Machine Learning with Applications 9 (2022) 100365

Contents lists available at ScienceDirect

Machine Learning with Applications

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

Time-to-event modeling for hospital length of stay prediction for COVID-19 patients

Yuxin Wen [a,1], Md Fashiar Rahman [b,1], Yan Zhuang [c,1], Michael Pokojovy [d,1], Honglun Xu [b,1], Peter McCaffrey [e,1], Alexander Vo [e,1], Eric Walser [e,1], Scott Moen [e,1], Tzu-Liang (Bill) Tseng [b,*,1]

[a] Dale E. and Sarah Ann Fowler School of Engineering, Chapman University, Orange, CA 92866, USA
[b] Department of Industrial, Manufacturing and Systems Engineering, The University of Texas at El Paso, El Paso, TX 79968, USA
[c] Department of Biomedical Engineering, Sichuan University, Chengdu, Sichuan 610065, China
[d] Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX 79968, USA
[e] The University of Texas Medical Branch, Galveston, TX 77550, USA

ARTICLE INFO

Keywords:
Length of stay
Survival analysis
Time-to-event modeling
Deep learning
COVID-19

ABSTRACT

Providing timely patient care while maintaining optimal resource utilization is one of the central operational challenges hospitals have been facing throughout the pandemic. Hospital length of stay (LOS) is an important indicator of hospital efficiency, quality of patient care, and operational resilience. Numerous researchers have developed regression or classification models to predict LOS. However, conventional models suffer from the lack of capability to make use of typically censored clinical data. We propose to use time-to-event modeling techniques, also known as survival analysis, to predict the LOS for patients based on individualized information collected from multiple sources. The performance of six proposed survival models is evaluated and compared based on clinical data from COVID-19 patients.

1. Introduction

Length of stay (LOS) refers to the cumulative duration of patient hospitalization between consecutive admission and discharge times over a given time horizon. Most hospitals face the challenges of providing timely patient care while maintaining optimal resource utilization, especially during the COVID-19 pandemic. According to an annual survey conducted by the American Hospital Association, admitted patients spent more than $1.16 trillion across all registered U.S. hospitals in 2019 (American Hospital Association, 2021). In the U.S., every 1-hour transfer delay is associated with an adjusted 3% increase in the odds of inpatient mortality (Churpek, Wendlandt, Zadravecz, Adhikari, Winslow, & Edelson, 2016). Moreover, from the medical perspective, prolonged LOS increases the risk of adverse events, such as poor nutritional levels, hospital-acquired infections, adverse drug events and other complications. From the hospital management perspective, prolonged LOS decreases the bed turnover rate, disrupts the patient flow and access to care due to bed shortages, which makes hospital strategic and operational management difficult. The ongoing COVID-19 pandemic has tremendously overloaded the healthcare systems leading to excessive demand for hospital beds. Therefore, quantifying and optimizing the LOS could improve the efficiency of hospital management, facilitate beneficial treatment, reduce wait times, and mitigate exposure to risks associated with hospitalization. Real-time demand capacity (RTDC) management has been considered a best practice to manage hospital capacity and improve patient flow (Resar, Nolan, Kaczynski, & Jensen, 2011). RTDC can be used to estimate the LOS and assist in performing adjustments by integrating four steps into bed management processes: predicting capacity, predicting demand, developing and evaluating a plan. However, this manual tool exhibits a number of flaws related to subjective outcomes and sole restriction to surgical departments. Past research has also attempted to group patients by their medical conditions. They assume that each disease or illness is associated with a recommended LOS, which refers to diagnosis-related group (DRG) systems. These systems assume that all patients who fall within the same DRG have identical LOS. However, LOS is a complex metric affected by many factors, including individual demographics, different treatment strategies and discharge planning, which may extend the LOS beyond the target range (Awad, Bader-El-Den, & McNicholas, 2017). Therefore, a personalized and accurate LOS prediction model is essential to improve hospital resources utilization and healthcare decision-making.

* Corresponding author.
E-mail addresses: yuwen@chapman.edu (Y. Wen), mrahman13@utep.edu (M.F. Rahman), 2550873360@qq.com (Y. Zhuang), mpokojovy@utep.edu (M. Pokojovy), hxu3@miners.utep.edu (H. Xu), pemccaff@utmb.edu (P. McCaffrey), ahvo@utmb.edu (A. Vo), emwalser@utmb.edu (E. Walser), stmoen@utmb.edu (S. Moen), btseng@utep.edu (T.-L. Tseng).
[1] All authors read and approved the final manuscript.

https://doi.org/10.1016/j.mlwa.2022.100365
Received 5 March 2022; Received in revised form 30 May 2022; Accepted 14 June 2022
Available online 18 June 2022

Over the past decade, the prediction of patient LOS with various diseases has been extensively investigated using several statistical and machine learning techniques, such as logistic regression, random forests, support vector machine (SVM), decision tree-based methods, etc. Luo et al. (Luo, Lian, Feng, Huang, & Zhang, 2017) proposed to use logistic regression and random forests to establish a model to predict the LOS of patients with chronic obstructive pulmonary disease. Tanuja et al. (Tanuja, Acharya, & Shailesh, 2011) compared multi-layer perceptron (MLP), naive Bayes, K-NN, and decision tree models to predict patients' LOS. Their results showed that MLP and naive Bayes models had the best classification accuracy of around 85%, while K-NN performed poorly with only 63.6% accuracy. Kulkarni et al. (Kulkarni, Thangam, & Amin, 2021) used MLP for the prediction of prolonged LOS in acute coronary syndrome patients. Dogu et al. (Dogu, Albayrak, & Tuncay, 2021) proposed an integrated approach combining fuzzy cognitive maps and artificial neural networks (ANN) for the LOS prediction for patients with chronic obstructive pulmonary disease. Barnes et al. (Barnes, Hamrock, Toerper, Siddiqui, & Levin, 2016) compared the prediction performance among random forests, logistic regression, and clinical decisions in application to inpatient LOS to support hospital discharge decisions. They found the machine learning models were more accurate than clinician predictions. A comprehensive review of patient LOS studies conducted by Awad et al. can be found in Awad et al. (2017). The emergence of advanced machine learning, in particular deep learning, has proved very powerful at distilling complex hidden relationships in the data and, thus, these methods typically demonstrate good prediction performance. The main advantage of deep learning approaches is their capability of automated extraction of complex data representations through end-to-end training from raw data, which significantly reduces the effort of manual feature engineering. Zebin & Chaussalet (Zebin & Chaussalet, 2019) proposed to use autoencoder dense neural networks to classify LOS into short (0–7 days) and long stays (>7 days) using the public MIMIC III dataset. Harerimana et al. (Harerimana, Kim, & Jang, 2021) proposed a hierarchical attention network to predict LOS and in-hospital mortality. The proposed model was able to leverage the patient anamnesis and free text diagnosis recorded on the first day for prediction purposes. Rajkomar et al. (2018) combined three different deep learning models and developed an ensemble model to predict hospital readmission and LOS.

Based on the aforecited literature, the existing LOS models can be grouped into two general categories: classification models and regression models. Classification models are usually used to predict categorical outcomes. In other words, the aim is to group the LOS into multiple classes, e.g., short stay, medium stay and long stay, based on the number of days that the patient stays in the hospital. However, several studies have demonstrated that the LOS distributions are highly skewed to the right (Harerimana et al., 2021; Ma, Yu, Ye, Yao, & Zhuang, 2020). This skewness indicates that the dataset becomes heavily imbalanced as only a few long LOS cases exist. This imbalance misleads the performance evaluation as classes with long LOS are deemed outliers by the model. Therefore, viewing the LOS task as a regression problem is a more appropriate and informative way of balancing the dataset by predicting the actual number of LOS days in lien of class labels. With this goal in mind, Caetano et al. (Caetano, Laureano, & Cortez, 2014) compared six regression techniques: taking average prediction, decision trees, multiple regression, ANN ensemble, random forests and SVM. They found that the best results were obtained by the random forest model. Rowan et al. (Rowan, Ryan, Hegarty, & O'Hare, 2007) found that ANNs could be an effective LOS stratification instrument in postoperative cardiac patients. Muhlestein et al. (Muhlestein, Akagi, Davies, & Chambless, 2019) presented a machine learning ensemble model to predict patient LOS after brain tumor surgery. Verburg et al. (Verburg, de Keizer, de Jonge, & Peek, 2014) compared eight regression models for modeling intensive care LOS and concluded that currently available models for ICU LOS are not suitable for predicting individual patient outcomes and should not be used as an indicator for ICU quality or efficiency. Moreover, Vekaria et al. (2021) pointed out that these data-driven methods failed to address the inherent uncertainty, complexity and heterogeneity in the healthcare field. It is frequently the case that data collected from clinical trials and cohort studies are high-dimensional, censored, heterogeneous by their nature and may have missing information, presenting additional challenges to traditional statistical analysis. For example, it is common in clinical studies to have subjects who did not experience the event of interest at the end of a study or dropped out before the event of interest occurs. These subjects are usually called to be right-censored. Although the data may seem to be incomplete for these subjects as the time-to-event is not actually observed, these subjects are highly valuable as the observation that they went a certain amount of time without experiencing an event is informative per se. Time-to-event modeling techniques, also referred to as survival analysis, have the capability to handle censored data that normally are disregarded by regular regression models. Several studies have applied time-to-event analysis in the clinical field (Schober & Vetter, 2018). The main advantage of using time-to-event data analysis is that such models can give a probability estimate instead of solely point estimates as conventional regression models. Moreover, survival models have the capability of incorporating censored data into models. The Cox proportional hazard (PH) model is the most frequently used technique for time-to-event analysis. The fundamental assumption in the Cox PH model is that the hazards are proportional by a linear combination of patient's covariates. In other words, the relative hazard remains constant over time with different predictor or covariate levels. However, this proportional assumption may limit its applicability to accounting for non-linear and complex relationships among various features in the dataset. Recently, a number of deep learning based time-to-event models have been developed, such as DeepSurv (Katzman, Shaham, Cloninger, Bates, Jiang, & Kluger, 2018), DeepHit (Lee, Zame, Yoon, & van der Schaar, 2018), Cox-CC (Kvamme, Borgan, & Scheel, 2019). Current research is predominantly focused on assessing the effect of factors, such as treatments that simultaneously allow one to control for the effects of other covariates and risk factors associated with prolonged LOS (Luth et al., 2021). To date, studies of COVID-19 mainly focused on epidemiological investigation, diagnosis and treatment, prevention and control (Wang et al., 2020). Fewer studies have investigated the COVID-19 patients' hospital LOS during the pandemic.

Driven by these issues, the primary goal of this study is to estimate the duration of hospital stay among patients admitted with COVID-19 using clinical data. In this paper, we systematically compare time-to-event models for individualized LOS prediction, which will be helpful to facilitate efficient medical resource allocation during the COVID-19 pandemic. Specifically, we apply six different time-to-event models, including Cox PH model, DeepSurv, Cox-CC, DeepHit, multi-task logistic regression (MTLR), random survival forest (RSF) and compare their performance.

The rest of this paper is organized as follows. Section 2 presents a brief review of time-to-event data-based LOS prediction. Section 3 summarizes the experimental settings, including data description and data preprocessing steps. The experimental findings, i.e., prediction performance results and discussion, is also reported in Section 3. The conclusions are given in Section 4.

2. Time-to-event models for LOS prediction

In this section, we provide a brief review of concepts employed in survival analysis and present models subsequently used in the paper. Survival analysis is a branch of statistics concerned with analyzing time-to-event data and predicting the probability of occurrence of an event. The event could be any format, such as recovery, relapse, or death, discharge, etc. The main benefit of time-to-event modeling is that the data related to participants who did not experience the event by the end of the study or were unavailable for follow-up (referred

to as data censoring) can still contribute to the analysis. However, in conventional regression models, censored data are typically discarded, which may introduce a bias into the model.

To estimate the LOS by using survival analysis tools, the objective is to model the LOS distribution as a function of time. Letting $f(t)$ and $F(t)$ denote the probability density function and the cumulative distribution function of the random time T, respectively, the goal is to find the distribution $P(T \le t) = \int_0^t f(u)\,du = F(t)$, which has a relationship with survival function $S(t)$ via $S(t) = P(T > t) = 1 - F(t)$. For patient i, define the associated event time $T_i = \min\{F_i, C_i\}$, where F_i denotes the event time and C_i denotes the censoring time. Let δ_i denote the indicator function taking value 1 if the event occurs for patient i at T_i, or taking value 0 if it is censored at T_i. Other predictors (e.g., demographics, vital signs, etc.) are denoted as a vector $\boldsymbol{Z}_i$. Now all observed data for patient i can be denoted as $\boldsymbol{D}_i = \{\boldsymbol{Z}_i, F_i, \delta_i\}$. Then the survival function for patient i $S_i(t) = P_i(T \ge t)$ represents the probability of "survival" up to time t. The hazard function is the instantaneous rate at which events occur for individuals who are "surviving" at time t

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \le t + \Delta t | T \ge t)}{\Delta t} \tag{1}$$

The term $\int_0^t h(u)\,du$ is then called cumulative hazard function and denoted by $H(t)$ (Kartsonaki, 2016). The hazard function is linked to survival function by $S(t) = \exp\left\{-\int_0^t h(u)\,du\right\}$. Accordingly, the full likelihood function, which are used for parameter estimation, can be written as the product of the survival and hazard functions

$$L = \prod_i L_i = h(t_i)^{\delta_i} S(t_i). \tag{2}$$

In addition, once we have $S(t)$, the mean or the expected value survival with boundary conditions $S(0) = 1$ and $S(\infty) = 0$ can be estimated by

$$\mu = \int_0^\infty S(t)\,dt \tag{3}$$

The one-dimensional integration can be easily performed numerically with usual approaches such as Gauss–Legendre quadrature method (Hildebrand, 1987).

In the context of the present paper, an important clarification needs to be made. When treating COVID-19 patients, any particular patient can and usually does have multiple admissions and discharges over any particular time frame, oftentimes due to transfer between different departments within the hospital or internal patient tracking peculiarities. Therefore, instead of counting time to the next discharge, a more adequate and useful LOS metric is the cumulative amount of time spent by a patient in a hospital over a certain time period (e.g., 30 days as we did in this paper) from a given encounter, which can encompass multiple non-overlapping stays at the hospital. Another advantage of this approach is that it is amenable to aggregation over population of patients and pragmatically useful for hospital management to understand predicted resource utilization over a reasonable time period and over a population of patients. Thus, the time T corresponds to the total time spent by a patient in the hospital within a certain window of time. Correspondingly, $F(t)$ describes the probability of T not exceeding t, μ is the expected time spent in the hospital and so on. The events of interest are discharged due to recovery or due to death. In sum, the model is used to predict the cumulative time spent in the hospital over the next 30 days rather than the time of the next discharge. Censoring effects can occur if the patient is transferred to another hospital and dies there.

A. Cox proportional hazard model

One class of prevailing survival models is given by the Cox PH model. Cox PH is a semiparametric model that quantifies the effects of observed covariates on the risk of an event occurring (e.g., death), which is one of the most popular methods for estimating the probability distribution of survival time based on one or more predictor variables in various fields (Wen, Guo, Son, & Wu, 2022). The functional form is defined as

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{Z}_i\right) \tag{4}$$

where $h_0(t)$ is a baseline hazard function shared by all individuals that can be in either a non-parametric or parametric form. $\boldsymbol{\omega}$ is a vector parameter to be estimated, representing the effect of covariates on the outcome. To estimate the unknown parameters $\boldsymbol{\omega}$, instead of using Eq. (2), partial likelihood is commonly used (Cox, 1972), which can be expressed as

$$PL(\boldsymbol{\omega}) = \prod_{i=1}^{N} \left(\frac{e^{\boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{Z}_i}}{\sum_{j \in R_i} e^{\boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{Z}_j}} \right)^{\delta_i} \tag{5}$$

where $R_i = \{1 \le j < N | T_j \ge T_i\}$ represents a risk set including all subjects at risk at time T_i. N is the total number of individuals. To maximize the Cox partial likelihood, Newton–Raphson's method is usually employed.

Lastly, for the baseline hazard estimation, Breslow approximations to the partial log-likelihood can be used (Breslow, 1972) as

$$\hat{h}_0(t_i) = \frac{d_i}{\sum_{j \in R_i} e^{\boldsymbol{\omega}^T \boldsymbol{Z}_j}}. \tag{6}$$

Where d_i is the number of events at time t_i. The model assumes that a patient's log-risk of failure is a linear combination of respective covariates. However, it may be too simplistic and limit the suitability of modeling non-linear interactions among prognostic factors. Therefore, more complex models are in demand to capture non-linear relationships. To overcome this well-known issue associated with Cox PH model, machine learning approaches, which facilitate the detection of relationships in complex datasets, have recently been employed for this purpose.

B. DeepSurv

ANNs have shown great potential when complex interactions or non-linear effects exist. Fig. 1 shows an example of a neural network-based Cox PH model architecture. The ANN architecture consists of inputs, one or more fully connected hidden layers, and a Cox-regression output layer. Replacing the exponential part $\boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{Z}$ of Eq. (4) by the output of ANN, the inputs are fused in a non-linear manner.

As an extension to the Cox PH model, DeepSurv is a deep feed-forward neural network, which involves modeling proportional hazard ratios over individuals using deep neural networks. It has the ability to learn non-linear hazard ratios (Katzman et al., 2018). In this model, the inputs to the network are the various patient information and diagnosis, and the hidden layers consist of a fully connected layer of nodes, followed by a dropout layer. The output layer is a single node with a linear activation which estimates the log-risk function in the Cox model. The loss function of the model is the average negative log-partial likelihood with regularization. Stochastic gradient descent optimization is used to find the optimal weights of the network.

C. Cox-CC

As shown in Eq. (5), to calculate the loss function, one needs to sum over all risk sets R_i, which can be extremely computationally expensive, especially for large datasets. To resolve this issue, Kvamme et al. (2019) proposed a case-control approximation to the loss function estimation. They fitted the Cox model with mini-batch stochastic gradient descent by approximating the risk set R_i with a reasonable portion subset $\widehat{R}_i$ and then weight the likelihood accordingly with weights w_i, which can be formulated as

$$PL(\boldsymbol{\omega}) = \prod_{i=1}^{N} \left(\frac{e^{\boldsymbol{\omega} Z_i}}{w_i \sum_{j \in \widehat{R}_i} e^{\boldsymbol{\omega} Z_j}} \right)^{\delta_i}. \tag{7}$$

The selected i's refer to *case* in case-control approximation. The weights w_i need to be determined to ensure the approximation is close to the

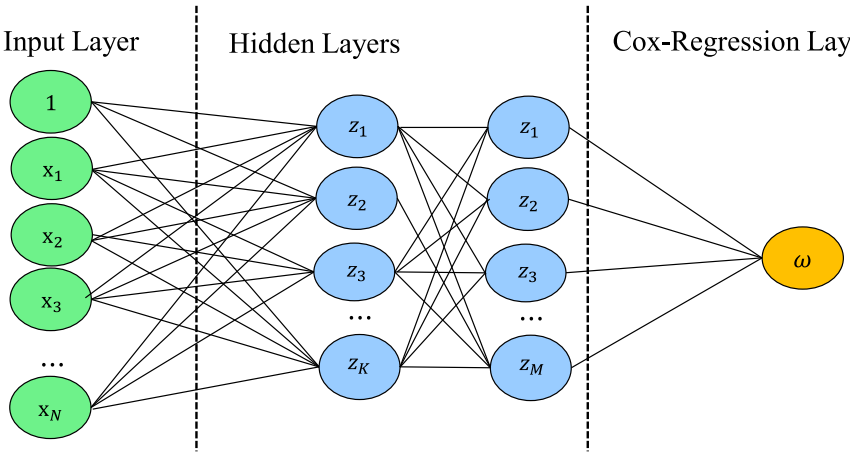

Fig. 1. Neural network-based Cox PH model.

full sum over R_i. The case-control design was originally developed by Goldstein and Langholz (Goldstein & Langholz, 1992). They showed that for the Cox partial likelihood, the sampled risk sets produce consistent parameter estimators equivalent to those using full risk sets.

D. Random survival forest (RSF)

Random forests are an ensemble learning method that performs by constructing a number of decision trees and taking their majority vote for classification and average in case of regression. Random survival forest (RSF), introduced by Ishwaran et al. (Ishwaran, Kogalur, Blackstone, & Lauer, 2008), is an extension of the random forest model to take censored data into account by constructing survival trees. Survival trees stem from the concept of regression trees, in which the observations are split into groups, and then maximize the difference in the response between groups using a metric. The time-to-event data are used as the response (Segal, 1988). The RSF algorithm constructs numerous survival trees from bootstrap samples of the data and utilizes the averaged predictions of each tree to construct an overall prediction of the survival time for each observation. The RSF method has become attractive as a non-parametric method with less restrictive model assumptions.

E. DeepHit

The aforementioned methods are continuous-time models. They need to define a functional form to learn the relationship between the covariates and the survival times. The following two models are known as discrete-time models.

DeepHit is a deep neural network that learns the distribution of survival times directly (Lee et al., 2018). To achieve this, it discretizes the survival times and treats the survival analysis problem as a multiclass classification problem over the discrete-time intervals. DeepHit employs a multi-task network architecture that consists of a single shared sub-network and a family of cause-specific sub-networks. Given the covariates $\boldsymbol{Z}_i$ of patient i, the DeepHit model tries to learn the probability $P(F_i = s, \delta_i = k|\boldsymbol{Z} = \boldsymbol{Z}_i)$,i.e., the probability that a (new) patient with covariates $\boldsymbol{Z}_i$ will experience the event k at time s. To train DeepHit, a loss function $L_{Total} = L_1 + L_2$ that is specifically designed to handle censored data is minimized, where L_1 is the log-likelihood of the joint distribution of the survival time and event of interest, while L_2 incorporates a combination of cause-specific ranking loss functions.

F. Multi-task logistic regression (MTLR)

The Multi-Task Logistic Regression (MTLR) model, developed by Yu et al. (Yu, Greiner, Lin, & Baracos, 2011), is a combination of a series of logistic regression models. Logistic regression can be viewed as modeling survival probability of individuals at a certain time point. By discretizing the time duration to disjoint multiple intervals, MTLR can be treated as a combination of logistic regression models that estimate the probability that the event of interest happens within each interval. MTLR enforces the dependency of the outputs by predicting the event status of an individual at each time snapshot jointly instead of independently.

3. Experiments

3.1. Data description and summary statistics

The data used in this study is obtained from the University of Texas Medical Branch at Galveston (UTMB). For this study, we analyzed 805 patient records with hospital admission times occurring from 03/16/2020 to 11/07/2020 and who were confirmed via PCR testing to have COVID-19. Each record consists of demographical data such as gender, ethnicity, age, vital indices such as pulse, temperature, BMI (body mass index) and ICD-10 diagnosis code. ICD-10 refers to the 10th revision of the International Statistical Classification of Diseases and Related Health Problems, which is a medical classification list provided by the World Health Organization. It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease. In the base ICD-10 classification system, the code set allows for more than 14,000 different codes to track various diagnoses. Considering the small size of our dataset, we filter out the barely used ICD-10 codes and only keep the codes with more than 60 occurrences that the number of patients has been diagnosed. This is necessary since the majority of the ICD-10 are never used. In this way, the dimension of the data can be significantly reduced. After excluding irrelevant ICD-10 codes, a total of 59 features are used for this study. In addition, there is a column to record the deceased status. As described in Section 2, rather than measuring the time to the next discharge, in this dataset, the LOS recorded corresponds to the cumulative amount of time (possibly, over several non-overlapping stays) a patient has spent in the hospital within 30 days of the encounter of interest. Table 1 and Fig. 2 summarize the distribution of LOS based on gender, ethnicity, and age. From Table 1 and Fig. 2 we can see that the distribution of LOS is heavily right-skewed, with 76.28% of LOS values below 7 days and 89.81% of LOS values below 14 days. The median and mean of LOS are 2.15 and 5.27, respectively. Interestingly, there are more male patients in almost all rows of Table 1 than female patients, which is also apparent from Fig. 2(b). The percentage of patients under 18 years old is extremely low with only 0.99% (8/805). Adults over 60 years of age represent 42.98% of hospitalizations. It indicates that older adults have a higher risk. Obviously, we can conclude that the risk for severe illness with COVID-19 increases with age. Another phenomenon we can observe is no significant difference between *Hispanic or Latino* and *not Hispanic or Latino* patients.

Table 1
Distribution of LOS.

LOS	Records (No./%)	Gender			Ethnicity			Age			
		Male	Female	Unknown	HL[a]	NHL[a]	Unknown	[0,18]	[18,60]	60+	Unknown
1	346/42.98%	158	188	0	153	191	2	8	235	103	0
2	47/5.84%	26	21	0	17	28	2	0	33	14	0
3	56/6.96%	38	17	1	21	30	5	0	27	28	1
4	56/6.96%	40	16	0	22	34	0	0	22	34	0
5	31/3.85%	23	8	0	17	13	1	0	17	14	0
6	52/6.46%	31	20	1	13	38	1	0	18	33	1
7	26/3.23%	19	7	0	12	13	1	0	12	14	0
8	32/3.98%	23	9	0	12	17	3	0	16	16	0
9	16/1.99%	13	3	0	5	10	1	0	7	9	0
10	7/0.87%	4	3	0	3	4	0	0	2	5	0
11	13/1.61%	9	4	0	2	11	0	0	6	7	0
12	25/3.11%	17	8	0	3	22	0	0	12	13	0
13	9/1.12%	7	2	0	4	5	0	0	1	8	0
14	7/0.87%	5	2	0	3	4	0	0	3	4	0
14+	82/10.19%	62	20	0	30	50	2	0	38	44	0
Total	805/100%	475	328	2	317	470	18	8	449	346	2

[a]H.L.: Hispanic or Latino; NHL: Not Hispanic or Latino.

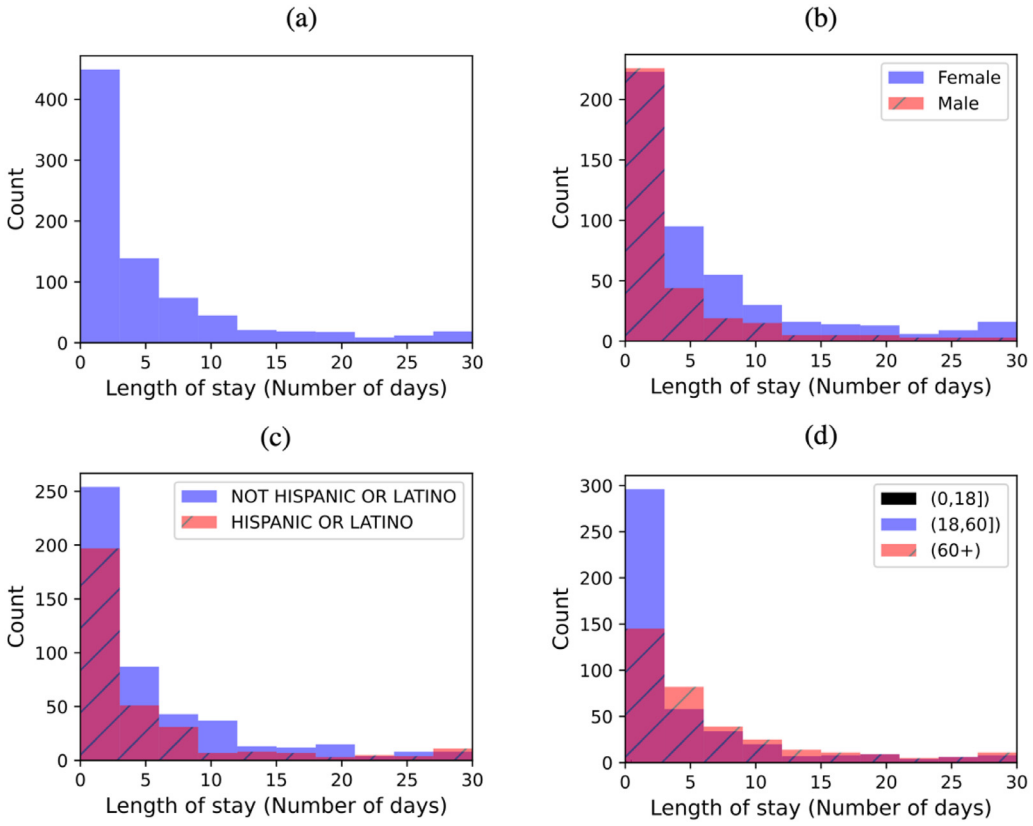

Fig. 2. Length of stay distribution (a) pooled; (b) grouped by gender; (c) grouped by ethnicity; (d) grouped by age.

3.2. Data preprocessing

It is quite common that clinic datasets have missing values which occur when the value of a variable of interest is not measured or recorded in the sample. Data can be missing for several reasons, including (i) loss of a patient to follow-up; (ii) patient refuses to respond to some questions; (iii) investigator or mechanical error; and (iv) physicians not ordering certain investigations for some patients (Austin, White, Lee, & van Buuren, 2020). In our dataset, missing values appear in the fields of age, temperature, vitals, etc. It was found that incomplete data lead to adverse effects on the outcome as it decreases the learning rate and accuracy of prediction and increases the variability in the evaluation metrics (Holmes & Bilker, 2002). It was also found that the effect relates to the percentage of the missing data as well as the type of the missing data. In our dataset, the majority of missing entries are from vitals. Among 805 records, the number of missing values for BMI, diastolic blood pressure, systolic blood pressure, pulse, pulse oximetry, respiration, temperature is 78, 3, 3, 1, 3, 6, 2, respectively. Considering that we only have a limited number of records, this is nonnegligible. The commonly used approach to address the presence of missing data is complete-case analysis, where subjects with missing data are excluded. Another method is called mean-value imputation, where missing values are replaced with the mean value of that variable in those subjects for whom it is not missing. However, in many settings, these approaches can lead to biased estimates (e.g., of regression coefficients) and/or confidence intervals that are artificially narrow (Austin et al., 2020).

To address this issue, in this paper, for each patient, we first extract the patient's vitals from the past two weeks and use Tukey's five-number summary for the pooled sample as the current vital information. Five-number summary refers to "minimum value", "lower hinge (or first quartile)", "median", "upper hinge (or third quartile)" and "maximum value". For example, instead of using one column for temperature, there are five columns used. Based on consultation with medical experts, this approach was deemed reasonable. In our paper, the missing values are handled by multivariate imputations by chained equations (MICE) based on random forests (Van Buuren & Groothuis-Oudshoorn, 2011).

After filling in missing entries, the data is preprocessed in further steps. For the categorical features (e.g., gender, ethnicity), one-hot encoding techniques are used. For continuous features, Min–Max scaling is used to normalize the data to reduce variation. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min–Max scaling, which is equalized as

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{8}$$

where x represents the original value. x_{max} and x_{min} are the maximum and the minimum values of the sampled feature values, respectively.

3.3. Evaluation metrics

Metrics for evaluating the prediction performance are summarized in this section. For survival models, various metrics are available to handle the censored values. Concordance index (C-index) and Brier score are the two most commonly used metrics for this purpose.

(1) C-index

The most commonly applied discriminative evaluation metric to evaluate the predictive ability of a survival model is C-index (Harrell, Califf, Pryor, Lee, & Rosati, 1982). The C-index is concerned with the order of the predictions, not the predictions themselves. The idea behind the C-index is that for a random pair of individuals in a dataset, the predicted event times of the two individuals have the same ordering as their true event times. A concordance index of 1 represents a model with perfect prediction, an index of 0.5 is equal to random prediction. The C-index shows the model's ability to correctly provide a reliable ranking of the survival times based on the individual risk scores.

(2) Brier Score

The Brier score rule is affected by both discrimination and calibration, which is defined by

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left(1_{T_i>t} - \hat{S}(t,x)\right)^2, \tag{9}$$

where N is the number of events under consideration. Brier scores calculated with the above formula lie between 0 and 1: A Brier score of 0 reflects perfect accuracy (i.e., there is no difference between event scores), and a Brier score of 1 reflects perfect inaccuracy. If the dataset contains right-censored data, it is necessary to adjust the core by weighting the squared distance using the inverse probability of censoring weights method (Graf, Schmoor, Sauerbrei, & Schumacher, 1999).

$$BS = \frac{1}{N}\sum_{t=1}^{N}\left(\frac{\left(0-\hat{S}(t,x)\right)^2 \cdot 1_{T_i\le t,\delta_i=1}}{\hat{G}(T_i^-)} + \frac{\left(1-\hat{S}(t,x)\right)^2 \cdot 1_{T_i>t}}{\hat{G}(t)}\right) \tag{10}$$

where $\hat{G}(t)$ is the estimator of the conditional survival function using the Kaplan–Meier method.

In addition, three frequently used performance metrics are applied to measure the prediction performance, i.e., mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). These metrics are defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|, \tag{11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{12}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left|\hat{y}_i - y_i\right|}{y_i}\cdot 100\right). \tag{13}$$

In the above equations, y_i and $\hat{y}_i$ represent observed LOS and predicted LOS, respectively. N is the total number of samples. All of these metrics are negatively oriented scores, which means a lower value indicates a better model. Both MAE and RMSE measure the average RUL prediction error of the model. The value of these two metrics could range between 0 and ∞. RMSE is the square root of the mean squared difference between true LOS and predicted LOS. RMSE is always nonnegative and zero RMSE means a perfect fit to the data, which is nearly impossible in practice and can also be a sign of overparameterization. MAPE is a variant of MAE, which is the absolute error normalized over the data. MAPE is widely used because it can be easily interpreted. For example, a MAPE of 30% means the model approximates the target value with an average accuracy of 70% (by subtracting from 100%).

3.4. Results and discussion

This section evaluates the performance of six survival models, which are considered for LOS prediction in this paper. All computations are conducted on a 2.3 GHz Quad-Core Intel Core i7 processor, and the code is written in Python. The percentage for the training and test sets are 80% and 20%, respectively. To train the survival models, the adaptive moment estimation (Adam) is chosen for optimization (Kingma & Ba, 2015). Adam is a stochastic first-order optimization algorithm that uses historical information about stochastic gradients and incorporates it in an attempt to estimate the second-order moment of stochastic gradients adaptively. This method requires little memory, which makes it computationally efficient. Thus, Adam has been proved well-suited for problems where data volume and number of parameters are comparatively larger. To determine the optimal network architecture for each survival model, i.e., the number of hidden layers and number of hidden nodes, five-fold cross-validation was used on the training data set (i.e., 20% of training records are used for validation). The model with the lowest validation error is eventually selected. The best configuration is listed in Table 2. To evaluate the prediction performance, we first randomly select two samples from test data and show the predicted LOS curve as depicted in Fig. 3. The vertical solid blue lines in Fig. 3(a) and (b) are the true LOS for each sample, which are 7.05 days and 12.41 days, respectively. As expected, for all models, the $S(t)$ monotonically decreases as the days of follow-up increase. It is observed that the predicted LOS curves are close to each other and to the true value in Fig. 3(a), whereas in Fig. 3(b), the predicted curve for each model varies. The center of the Cox-CC is closest to the true value, which indicates the best accuracy. To evaluate the performance quantitatively, the expected LOS is calculated using Eq. (3) for Cox-linear, DeepSurv, Cox-CC, DeepHit, MTLR, and RSF, respectively. They are 7.68, 5.51, 6.15, 14.15, 6.09, 8.30, 6.64 for (a) and 11.47, 19.85, 13.06, 27.18, 6.48, 14.58 for (b), respectively. Then we calculate the absolute error (i.e., $|predicted LOS - true LOS|$) for each model. The absolute errors yield 0.63, 1.54, 0.9, 7.1, 0.96, 0.41 for (a) and 0.94, 7.44, 0.65, 14.77, 5.93, 2.17 for (b). As we can see, Cox-linear and Cox-CC constantly show good predictive capability for both (a) and (b). On the other hand, DeepSurv and DeepHit significantly deviate from the true LOS in (b). From Fig. 3(b) we can see that the predicted LOS curves of both models have longer tails than others, so the predicted values are substantially larger than the true LOS.

Table 3 displays C-index and Brier scores for each model. In all performance comparison tables, the best performance values are marked bold. From Table 3 we observe that RSF shows the best C-index performance, and Cox-linear has the best Brier score. DeepHit, and MTLR show poor performance in terms of both C-index and Brier scores. As

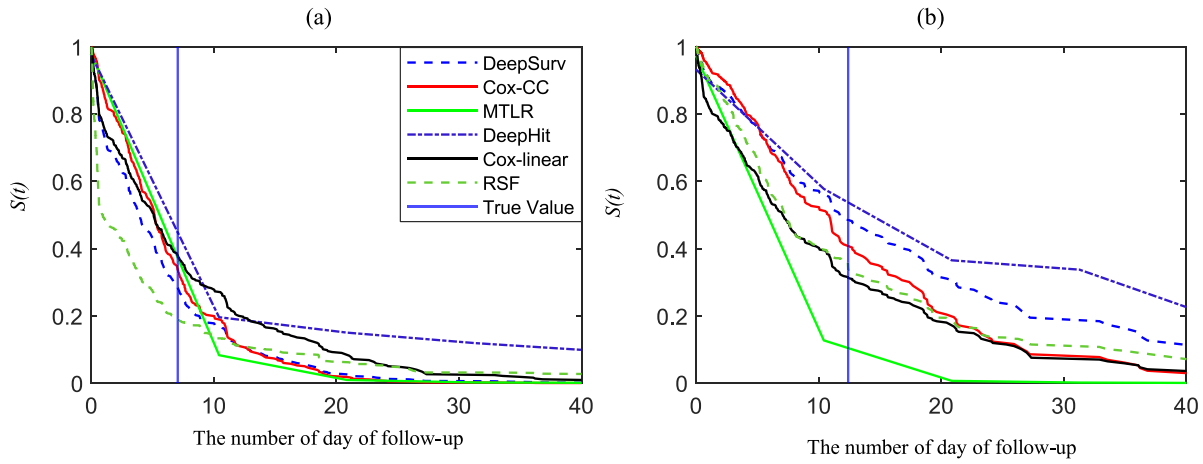

Fig. 3. Predicted LOS curve for two randomly selected samples.

Table 2
Parameter settings.

Models	Hidden layers	Batch size	Dropout
Cox-linear	–	–	–
DeepSurv	[64]	128	0.2
Cox-CC	[96]	128	0.1
DeepHit	[64,64]	128	0.1
MTLR	[128,128]	128	0.1
RSF	–	–	–

Table 3
Comparison of C-index and brier score.

Models	C-index	Brier score
Cox-linear	0.7335	**0.0741**
DeepSurv	0.7332	0.0755
Cox-CC	0.7066	0.0815
DeepHit	0.5540	0.1503
MTLR	0.4745	0.1021
RSF	**0.7388**	0.0793

we discussed earlier, these two models discretize the survival times and treat the survival analysis problem as a multiclass classification problem over the discrete-time intervals. However, the dataset is highly skewed, the majority of the LOS fall into less than one-day category and, thus, negatively affect the capability of these models.

To evaluate the overall prediction performance, we calculate the MAE, RMSE and MAPE in the 20% test data for each model. We also include the popular support vector regression (SVR) for comparison purposes. A support vector machine (SVM) constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional Hilbert space, which can be used for classification, regression, or other tasks like outlier detection. Because of their relative simplicity and flexibility, SVMs have become exceedingly popular in recent years. SVR uses the same principles as the SVM, with only a few minor differences to adapt for regression problems. We consider three different kernels in SVR, namely, linear, radial basis function (RBF) and polynomial. It should be noted that to run the SVR model, 60 censored samples had to be removed from the data. Table 4 shows the comparison results. It is observed that SVR with RBF kernel has the best MAE value. Cox-linear and SVR with linear kernel have the best RMSE and MAPE, respectively. It is somewhat surprising that deep learning models, including DeepSurv and DeepHit, perform worse than linear models. In this study, we only have 805 patient records with 59 selected features. For some patients, some features have missing entries, which makes the problem of data scarcity even more severe. Among 805 patient records, 644 records (80% of data) are used for training. Since the performance of deep learning models relies heavily on the size of training data, the former can be severely compromised when the number of training samples is small. We also observe that all discrete-time survival models, i.e., DeepHit, MTLR and RSF exhibit poor performance in terms of MAE, RMSE and MAPE. From Table 4 we can also see that the MAPE of discrete-time survival models is much bigger than their MAE and RMSE. Because of the skewness of the distribution of LOS, the MAPE increases quickly if a long LOS is erroneously predicted to be short or vice versa. In other words, a single mistake of the model may dominate the MAPE. It is surprising that the good C-index of RSF leads to a worse MAE, RMSE and MAPE. This method requires a larger number of samples to grow an accurate survival tree. With limited data, the LOS information learned from data through RSF is confined.

Table 4
Prediction performance comparison.

Models	MAE	RMSE	MAPE
Cox-linear	3.4387	**5.4834**	5.5368
DeepSurv	3.4468	5.6120	4.5242
Cox-CC	3.6175	5.8959	5.0749
DeepHit	8.2114	9.1956	27.1449
MTLR	4.4384	6.5393	9.9455
RSF	4.4477	5.9019	11.7549
SVR (linear kernel)	3.4059	6.0313	**4.1160**
SVR (rbf kernel)	**3.3073**	5.9660	4.6721
SVR (polynomial kernel)	3.8371	6.2235	6.6887

4. Conclusion

Predicting cumulative hospital LOS over a given time horizon allows hospitals to assess the overall patient load, which in turn allows improved scheduling of patient admissions leading to reduced variation of bed occupancies in hospitals. In this study, we investigated six different survival models for LOS prediction. Patient-specific LOS distributions can be learned by using survival models. Furthermore, we can make full use of censored clinical data. The results of our case study show that the selected survival models exhibited a variation in prediction performance depending on the model evaluation criteria. Continuous-time survival methods show good predictive capability compared with discrete-time models. In this research, we only consider patients' vital signs in the form of numeric data. However, features from the text and images could be very useful in improving the accuracy of LOS prediction, which could be an interest of the future research direction.

CRediT authorship contribution statement

Yuxin Wen: Wrote the draft manuscript. **Md Fashiar Rahman:** Conducted the study and analyzed the data. **Yan Zhuang:** Conducted

the study and analyzed the data. **Michael Pokojovy:** Supervised the study and revised the manuscript. **Honglun Xu:** Contributed to data collection, Digitalization and/or clinical interpretation/support. **Peter McCaffrey:** Supervised the study and revised the manuscript. **Alexander Vo:** Contributed to data collection, Digitalization and/or clinical interpretation/support. **Eric Walser:** Contributed to data collection, Digitalization and/or clinical interpretation/support. **Scott Moen:** Contributed to data collection, Digitalization and/or clinical interpretation/support. **Tzu-Liang (Bill) Tseng:** Supervised the study and revised the manuscript.

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.

Acknowledgments

This work was partially supported by the National Science Foundation, United States (ECR-PEER-1935454), (ERC-ASPIRE-1941524), and Department of Education (Award #P120A180101), (Award #P116S210004). We also acknowledge to the medical doctors from the University of Texas Medical Branch for their continuous support, information, and providing the proprietary dataset used in this research.

References

American Hospital Association, e. a. (2021). AHA hospital statistics. In *Fast facts on US hospitals*. American Hospital Association, Retrieved from https://www.aha.org/.

Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2020). Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology*.

Awad, A., Bader-El-Den, M., & McNicholas, J. (2017). Patient length of stay and mortality prediction: a survey. *Health Services Management Research*, *30*(2), 105–120.

Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., & Levin, S. (2016). Real-time prediction of inpatient length of stay for discharge prioritization. *Journal of the American Medical Informatics Association*, *23*(e1), e2–e10.

Breslow, N. E. (1972). Contribution to discussion of paper by DR Cox. *Journal of the Royal Statistical Society. Series B.*, *34*, 216–217.

Caetano, N., Laureano, R. M., & Cortez, P. (2014). A data-driven approach to predict hospital length of stay. In *Paper presented at the proceedings of the 16th international conference on enterprise information systems. Vol. 1*.

Churpek, M. M., Wendlandt, B., Zadravecz, F. J., Adhikari, R., Winslow, C., & Edelson, D. P. (2016). Association between intensive care unit transfer delay and hospital mortality: a multicenter investigation. *Journal of Hospital Medicine*, *11*(11), 757–762.

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

Dogu, E., Albayrak, Y. E., & Tuncay, E. (2021). Length of hospital stay prediction with an integrated approach of statistical-based fuzzy cognitive maps and artificial neural networks. *Medical & Biological Engineering & Computing*, *59*(3), 483–496.

Goldstein, L., & Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the cox regression model. *The Annals of Statistics*, 1903–1928.

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

Harerimana, G., Kim, J. W., & Jang, B. (2021). A deep attention model to forecast the length of stay and the in-hospital mortality right on admission from ICD codes and demographic data. *Journal of Biomedical Informatics*, *118*, Article 103778.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

Hildebrand, F. B. (1987). *Introduction to numerical analysis*. Courier Corporation.

Holmes, J. H., & Bilker, W. B. (2002). The effect of missing data on learning classifier system learning rate and classification performance. In *Paper presented at the international workshop on learning classifier systems*.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860.

Kartsonaki, C. (2016). Survival analysis. *Diagnostic Histopathology*, *22*(7), 263–270.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*(1), 1–12.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Kulkarni, H., Thangam, M., & Amin, A. P. (2021). Artificial neural network-based prediction of prolonged length of stay and need for post-acute care in acute coronary syndrome patients undergoing percutaneous coronary intervention. *European Journal of Clinical Investigation*, *51*(3), Article e13406.

Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, *20*(129), 1–30.

Lee, C., Zame, W., Yoon, J., & van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Paper presented at the proceedings of the AAAI conference on artificial intelligence*.

Luo, L., Lian, S., Feng, C., Huang, D., & Zhang, W. (2017). Data mining-based detection of rapid growth in length of stay on COPD patients. In *Paper presented at the 2017 IEEE 2nd international conference on big data analysis*.

Luth, E. A., Russell, D. J., Xu, J. C., Lauder, B., Ryvicker, M. B., Dignam, R. R., et al. (2021). Survival in hospice patients with dementia: the effect of home hospice and nurse visits. *Journal of the American Geriatrics Society*.

Ma, F., Yu, L., Ye, L., Yao, D. D., & Zhuang, W. (2020). Length-of-stay prediction for pediatric patients with respiratory diseases using decision tree methods. *IEEE Journal of Biomedical and Health Informatics*, *24*(9), 2651–2662.

Muhlestein, W. E., Akagi, D. S., Davies, J. M., & Chambless, L. B. (2019). Predicting inpatient length of stay after brain tumor surgery: Developing machine learning ensembles to improve predictive performance. *Neurosurgery*, *85*(3), 384–393.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*(1), 1–10.

Resar, R., Nolan, K., Kaczynski, D., & Jensen, K. (2011). Using real-time demand capacity management to improve hospitalwide patient flow. *The Joint Commission Journal on Quality and Patient Safety*, *37*(5), 217-AP213.

Rowan, M., Ryan, T., Hegarty, F., & O'Hare, N. (2007). The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. *Artificial Intelligence in Medicine*, *40*(3), 211–221.

Schober, P., & Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *Anesthesia and Analgesia*, *127*(3), 792.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, 35–47.

Tanuja, S., Acharya, D. U., & Shailesh, K. (2011). Comparison of different data mining techniques to predict hospital length of stay. *Journal of Pharmaceutical and Biomedical Sciences*, *7*(7).

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(1), 1–67.

Vekaria, B., Overton, C., Wiśniowski, A., Ahmad, S., Aparicio-Castro, A., Curran-Sebastian, J., et al. (2021). Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning. *BMC Infectious Diseases*, *21*(1), 1–15.

Verburg, I. W., de Keizer, N. F., de Jonge, E., & Peek, N. (2014). Comparison of regression methods for modeling intensive care length of stay. *PLoS One*, *9*(10), Article e109684.

Wang, Z., Ji, J. S., Liu, Y., Liu, R., Zha, Y., Chang, X., et al. (2020). Survival analysis of hospital length of stay of novel coronavirus (COVID-19) pneumonia patients in Sichuan, China. medRxiv.

Wen, Y., Guo, X., Son, J., & Wu, J. (2022). A neural network based proportional hazard model for IoT signal fusion and failure prediction. *IISE Transactions*, 1–15.

Yu, C.-N., Greiner, R., Lin, H.-C., & Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*, *24*, 1845–1853.

Zebin, T., & Chaussalet, T. J. (2019). Design and implementation of a deep recurrent model for prediction of readmission in urgent care using electronic health records. In *Paper presented at the 2019 IEEE conference on computational intelligence in bioinformatics and computational biology*. 9-11 2019.