INSTITUTE OF HEALTH INFORMATICS

Graduate Programme in Data Science for Research in Health & Biomedicine

Assessed Coursework

Student candidate number:	*****
Module:	CHMEGH39: Advanced Statistics for Records Research (now CHME0015)
Date due:	Tuesday, 7th May 2019, 12:00 midday
Word count: (excluding references, diagrams and appendices)	Word count: 2051
Disability or other medical condition for which UCL has granted special examination arrangements:	
Formative feedback:	Please address in formative feedback: Yes please. Please ignore in formative feedback:

"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?"

Methods

Data from the Framingham Heart Study, Massachusetts of 4215 individuals without prior coronary heart disease and diabetes free with a follow-up period of 24 years was analysed to investigate the association between blood glucose status (glucose) and probability of hospitalisation for myocardial infarction or death from coronary heart disease.

The exposure was blood glucose status (baseline) and the outcome was cardiovascular disease (CVD). Sex, age, body mass index (BMI), education status and smoking status were considered a priori confounders or effect modifiers.

Data exploration

Using Stata version 15SE, the dataset was explored to highlight missing values in the data and to inform decisions to recode clinically significant covariates for analysis and interpretation. Blood glucose status and BMI were found to have missing data values.

Recoding variables

Continuous variables were categorised into clinically meaningful categories where relevant, to aid interpretation, or for important potential confounders such as age.

Blood glucose status and BMI were categorised in accordance with accepted clinical ranges (Type 2 Diabetes guide. 2017), (NHS choices, 2016). Blood glucose levels were categorised into four levels, <70mg/dl (low blood sugar (hypoglycemia)), 70/100mg/dl (normal blood sugar), 101/125mg/dl (pre-diabetic blood sugar) and >126mg/dl (diabetic blood sugar) (Type 2 Diabetes guide. 2017). An alternative blood glucose variable was created which was categorised into two levels, <100mg/dl and >101mg/dl for sensitivity analysis.

BMI was categorised into underweight (min/18.5), healthy weight (18.5/24.9), overweight (24.9/29.9), obese (29.9/39.9), very obese (39.9/max) and missing (NHS choices, 2016).

The age of each participant at start of follow-up was determined by subtracting their date of birth (dob) from the baseline visit date and dividing by 365.25 to give the age in years. Based on the observed age distribution of study participants and frequently used clinical categories, age was categorised in to five categories: <45yrs, 45/50yrs, 50/55yrs, 55/60yrs and >60yrs.

The number of cigarettes per day were categorised as follows, <10 per day, 10/20 per day, 20/40 per day and >40 per day.

Education status was categorised in the following manner: 0-11yrs, high school, some college, and college graduates or higher qualification.

The date of death, date of coronary heart attack and date of last follow-up date were recoded to generate a new variable called 'enddate' to indicate end of follow-up for survival analysis.

Unadjusted analysis

An unadjusted logistic regression analysis was carried out showing the relationship between the outcome and each of the individual covariates: blood glucose, sex, age, BMI, current smoking status, cigarettes per day, education and prior hypertension. The unadjusted analysis was performed with covariates included as both continuous and categorical variables (where applicable). Non-linear relationships were explored by including continuous variables as quadratic and fractional polynomial terms. The association between the exposure and the outcome, was subsequently modelled using Cox regression. The test for proportional-hazards assumption (global ph test) was performed. The relevant results of this analysis have been presented in **Table 1**.

Adjusted logistic regression analysis: complete case analysis (CCA)

Prior to dealing with the missing values, a complete case analysis (CCA) using logistic regression was performed on the entire dataset to show the association between the outcome and the exposure with age, sex, BMI, current smoking status, cigarettes per day, education status, prior hypertension as covariates. This analysis was carried out for blood glucose levels modelled both as a continuous and categorical variable in turn. The initial model with blood glucose included as a continuous variable assumed a linear relationship between the exposure and outcome, **Table 2**.

Adjusted logistic regression analysis: multiple imputation (MI)

Overall, 0.91% of the glucose values and 0.31% of the BMI values were missing. Thus, multiple imputation was used for imputing the missing glucose and BMI covariate values (Missing data II: Multiple imputation, 2017), (Sterne J. et al. 2009). The following variables were used to impute the missing glucose and BMI values: age, sex, currently smoking smoke cigarettes per day, education status, prior hypertension and the CVD outcome.

Following multiple imputation, blood glucose was analysed both as a continuous and categorical variable in turn, **Table 2**.

Adjusted analysis: Cox regression

Cox regression was used as an alternative modelling approach to investigate the study hypothesis. A survival analysis was performed to explore the relationship between the outcome and both continuous and categorical blood glucose in turn (with the four categories as previously described). This was done using the MI dataset with the model including age, sex, BMI, current smoking status, cigarettes per day, education status, prior hypertension as covariates.

Sensitivity analysis

A sensitivity analysis was conducted using Weibull survival analysis as while the global test indicated the proportional hazards assumption was met, visual examination of the curves suggested that an alternative non-parametric method was worth exploring.

Further sensitivity analysis was carried out at various points in this analysis and will be referred to later.

Results

Of the total population (N=4215) it was found that 1810 were male and 2405 were female. By the last follow-up, 1388 participants had died. There was 602 individuals with an adverse CVD outcome.

Unadjusted analysis

In the CCA unadjusted analysis (**Table 1**), **glucose** modelled both as a continuous variable (OR 1.008 [1.005-1.011], p<0.001) and for individuals with >126 mg/dl blood glucose levels (OR 4.6 [2.84-7.44], p<0.001) in the categorical approach were strongly associated with increased CVD.

Age was found to be strongly associated to the CVD outcome for both continuous analysis (OR 1.04 [1.03-1.05], p<0.001) and categorical analysis of individuals >50yrs, with a 76% to 127% (p<0.001) increased likelihood of the adverse CVD outcome.

Males were 189% more likely to have an adverse CVD outcome (p<0.001). While individuals with a **BMI** categorised as obese had a 161% increased likelihood of CVD (p=0.046).

Those individuals who were **current smokers** were 34% more likely to have an adverse CVD outcome (p<0.001). Consumption of **cigarettes per day** (as a continuous variable) was strongly associated with the outcome (OR 1.02 [1.01-1.02], p<0.001). Categorisation showed that for 10/20 and 20/40 cigarettes per day there was a 51% (p<0.001) and 77% (p<0.001) increase in the likelihood of an adverse outcome, respectively.

Continuous **education** status (OR 0.95 [0.87-1.04], p=0.251) did not have an association with the outcome, however, in the categorical analysis, participants with some college were 29% less likely to have the adverse outcome (OR 0.71 [0.54-0.93], p=0.011).

Prior hypertension was found to be strongly associated with the outcome (OR 2.14 [1.80-2.56], p<0.001).

Analysis was carried out using **Cox regression** using two glucose categorisation approaches. The four and two category approaches, (see methods section). When blood glucose was categorized in to four levels, hazard ratios for individuals with >126 mg/dl blood glucose status was strongly associated with the outcome (HR 3.79 [2.66-5.41], p<0.001). While for the two category approach was also strongly associated with the outcome (HR 1.75 [1.37-2.25], p<0.001).

The **proportional hazards assumption** (four category approach, global ph test p=0.73 and the two category approach, global ph test p=0.45) was met, **Table 2**.

The quadratic and fractional polynomial analysis (Figure 2), demonstrated that there is an increasing likelihood of an adverse outcome until ~260 mg/dl of blood glucose. After this point, there is a decreasing likelihood of the outcome. Locally weighted scatterplot smoothing (lowess) was performed and confirmed the non-linear association between the exposure and outcome.

Adjusted analysis: complete case analysis (CCA)

In the adjusted CCA, for each mg/dl of glucose increase in continuous **glucose variable there was** a ~1% increase in the likelihood of an adverse outcome (1.007 [1.0003-1.009], p<0.001). While the adjusted CCA categorical glucose (>126mg/dl) showed a reduced likelihood of an adverse outcome (OR 3.14 [1.95-5.08], p<0.001) when compared to unadjusted CCA categorical glucose (OR 4.6 [2.84-7.44], p<0.001), (**Table 2**).

The CCA for both continuous and categorical blood glucose status were plotted presenting the predicted association of blood glucose status with adverse coronary heart disease as per **Figure 1(a)** and **Figure 1(b)**, respectively. 100 mg/dl of glucose was used as a reference as this represented the upper end of the normal blood glucose range. In both continuous and categorical plots, the odds of CVD increased with increasing blood glucose status.

The combined continuous and categorical logistic regression plots in **Figure 1(c)**, illustrates that with increasing blood glucose status there is a non-linear increase in the odds of an adverse outcome. The categorical plot shows that at ~125 mg/dl of glucose there is a sudden increase in the predicted odds of an adverse outcome, after which the predicted odds remain constant for the participants.

Adjusted analysis: multiple imputation (MI)

The same analysis was repeated, with missing glucose and BMI values generated using **multiple imputation (MI)**.

When treated as a continuous variable (**Table 2**), glucose was found to be unchanged in terms of likelihood of outcome, compared to the CCA adjusted continuous glucose. For MI adjusted categorical glucose, there was a slight difference (OR 3.12 [1.94-5.02], p<0.001) with the CCA approach (OR 3.14 [1.95-5.08], p<0.001).

Adjusted analysis: Cox and Weibull survival analysis (MI)

Adjusted **Cox** and **Weibull survival analysis** was used with continuous and categorical glucose. There was a strong association between the continuous glucose variable and the outcome (HR 1.07 [1.005-1.009], p<0.001). There was no change in hazard ratio between the Cox and the Weibull regression results for blood glucose modelled as a continuous variable.

For categorical glucose, the hazard ratios for individuals with >126 mg/dl blood glucose status was strongly associated with the outcome (HR 3.05 [2.42-5.10], p<0.001), which was consistent with the Weibull analysis for the >126 mg/dl blood glucose level (HR 3.53 [2.43-5.12], p<0.001), with the latter showing a slightly increased hazard ratio.

Discussion

Summary

The analysis demonstrated that the association between blood glucose and the outcome is non-linear. Blood glucose has an independent association with subsequent hospitalisation for myocardial infarction or death from coronary heart disease, after accounting for other measured risk factors. When compared to the baseline, there is a 212% increased likelihood of hospitalisation for blood glucose levels of >126 mg/dl. This is dependent on the blood glucose concentration.

For both the CCA and MI approaches, in **Figure 2** there is an increasing odds of an adverse outcome with each unit increase of glucose and after ~ 250 mg/dl of glucose the quadratic expression shows a reduction in the odds of an adverse outcome with increasing blood glucose status. This pattern is also apparent with the fractional polynomial analysis.

The findings of the CCA and MI approaches, using continuous, categorical, quadratic and fractional polynomial were consistent.

Missing data

Any instances of missing data in the dataset was explored using the tabulation and row chi commands in Stata. As previously mentioned, as both glucose and BMI had missing values, the MAR mechanism was assumed leading to the use of multiple imputation for both variables. Missing blood glucose values represented < 1% of the total glucose data.

Sensitivity analyses was conducted on the CCA and MI approaches for continuous, categorical and quadratic glucose. Keeping the reference (category unchanged) the ORs, 95% CIs and p-values were found to be the same between both the adjusted CCA continuous models, **Table 2**. The use of both adjusted Cox and MI Weibull analysis with multiple imputed data was treated as a sensitivity analysis as the unadjusted Cox regression global test was met for both the four and the two glucose level categories, demonstrating the robustness of the analysis, **Table 2**. The plot of the quadratic and fractional polynomial analysis of the CCA and MI approaches showed that both overlapped one another.

Limitations

Potential confounding by age, sex, currently smoking, smoke cigarettes per day, education status, blood glucose status, prior hypertension was accounted for via adjustment but the potential for residual confounding due to unmeasured confounders remains.

However, time varying exposures were not investigated as only baseline data was available, therefore, the current analysis does not account for variation across the duration of the 24 year study.

The age categorises used grouped those <45 years of age in to a single category representing 33.54% of the participants, while the remaining participants were split between four additional categories (<45yrs (33.54%), 45/50yrs (18.83%), 50/55yrs (16.35%), 55/60yrs (14.86%) and >60yrs (16.41 %)). This was done as it is accepted that there is an increased likelihood of the adverse outcome for older participants (Heart UK, 2017), (NHS choices, 2017).

References.

- Missing data II: Multiple imputation, 2017. Advanced Statistics for Records Research lecture. Available from: https://moodle.ucl.ac.uk/pluginfile.php/3311627/mod_resource/content/4/Missing%20data%20II.pdf [Accessed 03 April 2018].
- Sterne J. et al. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. British Medical Journal, 338(b2393), pp. 1-11
- NHS choices, 2016. What is the body mass index (BMI)?. Available from: https://www.nhs.uk/chq/Pages/3215.aspx?CategoryID=51 [Accessed 27 March 2018].
- Type 2 Diabetes guide. 2017. Conversion Chart for Blood Sugar Levels: mg/dL to mmol/L. Available from: http://www.type2diabetesguide.com/conversion-chart-for-blood-sugar-levels.shtml#.WrJpqP5LH3U [Accessed 27 March 2018].
- Heart UK, 2017. Risk charts. How to use the Cardiovascular Disease Risk Prediction Charts* for Primary Prevention. Available from: https://heartuk.org.uk/healthcare-professionals/resources-and-publications/risk-charts [Accessed 08 April 2018].
  NHS choices, 2017. Cardiovascular disease. Available from: https://www.nhs.uk/conditions/Cardiovascular-disease/ [Accessed 08 April 2018].

Table 1. Unadjusted analysis presenting the association of individual risk factors with Hospitalisation for myocardial infarction or death from coronary heart disease (CVD).

Variables	No. of individuals N=4215	No. of CVD cases (%) N=557	Unadjusted OR (95% CI), p-value
blood glucose (mg/dl)	.	.	.
continuous	3828	.	1.008 (1.005-1.011, p=**0.001**)
categorical:	.	.	.
<70mg/dl (low blood sugar (hypoglycemia))	892	112 (12.56)	1.0 (Ref)
70/100mg/dl (normal blood sugar)	2627	373 (14.20)	1.15 (0.92-1.45, p=0.219)
101/125mg/dl (pre-diabetic blood sugar)	226	39 (17.26)	1.45 (0.98-2.16, p=0.066)
>126mg/dl (diabetic blood sugar)	83	33 (39.76)	4.60 (2.84-7.44, p=**0.001**)
missing	387	45 (11.63)	0.92 (0.63-1.32, p=0.642)
age of participants (years)	.	.	.
continuous	4215	.	1.04 (1.03-1.05, p=**0.001**)
categorical:	.	.	.
<45yrs	1424	145 (10.18)	1.0 (Ref.)
45/50yrs	789	93 (11.79)	1.18 (0.89-1.55, p=0.244)
50/55yrs	699	116 (16.60)	1.76 (1.35-2.28, p=**0.001**)
55/60yrs	623	109 (17.50)	1.87 (1.43-2.45, p=**0.001**)
>60yrs	680	139 (20.44)	2.27 (1.76-2.92, p=**0.001**)
Sex	.	.	.
Female	2405	210 (8.73)	1.0 (Ref)
Male	1810	392 (21.66)	2.89 (2.41-3.46, p=**0.001**)
body mass index	.	.	.
continuous	4198	.	1.07 (1.05-1.09, p=**0.001**)
categorical:	.	.	.
underweight	59	5 (8.47)	1.0 (Ref.)
healthy weight	1820	181 (9.95)	1.19 (0.47-3.02, p=0.710)
overweight	1772	307 (17.33)	2.26 (0.90-5.70, p=0.083)
obese	519	101 (19.46)	2.61 (1.02-6.69, p=**0.046**)
very obese	28	4 (14.29)	1.8 (0.44-7.30, p=0.411)
missing	17	4 (23.53)	3.32 (0.78-14.13, p=0.104)
currently smoking	.	.	.
no	2127	266 (12.51)	1.0 (Ref)
yes	2088	336 (16.09)	1.34 (1.13-1.60, p=**0.001**)
cigarettes per day	.	.	.
continuous	4215	.	1.02 (1.01-1.02, p=**0.001**)
categorical:	.	.	.
<10 perday	2754	338 (12.27)	1.0 (Ref.)
10/20 per day	1001	175 (17.48)	1.51 (1.24-1.85, p=**0.001**)
20/40 per day	383	76 (19.84)	1.77 (1.34-2.33, p=**0.001**)
>40 per day	77	13 (16.88)	1.45 (0.79-2.66, p=0.229)

contd. on next page

Variables	No. of individuals N=4215	No. of CVD cases (%) N=602	Unadjusted OR (95% CI), p-value
education status	.	.	.
continuous	4215	.	0.95 (0.87-1.04, p=0.251)
categorical:	.	.	.
0-11yrs	1810	284 (15.69)	1.0 (Ref.)
high school	1246	162 (13.00)	0.80 (0.65-0.99, p=**0.039**)
some college	686	80 (11.66)	0.71 (0.54-0.92, p=**0.011**)
college grad or higher qual	473	76 (16.07)	1.03 (0.78-1.36, p=0.841)
prior hypertension	.	.	.
no	2916	326 (11.18)	1.0 (Ref.)
yes	1299	276 (21.25)	2.14 (1.80-2.56, p=**0.001**)
death	**.**	**.**	.
no	2827	160 (5.66)	1.0 (Ref.)
yes	1388	442 (31.84)	7.79 (6.41-9.47, p=**0.001**)

Note: p-values indicating an association with CVD are highlighted in bold.

Table 2. Adjusted OR logistic analysis presenting the association of multiple imputed continuous and categorical blood glucose with hospitalisation for myocardial infarction or death from coronary heart disease. Also presented, is the multiple imputed Cox and Weibull survival analysis and the unadjusted Cox regression tests of proportional-hazards assumption (global test).

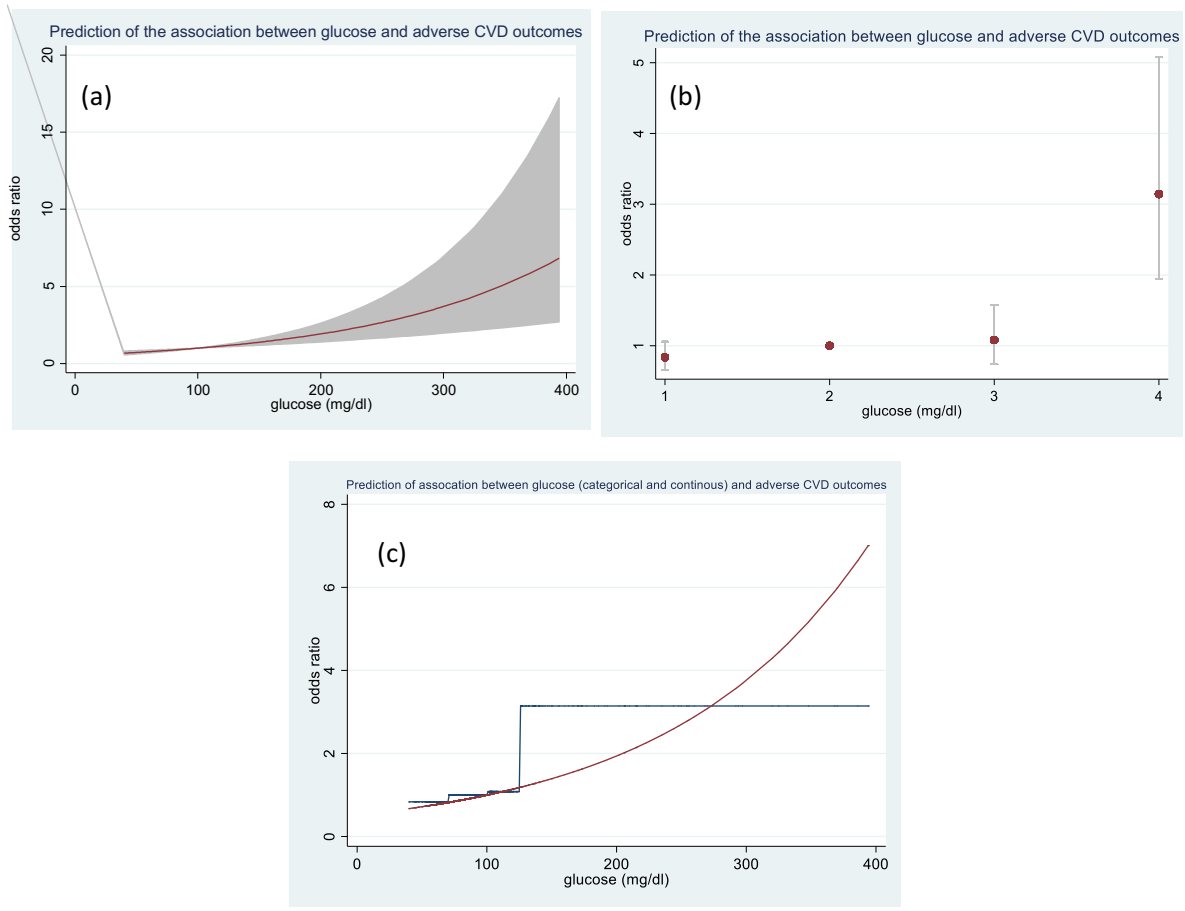
Variables	**(95% CI, p-value)**	
blood glucose (mg/dl)	**Adjusted CCA approach (OR)**	**Adjusted CCA approach with MI for glucose and BMI covariate (OR)**
continuous	1.006 (1.004-1.009, p=**0.001**)	1.006 (1.004-1.009, p=**0.001**)
categorical:	.	.
<70mg/dl (low blood sugar (hypoglycemia))	0.84 (0.66-1.06, p=0.136)	0.84 (0.66-1.08, p=0.173)
70/100mg/dl (normal blood sugar)	1.0 (Ref.)	1.0 (Ref.)
101/125mg/dl (pre-diabetic blood sugar)	1.08 (0.74-1.57, p=0.685)	1.08 (0.73-1.56, p=0.725)
>126mg/dl (diabetic blood sugar)	3.14 (1.95-5.08, p=**0.001**)	3.12 (1.94-5.02, **p=0.001**)

	Cox regression with MI applied (HR)	**Weibull survival analysis with MI applied (HR)**
continuous	1.07 (1.005-1.009, p<**0.001**)	1.07 (1.005-1.009, p<**0.001**)
categorical:	.	.
<70mg/dl (low blood sugar (hypoglycemia))	0.84 (0.67-1.05, p=0.122)	0.83 (0.67-1.04, p=0.112)
70/100mg/dl (normal blood sugar)	1.0 (Ref.)	1.0 (Ref.)
101/125mg/dl (pre-diabetic blood sugar)	1.05 (0.75-1.47, p=0.759)	1.05 (0.75-1.47, p=0.777)
>126mg/dl (diabetic blood sugar)	3.05 (2.42-5.10, p<**0.001**)	3.53 (2.43-5.12, p<**0.001**)

Test of proportional-hazards assumption (global test)	**p-value**
Unadjusted Cox regression:	
four glucose categories	0.73
two glucose categories	0.45

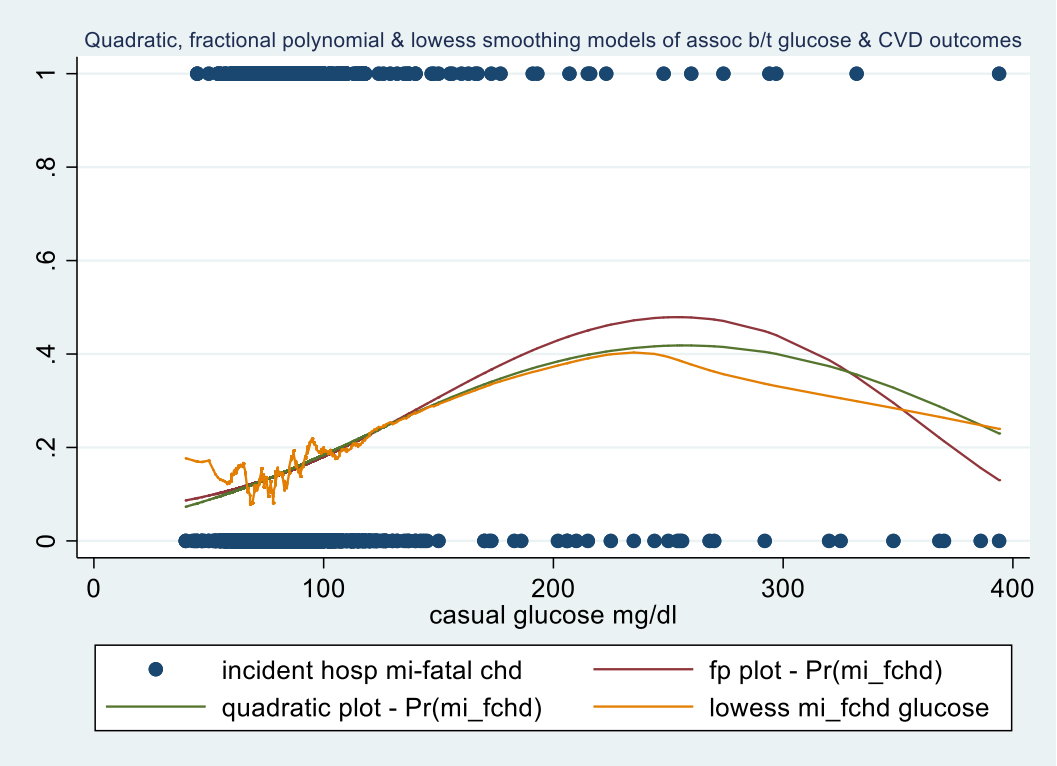
*****

Figure 1. Adjusted CCA plots of (a) continuous, (b) categorical and (c) combined categorical and continuous blood glucose association with hospitalisation for myocardial infarction or death from coronary heart disease.



Note: CVD (cardiovascular disease) is used as an umbrella term for hospitalisation for myocardial infarction or death from coronary heart disease in the plot headers.

Figure 2. Quadratic and fractional polynomial plots presenting the association between glucose and hospitalisation for myocardial infarction or death from coronary heart disease.



Note: CVD (cardiovascular disease) is used as an umbrella term for hospitalisation for myocardial infarction or death from coronary heart disease in the plot headers.

*****