Journal of Biomedical Informatics 125 (2022) 103959

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

AutoScore-Survival: Developing interpretable machine learning-based time-to-event scores with right-censored survival data

Feng Xie [a,b], Yilin Ning [b], Han Yuan [b], Benjamin Alan Goldstein [a,c], Marcus Eng Hock Ong [a,d], Nan Liu [a,b,e,f,*], Bibhas Chakraborty [a,b,c,g]

[a] Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore
[b] Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore
[c] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States
[d] Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore
[e] Institute of Data Science, National University of Singapore, Singapore, Singapore
[f] Health Services Research Centre, Singapore Health Services, Singapore, Singapore
[g] Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore

ARTICLE INFO

Keywords:
AutoScore
Interpretable machine learning
Score
Survival analysis
Time-to-event

ABSTRACT

Background: Scoring systems are highly interpretable and widely used to evaluate time-to-event outcomes in healthcare research. However, existing time-to-event scores are predominantly created ad-hoc using a few manually selected variables based on clinician's knowledge, suggesting an unmet need for a robust and efficient generic score-generating method.
Methods: AutoScore was previously developed as an interpretable machine learning score generator, integrating both machine learning and point-based scores in the strong discriminability and accessibility. We have further extended it to the time-to-event outcomes and developed AutoScore-Survival, for generating time-to-event scores with right-censored survival data. Random survival forest provided an efficient solution for selecting variables, and Cox regression was used for score weighting. We implemented our proposed method as an R package. We illustrated our method in a study of 90-day survival prediction for patients in intensive care units and compared its performance with other survival models, the random survival forest, and two traditional clinical scores.
Results: The AutoScore-Survival-derived scoring system was more parsimonious than survival models built using traditional variable selection methods (e.g., penalized likelihood approach and stepwise variable selection), and its performance was comparable to survival models using the same set of variables. Although AutoScore-Survival achieved a comparable integrated area under the curve of 0.782 (95% CI: 0.767–0.794), the integer-valued time-to-event scores generated are favorable in clinical applications because they are easier to compute and interpret.
Conclusions: Our proposed AutoScore-Survival provides a robust and easy-to-use machine learning-based clinical score generator to studies of time-to-event outcomes. It gives a systematic guideline to facilitate the future development of time-to-event scores for clinical applications.

1. Introduction

The interpretable predictive model is essential for supporting medical decision-making, where doctors can easily understand how the models make predictions in a transparent manner. There has been a growth in inherently interpretable machine learning models [1,2], where risk scoring systems were highly preferred in healthcare settings. Recently, Ustun et al. developed Risk-calibrated Supersparse Linear Integer Model (RiskSLIM) [3] and further improved it through the optimization of risk scores [4]. Besides, we previously provided a practical solution, AutoScore [5], as an interpretable machine learning-based automatic clinical score generator. Users can automatically generate a data-driven clinical score given a dataset in various clinical applications [6], facilitating automated machine learning (AutoML) [7] solutions in healthcare. However, those models were initially designed for binary outcomes, and extending them to time-to-event outcomes is of great value.

There are different regression and machine learning options for the

* Corresponding author at: Programme in Health Services and Systems Research Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore.
E-mail address: liu.nan@duke-nus.edu.sg (N. Liu).

https://doi.org/10.1016/j.jbi.2021.103959
Received 6 July 2021; Received in revised form 18 November 2021; Accepted 19 November 2021
Available online 23 November 2021

prediction of time-to-event outcomes. Typically, these models generate a probability of not having the event (i.e., surviving) at a specified time point (e.g., predicting 30-day mortality). However, like most predictive models, such algorithms fail to generate an indicative score for straightforward risk stratification. In comparison, scoring systems would be strongly preferred in healthcare since they are highly transparent and interpretable, based on addition, subtraction, and multiplication of a few sparse numbers, facilitating clinical practice even without the need for a computer. At present, this type of time-to-event score has been pervasively used in healthcare, such as survival prediction score [8], Palliative Prognostic Score [9], Respiratory ECMO Survival Prediction [10] across different clinical disciplines. They were developed to support treatment decisions by forecasting the time to patient outcome (e.g., death or disease progression) or by projecting the change in risk over time. However, these time-to-event scores were created ad-hoc via manual variable selection based on expert opinion, suggesting the unmet need for a robust and efficient generic method for deriving time-to-event scores.

Traditionally, survival data are analyzed using Cox regression, where variable selection is predominantly performed by stepwise selection (Akaike information criterion [AIC] [11,12] and the Bayesian information criterion [BIC] [13,14]) or by penalizing the partial likelihood [15] (i.e., least absolute shrinkage and selection operator [LASSO] [16]). However, such approaches are not efficient when working with big data, e.g., the electronic health records (EHR) [17]. Machine learning, such as random survival forest [18–20], XGBoosting [21], support vector machine (SVM) [22], and deep learning models (artificial neural networks) [23] have been applied for more efficiently handling high-dimensional survival data, but most of them are black boxes that are challenging to comprehend. Thus, there is an unmet need to develop a parsimonious survival prediction model with easy access to validation in the context of high-dimensional EHRs.

To address these challenges, we extended previously mentioned AutoScore [5,24] to survival data and systematically presented AutoScore-Survival, a generic method for developing parsimonious time-to-event scores. The proposed AutoScore-Survival framework can automatically generate a single indicative score for predicting patients' time-to-event outcomes and was demonstrated to build an actual score for survival prediction of intensive care unit (ICU) patients. We also compared the AutoScore-Survival-created scores with other standard baselines.

2. Methods

The AutoScore framework was developed to generate prediction scores for binary outcomes [5,24]. It consists of six modules: Module 1 ranks variables using machine learning methods, Module 2 categorizes continuous variables to deal with nonlinearity and simplify interpretation, Module 3 derives scores from a subset of variables using the logistic regression, Module 4 selects the best number of variables through parsimony plot, Module 5 allows fine-tuning of cut-offs for categorizing continuous variables for preferable interpretation and Module 6 performs the final performance evaluation of the score. Our proposed AutoScore-Survival method extends this framework to time-to-event data by modifying relevant modules. Fig. 1 illustrates the six-module framework of AutoScore-Survival, where the modules modified from AutoScore are highlighted in blue shape and elaborated in detail in the following subsections.

2.1. Variable ranking with random survival forests at Module 1

In real-world clinical applications, we split the data set into training, validation, and test sets. The training set is used to derive the score. The validation set is used for intermediate performance evaluation and parameter selection. The test set acts as an unseen dataset and is used to generate the final model performance. Let $(t_i, \delta_i, \boldsymbol{X}^i)$ denote the survival data for the *ith* individual in the training set. t_i denotes the time of the event if censoring indicator $\delta_i = 1$ and time of censoring if $\delta_i = 0$. $\boldsymbol{X}^i$ denotes the vector of p available predictor variables. Our goal is to rank all p available variables and select m parsimonious variables $(m < p)$ for the following score derivation. For simplicity of notation, we will omit i in the subscript and superscript when no confusion arises.

We use random survival forest (RSF) [18,19,25], an ensemble machine learning algorithm, to analyze survival data and rank variables. It

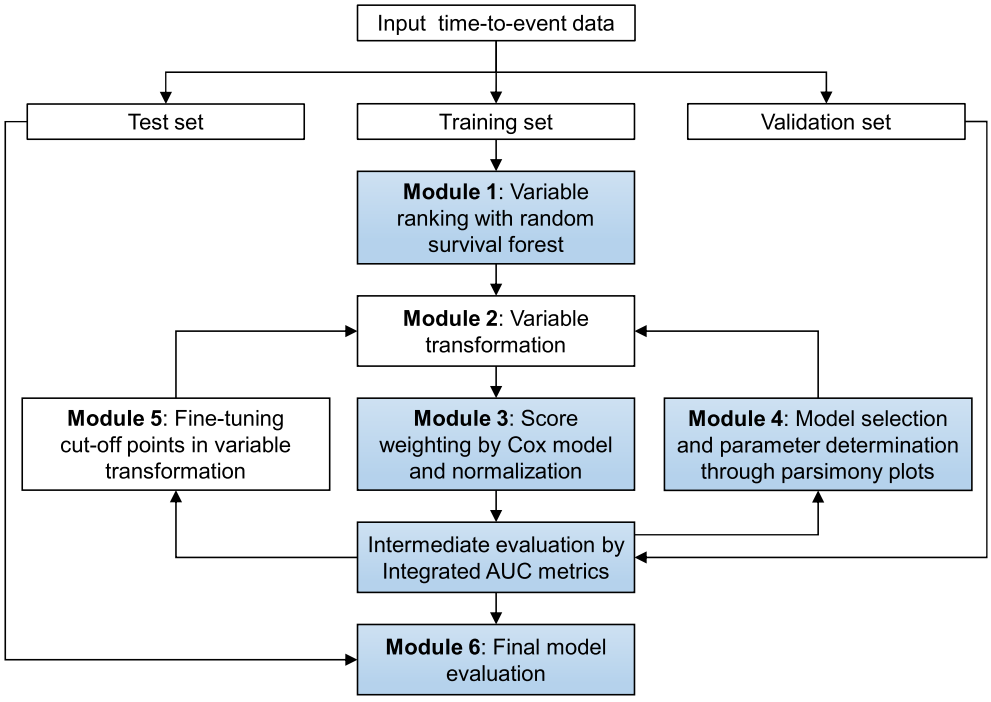

Fig. 1. Flowchart of the AutoScore-Survival framework. The blue shadow blocks are unique in the AutoScore-Survival compared with the original AutoScore. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

consists of a number of binary survival trees grown by a recursive splitting of tree nodes [26]. Central elements of the RSF algorithm are growing survival trees by maximizing survival difference (log-rank test statistic) [27] and estimating the survival probabilities based on the ensemble cumulative hazard function. RSF exerts two forms of randomization at the ensembling process: a bootstrapping sample of data and a randomly selected subset of variables. Averaging over survival trees, as well as the two forms of randomization, makes RSF much more accurate in prediction [20] and variable ranking [28,29]. The details of RSF are elaborated in eTextbox 1 of the Appendix.

Then, variable importance is calculated based on the corresponding reduction of predictive accuracy when the variable of interest is replaced with its random permutation value [25]. Unlike the traditional Cox regression, RSF does not assume proportional hazard or any functional form for the hazard function and works well for high-dimensional EHR data [30]. The ranking list will be used in subsequent modules for model building.

After variable selection, all selected variables are preprocessed for variable transformation [5,24]. That is, continuous variables are automatically converted into categorical ones through quantiles (e.g., 0%, *5%*, *20%*, *80%*, *95%*, 100%) or *k*-means clustering (e.g., $k = 5$) [31]. In this use case, quantiles were appropriate for most variables (such as common vital signs and laboratory test results), especially those with normal or near-normal distributions.

2.2. Score derivation by weighting and normalization at Module 3

Similar to AutoScore, in AutoScore-Survival models are built using the training set by selecting top-ranking variables from the ranking list, and continuous variables are transformed into categorical variables. With the selected and transformed variables, we create a time-to-event score for the survival data based on Cox Regression [32], with which the points can be easily interpreted:

$$h(t,X) = h_0(t) \times e^{-(\beta_1 X_1 + \cdots + \beta_m X_m)} \quad (1)$$

where t represents the survival time, $h(t,X)$ is the hazard function given variables $(X_1 \cdots X_m)$, $(\beta_1 \cdots \beta_m)$ are the coefficients for each variable, and $h_0(t)$ is the baseline hazard. The Cox regression does not make parametric assumptions on $h_0(t)$. Weibull and log-normal models can also be used as the weighting function, where $h_0(t)$ is assumed specific functional forms [33].

Based on equation (1), a partial score is assigned to each category of the variable, which is derived from the coefficients through a two-step procedure. The first step is to change the reference category in each variable to the category with the smallest β coefficient from the first-step regression such that all scores are non-negative. Next, the second-step regression is performed to generate new coefficients. The partial scores are derived from the second-step regression by dividing each coefficient by the minimum of all β's, and the results are rounded to the nearest integer. With a partial integer score associated with each category of a variable, the total score for each patient is computed by summing up all partial scores.

2.3. Model selection under the intermediate performance evaluation at Module 4

The validation set is used for the intermediate performance evaluation. We use a survival parsimony plot to visualize the change in model performance with an increasing number of variables, which helps us select a model that balances prediction accuracy and parsimony. For time-to-event outcomes, the time-dependent area under the curve or AUC(t) [34] is applied to measure model performance, which is an extension of the commonly used area under the curve (AUC) for measuring predictive accuracy of a score when studying binary outcomes. We chose the AUC(t) defined by cumulative sensitivity and dynamic specificity (C/D) as recommended by a comprehensive review [35], as this definition has more clinical relevance and has commonly been used by clinical applications [36]. This AUC(t) is introduced as a function of time, estimated through the Kaplan-Meier estimator of the survival function [34], to characterize how well the score can distinguish between subjects who had an event <= t and those who remained event-free at time t. To obtain a single overall performance metric in the parsimony plot, we derived the integrated AUC (iAUC), a weighted average of AUC(t) [37] over a follow-up period (i.e., from Day 1 to Day 90), summarizing the overall discrimination ability of the time-to-event score (see eTextbox 2 for details).

2.4. Final predictive performance evaluation at Module 6

We evaluate the final time-to-event score in the test set using multiple performance metrics. In addition to iAUC and AUC(t), we used the Harrell's concordance index (C-index) [38,39], which is the proportion of concordant pairs (i.e., when the observation with a longer survival time has a larger time-to-event score) in all pairs formed in the test set (see eTextbox 3 for details). Thus, the C-index is able to summarize risk, event occurrence, and survival time in a single number to distinguish between well-behaved scores and quasi-random ones [40].

2.5. Algorithm implementation and empirical validations

We implemented the AutoScore-Survival framework as an R package [41]. Given a new dataset with time-to-event outcomes and baseline covariates, the AutoScore-Survival package provides a pipeline of functions to split data and implement the six modules to generate the final scores that require minimal manipulation from users.

We demonstrated our AutoScore-Survival algorithm using the same dataset as our previous paper [5], including 44,918 ICU admissions from Beth Israel Deaconess Medical Center (BIDMC) [42] with 24 available variables of demographic information, vital signs, and lab tests at baseline (t = 0). The survival status and the date of death were additionally obtained from the database to derive the 90-day survival as the primary outcome. The baseline characteristics of the dataset were described through univariable and multivariable Cox regression. The Kaplan-Meier survival curves were generated for different risk groups stratified by the scores and compared through the log-rank test. Furthermore, we computed the 10th/25th and 50th percentile survival time and actual survival probabilities at different time points for each stratified group. To evaluate the performance of AutoScore-Survival, we compared it with several standard time-to-event prediction models. We considered the Cox model with (i) all variables and that with variables selected using (ii) stepwise and (iii) LASSO [43] approaches. The stepwise variable selection used AIC and considered both directions in each step. The regularization rate of LASSO was optimized through 10-fold cross-validation). We also built an RSF using all variables, with the widely-accepted default parameters [44] (i.e., 500 trees grown). Two widely used ICU-based clinical scores, such as Sequential Organ Failure Assessment (SOFA) [45] with eight variables, Simplified Acute Physiology Score (SAPS) [46] with 14 variables, were also involved in the comparison. Although these two scores were developed based on a binary outcome of ICU mortality instead of time-to-event outcomes, they consist of a comparable set of variables, including most vital signs and lab tests data collected in ICU, and thus, become reasonable comparators. Model performance was reported on the test set, and 100 bootstrapped samples were applied to calculate 95% confidence intervals (CI) [47].

3. Results

3.1. Cohort formation and basic covariates analysis

Overall survival probability was estimated by the Kaplan-Meier

Table 1
Univariable and multivariable survival analysis of all variables in the study cohort (N = 44,918).

	Unadjusted HR (95% CI)	p-Value	Adjusted HR (95% CI)	Adjusted p-Value
Age (years)	1.032 (1.031–1.034)	<0.001	1.027 (1.025–1.029)	<0.001
Gender				
Female	Baseline		Baseline	
Male	0.966 (0.922–1.011)	0.135	1.088 (1.037–1.142)	0.001
Ethnicity				
White	Baseline		Baseline	
Hispanic	0.368 (0.305–0.444)	<0.001	0.844 (0.695–1.026)	0.089
Asian	0.527 (0.480–0.578)	<0.001	0.968 (0.873–1.073)	0.533
African	0.482 (0.456–0.510)	<0.001	0.891 (0.834–0.953)	0.001
Others	0.502 (0.386–0.653)	<0.001	1.369 (1.045–1.795)	0.023
Insurance				
Medicare	Baseline		Baseline	
Government	1.432 (1.166–1.757)	0.001	1.146 (0.933–1.407)	0.194
Medicaid	2.717 (2.253–3.276)	<0.001	1.184 (0.975–1.439)	0.089
Private	1.311 (1.082–1.587)	0.006	1.056 (0.870–1.280)	0.582
Self-Pay	1.363 (0.989–1.879)	0.058	1.622 (1.176–2.237)	0.003
Heart rate (beats/min)	1.017 (1.015–1.018)	<0.001	1.016 (1.014–1.018)	<0.001
Systolic blood pressure (mmHg)	0.986 (0.985–0.988)	<0.001	0.988 (0.986–0.991)	<0.001
Diastolic blood pressure (mmHg)	0.978 (0.976–0.980)	<0.001	0.990 (0.985–0.995)	<0.001
Mean arterial pressure (MAP; mmHg)	0.977 (0.975–0.979)	<0.001	1.015 (1.009–1.021)	<0.001
Respiration rate (breaths/min)	1.097 (1.092–1.103)	<0.001	1.058 (1.052–1.065)	<0.001
Temperature (℃)	0.693 (0.666–0.721)	<0.001	0.796 (0.764–0.829)	<0.001
Peripheral capillary oxygen saturation (SpO2; %)	0.923 (0.916–0.931)	<0.001	0.981 (0.972–0.991)	<0.001
Glucose (mg/dL)	1.003 (1.003–1.004)	<0.001	1.000 (1.000–1.001)	0.577
Anion gap (mEq/L)	1.108 (1.102–1.114)	<0.001	1.036 (1.023–1.050)	<0.001
Bicarbonate (mmol/L)	0.961 (0.956–0.966)	<0.001	0.984 (0.973–0.995)	0.005
Creatinine (μmol/L)	1.095 (1.084–1.106)	<0.001	0.911 (0.893–0.930)	<0.001
Chloride (mEq/L)	0.970 (0.966–0.974)	<0.001	0.962 (0.953–0.972)	<0.001
Lactate (mmol/L)	1.243 (1.229–1.257)	<0.001	1.117 (1.100–1.134)	<0.001
Hemoglobin (g/dL)	0.849 (0.839–0.860)	<0.001	0.690 (0.657–0.724)	<0.001
Hematocrit (%)	0.956 (0.952–0.960)	<0.001	1.101 (1.082–1.119)	<0.001
Platelet (thousand per microliter)	1.000 (1.000–1.000)	0.014	0.999 (0.999–0.999)	<0.001
Potassium (mmol/L)	1.129 (1.090–1.169)	<0.001	0.886 (0.851–0.922)	<0.001
Blood urea nitrogen (BUN; mg/dL)	1.017 (1.016–1.018)	<0.001	1.013 (1.012–1.014)	<0.001
Sodium (mmol/L)	0.990 (0.985–0.996)	<0.001	1.020 (1.009–1.031)	0.001
White blood cells (thousand per microliter)	1.007 (1.006–1.007)	<0.001	1.005 (1.004–1.006)	<0.001

method (see Appendix eFig. 1). 37,462 (83.4%) admission episodes survived longer than 90 days and were censored at the end of the 90-day observation window. 7456 (16.6%) episodes died within 90 days, with a median survival time of 15 (IQR: 6–38) days and a mean survival time of 24.7 days (SD = 24.0). Table 1 summarizes the univariable and multivariable Cox analyses of all prognostic factors. All variables except gender got $P < 0.001$, making it hard to select a parsimonious model according to P values.

3.2. Parsimony plot and time-to-event scores

AutoScore-Survival selected seven variables by the parsimony plot (Fig. 2a) based on the validation set, as it achieved a good balance between model performance (i.e., iAUC) and complexity (number of variables, m). When more variables were added to the time-to-event score, the performance was not markedly improved. Fig. 2(b) based on the test set also demonstrated the trend.

The seven-variable time-to-event scores, derived from age, blood urea nitrogen, respiration rate, creatinine, anion gaps, lactate levels, and temperature, are tabulated in Table 2. The final score ranges from 0 to 100, where a smaller score indicates a higher survival probability. Table 3 shows different score intervals and their corresponding percentile survival time and survival probability estimated using the Kaplan-Meier method. The survival probability at 3, 7, 30, and 90 days decreases with increasing time-to-event scores, as expected. Scores larger than 60 correspond to a 90-day survival probability of lower than 50%. Table 3 and Fig. 3(a) offer a correspondence of scores and predicted survival probability based on the training set. As shown in Fig. 3 (b), the time-to-event score is able to accurately stratify patients in the test set into risk groups based on the Kaplan-Meier curve ($P < 0.0001$).

3.3. Performance evaluation and comparison

The performance of various methods evaluated in the unseen test set was reported in Table 4. The seven-variable AutoScore-Survival scores achieved an iAUC of 0.782 (95% CI: 0.767–0.794) and a C-index of 0.753 (95% CI: 0.740–0.762), comparable to the Cox regression with all 24 variables with an iAUC of 0.785 (95% CI: 0.768–0.798) and a C-index of 0.759 (95% CI: 0.748–0.769). LASSO and stepwise Cox regression achieved a comparable iAUC of 0.782 (95% CI: 0.766–0.795) and 0.785 (95% CI: 0.772–0.799) as well. But they selected 17 or 22 variables, respectively, failing to filter out redundant information efficiently to build up a parsimonious model for easy interpretation, compared with only seven variables of the AutoScore-Survival. Although the full RSF model achieved the highest iAUC and C-index in our experiment, consisting of a number of separate survival trees makes it become a black box and not interpretable enough for real-world applications. In terms of time-dependent AUC(t = 3,7,30,90), our seven-variable time-to-event score achieved comparable performances to 24-variable and stepwise Cox regression or LASSO. Furthermore, traditional ICU scores such as SOFA and SAPS achieved a much lower AUC(t = 3,7,30,90), even with 8 and 14 variables, respectively. As these two scores were not originally developed on time-to-event outcomes, iAUC and C-index were not calculated for them.

4. Discussion

In the present study, we developed AutoScore-Survival by extending the AutoScore method [5] to time-to-event outcomes and demonstrated its application by creating a time-to-event score using real-world data on 90-day survival in ICU. The score generated by the AutoScore-Survival was comparable with other standard survival prediction methods (i.e.,

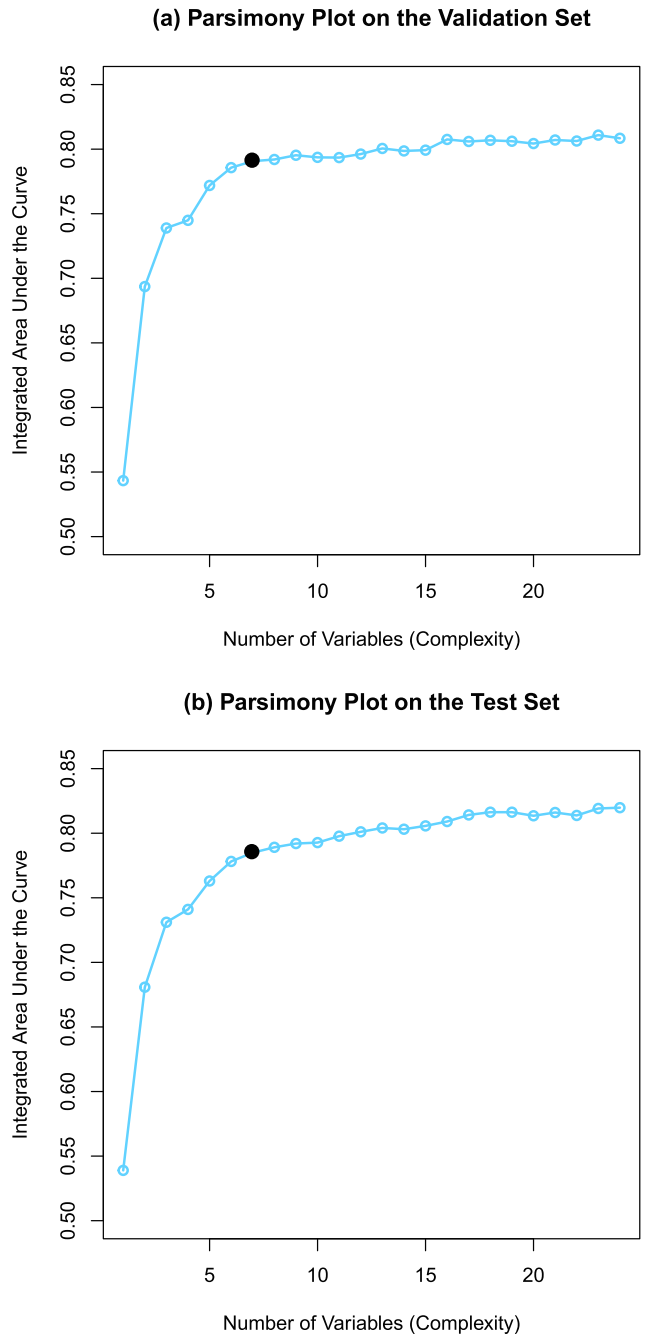

Fig. 2. Parsimony plot (model performance versus complexity) by the integrated area under the curve on the (a) validation set and (b) test set. The solid black dots show the selected point for achieving the model parsimony (i.e., number of variables $m = 7$).

Cox regression, LASSO, stepwise selection approach) and two clinical scores (i.e., SOFA and SAPS) in terms of discriminative capability. Although RSF outperformed at the model accuracy, more importantly, the AutoScore-Survival scores showed superior interpretability as a parsimonious point-based single indicative score for predicting patients' overall survival. This study's novelty is to integrate the advantages of RSF for robust variable selection and Cox regression in its accessibility for a generic methodology of quickly creating parsimonious time-to-event scores based on survival data. Future studies could apply it to various real-world clinical data (e.g., higher-dimensional or from different clinical settings) to develop useful time-to-event scores across diverse clinical backgrounds.

Table 2
Seven-variable AutoScore-Survival-derived scoring system.

	Variable and Interval	Partial Score
Age (years)	<30	0
	[30,48)	8
	[48,78)	15
	[78,85)	22
	>=85	25
Blood urea nitrogen (BUN; mg/dL)	<7.5	0
	[7.5,8.25)	17
	[8.25,12)	1
	>=12	8
Respiration rate (breaths/min)	<12	6
	[12,16)	0
	[16,22)	4
	>=22	11
Creatinine (mg/dL)	<0.5	14
	[0.5,0.8)	4
	[0.8,1.6)	0
	>=1.6	1
Anion Gap (mEq/L)	<15	0
	[15,20)	4
	>=20	7
Lactate (mmol/L)	<1	0
	[1,2.5)	3
	[2.5,4)	6
	>=4	15
Temperature (℃)	<36	11
	[36,36.5)	4
	[36.5,37.3)	0
	[37.3,38)	3
	>=38	6

The proposed AutoScore-Survival has several advantages in generating time-to-event outcome predictive scores. First, AutoScore-Survival could generate a parsimonious score by RSF-based variable selection, which has been shown to identify critical variables in high-dimensional data with multicollearity [48]. In our example, AutoScore-Survival achieved comparable performance with survival models of all 24 variables by selecting only seven variables, while LASSO and stepwise approaches failed to select a parsimonious list of variables. Second, AutoScore-Survival categorizes continuous variables to account for possible non-linear (e.g., U-shaped [49]) effects. Although some advanced regression methods handle nonlinearity [50], categorization is favorable in clinical and epidemiological applications for straightforward indication of identifying high-risk and low-risk values. Third, the scoring system is much easier to use and understand for healthcare professionals. Compared with other decimal predictors or probability outputs, integer time-to-event scores let users make quick predictions by simple arithmetic and gauge the effect of changing variables [40]. Our scores also gain high efficiency in terms of computational complexity even without the need for a computer, making it suitable for resource-challenged regions like rural areas. Thus, our scores have the advantage of accessibility and easy implementation, especially at the bedside. Furthermore, we derived scores from the Cox regression that is familiar to clinicians and does not make restrictive assumptions on the baseline hazard. Still, our R package [41] allows users to choose parametric survival models (Weibull and log-normal regression) as the weighting function. Besides right-censored data, left-censored and interval-censored Cox models [51] could be further extended by creating a different survival object using the *Surv* function.

We illustrated the application of AutoScore-Survival in acute care settings, where the better performance at the earlier times (e.g., the higher AUC(t = 3) and AUC(t = 7)) could help identify patients who need intensive care or extra medical attention. AutoScore-Survival is also useful for clinical decision-making on chronic diseases, e.g., cancer treatment and management [52]. For example, a small score or long survival time might indicate more aggressive and progressive cancer treatment. In contrast, a short survival time might be the indicator for palliative care [53] to optimize the quality of life and mitigate patients'

Table 3
Time-to-event score intervals and their corresponding percentile survival time or survival probability at different time points.

Score Value	Percent of patients	10th Percentile Survival Time (days)	25th Percentile Survival Time (days)	Median Survival Time (days)	Survival probability at three days (%)	Survival probability at seven days (%)	Survival probability at 30 days (%)	Survival probability at 90 days (%)
≤20	5.33%	90+	90+	90+	100.0%	99.8%	99.4%	98.7%
(20,30]	18.46%	90+	90+	90+	99.5%	99.2%	97.1%	95.4%
(30,40]	39.55%	87	90+	90+	99.2%	97.6%	93.9%	89.8%
(40,50]	24.16%	12	90+	90+	96.3%	92.5%	83.5%	76.6%
(50,60]	9.60%	5	19	90+	92.8%	85.6%	70.6%	58.6%
>60	2.91%	2	7	38	83.9%	72.8%	53.3%	43.7%

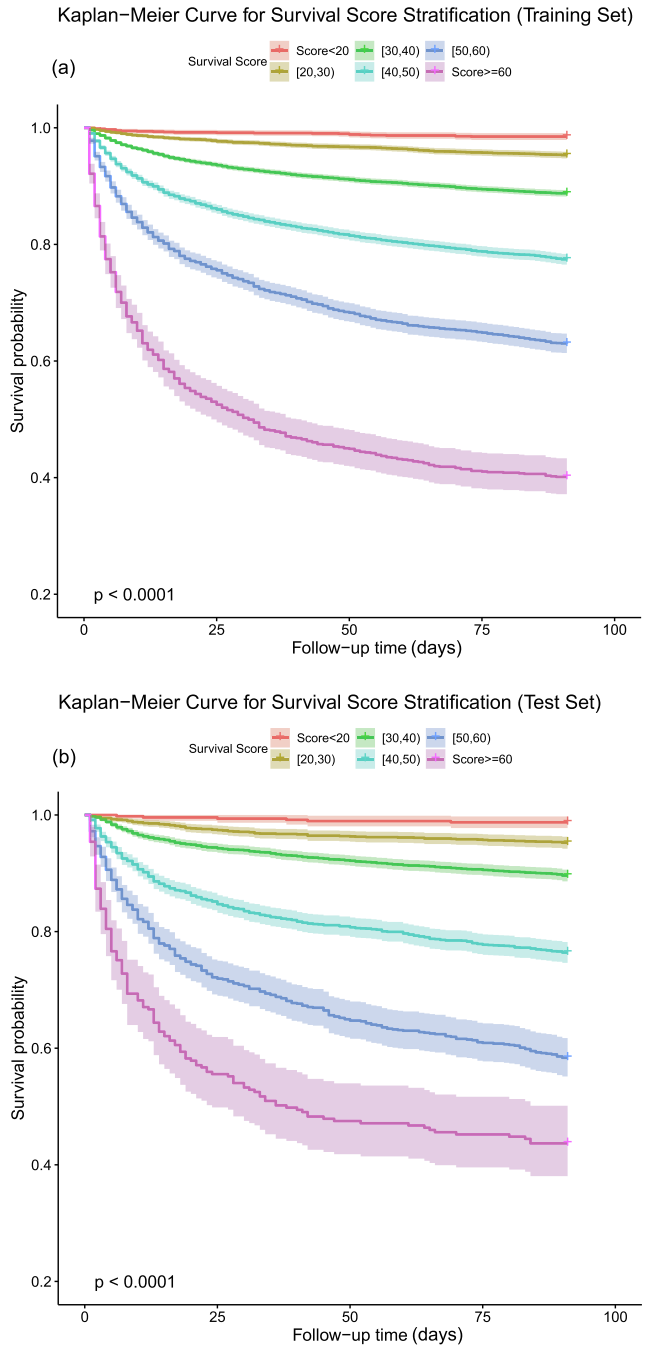

Fig. 3. Actual overall survival through risk stratification by AutoScore-Survival scores on the (a) training set and (b) test set (Kaplan-Meier estimates).

suffering. In our study, Table 3 and Fig. 3(a) linked the time-to-event scores with survival time and probability. It can help stratify patients into risk groups for appropriate allocation of therapeutic and care strategies [54].

There has been a growing interest in developing time-to-event scores in the clinical literature. For example, Kim et al. [55] recently built a scoring system to predict the overall survival of patients with advanced gastric cancer. Becker et al. [56] and Sharma et al. [57] also developed a time-to-event score for patients with general cancer and hepatocellular carcinoma, respectively. However, all of them were developed in an ad-hoc way. AutoScore-Survival provides a systematic guideline for automated development of time-to-event scores, contributing to data-driven research on various types of diseases. In comparison, traditional ICU-based scores (i.e., SAPS and SOFA) did not perform well in the test set, possibly due to the population shift over time. AutoScore could automatically update traditional clinical scores to make them keep excelling over time. In addition, AutoScore-Survival scores achieved good performances for multiple time points (t = 3,7,30,90), facilitating the generalizability of our scoring systems in different scenarios.

This study also has several limitations. First, we demonstrated the effectiveness of AutoScore-Survival for generating time-to-event scores using a single dataset extracted from the EHR, which includes 24 variables and more than 40,000 observations. While this dataset represents typical clinical data regarding the number of total variables and sample size for building up a new scoring system, a higher-dimensional dataset or EHR data from other clinical settings should be applied to evaluate and validate AutoScore-Survival in future research. Second, the scores represent the relative risk in the population if baseline hazard is ignored under Cox regression and adjustment is needed in different horizons. Third, computational efficiency for developing the scores might become a potential challenge in real-world implementation and was not compared among other methods, suggesting an area of possible future work to address this issue. At last, this is the initial development of AutoScore-Survival, where we selected commonly used methods to build our framework and demonstrated its usage using EHR data. Our demonstration using a large clinical dataset for survival prediction is aligned with our aim of developing a parsimonious time-to-event scoring system for clinical practices and provides an excellent reference for other clinical applications. Further development could extend the framework with advanced algorithms and apply it in various clinical use cases.

Funding

This study was supported by Duke-NUS Medical School, Singapore. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CRediT authorship contribution statement

Feng Xie: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing – original

Table 4
Performance of the AutoScore-Survival and other baseline models.

	AutoScore-Survival	Full Cox Regression	Full Random Survival Forest	Regularized Cox regression (LASSO)	Stepwise Cox regression	SOFA	SAPS
m	7	24	24	17	22	8	14
iAUC	0.782 (0.767–0.794)	0.785 (0.768–0.798)	0.843 (0.829–0.854)	0.782 (0.766–0.795)	0.785 (0.772–0.799)	–	–
C-index	0.753 (0.740–0.762)	0.759 (0.748–0.769)	0.808 (0.801–0.817)	0.755 (0.746–0.766)	0.759 (0.751–0.769)	–	–
AUC (t = 3)	0.805 (0.776–0.827)	0.781 (0.750–0.810)	0.852 (0.829–0.871)	0.782 (0.754–0.817)	0.782 (0.752–0.815)	0.738 (0.709–0.774)	0.785 (0.760–0.808)
AUC (t = 7)	0.787 (0.771–0.805)	0.787 (0.768–0.804)	0.843 (0.828–0.859)	0.785 (0.762–0.804)	0.788 (0.770–0.808)	0.705 (0.676–0.729)	0.744 (0.721–0.763)
AUC (t = 30)	0.773 (0.756–0.785)	0.786 (0.774–0.800)	0.841 (0.829–0.851)	0.780 (0.767–0.795)	0.785 (0.774–0.798)	0.685 (0.666–0.705)	0.714 (0.693–0.733)
AUC (t = 90)	0.763 (0.751–0.773)	0.778 (0.766–0.788)	0.829 (0.820–0.838)	0.774 (0.764–0.785)	0.778 (0.769–0.790)	0.664 (0.647–0.679)	0.691 (0.675–0.705)

iAUC, the integrated AUC(t); AUC(t), the time-dependent area under the curve
C-index, concordance index; LASSO, Least Absolute Shrinkage and Selection Operator for Cox regression
SOFA, Sequential Organ Failure Assessment
SAPS, Simplified Acute Physiology Score

draft, Writing – review & editing. **Yilin Ning:** Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Han Yuan:** Data curation, Investigation, Methodology, Writing – review & editing. **Benjamin Alan Goldstein:** Investigation, Methodology, Validation, Writing – review & editing. **Marcus Eng Hock Ong:** Investigation, Methodology, Validation, Writing – review & editing. **Nan Liu:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Bibhas Chakraborty:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103959.

References

[1] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: fundamental principles and 10 grand challenges, *arXiv preprint arXiv:210311251*, 2021.
[2] M.A. Ahmad, C. Eckert, A. Teredesai, Interpretable machine learning in healthcare, in: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018, 2018, pp. 559–560.
[3] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, Machine Learning 102 (3) (2016) 349–391.
[4] B. Ustun, C. Rudin, Learning optimized risk scores, J. Machine Learning Res. 20 (2019) 1–75.
[5] F. Xie, B. Chakraborty, M.E.H. Ong, B.A. Goldstein, N. Liu, AutoScore: a machine learning-based automatic clinical score generator and its application to mortality prediction using electronic health records, JMIR Med. Inform. 8 (10) (2020) e21798, https://doi.org/10.2196/21798.
[6] F. Xie, M.E.H. Ong, J.N.M.H. Liew, K.B.K. Tan, A.F.W. Ho, G.D. Nadarajan, L. L. Low, Y.H. Kwan, B.A. Goldstein, D.B. Matchar, B. Chakraborty, N. Liu, Development and assessment of an interpretable machine learning triage tool for estimating mortality after emergency admissions, JAMA Netw. Open 4 (8) (2021) e2118467, https://doi.org/10.1001/jamanetworkopen.2021.18467.
[7] J. Waring, C. Lindvall, R. Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, Artif. Intell. Med. 104 (2020) 101822, https://doi.org/10.1016/j.artmed.2020.101822.
[8] K. Angelo, A. Dalhaug, A. Pawinski, E. Haukland, C. Nieder, Survival prediction score: a simple but age-dependent method predicting prognosis in patients undergoing palliative radiotherapy, ISRN Oncol. 2014 (2014) 1–5.
[9] M. Maltoni, O. Nanni, M. Pirovano, E. Scarpi, M. Indelli, C. Martini, M. Monti, E. Arnoldi, L. Piva, A. Ravaioli, G. Cruciani, R. Labianca, D. Amadori, Successful validation of the palliative prognostic score in terminally ill cancer patients. Italian multicenter study group on palliative care, J Pain Symptom Manage 17 (4) (1999) 240–247.
[10] M. Schmidt, M. Bailey, J. Sheldrake, C. Hodgson, C. Aubron, P.T. Rycus, C. Scheinkestel, D.J. Cooper, D. Brodie, V. Pellegrino, A. Combes, D. Pilcher, Predicting survival after extracorporeal membrane oxygenation for severe acute respiratory failure. The respiratory extracorporeal membrane oxygenation survival prediction (RESP) score, Am. J. Respir. Crit. Care Med. 189 (11) (2014) 1374–1382.
[11] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.
[12] H. Liang, G. Zou, Improved AIC selection strategy for survival analysis, Comput. Stat. Data Anal. 52 (5) (2008) 2538–2548.
[13] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[14] C.T. Volinsky, A.E. Raftery, Bayesian information criterion for censored survival models, Biometrics 56 (2000) 256–262.
[15] J. Fan, G. Li, R. Li, An overview on variable selection for survival analysis, in: Contemporary Multivariate Analysis And Design Of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday, World Scientific, 2005, pp. 315–336.
[16] R. Tibshirani, The lasso method for variable selection in the cox model, Stat. Med. 16 (4) (1997) 385–395.
[17] B.A. Goldstein, A.M. Navar, M.J. Pencina, J.P. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 24 (2017) 198–208.
[18] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, Random survival forests, Ann. Appl. Statistics 2 (841–60) (2008) 20.

[19] H. Tin Kam, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995 14-16 Aug. 1995, vol. 1, 1995, pp. 278–282.
[20] S. Wongvibulsin, K.C. Wu, S.L. Zeger, Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis, BMC Med. Res. Methodol. 20 (2019) 1.
[21] A. Spooner, E. Chen, A. Sowmya, P. Sachdev, N.A. Kochan, J. Trollor, H. Brodaty, A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction, Sci. Rep. 10 (1) (2020), https://doi.org/10.1038/s41598-020-77220-w.
[22] V. Van Belle, K. Pelckmans, S. Van Huffel, J.A.K. Suykens, Support vector methods for survival analysis: a comparison between ranking and regression approaches, Artif. Intell. Med. 53 (2) (2011) 107–118.
[23] D.W. Kim, S. Lee, S. Kwon, W. Nam, I.H. Cha, H.J. Kim, Deep learning-based survival prediction of oral cancer patients, Sci. Rep. 9 (2019) 6994.
[24] F. Xie, Y. Ning, H. Yuan, S.E. Saffari, B. Chakraborty, N. Liu, Package 'AutoScore': An Interpretable Machine Learning-Based Automatic Clinical Score Generator. R package version, 2021. Available from: <https://cran.r-project.org/web/packages/AutoScore/AutoScore.pdf>.
[25] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
[26] M.R. Segal, Regression trees for censored data, Biometrics 44 (1) (1988) 35, https://doi.org/10.2307/2531894.
[27] M. Leblanc, J. Crowley, Survival trees by goodness of split, J. Am. Stat. Assoc. 88 (422) (1993) 457–467.
[28] O. Hamidi, J. Poorolajal, M. Farhadian, L. Tapak, Identifying important risk factors for survival in kidney graft failure patients using random survival forests, Iran. J. Public Health 45 (2016) 27–33.
[29] E. Hsich, E.Z. Gorodeski, E.H. Blackstone, H. Ishwaran, M.S. Lauer, Identifying important risk factors for survival in patient with systolic heart failure using random survival forests, Circ. Cardiovasc. Qual. Outcomes 4 (1) (2011) 39–45.
[30] H. Wang, G. Li, A selective review on random survival forests for high dimensional data, Quant. Biosci. 36 (2) (2017) 85–96.
[31] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, Oakland, CA, USA, 1967, pp. 281–297.
[32] D.R. Cox, Regression models and life-tables, J. Roy. Stat. Soc.: Ser. B (Methodol.) 34 (1972) 187–202.
[33] F.E. Harrell, Parametric survival models, in: J.F.E. Harrell (Ed.), Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Springer International Publishing, Cham, 2015, pp. 423–451.
[34] P.J. Heagerty, T. Lumley, M.S. Pepe, Time-dependent ROC curves for censored survival data and a diagnostic marker, Biometrics 56 (2000) 337–344.
[35] A.N. Kamarudin, T. Cox, R. Kolamunnage-Dona, Time-dependent ROC curve analysis in medical research: current methods and applications, BMC Med. Res. Methodol. 17 (2017) 53.
[36] J. Lambert, S. Chevret, Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves, Stat. Methods Med. Res. 25 (5) (2016) 2088–2102.
[37] P.J. Heagerty, Y. Zheng, Survival model predictive accuracy and ROC curves, Biometrics 61 (1) (2005) 92–105.
[38] M.J. Pencina, R.B. D'Agostino, Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation, Stat. Med. 23 (13) (2004) 2109–2123.
[39] F.E. Harrell, K.L. Lee, D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Stat. Med. 15 (4) (1996) 361–387.
[40] E. Longato, M. Vettoretti, B. Di Camillo, A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models, J. Biomed. Inform. 108 (2020) 103496, https://doi.org/10.1016/j.jbi.2020.103496.
[41] *AutoScore-Survival R package*. Available from: <https://github.com/nliulab/AutoScore-Survival>.
[42] A.E.W. Johnson, T.J. Pollard, L.u. Shen, L.-W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016), https://doi.org/10.1038/sdata.2016.35.
[43] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010) 1–22.
[44] P. Probst, Boulesteix A-LJTJoMLR, To tune or not to tune the number of trees in random forest 18 (2017) 6673–6690.
[45] J.-L. Vincent, A. de Mendonca, F. Cantraine, R. Moreno, J. Takala, P.M. Suter, C. L. Sprung, F. Colardyn, S. Blecher, Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on "sepsis-related problems" of the European Society of Intensive Care Medicine, Crit. Care Med. 26 (11) (1998) 1793–1800.
[46] J.-R. Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, D. Villers, A simplified acute physiology score for ICU patients, Crit. Care Med. 12 (11) (1984) 975–977.
[47] F. Xie, N. Liu, S.X. Wu, Y. Ang, L.L. Low, A.F.W. Ho, S.S.W. Lam, D.B. Matchar, M. E.H. Ong, B. Chakraborty, Novel model for predicting inpatient mortality after emergency admission to hospital in Singapore: retrospective observational study, BMJ Open 9 (9) (2019) e031382, https://doi.org/10.1136/bmjopen-2019-031382.
[48] S. Dietrich, A. Floegel, M. Troll, T. Kühn, W. Rathmann, A. Peters, D. Sookthai, M. von Bergen, R. Kaaks, J. Adamski, C. Prehn, H. Boeing, M.B. Schulze, T. Illig, T. Pischon, S. Knüppel, R. Wang-Sattler, D. Drogan, Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis, Int. J. Epidemiol. 45 (5) (2016) 1406–1420.
[49] D.-D. Yu, H. Dong, Z.-G. Wu, Y.-B. Xiao, C.-F. Zhou, Q.-Q. Wang, J. Cai, U-shaped relationship of age at diagnosis and cancer-specific mortality in primary urachal adenocarcinoma: a cohort study, Transl. Androl. Urol. 9 (3) (2020) 1073–1081.
[50] R. Andersen, Nonparametric methods for modeling nonlinearity in regression analysis, Ann. Rev. Soc. 35 (1) (2009) 67–85.
[51] D.M. Finkelstein, A proportional hazards model for interval-censored failure time data, Biometrics 42 (4) (1986) 845, https://doi.org/10.2307/2530698.
[52] A. Bashiri, M. Ghazisaeedi, R. Safdari, L. Shahmoradi, H. Ehtesham, Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review, Iran. J. Public Health 46 (2017) 165–172.
[53] C. Tournoux-Facon, X. Paoletti, J.-C. Barbare, O. Bouché, P. Rougier, L. Dahan, C. Lombard-Bohas, R. Faroux, J.L. Raoul, L. Bedenne, F. Bonnetain, Development and validation of a new prognostic score of death for patients with hepatocellular carcinoma in palliative setting, J. Hepatol. 54 (1) (2011) 108–114.
[54] M. Pirovano, M. Maltoni, O. Nanni, M. Marinari, M. Indelli, G. Zaninetta, V. Petrella, S. Barni, E. Zecca, E. Scarpi, R. Labianca, D. Amadori, G. Luporini, A new palliative prognostic score: a first step for the staging of terminally ill cancer patients. Italian multicenter and study group on palliative care, J. Pain Symptom Manage. 17 (4) (1999) 231–239.
[55] J. Kim, J.Y. Hong, S.T. Kim, S.H. Park, S.Y. Jekal, J.S. Choi, D.K. Chang, W.K. Kang, S.W. Seo, J. Lee, Clinical scoring system for the prediction of survival of patients with advanced gastric cancer, ESMO Open 5 (2) (2020) e000670, https://doi.org/10.1136/esmoopen-2020-000670.
[56] T. Becker, J. Weberpals, A.M. Jegg, W.V. So, A. Fischer, M. Weisser, F. Schmich, D. Rüttinger, A. Bauer-Mehren, An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort, Ann. Oncol. 31 (11) (2020) 1561–1568.
[57] S.A. Sharma, M. Kowgier, B.E. Hansen, W.P. Brouwer, R. Maan, D. Wong, H. Shah, K. Khalili, C. Yim, E.J. Heathcote, H.L.A. Janssen, M. Sherman, G.M. Hirschfield, J. J. Feld, Toronto HCC risk index: a validated scoring system to predict 10-year risk of HCC in patients with cirrhosis, J. Hepatol. 68 (1) (2018) 92–99.