Does **blood glucose have an independent association with subsequent hospitalisation for myocardial infarction or death from coronary heart disease**, after accounting for other measured risk factors?

1. How each variable is treated

The objective of this analysis is to appropriately describe, explore and analyse a subset of data from the Framingham heart study, to understand the nature of a potential independent association between our primary variable of interest – blood glucose and our outcome variable – the subsequent hospitalisation for myocardial infarction (MI) or death from coronary heart disease (CHD), after accounting for other measured risk factors. Approximately 14.3% of the individuals in the dataset experience our outcome of interest.

There are 15 variables considered as part of this dataset collected over a period of up to 24 years for 4,215 participants. 5 out of the 15 variables are date type variables, 3 variables are continuous (glucose, body mass index (BMI) and cigarettes per day), 1 variable is a random unique identifier for each individual and the remaining 6 variables where binary or categorical variables, where most individuals (if data was collected) were allocated to one of the discrete groups assigned to each variable.

In order to gather greater clinical significance from our investigative analysis for association it was decided two additional categorical variables for glucose and BMI would be generated. It was also decided not to categorise the variable – cigarettes per day, because under the knowledge available for this analysis, appropriate clinically significant groups could not be justified.

Using values referenced from Diabetes.co.uk (2018) it was decided to assign glucose values into two discrete categories those below 200mg/dl and those above. This is to differentiate those who have a normal random glucose test (<200mg/dl) and those who do not (≥200mg/dl) and therefore could be classified as showing early signs of a diabetic aetiology. This is useful as a categorisation as it may present an additional dimension with which to understand the potential independent association between our variable of primary interest and outcome. Allowing us to look at the risk difference between those who have clinically significant high glucose levels against those who do not.

For BMI (kg/mg^2), values were placed into the following categories (WHO, 2018) – Underweight (≤18.5), Normal weight (18.5≤25), Overweight (25≤30), Obese class I (30≤35) and Obese class II (>35). This is useful because it helps to understand potential trends associated with our survival outcomes for which cannot be observed if BMI were modelled only as a continuous variable. The challenge though with categorisation is the loss of information that may be further linked to a loss of power between groups, leading to results that are less statistically significant ($p<0.05$) and therefore arguable not as useful as a single statistically significant continuous variable result.

Lastly, 4 out of the 15 variables within our dataset demonstrated varying degrees of missing data. The subsequent methods and means by which this missingness is to be treated will be described further below.

2. Assumptions and modelling decisions relating measured baseline glucose to the outcome

Univariate regression analysis

As part of building an appropriate model, the first model decision taken is to understand which variables to include within the model in order to investigate our association of interest. The following univariate regression analysis (Table 1) aims to understand the single-variable relation of the potential predicator variables described above to our outcome. Although it might be argued such relations are not very informative by themselves, as potential interactions with respect to the outcome will be missing, such univariate regression analysis will help to appropriately limit the number of variables included in our model.

For categorical variables we will use the log-rank test of equality across each strata and for continuous variables we will use a univariate cox proportional hazard regression. We will consider including the predictor in the model if the p-value for the regression is $p>0.2$. A variable with a predicator, $p>0.2$ is highly unlikely to contribute much to a model that includes a number of other significant and often equally as clinical valid variables.

From Table 1 below it is clear than all the variables available to include in the model are significant for the outcome under an individual univariate regression analysis.

Cox multiple regression model

Participants as part of our dataset were followed up for varying amounts of time. Using a multiple regression model, the intention is to take these individual time linked observations into consideration whilst adjusting for the additional risk factors linked to our association of interest.

There are two types of multiple regression models, either semiparametric or parametric that could be applied to our dataset. The discussion below explains how our modelling decisions tested the overall fit for both models and comes a conclusion on the most appropriate model to use with regards to our association analysis. Our semiparametric model will be a Cox multiple regression model and our parametric model will be a Wiebull multiple regression model.

The nature of multiple regression models allows for testing of interactions between variables. From the nature of the clinical variables collected as part of our dataset it is difficult to identify all potential interactions that may exist and further consultation with clinical experts will be necessary to fully account for all interactions. From initial variable interaction testing of the model the most clinically relevant interaction presented in this model, from our given variables is that between sex and prior history of hypertension. This is likely to be relevant because research presents evidence that there are significant sex differences in hypertension (Gillis and Sullivan, 2016).

For our cox multiple regression model, the hypothesis test for the proportional hazard assumption showed overall evidence of a violation for both the continuous (p=0.011) and categorical glucose variable (p=0.0084). However, if we review the individual covariates in the model, there is a suggestion of a reasonably strong violation from the covariate -

currently smoking (p=0.04). Using a model stratified for currently smoking gives very similar hazard ratios to before but improves the overall evidence for the violation of the hypothesis test for the continuous glucose (p=0.17) and category glucose (p=0.15). Via stratification of this model, under the variable currently smoke, we are unable to measure the differences of our outcome between those who do and do not currently smoke (as reported at baseline measurement). As this covariate is not of primary interest, it arguably appropriate to stratify upon this variable in order to appropriately model our association of investigation without the constraint of proportionality.

Taking both this interaction and stratification of the model into account. The results below (Table 2) present the modelling effect for the semiparametric cox multiple regression model applied to our data.

Weibull multiple regression model

A semi-parametric and parametric proportional hazard model are similar in the interpretation of the results from the model yet differ via their individual assumptions made upon the nature of the hazard. The parametric model assumes the hazard follows a set statistical distribution when fitted to the data, whereas a semi-parametric model like our Cox regression model makes no such assumption.

For our data, the Weibull model provides very similar estimates to those of our Cox regression. In turn the extra assumptions made by the model are reasonably appropriate for our data. The likelihood ratio test for this model is also statistical significant at $p<0.05$ level ($p<0.0001$).

On review, it seems justified to select the Cox regression model as the model of choice to explain our analysis association. Both models produce significantly similar results. Yet the cox regression makes the fewest statistical assumptions upon our data and when stratified for the variable currently smoking, does not violate the proportional hazard assumption.

3. The relationship between glucose and the outcome

Under both multiple regression models, evidence exists to suggest that there is overall a linear relationship between our exposure and outcome variable. Our cox model described above estimates that a one-unit increase in glucose is associated with a 0.69% increase in the hazard of hospitalisation for myocardial infarction or death from coronary heart disease (HR=1.0069). Because the range of glucose measurement began at approximately 40mg/dl and extends to over 300mg/dl, a 10-unit increase in glucose may be a more appropriate clinical scale to describe this linearity. To this interpretation, a 10-unit increase in glucose is associated with an approximate 7.14% increase in hazard of hospitalisation for myocardial infarction or death from coronary heart disease (HR=1.071).

A similar hazard ratio for continuous glucose measurements (HR=1.0069) under the Weibull regression model further strengthens the overall linearity described by the cox regression. The Weibull model makes more assumptions regarding the nature of the relationship between our exposure and outcome. Therefore this similarity of hazard ratio suggests the association between our exposure and outcome is linked with a strong linear relationship.

If we are also to review the hazard ratio under glucose as a categorical variable we continue to observe that a linear relationship between those with a normal random glucose measurement (<200mg/dl) and those with a random glucose measurement suggestive of diabetes (≥200mg/dl). Our results suggest that for every one unit increase in glucose measures those with a glucose measure suggestive of having diabetes (≥200mg/dl), have a 25.2% increase hazard of hospitalisation for myocardial infarction or death from coronary heart disease (HR=1.25).

Having said this, upon carrying out further tests for linearity, it becomes clear that the relationship between glucose and the outcome is in fact non-linear and more likely quadratic. Carrying out a likelihood ratio test between a cox regression model which includes a quadratic term for glucose and a model without, we obtain a p-value that is significant at the $p<0.05$ level ($p=0.035$). Taking this quadratic relation into account in turn alters the hazard

ratio for glucose (HR = 1.016). Therefore suggesting a one-unit increase in hazard at 1.6% rather than an 0.69%.

For clarity of analysis, beyond understanding linearity of association the quadratic term for glucose has been omitted from further analysis of association.

4. How missing data is treated

According to initial investigation of the data the following four variables have the following level of missing observations – BMI (17 missing observations), Glucose (387 missing observations), Date of death (2,877 missing observations), Date of hospitalised MI or death from CHD (3,613 missing observations). To handle the continuous data of BMI and glucose, the statistical method of multiple imputation will be used to impute the missing gaps in the data. This imputation technique will be used under the missingness assumption that data is most likely missing at random. Under this assumption multiple imputation does well to produce data that is broadly consistent with the nature of distribution that already exists within the dataset. Therefore the significance of our cox model may be improved without compromising the skewness of data or increased variation. We will not be imputing data for the missing dates values using multiple imputation because it is not a continuous variable and therefore missing values cannot be drawn appropriately from a distribution.

Our results upon running our original cox regression model using imputed glucose and BMI values (Table 3) show that the hazard ratio has not changed much. Yet there is a small decrease in the standard error which is somewhat to be expected because the level of missingness for glucose and BMI are not extensive at 0.4% and 9.1% respectively. A similar pattern though cannot be said for glucose as a categorical variable, although the hazard ratio are similar the standard error seems to increase. If we refer to the wider 95% CI for the categorical HR compared to the continuous HR, it is possible to expect greater variation between different analysis for the categorical glucose model compared to the continuous glucose model.

Finally with reference to results in Table 4, we can be highly confident that the missingness at random assumption is highly plausible for our data. Small changes if any in the standard deviation or mean along rounds of imputation suggests a strong robustness of this assumption and further justifies multiple imputation as an appropriate statistical technique to deal with continuous missing data within our dataset.

5. Overall interpretation of the results and conclusion

Overall from our results we can be confident to conclude that there is an independent non-linear association between blood glucose and subsequent hospitalisation for myocardial infarction or death from coronary heart disease, after accounting for other measured risk factors.

Using our fitted cox regression model, stratified for currently smoking, having taken into account the interaction between sex and prior history of hypertension, we observe a 7.14% increase in hazard of hospitalisation for myocardial infarction or death from coronary heart disease (HR=1.071) for every 10-unit increase in glucose. Although the model is appropriate as the proportionality of the hazard assumption is not violated, the clinical significance of this result would need to be questioned in consultation with clinical experts to understand the true significance of this association.

Although these results are in line with a potential clinical explanation, additional variables from our model including age of each individual and ethnicity were missing from the data available for our analysis. Taking into account that further assumptions were made using multiple imputation (although reasonably justified), it needs to be considered that there are tangible limitations to our model and in turn the nature of the independent association described above.

Table 1: Univariate analysis results

Categorical Variables	**Log rank test of equality across strata p-value**
Sex (1=male, 0=female)	P<0.0001
BMI category (Underweight, Normal weight (reference group), Overweight, Obese class I, Obese class II)	P<0.0001
Currently Smoking (0 = No, 1 = Yes)	P<0.0001
Education Level (1=0-11 years, 2=high school, 3=some college, 4=college graduate or higher)	P=0.12
Glucose category (0 = Normal, 1 = Diabetic)	P=0.0001
Hypertension prior to baseline (0 = No, 1 = Yes)	P<0.0001
Continuous Variables	**Chi-squared test p-value**
BMI (kg/mg^2)	P<0.0001
Number of cigarettes per day	P<0.0001
Glucose (mg/dl)	P<0.0001

Table 2: Adjusted* multiple linear regression model results (without the glucose quadratic term)

	Cox regression (stratified for the variable - 'currently smoking')			Weibull regression		
Covariate	Hazard ratio (HR)	95% CI	P-value	Hazard ratio (HR)	95% CI	P-value
Glucose (continuous)	1.0069	1.0046 – 1.0093	<0.0001	1.0069	1.0046 – 1.0092	<0.0001
Glucose (categorical)						
- <200 mg/dl	1	(base)		1	(base)	
- >200mg/dl	1.25	1.10– 1.42	<0.0001	1.25	1.10– 1.42	<0.0001
*adjusted for Cigarettes per day, Currently smoking (omitted from cox regression), BMI category, Education, the interaction between Sex and prior history of hypertension before baseline						

Table 3: Adjusted* multiple Cox regression model results with and without imputed data (without the glucose quadratic term)

Glucose	HR for mortality	95% CI	SE of Log HR
Without imputation			
Continuous	1.0069	1.0046 – 1.0093	0.0012
Categorical			
- <200 mg/dl	1		
- ≥200mg/dl	1.25	1.10– 1.42	0.064
With imputation of Glucose and BMI			
Continuous	1.0070	1.0049 – 1.0093	0.0011
Categorical			
- <200 mg/dl	1		
- ≥200mg/dl	1.26	1.06 – 1.50	0.088
*adjusted for Cigarettes per day, Currently smoking (omitted from cox regression), BMI category, Education, the interaction between Sex and prior history of hypertension before baseline			

Table 4: Sensitivity analysis of glucose and BMI multiple imputation rounds

Glucose					
Interaction round	Observations	Mean	Standard deviation	Minimum value	Maximum value
m = 0	3,828	81.93	23.90	40	394
m = 1	4,215	82.03	23.95	16.46	394
m = 10	4,215	81.98	23.87	10.67	394
BMI					
m = 0	4,198	25.79	4.06	15.54	56.8
m = 1	4,215	25.79	4.05	15.54	56.8
m = 10	4,215	25.80	4.06	15.54	56.8

References

Diabetes.co.uk. (2018). *Normal and Diabetic Blood Sugar Level Ranges - Blood Sugar Levels for Diabetes*. [online] Available at: https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html [Accessed 8 Apr. 2018]

WHO. (2018). *Body mass index - BMI*. [online] Available at: http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi [Accessed 8 Apr. 2018].

Gillis, E. and Sullivan, J. (2016). Sex Differences in Hypertension. *Hypertension*, 68(6), pp.1322-1327.