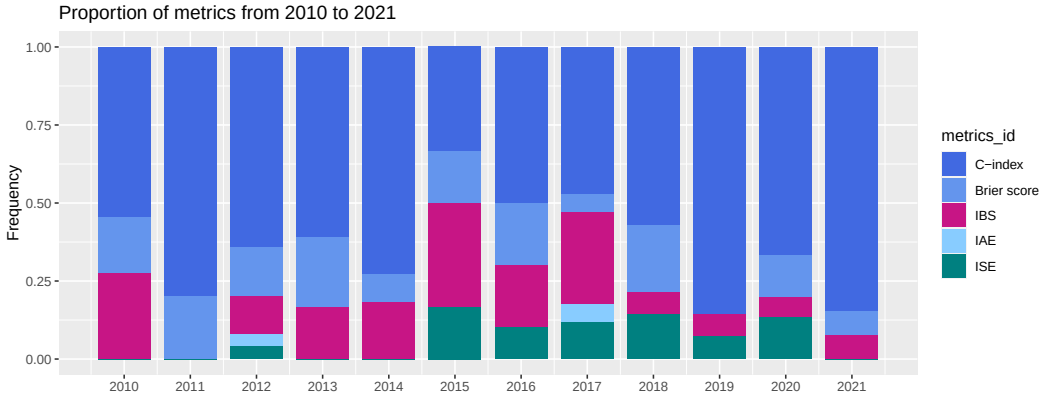

Figure 2: This graphic shows the percentage of metrics appearing from 2010 to 2021 in journals such as "*Annals of Statistics*", "*Biometrika*", "*Journal of the American Statistical Association*", "*Journal of the Royal Statistical Society, Series B*", "*Statistics in Medicine*", "*Artificial Intelligence in Medicine*", "*Lifetime Data Analysis*". It indicates that the use of C-index is steadily increasing and it has become a dominant predictive measure recently. Meanwhile, other metrics do not have an adequate seat in survival prediction practices.

From Figure 1, we can see that the two most frequently used metrics are C-index and IBS, while the other metrics are less used. This may be partly due to the fact that the available predictive metrics are scattered across several R packages, and differences in input data of these tools make it difficult for non-specialists to use or compare the survival models (Schröder et al., 2011; Peters and Hothorn, 2021). Hence, software which provides easy input data formats and includes most popular metrics is desirable (Schröder et al., 2011; Wang et al., 2019).

Figure 2 indicates that the use of C-index is steadily increasing and it has become a dominant predictive measure recently (Jing et al., 2019; Amico et al., 2021). On the one hand, this implies that the lack of handy evaluation metrics tools since other measures are also very important in survival analysis but do not have an adequate seat in survival prediction practices (Nemati et al., 2021; Li et al., 2021; Ensor et al., 2021). On the other hand, it reminds us that providing an accurate C-index measure which can take various situations into account is extremely important (H Ea Gerty and Zheng, 2005; Kang et al., 2015; Zadeh and Schmid, 2020; Zhang et al., 2021).

Different from common classification or regression problems where tools to evaluate predictive performance are abundant, there are few R packages focusing on evaluating the predictive power of survival models (Harrell Jr, 2021; Peters and Hothorn, 2021). To make things worse, most of these available packages have implemented only one or two evaluation metrics and/or some of them often throw errors in real survival problems (Peters and Hothorn, 2021). Take the C-index for an example, if the user wants to compare two samples with the same survival time, this scenario returns NA in **Hmisc**, **survivals**, and **survcomp** packages. However, as described by Ishwaran et al. (2008), "for sample pairs, where observed time $X_i = X_j$ and both are deaths, count 1 if predicted outcomes are tied." This error is mainly due to the fact that the above packages ignore this possible scenario in practice. Omitting such scenarios can result in the loss of information from survival data, especially when the sample size is small with high data collection costs. In **SurvMetrics**, we have implemented the method described by Ishwaran et al. (2008) to cover all tied scenarios. There are also functions available in the **ipred** package to calculate BS and IBS values, but a list of *survfit* objects is required in the calculation process (Peters and Hothorn, 2021). However, for most survival models, only a survival probability vector or matrix is provided, making calculations of such values a challenging problem (Ishwaran et al., 2008). It is not easy for non-specialists to get IBS using the **ipred** package directly on the survival model results.

We argue that an appealing survival evaluation metric tool should satisfy the following two properties. First, the evaluation metrics should not depend on any specific model and they can be applied to evaluating survival models without specifying a specific model form (Schröder et al., 2011). Second, the metrics should be computationally tractable and user has the choice to input only the

fitted model without any additional processing, especially for the non-specialist (Harrell et al., 1982; Graf et al., 1999; HooraMoradian et al., 2017; Ishwaran et al., 2008).

In this research, we are trying to develop a comprehensive and effective R package with the above two properties. In the proposed **SurvMetrics** package, the calculation process of most metrics does not depend on the specific model form and only predicted survival probability vector (or matrix) is needed. This feature is particularly desirable for non-statistical researchers such as clinical and financial practitioners (Ali et al., 2021). To illustrate the effectiveness of our tool, we choose two popular survival models, namely, a semi-parametric Cox model and a non-parametric random survival forest model (RSF) to check and test the provided functionalities (Cox, 1972; Ishwaran et al., 2008).

2 Predictive evaluation metrics in survival analysis

In this section, we will present the survival model evaluation metrics and some implementing details. The following Table 1 presents some popular R packages and the metrics they provide.

Packages	C-index	BS	IBS	IAE	ISE	MAE
survival	✓	✗	✗	✗	✗	✗
Hmisc	✓	✗	✗	✗	✗	✗
survcomp	✓	✓	✓	✗	✗	✗
ipred	✗	✓	✓	✗	✗	✗
mlr	✗	✗	✗	✗	✗	✓
SurvMetrics	✓	✓	✓	✓	✓	✓

Table 1: R packages that are commonly used to evaluate survival predictions.

As can be seen in Table 1, the proposed **SurvMetrics** package implements all popular evaluation metrics, while other packages mostly implement just one or two. With our **SurvMetrics** package, one can compare and evaluate different survival models in a convenient and comprehensive way.

C-index

The most popular evaluation metric in survival predictions is C-index (Harrell et al., 1982; Li et al., 2016; Lee et al., 2018; Devlin and Heller, 2021; Subramanian et al., 2020; Hsu and Lin, 2020; Zadeh and Schmid, 2020; Zhang et al., 2021; Wang and Zhou, 2017; Wang and Li, 2019), which is a generalization of the ROC curve (H Ea Gerty and Zheng, 2005; Obuchowski, 2006; Kang et al., 2015; Li et al., 2016). C-index intends to measure the proportion of predictions of the binary survival status that agree with the true survival status. If the study is concerned with the comparison of survival probability between different samples, for example, which component in the device shows damage earlier, the C-index can be used.

To calculate the values of C-index from different models, **Hmisc**, **survcomp** and **survival** R packages are frequently used in practice (Harrell et al., 1996; Schröder et al., 2011; Therneau, 2021; Harrell Jr, 2021). All the above-mentioned packages in their current versions do not consider the cases of tied survival data, i.e., samples with the same survival time. As shown in Table 3, they return 0 or NA for all cases of survival data tied. However, in practice, cancer patients are usually surveyed annually or monthly after surgery for survival status and with the advent of big data, tied samples are becoming commonplace. Simply ignoring the presence of tie data in calculating C-index will result in evaluation bias among survival models.

Meanwhile, in the calculation of the C-index, not all sample pairs are comparable, e.g., for those pairs whose shorter survival time is censored. So it is necessary to filter the sample pairs (i, j) that can be compared in the calculation, i and j denote any two samples in the data set, respectively. By defining the comparable sample pair in the following way, when $np_{ij} = 1$, all comparable sample pairs described by Ishwaran et al. (2008) can be covered.

In the proposed R package, we will take a similar strategy adopted in Ishwaran et al. (2008). δ_i denotes the survival status of sample i (0 means censoring, 1 means event), Y_i and X_i represent the predicted survival time and the observed survival time, respectively. By defining *csign* and *sign*, we can easily distinguish which sample pairs meet the concordance and which are not.

(a) The comparable sample pair is defined as:

$$np_{ij}\left(X_i, \delta_i, X_j, \delta_j\right) = \max(I\left(X_i \geqslant X_j\right)\delta_j, I\left(X_i \leqslant X_j\right)\delta_i). \tag{1}$$

(b) The complete concordance (CC) is defined as:

$$\mathrm{CC} = \sum_{i,j} I(\mathrm{sign}(Y_i, Y_j) = \mathrm{csign}(X_i, X_j) | np_{ij} = 1), \tag{2}$$

where

$$\mathrm{sign}(Y_i, Y_j) = I\left(Y_i \geqslant Y_j\right) - I\left(Y_i \leqslant Y_j\right) \tag{3}$$

and

$$\mathrm{csign}(X_i, \delta_i, X_j, \delta_j) = I(X_i \geqslant X_j)\delta_j - I\left(X_i \leqslant X_j\right)\delta_i. \tag{4}$$

(c) We derive the partial concordance (PC):

$$\begin{aligned} \mathrm{PC} = \sum_{i,j} & I\left(Y_i = Y_j \mid np_{ij} = 1, X_i \neq X_j\right) \\ & + I\left(Y_i \neq Y_j \mid np_{ij} = 1, X_i = X_j, \delta_i = \delta_j = 1\right) \\ & + I\left(Y_i \geq Y_j \mid np_{ij} = 1, X_i = X_j, \delta_i = 1, \delta_j = 0\right). \end{aligned} \tag{5}$$

(d) The C-index value can be given by:

$$C\text{-}index = \frac{\text{concordance}}{\text{permissible}} = \frac{CC + 0.5 \times PC}{\sum_{i,j} np_{ij}(X_i, \delta_i, X_j, \delta_j)}, \tag{6}$$

From the above, we know C-index lies between 0 and 1 and usually the larger the C-index, the more accurate prediction the model can get.

To further illustrate the proposed `Cindex()` function in **SurvMetrics** with other packages, we first give a toy example in Table 2.

sample	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
time	1	1	2	2	2	2	2	2
status	0	1	1	0	1	1	0	1
predicted	0.2	0.3	0.3	0.3	0.4	0.2	0.4	0.3

Table 2: A simulation data set which can include all the possible sample pairs, where time and status are the observed survival time and survival status of samples $S_1 ... S_8$, respectively, and predicted is the survival probability predicted by survival models.

Based on the above information, different R packages will take different ways to deal with these predictions. We summarize these calculation differences behind these R packages in Table 3.

sample pair	(S_1, S_2)	(S_3, S_5)	(S_3, S_6)	(S_3, S_8)	(S_4, S_5)	(S_5, S_6)	(S_5, S_8)	(S_6, S_8)
Hmisc	0	NA	NA	NA	0	NA	NA	NA
survival	0	NA	NA	NA	0	NA	NA	NA
survcomp	NA	NA	NA	NA	NA	NA	NA	NA
SurvMetrics	0.5	0.5	0.5	1	0.5	0.5	0.5	0.5

Table 3: Differences of C-index calculated by the **Hmisc** package, the **survival** package, the **survcomp** package and the **SurvMetrics** package.

From the above, we can find that when ignoring part or all of the information from tied sample pairs, the C-index values obtained from **Hmisc** and other two packages cannot reflect the predictive powers of survival models in a true manner. In this case, one can turn to the `Cindex()` function in **SurvMetrics** package for help.

BS

BS is another commonly used metric in survival analysis (Steyerberg et al., 2010; Chicco et al., 2021; Imani et al., 2019; Pakbin et al., 2018), which measures the mean squared error between the observed progression status and the predicted survival probability at time t^*. Thus, a lower BS usually indicates a better prediction model. BS focuses on the residuals between the predicted survival probability and the survival status at a fixed time point. If the study is concerned with the prediction probability of a model at a fixed time point, BS can be chosen as the metric. For example, the 10-year average survival probability of cancer patients.

In the **SurvMetrics** R package we will take the most classical method to calculate the BS by weighting the prediction residuals with inverse probability censoring weights (IPCW) proposed by Graf et al. (1999):

$$BS(t^*) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(\hat{S}(t^*|z_i))^2}{\hat{G}(X_i)}\cdot I(X_i < t^*, \delta_i = 1) + \frac{(1-\hat{S}(t^*|z_i))^2}{\hat{G}(t^*)}\cdot I(X_i \geq t^*)\right], \tag{7}$$

where t^* is the time point at which BS is to be calculated, N is the sample size, z_i is the covariates of instance i, $\hat{S}(\cdot)$ is the survival function predicted by the model, $\hat{G}(\cdot)$ denotes the weight for the instance which is estimated by the KM(Kaplan–Meier) estimator of the censoring distribution.

According to the definition of BS, its value depends on the selection of t^* provided by the user and different choices of t^* will always lead to different BS values (Chew et al., 2001; Li et al., 2021; Zhu and Kosorok, 2012; Ji et al., 2020). In the `Brier()` function of our **SurvMetrics** R package, t^* is set to median survival time as default.

IBS

As we can see above, BS depends partly on the choices of a user-specified time point. It makes the comparison between models somewhat difficult when we want to know an average prediction performance over all prediction times. In practice, the integral of BS or IBS, which does not depend on the selection of one time point, is more widely used when we are interested in the entire time interval (Zhu and Kosorok, 2012; Periáñez et al., 2016; Bertens et al., 2017). IBS is more concerned with the residuals at all observed time points. When the time of interest is no longer a specific time point, IBS can provide more information than BS. For example, the probability of survival of a cancer patient for each year after the disease.

The definition of IBS is straightforward:

$$IBS = \frac{1}{\max\limits_{i}(X_i)}\int_0^{\max\limits_{i}(X_i)} BS(t)dt. \tag{8}$$

But the calculation of IBS using `sbrier()` function from the **ipred** package does not always go smoothly (Peters and Hothorn, 2021) . If a list of `survfit` objects is incorrectly specified, then the package currently returns a particular error:

`'Error in switch(ptype,survfit = { : EXPR must be a length 1 vector} '`

In our experience, this kind of error is common, especially for the non-specialists (Peters and Hothorn, 2021).

In the **SurvMetrics** package, we will make life easier for non-specialists. Users only need to input survival time, survival status, the predicted survival probability matrix and the range of integration to the `IBS()` function, and our program will take care of all the rest work and give a correct output. Similar to BS, a smaller IBS value usually implies a more accurate survival model.

IAE and ISE

Another two evaluation metrics, namely, IAE and ISE are also occasionally used to compare the difference between the estimated survival function $\hat{S}(\cdot)$ and the true survival function $S(\cdot)$ in terms of L_1 and L_2 paradigms, respectively. When the purpose of the model is to approximate the theoretical distribution function, IAE and ISE can be selected as the metrics. HooraMoradian et al. (2017); Zou et al. (2021). The IAE and ISE can be defined as:

$$IAE = \int_t |S(t) - \hat{S}(t|X)|dt \tag{9}$$

and

$$ISE = \int_t (S(t) - \hat{S}(t|X))^2 dt, \tag{10}$$

where X is the covariate of the training set.

Since the true survival functions are usually unknown beforehand, traditional IAE and ISE methods are only applicable to simulation scenarios. Here, we propose to approximate $S(t)$ using the non-parametric KM estimator in the `IAEISE()` function and this makes IAE and ISE also suitable for real data study. Similar to IBS, a smaller IAE and ISE values lead to a more accurate model.

MAE

The last evaluation metric presented here is MAE (Schemper, 1992), which can be used to measure the residuals of observed survival time versus predicted survival time, for example, to predict the time to damage of the device. MAE is a better choice than BS if predicted survival time is a model's output. The definition of MAE is shown below:

$$MAE = \frac{1}{n}\sum_{i=1}^{N}\left(\delta_i \left|Y_i - X_i\right|\right), \tag{11}$$

where n is the event sample size.

MAE only estimates the average absolute error between the predicted survival time and the true survival time in the uncensored samples and is rarely used in practice. For consistence, we also provide this metric in the `MAE()` function. Similar to MSE, the lower the MAE, the higher the accuracy of prediction.

3 Simulations and examples

In this section, we will use some simulated and real survival data sets to illustrate the effectiveness of the metrics provided in the **SurvMetrics** package.

The performance of SurvMetrics on simulation data sets

First, Cox proportional hazard models will be fitted on three simulated scenarios which are very similar to Settings 1–4 by Steingrimsson et al. (2019). Scenario1 satisfies the proportional hazards assumption and others violate it.

Scenario1: This data set is created using N independent observations, where the covariate vector $(W_1, ..., W_p)$ is multivariate normal with mean zero and a covariance matrix having elements (i, j) equal to $0.9^{|i-j|}$. Survival times are simulated from an exponential distribution with mean $\mu = e^{0.1\sum_{i=[p/2]+1}^{p} W_i}$ (i.e., a proportional hazards model) and the censoring distribution is exponential with mean c_{mean} which is chosen to control the censoring rate. Here, $[x]$ denotes the largest integer no more than x.

Scenario2: This data set is created using N independent observations where the covariates $(W_1, ..., W_p)$ are multivariate normal with mean zero and a covariance matrix having elements (i, j) equal to $0.75^{|i-j|}$. Survival times are gamma distributed with shape parameter $\mu = 0.5 + 0.3|\sum_{i=[2p/5]}^{[3p/5]} W_i|$ and scale parameter 2. Censoring times are uniform on $[0, u_{max}]$, and u_{max} is chosen to control the censoring rate. Here, the proportional hazards assumption is violated.

Scenario3: This data set is created using N independent observations where the covariates $(W_1, ..., W_p)$ are multivariate normal with mean zero and a covariance matrix having elements (i, j) equal to $0.75^{|i-j|}$. Survival times are simulated according to a log-normal distribution with mean $\mu = 0.1|\sum_{i=1}^{[p/5]} W_i| + 0.1|\sum_{i=[4p/5]}^{p} W_i|$. Censoring times are log-normal with mean $\mu + c_{step}$ and scale parameter one, where c_{step} is chosen to control the censoring rate. Here, the underlying censoring distribution depends on covariates and the proportional hazards assumption is also violated.

From Figure 3, one may see that the prediction performance of the Cox model keeps decreasing in three scenarios as expected. However, the difference in terms of C-index and BS are not significant for Scenario2 and Scenario3 while in terms of IBS, IAE and ISE, a sharp difference and a clear trend can be observed.

As we know, C-index evaluates the model from a ranking perspective in a rough matter and BS only considers a fixed time point. Both metrics may not distinguish models described above or in other complex settings. In this case, evaluation metrics which consider model performance in a global way, such as IBS, IAE, and ISE, may be useful alternatives. This also justifies why a package such as **SurvMetrics** that can evaluate survival models from multiple perspectives is strongly needed in practice.

An example of SurvMetrics

In this section, a kidney dataset from the **survival** R package is used to illustrate the usage of the **SurvMetrics** package using popular survival models widely used in biostatics and biomedical study.

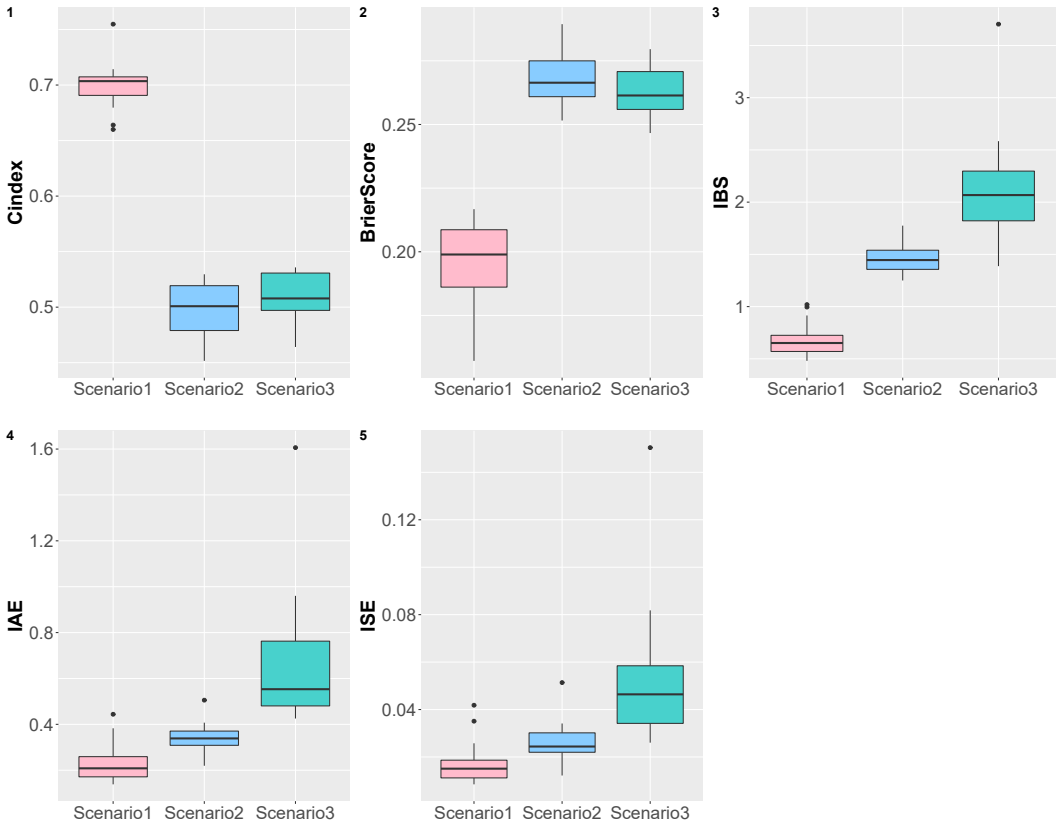

Figure 3: This graphic shows the Cox model prediction accuracy based on three different scenarios via 5 evaluation metrics by the **SurvMetrics** package. The sample size is 300, the dimension is 25, and control the censoring rate about 30%. The prediction performance of the Cox model keeps decreasing in three scenarios, which satisfies the data generated methods. However, in some complex situations where C-index and BS cannot distinguish models, IBS, IAE, and ISE, may be useful alternatives.

In the following, we first cut the kidney dataset into a training set and a testing set. Then, two popular models, namely, Cox model and random survival forest model are constructed based on the training set to show how different functions are used in **SurvMetrics**. Finally, we will show how to evaluate the predictive performance of these two models on the testing set using the proposed **SurvMetrics** R package. In the latest version of **SurvMetrics**, user has the choice to input only the standard survival models and testing sets. Meanwhile, it is still useful to deal with predicted survival probabilities from non-standard models. The following example will contain these two input forms.

The corresponding R code is provided here:

```
#1. data preparation
library(survival)        # to fit a Cox model
library(randomForestSRC) # to fit an RSF model
library(SurvMetrics) # to get all the metrics
library(pec)         # to make predictions based on Cox model
set.seed(1)
mydata <- kidney[, -1]
train_index <- sample(1:nrow(mydata), 0.7 * nrow(mydata))
train_data <- mydata[train_index, ]
test_data <- mydata[-train_index, ]

#2. fit the RSF model and Cox model to predict the testing set
#2.1 RSF model
fit_rsf <- rfsrc(Surv(time,status)~., data = train_data)  #fit the RSF model
distime <- fit_rsf$time.interest  #get the survival time of events
med_index <- median(1:length(distime))  #the index of median survival time of events
mat_rsf <- predict(fit_rsf, test_data)$survival  #get the survival probability matrix
vec_rsf <- mat_rsf[ ,med_index]  #median survival probability of all samples
```

```
#2.2 Cox model
fit_cox <- coxph(Surv(time,status)~., data = train_data, x = TRUE)  #fit the Cox model
mat_cox <- predictSurvProb(fit_cox, test_data, distime)  #get the survival probability matrix
vec_cox <- mat_cox[ ,med_index]

#3. get all the metrics by SurvMetrics
#3.1 CI BS IBS IAE ISE based on RSF model: standard model input methods
Cindex_rsf <- Cindex(fit_rsf, test_data)
BS_rsf <- Brier(fit_rsf, test_data, distime[med_index])
IBS_rsf <- IBS(fit_rsf, test_data)
IAE_rsf <- IAEISE(fit_rsf, test_data)[1]
ISE_rsf <- IAEISE(fit_rsf, test_data)[2]

#CI BS IBS IAE ISE based on Cox model: standard model input methods
Cindex_cox <- Cindex(fit_cox, test_data)
BS_cox <- Brier(fit_cox, test_data, distime[med_index])
IBS_cox <- IBS(fit_cox, test_data)
IAE_cox <- IAEISE(fit_cox, test_data)[1]
ISE_cox <- IAEISE(fit_cox, test_data)[2]

#3.2 CI BS IBS IAE ISE based on RSF model: Non-standard model input methods
times <- test_data$time
status <- test_data$status
Cindex_rsf <- Cindex(Surv(times, status), vec_rsf)
BS_rsf <- Brier(Surv(times, status), vec_rsf, distime[med_index])
IBS_rsf <- IBS(Surv(times, status), mat_rsf, distime)  # distime can be replaced by range(distime)
IAE_rsf <- IAEISE(Surv(times, status), mat_rsf, distime)[1]
ISE_rsf <- IAEISE(Surv(times, status), mat_rsf, distime)[2]

#CI BS IBS IAE ISE based on Cox model: Non-standard model input methods
Cindex_cox <- Cindex(Surv(times, status), vec_cox)
BS_cox <- Brier(Surv(times, status), vec_cox, distime[med_index])
IBS_cox <- IBS(Surv(times, status), mat_cox, distime)
IAE_cox <- IAEISE(Surv(times, status), mat_cox, distime)[1]
ISE_cox <- IAEISE(Surv(times, status), mat_cox, distime)[2]
```

models	C-index	BS	IBS	IAE	ISE
Cox	0.751185	0.18133	0.08842	77.90947	19.65131
RSF	0.729858	0.221523	0.10920	105.6936	28.80941

Table 4: This table shows the prediction accuracy by using **SurvMetrics** package to compare the Cox and RSF models based on the kidney data set from **survival** R package. In this case, the values of C-index are close, and the values of IAE and ISE are far different. When using C-index alone is hard to differentiate the predictive power of the two models, alternative measures provided in **SurvMetrics** may help give a more accurate result.

The results presented in Table 4 and Figure 4 show that using C-index alone is hard to differentiate the predictive power of the two models. In this case, alternative measures provided in the proposed package may help if further evaluation is needed. This result is also consistent with the simulated scenario mentioned above.

4 Summary

Assessing the predictive performance of survival models is complex due to the lack of standards regarding the best criterion to use in survival analysis. The available metrics are scatter across different R packages that use heterogeneous interfaces, which makes it difficult for the non-specialist to use or compare the performance of various survival models.

In this paper, we try to fill the gap by providing an "all-in-one" R package called **SurvMetrics** which provides a uniform interface to an extensive set of performance assessment and statistical comparison methods. Practitioners can easily implement comparative studies and identify the best model(s) using this package. In the current version of the **SurvMetrics** package, six evaluation metrics

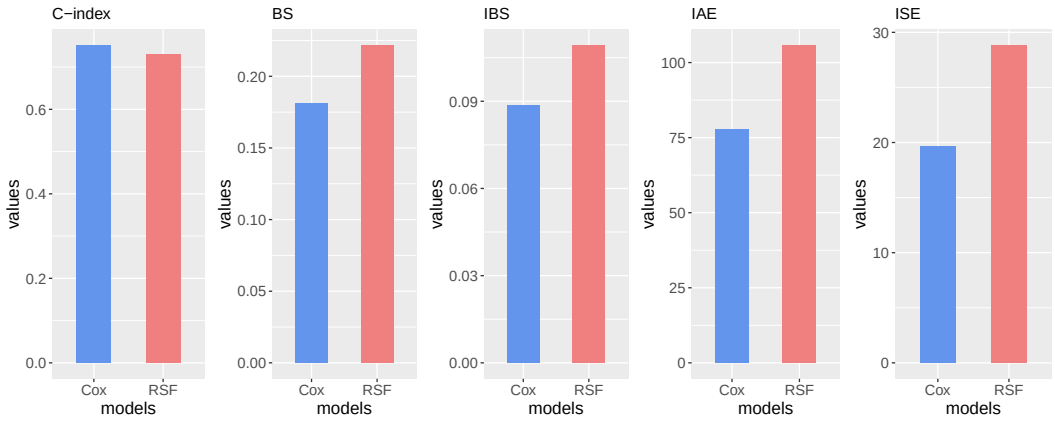

Figure 4: This graphic shows the prediction accuracy by using **SurvMetrics** package to compare the Cox and RSF models based on the kidney data set from **survival** R package. In this case, using IBS, IAE and ISE can give a more clear conclusion that Cox model performs better.

are present. Evaluation metrics used in more complicated settings such as time-dependent AUC for joint modelling of longitudinal and survival data, and concordance index for competitive risks are still being developed and will be added to the **SurvMetrics** package in the future. Meanwhile, we are also working to provide a more user-friendly interface to facilitate both statistical and non-statistical clinical research workers in evaluating survival models.

5 Acknowledgments

This research is support in part by National Statistical Scientific Research Project of China (No.2022LZ28), Changsha Municipal Natural Science Foundation (No.kq2202080), The Open Research Fund from the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen(No. B10120210117-OF04) and the Postgraduate Scientific Research Innovation Project of Hunan Province, China (CX20210155).

Bibliography

M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, S. Sharifzadeh, V. Tresp, and J. Lehmann. Pykeen 1.0: A python library for training and evaluating knowledge graph embeddings. *Journal of Machine Learning Research*, 22(82):1–6, 2021. [p3]

M. Amico, I. Van Keilegom, and B. Han. Assessing cure status prediction from survival data using receiver operating characteristic curves. *Biometrika*, 108(3):727–740, 2021. [p2]

A. Bender, D. Rügamer, F. Scheipl, and B. Bischl. A general machine learning framework for survival analysis. In F. Hutter, K. Kersting, J. Lijffijt, and I. Valera, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 158–173, Cham, 2021. Springer International Publishing. ISBN 978-3-030-67664-3. [p1]

P. Bertens, A. Guitart, and Á. Periáñez. Games and big data: A scalable multi-dimensional churn prediction model. In *2017 IEEE conference on computational intelligence and games (CIG)*, pages 33–36. IEEE, 2017. [p5]

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1950. [p2]

Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. [p1]

D. P. Chew, D. L. Bhatt, M. A. Robbins, M. S. Penn, J. P. Schneider, M. S. Lauer, E. J. Topol, and S. G. Ellis. Incremental prognostic value of elevated baseline c-reactive protein among established markers of risk in percutaneous coronary intervention. *Circulation*, 104(9):992–997, 2001. [p5]

D. Chicco, M. J. Warrens, and G. Jurman. The matthews correlation coefficient mcc is more informative than cohen's Kappa and Brier score in binary classification assessment. *IEEE Access*, 2021. [p4]

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. [p3]

S. M. Devlin and G. Heller. Concordance probability as a meaningful contrast across disparate survival times. *Statistical methods in medical research*, 30(3):816–825, 2021. [p3]

J. Ensor, K. I. Snell, T. P. Debray, P. C. Lambert, M. P. Look, M. A. Mamas, K. G. Moons, and R. D. Riley. Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model. *Statistics in Medicine*, 2021. [p2]

J. Fan, Y. Wu, M. Yuan, D. Page, J. Liu, I. M. Ong, P. Peissig, and E. Burnside. Structure-leveraged methods in breast cancer risk prediction. *Journal of Machine Learning Research*, 17(1):2956–2970, 2016. [p1]

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999. [p2, 3, 5]

P. J. H Ea Gerty and Y. Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1): 92–105, 2005. [p2, 3]

F. Harrell, K. Lee, and D. MARK. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15: 361–387, 02 1996. [p3]

F. E. J. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *JAMA The Journal of the American Medical Association*, 247(18):2543–2546, 1982. [p2, 3]

F. E. Harrell Jr. *Hmisc: Harrell Miscellaneous*, 2021. URL https://CRAN.R-project.org/package=Hmisc. R package version 4.6-0. [p2, 3]

HooraMoradian, DenisLarocque, and FranoisBellavance. L1 splitting rules in survival forests. *Lifetime Data Analysis*, 23(4):671–691, 2017. [p2, 3, 5]

T.-C. Hsu and C. Lin. Generative adversarial networks for robust breast cancer prognosis prediction with limited data size. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5669–5672. IEEE, 2020. [p3]

F. Imani, R. Chen, C. Tucker, and H. Yang. Random forest modeling for survival analysis of cancer recurrences. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pages 399–404. IEEE, 2019. [p4]

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008. [p2, 3]

G.-W. Ji, F.-P. Zhu, Q. Xu, K. Wang, M.-Y. Wu, W.-W. Tang, X.-C. Li, and X.-H. Wang. Radiomic features at contrast-enhanced ct predict recurrence in early stage hepatocellular carcinoma: a multi-institutional study. *Radiology*, 294(3):568–579, 2020. [p5]

B. Jing, T. Zhang, Z. Wang, Y. Jin, and C. Li. A deep survival analysis method based on ranking. *Artificial intelligence in medicine*, 98:1–9, 2019. [p2]

L. Kang, W. Chen, N. A. Petrick, and B. D. Gallas. Comparing two correlated c indices with right-censored survival outcome: a one-shot nonparametric approach. *Statistics in Medicine*, 34(4), 2015. [p2, 3]

C. Lee, W. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. [p3]

Y. Li, J. Wang, J. Ye, and C. K. Reddy. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1715–1724, 2016. [p3]

Z. Li, L. Yuan, C. Zhang, J. Sun, Z. Wang, Y. Wang, X. Hao, F. Gao, and X. Jiang. A novel prognostic scoring system of intrahepatic cholangiocarcinoma with machine learning basing on real-world data. *Frontiers in Oncology*, 10:3146, 2021. ISSN 2234-943X. doi: 10.3389/fonc.2020.576901. [p2, 5]

M. Nemati, H. Zhang, M. Sloma, D. Bekbolsynov, H. Wang, S. Stepkowski, and K. S. Xu. Predicting kidney transplant survival using multiple feature representations for HLAS. In *International Conference on Artificial Intelligence in Medicine*, pages 51–60. Springer, 2021. [p2]

N. A. Obuchowski. An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Statistics in Medicine*, 25(3):481–493, 2006. [p3]

A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M. H. Krumholz, and J. B. Mortazavi. Prediction of icu readmissions using data at patient discharge. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4932–4935. IEEE, 2018. [p4]

Á. Periáñez, A. Saas, A. Guitart, and C. Magne. Churn prediction in mobile social games: Towards a complete assessment using survival ensembles. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 564–573. IEEE, 2016. [p5]

A. Peters and T. Hothorn. *ipred: Improved Predictors*, 2021. URL https://CRAN.R-project.org/package=ipred. R package version 0.9-12. [p2, 5]

M. Schemper. The explained variation in proportional hazards regression. *Biometrika*, 79(1):202–204, 1992. [p2, 6]

M. S. Schröder, A. C. Culhane, J. Quackenbush, and B. Haibe-Kains. survcomp: an R/bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, 27(22): 3206–3208, 2011. [p2, 3]

J. A. Steingrimsson, L. Diao, and R. L. Strawderman. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114, 2019. [p6]

E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. A. Gerds, M. Gonen, N. Obuchowski, M. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010. [p1, 4]

V. Subramanian, M. N. Do, and T. Syeda-Mahmood. Multimodal fusion of imaging and genomics for lung cancer recurrence prediction. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 804–808. IEEE, 2020. [p3]

T. M. Therneau. *A Package for Survival Analysis in R*, 2021. URL https://CRAN.R-project.org/package=survival. R package version 3.2-13. [p3]

H. Uno, T. Cai, M. Pencina, R. D'Agostino, and L. Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30: 1105–17, 05 2011. [p1]

H. Wang and G. Li. Extreme learning machine cox model for high-dimensional survival analysis. *Statistics in medicine*, 38(12):2139—2156, 2019. [p3]

H. Wang and L. Zhou. Random survival forest with space extensions for censored data. *Artificial Intelligence in Medicine*, 79:52–61, 2017. [p3]

P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019. [p1, 2]

J. Wiens, J. Guttag, and E. Horvitz. Patient risk stratification with time-varying parameters: a multitask learning approach. *Journal of Machine Learning Research*, 17(1):2797–2819, 2016. [p1]

N. Wijethilake, D. Meedeniya, C. Chitraranjan, I. Perera, and H. Ren. Glioma survival analysis empowered with data engineering—a survey. *IEEE Access*, 9:43168–43191, 2021. [p1]

S. G. Zadeh and M. Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 2020. [p2, 3]

L. Zhang, D. Dong, L. Zhong, C. Li, C. Hu, X. Yang, Z. Liu, R. Wang, J. Zhou, and J. Tian. Multi-focus network to decode imaging phenotype for overall survival prediction of gastric cancer patients. *IEEE Journal of Biomedical and Health Informatics*, 2021. [p2, 3]

R. Zhu and M. R. Kosorok. Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340, 2012. [p5]

Y. Zou, G. Fan, and R. Zhang. Integrated square error of hazard rate estimation for survival data with missing censoring indicators. *Journal of Systems Science and Complexity*, 34(2):735–758, 2021. [p5]

Hanpu Zhou
Central South University
School of Mathematics & Statistics
Changsha, Hunan Provinice,410075
China
zhouhanpu@csu.edu.cn

*Hong Wang**
Central South University
School of Mathematics & Statistics
Changsha, Hunan Provinice,410075
China
(Corresponding Author)
wh@csu.edu.cn

Sizheng Wang
Central South University
School of Mathematics & Statistics
Changsha, Hunan Provinice,410075
China
wszhchina@163.com

Yi Zou
Central South University
School of Mathematics & Statistics
Changsha, Hunan Provinice,410075
China
zy6868@csu.edu.cn