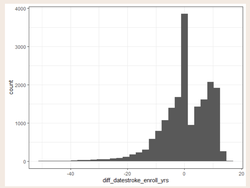
Measuring Brain Health with Survival Modeling

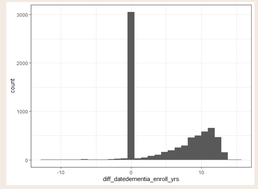
Hannah Ramcharan, Rui-Jie Yew, Livia Parodi, Madhur Nayan, Puneet Batra

Background

We applied survival modeling methods on UK Biobank data to evaluate the features in the Framingham Heart Study and the McCance Brain Care Score (M-BCS), recently developed (but not yet used) by the McCance Center for Brain Health (Massachusetts General Hospital, Boston) to predict risk of stroke and cognitive decline. This project extends the features recognized as important as part of the Framingham heart study and validates the features that are included as part of the McCance brain care score.

Time between Enrollment and censor date





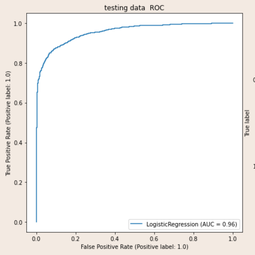
Status	n
Controls	483,780
Stroke	18,705
NA	19

Status	n
Controls	495,354
Dementia	7,131
NA	19

Figure 1. Total number of healthy controls and stroke (top left) and dementia(top right) patients included in UKBB. Counts of cases and controls for stroke (bottom left) and dementia (bottom right)

Predictive Analysis with Logistic Regression

Testing Data ROC



Testing Data Confusion Matrix

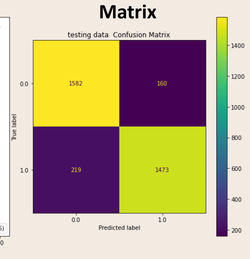


Figure 2. Training performance of Logistic model to predict stroke outcome.

A logistic regression analysis was performed on the data to predict disease outcome of samples given their features. The model was trained using L2 regularization. The training AUC is 0.889. The feature importance was calculated by extracting the coefficients. The results shown in figure 3 are from the stroke subset. Feature 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.

Feature **Coefficients**

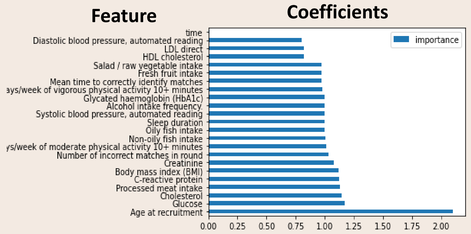


Figure 3. Feature importance by coefficient

Defining Risk with CoxPH modeling

CoxPH Model Performance on Difference Datasets

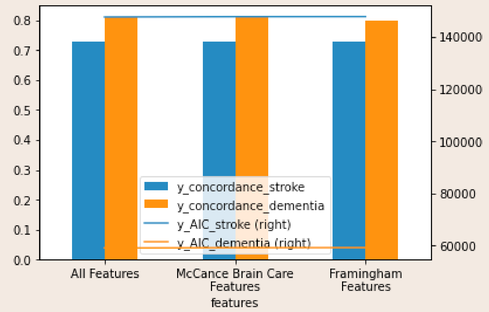
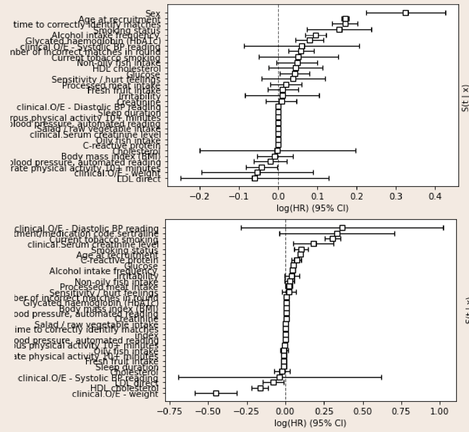


Figure 4. CoxPH Model Performance on features from the UKBB McCance Brain care and Framingham studies was measured by looking at the concordance and the Akaike information criterion (AIC). We compared these scores for the features in the UKBB, MCB and Framingham datasets.

Comparison of CoxPH Coefficients



CoxPH Survival Curves

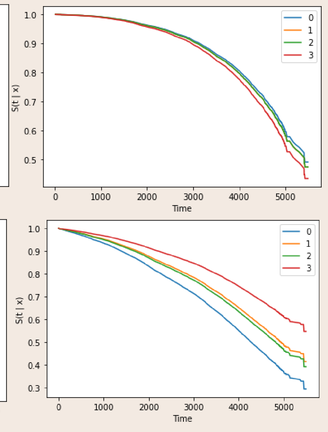


Figure 5. Significant coefficients for model ran on Dementia (top left) and Stroke samples (Bottom left). CoxPH survival curves for stroke (bottom right) and dementia samples (top right)

Survival with Stratification

Stratified Kaplan Meier Plot

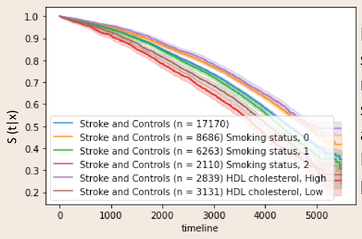


Figure 6. Sample stratified by median of smoking status, and cholesterol Kaplan Meier plots by

Log rank test P - values

Feature	P- Value
Smoking status	0.004
HDL cholesterol	0.0172
Age	0.0004

Table 1. Log rank test results

Equation 1. Log Rank test: O1 and O2 are the total numbers of observed events in groups 1 and 2, respectively, and E1 and E2 the total numbers of expected events.

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

We stratify the population by smoking status and cholesterol and generate Kaplan Meier survival models. A log rank test as shown in Equation 3 is then performed to understand if there is a difference in survival between groups. The p value was calculated with a Chi squared statistical.

Evaluating Constand Hazard Assumption

Cumulative Hazard functions

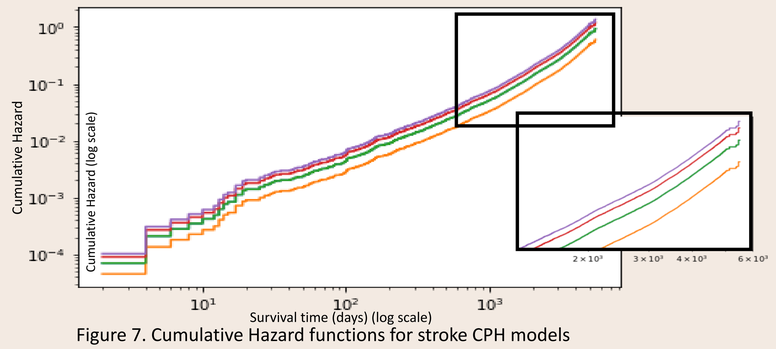


Figure 7. Cumulative Hazard functions for stroke CPH models

Scaled Schoenfeld Over Time for Fresh Fruit Intake and Age

Figure 8. Scaled Schoenfeld Over Time for Fruit Intake and Age. P values shown.

The Cox Proportional Hazards model relies on the assumption that the proportional hazards Eq. 1 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 hold 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 as it appears to be. We then used the Schoenfeld residuals test to further check the assumption for each variable to see if residuals change over time.

Equation 1. Hazard Function

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

Equation 2. Cumulative Hazard Function

$$H(t) = -\log S(t)$$

P Vales for Scaled Schoenfeld Residuals test

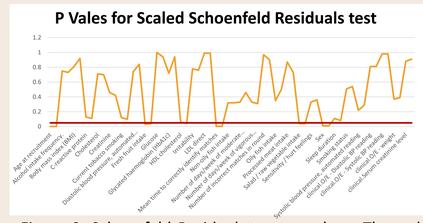
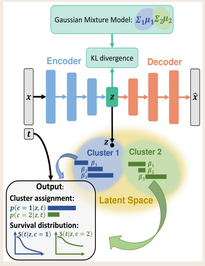


Figure 9. Schoenfeld Residuals test p- values. The red line shows a cut-off value of .05. Most values fall above this threshold.

Conclusions & Future Directions



- Survival models provide a useful way to predict disease risk.
- Clustering survival data using Variational Autoencoders.
- Generate clusters of samples based on features and survival score
- Deep survival models that provide transparency regarding feature importance.
- Discrete time models

Scan for References and Code: