Data

The Framingham Heart Study is a long-term cohort study investigating cardiovascular disease in residents of Framingham, Massachusetts.The data used in this report is based on 400 non-diabetic participants without previous coronary heart disease. For this analysis, the 'outcome' is whether or not a participant was hospitalised for myocardial infarction or died from coronary heart disease, in the subsequent 24 year follow-up period. Study data may be found at https://www.framinghamheartstudy.org/fhs-for-researchers/data-available-overview/.

In this report, the variable being predicted will be referred to as the 'outcome'. The variables making the predictions will be referred to as 'features'.

Continuous Variables

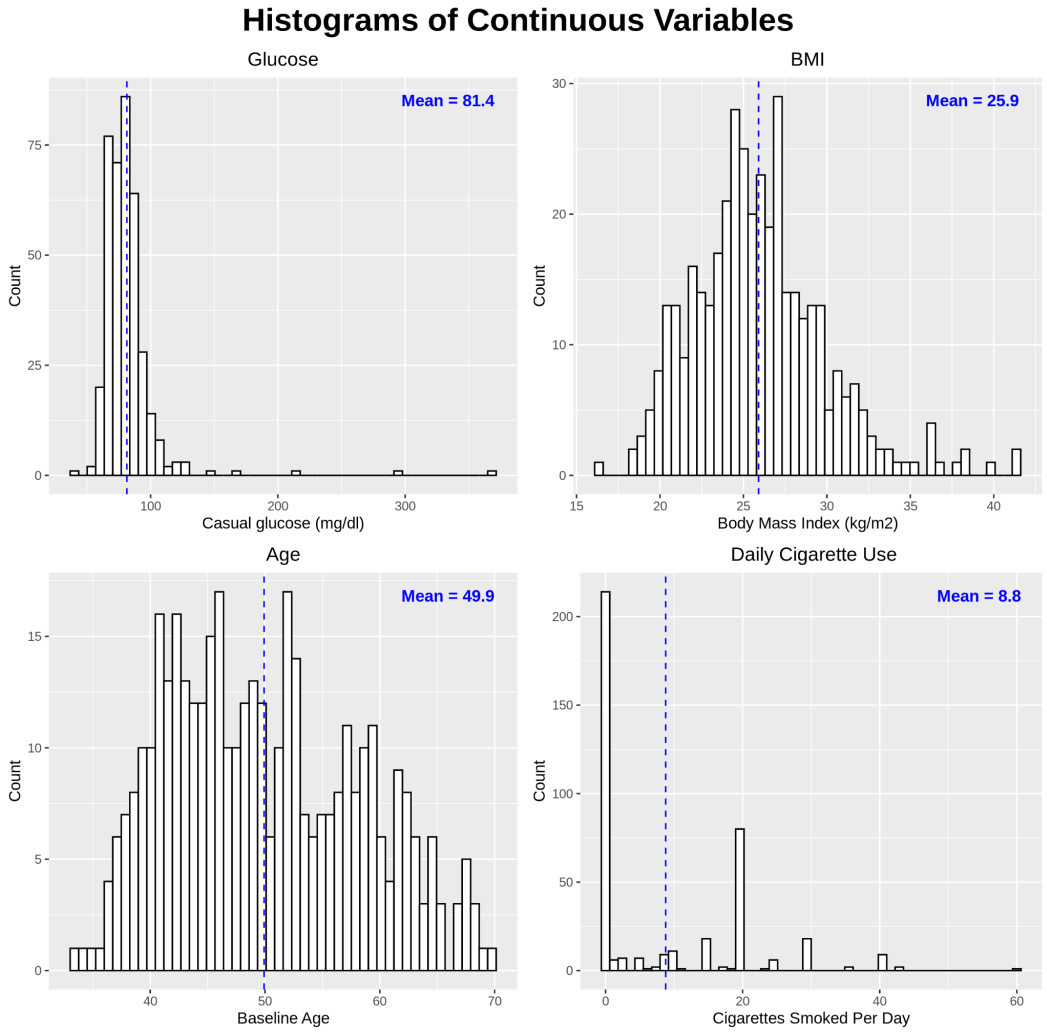

Figure 1.

There are some notably large outlying values for glucose, BMI and cigarettes per day (figure 1). These will not be removed immediately as they may be strongly associated with a particular outcome (e.g the variable mi_fchd). The youngest participant is 33, so relationships drawn between variables may lack applicability to younger demographics, as children, teens and young adults may have different physiological profiles to older demographics. Similarly, the models built may not generalise well to younger demographics due to having not been trained on these ages.

A matrix scatter plot (appendix, figure 5) helped assess whether there were any clear non-linear relationships between these variables. This plot also helped identify whether variables were strongly correlated, which could otherwise lead to regression models whose coefficients are:

- Diluted across the correlated features.
- Unstable which can lead to poor predictive performance.

In this case there were no clear non-linear relationships between the variables. The correlations, although significant (excluding glucose vs cigpday), were not large enough to suggest any variables need removing prior to modelling.

Discrete Variables

Only 18% of participants have the label mi_fchd = 1. Since the dataset (n=400) is relatively small this could mean there are large regions in space (where each variable is plotted along an axis) where this minority class is underrepresented. This may cause problems when building a model of the form $p(mifchd \mid X)$ (i.e the probability of death/hospitalisation for cardiac problems given a column vector of features X) as the parameters may not have been properly specified meaning the model does not generalise well[1].

This class imbalance means it is more sensible to compare suitability of models via precision recall curves as opposed to ROC (receiver operator characteristic) curves. ROC curves analyse the tradeoff between true and false positive rates, but will not always discriminate well between models of differing precision. From a practical standpoint however, a model predicting whether a patient is at risk of death/hospitalisation has to be sufficiently precise to be useful. That is, it should ideally only predict mi_fchd = 1 for those who will actually die/be hospitalised in the future. Therefore precision is central to the question the model aims to answer. Ignoring this may lead to suboptimal model selection.

[1] 'educ' also has some minority classes (Some College = 14% and College Grad + = 12%). This could have conceivably caused the outcome model to not generalise particularly well to individuals belonging to these minority classes. However, education was not a significant feature in the model anyway (see question 3).

Variable	Change	Reason
randid	Dropped	Randomly assigned ID has no predictive power, so is removed before modelling.
bl_age = baseline_visit - dob	Added	Baseline age may be a useful feature for making predictions.
baseline_visit		
dob	Dropped	Relevant age information extracted.
sex		
bmi		
cursmoke		
cigpday	Kept	No changes made.
educ_1 (1 = 0-11 years, 0 otherwise)		A) Breaking down 'educ' allows each class to be it's own feature. This gives the model more flexibility since each new feature can have its own parameter. B) It is not necessarily true the numeric ordering should be 0-11 years < high school or GED < some college < college graduate +. This ordering assumes contributions from this variable should scale (linearly in the case of linear regression models) with level of education - this may not be true.
educ_2 (1 = high school or GED, 0 otherwise)		
educ_3 (1 = some college, 0 otherwise)		
educ_4 (1 = college graduate +, 0 otherwise)	Added	
educ	Dropped	Classes extracted as binary features.
educ_4	Dropped	Any of the newly defined 'educ_i' variables could have been dropped. Dropping educ_4 allows the model to be uniquely specified, as any model (with an intercept term) including educ_4 can be equivalently written without educ_4 by substituting educ_4 = 1 - (educ_1 + educ_2 +educ_3).
glucose		
prior_hyp		
death		
mi_fchd	Kept	No changes made.

Table 1.

BMI/Glucose Relationship:

Red data points in figure 2 have a cook's distance > $4/(sample\ size)$, upon fitting a linear regression model with outcome = glucose and feature = BMI[2]. To determine what *general* trend exists for the majority of participants, these outliers are removed. This results in a regression line with a slightly shallower gradient than if the outliers were kept.

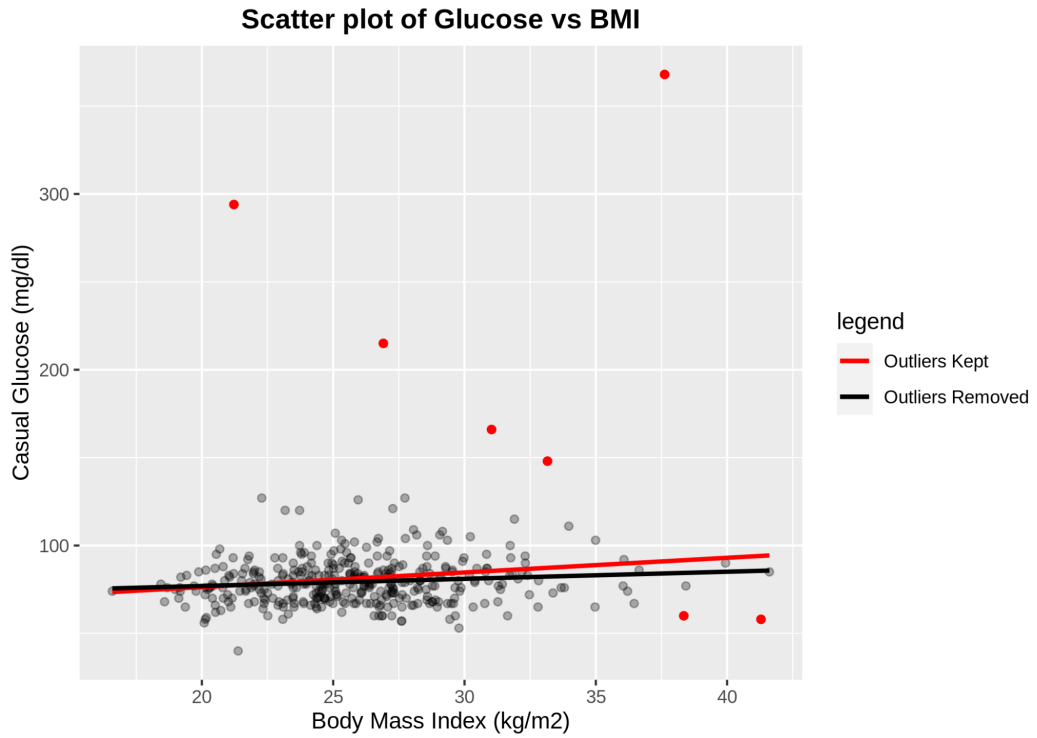

Figure 2.

Before estimating regression parameters, the 24 missing pieces of data are imputed. In particular, BMI data is missing from 3 participants and glucose from 16. Simply removing these participants risks biassing the sample, as missingness may not be completely random. The alternative is to assume the probability of missingness depends on what subset of the cohort is selected, and therefore can be (at least partially) corrected for by building an imputation model on the known data.

Broadly speaking, multiple imputation creates copies of the original dataset and imputes values to each, adding some random noise in the process. Estimates based on each dataset individually can then be pooled together and averaged, allowing for analysis which has incorporated the uncertainty associated with imputation. Single

[2] This initial plot excludes rows with missing glucose/BMI data. The sample size used in the cook's distance calculation reflects this.

imputation methods fail to account for this uncertainty[3]. The default imputation model in the R package MICE is based on predictive mean matching, which imputes missing values by copying the observed value of a similar data point. In this context ‘similar’ means predicted (linear/logistic) regression values are close. This method allows for sensible imputations since imputations come from known values. Imputing across 5 different datasets and pooling the regression estimates gives:

Term	Estimate	Lower 95%	Upper 95%	P-value
Intercept	68.62	59.86	77.39	0.00
BMI	0.42	0.08	0.75	0.01

Table 2.

MICE does not allow for pooled F-tests. However, since there is only one feature in this case, the T-test on the BMI coefficient will be equivalent to an F-test. The significant T-tests indicate each term individually improves the fit of the model. The BMI coefficient (i.e the gradient), says each unit increase in BMI gives an average increase in glucose of 0.42 (mg/dl). This relationship however, is very crude. This is clear from the spread of the data about the trend line in figure 2. Numerically this translates to an R^2 value of 1.31%, meaning only 1.31% of the variation in glucose is explained by BMI (relative to an intercept only model).

This is not a perfectly fitting model. The left plot in figure 3 seems to indicate there is less variability for lower fitted values than larger ones suggesting homoscedasticity

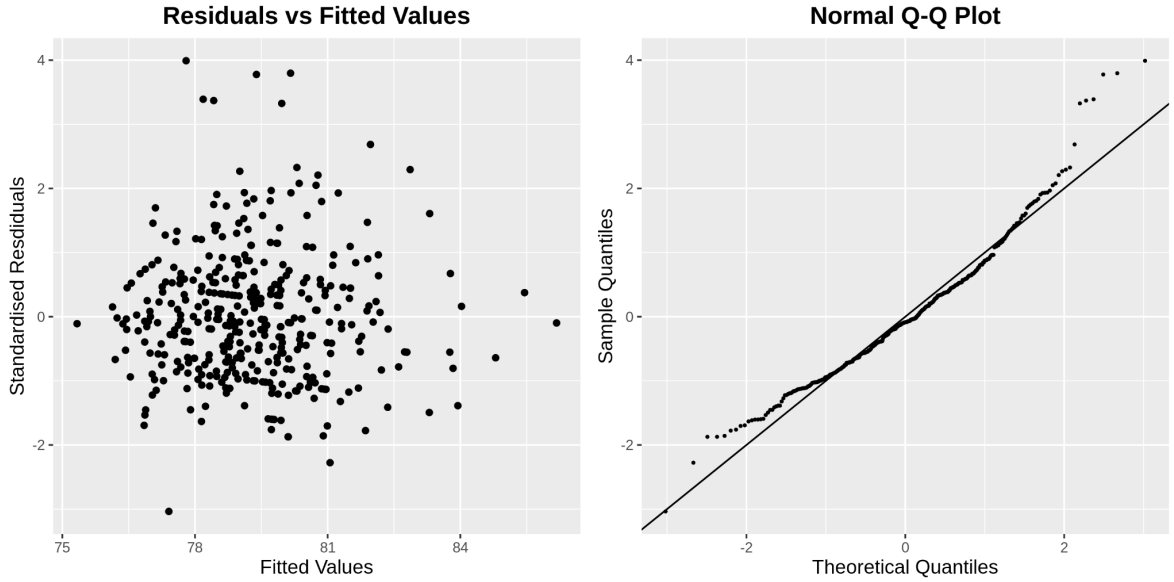

Figure 3.

[3] There aren’t many missing values, so the advantages of multiple imputation over single imputation probably won’t be realised in this case. Nevertheless, it is good practice to apply this approach.

is somewhat violated. The Q-Q plot also suggests the residuals are slightly right skewed. Although not completely inducing homoscedasticity, log transforming glucose does seem to dampen the right skew, indicating this transformed model may fit better (appendix, figure 6). Overall however, I think this risks over-interpreting the relationship between the variables. The key takeaway should be that there is a small but statistically significant positive correlation between the variables.

To see whether the BMI coefficient changes when other variables are accounted for, a multivariable regression is run with all features except 'death' and 'mi_fchd', as these are not known at baseline. VIF's help check whether the $i'th$ feature has linear dependencies on the other features by fitting a regression model and computing $VIF(i) = 1/(1 - R_i^2)$. Therefore VIF's help identify *multi*collinearity (vs pairwise correlations which test for collinearity). The higher the VIF the more the feature can be described via a linear combination of the others. A typical cutoff is a VIF > 5. In this instance however, the max VIF for a given variable across any of the (5 imputed) datasets was 3.4 (result in R code).

The T-tests for all coefficients except for the intercept and cigpday were not significant. This likely means the small effect of BMI (p-value now 0.27) is being washed out by the other variables. Therefore, no new insights are gained via the multiple regression analysis.

Glucose/Outcome Model:

74 participants (18.5%) had death = 1 and mi_fchd = 0, meaning their outcome, had they survived to final follow-up, is unknown. There are arguments for keeping and removing these participants. On one hand, perhaps many of the deaths were a result of generally poor health. In this case, they may have been predisposed to have mi_fchd = 1 (had they survived) and therefore might be expected to have similar features to those with mi_fchd =1. But now these regions of space will be diluted by 'fake' 0's, meaning the probability of mi_fchd = 1 in these regions will be underestimated. On the other hand, perhaps lots of these deaths were simply of old age, and most would not have suffered cardiac problems, had they survived to final follow-up. In this case, removing these data points could bias the model towards over predicting mi_fchd = 1. It helps instead to consider what question the model is trying to answer. Ultimately, a patient wants to know the the probability they will be hospitalised/die of cardiac problems *within the next Y years*. Implicit in this information, is the assumption that only two outcomes are being considered - either being hospitalised/dying (of cardiac problems) within the Y year time frame or surviving Y years without any such problems. That is, it ignores the possibility of dying of something else within the next Y years. Removing participants with (death, mi_fchd) = (1,0) leaves a cohort who either realised the outcome within 24 years, or

survived the 24 years to final follow-up[4], therefore allowing this question to be answered for the special case $Y = 24$. These participants have been removed from the following analysis, to allow the resulting probabilities of the logistic regression model below, to be interpreted in this context.

A similar analysis of the cook's distances, now using a logistic regression model with outcome = mi_fchd and feature = glucose, locates 17 outliers (appendix, figure 7, red data points). In this instance however, the 'outliers' should *not* be removed, as 16/17 are associated with mi_fchd = 1.

Imputing[5] and pooling estimates for this single variable model gives a glucose parameter estimate of 0.02428 (95% CI = [0.00746,0.04110], p-value 0.005), meaning each unit increase in glucose increases the odds of hospitalisation/death by a factor of exp(0.02428) = 1.025 (i.e by 2.5%). To account for other variables, a multiple regression model is run. Again, VIFs were used to detect multicollinearity. As before, the maximum VIF for any given variable across any of the datasets was less than 5 (result in R code). Fitting a logistic regression model gives:

Term	Estimate	Lower 95%	Upper 95%	P-value
Intercept	-10.26	-13.93	-6.59	0.00
Glucose	**0.02313**	**0.00666**	**0.03960**	**0.006**
Sex	0.87	0.19	1.55	0.01
BMI	0.07	-0.02	0.15	0.11
Cursmoke	-0.06	-1.16	1.03	0.91
Cigpday	0.04	0.00	0.09	0.04
Prior_hyp	0.74	0.03	1.46	0.04
bl_age	0.09	0.05	0.13	0.00
educ_1	-0.70	-1.66	0.25	0.15
educ_2	-0.36	-1.32	0.61	0.47
educ_3	0.24	-0.91	1.40	0.68
Significant terms (p-value <5%) highlighted in green.				

Table 3.

Glucose remains significant after accounting for other variables. Now a unit increase in glucose (holding the other features constant) increases odds of

[4] The assignment states participants were followed for 'up to' 24 years for outcomes. An assumption has been made that this means: Scheduled final follow-up age = baseline age + 24 years, and death = 1 means the participant died before reaching this final follow-up age. Of course, some of those with death = 0 may have dropped out of the study. In the absence of such information however, it is assumed they made it to final follow-up.

[5] In this case, predictive mean matching may allow for more sensible imputations than directly imputing regression outputs, as these may be thrown off if the 'outliers' misspecify the regression model.

hospitalisation/death by a factor of exp(0.02313) = 1.023 (i.e by 2.3%). The model also says the odds of death/hospitalisation are:

- 2.38 times greater for men.
- 1.04 times greater for each cigarette/day reported at baseline.
- 2.1 times greater for those with prior hypertension.
- 1.09 times greater for each additional year in baseline age.

Interpreting the non-significant variables would be unreliable given the 95% confidence intervals include positive and negative values (i.e could correspond to increases or decreases in the odds for each unit increase), and there is no conclusive evidence these terms have any effect.

Glucose/Outcome Relationship:

The multivariable logistic regression model take the form:

$$p \;=\; p(mifchd \mid X) \;= \left\{1 \;+\; exp(-\;(\beta_0 + \beta^T X))\right\}^{-1}$$

where X and β are the column vectors of the features and their coefficients respectively and β_0 represents the intercept term. Rearranging gives:

$$\beta_0 + \beta^T X \;=\; ln(p/\,(1 \;-\; p)).$$

Therefore the log odds of being hospitalised/dying is a linear combination of the features. This means each unit increase in glucose linearly increases the log odds by 0.02313. Note however, this is the interpretation of the fitted model and is not necessarily a reflection of the true relationship.

One way to check the relative validity of this linearity assumption is to measure performance against models which incorporate non-linearity. As discussed previously, ‘performance’ will be how much tradeoff there is between precision (% of positive *predictions* which are true positives) and recall (true positive rate) as the threshold for predicting mi_fchd = 1 is lowered from $p = 50\%$ to $p = 0\%$. A logistic model with second order interaction terms between each feature is used as a non-linear comparison. To prevent this more flexible model from overfitting, lasso regression was applied with λ (the regularisation parameter) being tuned via 10-fold cross validation over the training data[6]. Performance over the test set is summarised in figure 4.

[6] 70% of data was used for training and 30% for testing.

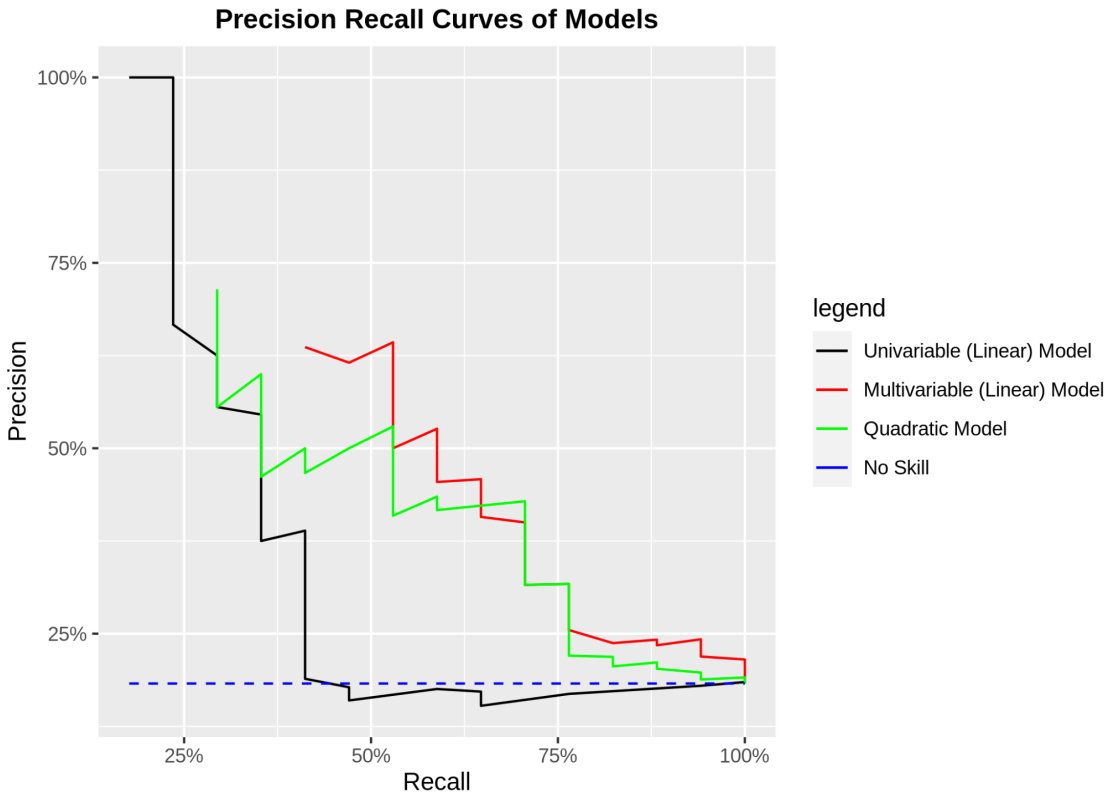

Figure 4.

The quadratic model did not perform any better than the multivariable (linear) model over the test set, indicating the linearity assumption is not restricting performance. However, given the test set contains only 93 data points[7] and 17 positives, this is *not* a particularly reliable finding. Also, the quadratic model will only capture certain forms of non-linearity. It's possible there is some non-linear relationship to be captured, just not by polynomial basis functions.

Interpretation of Results & Conclusions:

Limitations with the analysis have been discussed throughout the report. Additional comments:

Glucose vs BMI

Participants in the top right corners of the Q-Q plots in figures 3 and 6 (deviating away from the 45 degree lines) did not seem to share characteristics explaining their higher glucose levels. Although this meant the homoscedasticity and normality assumptions were not completely met, it's generally unrealistic to expect real world data to perfectly adhere to these simplifying assumptions.

[7] Data without imputations was used, as imputing and then splitting may introduce bias to the test set.

The Logistic Models

The logistic models (especially the univariable one) all have a reasonably large decrease in precision as recall increases. This means, for some given threshold, if the model predicts a ‘1’, then it can’t be said with much confidence whether this outcome will actually be realised. This is likely a feature of the data being modelled. Figure 7 (appendix) shows there is significant overlap between mi_fchd = 0 and mi_fchd = 1, for most glucose values, meaning any model is going to struggle to separate the classes based on this feature alone. Although adding more features significantly improved the precision-recall plot, the tradeoff still suggests the classes aren’t overly well separated even in this higher dimensional space. However, since these are probabilistic models, it makes more sense instead to report the *probability* of hospitalisation/death to the patient. This avoids having to make a definitive and crude statement on future outcomes and gives the patient a more practical measure of risk.

Survival Analysis

Presumably more time data on participants was recorded, such as hospitalisation/death dates. Survival analysis could have used this information to model probability of survival any number of years from baseline (vs the current model which only predicts outcomes 24 years from baseline). This would have been particularly useful for older patients, where short term information (e.g. the likelihood of death/hospitalisation for cardiac problems one year down the line) is much more practical than a 24 year forecast. The kaplan-meier survival estimator would have incorporated those who died prior to final follow-up without having a realised outcome, by adjusting the estimator to reflect fewer participants as the study progressed.

Appendix

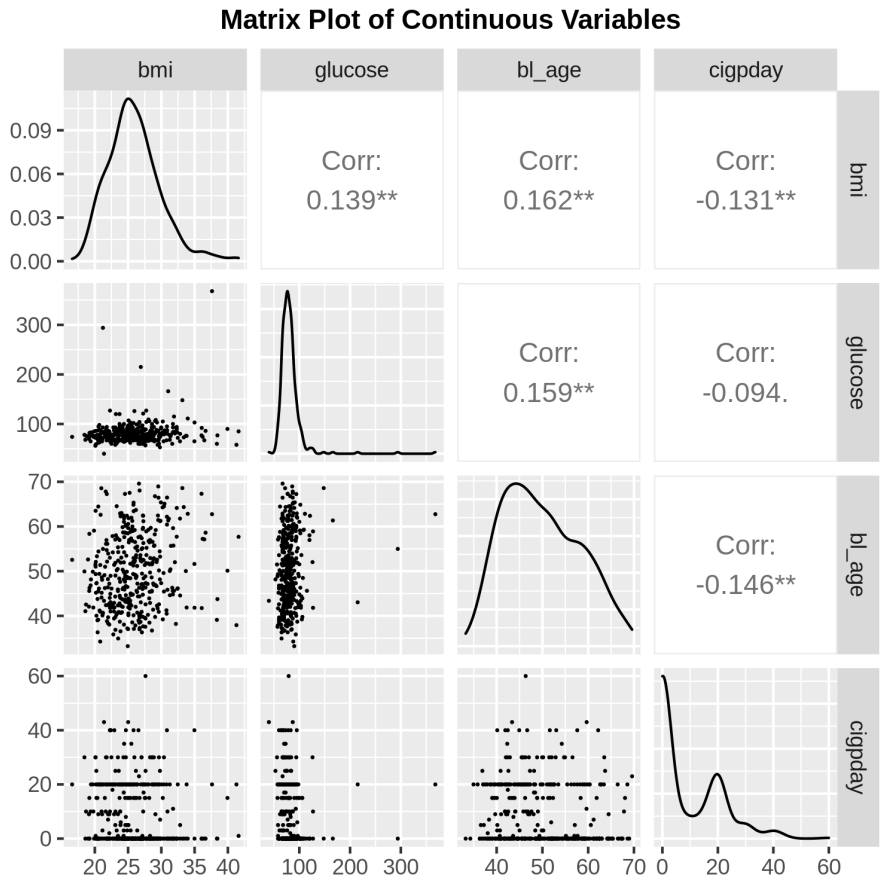

Figure 5.
*** Indicates p < 0.01*

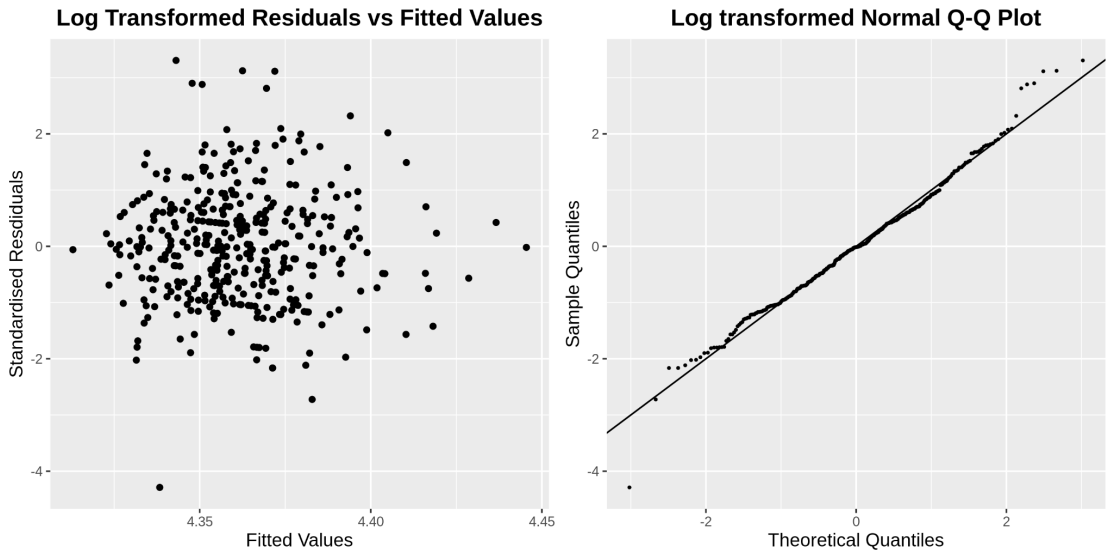

Figure 6.

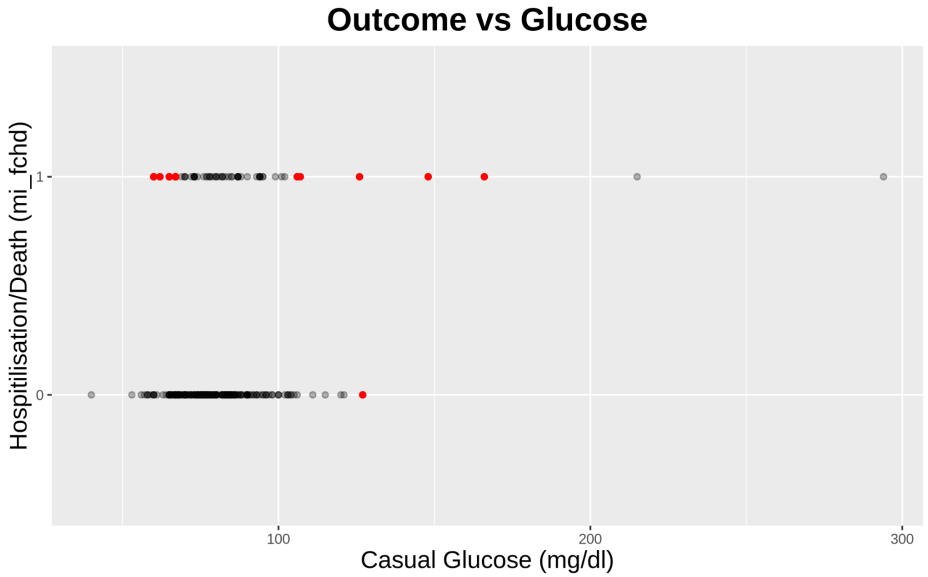
Outcome vs Glucose
Hospitilisation/Death (mi_fchd)
1
0
100
200
300
Casual Glucose (mg/dl)

Figure 7.