Class	4
Full Name	Teresa Zhang Han Yu
Matriculation Number	U2022886C

Declaration of Academic Integrity

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

[X] I have read and accept the above.

Table of Contents

Answer to Q1:

1(a)

```
#1a.
#import csv dataset
data1<-fread("data01.csv")

#check the class types of all the variables
for(var in colnames(data1)){
  print(var)
  print(class(data1[[var]]))
}

#all the variables with class type integer are categorical except for ID, age and EF
#make the class types of all those variables categorical
for(var in colnames(data1)){
  if(class(data1[[var]]) == "integer" & var != "ID" & var != "age" & var != "EF")
    data1[[var]] <- as.factor(data1[[var]])
}
```

1(b)

The Derivation group is the trainset where machine learning models are trained on, and the Validation group is the testset to test the model to ensure that there is no overfitting.

The Derivation group and Validation group are reflected in the group variable. group = 1 is the derivation group, and group = 2 is the validation group.

1(c)

	Variable.Name	Data.Type	NA.Count
1	outcome	factor	1
2	BMI	numeric	215
3	heart rate	numeric	13
4	Systolic blood pressure	numeric	16
5	Diastolic blood pressure	numeric	16
6	Respiratory rate	numeric	13
7	temperature	numeric	19
8	SP O2	numeric	13
9	Urine output	numeric	36
10	Neutrophils	numeric	144
11	Basophils	numeric	259
12	Lymphocyte	numeric	145
13	PT	numeric	20
14	INR	numeric	20
15	Creatine kinase	numeric	165
16	glucose	numeric	18
17	Blood calcium	numeric	1
18	PH	numeric	292
19	Lactic acid	numeric	229
20	PCO2	numeric	294

Figure 1.1: Table of missing value counts

1(d)

Data1 is separated into continuous and categorical data for data exploration (exclude group and ID).

Interesting finding 1:

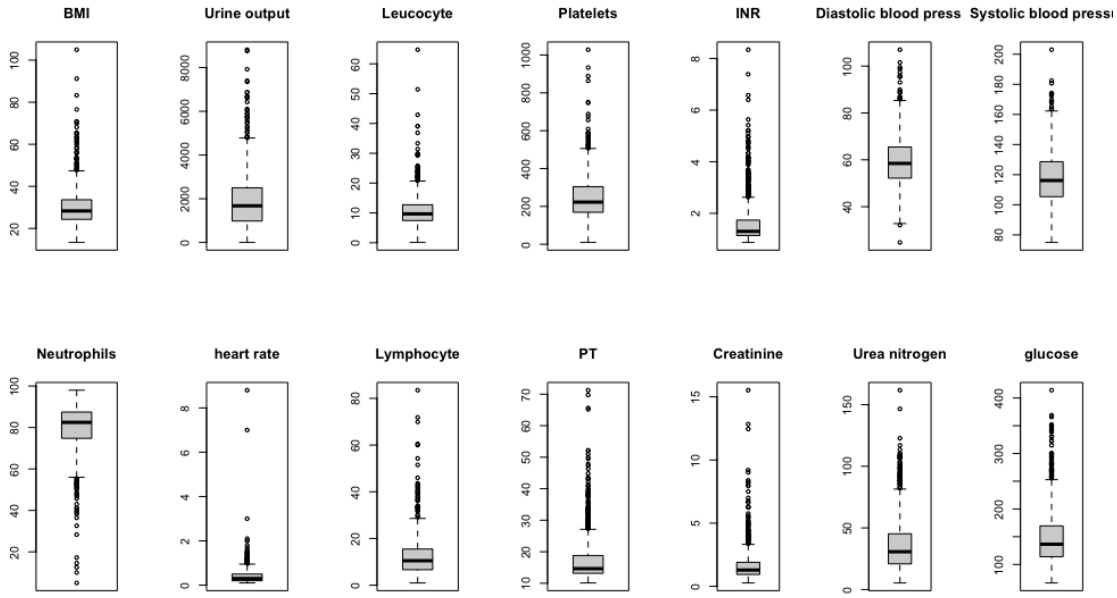


Figure 1.2: Boxplots for a few continuous variables in data1

Since there are so many continuous variables, a few variables that have extreme outliers (look at the summary) are chosen for the boxplot.

A normal diastolic blood pressure is less than 80mmHg, while a normal systolic blood pressure is less than 129mmHg. Anything higher than that will result in high blood pressure or even hypertensive crisis in worse scenarios (Appendix B). As seen from the figure above, there are several ICU patients with diastolic blood pressure above 80mmHg and systolic blood pressure above 129mmHg. The number is so significant that even the upper limit of the boxplot for diastolic blood pressure is above 80mmHg. It is even worse for systolic blood pressure, with the upper limit significantly above 129mmHg.

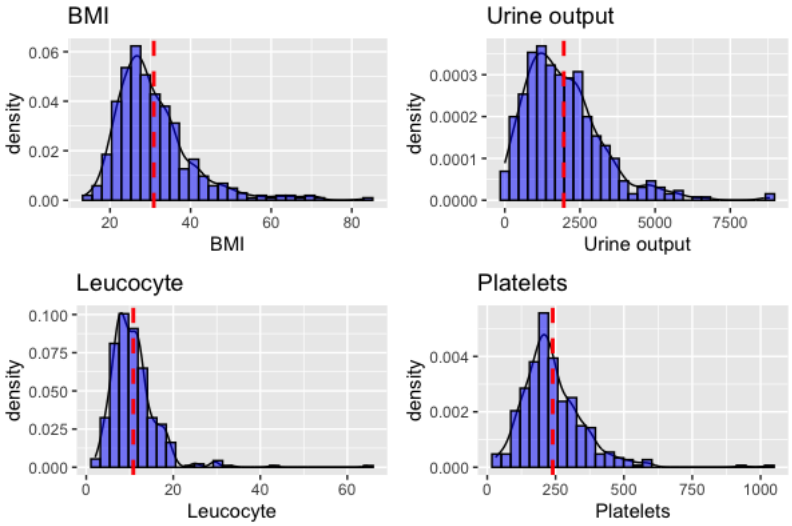


Figure 1.3: Histogram for a few continuous variables in data1

Most of the histograms of the continuous variables show that they are generally normally distributed with a positive skew due to extreme outliers with higher values. Some might have a negative skew like Neutrophils due to extreme outliers with lower values.

The boxplots and histograms suggest that there are a lot of outliers for the continuous variables, which shows that many ICU patients have extreme values for the clinical data.

Interesting finding 2:

Three variables are chosen to do a bivariate exploration with outcome to gauge their significance as a predictor for outcome. BMI is chosen as both GWTG and Nomogram did not find it significant to include, Urea nitrogen is chosen as both GWTG and Nomogram find it significant to include, and heart rate is chosen as most risk scoring systems, including GWTG included it, but Nomogram found it to be insignificant.

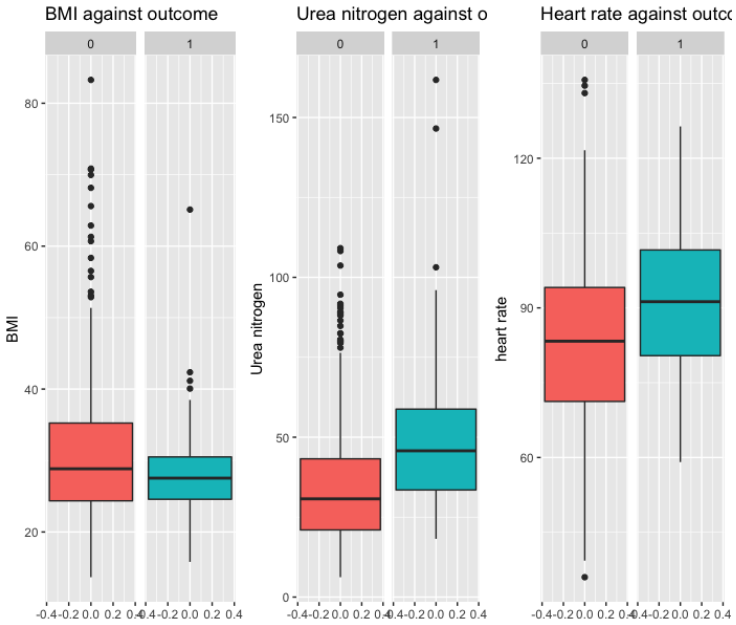


Figure 1.4: Facetted boxplot for BMI, Urea nitrogen and heart rate against outcome

From the boxplot of the continuous variables facetted by outcome, it is not clear if a higher or lower BMI will result in an outcome = 1, suggesting that it is an insignificant predictor for outcome, thus not included in both GWTG and Nomogram. Both boxplots for Urea nitrogen and heart rate show that having higher Urea nitrogen and heart rate will generally result in an outcome = 1, with the results shown in the boxplot for Urea nitrogen being clearer.

Interesting finding 3:

Univariate

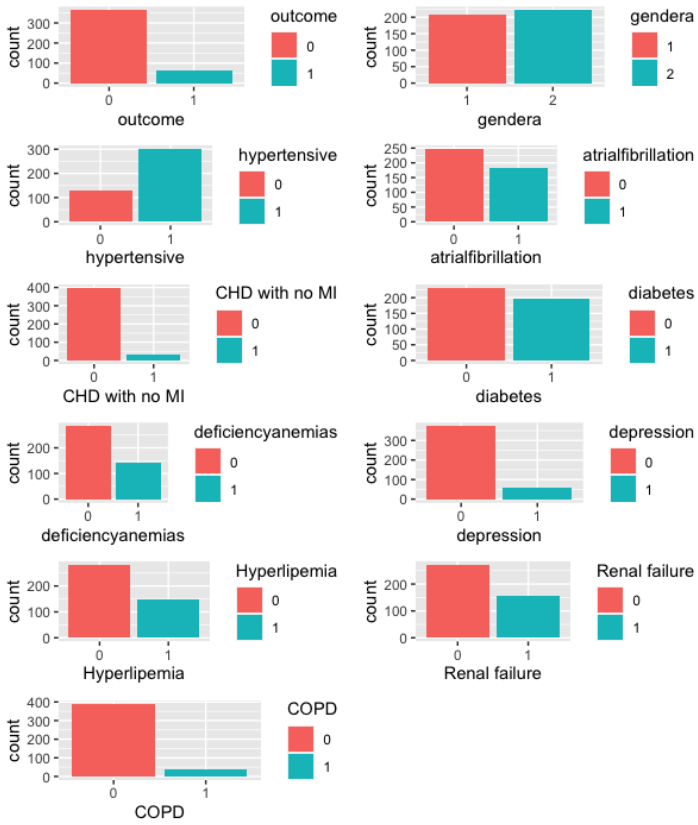


Figure 1.5: Bar charts for all categorical variables in data1

From the bar plots of all the categorical variables, CHD with no MI, depression, and COPD are generally uncommon even for ICU patients. The number of ICU patients who died are also rare, making the dataset unbalance.

Bivariate

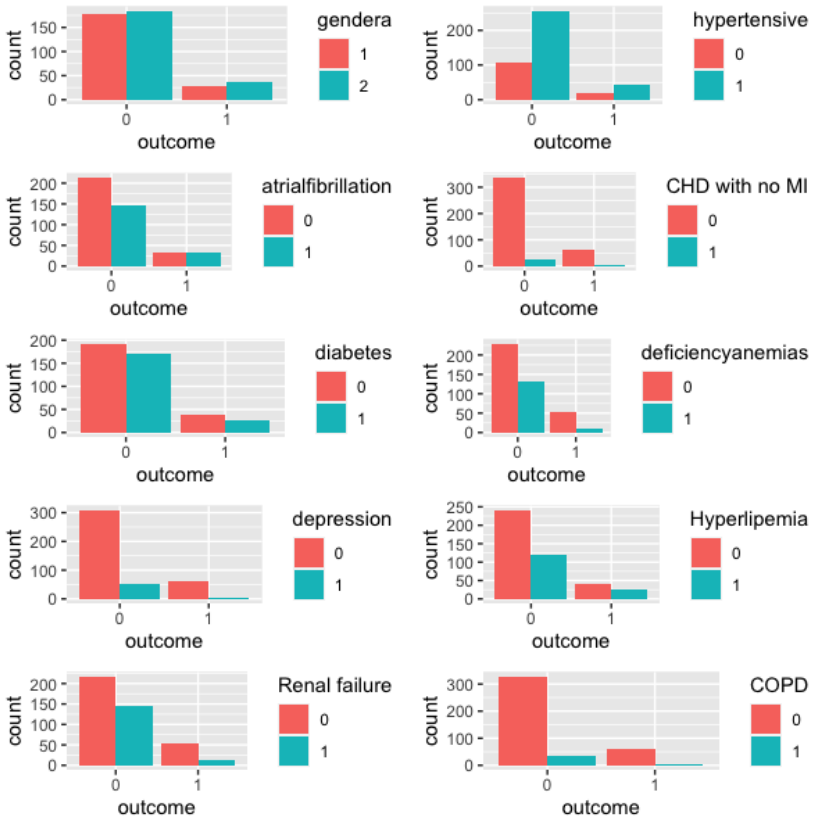


Figure 1.6: Stacked bar charts for all categorical variables with outcome

Most of the categorical variables have a categorical that has a higher number in both outcome categories because the dataset is unbalance. However, it is interesting to note that even though there are fewer ICU patients with atrial fibrillation in total, there are still more patients with atrial fibrillation that died than those without, suggesting that there is a higher chance that having atrial fibrillation will lead to death.

Answer to Q2:

1(a)

```
#copy of data1
data2 <- data1

#replace all missing values for continuous variables with median
for(var in colnames(data2)){
  if(class(data2[[var]]) != "factor"){
    data2[[var]][is.na(data2[[var]])] <- median(data2[[var]], na.rm = T)
  }
}

#replace all missing values for categorical variables with mode
mode <- function(x) {
  unique_x <- unique(x)
  tabulate_x <- tabulate(match(x, unique_x))
  unique_x[tabulate_x == max(tabulate_x)]
}
for(var in colnames(data2)){
  if(class(data2[[var]]) == "factor"){
    data2[[var]][is.na(data2[[var]])] <- mode(data2[[var]])
  }
}
```

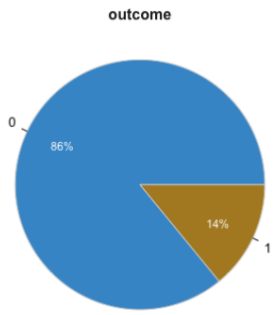
```
> #check that data2 has no missing value
> sum(is.na(data2))
[1] 0
```

1(b)

```
#produce a trainset
trainset <- data2[group == 1]

#remove "group" and "ID"
trainset[, c("group", "ID"):=NULL]

#show proportion of died vs alive
PieChart(outcome, data = trainset, hole = 0)
```

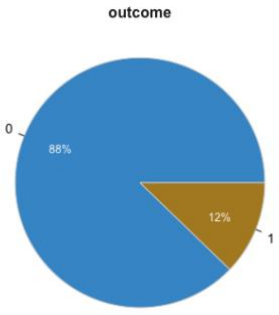


1(c)

```
#produce a testset
testset <- data2[group == 2]

#remove "group" and "ID"
testset[, c("group", "ID"):=NULL]

#show proportion of died vs alive
PieChart(outcome, data = testset, hole = 0)
```



Answer to Q3:

Systolic BP	Points	BUN	Points	Sodium	Points	Age	Points
50-59	28	≤9	0	≤130	4	≤19	0
60-69	26	10-19	2	131	3	20-29	3
70-79	24	20-29	4	132	3	30-39	6
80-89	23	30-39	6	133	3	40-49	8
90-99	21	40-49	8	134	2	50-59	11
100-109	19	50-59	9	135	2	60-69	14
110-119	17	60-69	11	136	2	70-79	17
120-129	15	70-79	13	137	1	80-89	19
130-139	13	80-89	15	138	1	90-99	22
140-149	11	90-99	17	≥139	0	100-109	25
150-159	9	100-109	19			≥110	28
160-169	8	110-119	21				
170-179	6	120-129	23				
180-189	4	130-139	25				
190-199	2	140-149	27				
≥200	0	≥150	28				

Heart Rate	Points	Black Race	Points	COPD	Points
≤79	0	Yes	0	Yes	2
80-84	1	No	3	No	0
85-89	3				
90-94	4				
95-99	5				
100-104	6				
≥105	8				

Total Score	Probability of Death
0-33	**<1%**
34-50	**1-5%**
51-57	**>5-10%**
58-61	**>10-15%**
62-65	**>15-20%**
66-70	**>20-30%**
71-74	**>30-40%**
75-78	**>40-50%**
≥79	**>50%**

Figure 3.1: GWTG-HF risk score

- In Peterson P.N. (2009) research paper, under predictors of mortality, age, admission systolic blood pressure, admission BUN, admission serum sodium, admission heart rate, nonblack race, and the presence of COPD were the chosen independent predictors of in-hospital death for GWTG, thus these variables from the trainset are included in gwtg_train, including the outcome variable.
- Using the cut function, the variables are segmented according to the range given in the diagram above (using the parameter, breaks), their respective points are assigned (using the parameter, labels) to each variable. The parameter include.lowest is set to TRUE, to include the lower bound (R Coder, 2020).
- The ICU patients in the dataset are assumed to be non-black, therefore 3 points are assigned to all race values.
- Make all the points assigned to the predictors numeric and tabulate the total risk score by adding up all of them.
- For total risk score more than or equal to 79, the probability of death is more than 50%, thus the predicted outcome is 1, for total risk score less than 79, the predicted outcome is 0.

Assumptions:

- BP values between 59 and 60 (not inclusive of 60) will be included in BP values between 50 to 59 and will be assigned 28 points. This will be the same for other such values and variables.

Answer to Q4:

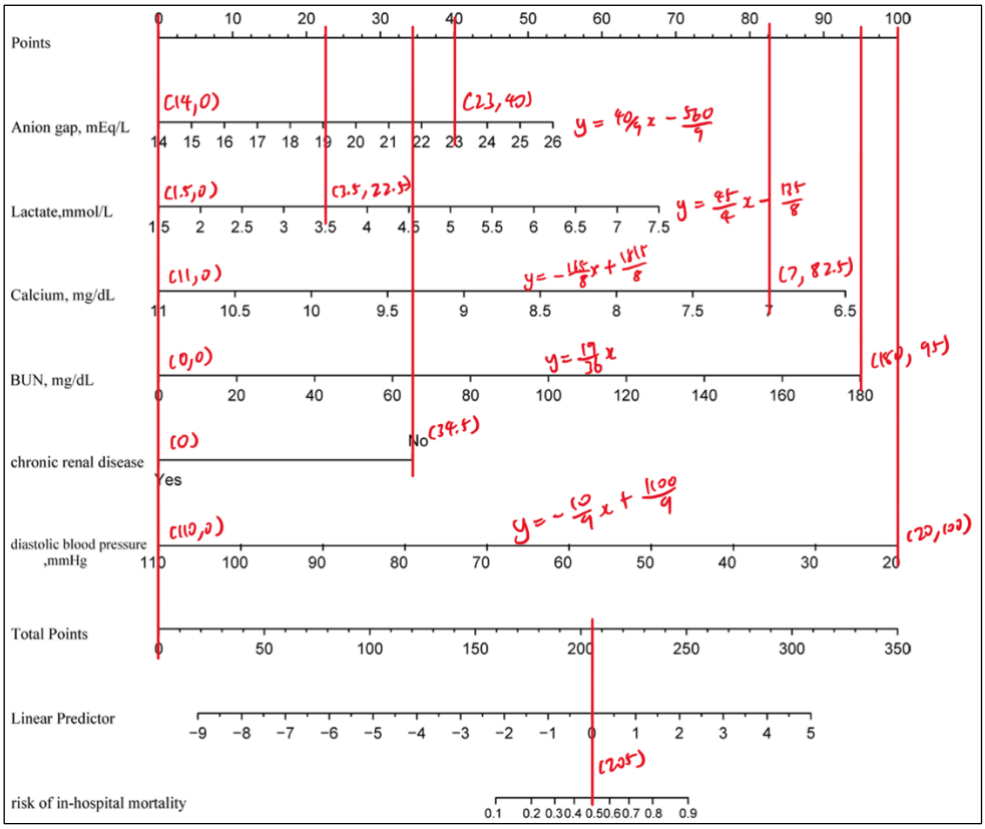


Figure 3.2:Developed risk of in-hospital mortality nomogram

- In Li F., et al (2021) research paper, the chosen predictors for mortality for Nomogram are Anion gap, Lactate, Calcium, chronic renal disease, diastolic blood pressure, thus these variables from the trainset are included in nomo_train, including the outcome variable.
- The nomogram was based on proportionally converting each regression coefficient in multivariate logistic regression to a 0–100-point scale, thus to assign the points, generate an equation for each variable based on the diagram above.
- Make all the points for the predictors numeric and tabulate the total risk score by adding up all of them.
- For total risk score more than or equal to 205, the risk of in-hospital mortality is more than 50%, thus the predicted outcome is 1, for total risk score less than 205, the predicted outcome is 0.

Assumptions:

- The straight lines drawn in the diagram above are generally accurate.
- If values fall outside the line, the line is extrapolated, since the scale is linear.

Answer to Q5:

Model	FPR	FNR	Err
Logistic Reg(BE)	0.0225670	0.6034483	0.1042424
Random Forest	0.0000000	0.0000000	0.0000000
GWTG	0.0000000	1.0000000	0.1406061
Nomogram	0.0112835	0.7413793	0.1139394

Figure 5: Table of trainset errors

As seen from Figure 5, Random Forest has the most accurate model with zero false positive and false negative (all trainset values are predicted correctly). The GWTG model predicts that no ICU patients will die, thus having zero false positives and all false negatives, hence not very accurate with the highest overall error. The Logistic Regression model with Backward Elimination is the second most accurate with the second lowest overall error, followed closely by the Nomogram model. Logistic Regression has more false positives while Nomogram has more false negatives.

Answer to Q6:

Model	FPR	FNR	Err
Logistic Reg(BE)	0.03883495	0.6744186	0.1164773
Random Forest	0.00000000	0.9069767	0.1107955
GWTG	0.00000000	1.0000000	0.1221591
Nomogram	0.01618123	0.7441860	0.1051136

Figure 6: Table of testset errors

All the models did generally well for the testset with a small overall error. For the testset, Random Forest has a high false negative rate even though there was no error for the trainset, suggesting that the model might have been slightly overfitted to the trainset, it might also be due to the unbalanced trainset. Similar to the predictions in the trainset, the GWTG model predicts that no ICU patients will die, thus having zero false positives and all false negatives, hence not very accurate with the highest overall error. The Logistic Regression model with Backward Elimination has the most false positives but the least false negatives. The Nomogram model is the most accurate model with the smallest overall error.

Answer to Q7:

Model	FPR	FNR	Err
Logistic Reg(BE)	0.3042071	0.3255814	0.3068182
Random Forest	0.2880259	0.2325581	0.2812500

Figure 7: Table for testset errors for models trained on a balanced trainset

After balancing the trainset, the Random Forest model performs better on the testset than the Logistic Regression model with a lower value for all three errors. It is important to note that even though the false positive rate and overall errors increased significantly, the false negative rate also decreased significantly. To achieve a more accurate model for a small dataset, it is important to balance the trainset so the result will not be skewed to one outcome, in this case, towards outcome = 0.

For an unbalanced trainset, frequent cases will get fewer misclassifications (Stack Exchange, 2018), while the seldom cases will get more misclassifications, as shown in Table 6, where there is a much higher false negative rate compared to the false positive rate as the trainset is dominated with negatives. However, ultimately, the decision to balance the trainset, to get fewer misclassifications for seldom but more misclassifications for frequent cases, depends on the context.

Answer to Q8:

Model	FPR	FNR	Err
Logistic Reg(BE)	0.3042071	0.3255814	0.3068182
Random Forest	0.2880259	0.2325581	0.2812500
RF VarImpt into Logistic Reg	0.3236246	0.3953488	0.3323864

Figure 8: Table for testset errors for models trained on a balanced trainset including RF VarImpt into Logistic Regression

This model is inferior to both the stand-alone Logistic Regression with backward elimination and Random Forest, with the highest values for all three errors. The Random Forest model still performs the best with the lowest values for all three errors.

Answer to Q9:

GWTG

The GWTG-HF risk score uses commonly available clinical variables to predict in-hospital mortality and provides clinicians with a validated tool for risk stratification that is applicable to a broad spectrum of patients with heart failure, including those with preserved left ventricular systolic function (Peterson, 2009).

Nomogram

The nomogram was formulated based on the results of multivariate logistic regression analysis. Multivariate logistic regression analysis was used to build prediction models. The best performing model, XGBoost, was chosen to build a nomogram. The nomogram was based on proportionally converting each regression coefficient in multivariate logistic regression to a 0–100-point scale (Li, 2021).

GWTG & Nomogram vs. Other Risk Scoring System

The Acute Decompensated Heart Failure National Registry (ADHERE) model is appealing because it uses only 3 variables to classify patients as low, intermediate, or high risk. However, it does not allow more precise characterization of individual risk, and it does not include all variables that significantly inform outcomes (Li, 2009), unlike GWTG and Nomogram.

A prognostic model for in-hospital mortality from OPTIMIZE-HF was recently published with some overlapping variables but did not have a separate derivation and validation cohort and did not include data on admission BUN (Li, 2009).

Thus, GWTG and Nomogram is a better risk scoring system to assess ICU patient mortality.

GWTG vs. Nomogram

The predictors both models used for outcome are different. GWTG uses Systolic blood pressure, Urea nitrogen, Blood sodium, age, heart rate, and COPD as its predictors while Nomogram uses Anion gap, Lactic acid, Blood calcium, Urea nitrogen, Renal failure, and Diastolic blood pressure. Nomogram uses the predictors from the XGBoost model, which has a considerably better predictive effectiveness than the GWTG-HF risk score model, suggesting that the predictors used in Nomogram are better than those used in GWTG.

However, looking at Figure 1.1, all of the variables that GWTG uses have very few NA values, while Lactic acid, a variable used by Nomogram has 229 NA values. The use of available clinical information makes the GWTG-HF score less susceptible to missing data (Peterson, 2021). However, in the context of technologically advanced Singapore, collecting data like Lactic acid will be less of a problem. Thus, this issue can be mitigated for hospitals in Singapore.

For GWTG, a weighted integer was assigned to each independent predictor, based on the predictor's coefficient in the reduced regression model, while the nomogram was based on proportionally converting each regression coefficient in multivariate logistic regression to a 0–100-point scale.

Assigning points using a continuous scale is more accurate than segmenting the values into different groups and assigning the same points to values that belong to a certain range. Categorization assumes that the relationship between the predictor and the response is flat within intervals; this assumption is far less reasonable than a linearity assumption in most cases (Frank, 2020) and results in a loss in precision.

As such, because of better predictors used, and a more accurate method of assigning points to variables, Nomogram is a more accurate model as compared to GWTG. This can be proven from Figure 6. Nomogram has a lower false negative rate and a lower overall error as compared to GWTG.

Therefore, I will recommend Nomogram as a risk scoring system to assess ICU patient mortality to a hospital in Singapore.

Answer to Q10:

Limitations of Nomogram and Solutions

Nomogram has some limitations. First, the data extracted from Medical Information Mart for Intensive Care (MIMIC) database is spread across several years (2001–2012), during which the treatment of heart failure had changed greatly, which may weaken the application of the model (Li, 2021). Machine learning models often become stale over time for various reasons, and thus models that might have been a good match for the task in the past may stop being relevant (Itay, 2021). Therefore, it is essential to retrain the Nomogram model whenever a significant amount of additional data becomes available or there is a breakthrough in the medical field, to ensure optimal performance without excessive overhead, thus improving the accuracy of the model.

Second, MIMIC is a database of de-identified electronic health records (EHR) associated with patients who stayed in intensive care units (ICU) at the Beth Israel Deaconess Medical Centre in Boston (Front., 2021). Thus, as a single-centre study, the population was relatively small. Although the robustness of Nomogram was tested extensively with internal validation using bootstrap testing, it remains uncertain whether the results of this study can be applied to other populations (Li, 2021). To improve the accuracy of Nomogram for a hospital in Singapore, it should be trained on data provided by hospitals in Singapore instead as the results might differ due to differing factors such as technology.

Additional improvements that can be made to Nomogram

In the case of predicting ICU patient mortality, false negatives are worse than false positives. If the outcome of an ICU patient is supposed to be predicted as 1 is predicted as 0, preventive measures will not be taken by the hospital, thus resulting in the predicted outcome being true. Thus, to get a better Nomogram model, the false negative rate should not be as high as the values shown in the table for question 6. As proven and explained in question 7, for an unbalanced trainset, frequent cases will get fewer misclassifications, while the seldom cases will get more misclassifications, thus, to decrease the misclassification seldom cases, in this case, the ICU patients who died, Nomogram should be trained on a balanced trainset.

However, it is important to note that the class imbalance problem is caused by there not being enough patterns belonging to the minority class. Generally, if there is enough data, the "class imbalance problem" doesn't arise (Stack Exchange, 2018). Thus, Nomogram should be trained on a significant amount of data so that there is no need to balance the trainset.

These ideas can improve the accuracy of the Nomogram model so that preventive measures can be taken for more ICU patients that are predicted to die and save their lives.

Appendix A: List of assumptions on the data or analysis

[If necessary, you may add your assumptions into the list.]

1. Race is assumed to be non-black in the data for GWTG Risk Scoring Model as race/ethnicity is not available in the data but Li F., et al (2021) shows that non-blacks is the great majority (approx. 86%).

Appendix B: Blood Pressure readings

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

(American Heart Association, 2021)

Reference:

APTA. (2021, April). *ADULT VITAL SIGN INTERPRETATION IN ACUTE CARE GUIDE 2021.* Retrieved From.

https://cardiopt.memberclicks.net/assets/docs/CPG/Joint%20Vital%20Sign%20Booklet.pdf

R Coder. (2020, June 26). *Cut in R.* Retrieved From.

https://r-coder.com/cut-r/

Frank Harrell. (2020, June). *Categorizing Continuous Variables.* Retrieved From.

https://discourse.datamethods.org/t/categorizing-continuous-variables/3402

Front. Artif. Intell. (2021, May 31). *Medical Information Mart for Intensive Care: A Foundation for the Fusion of Artificial Intelligence and Real-World Data.* Retrieved From.

https://www.frontiersin.org/articles/10.3389/frai.2021.691626/full

Itay Gabbay. (2021, July 14). *How Often Should You Retrain Your Machine Learning Model?* Retrieved From.

https://deepchecks.com/how-often-should-you-retrain-your-machine-learning-model/

Stack Exchange (2018, October 17). *When should I balance classes in a training data set?* Retrieved From.

https://stats.stackexchange.com/questions/227088/when-should-i-balance-classes-in-a-training-data-set