Risk Score from Survival Random Forest Predicts Lower ICU Admission Among Patients

Abstract

ICU (Intensive Care Unit) is a hospital ward specifically used to assist patients who are at the highest risk of death. The COVID-19 pandemic has taught us once again the need to best handle situations where the number of patients in need of intensive care far exceeds the number of ICU beds available. Physicians must be able to make decisions based on some type of "risk" analysis method to prioritize patients needing the ICU beds the most. This study aimed to obtain an "optimal" score that best identifies subjects with a high risk of being admitted into ICU using a state-of-the-art machine learning tool. Fourteen Survival random forests were built to simulate patients' chances of survival each day. A new risk score was constructed by taking a complement of an aggregate survival chance for each observation based on each iteration of the model. The analysis revealed the new risk score to illustrate milder risk than the existing risk score with most of the patients being deemed only moderately at risk. The prediction accuracies for each iteration of the forest also ranged between 53% to 94%. The study also explored which health biomarkers more useful in predicting risk based on the survival random forest. The result produced by the models revealed factors such as temperature, age, heart rate, oxygen saturation, and respiration rates to be most significant in accurately predicting patients' risk of ICU admission.

Introduction

The term ICU has grown familiar to us over the past two year ever since COVID-19 was first declared a pandemic in March 2020. ICU stands for Intensive Care Unit, and it is an equipment most often used as a last resort method to keep patients alive when their body fails to do so. One of the biggest concerns during the height of the pandemic was the massive shortage of ICU beds due to an exponential rise of cases worldwide. During such cases, doctors must decide which patients should be given priority based on their "risk score". Risk score, in our context, is thus defined as a metric that determines a patient's current/future likelihood of requiring a serious medical attention (ICU admission). Knowing patients' risk score can not only inform doctors on who needs most urgent attention, but it can also act as an intervention against future hospital visits.

This study aimed to explore an "optimal" risk score that best identified patients with an increased risk of ICU admission. It has also identified which health biomarkers are most significant in predicting patient risk as well as comparing the existing risk score with the new risk score. The study has constructed a risk score framework using aggregated survival functions generated from a machine learning technique called survival random forest. The method section has outline the logistic of the survival random forest used including its formula, cross-validation approaches and risk score construction. The result section then showed that the survival curves produced from the forest are consistent with our findings in the EDA, visually compared new risk score versus old risk score, and illustrated covariates deemed important by the model. Finally, the discussion section gave explanation for the behavior shown by the new risk score and explained possible limitations.

Methods

The data was collected from 900 patients over a course of zero to thirteen days until the patients were either discharged from the hospital or admitted to the ICU. Status of the patients' health indicators were recorded every four hours. Recorded health indicators used for the study include patients' oxygen level (%), number of breaths per minute, whether the patient required supplementary oxygen (supplementary oxygen = 1), blood pressure (mm Hg), heart rate, patients' consciousness, temperature (Celsius), and whether the patient had been admitted to the ICU beds (status = 1). Other factors such as age, sex (female = 1, male = 0), risk score used to measure current patients' risk, and duration of stay were also included in the study.

Table 1: List of variables involved

variable	explaination
ID	Patient ID
Time	Time of record
Respiration Rate	Number of breaths per minute
Oxygen Saturation (Percentage)	Percentage of oxygen travelling throughout the body
Supplementary Oxygen	Whether the patient is breathing normally
Systolic Blood Pressure (mm Hg)	Blood pressure measured during heart beats
Heart Rate	Number of heart beats per minute
Level of Consciousness	Whether the patient is conscious or unscious
Temperature (Celsius)	Patient temperature in celsius
Risk Score	Risk score derived from the variables combined
Sex	Sex of the patient
Age	Age of the patient
Event Time	Duration of patient stay in hopsital bed in days
Status	Whether the patient has experienced a serious event

All missing values have also been replaced to reflect that of their previous value. For example, if one of the observations was missing the patient's sex at a particular time stamp due to a technical error, then the missing value would be replaced with the value of its previous time stamp. This decision was made knowing that observations were frequently updated (every four hours), and thus the missing value will most likely be identical to its previous observation. In addition, no missing values were ever found within patients' first observations meaning that current imputation method is sufficient for the given data.

Data on patients' health indicators were also combined to better illustrate the change in patients' health over time. Rows containing Patients' health status were reduced based on time in which the records were updated and the data on patients' age, sex and total duration of stay were also combined with the patients' health data.

A survival random forest was used in the study to answer the aforementioned research questions. By definition, survival random forest is a non-parametric machine learning method used to build prediction models using an ensemble of simulated survival trees (Mogensen et al., 2012). The following steps were taken to build a survival random forest in context to the ICU data:

1. First, B bootstrap samples were selected at random with replacement.
2. A survival tree is built using one of the bootstrap samples with a random subset of covariates (e.g., temperature, sex, age, etc.) as nodes
    - when a tree is built, the node splits are determined with regards to the right-censored data (e.g., log-rank test) (Mogensen et al., 2012).
3. This step is repeated B times using a new bootstrap sample each time.

Right censoring is appropriate for this study since we do not know what happened after the last recorded stay of the patient. In this usage of survival random forest, 500 bootstrap samples were taken ($B = 500$) to build a random forest. A survival prediction ensemble is then calculated using the aggregate information from the leaves of the trees (Mogensen et al., 2012). In each leaf, a conditional cumulative hazard function (CHF) is estimated using the Nelson-Aalen estimator from the bootstrapped data.

$\widehat{H}_b(t|x) = \int_0^t \frac{\tilde{N}_b(ds,x)}{\tilde{Y}_b(s,x)}$

Where:

- b stands for the b^{th} bootstrap
- x stands for the predictor
- $\tilde{N}_b(ds, x)$ represents the uncensored event until time s
- $\tilde{Y}_b(s, x)$ represents the number at risk at time s

Lastly, the ensemble survival function is then obtained from aggregating the Nelson-Aalen estimate from all the leaves (i.e., $\widehat{H}_b(t|x)$).

$\widehat{S}^{rsf}(t|x) = exp(-\frac{1}{B}\sum_{b=1}^{B} \widehat{H}_b(t|x))$

This function is then used to predict the response of interest (i.e., the likelihood of experiencing a serious event or Status).

Survival random forest was used in this study for the following reasons: Firstly, random forest is a time-efficient tool that also produces highly accurate predictions. In particular, it offers flexibility for data such as survival data wherein restrictions in alternative models, such as Cox-proportional hazard models, may pose modelling constraints on the data (Ishwaran et al., 2008). Secondly, the research question at hand is inherently a survival problem. In order to obtain an optimal risk score that best predicts patients' ICU admission, we must use the *status* variable which describes whether a patient has been admitted to ICU or not (0 = no serious event and 1 = serious event/ICU admission). This is identical to a standard survival analysis question with the DV being patients' survival (0 = alive, 1 = death). Therefore, using a survival random forest offers us a flexible way to handle a survival analysis question without worrying about modelling constraints.

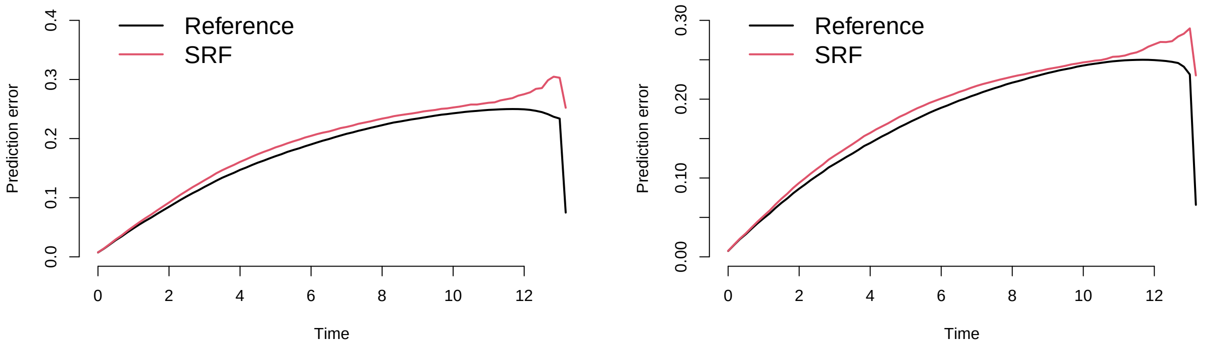
Since survival random forest was used to train the model for the analysis, the response was also augmented to reflect the change. Instead of using the basic response variable given in the data, the *Status* was modified using survival analysis techniques. In survival analysis, an outcome (DV) is described as time until an event happens, with the event usually being a binary response of yes or no. As for *Status*, a survival object was used to illustrate the time it takes until a patient would be either: admitted to the ICU (Status = 1) or be discharged (Status = 0). In other words, Status was modified to focus how time affects potential risk of patient being admitted into the ICU.

Diagnostic: Random forest is a powerful predictive tool because it is able to make accurate predictions despite the missing values, large number of covariates, and nonlinearity aspect of the models (McAlexander and Mentch, 2020). Consequently, there are no formal assumptions for random forests (Richmond, 2016). This does not mean, however, that random forest works perfectly in every scenarios. For example, although random forest-based models handle multicollinearity sufficiently well (Chowdhury et al., 2021), they can still skew the result if the effect of multicollinearity is strong enough. Therefore, I decided to remove risk score, a highly correlated covariate from the model to prevent the possibility of multicollinearity on the model. Furthermore, out-of-bag error rate was observed to determine the accuracy of the prediction made by the survival random forest. When bootstrapping the data, parts of the observations can be left out of the sampling. The collection of these observations that are unconsidered are known as out-of-bag (OOB) samples. On average, up to 37% of the overall data can be left out as an OOB sample in a single bootstrap (Ishwaran et al., 2008). Much like cross-validation, the OOB samples can be used as the test set to examine the accuracy of the model. OOB error rate is referred to the likelihood of the model failing when OOB sample is used as the test set. For survival random forest specifically, one minus Harrell's C-index is used to determine the out-of-bag (prediction) error rate (Wright et al., 2021). Harrell's C-index offers a way measure performance of risk predictions within survival analysis context (Schmid et al., 2016). The C-index is given as

$$C = \frac{\sum_{i,j} I(\widehat{T}_i > \widehat{T}_j) \cdot I(\eta_j > \eta_i) \cdot \Delta_j}{\sum_{i,j} I(\widehat{T}_i > \widehat{T}_j) \cdot \Delta_j}, 0 \leq C \leq 1$$

Where the numerator represents the number of concordant pairs (when shorter time until event, T, is attributed to higher risk score η) and the denominator represents the sum of concordant pairs and discordant pairs (when shorter time until event, T, is attributed to lower risk score η) (Tay, 2019). Simply put, higher C-index implies the risk score is highly accurate in predicting survival event while C-index closer to 0 implies the risk score is poor in predicting survival event. Therefore, one minus C-index is chosen to represent the out of bag prediction errors for survival random forest.

Figure 1: Prediction Error Curve for the Survival Random Forest



Although the process in which OOB error rate is obtained is identical to that of cross-validation, OOB is obtained as the forest is built and thus may be prone to possible biases. Therefore, I have conducted a cross-validation to validate the survival random forest post construction. A 2-fold cross-validation was used in reference to a process done by Thomas Gerlach (2018). Due to computational limitation being that the data used was too large, only 2-fold cross-validation was conducted. As seen from figure 1 above, the prediction error comparison between the trained curve and the true curve indicates minimal deviations; implying the predictive performance of the trained model and the reference model are identical. In conclusion, this shows that the survival random forest performs adequately with minimal predictive errors.

A new risk score was also proposed by taking the complement of the aggregate survival functions (likelihood of not being admitted to the ICU) for each patient at each time stamp. As mentioned above, the survival curves produced from the survival random forest illustrates the patient's chances of survival overtime (figure 2). However, because each observation in the data represents a patient's health statuses at a given *time stamp*, calculating the patient's chances of survival overtime at a specific time stamp would not make sense. Therefore, an aggregate survival chance was computed for each patient at a given time stamp by taking an average of the simulated survival chance overtime for each observation. This was then used to represent the patient's overall chances of survival at given time stamp. The complement (one minus) of the survival chance were then calculated to determine the likelihood of ICU admission (0 = 0% ICU admission & 1 = 100% ICU admission). This choice was made as the pre-existing risk score was illustrated the same way with there being a risk score for each patient at a given time stamp and this similarity would be useful in comparing the two risk scores. To accommodate for the effect of time, several iterations of the survival random forest were run to obtain a new risk score each day (n = 14). For each iteration, only the patients who were still in the hospital at the time were included in the present iteration. For patients who remained in the hospital, their cumulative records (from day 0 to present) were used in risk construction to better track the changes in patient's health over time.

Result

Result: After building the survival random forest, 7 patients were chosen through judgement sampling technique. The survival curves based on the patient's health status were averaged over the their duration of stay to illustrate the patient's overall risk of being admitted into ICU over time. As seen from figure 2, we can see that some samples illustrate steeper downward trend (indicating that risk of being admitted into ICU increases at a faster rate as time progresses OR the chance of survival decreases as time progresses), but other samples indicate a relatively shallow trend. The average of all simulated responses was also included to observe the overall trend across all patients, and once again we can see that the overall risk of being admitted into ICU increases at a slower rate as time progresses. In other words, the plot suggests that whether a patient is likely going to be admitted into ICU is usually determined during the first couple days, and once this window passes, the patients are not any more likely to be admitted to ICU than before. This phenomenon is consistent with the recorded differences in appendix a; where larger number of patients were admitted into ICU (Status = 1) during the first couple days and as time progressed, there were greater number of patients being discharged (Status = 0) versus patients being admitted into the ICU.

The OOB prediction error based on Harrell's C-index also revealed that the model is accurate 3/4 times.

Figure 2: Chances of Patient Discharge

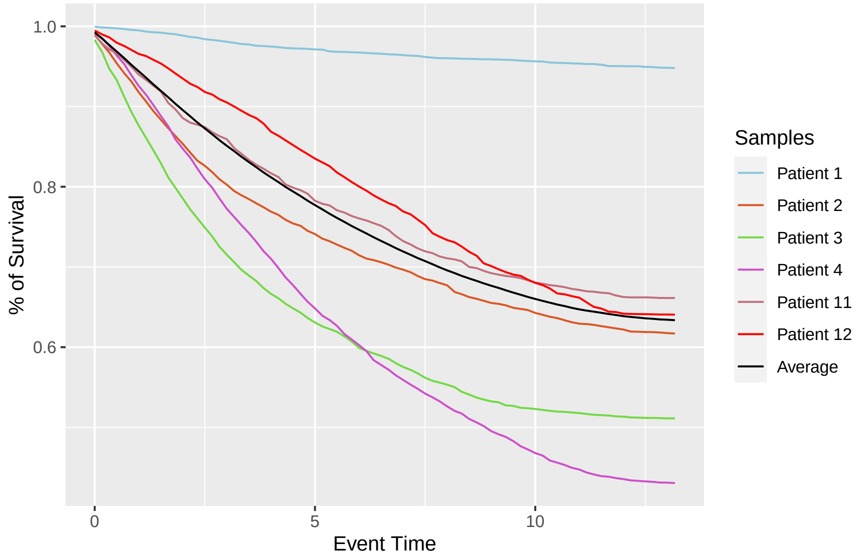
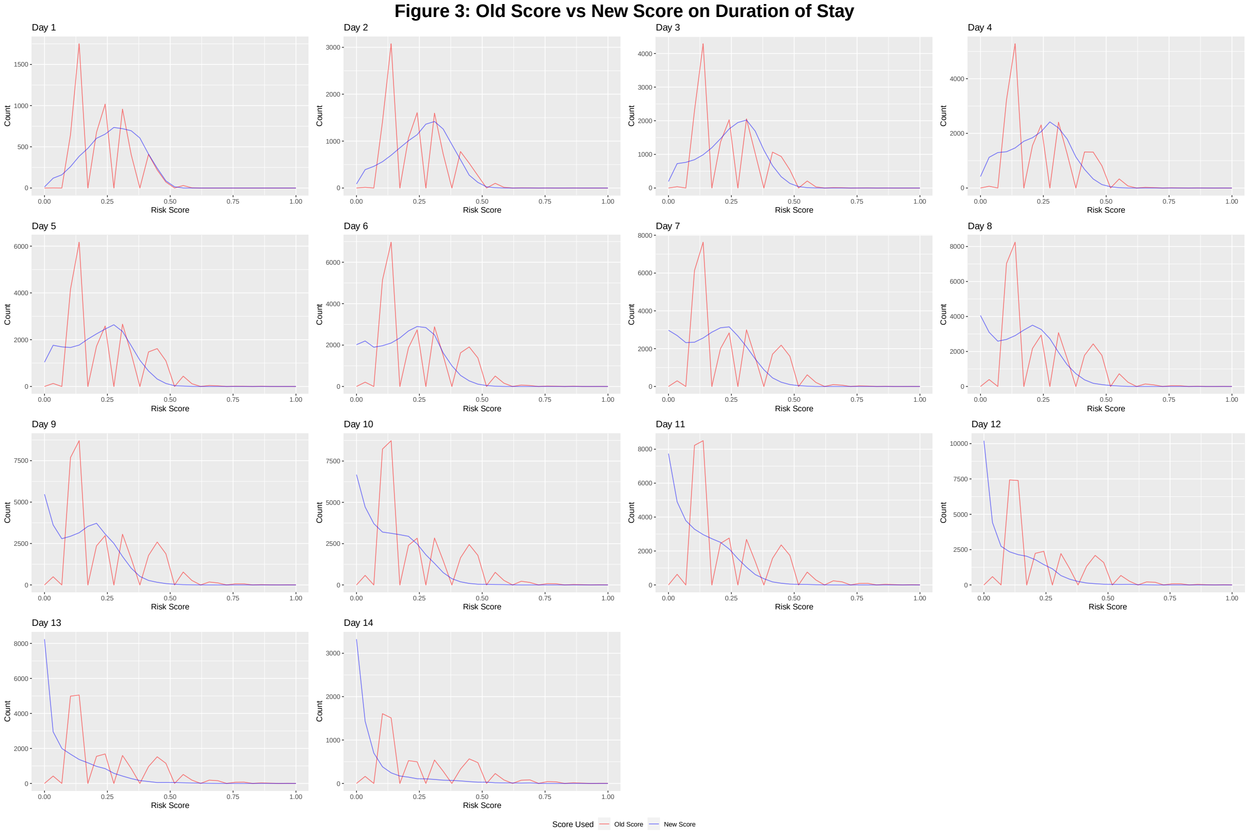


Table 2: Comparison of existing risk score and new risk score

ID	Time	Event Time	Old Score	New Score
1	2021-06-21 12:00:00	0.00	0.15	0.221
1	2021-06-21 16:00:00	0.17	0.15	0.242
1	2021-06-21 20:00:00	0.33	0.15	0.156
1	2021-06-22 00:00:00	0.50	0.20	0.109
1	2021-06-22 04:00:00	0.67	0.10	0.096
1	2021-06-22 08:00:00	0.83	0.15	0.235
1	2021-06-22 12:00:00	1.00	0.25	0.220
2	2021-08-12 20:00:00	0.00	0.30	0.403
2	2021-08-13 00:00:00	0.17	0.20	0.344
2	2021-08-13 04:00:00	0.33	0.30	0.125

Table 2 compares the existing risk score versus the new risk score. To allow for easier comparison, the existing risk score was converted to proportions to match that of the new risk score by taking the current old score over the highest possible score (20). Figure 3 better illustrates the distributions between two scores. Total of 14 plots are shown, each plot representing a day in the hospital stay. As seen from figure 3, it is clear that the new risk score demonstrate more conservative risk scoring than the existing risk score, with most of the scores clustering below 0.5 risk. We can also observe a similar pattern as figure 2 where the highest risk occurs during the first couple days and the overall risk of admission decreasing overtime.

Figure 3: Old Score vs New Score on Duration of Stay



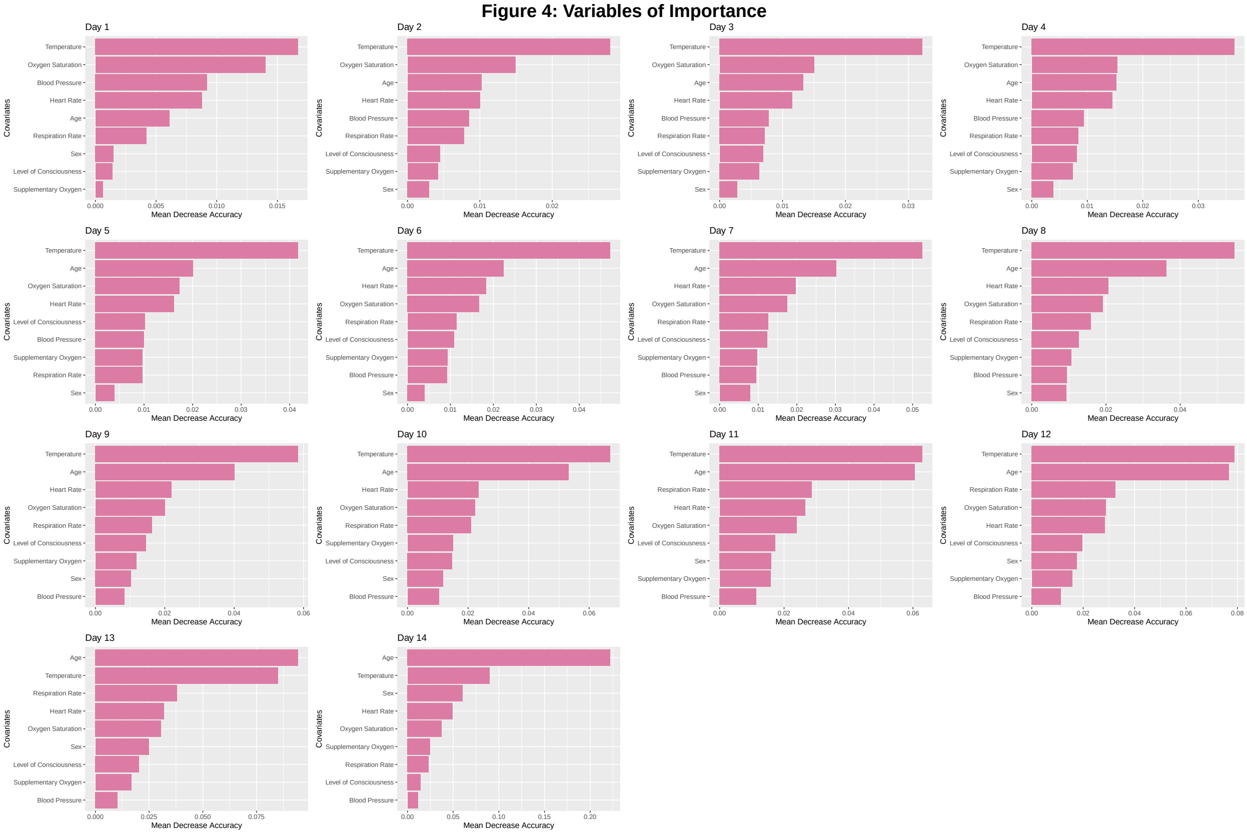
Another version of the risk score used isolated records of the patients restricted within that day as opposed to cumulative records in figure 3 to check for differences if only the records for that day were used to build the model. The result (shown in appendix g) illustrated a relatively similar trend as figure 3 which shows that either methods are viable. Figure 3 was chosen in the end, however, as the OOB predictive errors were much more robust in this version of the model as seen from table 3.

Table 3: Prediction errors for each iteration of the model

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14
0.469	0.4	0.372	0.343	0.313	0.295	0.271	0.254	0.238	0.207	0.185	0.152	0.131	0.069

Figure 4 illustrates which covariates are considered significant predictors of ICU admission. Typically, when a random forest is built, it ranks the covariates from the highest predictive power to the lowest. Figure 4 shows all variables of importance for each iteration of the model. From this we see that temperature is the most significant contributor to the model prediction at all times. This is followed by oxygen saturation, but age overtakes oxygen saturation in subsequent days. Heart rate and respiration rate also become more prevalent in later iterations of the model.

Figure 4: Variables of Importance



Discussion

So far, we have shown in our report a newly proposed model simulating patients' risk of ICU admission over time and we have also tested the model validity using methods such as cross-validation and OOB prediction errors. A simple illustration of the simulated survival curves have also shown to be consistent with our data as shown in appendix a and figure 3. Next, a new risk score was proposed using the complement of an aggregated survival function to represent the chances of ICU admission (table 2). As mentioned earlier, while the newly proposed risk score does illustrate a milder risk than its former, its validity can be assured as seen from their prediction errors (table 3). We see that the models can accurately predict 53% to 93% of the data with their accuracies increasing with each iteration. As of its current version, the risk score can be used as a substitute of the old risk score given their similarities. Since the new risk score is stored within the data of their respective n-th iteration, the health experts only need to access the data containing the n-th iteration to obtain patients' risk score at day n. This method is also efficient as the risk construction is mostly done by the survival random forest automatically and thus the health experts only need to use the given risk score from the table without further modification.

With regards to the variables of importance, we can see that temperature, age, oxygen saturation, heart rate, and respiration rate are the five most important contributors predicting the risk of a patient being admitted into the ICU. This result in figure 4 is consistent with the plots from appendix b, c, e, and f as seen from the noticeable differences between patients admitted into ICU and patients not admitted into ICU within these health statuses. It also makes intuitive sense that age will play a part in predicting patient's likelihood of ICU admission as we expect older patients to be at higher risk of experiencing a serious event than younger patients.

However, the newly proposed risk score is not without its limitations. Most noticeable challenge with the new risk score is that the majority of the calculated risk scores do not exceed 0.5. Although the existing risk score can be seen to exceed this threshold, this is not the case for the new risk score. This can be problematic as the risk score may incorrectly classify at-risk patients as low risk. One reason why this is occurring may be due to the averaging effect when the survival functions are aggregated. As explained earlier, the survival functions are averaged over time to illustrate an overall survival chance at a given time stamp. However, the problem with averages/summary statistics is that they tend to oversimplify the end result as it is illustrated as a point estimate. Therefore, future replications of the study may observe more accurate results by using an estimate that better captures true risk such as an interval estimate.

References

Chowdhury, S., Lin, Y., Liaw, B., & Kerbyd, L. (2021, November 5). *Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance.* Retrieved February 7, 2022, from https://arxiv.org/pdf/2111.02513.pdf

Mogensen UB, Ishwaran H, Gerds TA.*Evaluating Random Forests for Survival Analysis using Prediction Error Curves.* J Stat Softw. 2012;50(11):1-23. doi:10.18637/jss.v050.i11

McAlexander, R. J., & Mentch, L. (2020). *Predictive inference with random forests: A new perspective on classical analyses.* Research & Politics. https://doi.org/10.1177/2053168020905487

Richmond, S. (2016, March 21). *Algorithms exposed: Random Forest.* BCCVL. Retrieved February 7, 2022.

Rickert, J. (2017, September 25). *Survival analysis with R.* R Views. Retrieved February 7, 2022, from https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, Michael S. Lauer "Random survival forests," *The Annals of Applied Statistics, Ann. Appl. Stat.* 2(3), 841-860, (September 2008)

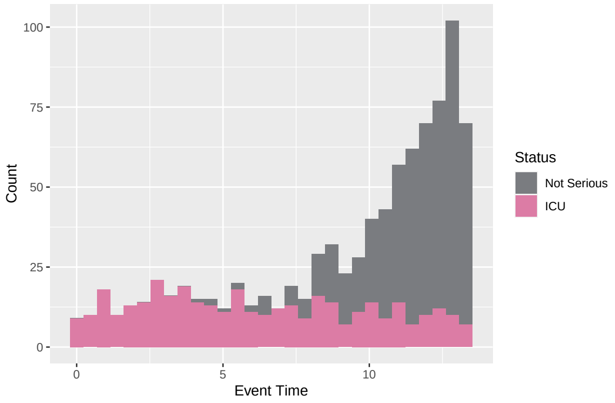
Gerlach, T. (2018). Random survival forest example, R, Package Ranger. GithubGist. Retrieved March 18, 2022, from https://gist.github.com/thomasmooon/6eb87964ea663f4a7441cc2b2b730bd4

Tay, K. (2019, October 26). What is Harrell's C-index? Statistical Odds & Ends. Retrieved April 6, 2022, from https://statisticaloddsandends.wordpress.com/2019/10/26/what-is-harrells-c-index/

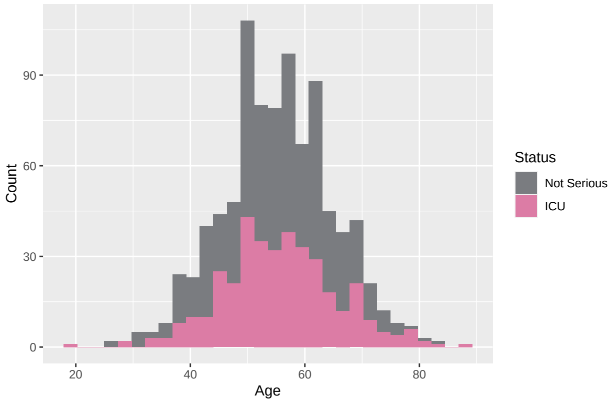
Schmid, M., Wright, M. N., & Ziegler, A. (2016). On the use of Harrell's C for clinical risk prediction via random survival forests. Expert Systems with Applications, 63, 450–459. https://doi.org/10.1016/j.eswa.2016.07.018

Appendix

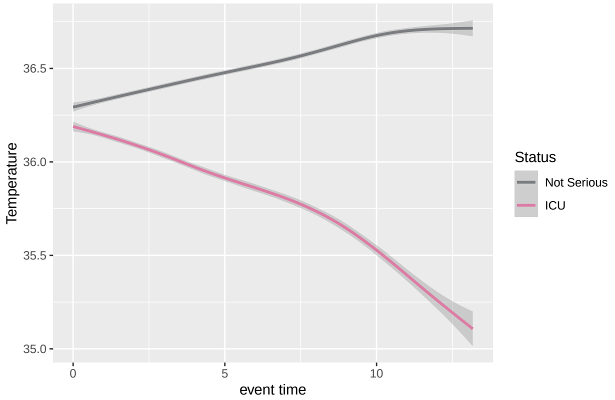
Appendix a: ICU Admission/Discharges Overtime



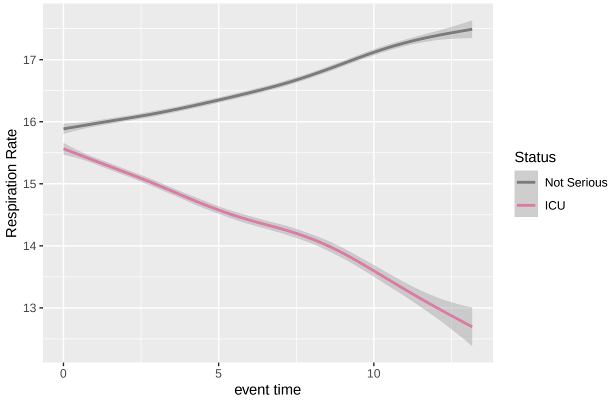
Appendix d: Age vs Status



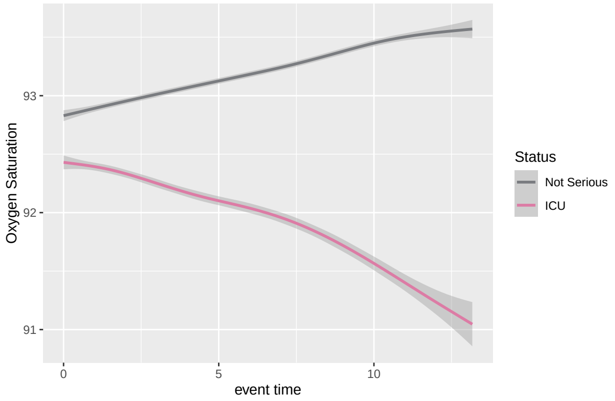
Appendix b: Temperature Changes over Time



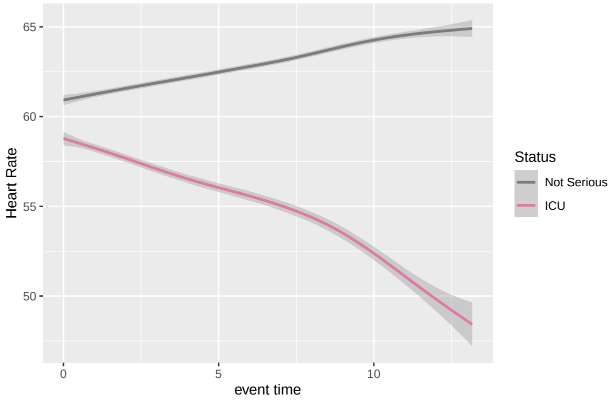
Appendix c: Respiration Rate over Time



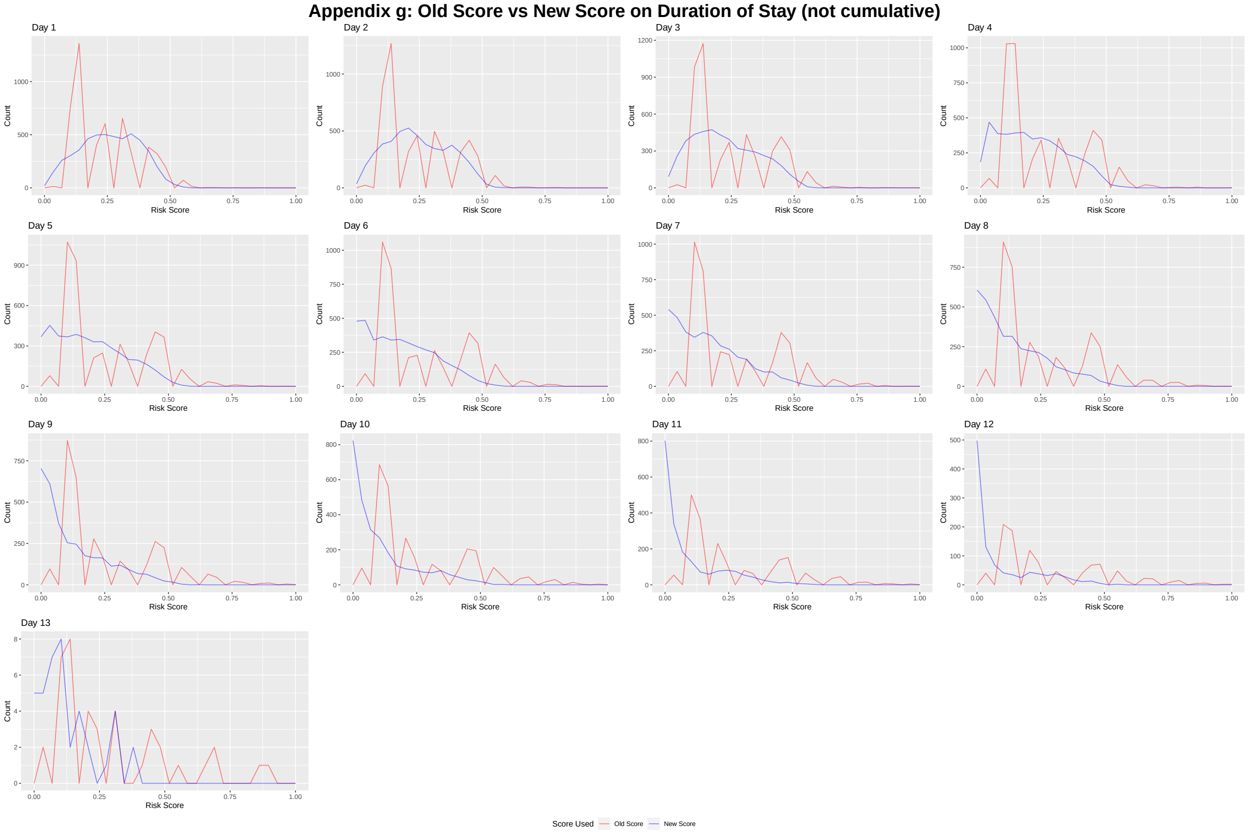
Appendix e: Oxygen Saturation over Time



Appendix f: Heart Rate over Time



Appendix g: Old Score vs New Score on Duration of Stay (not cumulative)


Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Day 7
Day 8
Day 9
Day 10
Day 11
Day 12
Day 13
Count
Risk Score
Score Used
Old Score
New Score