Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.022647

ARTICLE

A Two-Step Algorithm to Estimate Variable Importance for Multi-State Data: An Application to COVID-19

Behnaz Alafchi[1], Leili Tapak[1,*], Hassan Doosti[2], Christophe Chesneau[3] and Ghodratollah Roshanaei[1]

[1]Department of Biostatistics, School of Public Health, Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran

[2]School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia

[3]Department of Mathematics, LMNO, University of Caen-Normandie, Caen, France

*Corresponding Author: Leili Tapak. Email: l.tapak@umsha.ac.ir

Received: 18 March 2022 Accepted: 07 July 2022

ABSTRACT

Survival data with a multi-state structure are frequently observed in follow-up studies. An analytic approach based on a multi-state model (MSM) should be used in longitudinal health studies in which a patient experiences a sequence of clinical progression events. One main objective in the MSM framework is variable selection, where attempts are made to identify the risk factors associated with the transition hazard rates or probabilities of disease progression. The usual variable selection methods, including stepwise and penalized methods, do not provide information about the importance of variables. In this context, we present a two-step algorithm to evaluate the importance of variables for multi-state data. Three different machine learning approaches (random forest, gradient boosting, and neural network) as the most widely used methods are considered to estimate the variable importance in order to identify the factors affecting disease progression and rank these factors according to their importance. The performance of our proposed methods is validated by simulation and applied to the COVID-19 data set. The results revealed that the proposed two-stage method has promising performance for estimating variable importance.

KEYWORDS

Multi-state data; deviance residual; martingale residual; gradient boosting; random forest; neural network; variable importance; variable selection

1 Introduction

In longitudinal health studies, a patient may experience a sequence of clinical progression events. For instance, the local recurrence may be followed by a distant recurrence and then death [1]. Also, in the renal transplantation process of patients with end-stage kidney disease, renal allograft failure is considered an intermediate event that can affect the whole survival of the patient [2]. For another instance, the state of the COVID-19 patients who are admitted to the intensive care unit (ICU) at the time of admission can influence their survival time. A patient may need invasive or non-invasive ICU ventilation and extracorporeal membrane oxygenation and may be discharged from the hospital or die eventually [3]. Moreover, the need for ICU before death or discharge from the hospital among COVID-19 patients who are admitted to the ward could be another state. In such studies, the progression of the disease is considered as a multi-state process, and multi-state models (MSM) are used to assess the impact of different covariates on the transition between different health states [1,4]. Ferrer et al. [1] developed a multi-state model for the analysis of prostate cancer. Beesley et al. [4] also applied the multi-state approach for modeling the progression of prostate cancer.

One main objective in the MSM framework is variable selection, where attempts are made to identify the risk factors associated with the transition hazard rates or probabilities of the disease progression. The classic approach is to use stepwise methods, where steps are taken by sequentially adding (forward) or removing (backward) variables at each step. However, the stepwise method suffers from some drawbacks, such as instability of the selected variables. In the context of MSMs, few attempts have been made at the variable selection. For example, Dang et al. [5] suggested L1-Regularized multi-state models using a Least Absolute Shrinkage and Selection Operator (LASSO) penalty for parameter estimation and variable selection, simultaneously. The LASSO method has several advantages for variable selection, including sparseness avoiding overfitting and a lower burden of computations. It is also applicable in the presence of a large number of covariates. Nevertheless, the selected model by the LASSO is unstable in terms of selected features. Moreover, for a set of highly correlated features, the LASSO selects one of them randomly. Penalized methods neglect the interaction between variables and their complex relationships or their unknown functional form. Most importantly, these techniques were unable to provide a quantitative assessment of their significance [6]. While some of the features measured in a study may be associated with the sojourn times in the entry states of each transition, they may have different degrees of importance, and some of them may be more important for some transitions than others [7]. The variable importance can be estimated as a real-valued parameter using an optimal estimating function [8].

In the context of COVID-19, several studies have investigated factors associated with various outcomes using machine learning models through variable importance. For example, Stachel et al. utilized machine learning methods for variables associated with mortality among COVID-19 patients [9]. Also, Snider et al. investigated variables affecting mortality in COVID-19 patients using artificial intelligence methods [10]. However, most of these studies considered binary outcomes such as survival or death. In various studies, random survival forest has been used to predict mortality in hospitalized COVID-19 patients. They considered time to hospital discharge and mortality as the outcomes of interest and competing risks, respectively. More complex structures were also considered by some authors. Hazard et al. [11] considered a multistate structure for joint analysis of the duration of ventilation, length of intensive care, and mortality of COVID-19 patients. To our knowledge, no study has dealt with estimating variable importance in multi-state data structures such as COVID-19 data.

Breiman [12] introduced the random forest technique for classification and regression problems, where it has been widely used for variable selection. Gradient boosting [13] is also a commonly used method for variable selection, showing satisfactory performance in many problems. Negassa et al. [14] investigated model selection in tree-structured subgroup analysis using the RECursive Partition and Amalgamation algorithm. They found that there is no single model selection criterion with uniformly superior performance, and they proposed a two-stage approach for model selection with promising performance in variable selection. Recently, Duan et al. [6] proposed a machine learning-based approach to estimate variable importance using martingale and deviance residuals and their standardized counterparts which resulted in promising performances based on simulation studies. In the present study, we proposed a two-step algorithm to estimate the importance of variables for multi-state data based on three different machine learning approaches: random forest, gradient boosting, and neural network as the most widely used method. Our main contribution to this study is to extend the proposed approach to multi-state data to estimate the variable importance. The remaining sections of this paper are organized as follows: In Section 2, we describe the multi-state model and its related residuals. In Section 3, we define three risk indices based on the residuals mentioned in Section 2. In Section 4, we describe our proposed two-step algorithm to evaluate the variable's importance for multi-state data and introduce the loss function criteria to compare the performance of different methods. In Section 5, the performance of our algorithm is evaluated via simulation studies, and we apply our method to the COVID-19 data set through an illness-death multi-state model, and a brief discussion is finally given in Section 5.

2 Methods

2.1 Multi-State Model

A multi-state model is a stochastic process ($\{X(t), t \in \varsigma\}$) with a finite state space $S = \{0, 1, \ldots, p\}$, where $\varsigma = [0, \tau]$ for $\tau \leq +\infty$. The multi-state process is assumed to be continuous-time with the right-continuous sample paths, $X(t^+) = X(t)$. The transition probabilities are given as follows:

$$P_{hk}(s,t) = P(X(t) = k|X(s) = h, \chi_{s^-}),$$

for h, k $\in S$, $s, t \in \varsigma, s \leq t$, and χ_s be a σ-algebra that stands for everything that happened up to time s. Then, the transition intensities are defined as follows:

$$\lambda_{hk}(t) = \lim_{\Delta t \to 0} \frac{P_{hk}(t, t+\Delta t|\chi_{s^-})}{\Delta t}.$$

Suppose, there are n subjects and let $T_i = (T_{i1}, \ldots, T_{im_i})$ represents the vector of the $m_i \geq 1$ observed transition times for the ith subject, with $T_{ir} < T_{i(r+1)}$, $r \in \{1, \ldots, m_i\}$. When the last observed state is an absorbing state (i.e., a state that once entered cannot be never left; e.g., death), then the ith subject will experience m_i direct transitions. Otherwise, T_{im_i} is equal to right censoring time C_i and there are $m_i - 1$ direct transitions.

The counting process for the subject i ($i = 1, \ldots, n$) can be defined as (a multivariate counting process) by $\{N_{hki}(t), h, k \in S, h \neq k, t \leq C_i\}$ counting the number of direct transitions from the state h to the state k for the subject i over the interval $[0, t]$ and $C_i (\leq \tau)$. The intensity function for the subject i based on the Cox regression model, by using counting process formulation of the model, is given as follows:

$$\begin{aligned}\lambda_{hki}(t) &= \lim_{\Delta t \to 0} \frac{P\{N_{hki}(t, t+\Delta t) = 1|\chi_{s^-}\}}{\Delta t} \\ &= Y_{hi}(t)\,\lambda_{hk0}(t)\exp\left\{{\beta_{hk}}^T Z_{hki}(t)\right\}.\end{aligned} \tag{1}$$

Here, $Y_{hi}(t) = I(X_i(t^-) = h)$ is an indicator specifying whether $X_i(.)$ is in the state h at time t^- (some h and k combinations may not be possible). Specifically, $Y_{hi}(t) = 1$ means that the individual i is at risk for transition from h to k immediately before time t. The $N_{hki}(t)$ only jumps when $Y_{hi}(t) = 1$. Also, in the above equation $\lambda_{hk0}(.)$ is the baseline intensity function, and $Z_{hki}(t)$ is the vector of predictors associated with the vector of coefficients β_{hk} for the transition from h to k.

Note that, under the Markov assumption, future evolution of the process only depends on the current state and, therefore, χ_{s^-} has been omitted from the formulas.

2.2 Risk Indices

2.2.1 Martingale Residuals

Using the Doob–Meyer decomposition [15], in multi-state models, the martingale residual for the subject i on the time interval of $[0, t]$ is defined as follows:

$$\begin{aligned} M_{hki}(t) &= N_{hki}(t) - \Lambda_{hki}(t) \\ &= N_{hki}(t) - \int_0^t Y_{hi}(u) \exp\{\beta_{hk}{}^T Z_{hki}(u)\} d\Lambda_{hk0}(u). \end{aligned}$$

In the above equation, $\Lambda_{hki}(t)$ represents the cumulative intensity function and is a predictable process called the compensator of $N_{hki}(t)$. Also, $\Lambda_{hk0}(t) = \int_0^t \lambda_{hk0}(u)du$ indicates the cumulative baseline intensity function. The β_{hk} and $\Lambda_{hk0}(t)$ are unknown parameters that should be estimated. Let $\widehat{\beta}_{hk}$ and $\widehat{\Lambda}_{hk0}\,(t)$ represent the maximum likelihood estimator of β_{hk} [16,17] and the Breslow estimator of $\Lambda_{hk0}(t)$ [17,18], respectively. Then, the martingale residual for the transition $h \rightarrow k$ is defined as follows:

$$\widehat{M}_{hki}\,(t) = N_{hki}\,(t) - \int_0^t Y_{hi}\,(u) \exp\left\{\widehat{\beta}_{hk}^T Z_{hki}\,(u)\right\} d\widehat{\Lambda}_{hk0}\,(u)\,.$$

In the multi-state data, suppose that the last observed time of transition from the state h to the state k (h → k) for the subject i is called τ_{ki}, (the minimum of the administrative censoring time C_a, the individual censoring time (loss to follow-up) C_i and the time of transition to the state k, $\forall k \in \{2, \ldots, m_i\}$). Therefore, the martingale residual for the subject i at time τ_{ki} has the following form:

$$\widehat{M}_{hki}\,(\tau_{ki}) = N_{hki}\,(\tau_{ki}) - \exp\left\{\widehat{\beta}_{hk}^T Z_{hki}\right\} \widehat{\Lambda}_{hk0}\,(\tau_{ki})\,. \qquad (2)$$

Note that, we use $\widehat{M}_{hki}$ to represent $\widehat{M}_{hki}\,(\tau_{ki})$.

2.2.2 Deviance

The deviance defined in Therneau et al. [19] is as follows:

D = 2{log likelihood (saturated model)-log likelihood ($\widehat{\beta}$)}.

The saturated model is a model with a parameter for every observation. Here, we extend the definition of the deviance for univariate survival data to the multi-state data. Under mild regularity conditions, the log-likelihood contribution for each individual can be obtained using counting process theory as follows:

$$L_i = \sum_{r=1}^{m_i} \left[\int_0^{T_{ir}} \log(\lambda_{X_i(T_{ir}),X_i(T_{ir}),i}(t)) dN_{hki}(t) - \int_{T_{i(r-1)}}^{T_{ir}} \lambda_{X_i(T_{i(r-1)}),X_i(T_{ir}),i}(t)\, Y_{hi}\,(t)\, dt \right].$$

Note that $\lambda_{hhi}(t) = -\Sigma_k \lambda_{hki}(t)$ for $k \neq h$, and δ_{ir} is a transition indicator (for the subject i) which equals one if a direct transition is observed at time T_{ir} and zero otherwise.

Then, the definition of deviance for multi-state data with the intensity function in Eq. (1), which is an extension of the deviance provided by Therneau et al. [19] for survival data, is as follows:

$$\begin{aligned} D_{hk} = 2\sup_{\gamma_{hk}} \sum_{i=1}^{n} \Bigg[\int Y_{hi}(t) &\left\{ \gamma_{hki}{}^T Z_i - \widehat{\beta}_{hk}^T Z_i \right\} dN_{hki}\,(t) \\ - \int Y_{hi}\,(t) &\left\{ \exp\left(\gamma_{hki}{}^T Z_i\right) - \exp\left(\widehat{\beta}_{hk}^T Z_i\right) \right\} d\Lambda_{hk0}\,(t) \Bigg]. \end{aligned} \qquad (3)$$

Let γ_{hki} be the per-subject estimates of β_{hk}. By taking the first derivation of the above summation part with respect to γ_{hki} and setting it equal to 0, we have:

$$\int Y_{hi}(t)\exp(\gamma_{hki}{}^{T}Z_i)d\Lambda_{hk0}(t) = \int dN_{hki}(t). \tag{4}$$

By subtracting Eqs. (3) and (4), D_{hki} has the following form:

$$D_{hki} = -2\left[M_{hki} + N_{hki}\,(t)\log\left(\frac{N_{hki}(t) - M_{hki}}{N_{hki}(t)}\right)\right].$$

As mentioned before, we supposed τ_{ki} is the observed time of transition to the state k (for the transition from state h to state k). Therefore, the deviance at time τ_{ki} has the following form:

$$D_{hki} = -2\left[\widehat{M}_{hki} + N_{hki}\,(\tau_{ki})\log\left(\frac{N_{hki}\,(\tau_{ki}) - \widehat{M}_{hki}}{N_{hki}\,(\tau_{ki})}\right)\right]. \tag{5}$$

Note that $N_{hki}(\tau_{ki}) \neq 0$, and $D_{hki} = -2\widehat{M}_{hki}$ when $N_{hki}(\tau_{ki}) = 0$.

2.2.3 Deviance Residuals

The deviance residuals are defined as follows:

$$D^{res}_{hki} = sign\left(\widehat{M}_{hki}\right) \times \sqrt{D_{hki}}. \tag{6}$$

The deviance residual D^{res}_{hki} is equal to zero if the martingale residual is zero ($M_{hki} = 0$). The variables are considered to be fixed over time (not time-dependent). In this paper, the martingale residuals, the deviances, and the deviance residuals are called risk indices.

2.3 Variable Importance for Multi-State Data

In this section, we introduce an algorithm to evaluate the variable's importance for multi-state data. Therneau et al. [19] mentioned that using martingale residuals from a null Cox model as the input of classification and regression trees worked well for survival data. Then, Duan et al. [6] extended their proposed algorithm to recurrent event data. Here, we extend their approach to multi-state data.

2.3.1 Two-Step Variable Selection Algorithm

The proposed algorithm has two steps. In the first step, a null multi-state model is fitted as follows:

$$\lambda_{hki}(t) = Y_{hi}(t)\lambda_{hk0}(t).$$

This model does not include any variables, and only the baseline intensity functions should be estimated, which can be estimated using the Breslow estimator [17,18]. Then, the martingale residuals, the deviances, and the deviance residuals are obtained as three risk indices for each subject. For a multi-state model with three states (depicted in Fig. 1) ($m_i = 3$), the martingale residuals are defined as follows:

$$\widehat{M}_{12i}\,(\tau_{2i}) = N_{12i}\,(\tau_{2i}) - \widehat{\Lambda}_{120}\,(\tau_{2i})\,,$$

$$\widehat{M}_{13i}\,(\tau_{3i}) = N_{13i}\,(\tau_{3i}) - \widehat{\Lambda}_{130}\,(\tau_{3i})\,,$$

$$\widehat{M}_{23i}\left(\tau^{*}_{3i}\right) = N_{23i}\left(\tau^{*}_{3i}\right) - \widehat{\Lambda}_{230}\left(\tau^{*}_{3i}\right),$$

where $\tau_{2i} = T_{i2}$ is the observed time of the transition from the state 1 to the state 2, $\tau_{3i} = T_{i3}$ is the observed time of the transition from the state 1 to the state 3, and $\tau^*_{3i} = T_{i3} - T_{i2}$ is the observed time of the transition from the state 2 to the state 3. Then, the deviances are as follows:

$$D_{12i} = -2\left[\widehat{M}_{12i} + N_{12i}(\tau_{2i})\log\left(\frac{N_{12i}(\tau_{2i}) - \widehat{M}_{12i}}{N_{12i}(\tau_{2i})}\right)\right],$$

$$D_{13i} = -2\left[\widehat{M}_{13i} + N_{13i}(\tau_{3i})\log\left(\frac{N_{13i}(\tau_{3i}) - \widehat{M}_{13i}}{N_{13i}(\tau_{3i})}\right)\right],$$

$$D_{23i} = -2\left[\widehat{M}_{23i} + N_{23i}\left(\tau^*_{3i}\right)\log\left(\frac{N_{23i}\left(\tau^*_{3i}\right) - \widehat{M}_{23i}}{N_{23i}(\tau^*_{3i})}\right)\right],$$

and the deviance residuals are defined as follows:

$$D^{res}_{12i} = sign\left(\widehat{M}_{12i}\right) \times \sqrt{D_{12i}},$$

$$D^{res}_{13i} = sign\left(\widehat{M}_{13i}\right) \times \sqrt{D_{13i}},$$

$$D^{res}_{23i} = sign\left(\widehat{M}_{23i}\right) \times \sqrt{D_{23i}}.$$

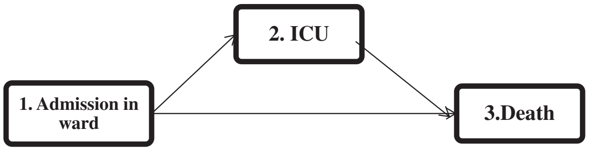

Figure 1: The multi-state representation of the disease progression of COVID-19 hospitalized patients

In the second step, the random forest, the gradient boosting, and the neural network methods are applied to the risk indices generated in the first step to evaluate the importance of each variable in different transitions.

2.4 Machine Learning Models

2.4.1 Random Forest

The random forest method is a supervised ensemble learning algorithm developed by Brieman in 2001. It is constructed by combining several decision tree algorithms to create solutions for complex problems. In this method, a series of simple unpruned regression trees are constructed by using random bootstrapped samples that are obtained from the original data sample. The results of the simple trees are then accumulated to produce a final prediction of the response for the subjects in regression problems, which is the average of the predictions given by all trees. In the random forest algorithm, in the first step, for b = 1 to B, a bootstrapped sample K∗ of size N (the original sample size) is drawn from the training data. Each tree (T_b), in the random forest, is grown for the bootstrapped data by repeating the following recursive steps for each of the terminal nodes in the tree: a) A set of m variables is randomly selected from a set of p original variables; b) the best variable/split-point is selected from the set of m variables chosen in Step (a); and c) the node is split into two daughter nodes. These three steps are repeated until the minimum predefined node size is achieved (n_{min}). In the second step, the output of all trees (B) is aggregated. A prediction for a new observation is obtained by calculating $\frac{1}{B}\sum_1^B T_b(x)$.

2.4.2 Gradient Boosting Machine

Gradient boosting machines (GBMs) consist of a group of powerful machine-learning methods with substantial accomplishments in many practical applications. GBMs can be very adaptable to the needs of the application. The boosting works based on sequentially adding new models to the ensemble, so that in each particular iteration, "a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far" [20]. A GBM has been established to connect boosting with a statistical framework [21–23], which provides the required justification for the hyperparameters in the model and the methodological foundation for developing further gradient boosting models [20].

In GBMs, successive new models are fitted based on the chosen learning technique; so that more accurate estimates of the outcome are produced. The main idea underlying this algorithm is to create "the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble" [20]. The loss functions used can be researcher-defined or standard loss functions derived through trial and error in the past.

2.4.3 Neural Network

Artificial neural networks (ANNs) are machine learning methods that consist of an input layer of neurons (or nodes, units), at least one hidden layer of neurons, and a layer of output neurons. Connections between layers are illustrated by numeric values (e.g., weights) through activation functions. The most widely used form of ANN is the multi-layer perceptron, where the data flows in a forward direction from the input layer to the output layer. So, the neurons are trained with the back propagation learning algorithm. In this study, the regression weights were estimated using a tangent hyperbolic activation function and an identity activation function, which provided better results than other settings [24].

2.5 Performance Measurement

Here, the loss function (introduced by Cox [17,18]) was considered to assess the performance of the different variable importance evaluation methods for multi-state data. Suppose that there are q standardized covariates $x_{hk1}, \ldots, x_{hkq}$ for the transition from state h to state k that are associated with the coefficients $\eta_{hk1}, \ldots, \eta_{hkq}$, respectively. The absolute values of the coefficients are related to their importance, such that, for the transition $h \to k$, if the ranks of the absolute value of the coefficients are 1, 2, . . . , q ($|\eta_{hk1}|$ has the maximum value), the rank of their corresponding variable importance values should be 1, 2, . . . , q (from the most to the least important). However, the estimated rank may not be absolutely correct. In such a situation, if $b_{hk1}, \ldots, b_{hkq}$ represent the estimated rank of the covariates for the transition $h \to k$, at least one pair of parameters satisfies the $b_{hkj} > b_{hki}$ for $j < i$. The loss function of the ranking results for the transition $h \to k$ is defined as follows:

$$Loss\left(b_{hk1}, \ldots, b_{hkq}; \eta_{hk1}, \ldots, \eta_{hkq}\right) = \sum_{i=1}^{q} \sum_{j<i} I\left(b_{hkj} > b_{hki}\right) |\eta_{hkj} - \eta_{hki}|.$$

Here, $I(b_{hkj} > b_{hki})$ indicates whether there is an incorrect ranking in the transition $h \to k$, and $|\eta_{hkj} - \eta_{hki}|$ is a weighted function that takes into account the severity of the ranking mistake as mentioned by Cox et al. [17,18].

3 Simulation Study and Application

3.1 Simulation Study

Simulation studies are presented in this section to evaluate the model performance and to compare different proposed algorithms.

3.1.1 Data Generation

Mimicking the structure of the real data (depicted in Fig. 1), we consider a three-state process ($h, k \in \{1, 2, 3\}$) with three possible transitions. The vector of the observed times of transitions of $T_i = (T_{i,12}, T_{i,13}, T_{i,23})$, was generated according to similar studies [25,26]. For the *i*th subject, the censoring time was generated according to a uniform distribution, $C_i \sim \text{Uniform}(0, T)$, where T is the largest time-point in the study (here, two different values of 3 and 12 months were assumed).

The vector of $T_i^* = (T_{i,12}^*, T_{i,13}^*, T_{i,23}^*)$ (indicating true transition times) was generated as follows:

I. Three random numbers of $u_{i,12}, u_{i,13}, u_{i,23}$ were generated from standard uniform distribution.

II. The values of $T_{i,12}^*$ and $T_{i,13}^*$ were generated by solving the following equations: $\int_0^{T_{i,1k}^*} \lambda_{i,1k}(\upsilon_{1k}) d\upsilon_{1k} + \log(u_{i,1k}) = 0$, for k = 2, 3. The Brent's root funding method [27] was employed.

III. Finally, the value of $T_{i,23}^*$ was generated by solving $\int_{T_{i,12}^*}^{T_{i,23}^*} \lambda_{i,23}(\upsilon_{23}) d\upsilon_{23} + \log(u_{i,23}) = 0$.

Eventually, the observed transition times T_i were considered by the following relationships:

- $T_{i,12} = \min(T_{i,12}^*, C_i)$,
- $T_{i,13} = \min(T_{i,13}^*, C_i)$,
- $T_{i,23} = \begin{cases} T_{i,23}^* & \textit{if } \left(T_{i,12}^* + T_{i,23}^* \le C_i\right) \\ C_i & \textit{if } \left(T_{i,12}^* + T_{i,23}^* > C_i\right) \end{cases}$.

The transition probability from the first state to the kth ($k = \{2, 3\}$) state was calculated using $P_{i,1k}(T) = \dfrac{\lambda_{i,1k}(T)}{\lambda_{i,12}(T) + \lambda_{i,13}(T)}$, and patients were eventually transferred to each of these states with a greater probability.

Software Source codes of this paper are available on Github:

https://github.com/BehnazAlafchi/Variable-importance-for-multi-state-data.

3.1.2 Simulation Result

The performance of our proposed two-step algorithm was evaluated via simulation studies utilizing the combination of different risk indices with the gradient boosting and random forest algorithms. We also used the original generated transition times at the first step and then applied the neural network, gradient boosting, and random forest algorithms at the second step to evaluate whether it could be useful to use the original survival times instead of risk indices for the variable selection. We considered 10 covariates in each possible transition, but only three of them were effective. All of the effective covariates were generated from an uniform distribution $U[0, 1]$. We simulated multi-state data with $(\lambda_{012}, \lambda_{013}, \lambda_{023}) = (0.15, 0.1, 0.1)$. The maximum time of follow-up was 3 and 12 months in different simulation studies. The results were based on two sample sizes, n = 500 and n = 1000, and each simulation run was based on 1000 iterations.

Tables 1 and 2 show the values of the loss function provided by different methods. Table 1 provides the results for a 1-year study, and Table 2 provides the results for a 3-month study. In these tables, the first column indicates the type of transition, the second column gives the values of the three effective parameters, and the subsequent columns give the mean value of the loss function provided by different methods. Columns 3 and 5 give the results for D, columns 6 and 8 show the results for DR, and columns 9 and 11 show the results for MR. As the simulation results show, the gradient boosting method on MR and DR provided the smallest loss values in almost all simulations for all transitions, respectively. The results also revealed that the proposed two-stage method has better performance for larger sample sizes. However, its performance was almost similar across different lengths of studies. The last column also shows the performance of the traditional multi-state model fitted to the generated data in simulations. According to the results, machine learning models outperformed the traditional model based on the Cox proportional hazards model.

Table 1: The values of loss function for different classification methods and different residuals/time-to-event with follow-up time 365 days

	Parameters	D. ANN	D. RF	D. GBM	DR. ANN	DR. RF	DR. GBM	MR. ANN	MR. RF	MR. GBM	SMSM
n = 500											
Transition 1 → 2	2/1.7/0.8	10.54	11.15	12.04	27.72	3.18	**0.34**	28.26	2.19	**0.18**	29.33
Transition 1 → 3	2/1.9/0.1	9.50	10.49	12.12	7.41	4.07	**0.48**	24.91	1.87	**0.26**	27.45
Transition 2 → 3	1/0.9/0.1	6.28	5.87	6.54	3.75	2.26	**0.52**	12.42	2.89	**0.39**	13.50
Transition 1 → 2	2/1/0.1	9.09	8.63	9.05	20.90	2.54	**0.67**	21.18	1.61	**0.30**	22.14
Transition 1 → 3	2/0.8/0.2	9.46	8.29	8.56	6.11	4.06	**2.19**	18.62	3.06	**1.13**	20.97
Transition 2 → 3	1.5/0.5/0.3	7.88	7.15	8.17	5.71	3.32	**2.07**	15.79	3.54	**1.60**	15.45
Transition 1 → 2	1/0.9/0.1	6.31	6.34	3.37	12.37	**0.74**	3.12	12.34	2.60	**0.36**	13.47
Transition 1 → 3	1/0.6/0.1	6.13	5.77	5.65	**2.79**	4.45	3.09	8.80	3.43	**1.20**	11.21
Transition 2 → 3	1/0.9/0.3	6.30	6.57	7.10	3.52	3.08	**0.82**	12.92	3.39	**0.67**	14.16
n = 1000											
Transition 1 → 2	2/1.7/0.8	8.51	9.01	10.14	27.08	1.05	**0.08**	28.79	0.74	**0.03**	29.38
Transition 1 → 3	2/1.9/0.1	8.57	8.12	9.23	6.46	1.62	**0.26**	25.54	0.65	**0.08**	27.44
Transition 2 → 3	1/0.9/0.1	6.04	5.63	5.49	3.31	1.03	**0.32**	13.22	1.31	**0.16**	13.52
Transition 1 → 2	2/1/0.1	6.99	7.94	8.30	21.26	1.52	**0.35**	21.74	1.02	**0.04**	22.13
Transition 1 → 3	2/0.8/0.2	8.23	6.74	7.62	6.65	2.52	**1.28**	19.74	**0.41**	1.66	21.18

(Continued)

Table 1 (continued)

	Parameters	D. ANN	D. RF	D. GBM	DR. ANN	DR. RF	DR. GBM	MR. ANN	MR. RF	MR. GBM	SMSM
Transition 2 → 3	1.5/0.5/0.3	6.07	6.34	6.13	5.07	**0.72**	1.86	15.14	1.88	**0.43**	15.89
Transition 1 → 2	1/0.9/0.1	6.56	6.03	5.89	13.05	1.48	**0.41**	12.85	0.99	**0.18**	13.52
Transition 1 → 3	1/0.6/0.1	6.19	5.21	5.41	2.97	3.35	**1.93**	10.05	2.18	**0.55**	11.58
Transition 2 → 3	1/0.9/0.3	5.64	6.37	6.43	3.74	1.47	**0.55**	13.69	1.63	**0.38**	14.36

Note: MR.ANN: Artificial neural network based on martingale residual MR.RF: Random forest based on martingale residual MR.GBM: Gradient boosting machine based on martingale residual D.ANN: Artificial neural network based on deviance D.RF: Random forest based on deviance D.GBM: Gradient boosting machine based on deviance DR.ANN: Artificial neural network based on deviance residual DR.RF: Random forest based on deviance residual DR.GBM: Gradient boosting machine based on deviance residual SMSM: Stepwise transition-specific proportional intensities multi-state model For each row, two least values are indicated in bold.

Table 2: The values of loss function for different classification methods and different residuals/time-to-event with follow-up time 90 days

	Parameters	D. ANN	D. RF	D. GBM	DR. ANN	DR. RF	DR. GBM	MR. ANN	MR. RF	MR. GBM	SMSM
n = 500											
Transition 1 → 2	2/1.7/0.8	9.62	11.02	1.77	26.47	3.02	**0.41**	27.73	2.46	**0.24**	29.37
Transition 1 → 3	2/1.9/0.1	10.28	9.37	10.92	7.87	4.14	**0.42**	24.67	1.78	**0.25**	27.49
Transition 2 → 3	1/0.9/0.1	6.07	6.10	6.78	3.96	2.72	**0.61**	12.44	3.09	**0.50**	13.51
Transition 1 → 2	2/1/0.1	8.77	8.44	9.59	20.51	2.04	**0.62**	20.99	1.76	**0.19**	22.11
Transition 1 → 3	2/0.8/0.2	9.09	8.38	9.44	5.33	4.17	**2.09**	18.58	2.79	**1.12**	21.01
Transition 2 → 3	1.5/0.5/0.3	7.32	6.30	6.82	4.05	2.76	**1.16**	14.37	2.95	**1.10**	15.56
Transition 1 → 2	1/0.9/0.1	6.15	5.88	6.53	12.50	3.06	**0.87**	12.25	2.63	**0.36**	13.49
Transition 1 → 3	1/0.6/0.1	6.36	5.43	5.77	3.49	4.15	**2.61**	8.78	3.04	**1.26**	11.02
Transition 2 → 3	1/0.9/0.3	6.85	6.36	6.38	4.39	3.36	**1.01**	12.75	3.80	**0.72**	14.02
n = 1000											
Transition 1 → 2	2/1.7/0.8	8.06	10.45	10.42	27.97	1.07	**0.06**	28.83	1.01	**0.02**	29.38
Transition 1 → 3	2/1.9/0.1	7.89	8.61	10.24	9.72	1.08	**0.32**	26.37	0.51	**0.08**	27.45
Transition 2 → 3	1/0.9/0.1	6.26	5.93	6.54	4.85	1.24	**0.33**	13.15	1.57	**0.19**	13.53

(Continued)

Table 2 (continued)

	Parameters	D. ANN	D. RF	D. GBM	DR. ANN	DR. RF	DR. GBM	MR. ANN	MR. RF	MR. GBM	SMSM
Transition 1 → 2	2/1/0.1	6.02	7.45	8.42	21.32	1.35	**0.28**	21.75	0.87	**0.03**	22.17
Transition 1 → 3	2/0.8/0.2	8.74	6.99	8.21	6.27	2.77	**1.28**	19.97	1.95	**0.47**	21.18
Transition 2 → 3	1.5/0.5/0.3	6.10	6.31	6.32	4.72	1.68	**0.85**	14.98	1.94	**0.58**	15.89
Transition 1 → 2	1/0.9/0.1	6.31	5.84	6.29	12.97	1.36	**0.44**	12.97	1.14	**0.19**	13.49
Transition 1 → 3	1/0.6/0.1	6.32	5.51	5.70	3.84	3.72	**1.48**	10.12	2.49	**0.56**	11.58
Transition 2 → 3	1/0.9/0.3	5.81	6.22	6.41	4.87	1.85	**0.60**	13.96	2.25	**0.43**	14.31

Note: MR.ANN: Artificial neural network based on martingale residual MR.RF: Random forest based on martingale residual MR.GBM: Gradient boosting machine based on martingale residual D.ANN: Artificial neural network based on deviance D.RF: Random forest based on deviance D.GBM: Gradient boosting machine based on deviance DR.ANN: Artificial neural network based on deviance residual DR.RF: Random forest based on deviance residual DR.GBM: Gradient boosting machine based on deviance residual SMSM: Stepwise transition-specific proportional intensities multi-state model For each row, two least values are indicated in bold.

4 Application

We apply our proposed two-step variable selection algorithm to a dataset of COVID-19 hospitalized patients. Each run took 30 s using an HP i5-laptop, RAM 8. The information of 2943 hospitalized patients with COVID-19 from February 20, 2020, to June 02, 2021 in Farshchian Medical Center and Shahid Beheshti Medical Center in Hamadan province, the west of Iran, was enrolled (Table A in the Appendix A). All of the patients were admitted to the ward (state 1). Then a patient may be transferred to the ICU (state 2) or die (state 3). The outcomes of interest were time to transfer to the ICU, time to death, and time to death after admission to the ICU. All patients who were alive at the end of the study were censored for death, and those who did not need to transfer to the ICU during the study were censored for admission to the ICU. The multi-state structure of the data is depicted in Fig. 1. The matrix below shows the number of observed transitions between different health states:

$$\Omega = \begin{array}{cc} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 2016 & 852 & 75 \\ 0 & 438 & 414 \\ 0 & 0 & 489 \end{pmatrix} \end{array}$$

This matrix gives the number of direct transitions between health states. In total, 852 patients were transferred to the ICU. A total of 489 patients died during the follow-up; among them, 414 patients died in the ICU and 75 patients died in the ward. In addition, 2454 patients were recovered and were considered as censored. Among them, a number of 2016 patients were censored for both transfer to ICU and death events, and a number of 438 patients were censored for death.

The mean (SD) and median age of patients were 60.2 (17.11) and 61.0 years, respectively. The clinical and demographic information of the patients are given in Appendix A. Here, we applied our proposed two-step algorithm to detect associated covariates with the risk of transition between different health states. We have used MR at the first step and gradient boosting at the second step.

Figs. 2a–2c depicts the variable importance associated with Admission in ward → ICU, Admission in ward → Death, and ICU → Death, respectively. In this figure, the variables are ranked according to their importance. Based on Fig. 2a, the most important variables for the time from admission in the ward to transfer to the ICU were saturation of peripheral oxygen (SPO2), age, lymphocyte count (LYM), lactate dehydrogenase (LDH), hemoglobin (Hb), blood sugar (BS), initial heart rate, body mass index (BMI), erythrocyte sedimentation rate (ESR), and Oseltamivire, respectively. The most important variables for the transition from ward to death were age, SPO2, LYM, LDH, Hb, initial heart rate, ESR, heart disease, BMI, and cancer, respectively (Fig. 2b). Moreover, as given in Fig. 2c, SPO2, age, Hb, hydroxychloroquine, LDH, LYM, BS, ESR, initial heart rate, BMI, and Koltra were the most important variables at the time of transition from ICU to death. Among many others, these variables were chosen as the most important (presented in Appendix A).

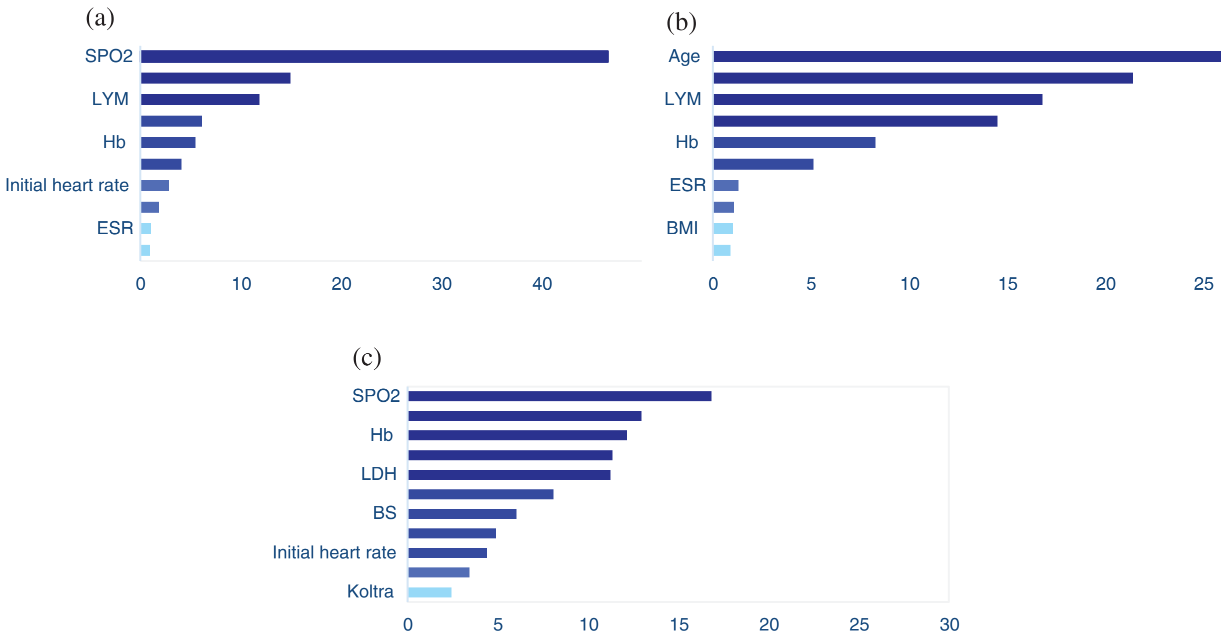

Figure 2: The results of variable importance based on martingale residual and gradient boosting algorithm for transitions from (a) Admission in ward to ICU, (b) Admission in ward to death, and (c) ICU to death, in hospitalized COVID-19 patients

Our case study revealed that SPO2 was an important predictor of the time of transition to the ICU and the time of death (among patients who were transferred to the ICU or not). Other studies have revealed that SPO2 affects the survival time and the length of hospital stay in COVID-19 patients [28]. According to Zhao et al. [29] a lower level of oxygen saturation at the time of admission is associated with a longer length of stay in the hospital.

Consistent with other studies, age was one of the most important predictors of the time of transition to ICU and the time of death among patients who were either transferred to ICU or not [28–31]. Several studies revealed that older patients were more likely to be transferred to the ICU and die [28,29]. Moreover, it has been shown that the age of the patients can influence the effects of other risk factors such as LYM on COVID-19 outcomes [32], which is an important risk factor for the disease progression [31,33]. These findings are consistent with the results of previous studies proposing that have shown that lower levels of LYM are related to the higher risk of admission to the ICU [32,34], such that ICU admitted patients had a decreased fraction of LYM compared to the other patients who were admitted to the ward [29]. Moreover, lower levels of LYM in peripheral blood were observed among patients who died [29].

The results also showed that the level of Hb is another important predictor for all three transitions, especially for mortality among ICU admitted patients. Other researchers have conducted different studies to evaluate the association between the severity of the disease or mortality and anemia among COVID-19 patients and they have received controversial results [30,35–39]. A meta-analysis study has shown that the levels of Hb were considerably lower than the normal level in patients with severe disease [40]. Several studies, mostly conducted in China, have shown an association between anemia and poor outcomes in hospitalized patients, and this may be because of its impact on immunity [38,39].

5 Discussions and Conclusions

Estimating variable importance in a multi-state process can be a challenging issue due to the complex relationships between variables. Classic variable selection methods like stepwise proportional hazards regression or penalized methods fail to provide an estimation of variable importance or considering complex/non-linear relationships between the inputs and outputs as well as the interaction between covariates. Therneau et al. [19] suggested that the martingale residual obtained from a null Cox model can be used as the outcome variable, so that usual regression analysis or classification methods like machine learning techniques can be applied to the new outcome variable (martingale residuals), and they provide very good results for survival data. This approach has also been applied for analyzing multivariate survival data like recurrent events. Nevertheless, few attempts have been made in relation to the multi-state data. In this paper, we proposed a two-step algorithm to evaluate the variable's importance for MSMs. We applied neural network, random forest, and gradient boosting algorithms to the martingale residuals, deviance, and deviance residuals made from an MSM and compared their results. The simulation studies revealed that using gradient boosting on the martingale residuals and deviance residuals outperforms other algorithms. This may be due to the fact that the gradient boosting is trained sequentially, so that in the training process the errors in the previous steps are corrected. This is in contrast to methods like random forest, where the trees are parallel and are made independently. Gradient boosting is also able to capture complex patterns in the data. Moreover, the individual predictions, obtained from several independent trees (that are determined in any order) in the random forest, are aggregated (by the principal of the majority vote or the average value), while the sequence of gradient boosting does not change (it runs in a fixed order). Boosted trees are prone to overfitting and begin modeling the noise in the presence of noisy data, despite the benefits of gradient boosting.

In this study, we assumed a continuous and Markov multi-state process. Nevertheless, in other contexts, a semi-Markov or non-Markov process could be defined as well. In addition to considering ensemble methods, other model selection methods like support vector machines and deep learning methods can be introduced into the proposed residuals, which can be considered as a future work. Also, optimization of tuning parameters using heuristic algorithms like genetic algorithms or Bayesian optimization methods is worth investigating in future studies.

Here, we used our proposed method to identify the important variables at the time of transition from the ward to the ICU and death among COVID-19 patients who were admitted to the hospital. It should be noted that our goal in this case was simply to identify the most important variables influencing the risk of transitioning between different health states. To assess the direction and impact of the selected variables on the risk of different transitions, the use of classical multi-state models can be used as a complementary method. It is noteworthy that while there are too many studies that have utilized machine learning methods in COVID-19 data sets, no study has hybridized machine learning and multi-state methods, especially in analyzing COVID-19 data sets. This highlights the novelty aspect of this study.

Limitations

There were some limitations to the present study. In the data used in this study, information on only three states was available, including admission to the ward, ICU, and death. Although, there were patients who were retransferred to the ward from the ICU, their information was not available. So, our example had the illness-death multi-state structure. It is suggested to analyze more complex multi-state data sets with the provided model in this study. Another limitation was that dealing with time-dependent covariates is only possible by using martingale residual and adjusting martingale residual. Despite these limitations, the proposed method can be easily applied to the context of high-dimensional data like genome-wide association studies and medical image data to detect the most important genes or brain regions associated with survival outcomes more accurately than classic statistical methods.

Acknowledgement: We would like to appreciate the Vice-chancellor of Education of the Hamadan University of Medical Science for technical support for their approval and support of this work.

Ethics Approval and Consent to Participate: The data were collected from the patients' medical recodes that have already been discharged and were not accessible for giving informed consent. A waiver of informed consent was awarded for the analysis conducted in this study by the Ethical Committee of the Hamadan University of Medical Sciences. All methods were carried out in accordance with relevant guidelines and regulations, and the study was approved by the Ethical Committee of the Hamadan University of Medical Sciences (IR.UMSHA.REC.1401.251; No. 140104072334).

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Ferrer, L., Rondeau, V., Dignam, J., Pickles, T., Jacqmin-Gadda, H. et al. (2016). Joint modelling of longitudinal and multi-state processes: Application to clinical progressions in prostate cancer. *Statistics in Medicine, 35(22),* 3933–3948. DOI 10.1002/sim.6972.
2. Mirzaee, M., Mohammad, K., Mahmoodi, M., Zeraati, H., Ebadzadeh, M. R. et al. (2014). Multi-state survival analysis in renal transplantation recipients. *Iranian Journal of Public Health, 43(3),* 316–322.
3. Ursino, M., Dupuis, C., Buetti, N., de Montmollin, E., Bouadma, L. et al. (2021). Multistate modeling of COVID-19 patients using a large multicentric prospective cohort of critically ill patients. *Journal of Clinical Medicine, 10(3),* 544. DOI 10.3390/jcm10030544.
4. Beesley, L. J., Morgan, T. M., Spratt, D. E., Singhal, U., Feng, F. Y. et al. (2019). Individual and population comparisons of surgery and radiotherapy outcomes in prostate cancer using Bayesian multistate models. *JAMA Network Open, 2(2),* e187765. DOI 10.1001/jamanetworkopen.2018.7765.
5. Dang, X., Huang, S., Qian, X. (2021). Risk factor identification in heterogeneous disease progression with L1-regularized multi-state models. *Journal of Healthcare Informatics Research, 5(1),* 20–53. DOI 10.1007/s41666-020-00085-1.

6. Duan, R., Fu, H. (2015). Estimate variable importance for recurrent event outcomes with an application to identify hypoglycemia risk factors. *Statistics in Medicine, 34(19),* 2743–2754. DOI 10.1002/sim.6516.
7. Sennhenn-Reulen, H., Kneib, T. (2016). Structured fusion lasso penalized multi-state models. *Statistics in Medicine, 35(25),* 4637–4659. DOI 10.1002/sim.7017.
8. van der Laan, M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics, 2(1),* 1–31. DOI 10.2202/1557-4679.1008.
9. Stachel, A., Daniel, K., Ding, D., Francois, F., Phillips, M. et al. (2021). Development and validation of a machine learning model to predict mortality risk in patients with COVID-19. *BMJ Health & Care Informatics, 28(1),* e100235. DOI 10.1136/bmjhci-2020-100235.
10. Snider, B., McBean, E. A., Yawney, J., Gadsden, S. A., Patel, B. (2021). Identification of variable importance for predictions of mortality from COVID-19 using AI models for Ontario, Canada. *Frontiers in Public Health, 9,* 6757–6766.
11. Hazard, D., Kaier, K., von Cube, M., Grodd, M., Bugiera, L. et al. (2020). Joint analysis of duration of ventilation, length of intensive care, and mortality of COVID-19 patients: A multistate approach. *BMC Medical Research Methodology, 20(1),* 1–9. DOI 10.1186/s12874-020-01082-z.
12. Breiman, L. (2001). Random forests. *Machine Learning, 45(1),* 5–32. DOI 10.1023/A:1010933404324.
13. Breiman, L. (1996). Bagging predictors. *Machine Learning, 24(2),* 123–140. DOI 10.1007/BF00058655.
14. Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., Boivin, J. F. (2005). Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistics and Computing, 15(3),* 231–239. DOI 10.1007/s11222-005-1311-z.
15. Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer-Verlag.
16. Andersen, P. K., Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics, 10(4),* 1100–1120. DOI 10.1214/aos/1176345976.
17. Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34(2),* 187–202.
18. Cox, D. R. (1972). Breslow's commons on regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 23,* 187–220.
19. Therneau, T. M., Grambsch, P. M., Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika, 77(1),* 147–160. DOI 10.1093/biomet/77.1.147.
20. Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7,* 21. DOI 10.3389/fnbot.2013.00021.
21. Freund, Y., Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55(1),* 119–139. DOI 10.1006/jcss.1997.1504.
22. Friedman, J., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics, 28(2),* 337–407. DOI 10.1214/aos/1016218223.
23. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29(5),* 1189–1232. DOI 10.1214/aos/1013203451.
24. Raj, P., Evangeline, P. (2020). *The digital twin paradigm for smarter systems and environments: The industry use cases*. London, UK: Academic Press.
25. Beyersmann, J., Allignol, A., Schumacher, M. (2012). *Competing risks and multistate models with R*. New York, NY: Springer Science & Business Media.
26. Crowther, M. J., Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine, 32(23),* 4118–4134. DOI 10.1002/sim.5823.
27. Brent, R., Richard, P. (1973). *Algorithms for minimization without derivatives*. New York: Dover Publications.

28. Nasir, M., Perveen, R. A., Ahmad, S. N., Nazneen, R., Ahmed, S. M. P. (2021). Outcome of instrumental oxygen therapy in COVID-19: Survivors versus non-survivors in Bangladeshi cohort. *American Journal of Internal Medicine, 9(1),* 52–57. DOI 10.11648/j.ajim.20210901.18.
29. Zhao, Z., Chen, A., Hou, W., Graham, J. M., Li, H. et al. (2020). Prediction model and risk scores of ICU admission and mortality in COVID-19. *PLoS One, 15(7),* e0236618. DOI 10.1371/journal.pone.0236618.
30. Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine, 382(18),* 1708–1720. DOI 10.1056/NEJMoa2002032.
31. Lu, J., Hu, S., Fan, R., Liu, Z., Yin, X. et al. (2020). ACP risk grade: A simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China. DOI 10.1101/2020.02.20.20025510.
32. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet, 395(10223),* 497–506. DOI 10.1016/S0140-6736(20)30183-5.
33. Hou, W., Zhang, W., Jin, R., Liang, L., Xu, B. et al. (2020). Risk factors for disease progression in hospitalized patients with COVID-19: A retrospective cohort study. *Infectious Diseases, 52(7),* 498–505. DOI 10.1080/23744235.2020.1759817.
34. Ng, D. H., Choy, C. Y., Chan, Y. H., Young, B. E., Fong, S. W. et al. (2020). Fever patterns, cytokine profiles, and outcomes in COVID-19. *Open Forum Infectious Diseases, 7(9),* ofaa375. DOI 10.1093/ofid/ofaa375.
35. Bellmann-Weiler, R., Lanser, L., Barket, R., Rangger, L., Schapfl, A. et al. (2020). Prevalence and predictive value of anemia and dysregulated iron homeostasis in patients with COVID-19 infection. *Journal of Clinical Medicine, 9(8),* 2429. DOI 10.3390/jcm9082429.
36. Tao, Z., Xu, J., Chen, W., Yang, Z., Xu, X. et al. (2021). Anemia is associated with severe illness in COVID-19: A retrospective cohort study. *Journal of Medical Virology, 93(3),* 1478–1488. DOI 10.1002/jmv.26444.
37. Young, B. E., Ong, S. W. X., Kalimuddin, S., Low, J. G., Tan, S. Y. et al. (2020). Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore. *JAMA, 323(15),* 1488–1494. DOI 10.1001/jama.2020.3204.
38. Ryan, A. S. (1997). Iron-deficiency anemia in infant development: Implications for growth, cognitive development, resistance to infection, and iron supplementation. *American Journal of Biological Anthropology, 104(S25),* 25–62. DOI 10.1002/(ISSN)1096-8644.
39. Dinevari, M. F., Somi, M. H., Majd, E. S., Farhangi, M. A., Nikniaz, Z. (2021). Anemia predicts poor outcomes of COVID-19 in hospitalized patients: A prospective study in Iran. *BMC Infectious Diseases, 21(1),* 1–7.
40. Lippi, G., Mattiuzzi, C. (2020). Hemoglobin value may be decreased in patients with severe coronavirus disease 2019. *Hematology, Transfusion and Cell Therapy, 42,* 116–117.

Appendix A. Characteristics of COVID-19 patients

Table A: Characteristics of the study population infected with COVID-19 (n = 2943)

Variable	Category	Number	Percent
Sex	Male	1565	53.2
	Female	1378	46.8
Marital status	Married	2560	87.4
	Single	368	12.6

(Continued)

Table A (continued)

Variable	Category	Number	Percent
Resident	Urban	2452	83.3
	Ruler	490	16.7
Smoking	No	2727	92.7
	Yes	216	7.3
Substance abuse	No	2740	93.1
	Yes	203	6.9
Lung disease	No	2641	89.7
	Yes	302	10.3
Heart disease	No	2487	84.5
	Yes	456	15.5
Diabetes	No	2384	81.0
	Yes	559	19.0
Kidney disease	No	2821	95.9
	Yes	122	4.1
Hypertension	No	1942	66.0
	Yes	1001	34.0
Liver disease	No	2919	99.2
	Yes	24	0.80
Cancer	No	2884	98.0
	Yes	59	2.0
Weakness, lethargy, and fatigue	No	2620	89.0
	Yes	323	11.0
Lack of smell	No	2884	98.0
	Yes	59	2.0
Vomiting	No	2219	75.4
	Yes	724	24.6
Diarrhea	No	2654	90.2
	Yes	289	9.8
Fever	No	1417	48.1
	Yes	1526	51.9
Shortness of breath	No	1220	41.5
	Yes	1723	58.5
Muscle pain	No	1610	54.7
	Yes	1333	45.3
Stomachache	No	2926	99.4
	Yes	17	0.6
Chest pain	No	2881	97.9
	Yes	62	2.1
Loss of consciousness	No	2919	99.2
	Yes	24	0.8
Sore throat	No	2837	96.4

(Continued)

Table A (continued)

Variable	Category	Number	Percent
	Yes	106	3.6
Cough	No	1173	39.9
	Yes	1770	60.1
Headache	No	2425	82.4
	Yes	518	17.6
Blood pressure	Normal	2776	94.9
	Abnormal	149	5.1
Oxygen therapy	No	123	4.2
	Yes	2820	95.8
Mechanical ventilation	No	2365	80.4
	Yes	578	19.6
Oseltamivir	No	2680	91.1
	Yes	263	8.9
Azithromycin	No	1725	58.6
	Yes	1218	41.4
Kaltura	No	1217	41.4
	Yes	1726	58.6
Hydroxychloroquine	No	1490	50.6
	Yes	1453	49.4
Variable		**Mean**	**SD**
Age		60.2	17.11
BMI		26.64	4.49
Initial body temperature		36.91	1.72
Initial heart beat		92.17	15.44
SPO2		84.41	10.89
ESR		43.17	29.20
LDH		597.00	276.13
BS		146.87	74.10
Hb		13.86	2.13
LYM		22.52	11.68

Note: BMI: body mass index; SPO2: saturation of peripheral oxygen; ESR: erythrocyte sedimentation rate; LDH: lactate dehydrogenase; BS: blood sugar; Hb: hemoglobin; LYM: lymphocyte count.