# Identifying Characteristics of Brain Health with Survival Modelling

**Hannah Ramcharan** HKRAMCHA@MIT.EDU
*Computational Biology*
*MIT*
*Cambridge, MA*

**Rui-Jie Yew** RJY@MIT.EDU
*Technology and Policy Program*
*MIT CSAIL*
*Cambridge, MA*

**Livia Parodi** PARODIL@MIT.EDU
*Department of Neurology*
*Massachusetts General Hospital*
*Boston, MA*

## Abstract

In this work, we apply survival modeling methods on UK Biobank data (UKBB) to evaluate and optimize the McCance Brain Care Score (M-BCS), a new tool recently developed by the McCance Center for Brain Health (Massachusetts General Hospital, Boston) to predict risk of stroke and cognitive decline. We use survival models to gain insight about longitudinal risk for developing dementia and stroke. Applying a non-paramterized Kaplan Meier estimator on the stratified data elucidates how single features could contribute to risk of developing either disease over time. Additionally, applying a semi-parameterized Cox Proportional Hazards model (CoxPH) allow for personalized patient information to be captured and used to predict time to developing disease. *Our results validate the utility of the M-BCS in predicting risk of stroke/dementia and identifies additional predictive features that could potentially be included in a newly optimized score.*

## 1. Introduction

### 1.1. Clinical Importance

Brain health is defined as the preservation of optimal brain integrity and cognitive function (Wang et al., 2020). Neurological manifestations such as cerebrovascular disorders and dementias are the largest contributors to brain health deterioration, affecting 523(Roth et al., 2020) and 443 million(Cooper, 2019) people globally(World Health Organization et al., 2004). Effective primary prevention strategies could greatly reduce the impact caused by brain health conditions on society and healthcare (Cooper, 2019; Virani et al.; Wimo et al., 2017), since up to 40% of all cases of dementia (Livingston et al., 2020) and 80% of all strokes (Boehme et al., 2017) can be attributed to modifiable risk factors.

The increasing availability of large clinical datasets and the use of innovative Machine Learning (ML) techniques in biomedicine provide a unique opportunity to boost the efficacy of preventive medicine strategies, e.g., through better preemptive evaluations of brain

health. ML has recently been employed to improve upon traditional risk assessment methods in various disorders, including cerebrovascular diseases (Heo et al., 2020; Ambale-Venkatesh et al., 2017; Jamthikar et al., 2019), dementia, and cognitive decline (Aschwanden et al., 2020; Danso et al., 2021; Jia et al., 2020).

*In this project, we focus on the application of ML methods to the evaluation and optimization of established risk scores for brain health. This has clinical significance in encouraging preemptive measures in the maintenance and improvement of brain health.*

### 1.2. Background and Motivation

Leveraging the richness of the features and patient population in the UK Biobank (UKBB) dataset, we evaluate the utility of the newly developed McCance Brain Care Score (M-BCS) (Massachusetts General Hospital McCance Center for Brain and Health, 2020) by comparing its performance to the Framingham Stroke Risk Score, which is a widely established tool to predict stroke and dementia (Belanger, 1991; Kaffashian et al., 2013). In this section, we provide a general overview of the UK BioBank dataset, the two risk scores, and motivate our work. Designed to be transferable to any health care setting, the M-BCS (Massachusetts General Hospital McCance Center for Brain and Health, 2020) is an 11-item score including modifiable features that can be grouped into three main categories - Physical, Lifestyle, and Social Emotional. The BCS targets well-known risk factors for both cardiovascular disorders and cognitive decline. Along with biological measures indicating the presence of pre-existing disorders, the M-BCS includes factors related to individuals' lifestyles and social status. With a total score ranging from 0 to 21, higher M-BCS scores represent optimal brain health care, providing positive feedback that could stimulate patients' motivation to keep taking care of their brain health, increasing the score over time.

The UK Biobank (UKBB) is a large population-based prospective study including 500,000 participants aged 40-69 years when recruited in 2006-2010 (Ollier et al., 2005; UK Biobank, 2006). Health status, physical measures, hematological assays, and a questionnaire addressing, among others, lifestyle, cognitive function, psychosocial and sociodemographic factors, are only some of the exposures collected at baseline assessment, for each participant enrolled in the study. Data derived from follow-up visits performed every 2 years are also available. Along with extensive and precise assessment of exposures, UKBB data include characterization of various algorithmically-derived outcomes, including stroke and dementia (Biobank, 2022).

The UKBB data offers an opportunity to build and validate a score comprised of many of the ideal BCS features, some of which may ultimately have proxies that can be labeled from EHR data in future iterations. The combination of this rich dataset with survival analysis and ML approaches we learned in class hold opportunities to generate an optimized BCS version, the UKBB-BCS. This new UKBB-BCS will potentially include new, more easily quantifiable exposures, improving the overall score's predictive power in maintaining brain health.

### 1.3. Contributions

Our results extend the features recognized as important as part of the established Framingham stroke risk score, validates some of the features that are included as part of the

McCance Brain Care Score, and point to future directions. We found several significant features that were not previously included in the Framingham risk score, and some significant features that were not included in the BCS. As expected (Belanger, 1991), we found that the features included as part of the Framingham Risk Score are a good predictor of dementia and stroke. Additionally, we found that some of the additional features included as part of the M-BCS are also good predictors of stroke. We also found some features not included in the BCS that were significant. Through an exploratory analysis with the UK Biobank Data, our results validate the features included in the M-BCS and suggest the possibly increased utility to the brain care score in the exploration of additional features.

## 2. Methods

### 2.1. Predictive Analysis with Logistic Regression

A logistic regression analysis was performed on the data to predict onset of disease given patient features. Disease outcome was measured by a binary variable indicating if the individual developed the disease or not. To reduce redundancy, a Pearson's Correlation test was performed to test for highly correlated features. To fairly evaluate the performance, the data was randomly split into training and testing data with an 80%:20% ratio. It is important to transform the features to a consistent scale. Both the mean and standard deviations for the numerical values in each feature are calculated. These values are thus used to transform the data by subtracting the mean and scaling by the variance such that the transformed values have a mean of 0 and a variance of 1. After applying the scaling transform to the training dataset, it was then applied to the test set. This transform was applied using `scikit-learn`'s `preprocessing.StandardScaler` (Pedregosa et al., 2011). The model was trained using an L2 penalty, 0.0001 for tolerance for early stopping, and a Limited-memory BFGS solver. Model performance was assessed by performing a K-fold cross validation. The data was split into 5 equal seized folds and for each iteration we train on four folds and test on one. Separate experiments were done for stroke and dementia samples. The cross validated AUC for our logistic model was 0.73 with a standard deviation of .007 and 0.812 with a standard deviation of 0.006 and of for stroke and dementia respectively. To understand how each feature contributed to risk of developing disease, the model coefficients were extracted.

### 2.2. Modeling Risk with Kaplan Meier and Stratification

$$\chi^2(\text{log rank}) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{1}$$

We generated Kaplan Meier (KM) curves to assess the time to developing disease risk for stroke and dementia. KM plots are a simple way to measure the time varying survival outcomes of the data considering small time intervals to the next event. Time to event was calculated by taking the difference between the censor date and enrollment date. The event that was measured in this analysis was if the individual developed the disease or the time that they left the study which is the censored date.

Next, we wanted to quantify the effect of individual features on survival. Using the Kaplan Meier estimator, this can be done for one feature at a time by generating KM curves on

the population stratified by the feature of interest and evaluating the statistical difference of the curves using a log rank test as shown in (1). The most positive or negative predictive features from the logistic regression analysis and the CoxPH models (as discussed below) were used as a proxy to stratify the population. Features that were not discrete and binary required more deliberation about how to stratify. Continuous and non binary categorical ordinal variables were split using the upper(.8) and lower(.2) quantiles of the distribution. We decided to use the samples that had values in the upper and lower extremes of the feature distribution to magnify the discriminatory affect on the continuous variable. The KM plots were then generated for these strata and a log rank test was performed to evaluate each variable's effect on survival. The null hypothesis is that there is no statistically significant difference between the two survival curves (Clark et al., 2003). $\chi^2$ value is calculated as shown in Equation (1) where the variables $O_1$ and $O_2$ are the total number of events that are observed and the variables $E_1$ and $E_2$ are the number of expected events. This value is then compared to a $\chi^2$ distribution with one degree of freedom (Bewick et al., 2004). These calculations were performed using the `survival_difference_at_fixed_point_in_time_test` from the `lifelines` package (Davidson-Pilon, 2019). It is important to note that an assumption of the log rank test is that of proportional cumulative hazards, which will be discussed in a latter section.

### 2.3. Defining Survival Risk With CoxPH Modeling

Using the log rank test with the Kaplan Meier estimate is one way to measure the effect of patients features on survival. However, this method only allows us to evaluate the differences between groups of samples based on one feature at a time. The Cox Proportional Hazards model can use both patient features and time to event to model survival. It is preferred over a parametric model because it approximates functional form with an unknown baseline hazard (Benítez-Parejo et al., 2011). A maximum likelihood estimation is used to obtain regression coefficients as well as a $p$ value which indicates the significance of any interaction that was found between the feature and survival. We wanted to test the effect of different features from the UKBB compared to other data sets like the McCance Brain care Features. We fit each of these data sets using the CoxPHFitter from the lifelines package to calculate the regression coefficients as well as their corresponding $p$ values (Davidson-Pilon, 2019). One way to compare model performance is to look at the concordance index which measures the probability of the model correctly ranking the risks of two randomly sampled patients. Concordance is a probability score so that value ranges from zero to one. Another metric to gauge model performance is the Akaike Information Criterion which is (AIC) which measures the amount of information needed to make precise predictions. Thus, a lower AIC means a better model. This value is calculated by using the maximum likelihood estimate and number of features (Kallogjeri et al., 2012). Both the concordance and the AIC score are calculated using the CoxPHFitter method from the lifelines package (Davidson-Pilon, 2019).

$$h(t) = -\frac{d}{dt}[\log S(t)] \tag{2}$$

$$H(t) = -\log S(t) \tag{3}$$

The Cox Proportional Hazards (CoxPH) model relies on the assumption that the hazards for any two individuals in the data stay constant or proportional over time (Bewick et al., 2004). This is a key assumption of the CoxPH model and any minor violations can result in poor model performance or information loss. We graphically tested this assumption by first looking at the hazard function that is generated by the CoxPH survival model in Equation (2). Computing this hazard function is a challenging mathematical problem thus, the cumulative hazards function is used instead. This is calculated by enumerating the area under the hazards curve, or by taking its integral with respect to time (Bewick et al., 2004). The result of this calculation yields Equation (3) which is easy to compute once we know that survival function, s(x). Using this mathematical simplification is beneficial because it allows for the computation of the hazard using s(x) which is readily available. If the constant hazards assumption holds true, then the hazards for any two individuals with different features does not change over time. We show this graphically when plotting equation 2. If the slopes between the cumulative hazards functions are parallel then the assumption holds true (Bewick et al., 2004).

Another idea that supports the constant cumulative hazards assumption is that of time invariant covariates. To further test the cumulative hazards assumption, we then applied a Schoenfeld residuals test to test how each variable changes over time. This plot and calculation was done by using `check_assumptions()` from the `lifelines` package on the fitted Cox model (Davidson-Pilon, 2019).

## 3. Cohort

In this section, we describe the composition of our dataset and the features we had access to. Namely, we provide descriptions of how we chose our controls and event patient groups, as well as how the features in the UKB compare to our features in existing risk scores.

### 3.1. Cohort Selection

We split our cohort into three main groups:

1. **Controls** ($n = 450379$): The controls never had dementia or stroke and did not die during the study. We made the decision to not include patients who died, because we had a lot of controls and the tool that we tried to develop is a preventative tool. We might want to keep optimizing the preventative tool. By keeping only the people who are alive, there is the potential to add to the same data, and have follow-up to the same patients.

2. **Dementia** ($n = 3990$): We specified the dementia event group such that it includes patients who were diagnosed with incident dementia (diagnosis made after enrollment).

3. **Stroke** ($n = 8585$): We specified the stroke event group such that it includes patients who were diagnosed with incident stroke (diagnosis made after enrollment).

We ran our Cox Proportional-Hazards (CoxPH) model with all the groups. However, for the Kaplan-Meier curves we took a different approach.

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) x \hat{Pr}(T > t_{(f)} | T \geq t_{(f)}) \tag{4}$$

Based on Equation 4, we are calculating points in the Kaplan-Meier curve by looking at the number of people still at risk in the denominator. The large number of controls makes the number of people still at risk very large at every time point. This makes the the effect of people getting diagnosed with dementia or stroke (shrinking numerator) very small or negligible and made it seem like there was a high survival rate. We care about modeling time to event for patients so we wanted to ensure that the effect of people getting the disease was observable. *For this reason, we randomly sampled from the controls so that we had an equal number of cases and controls in the CoxPH analysis.*

### 3.2. Data Extraction

| | | Original BCS | | UKBB BCS | Framingham Stroke Risk Score |
|---|---|---|---|---|---|
| | | Age | | Age | Age |
| | | Sex | | Sex | Sex |
| Physical | Blood Pressure | Diastolic BP | | Diastolic BP | Diastolic BP |
| | | Systolic BP | | Systolic BP | Systolic BP |
| | Blood sugar | Hemoglobin A1c | | Hemoglobin A1c | Diabetes |
| | | | | Glucose | |
| | Cholesterol | Cholesterol | | HDL cholesterol | Total cholesterol + HDL |
| | | | | LDL cholesterol | |
| | BMI | BMI | | BMI | |
| | | | | ***C-reactive protein*** | |
| | | | | ***Creatinine*** | |
| Lifestyle | Nutrition | Dietary habits (adherence to Mediterranean diet) | | ***Salad / raw vegetable intake*** | |
| | | | | ***Fresh fruit intake*** | |
| | | | | ***Oily fish intake*** | |
| | | | | ***Non-oily fish intake*** | |
| | | | | ***Processed meat intake*** | |
| | Alcohol consumption | N of alchoholic drinks per week | | Alcohol intake frequency | |
| | Smoking | Current smoker | | Current tobacco smoking | Current smoking |
| | | | | Smoking status | |
| | Aerobic activities | Weekly moderate/high intensity physical activity | | Number of days/week of moderate physical activity 10+ minutes | |
| | | | | Number of days/week of vigorous physical activity 10+ minutes | |
| | Sleep | Sleep disorder/disturbance | | Sleep duration | |
| Social Emotional | Stress | Stress levels | | Irritability', 'Sensitivity / hurt feelings | |
| | Social relationships | N of meaningful social connections | | | |
| | Meaning in Life | Struggle to find value in life | | | |
| | | | Presence of cognitive decline | ***Mean time to correctly identify matches*** | |
| | | | | ***Number of incorrect matches in round*** | |

Figure 1: Features present in the original brain care score (M-BCS), features present in the UKB-BCS, and features present in the Framingham stroke risk score.

To evaluate the M-BCS predictive power and to generate an optimized version, we extracted features already included in the brain care score, or their proxies, as well as new features potentially belonging to the 3 main BCS categories (physical, lifestyle, social emotional). Since the M-BCS can be considered as an extension of the Framingham score, all the features included in the latter were also exported. Figure 1 provides an overview of the features that are included as part of the UKB, and how those features compare to the original BCS (M-BCS) and the Framingham Stroke Risk Score.

### 3.3. Feature Choices

For the continuous/ numeric variables, we imputed the data using the mean of the values for each variable in the dataset. For the categorical variables, we imputed the data using the median of the values for each variable in the dataset. To do this, we used `simpleimputer`, with `strategy = "most_frequent"` for the categorical variables and `strategy = "mean"` for the continuous variables. In future work, more sophisticated imputation techniques can be utilized, such as imputing missing weight values using mean values based on sex.For the purposes of an exploratory analysis, we opted to perform general imputing on the data.

## 4. Results

### 4.1. Logistic Regression Predicts Disease Outcome

A logistic regression model was applied to our data to predict disease outcome given patient features. The model coefficients for the logistic regression were extracted and used to quantify the predictive ability of each feature shown in Figure 2. Coefficients other than age have a similar value which means that they all contribute similarly to risk of developing stroke. Age would make sense in this context to have a large coefficient because the risk of developing disease such as Stroke and Dementia increases over time (Wang et al., 2020). Although the model succeeds in providing meaningful predictions for if an individual will develop a disease, it is oftentimes more useful to predict when the disease will occur in one's lifetime. For this reason we then took a survival modeling approach.

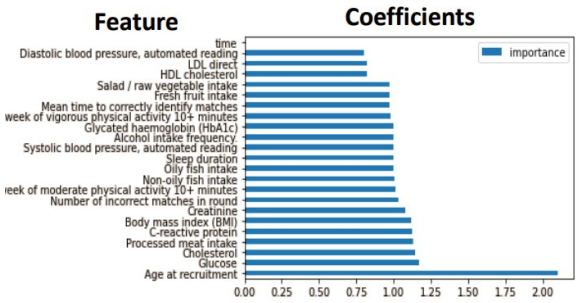


Figure 2: Feature Importance by Coefficient from the Logistic Regression Model.

### 4.2. Modeling Time to Developing Stroke or Dementia

First, we modeled survival using the Kaplan Meier (KM) estimator which is a non-parametric time discretized model. We generated a KM plot for the Dementia patients as well as the stroke patients, pictured in Figure 3. When looking at the curve generated for dementia, it appears that the slope decreased quicker than that of stroke. This means that the survival rate for dementia is much lower than that of stroke. This is a finding that aligns with previous literature that claims that patients can typically recover from stroke in a finite time, whereas the effects of dementia are more pronounced (Isaacson et al., 2019).

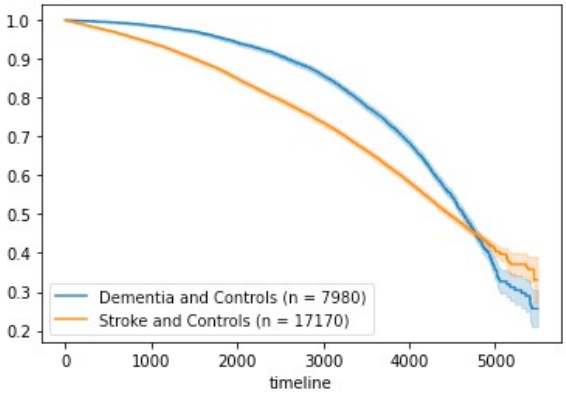


Figure 3: Kaplan-Meier curves for dementia and stroke.

### 4.3. Identifying Significant Features and Defining Risk with CoxPH Modeling

Table 1: Significant Features with CoxPH Modelling

| **Features** | **Event** |
|---|---|
| Sex | Dementia, Stroke |
| Age | Dementia, Stroke |
| Smoking Status | Dementia, Stroke |
| Current Tobacco Smoking | Stroke |
| **Alcohol Intake Frequency** | Dementia |
| **C-Reactive Protein** | Stroke |
| **Creatinine** | Stroke |
| Glucose | Stroke |
| Glycated haemoglobin | Dementia, Stroke |
| HDL cholesterol | Stroke |
| LDL direct | Stroke |
| Diastolic Blood Pressure | Stroke |
| Weight | Stroke |

Modelling survival using multiple features at time in order to find significant features, we used CoxPH. We fit the model on both the stroke and dementia samples separately using both continuous and categorical features. The significant features are detailed in Table 1.

*Importantly, we find possible correlative significance for dementia diagnosis in* ***Creatinine****, and* ***C-Reactive protein****. Additionally we find possible correlative significance for stroke diagnosis in* ***Alcohol intake frequency****. While alcohol is considered in BCS, it is not considered in Framingham. Thus, our results also validate the existence of alcohol in the BCS score. The significance of Creatinine and C-Reactive protein measures are neither included in the BCS nor in the Framingham risk score.* There is additionally research in the medical community that points to the relationship between c-reactive protein and cre-

atinine and brain health (Bettcher et al., 2012; Roschel et al., 2021). A table detailing the full results can be found in Figure 12 and Figure 13.

The survival plot for the first four stroke patients are shown in Figure 14. We observe that the risk for these patients could vary across individuals. The variability seen here can be attributed to the patient's features. In contrast, when looking at the CoxPH survival curve for the first four dementia patients appear to be very close to each other Figure 14. Despite individual values for the given features, survival is not affected that much across individuals for dementia time to event as compared to that of stroke.

A similar pattern can be observed when regression coefficients for the CoxPh model. For stroke the highest coefficients were sex, age, and smoking status($p$-value $< .005$). These features had the most impact on differentiating the survival outcome for stroke patients. When looking at the dementia cohort, most of the coefficients are close to 0. Among the non-zero regression coefficients recorded were for age, BMI, clinical weight, and smoking status ($p$-value $< .005$).

To evaluate the model performance on the data we measured the concordance index and the AIC shown in Figure 15. We additionally compared the performance of the CoxPH model across two different cohorts, the McCance Brain Care (MBC) and Framinham Cohorts (FC). The UKBB features perform at the approximately at level of the MCB and FC with a concordance and AIC of .81 and 59169 for dementia samples respectively and .73 and 147626 on the stroke samples respectively.

### 4.4. Evaluating Survival Risk with Stratification

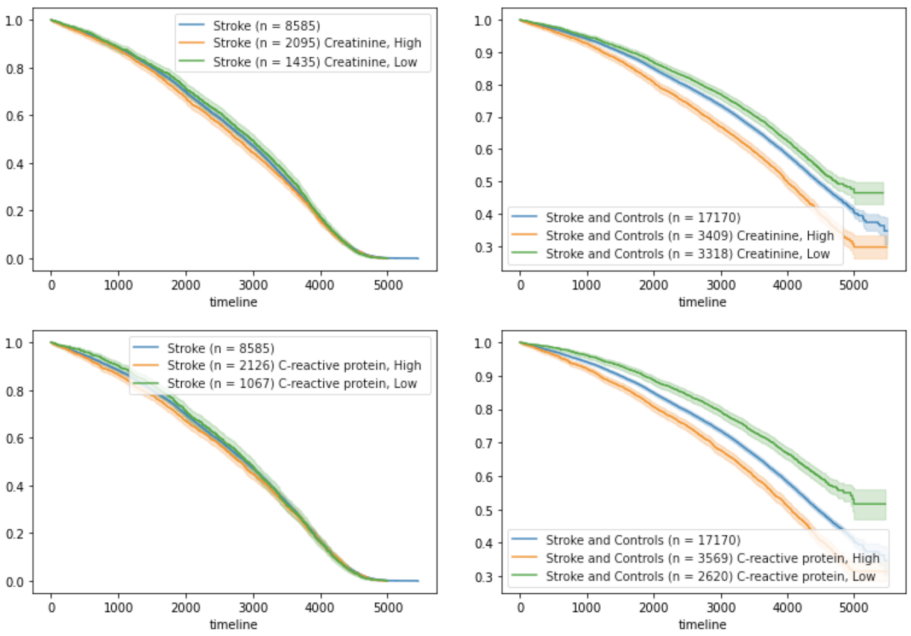


Figure 4: Stratification for new dementia features. CoxPH (left) KM (right).

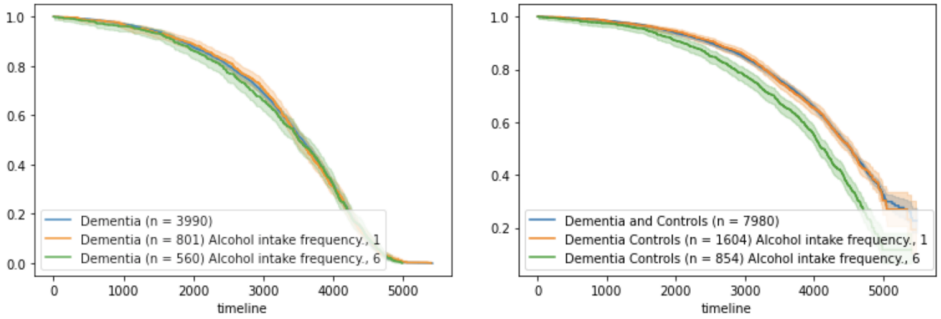


Figure 5: Stratification for new stroke features. CoxPH (left) KM (right).

To understand how the patient's features could affect time to developing dementia or stroke, the cases and controls were split into groups depending on their value of a particular variable. We show the Kaplan-Meier stratifications for the new features discussed in Section 4.3 in Figures 4 and 5.

For stroke samples, we stratified on alcohol intake frequency. A KM fitter was applied pictured in Figure 4 to the groups and the resulting log rank test was performed with a resulting $p$ value of $.330 > .05$. The same steps were performed to stratify dementia patients shown in Figure 5 and controls for creatinine (p-value $0.004 < .05$) and, C-reactive protein: (p-value $0.023 < .05$). We can reason with confidence that that patients with high levels of both creatinine and c-reactive protein had a higher probability of developing stroke within a fixed time range. The CoxPH models are also shown for the subgroups discussed above in Figures 4 and 5.

### 4.5. Evaluating the Constant Proportional Hazards Assumption

The Cox Proportional Hazards model relies on the assumption that the proportional hazards equation stay constant over time. The cumulative hazard function is used which can be calculated by taking the derivative of the hazard function . To test if this assumption holds for our data, we calculated the cumulative hazards as shown and plot it against the survival time. If there is a constant hazard over time, then the slopes of these functions should be parallel. Figure 7 shows these functions with the x and y axes on a log scale. The cumulative hazards functions appear to be roughly parallel over time. We then used the Schoenfeld residuals test to further check the constant hazards assumption for each variable to see if residuals change over time. The result of the Schoenfeld residuals test for stroke for fresh fruit intake and age are shown in Figure 18. We see that for age, there is a correlation ($p$ value $< .005$) between the Schoenfeld residuals and time as shown by the slight slope. This means that the hazard function for developing dementia changes with time. We can observe that the correlation ($p$ value .0420) of fresh fruit intake and the Schoenfeld residuals over time is not as considerable. Figure 6 shows $p$ values of the Schoenfeld residuals test where we can see that there are only a small handful of variables that violate the constant hazards function such as 'mean time to correctly identify matches', 'cholesterol', and 'salad intake'.

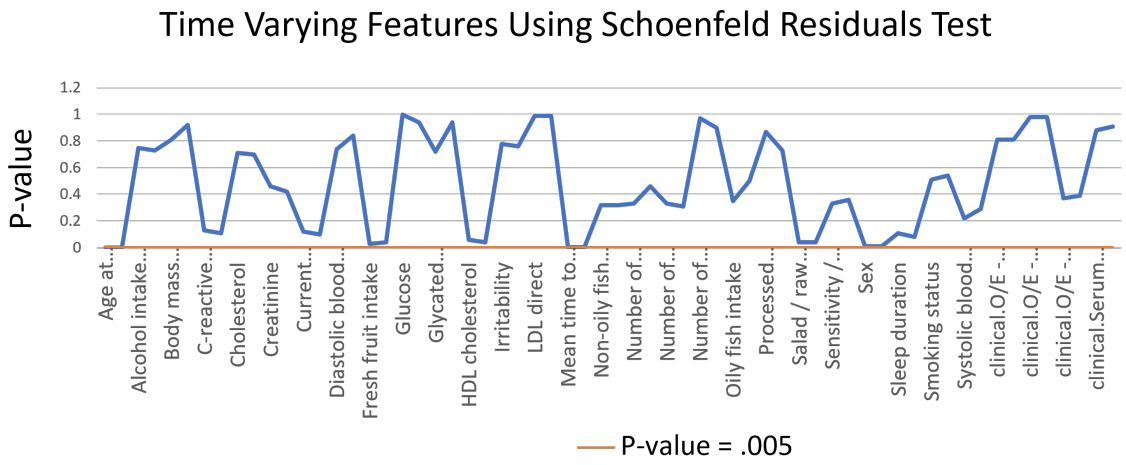


Figure 6: P values for the Schoenfeld Residuals Test

## 5. Discussion

Our analysis is a first attempt at validating/evaluating the BCS in a cohort of stroke/dementia patients. Leveraging rich datasets such as the UKBB can allow to further optimize risk scores such as the BCS, adding new features. Our preliminary results are in support of C-Reactive protein and Creatinine inclusion, in a future, optimized version of the BCS. Understanding the contribution to risk of modifiable features can have a positive impact understanding ways to improve one's health outcome. In this study we show how such modifiable features can be used to assess risk of developing Stroke and Dementia. Leveraging rich datasets such as the UKBB can allow to further optimize the risk score such as the BCS, adding new features. We show that although a logistic regression model is useful to predict onset of disease, it does not tell us information about longitudinal health. Survival models are a powerful tool that models survival using two key components, patient data and time to event.

From this exploratory analysis, we find that the M-BCS is a useful tool to predict stroke/dementia, its performance is comparable to the Framingham (considered a gold standard within the medical community) but it also adds new modifiable risk factors (e.g. alcohol, glucose/hemoglobin, BMI/weight) that are predictive of stroke/dementia risk. Thus, it expands on the Framingham features. Creatinine is a metabolite that can be used to measure renal health. Studies show that renal health dysfunction is associated with cognitive decline which may suggest that creatinine levels is a key feature that can be used for dementia diagnosis. (Elias et al., 2009). Furthermore, a logistic regression method revealed that c-reactive protein is associated with developing dementia in the elder population (Kravitz et al., 2009).

By applying the KM-fitter, we wanted to gauge an idea for the overall survival function for developing dementia and stroke. There were notable differences in the survival plots for the two diseases. A good next step here would be to develop metrics to evaluate the differences between survival functions of dementia and stroke cohorts. Additionally, we found

that stratifying the population yielded interesting information about feature contribution to time to event for developing disease. From this analysis we were able to conclude that individuals with differing levels of c-reactive protein and creatinine had different probabilities of developing dementia within a particular time span.

A number of additional modifiable risk factors, such as smoking status, were found to be significant for predicting the time to developing stroke thus validating previous brain care scores.

Survival modeling provides a powerful tool to understand how modifiable features can contribute to time to developing Stroke and Dementia.

**Limitations and Future Directions** There are several limitations to our approach. In some instances in the data, there were the absence of multiple follow-ups, e.g., there were no follow-ups before the event occurred, so we could not check the selected features' variation over time and how this variation can impact survival. Additionally, newly developed clustering methods provide an important avenue in determining significant features. We document our preliminary approach in appendices B and C. Clustering allows us to better understand how the population is clustered and how each feature (e.g. what ranges) influences the survival.

Additionally, the CoxPH model relies on the assumption that time varying variables cannot be used to predict survival risk. However these types of variables can provide useful information to understanding dementia and stroke. Next steps would be to explore a model that can handle time varying variables such as the extended CoxPH models.

## 6. Member Contributions

Hannah Ramcharan: Researched methods for survival modeling, ran PCA analysis and logistic regression, ran survival models, conducted experiments to test constant hazards assumption, analyzed and wrote up results, created and designed poster.
Rui-Jie Yew: Set up Jupyter notebook on VM machine, cleaned and provided data, performed data analysis for checkpoint 1, ran KM and CoxPH analyses, attempted implementation of clustering algorithms, created Overleaf document and worked on LaTeX formatting.

## 7. Acknowledgements

Livia Parodi: Ran and validated results in R, provided data with features, wrote clinical introduction of report, and created table of variables for this write-up. **Thank you to Livia Parodi for being such a present supportive project supervisor!**

## References

Bharath Ambale-Venkatesh, Xiaoying Yang, Colin O Wu, Kiang Liu, W Gregory Hundley, Robyn McClelland, Antoinette S Gomes, Aaron R Folsom, Steven Shea, Eliseo Guallar, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circulation research*, 121(9):1092–1101, 2017.

Damaris Aschwanden, Stephen Aichele, Paolo Ghisletta, Antonio Terracciano, Matthias Kliegel, Angelina R Sutin, Justin Brown, and Mathias Allemand. Predicting cognitive

impairment and dementia: A machine learning approach. *Journal of Alzheimer's Disease*, 75(3):717–728, 2020.

Wolf PA D'Agostino RB Belanger. Aj kannel wb. *Probability for stroke: a risk profile from the Framingham Study. Stroke*, 22:312–8, 1991.

N Benítez-Parejo, MM Rodríguez del Águila, and S Pérez-Vicente. Survival analysis and cox regression. *Allergologia et immunopathologia*, 39(6):362–373, 2011.

Brianne Magouirk Bettcher, Reva Wilheim, Taylor Rigby, Ralph Green, Joshua W Miller, Caroline A Racine, Kristine Yaffe, Bruce L Miller, and Joel H Kramer. C-reactive protein is related to memory and medial temporal brain volume in older adults. *Brain, behavior, and immunity*, 26(1):103–108, 2012.

Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 12: survival analysis. *Critical care*, 8(5):1–6, 2004.

UK Biobank. Algorithmically-defined outcomes. 2022.

Amelia K Boehme, Charles Esenwa, and Mitchell SV Elkind. Stroke risk factors, genetics, and prevention. *Circulation research*, 120(3):472–495, 2017.

Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

Cyrus Cooper. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 18(4):357–375, 2019.

Samuel O Danso, Zhanhang Zeng, Graciela Muniz-Terrera, and Craig W Ritchie. Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms. *Frontiers in big Data*, 4: 21, 2021.

Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019. doi: 10.21105/joss.01317. URL https://doi.org/10.21105/joss.01317.

Merrill F Elias, Penelope K Elias, Stephen L Seliger, Sriram S Narsipur, Gregrory A Dore, and Michael A Robbins. Chronic kidney disease, creatinine and cognitive functioning. *Nephrology Dialysis Transplantation*, 24(8):2446–2452, 2009.

Jaehyuk Heo, Sang Jun Park, Si-Hyuck Kang, Chang Wan Oh, Jae Seung Bang, and Tackeun Kim. Prediction of intracranial aneurysm risk using machine learning. *Scientific Reports*, 10(1):1–10, 2020.

Richard S Isaacson, Hollie Hristov, Nabeel Saif, Katherine Hackett, Suzanne Hendrix, Juan Melendez, Joseph Safdieh, Matthew Fink, Madhav Thambisetty, George Sadek, et al. Individualized clinical management of patients at risk for alzheimer's dementia. *Alzheimer's & Dementia*, 15(12):1588–1602, 2019.

Ankush Jamthikar, Deep Gupta, Narendra N Khanna, Luca Saba, Tadashi Araki, Klaudija Viskovic, Harman S Suri, Ajay Gupta, Sophie Mavrogeni, Monika Turk, et al. A low-cost machine learning-based cardiovascular/stroke risk assessment system: integration of conventional factors with image phenotypes. *Cardiovascular diagnosis and therapy*, 9(5): 420, 2019.

Yichen Jia, Chung-Chou H Chang, Tiffany F Hughes, Erin Jacobsen, Shu Wang, Sarah Berman, M Ilyas Kamboh, and Mary Ganguli. Predictors of dementia in the oldest old: A novel machine learning approach. *Alzheimer disease and associated disorders*, 34(4): 325, 2020.

Sara Kaffashian, Aline Dugravot, Alexis Elbaz, Martin J Shipley, Séverine Sabia, Mika Kivimäki, and Archana Singh-Manoux. Predicting cognitive decline: A dementia risk score vs the framingham vascular risk scores. *Neurology*, 80(14):1300–1306, 2013.

Dorina Kallogjeri, Jay F Piccirillo, Edward L Spitznagel Jr, and Ewout W Steyerberg. Comparison of scoring methods for ace-27: simpler is better. *Journal of geriatric oncology*, 3(3):238–245, 2012.

B Adar Kravitz, Maria M Corrada, and Claudia H Kawas. Elevated c-reactive protein levels are associated with prevalent dementia in the oldest-old. *Alzheimer's & Dementia*, 5(4): 318–323, 2009.

Gill Livingston, Jonathan Huntley, Andrew Sommerlad, David Ames, Clive Ballard, Sube Banerjee, Carol Brayne, Alistair Burns, Jiska Cohen-Mansfield, Claudia Cooper, et al. Dementia prevention, intervention, and care: 2020 report of the lancet commission. *The Lancet*, 396(10248):413–446, 2020.

Laura Manduchi, Ričards Marcinkevičs, Michela C Massi, Thomas Weikert, Alexander Sauter, Verena Gotta, Timothy Müller, Flavio Vasella, Marian C Neidert, Marc Pfister, et al. A deep variational approach to clustering survival data. *arXiv preprint arXiv:2106.05763*, 2021.

Massachusetts General Hospital McCance Center for Brain and Health. Mccance brain care score. 2020.

Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. *arXiv preprint arXiv:2204.07276*, 2022.

William Ollier, Tim Sprosen, and Tim Peakman. Uk biobank: from concept to reality. 2005.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Hamilton Roschel, Bruno Gualano, Sergej M Ostojic, and Eric S Rawson. Creatine supplementation and brain health. *Nutrients*, 13(2):586, 2021.

Gregory A Roth, George A Mensah, Catherine O Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M Baddour, Noël C Barengo, Andrea Z Beaton, Emelia J Benjamin, Catherine P Benziger, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American College of Cardiology*, 76(25):2982–3021, 2020.

UK Biobank. Protocol for a large-scale prospective epidemiological resource. 2006.

Salim S Virani et al. Heart disease and stroke statistics—update: A report from the american heart association. *Circulation*, 141(9).

Yongjun Wang, Yuesong Pan, and Hao Li. What is brain health and why is it important? *bmj*, 371, 2020.

Anders Wimo, Maëlenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, A Matthew Prina, Bengt Winblad, Linus Jönsson, Zhaorui Liu, and Martin Prince. The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's & Dementia*, 13(1):1–7, 2017.

World Health Organization et al. *Atlas: country resources for neurological disorders 2004: results of a collaborative study of the World Health Organization and the World Federation of Neurology*. World Health Organization, 2004.

## Figures

### .1. Evaluating Constant Hazard Assumption

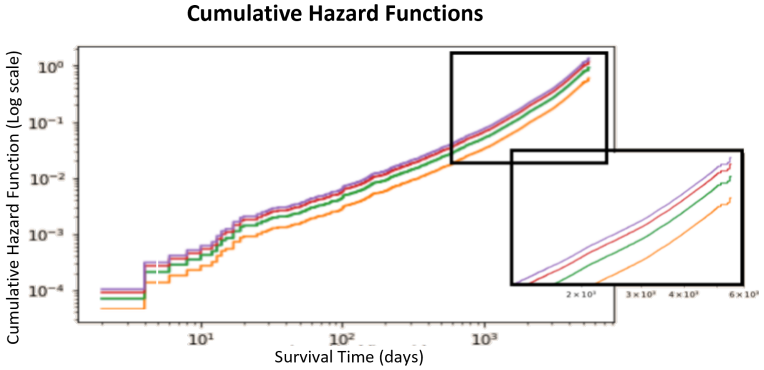


Figure 7: Cumulative Hazards Function Over Time.

| Feature | P-Values |
|---|---|
| Current tobacco smoking | 0.0601 |
| Smoking status | 0.0049 |
| HDL cholesterol | 0.0171 |
| age | 0.0004 |

Figure 8: Log Rank Test P values for stroke.

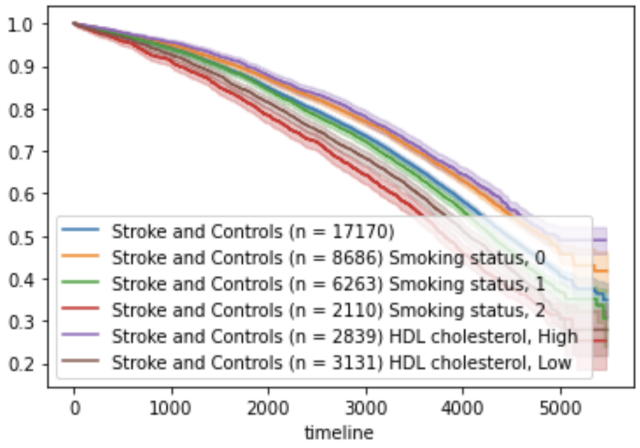


Figure 9: Stratified Kaplan-Meier curves for stroke.

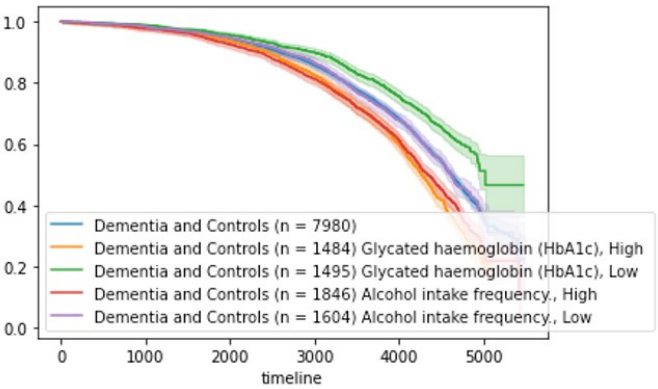


Figure 10: Stratified Kaplan-Meier curves for dementia.

| Feature | P-Values |
|---|---|
| Number of incorrect matches in round | 0.5045 |
| Glycated hemoglobin | 0.9857 |
| Smoking status | 0.3359 |
| Alcohol intake frequency | 0.2588 |
| HDL cholesterol | 0.6428 |

Figure 11: Log Rank Test P values for dementia.

| | coef | exp(coef) | se(coef) | coef lower 95% | coef upper 95% | exp(coef) lower 95% | exp(coef) upper 95% | cmp to | z | p | -log2(p) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sex** | 0.32 | 1.38 | 0.05 | 0.22 | 0.43 | 1.25 | 1.53 | 0.00 | 6.26 | <0.005 | 31.27 |
| **Age at recruitment** | 0.17 | 1.19 | 0.00 | 0.16 | 0.18 | 1.18 | 1.20 | 0.00 | 40.60 | <0.005 | inf |
| **Body mass index (BMI)** | -0.01 | 0.99 | 0.02 | -0.05 | 0.04 | 0.95 | 1.04 | 0.00 | -0.36 | 0.72 | 0.48 |
| **Mean time to correctly identify matches** | 0.17 | 1.19 | 0.02 | 0.14 | 0.20 | 1.15 | 1.22 | 0.00 | 10.39 | <0.005 | 81.56 |
| **Number of incorrect matches in round** | 0.06 | 1.06 | 0.02 | 0.03 | 0.09 | 1.03 | 1.09 | 0.00 | 3.56 | <0.005 | 11.40 |
| **Diastolic blood pressure, automated reading** | -0.02 | 0.98 | 0.02 | -0.06 | 0.02 | 0.94 | 1.02 | 0.00 | -0.91 | 0.36 | 1.47 |
| **Systolic blood pressure, automated reading** | 0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| **Number of days/week of moderate physical activity 10+ minutes** | -0.04 | 0.96 | 0.02 | -0.08 | -0.00 | 0.92 | 1.00 | 0.00 | -2.03 | 0.04 | 4.56 |
| **Number of days/week of vigorous physical activity 10+ minutes** | 0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| **Smoking status** | 0.16 | 1.17 | 0.04 | 0.07 | 0.24 | 1.07 | 1.27 | 0.00 | 3.67 | <0.005 | 12.00 |
| **Current tobacco smoking** | 0.05 | 1.05 | 0.05 | -0.05 | 0.15 | 0.95 | 1.16 | 0.00 | 1.00 | 0.32 | 1.65 |
| **Sleep duration** | 0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| **Irritability** | 0.01 | 1.01 | 0.05 | -0.08 | 0.10 | 0.92 | 1.11 | 0.00 | 0.21 | 0.83 | 0.27 |
| **Sensitivity / hurt feelings** | 0.04 | 1.04 | 0.04 | -0.04 | 0.12 | 0.96 | 1.13 | 0.00 | 0.92 | 0.36 | 1.49 |
| **Salad / raw vegetable intake** | -0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | -0.00 | 1.00 | 0.00 |
| **Fresh fruit intake** | 0.01 | 1.01 | 0.02 | -0.03 | 0.05 | 0.97 | 1.05 | 0.00 | 0.57 | 0.57 | 0.81 |
| **Oily fish intake** | -0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | -0.00 | 1.00 | 0.00 |
| **Non-oily fish intake** | 0.05 | 1.05 | 0.03 | -0.00 | 0.10 | 1.00 | 1.11 | 0.00 | 1.79 | 0.07 | 3.77 |
| **Processed meat intake** | 0.02 | 1.02 | 0.02 | -0.02 | 0.06 | 0.98 | 1.06 | 0.00 | 1.01 | 0.31 | 1.69 |
| **Alcohol intake frequency.** | 0.09 | 1.10 | 0.01 | 0.07 | 0.12 | 1.07 | 1.13 | 0.00 | 7.00 | <0.005 | 38.47 |
| **C-reactive protein** | -0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | -0.00 | 1.00 | 0.00 |
| **Cholesterol** | -0.00 | 1.00 | 0.10 | -0.20 | 0.20 | 0.82 | 1.22 | 0.00 | -0.02 | 0.98 | 0.02 |
| **Creatinine** | 0.01 | 1.01 | 0.02 | -0.03 | 0.05 | 0.97 | 1.05 | 0.00 | 0.42 | 0.67 | 0.57 |
| **Glucose** | 0.04 | 1.04 | 0.02 | 0.01 | 0.08 | 1.01 | 1.08 | 0.00 | 2.22 | 0.03 | 5.25 |
| **Glycated haemoglobin (HbA1c)** | 0.08 | 1.08 | 0.02 | 0.04 | 0.11 | 1.04 | 1.12 | 0.00 | 4.33 | <0.005 | 16.05 |
| **HDL cholesterol** | 0.04 | 1.04 | 0.04 | -0.02 | 0.11 | 0.98 | 1.12 | 0.00 | 1.25 | 0.21 | 2.24 |
| **LDL direct** | -0.06 | 0.94 | 0.10 | -0.25 | 0.13 | 0.78 | 1.14 | 0.00 | -0.63 | 0.53 | 0.92 |
| **clinical.O/E - Systolic BP reading** | 0.06 | 1.06 | 0.07 | -0.09 | 0.21 | 0.92 | 1.23 | 0.00 | 0.81 | 0.42 | 1.26 |
| **clinical.O/E - Diastolic BP reading** | 0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| **clinical.Serum creatinine level** | -0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | -0.00 | 1.00 | 0.00 |
| **clinical.O/E - weight** | -0.05 | 0.95 | 0.07 | -0.20 | 0.09 | 0.82 | 1.09 | 0.00 | -0.74 | 0.46 | 1.12 |

| | |
|---|---|
| **Concordance** | 0.81 |
| **Partial AIC** | 59077.22 |
| **log-likelihood ratio test** | 3321.84 on 31 df |
| **-log2(p) of ll-ratio test** | inf |

Figure 12: CoxPH Model Dementia Features.

| | coef | exp(coef) | se(coef) | coef lower 95% | coef upper 95% | exp(coef) lower 95% | exp(coef) upper 95% | cmp to | z | p | -log |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sex** | 0.32 | 1.37 | 0.03 | 0.26 | 0.38 | 1.29 | 1.46 | 0.00 | 10.28 | <0.005 | 79 |
| **Age at recruitment** | 0.10 | 1.10 | 0.00 | 0.09 | 0.10 | 1.10 | 1.11 | 0.00 | 44.53 | <0.005 | |
| **Body mass index (BMI)** | 0.01 | 1.01 | 0.00 | 0.00 | 0.01 | 1.00 | 1.01 | 0.00 | 2.79 | 0.01 | 7 |
| **Mean time to correctly identify matches** | 0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 1.91 | 0.06 | 4 |
| **Number of incorrect matches in round** | 0.00 | 1.00 | 0.00 | -0.00 | 0.01 | 1.00 | 1.01 | 0.00 | 0.94 | 0.35 | 1 |
| **Diastolic blood pressure, automated reading** | -0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | -0.03 | 0.97 | 0 |
| **Systolic blood pressure, automated reading** | 0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 1.93 | 0.05 | 4 |
| **Number of days/week of moderate physical activity 10+ minutes** | -0.01 | 0.99 | 0.01 | -0.02 | 0.00 | 0.98 | 1.00 | 0.00 | -1.79 | 0.07 | 3 |
| **Number of days/week of vigorous physical activity 10+ minutes** | -0.00 | 1.00 | 0.01 | -0.02 | 0.01 | 0.98 | 1.01 | 0.00 | -0.50 | 0.62 | 0 |
| **Smoking status** | 0.08 | 1.08 | 0.03 | 0.02 | 0.13 | 1.03 | 1.14 | 0.00 | 2.85 | <0.005 | 7 |
| **Current tobacco smoking** | 0.34 | 1.41 | 0.03 | 0.28 | 0.40 | 1.33 | 1.49 | 0.00 | 11.29 | <0.005 | 95 |
| **Sleep duration** | -0.01 | 0.99 | 0.01 | -0.03 | 0.01 | 0.97 | 1.01 | 0.00 | -1.37 | 0.17 | 2 |
| **Irritability** | 0.02 | 1.02 | 0.03 | -0.04 | 0.08 | 0.96 | 1.09 | 0.00 | 0.73 | 0.47 | 1 |
| **Sensitivity / hurt feelings** | 0.05 | 1.05 | 0.03 | 0.00 | 0.11 | 1.00 | 1.11 | 0.00 | 1.98 | 0.05 | 4 |
| **Salad / raw vegetable intake** | 0.00 | 1.00 | 0.01 | -0.01 | 0.01 | 0.99 | 1.01 | 0.00 | 0.23 | 0.82 | 0 |
| **Fresh fruit intake** | -0.01 | 0.99 | 0.01 | -0.02 | 0.01 | 0.98 | 1.01 | 0.00 | -0.84 | 0.40 | 1 |
| **Oily fish intake** | -0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | -0.01 | 0.99 | 0 |
| **Non-oily fish intake** | 0.02 | 1.02 | 0.02 | -0.01 | 0.06 | 0.99 | 1.06 | 0.00 | 1.32 | 0.19 | 2 |
| **Processed meat intake** | 0.02 | 1.02 | 0.01 | -0.00 | 0.05 | 1.00 | 1.05 | 0.00 | 1.73 | 0.08 | 3 |
| **Alcohol intake frequency.** | 0.04 | 1.04 | 0.01 | 0.03 | 0.06 | 1.03 | 1.06 | 0.00 | 5.03 | <0.005 | 20 |
| **C-reactive protein** | 0.01 | 1.02 | 0.00 | 0.01 | 0.02 | 1.01 | 1.02 | 0.00 | 6.63 | <0.005 | 34 |
| **Cholesterol** | -0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | -0.01 | 0.99 | 0 |
| **Creatinine** | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 7.14 | <0.005 | 39 |
| **Glucose** | 0.05 | 1.05 | 0.01 | 0.04 | 0.07 | 1.04 | 1.07 | 0.00 | 7.37 | <0.005 | 42 |
| **Glycated haemoglobin (HbA1c)** | 0.01 | 1.01 | 0.00 | 0.00 | 0.01 | 1.00 | 1.01 | 0.00 | 5.07 | <0.005 | 21 |
| **HDL cholesterol** | -0.16 | 0.85 | 0.03 | -0.22 | -0.11 | 0.80 | 0.90 | 0.00 | -5.57 | <0.005 | 25 |
| **LDL direct** | -0.08 | 0.92 | 0.01 | -0.10 | -0.06 | 0.90 | 0.94 | 0.00 | -6.74 | <0.005 | 35 |
| **clinical.O/E - Systolic BP reading** | 0.00 | 1.00 | 0.00 | -0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 0.01 | 0.99 | 0 |
| **clinical.O/E - Diastolic BP reading** | 0.30 | 1.35 | 0.05 | 0.20 | 0.40 | 1.23 | 1.49 | 0.00 | 5.99 | <0.005 | 28 |
| **clinical.Serum creatinine level** | 0.00 | 1.00 | 0.00 | -0.01 | 0.01 | 0.99 | 1.01 | 0.00 | 0.01 | 1.00 | 0 |
| **clinical.O/E - weight** | -0.27 | 0.76 | 0.05 | -0.37 | -0.17 | 0.69 | 0.84 | 0.00 | -5.46 | <0.005 | 24 |

| | |
|---|---|
| **Concordance** | 0.73 |
| **Partial AIC** | 147626.51 |
| **log-likelihood ratio test** | 4104.10 on 31 df |
| **-log2(p) of ll-ratio test** | inf |



Figure 13: CoxPH Model Stroke Features.

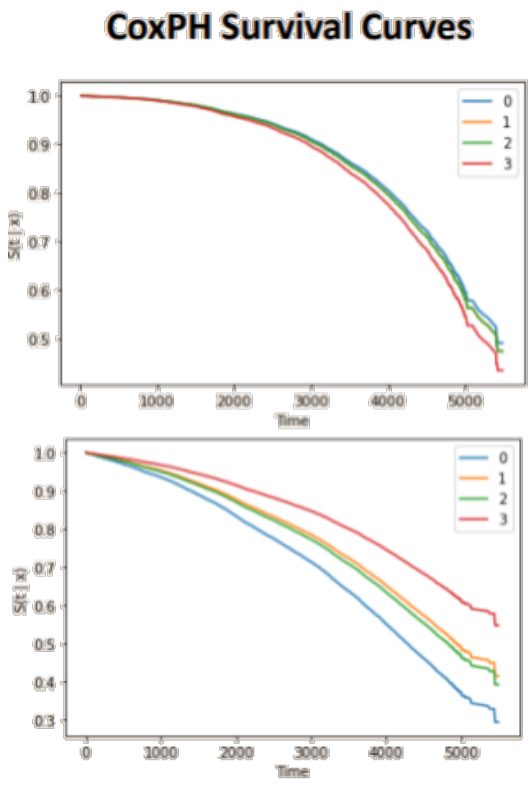


Figure 14: CoxPh Survival Curves for Dementia(top) and Stroke(bottom)

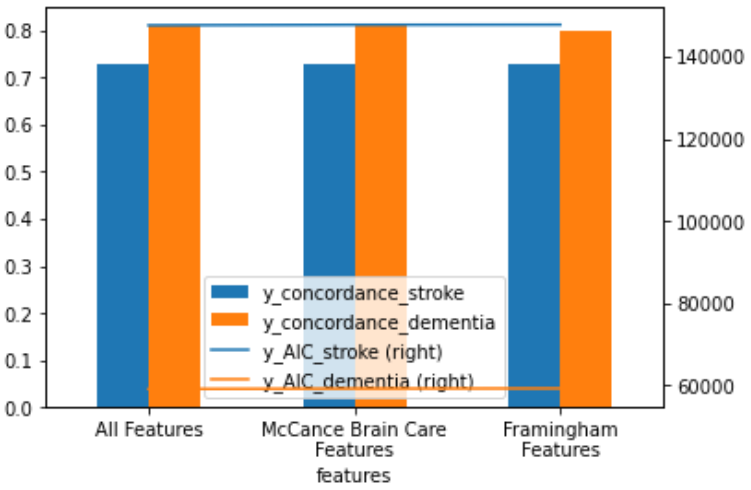


Figure 15: CoxPH Model Performance Against Multiple Datasets.

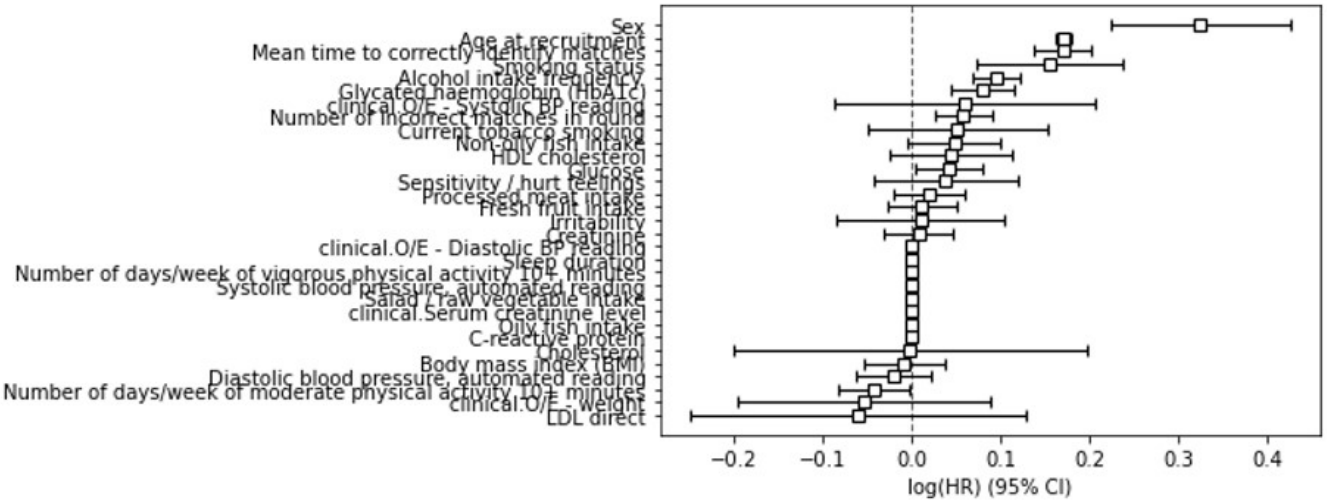


Figure 16: Coefficients for the CoxPH Model for Dementia.

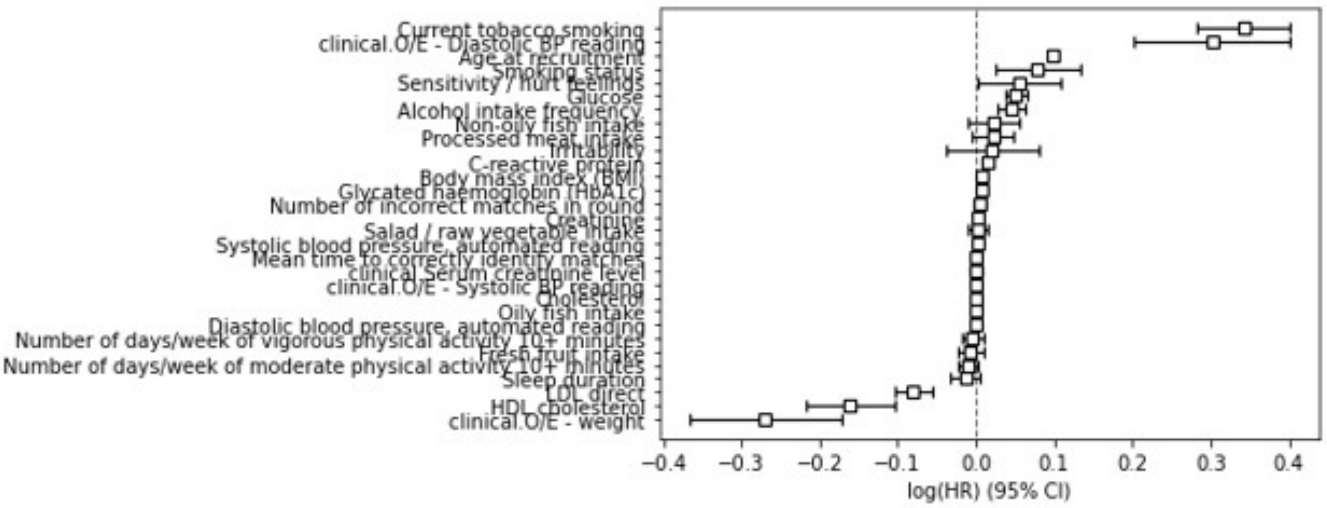


Figure 17: Coefficients for the CoxPH Model for Stroke.

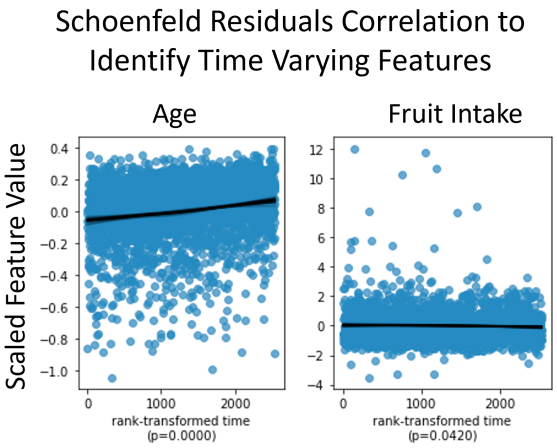


Figure 18: Schoenfeld Residuals Test

## Appendix A. Code / Approach Documentation

In this section, we detail the set up of our code and provide useful starter code to perform Kaplan-Meier and CoxPH analyses. We set up a Jupyter Notebook on the Virtual Machine to access the UK BioBank data on Google Cloud Platform's Virtual Machine instance. This is a helpful resource to get started with setting up your Jupyter Notebook on a virtual machine.

### A.1. Data Pre-Processing

As part of data preprocessing, we were provided with a csv of features that would be important to consider and csvs of patients with dementia and stroke. We merged these files using the unique identifier: `eid`.

```
df_bcs_features = pd.read_csv("../../features_for_bcs.csv", delimiter = ",")
dementia = pd.read_csv("../../Labels_ZY/UKBB_Dementia_202109.tsv", delimiter = "\t")
stroke = pd.read_csv("../../Labels_ZY/UKBB_Stroke_202109_v2.tsv", delimiter = "\t")
bcs_features = pd.read_excel("../../fields_to_be_kept.xlsx")['eid'].tolist()
bcs_features.append('eid')
df_bcs_features = df_bcs_features[bcs_features]
df_dementia = pd.merge(df_bcs_features, dementia, left_on=['eid'], right_on=['sample_id'],
                how='inner')
df_all = pd.merge(df_dementia, stroke, left_on=['eid'], right_on=['sample_id'],
                how='inner', suffixes=('_dementia','_stroke'))
```

### A.2. Splitting the Data

We split the data as detailed in Section 3.1, and we provide the calculation for time to event (time from when patient was enrolled in the study to diagnosis) and the value for whether the event occurred (whether the patient was diagnosed with dementia or stroke).

```
controls_bool = (df_all["incident_disease_stroke"] == 0) &
                (df_all["prevalent_disease_dementia"] == 0) &
                (df_all["incident_disease_dementia"] == 0) &
                (df_all["prevalent_disease_stroke"] == 0)&
                (df_all["has_died_stroke"] == 0) &
                (df_all["has_died_dementia"] == 0)
dementia_bool = (df_all["incident_disease_dementia"]==True)
stroke_bool = (df_all["incident_disease_stroke"] == True)

df = df_all[(controls_bool) | (dementia_bool) | (stroke_bool)]

df.loc[(controls_bool), "time_to_event_stroke"]
= pd.to_numeric(pd.to_datetime(df[(controls_bool)]["censor_date_stroke"]) -
pd.to_datetime(df[(controls_bool)]["enroll_date_stroke"]))/1000000000/60/60/24
df.loc[(controls_bool), "time_to_event_dementia"]
= pd.to_numeric(pd.to_datetime(df[(controls_bool)]["censor_date_dementia"]) -
pd.to_datetime(df[(controls_bool)]["enroll_date_dementia"]))/1000000000/60/60/24
df.loc[(controls_bool), "dementia_status"] = 0
df.loc[(controls_bool), "stroke_status"] = 0

df.loc[(dementia_bool), "time_to_event_stroke"]
= pd.to_numeric(pd.to_datetime(df[(dementia_bool)]["censor_date_stroke"]) -
pd.to_datetime(df[(dementia_bool)]["enroll_date_stroke"]))/1000000000/60/60/24
df.loc[(dementia_bool), "time_to_event_dementia"]
= pd.to_numeric(pd.to_datetime(df[(dementia_bool)]["censor_date_dementia"]) -
pd.to_datetime(df[(dementia_bool)]["enroll_date_dementia"]))/1000000000/60/60/24
df.loc[(dementia_bool), "dementia_status"] = 1
df.loc[(dementia_bool), "stroke_status"] = 0

df.loc[(stroke_bool),"time_to_event_stroke"]
= pd.to_numeric(pd.to_datetime(df[(stroke_bool)]["censor_date_stroke"]) -
pd.to_datetime(df[(stroke_bool)]["enroll_date_stroke"]))/1000000000/60/60/24
df.loc[(stroke_bool),"time_to_event_dementia"]
= pd.to_numeric(pd.to_datetime(df[(stroke_bool)]["censor_date_dementia"]) -
pd.to_datetime(df[(stroke_bool)]["enroll_date_dementia"]))/1000000000/60/60/24
df.loc[(stroke_bool),"dementia_status"] = 0
df.loc[(stroke_bool),"stroke_status"] = 1

df.loc[(stroke_bool & dementia_bool),"dementia_status"] = 1
df.loc[(stroke_bool & dementia_bool),"stroke_status"] = 1
```

### A.3. Code for Kaplan-Meier Analyses

For the KM analysis, we randomly sample from the controls group per the rationale described in Section 3.1. Then, we plot the KM curves using the `lifelines` package and also evaluate the $p$-values. Below, we provide the code and a sample call to the function.

```
from statistics import median, mean, mode
from lifelines.statistics import survival_difference_at_fixed_point_in_time_test

# kaplan meier analysis
dementia_df_km  = pd.concat([df[dementia_bool][all_features+["time_to_event_dementia",
                                                             "dementia_status"]],
                             df[controls_bool][all_features+["time_to_event_dementia",
                             "dementia_status"]].sample(n= len(df[dementia_bool]))])

stroke_df_km = pd.concat([df[stroke_bool][all_features+["time_to_event_stroke",
                                                        "stroke_status"]],
                             df[controls_bool][all_features+["time_to_event_stroke",
                             "stroke_status"]].sample(n= len(df[stroke_bool]))])

def dementia_kms(features, categorical, dementia_df_km):
    print("Dementia Statistics:")
    D = {}
    kmf_dementia = KaplanMeierFitter(label="Dementia and Controls (n = " +
                                     str(len(dementia_df_km)) + ')')
    kmf_dementia.fit(dementia_df_km['time_to_event_dementia'],
                     dementia_df_km["dementia_status"])
    kmf_dementia.plot()
    for feat, cate in zip(features, categorical):
        D[feat] = []
        if cate:
            for c in cate:
                high = dementia_df_km[dementia_df_km[feat] == c]
                kmf_dementia_high = KaplanMeierFitter(label= "Dementia and Controls
                    (n = " + str(len(high)) + ') ' + feat + ', ' + str(c))
                kmf_dementia_high.fit(high['time_to_event_dementia'],
                        high["dementia_status"])
                kmf_dementia_high.plot()
                D[feat].append(kmf_dementia_high)
        else:
            q = list(dementia_df_km[feat].quantile([0.2,0.8]))
            med_high_feat = q[-1]
            med_low_feat = q[0]

            high = dementia_df_km[dementia_df_km[feat] >= med_high_feat]
            kmf_dementia_high = KaplanMeierFitter(label="Dementia and Controls
```

```
            (n = " + str(len(high)) + ') ' + feat + ', High ')
            kmf_dementia_high.fit(high['time_to_event_dementia'],high["dementia_status"])
            kmf_dementia_high.plot()
            D[feat].append(kmf_dementia_high)


            low = dementia_df_km[dementia_df_km[feat] <= med_low_feat]
            kmf_dementia_low = KaplanMeierFitter(label="Dementia and Controls
            (n = " + str(len(low)) + ') ' + feat + ', Low ')
            kmf_dementia_low.fit(low['time_to_event_dementia'],low["dementia_status"])
            kmf_dementia_low.plot()
            D[feat].append(kmf_dementia_low)

    for feat in D:
        results = survival_difference_at_fixed_point_in_time_test(point_in_time=100,
                        fitterA=D[feat][0], fitterB=D[feat][-1])
        print(feat + ' Chi-squared(1) Test statistic='+str(results.test_statistic) + '
                        p-value='+str(results.p_value))

features, categories = ["Smoking status", 'HDL cholesterol'], [[0,1,2], False]
dementia_kms(features, categories, dementia_df_km)
```

### A.4. CoxPH Models

To prepare the data for CoxPH, we split the data in the same way as PSET 5 used `simpleimputer`, with `strategy = "most_frequent"` for the categorical variables and `strategy = "mean"` for the continuous variables. Then, we specify the features for the training, validation, and test sets as follows:

```
dementia_train = dementia_train[FEATURES+['time_to_event_dementia', 'dementia_status']]
dementia_valid = dementia_valid[FEATURES+['time_to_event_dementia', 'dementia_status']]
dementia_test = dementia_test[FEATURES+['time_to_event_dementia', 'dementia_status']]

stroke_train = stroke_train[FEATURES+['time_to_event_stroke', 'stroke_status']]
stroke_valid = stroke_valid[FEATURES+['time_to_event_stroke', 'stroke_status']]
stroke_test = stroke_test[FEATURES+['time_to_event_stroke', 'stroke_status']]
```

## Appendix B. vadesc

In this section, we document our attempt to use `vadesc` (Manduchi et al., 2021) to cluster the data. We noticed that the UK Biobank is most similar to the Support dataset provided in the Github Repository. We wrote code to generate the initial clusters based on the code to generate the initial clusters for the Support dataset (also patient survival data), clustering intially based on patients who got dementia versus patients who did not get dementia. At times, we were able to get the code working, but it was extremely unstable with convergence issues. Additionally, with the only notebook example for cluster generation in the Github

Repository being image data, we were unable to interpret how that translates for our data. Since this work is extremely new, perhaps, with increased documentation and better data cleaning on our end, it would be possible to incorporate results from the repository in the future.

The following are the packages that we used to get `vadesc` to work, and the code we wrote the generate the intial clusters:

```
# install necessary packages :
!pip install lifelines
!pip install torchtuples
!pip install pycox
!pip install torch
!pip install utils
!pip install tensorflow_probability
!pip install tensorflow
!pip install progressbar
!pip install opencv-python
!pip install pydicom
!pip install pydicom_seg
!pip install pyradiomics
!pip install openTSNE
!pip install scikit-learn==0.22.2
!pip install tqdm
```

```
def generate_data(data_frame, seed=42):
    np.random.seed(seed)
    to_drop = ['time_to_event_dementia', 'dementia_status']
    one_hot_encoder_list = cat_feats
    data_frame = one_hot_encoder(data=data_frame, encode=one_hot_encoder_list)

    t_data = data_frame[['time_to_event_dementia']]
    e_data = data_frame[['dementia_status']]

    # dzgroup roughly corresponds to the diagnosis; more fine-grained than dzclass
    c_data = data_frame[['dementia_status']]
    c_data['dementia_status'] = c_data['dementia_status'].astype('category')
    c_data['dementia_status'] = c_data['dementia_status'].cat.codes

    x_data = data_frame.drop(labels=to_drop, axis=1)
    print(x_data.columns)
    encoded_indices = one_hot_indices(x_data, one_hot_encoder_list)
    include_idx = set(np.array(sum(encoded_indices, [])))
    mask = np.array([(i in include_idx) for i in np.arange(x_data.shape[1])])
    covariates = np.array(x_data.columns.values)
    x = np.array(x_data).reshape(x_data.shape)
    t = np.array(t_data).reshape(len(t_data))
```

```
e = np.array(e_data).reshape(len(e_data))
c = np.array(c_data).reshape(len(c_data))

idx = np.arange(0, x.shape[0])

np.random.shuffle(idx)
x = x[idx]
t = t[idx]
e = e[idx]
c = c[idx]

# Normalization
t = t / np.max(t) + 0.001
scaler = StandardScaler()
scaler.fit(x[:, ~mask])
x[:, ~mask] = scaler.transform(x[:, ~mask])

end_time = max(t)

num_examples = int(0.80 * len(e))

train_idx = idx[0: num_examples]
split = int((len(t) - num_examples) / 2)

test_idx = idx[num_examples: num_examples + split]
valid_idx = idx[num_examples + split: len(t)]


imputation_values = get_train_median_mode(x=x[train_idx], categorial=encoded_indices)

preprocessed = {
    'train': formatted_data(x=x, t=t, e=e,
    idx=train_idx, imputation_values=imputation_values),
    'test': formatted_data(x=x, t=t, e=e,
    idx=test_idx, imputation_values=imputation_values),
    'valid': formatted_data(x=x, t=t, e=e,
    idx=valid_idx, imputation_values=imputation_values)
}

preprocessed['train']['c'] = c[train_idx]
preprocessed['valid']['c'] = c[valid_idx]
preprocessed['test']['c'] = c[test_idx]

return preprocessed
```

```
def generate_ukb(df, seed=42):
    preproc = generate_data(df, seed)
    x_train = preproc['train']['x']
    x_valid = preproc['valid']['x']
    x_test = preproc['test']['x']

    t_train = preproc['train']['t']
    t_valid = preproc['valid']['t']
    t_test = preproc['test']['t']

    d_train = preproc['train']['e']
    d_valid = preproc['valid']['e']
    d_test = preproc['test']['e']

    c_train = preproc['train']['c']
    c_valid = preproc['valid']['c']
    c_test = preproc['test']['c']

    return x_train, x_valid, x_test, t_train, t_valid, t_test,
    d_train, d_valid, d_test, c_train, c_valid, c_test
```

## Appendix C. auton

We attempted to use `auton` (Nagpal et al., 2022) to cluster the data. With this GitHub repository, we looked at the phenotyping censored time to events notebook included in the repository. Similarly there were convergence issues that we could not resolve. In running the deepsurv model on our data, we did not get improved concordance indices, and as far as we know, there is not a straightforward way to extract feature significance.