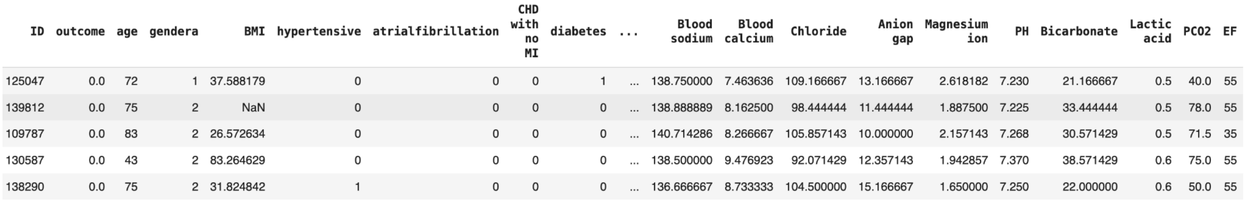
Data 102 Mortality Project

Yuanrui Zhu, Deborah Chang, Ryan Jefferson Soohoo, Haixin Guo

Section 1: Data Overview

For our research, we used the dataset "In Hospital Mortality Prediction." We found the dataset on this website. The website has a download button through which we downloaded the dataset. Figure 1 shows a preview of the data.



ID	outcome	age	gendera	BMI	hypertensive	atrialfibrillation	CHD with no MI	diabetes	...	Blood sodium	Blood calcium	Chloride	Anion gap	Magnesium ion	PH	Bicarbonate	Lactic acid	PCO2	EF
125047	0.0	72	1	37.588179	0	0	0	1	...	138.750000	7.463636	109.166667	13.166667	2.618182	7.230	21.166667	0.5	40.0	55
139812	0.0	75	2	NaN	0	0	0	0	...	138.888889	8.162500	98.444444	11.444444	1.887500	7.225	33.444444	0.5	78.0	55
109787	0.0	83	2	26.572634	0	0	0	0	...	140.714286	8.266667	105.857143	10.000000	2.157143	7.268	30.571429	0.5	71.5	35
130587	0.0	43	2	83.264629	0	0	0	0	...	138.500000	9.476923	92.071429	12.357143	1.942857	7.370	38.571429	0.6	75.0	55
138290	0.0	75	2	31.824842	1	0	0	0	...	136.666667	8.733333	104.500000	15.166667	1.650000	7.250	22.000000	0.6	50.0	55

Figure 1. Head data of In Hospital Mortality Prediction

The "In Hospital Mortality Prediction" dataset itself is derived from the Medical Information Mart for Intensive Care (MIMIC-III) database, a large public database consisting of de-identified health-related data associated with over 40,000 patients who stayed in the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The study itself is a simple study, specifically a retrospective cohort study in which researchers gathered and analyzed data from a group of participants from the MIMIC-III database — specifically 1177 heart failure patients admitted to intensive care units (ICUs) from the original database. The data extracted is a sample, not a census, since it reflects a subset of not only the MIMIC-III database but also of all the possible hospital admissions that could occur. The data was collected from this sample to develop and validate a prediction model for all-cause in-hospital mortality among ICU-admitted heart failure patients. We've noticed that this study does not aim to collect information from every heart failure patient in ICUs (census) but rather from a subset (sample) to make generalizations about the larger population.

Since the data is a sample, we set out to compare the distribution of age in the dataset to the age structure of the population. We found that age in the dataset is drastically different from that of the population since our age distribution skewed left: there's very few young people in our dataset. This significantly affects the generalizability of our result. We would limit the generalizability of our analysis specifically to the same or similar population or similar settings as this study. More specifically, we would have to generalize the result to the population of heart failure patients who were admitted to ICUs.

The dataset contains highly granular data that encompasses a range of information from demographic characteristics and vital signs to laboratory values and comorbidities, with each row representing the information and health indexes of individual patients. Since each row corresponds to a single patient, the analysis and conclusions drawn from the study will be at the individual patient level. This allows researchers to identify patterns, risk factors, and characteristics associated with in-hospital mortality for heart failure patients in the specific context of ICU admissions. These data points were captured at various times during the patients' hospital stay, with demographic characteristics and vital signs recorded within the first 24 hours of admission and laboratory variables measured throughout the ICU stay.

Given the nature of the MIMIC-III database from which this dataset was extracted, it is important to note that the data used were de-identified to protect patient privacy. It is reported that consent for the collection of data in the MIMIC-III database itself was waived because the project did not impact clinical care and all protected health information was de-identified. Therefore, it was modified for differential privacy. In the documentation of the dataset, it is recorded that the data was "deidentified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting." The process that is listed then details removal of personal identifiable data like names and phone numbers, while "shifting" data by a random offset for each individual in a consistent manner, including dates, time, and dates of birth.

In the dataset, some important features that we wished to have were financial situations and education level. We think that these could not only be confounders we need to consider when performing causal inference, but also important features when making predictions on a person's final outcome.

Some biases or problems with the dataset that could affect the conclusions have to do with the nature of the data itself. Some examples are listed below:

- Selection Bias:
    - The MIMIC-III database only includes data from one hospital in a specific location - the Beth Israel Deaconness Medical Center in Boston, USA - and therefore may not generalize to a broader, more diverse audience and/or population.
    - Since the database is derived from critical care, it may not actually encompass extraneous variables that may affect our conclusions of mortality such as access to treatment or lifestyle. Therefore, it is important to understand that conclusions drawn from this dataset are most applicable in the context of critical care, and may not be generalizable to a larger public.
- Measurement Error:
    - Measurement error is always a concern in datasets if there are inconsistencies in the recorded data or variations in the data collection methods. Although the study and the documentation of the MIMIC-III database do not state the existence of this type of error, more investigation should be conducted in the future.
- Convenience Sampling:
    - There is no evidence of convenience sampling, but the sampling method could involve selecting patients who meet certain criteria or conditions, rather than a random or systematic approach. One issue we could think of is that it doesn't include financially incapable individuals who cannot afford the expensive payment of ICU.

We noticed that there were a lot of columns with missing data in the dataset. We believe the reason such data are missing is because either the patient was not willing to share, or there was difficulty collecting data from specific patients. We decided to drop all rows with missing data after we chose our columns of interest.

In our cleaning/pre-processing, we dropped outliers from our columns of interest and we transferred categorical data to ordinal numerical numbers. We believed that doing so could improve the quality of our data analysis and machine learning. Specifically when we performed GLM and non-parametric modeling, we did oversampling and train-test-split to make sure we have balanced data with respect to our outcome and that our models are robust.

Section 2: Research Questions

Our research question #1 is: Under the mortality dataset setting, does having diabetes cause a higher BMI? This question can be applicable to real-world decisions on designing treatment and/or intervention strategies that target specifically individuals with diabetes. For example, this may involve directing more resources to be allocated towards diabetes management programs, counseling services for patients, or overall improving more effective healthcare solutions.

For this research question, we used causal inference because it allows us to establish a causal relationship between diabetes and BMI by controlling the confounders under an established, rigid framework of outcome regression and inverse propensity weighting. We plan to use outcome regression and Inverse propensity weighting (IPW) to calculate the average treatment effect (ATE) of diabetes on BMI controlling for confounders. We also tried to implement matching and approximate matching algorithms, but since we have so many variables to consider, it was hard for us to find individuals with similar conditions on the health indexes we controlled. For outcome regression, the limitation is that it relies on an assumption of linearity between covariates and the outcome. If the model assumption is violated, the model may not accurately capture the true relationship and leads to a biased estimate. Furthermore, if important confounders are omitted from the outcome regression model, the estimated treatment effect may be confounded and the result could be biased. For inverse propensity weighting, it relies on a positive probability of receiving each treatment level for every combination of covariates, the lack of outliers in the covariates, and a low variance on weight. Violating any criteria specified above could lead to the model generating biased results.

Our research question #2 is: Under the mortality dataset setting, what's the overall accuracy of using GLM and non-parametric methods to predict the outcome? By answering this question, we can have early intervention using the predicted outcome to identify individuals at higher risk of mortality and allocate more resources beforehand.

Both logistic regression and random forest fits well for our target question on predicting a binary outcome based on several explanatory variables. Logistic regression has a good interpretability and is good for statistical inference as a next step (using hypothesis testing and confidence interval estimation). Random forest has good flexibility and is able to handle non-linearity and outliers among explanatory variables. It’s also more robust to overfitting since it makes decisions by combining the result of multiple decision trees. Getting the accuracy for both models could inform future modeling practices on similar datasets. There are some limitations of the method we chose. For logistic regression, it assumes a linear relationship between predictors and the log-odds of the response variable, and it also assumes an independence of all observations. Similar to other GLMs, logistic regression can suffer from overfitting and underfitting depending on the model complexity and the data it fits. For random forest, it is harder to interpret than logistic regression and has a higher computational cost. In both cases, we need to ensure that data is of high quality and we don't have missing values and outliers; if not, we may not be able to do well on answering this research question.

Section 3: Exploratory Data Analysis

Figure 2 compares the BMI across two groups - those with diabetes and those without. This visualization is critical because it motivates asking the causal question of whether a higher BMI leads to the diagnoses of diabetes. From our histogram, we can observe there exists a difference between the distribution of BMI for diabetes and non-diabetes group, with the diabetes group generally having higher BMI than non-diabetes group. This trend suggests a potential causal influence of BMI on diabetes.

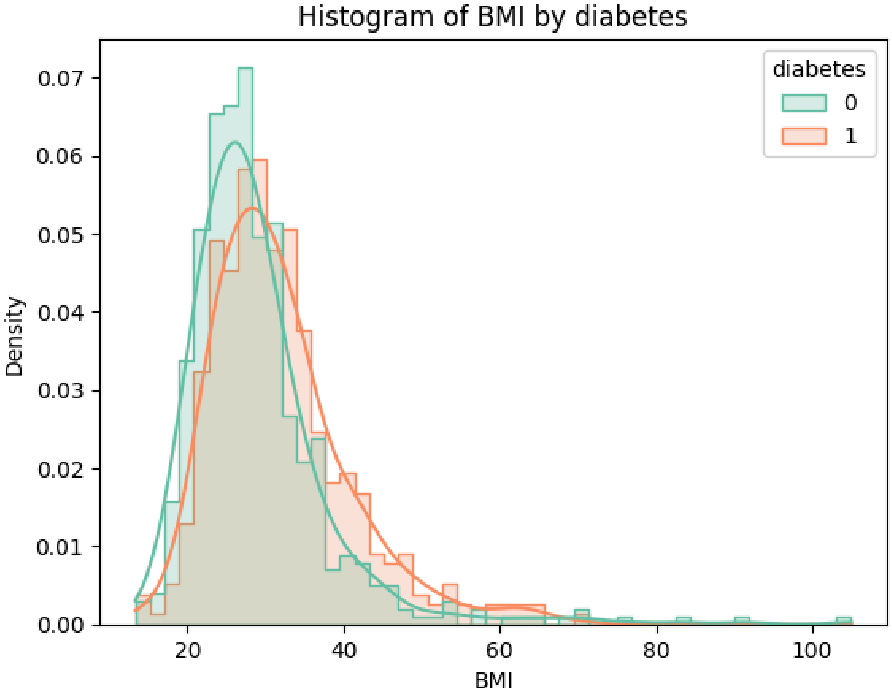


Fig. 2 Histogram of BMI by diabetes

We built Figure 3 and 4 to try and identify potential confounders in the causal relationship specified above. We hypothesize that gender could be a possible confounder that affects both Diabetes and BMI, so we plotted gender in relation to Diabetes and BMI respectively. From Figure 3, we found that there is a higher prevalence of diabetes among males compared to females and a lower incidence of males without diabetes, which indicates a potential association between gender and diabetes. Moreover, the histogram in Figure 4 demonstrates a higher density of younger males than females, which possibly indicates an association between gender and BMI as well.

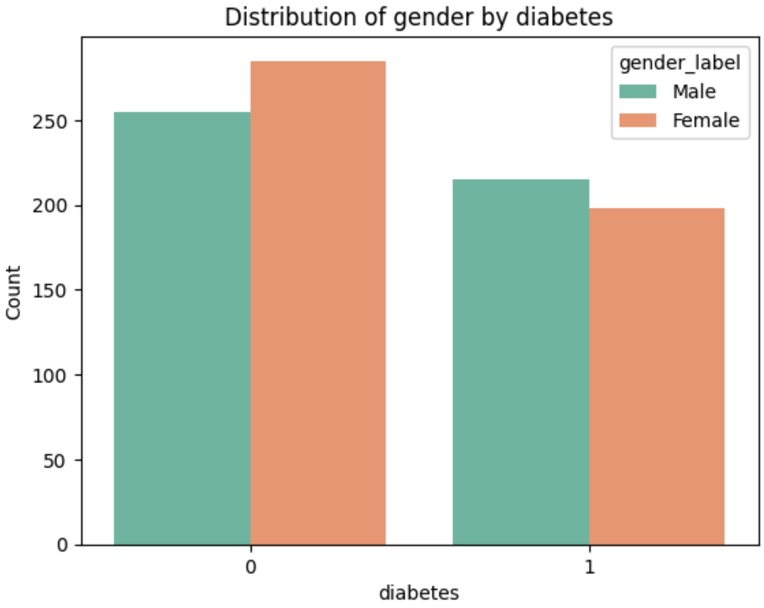


Fig. 3 Distribution of gender by diabetes

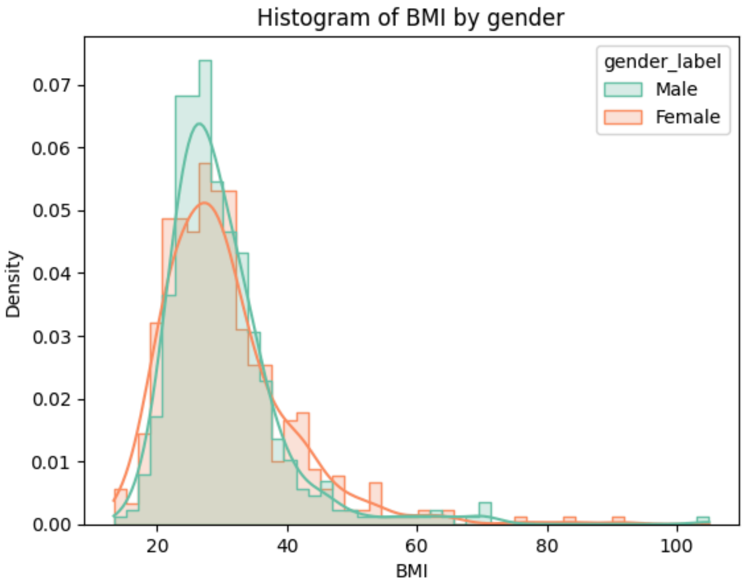


Fig. 4 Histogram of BMI by gender

Figure 5 compares the ages of individuals who passed away and those who did not. This is because we are interested in using multiple linear regression on predicting mortality based on other variables. We plotted the histogram of age separated by outcome, and we found that a higher density of older individuals for those who died compared to those who survived. This indicates a potential predictive power of age as a variable within the logistic regression.

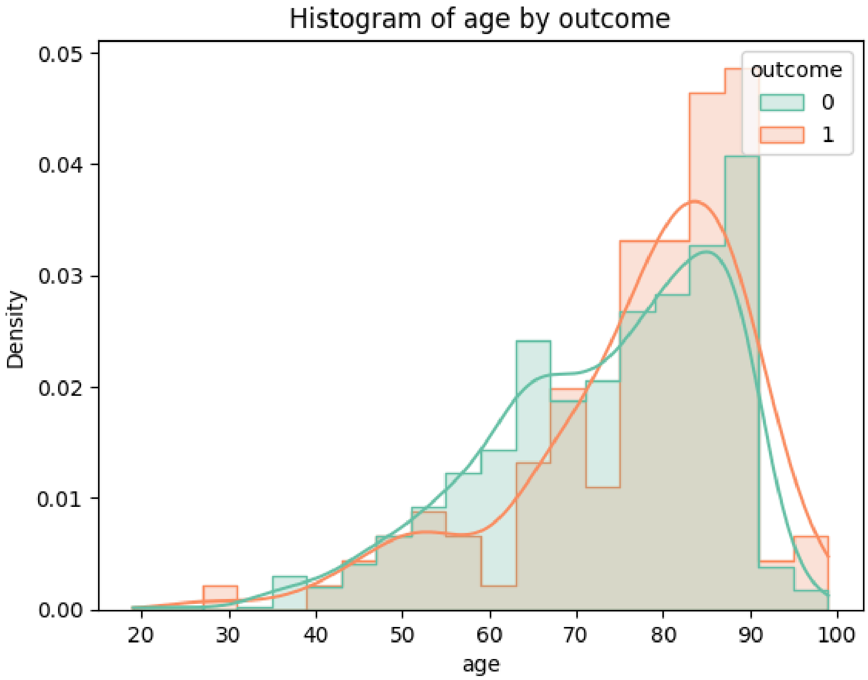


Fig. 5 Histogram of age by outcome

Additionally, we need to confirm that the class labels are balanced within our dataset. This ensures the accuracy of our model and we minimize the bias within our prediction. In Figure 6, we do see that people who died (about 100 in total) are much fewer than the people who survived (about 800 in total), which indicates that techniques such as resampling would be useful.

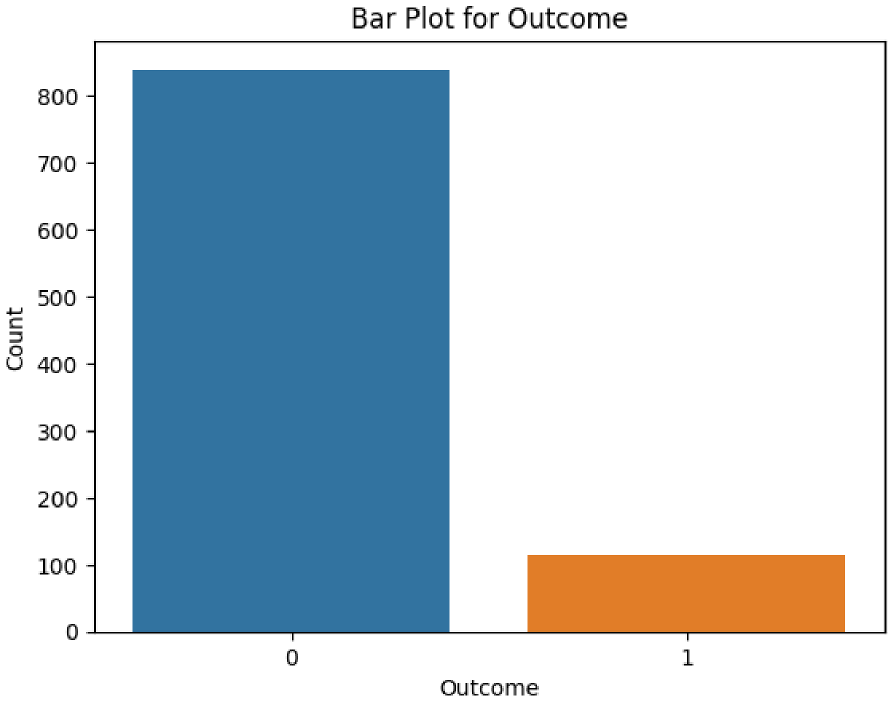


Fig. 6 Bar Plot for Outcome

Section 4: Prediction with GLMs and nonparametric methods

Methods

For this subpart, we want to predict the final outcome (outcome = 0 or outcome = 1) from the explanatory variables from our dataset. We firstly kept the features which are commonly known by people for explanatory purposes. Then we ran logistic regression on every combination of features and recorded their AIC and BIC score. We finally chose the features with the lowest AIC score, which are 'age', 'Systolic blood pressure', 'Respiratory rate', 'glucose', 'depression', 'Hyperlipemia', and 'Renal failure'.

We used logistic regression for our GLM. The reason for that is because we're performing binary classification and we can get a class weight for each relevant feature we are interested in. In terms of the assumptions, we assume (1) that our observations are independent which is true from the data collection mechanisms observed, (2) that we have a large sample size without outliers, and (3) that the relationship between the log-odds of the outcome and each explanatory variable is assumed to be linear which is true according to the partial regression plot (Figure 7).

We used accuracy scores to evaluate our model’s performance. The accuracy score was calculated based on our validation set which was splitted and left out from our entire dataset before training the model.

We used Random Forest for our nonparametric method. The reasons lie upon the several merits of Random Forest, including its robustness to overfitting and outliers and its ability to handle non-linear relationships between variables and outcome. There are no specific assumptions we made before choosing this model. Also, we splitted the whole dataset into training and test sets according to 0.8-0.2 proportion and oversampled the minority class (outcome = 1). The accuracy score was calculated based on our validation set which was splitted and left out from our entire dataset before training the model.

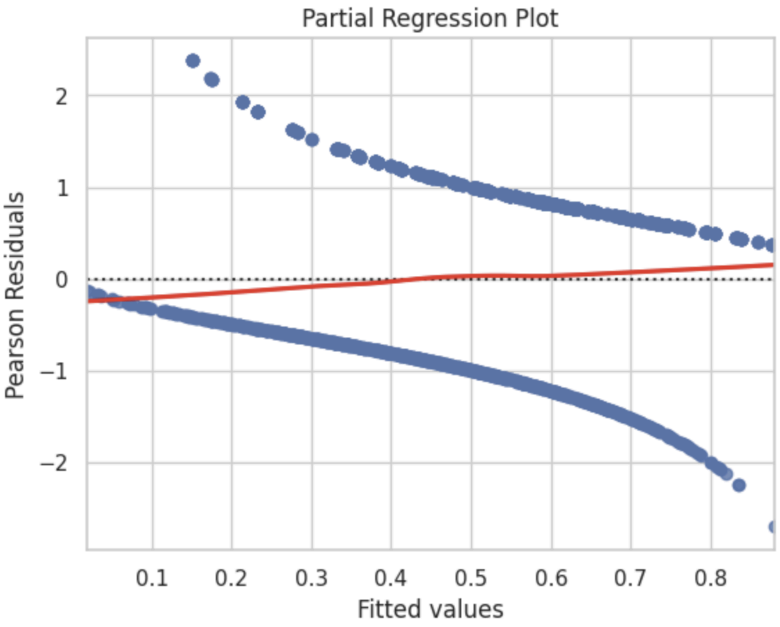


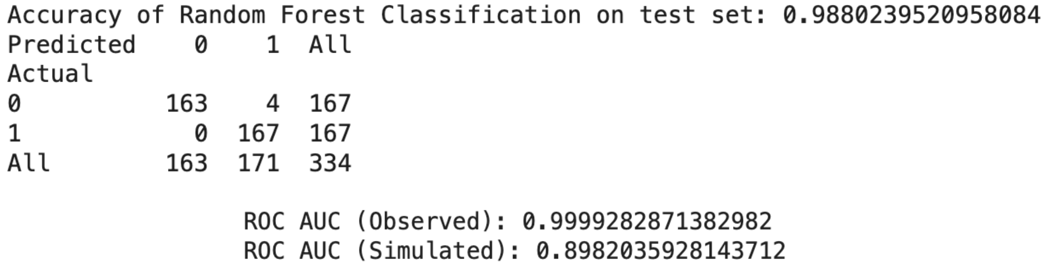
Figure 7: Partial Regression Plot between all predictors and outcome

For logistic regression, it assumes a linear relationship between predictors and the log-odds of the response variable, and it also assumes an independence of all observations. Similar to other GLMs, logistic regression can suffer from overfitting and underfitting depending on the model complexity and the data it fits. For random forest, it is harder to interpret than logistic regression and has a higher computational cost. In both cases, we need to ensure that data is of high quality and lack of missing values and outliers; if not, we may not be able to do well on answering this research question.

Results

Random Forest achieved an overall high accuracy of around 98.8%; for logistic regression, it achieved an overall moderate accuracy of around 67.1%. The low accuracy of the logistic regression model indicates that it may not have captured the underlying patterns within the dataset as good as Random Forest.

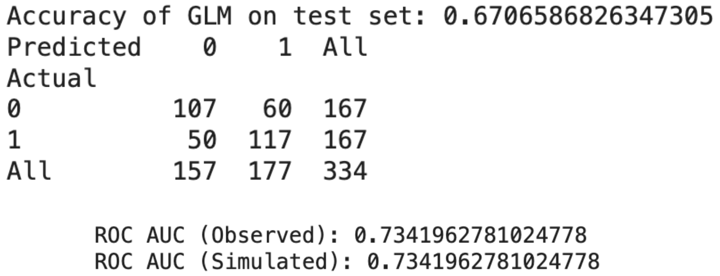
Since we have binary outcome for our prediction task, we chose ROC AUC as a measure of fit for binary classification tasks instead of comparing histograms for posterior predictive distributions as we did in the class. For the logistic regression model, we achieved both observed and simulated ROC AUC of around 0.7342, which indicates a moderate discriminative ability. Since the two AUC is consistent, it suggests that the logistic regression model is reasonably calibrated. For the random forest model, we achieved an observed ROC AUC of 0.99 and a simulated ROC AUC of 0.89. The nearly perfect observed ROC AUC may suggest an overfitting issue with Random Forest, but a simulated ROC AUC of 0.89 could indicate that the model is more realistic when accounting for uncertainty. While the discrepancy between observed and simulated AUC values suggest some uncertainty in our model's predictions, given the random forest is generally unlikely to overfit and the high validation accuracy we got earlier, we conclude that our random forest model has a strong discriminative ability.



```
Accuracy of Random Forest Classification on test set: 0.9880239520958084
Predicted    0    1  All
Actual
0          163    4  167
1            0  167  167
All        163  171  334

                 ROC AUC (Observed): 0.9999282871382982
                 ROC AUC (Simulated): 0.8982035928143712
```

Figure 8: Results of Random Forest



```
Accuracy of GLM on test set: 0.6706586826347305
Predicted    0    1  All
Actual
0          107   60  167
1           50  117  167
All        157  177  334

      ROC AUC (Observed): 0.7341962781024778
      ROC AUC (Simulated): 0.7341962781024778
```

Figure 9: Results of Logistic Regression

Discussion

Random forest performs better than logistic regression as it achieves nearly perfect accuracy on our validation set. However, we are moderately confident to apply this to future datasets and we need to account for the complexity of datasets and adjust our model correspondingly to make the best of our model.

Our results show random forest fits our data better. Logistic regression and random forest have their strengths and trade-offs. Logistic regression provides interpretability but may struggle with complex non-linear relationships. Random Forest excels in capturing non-linear patterns but requires careful consideration of overfitting. Regular monitoring, validation on new data, and adherence to model assumptions are essential for both models.

Other than results included in the previous sections, we also computed the precision, recall, and F1 Scores for both outcomes for both models. Precision reflects the accuracy of positive predictions, while recall indicates the ability to capture positive instances. F1 score provides a balance between precision and recall. In this context, the Random Forest model outperforms the GLM model in terms of precision, recall, and F1 score.

Incorporating additional data could provide relevant and informative features. For example, time-series data related to people's health outcome or physiological measurements can capture trends over time, and environmental factors such as air quality and pollutants could also provide a more holistic understanding of health risks associated with patients.

We observed a relatively low uncertainty in our results, but there is still some additional space for improvement.

- For dataset size, our data is relatively small (containing only about 1000 individuals), and we believe the predictive models could achieve a better outcome if we train on more relevant data, possibly through data collection and augmentation.
- For noisy data, we believe there exists some noise within our dataset, clearing and preprocessing data to handle outliers and errors can help us ensure the quality of our data, and thus mitigate the impact of noisy data. If we have more data available, we are able to remove more outlier data without giving too much reduction on the size of our dataset.
- For feature quality, although we had some tools to ensure a relatively high quality of feature such as truncating our dataset with AIC and BIC scores, it could be much better if more domain expertise is involved in selecting informative features which could not only enhance model performance but also reduce uncertainty.

Section 5: Causal Inference

Methods

For our causal inference question, our treatment variable is diabetes (binary 0/1), and our outcome variable is BMI (continuous). We recognized that there are many confounders in our dataset, including age, gender, hypertensive, Systolic blood pressure, Diastolic blood pressure, Respiratory rate, glucose, depression', Hyperlipidemia, Renal failure, COPD. While we've considered all these confounders, the unconfoundedness assumption does not hold since there are other factors not included in the dataset that we're not able to include. For example, financial situations could affect both a person's diabetes situation and BMI (a better investment on personal health or regular body check). To adjust for confounders, we used outcome regression that controlled for observed confounders in our dataset. More specifically, we run OLS on dependent variables on treatment and all controlled variables. Additionally, Inverse Propensity Weighting (IPW) in itself also controls for confounders by trying to undo or compensate for the effect of confounders by reweighting the outcome variables based on the treatment and confounders.

The following rationale is listed for each confounder:

- Age: age is often associated with both BMI and diabetes. As people age, they may experience changes in body composition and an increased risk of developing diabetes
- Gender: gender differences exist in BMI distribution and prevalence of diabetes. Hormonal and metabolic variations between males and females could influence the causal relationship
- Hypertension/Systolic/Diastolic blood pressure: high blood pressure is a common comorbidity with diabetes and may also be associated with BMI (ex: diet of unhealthy, high-fat foods => higher BMI and higher blood pressure)

- Respiratory rate: obesity, represented by higher BMI, can affect respiratory function, and diabetes may have an impact on respiratory health
- Glucose levels: glucose levels directly related to diabetes (diabetes is when glucose is too high), BMI may influence glucose metabolism.
- Depression: people with depression may have changes in appetite, physical activity levels/general lifestyle, influencing BMI and diabetes
- Hyperlipemia: higher lipid levels are associated with both higher BMI and diabetes (especially type2). In patients with diabetes, coronary artery disease is most common cause of death
- Renal failure: more likely to develop kidney disease if blood glucose or blood pressure is too high. Diabetes and bmi associated with higher glucose and blood pressure
- COPD: increases risk of metabolic syndrome or type 2 diabetes
- Key: model is biased because there are more confounders that we don't have access to (ex: ethnicity, income, education, etc.)

Since colliders are variables that are influenced by at least two other variables, there is a possibility of colliders in the mortality dataset, but since we're not experts on public health, we cannot say for certain. Of the variables that we mainly used though (gender, BMI, Systolic blood pressure, Diastolic blood pressure, diabetes), we do not believe there are any colliders. Here is a DAG for our variables:

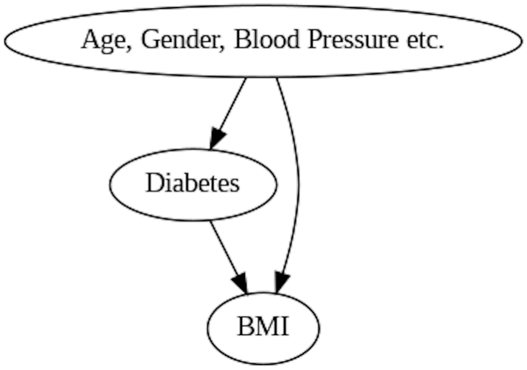


Figure 9: DAG

Results

Before running the outcome regression, we used a partial regression plot (Figure 10) to show there exists an approximate linear relationship between BMI and all covariates.

For outcome regression, the coefficient with respect to diabetes reveals our ATE, which is 1.809. The p-value = 0.005 < 0.05 indicates that it is a statistically significant result. For IPW, the coefficient with respect to diabetes reveals our ATE, which is 1.762. The p-value = 0.001 < 0.05 also indicates that it is a statistically significant result. Under the condition that we've accounted for all confounders, given the small gap between the results of the two methods, we should be quite certain about our concluded ATE of having diabetes causes BMI to increase by around 1.8. We further confirmed that there is sufficient overlap between the propensity score of the two groups (Figure 10). The uncertainty lies upon the fact that we may not be accounting for all confounders within this causal inference scenario, as previously discussed. There could be more confounders in this setting than we previously expected. Moreover, it could be the case that some confounders we identified are also colliders (in this case becomes “bad control”), which makes the problem more intricate to solve. Some ways to account for these two challenges is by using instrumental variables, if we can find a valid IV after careful consideration and validate its assumptions are also satisfied.

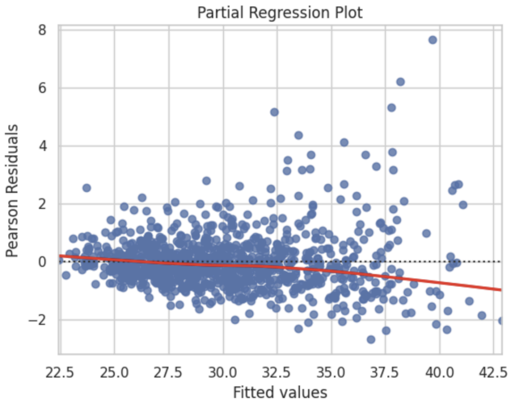


Figure 10: Partial regression plot between all covariates and BMI

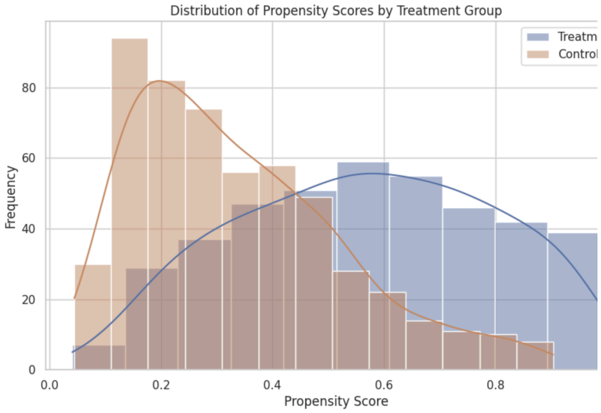
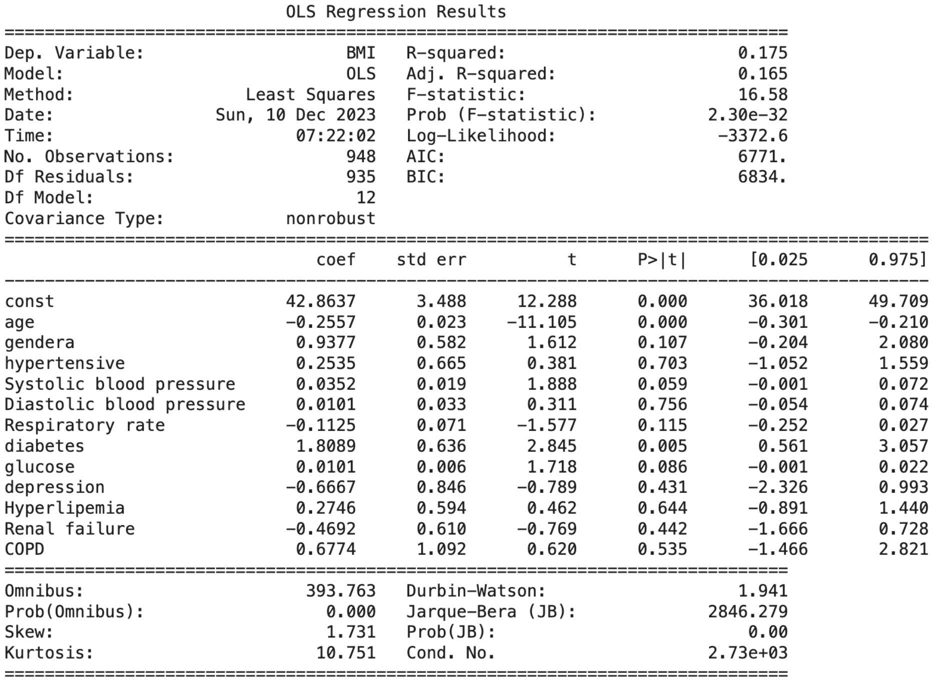
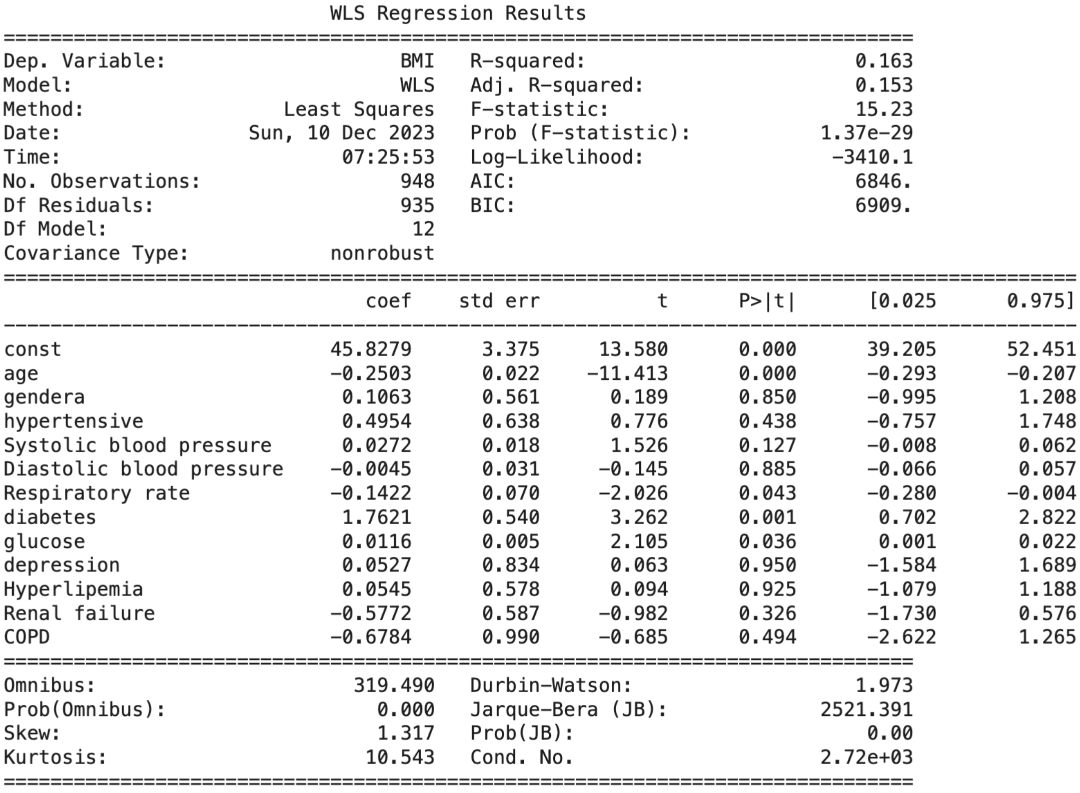


Figure 11: Distribution of Propensity Scores by Treatment



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BMI   R-squared:                       0.175
Model:                            OLS   Adj. R-squared:                  0.165
Method:                 Least Squares   F-statistic:                     16.58
Date:                Sun, 10 Dec 2023   Prob (F-statistic):           2.30e-32
Time:                        07:22:02   Log-Likelihood:                -3372.6
No. Observations:                 948   AIC:                             6771.
Df Residuals:                     935   BIC:                             6834.
Df Model:                          12
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       42.8637      3.488     12.288      0.000      36.018      49.709
age                         -0.2557      0.023    -11.105      0.000      -0.301      -0.210
gendera                      0.9377      0.582      1.612      0.107      -0.204       2.080
hypertensive                 0.2535      0.665      0.381      0.703      -1.052       1.559
Systolic blood pressure      0.0352      0.019      1.888      0.059      -0.001       0.072
Diastolic blood pressure     0.0101      0.033      0.311      0.756      -0.054       0.074
Respiratory rate            -0.1125      0.071     -1.577      0.115      -0.252       0.027
diabetes                     1.8089      0.636      2.845      0.005       0.561       3.057
glucose                      0.0101      0.006      1.718      0.086      -0.001       0.022
depression                  -0.6667      0.846     -0.789      0.431      -2.326       0.993
Hyperlipemia                 0.2746      0.594      0.462      0.644      -0.891       1.440
Renal failure               -0.4692      0.610     -0.769      0.442      -1.666       0.728
COPD                         0.6774      1.092      0.620      0.535      -1.466       2.821
==============================================================================
Omnibus:                      393.763   Durbin-Watson:                   1.941
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2846.279
Skew:                           1.731   Prob(JB):                         0.00
Kurtosis:                      10.751   Cond. No.                     2.73e+03
==============================================================================
```

Figure 12: Result of Outcome Regression



```
                            WLS Regression Results
==============================================================================
Dep. Variable:                    BMI   R-squared:                       0.163
Model:                            WLS   Adj. R-squared:                  0.153
Method:                 Least Squares   F-statistic:                     15.23
Date:                Sun, 10 Dec 2023   Prob (F-statistic):           1.37e-29
Time:                        07:25:53   Log-Likelihood:                -3410.1
No. Observations:                 948   AIC:                             6846.
Df Residuals:                     935   BIC:                             6909.
Df Model:                          12
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                       45.8279      3.375     13.580      0.000      39.205      52.451
age                         -0.2503      0.022    -11.413      0.000      -0.293      -0.207
gendera                      0.1063      0.561      0.189      0.850      -0.995       1.208
hypertensive                 0.4954      0.638      0.776      0.438      -0.757       1.748
Systolic blood pressure      0.0272      0.018      1.526      0.127      -0.008       0.062
Diastolic blood pressure    -0.0045      0.031     -0.145      0.885      -0.066       0.057
Respiratory rate            -0.1422      0.070     -2.026      0.043      -0.280      -0.004
diabetes                     1.7621      0.540      3.262      0.001       0.702       2.822
glucose                      0.0116      0.005      2.105      0.036       0.001       0.022
depression                   0.0527      0.834      0.063      0.950      -1.584       1.689
Hyperlipemia                 0.0545      0.578      0.094      0.925      -1.079       1.188
Renal failure               -0.5772      0.587     -0.982      0.326      -1.730       0.576
COPD                        -0.6784      0.990     -0.685      0.494      -2.622       1.265
==============================================================================
Omnibus:                      319.490   Durbin-Watson:                   1.973
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2521.391
Skew:                           1.317   Prob(JB):                         0.00
Kurtosis:                      10.543   Cond. No.                     2.72e+03
==============================================================================
```

Figure 13: Result of Inverse Propensity Weighting (IPW)

Another IPW strategy we used was to use logistic regression to calculate the propensity scores of each individual - basically calculating the probability of having diabetes given the covariates age, gender, Systolic blood pressure, and Diastolic blood pressure - and calculating the adjusted “weights” of individuals based on the scores. Afterwards, the weighted means of the two groups were compared, which gave an IPW estimate of 2.14. This estimate is fairly close to the estimates given by outcome regression and IPW (1.8). The interpretation from this strategy would be that the estimated effect of diabetes on BMI is 2.14, or the IPW estimate. In other words, on average, individuals with diabetes have a BMI that is 2.14 points higher than those without diabetes, after accounting for differences in age, gender, and blood pressure.

Discussion

The limitations of our study, as above, is that may not account for some confounders. We believe adding more data points could make our estimation more robust and free of bias, and we also believe that including some personal information other than health indexes (i.e. financial situation, education level etc.) could help us to account for more confounders, thus making us more confident on satisfying the unconfoundedness assumption. We are quite confident that there exists a causal relationship between diabetes and BMI (not only because we believe we’ve accounted for enough confounders that we should be close on satisfying the unconfoundedness assumption, but also because the 95% confidence interval does not include 0 for diabetes coefficient). However, we still believe there is space for improvement based on the aforementioned points we discussed.

Section 6: Conclusions

Our key finds in this project, derived from answering our research questions, are (1) Random Forest performs better when predicting the outcome of patients based on the explanatory variables of our choice, which reaches a 98.8%, and (2) The ATE of having diabetes causes BMI to increase by about 1.8 points if we assume the unconfoundedness assumption holds true. As discussed in the first section, our result may not be quite generalizable to the general population and should be limited to apply to similar target patients and similar study settings. A real world decision could be made by designing treatment and intervention strategies targeting specifically for individuals with diabetes given our causal relationship. We can also have early intervention using the predicted outcome to identify individuals at higher risk of mortality and allocate more resources beforehand.

We didn't merge different data sources, since it was hard for us to find relevant information for the patients in the dataset from other data sources. Benefits for combining different sources could be that we have more information on the individuals thus allowing us to have more confounders considered and have more sources to make predictions. However, combining the information from different sources often requires primary key matches. Given the necessity for the health care dataset to protect personal information, it is often hard and unethical to perform the dataset combination.

One limitation as mentioned above is we cannot get other personal information other than health indexes, and we have not very many observations that may influence the accuracy of our model. More intrinsically, our data comes from a mortality dataset in which data was collected from "patients with a diagnosis of HF (Heart Failure), identified by manual review of ICD-9 codes, and who were >15 years old at the time of ICU admission" (Li et al., 2021). This target was chosen because they could be considered vulnerable patients who could need to receive more treatment and care. More considerations and studies are needed to generalize our findings to a wider population.

There are several potential pathways for future research:

- Ensemble methods: we can conduct comprehensive comparisons on a wider range of predictive methods and explore ensemble methods that combine strengths of different models to improve the predictive performance.
- Temporal analysis: it could be interesting to explore how mortality prediction models perform over time, taking into account the seasonality and changes in healthcare practices.
- Patient engagement: encouraging more patients and hospitals to involve in the study which could give more high-quality data and validate our prediction outcome
- Existing risk scores: we can learn how risk scores are derived from existing health data and whether our model can have additional improvement on existing benchmarks.
- Involving health experts into the study

This project has been an invaluable learning experience, allowing us to bridge theoretical concepts from our Data 102 coursework to practical applications with a real-world dataset. The process of working with the dataset enabled us to deepen our understanding of causal inference and machine learning methodologies. We found ourselves drawing on the foundations laid in various data science and statistics courses, seamlessly integrating and applying that knowledge to address complex challenges presented by the dataset. This hands-on experience not only solidified our grasp of fundamental principles but also illuminated the practical nuances of conducting meaningful analyses. We are grateful for the continuous support and guidance from the course staff, whose assistance proved instrumental in navigating the intricacies of this project. This endeavor has not only enriched our skill set but also reinforced the interdisciplinary nature of data science, where theoretical understanding transforms into actionable insights.

References

Medical News Today. (n.d.). Hypertensive.
https://www.medicalnewstoday.com/articles/317220#:~:text=According%20to%20a%202018%20article,those%20with%20typical%20blood%20pressure.

Centers for Disease Control and Prevention (CDC). (n.d.). Age, Gender, Race, and Ethnicity.
https://www.cdc.gov/diabetes/basics/risk-factors.html

Diabetes.co.uk. (n.d.). Respiratory Rate.
https://www.diabetes.co.uk/body/respiratory-system.html#:~:text=Rapid%20or%20laboured%20breathing%2C%20known,of%20ketones%20in%20the%20blood.

Hackensack Meridian Health. (2021, December 14). Can You Get Diabetes from Eating Too Much Sugar?
https://www.hackensackmeridianhealth.org/en/healthu/2021/12/14/can-you-get-diabetes-from-eating-too much-sugar#:~:text=Since%20glucose%20levels%20are%20elevated,insulin%20resistance%20and%20pancreatic%20failure.

Diabetes UK. (n.d.). Depression.
https://www.diabetes.org.uk/guide-to-diabetes/emotions/depression#:~:text=The%20shared%20genes%20play%20a,risk%20of%20type%202%20diabetes.

American Family Physician (AAFP). (1999, March 15). Hyperlipemia.
https://www.aafp.org/pubs/afp/issues/1999/0315/p1666.html

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (n.d.). Diabetic Kidney Disease.
https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-kidney-disease#:~:text=or%20diabetic%20nephropathy.-,How%20does%20diabetes%20cause%20kidney%20disease%3F,can%20also%20damage%20your%20kidneys.

UC San Diego Health. (n.d.). Chronic Obstructive Pulmonary Disease (COPD).
https://myhealth.ucsd.edu/Library/HealthSheets/3,S,60033#:~:text=If%20you%20have%20chronic%20obstructive,syndrome%20or%20type%202%20diabetes.

MIMIC-III Clinical Database
https://physionet.org/content/mimiciii/1.4/

Li, F., Xin, H., Zhang, J., Fu, M., Zhou, J., & Lian, Z. (2021). Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. BMJ open, 11(7), e044779.