A machine learning approach to identify and compare heterogeneous subgroups of patients treated for alcohol use disorder, over time

Apeksha Kumar
MS in Computer Science
University of Southern California
avkumar@usc.edu

Daniel Ley
MS in Operations Research Engineering
University of Southern California
dley@usc.edu

Katie Foss
MS in Computer Science
University of Southern California
katiefos@usc.edu

Rafael V. Sanchez-Romero
MS in Computer Science
University of Southern California
rs06167@usc.edu

Abstract

Much of the research in the addiction sciences field is centered around opioid use or substance use disorder (SUD) in general. Alcohol use disorder (AUD) which is the most prevalent substance use disorder worldwide is less commonly studied and even less so in the context of the intersection between social work and computer science. In this paper, we use a longitudinal cohort of (n = 3290) individuals across a 12 month treatment study specific to AUD to identify heterogeneous subgroups within the homogeneous population and determine how these subgroups differ with respect to relapse events. To do this we use clustering analysis to determine the unique characteristics of patients undergoing AUD treatment and survival modeling to determine patients' time to relapse and predictors which contribute significantly to it. Based on these results, we can better understand the factors that affect patient relapse within each individual subgroup to inform better treatment programs tailored to individuals in a given subgroup.

1 Introduction

1.1 Motivation

The consumption of alcohol is common in many countries and cultures across the globe. Although alcohol is glamorized in the media, it is one of the most dangerous and compulsive substances accounting for approximately 3 million deaths each year. Moreover, it is a "psychoactive substance with dependence-producing properties in addition to being a causal factor in more than 200 diseases, injuries and other health conditions" [1].

Alcohol use disorder (AUD) is defined as the "impaired ability to stop or control alcohol use despite adverse social, occupational, or health consequences" [2]. Understanding AUD along with its health and socioeconomic effects is strongly related to the United Nations (UN) Sustainable Development Goals (SDG) goal number three *Good Health*, to "ensure healthy lives and promote well-being for all at all ages" [3].

1.2 Problem statement

Treatment for AUD is a unique process for each individual, therefore it is highly important to understand an individual's trajectory of consumption if and when a relapse occurs. One way to measure a patient's treatment outcome for AUD is abstinence, another way is measuring the level of function (high, low) a patient has, post-treatment. In a study by Witkiewitz et al. [4], it was determined that heterogeneous patient subgroups can be used to predict what level of functionality a person will have post-treatment. In this project, we will expand on this idea by developing a framework to answer the overarching question of how to classify heterogeneous subgroups of patients being treated for AUD and how each of those subgroups differs from the overall population over time.

1.3 Project outline

To successfully achieve the objectives of this project, we will follow several steps detailed in the subsequent sections of this report.

First, we will perform data evaluation and cleaning. Evaluation of all features associated with a patient entering a treatment facility will be performed to determine candidate features to train a survival model. This will be done in exploratory data analysis (EDA) with data visualization, and univariate and bivariate analyses of each feature, finally we will impute missing data where necessary and study the feature correlations.

Second, we will cluster patients using their background features recorded over the intake survey. Moreover, we will use survival modeling to predict the time to first drink (relapse). We will also determine the feature importance of the input features to the survival model.

Third, expert defined groups will be determined based on sets of one or two features. Then a different survival model will be trained for each group. These survival models will be combined to create an ensemble aiming to get better results rather than using only one general model over the whole dataset.

2 Related Work

Survival analysis is a common method used in time-to-event predictions. While survival analysis typically uses statistical models like the Cox proportional hazard model, several machine learning and AI approaches have been developed to improve the accuracy of predictions.

In [5] authors used survival models such as the common Cox proportional hazard (Cox-PH) model to predict the length of stay for COVID-19 patients and compared their results to Deep Learning (DL) models such as DeepSurv and DeepHit . For the dataset used by the authors (~ 800 data points), the Deep Learning methods did not perform as well as the Cox-PH models because these models did not get sufficient training data to model the non-linear relationships. Random Survival Forests (RSF), another method used by the authors, provided a good C-index score, however, it failed to perform as well when it came to other metrics like Brier score, also owing to the small dataset size. We plan to use RSF and other Deep Learning models from this paper for our use case, due to the larger AUD dataset.

In [6] the main novel contribution is the use of deep learning models to predict time to mortality for patients with prostate cancer on both longitudinal and cross sectional data using the Recurrent Deep Survival Machine (RDSM) model and the Deep Survival Machine (DSM) model respectively. Given that the authors had over 100k samples it was possible for them to try deep learning survival models. Authors defined two different tasks: prediction of patient death and prediction of patient developing the disease given the past history. For both tasks, deep learning models outperformed other baseline machine learning models such as Random Survival Trees, Cox Regression or Gradient Boosting Machines.

Time-to-event modeling is also commonly used in the field of addiction sciences to predict the time to relapse or to simply understand the rank-order predictors (key factors) that lead to an event such as relapse. This is the main contribution in Davis et.al. [7], which looks at determining the individual and environmental predictors of opioid use among adolescents using Lasso regression and random survival forests. Based on their analysis of the rank order predictors, a mix of individual and environmental variables proved to be strong predictors of substance use for all severities (low, medium, high usage). Each severity's predictor was influenced by distinct features which provides further indication that SUD treatment should be tailored to the severity as well as the substance. We plan to utilize a similar technique to combine important features and hazard ratios to better quantify the effects of predictors on AUD.

Lastly, Witkiewitz et al. [4], used mixture models to classify AUD patients into 4 subgroups and predict the probability that a patient will endorse having anxiety or depression symptoms. This work was the motivation to consider the effects of heterogeneous groups on predictions of AUD treatment outcomes.

Building on some of the characteristics of these previous publications, we aim to develop a methodology to detect via clustering possible subgroups in the heterogeneous study population that can add insight on how different people are affected by AUD. A series of survival models will be used to model and evaluate the time to relapse task. Finally, through feature importance tasks, we will aim to obtain an explanation on how several features can affect (and to what degree) the treatment of AUD in patients.

3 Data

3.1 Data Overview

The full data set used in this analysis is a fusion of multiple data sources which includes factors on demographics, mental health, treatment, substance use, and social and environmental indicators for anonymized patients being treated for AUD. Additionally, the full data consists of longitudinal samples meaning that for each patient, there are multiple observations (recording the same features) collected across the year. This longitudinal data structure, which differs from cross-sectional data with random samples, enables the use of survival analysis allowing for a more complete picture of how various factors contribute to relapse with a certain probability over time.

3.2 Data sources

Two data sources are used in the creation of the final dataset. The first data source, the Global Appraisal of Individual Needs (GAIN) encodes data gathered from patients aged 12 and up, before, during and after undergoing treatment for a substance use disorder (SUD) [8] from 137 treatment centers [7]. For our purposes, we will only examine patients being treated with AUD. Each participant completed a baseline assessment at treatment entry, received treatment, and then completed follow-up assessments at 3, 6, and 12 months [7]. It is important to note that treatment lasts for a total of 90 days (3 months). Assessments included eight core topics, where each topic contains questions on the recency of problems, breadth of symptoms, as well as frequency of substance use [9].

The second data source adds socioeconomic context to each patient based on the Census tract location of the treatment center and the year they visited. Census tract data (including population and unemployment statistics) comes from the American Community Survey.

3.3 Data description

The final dataset used in the analysis has a total of 3290 samples from individuals across a one year time frame with regular surveys at the start of treatment (month 0), at the end of treatment (month 3),

Table 1: Details of censor variable levels. Note in the final analysis only censored/uncensored binary levels are used.

Censored/Uncensored	Censor Value	Days to Relapse	Description
Uncensored	0	*d_0 = 0,...,365*	Relapsed on day d_0
Censored	1	*d_1 = 300,...,365*	Relapse not observed. Patient made it to the end of the study period with no relapse, outcome after study not known.
	2	*d_2 = 0,...,365*	Relapse not observed. Lost to followup on day d_2

and two periods after treatment has concluded (months 6 and 12). At each followup session, patients were asked the number of days they had consumed alcohol within a given period. Patients were also surveyed on the number of days they remained sober before consuming alcohol. In other words, the number of days before experiencing a relapse event. This information is used as the primary target variable for our analysis on which the survival models are built.

Common to clinical datasets, patients do not always follow up at the specified time frames. A patient's inability to follow up could be due to many reasons such as death, refusal to participate in treatment or the lack of resources to receive treatment for their disorder. To account for these cases, an added variable –called a censoring variable– allows us to understand whether or not a patient has fallen out of the study. This variable takes on three different values 0, 1, 2 which are described in Table 1. In the final dataset, the censor variable has been modified to indicate the event that an individual has not relapsed (1) and the event that an individual has failed to followup in the study (2) with a value of True (censored) and those who have accurate and up to date information concerning relapse (0) are denoted with a value of False (relapse event occurred). In many cases, standard machine learning regression tasks are used to predict the time to an event, however, these standard regression tasks do not take into account censored data since there is no disposition on the outcome of such a data point. In the case of survival analysis, censored data proves to be a useful piece of information as it can be included in prediction of the probability distribution an event occurs within a certain time frame.

The variables *days until first drink (relapse)* and the *censor indicator* are used as target variables in the analysis. However, as mentioned earlier the dataset consists of features concerning an AUD patient's substance abuse, treatment, environmental factors and more. Some examples include age group (AgeGroup), race group (White, Hispanic, Black, etc.), region (Midwest, West, etc.), and many additional features corresponding to pertinent medical questions associated with determining a patient's substance use and desire for treatment. A breakdown of the different variable types as well as some examples of each variable type are provided in Table 2. A more detailed description of each variable's significance as well as the variable name in the dataset, its corresponding English name, and the variable type can be found in Appendix A.

For each variable in the final dataset, a correlation value was calculated and plotted in Figure 1. It is important to note that the variables describing the demographics of region and race are separated into one-hot encoded columns in the final dataset. For readability and interpret-ability in the correlation plot, the 5 region variables and the 4 race variables were grouped into a single categorical column for each variable, region and race. As most of the variables are categorical or binary (True/False) standard correlation metrics such as Pearson's or Spearman's correlation values cannot be computed. To understand how these categorical variables are correlated, a variation of the chi-squared statistical test called Cramer's V is used. For all continuous variables, standard Pearson correlation is used.

Based on the coloring shown in the correlation heat map in Figure 1, we can see that there is little correlation among variables with most values around zero. Small patches of positively correlated values cover the heat map for variables within the same type (i.e. demographics, mental health, etc.). The positive correlation which stands out most is the Alcohol_Days (DaysToRelapseAUD) variable and its corresponding censor variable (CensorVariable). Additionally, features such as MedianHomeValue and MedianGrossRent appear to have a high correlation due to the obvious correlation in housing. Other variables such as MedianFamilyIncome and percentage below the poverty line

Table 2: Breakdown of variable types in the final dataset.

Variable Type	Count of Variables	Example
Demographic	21	Race Group, Region
Environment	1	Alcohol Access Points
Mental Health	8	Depression, Suicidal Thoughts
Social Factors	5	Not Close to Anyone in Recovery
Substance Use	6	Withdrawal Symptoms
Treatment	4	Prior Substance Use, Treatment Motivation
Target	3	Days Until First Drink, Alcohol Censor
Total Count	48	

(PctPoverty) show a strong positive correlation as they pertain to a patient's livelihood. Finally the AgeGroup variable is highly correlated with unemployment (CurrentlyUnemployed) due to the fact that many patients are not of employment age. The other target variables being the number of drinks (DaysConsumedThreeMonths, etc.) within each follow up period show a negative correlation with the DaysToRelapseAUD variable. As these are the only strongly negatively correlated variables we can infer that patients who have relapsed early on in the trial have larger values for the days of consumption between periods and conversely patients who do relapse early on have no consumption days in the periods prior to their relapse and perhaps low consumption in the periods following their relapse.

3.4 Pre-processing

The primary dataset was provided by Davis et al., who previously did the work to process and join the two data sources in [7]. Being that the dataset provided contains treatment data for various types of SUD treatments, the dataset was filtered to contain only patients being treated for AUD. Furthermore, the dataset was filtered to only contain relevant data where a patient's time to relapse could be determined.

Aside from these filters, standard data cleaning practices were used such as the removal of duplicate observations across the same patient ID, as well as the conversion of numerical data into terciles (low, medium, high values) where applicable. The conversion to terciles is a common feature engineering approach in survival modeling and most of the features are provided in this structure as a result of surveyed information. The data contained few Null values with Nulls only present in 4 columns. The columns containing Null values were the followup survey questions for number of days where alcohol was consumed in each period. This result is expected since many people in the data are lost due to various reasons. Therefore Null values were simply kept in the data. For each follow up period 0, 3, 6, 12 months, the amount of Null values was 6 (0.18%), 474 (14.4%), 640 (19.45%), and 1715 (52.13%) respectively (Appendix B).

3.5 Visualization

To better understand the characteristics of relapse among patients being treated for AUD, the density of the DaysToRelapseAUD variable was plotted for both censored and uncensored patients in Figure 2.

In the plot we can see that, of those patients who are uncensored, the majority experienced a relapse event within the first 50 days. As the number of days increases the density of patients experiencing a relapse event declines. Slightly before 200 days, there is a small uptick in patients experiencing a relapse event.

For those patients whose data is censored (i.e. they did not followup for any given reason), we can gain some interesting insights from the plot as well. The density plot for censored patients appears to be tri-modal around the followup periods of 3 months, 6 months, and 12 months. This

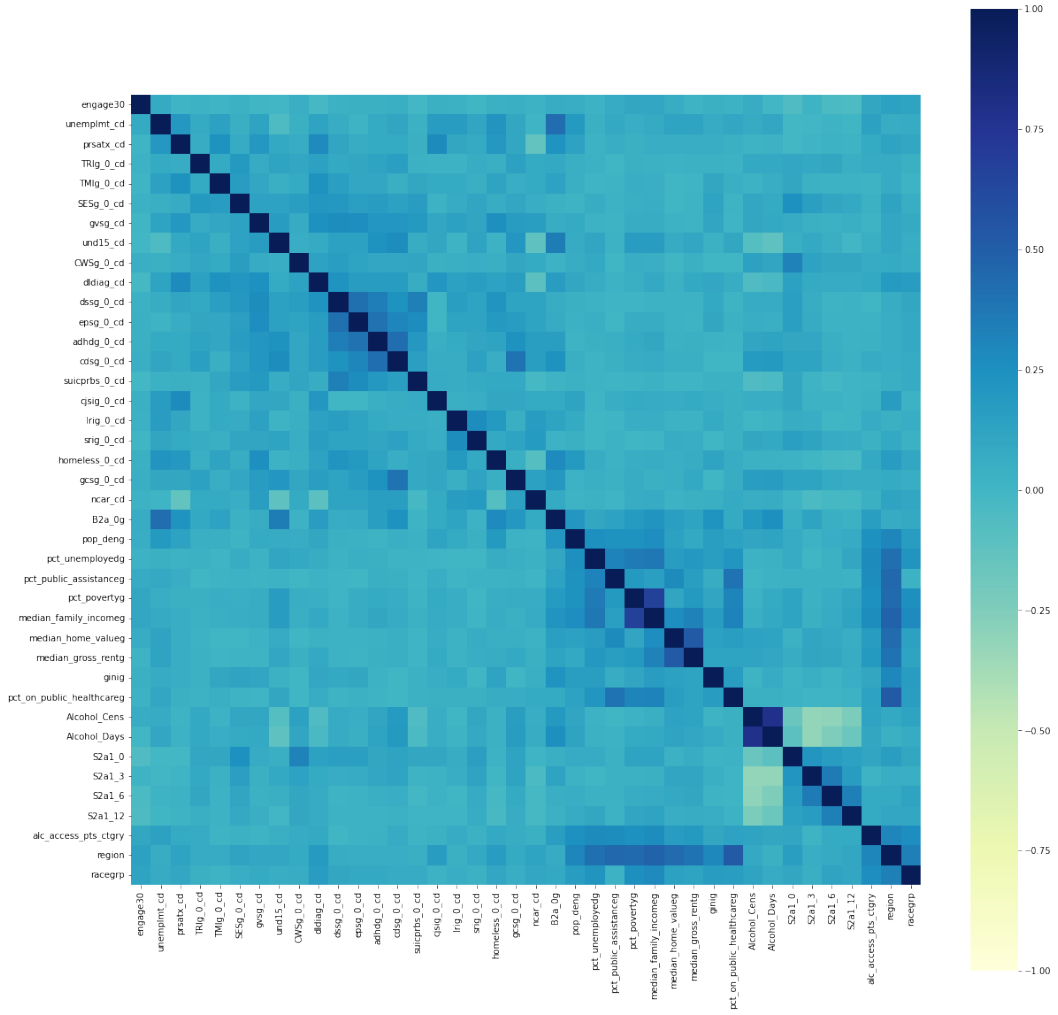


Figure 1: Correlation matrix for all features.

gives indication as to which followup period the majority of patients stopped participating in the treatment program. As time moves on from treatment, we can see that the peaks for the censored density distribution get higher meaning more patients do not follow up. This sort of insight might be useful in further developing the treatment program to keep patients engaged in treatment in the periods between followup.

In Figures 3a and 3b we can see the ground-truth time-to-event plotted as a cumulative distribution function for percentage of patients not relapsed and percentage of patients who have relapsed. Note that both plots are inverses of one another meaning the same information is shown in both plots from a different perspective. In Figure 3a we can see a steep decline as time progresses and by day 100 (only 10 day after the treatment program has completed) only around 10 percent of patients undergoing AUD treatment have not relapsed.

The descriptive statistics for censored and uncensored patients can be found in Table 3. For the information in Table 3 we can see that those who have relapsed (uncensored) have a low average days to relapse of only 42 days (SD = 46.3 days). This is juxtaposed with an average of 244.271 days (SD = 105.964 days) to relapse for censored patients, (Note that these patients did not actually relapse but either did not continue treatment or successfully completed the treatment program treatment).

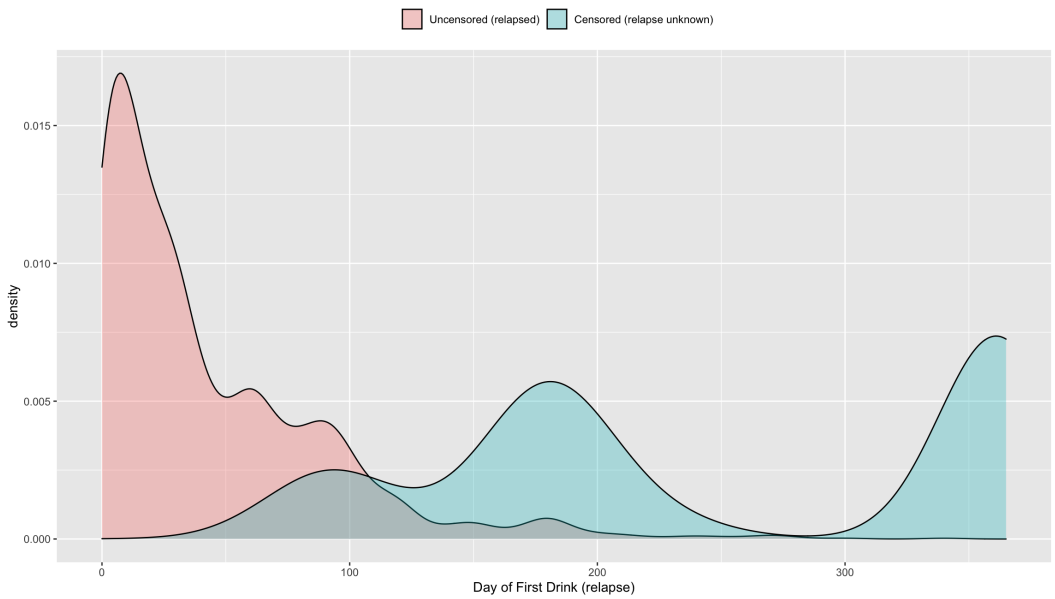
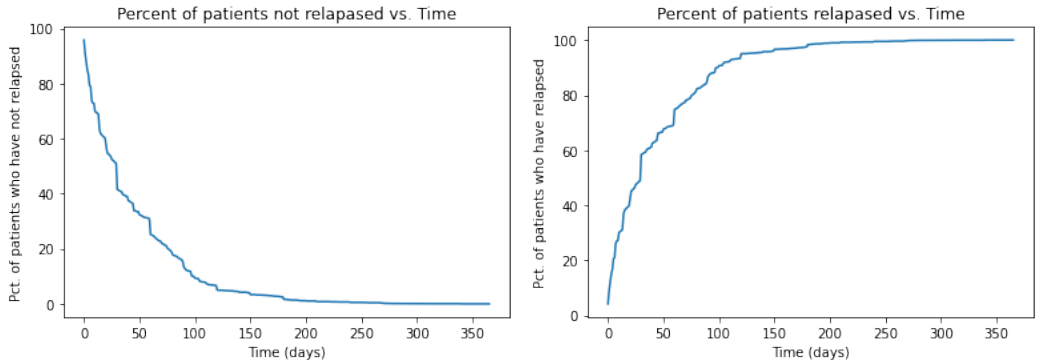


Figure 2: Density of days to first drink (relapse) grouped by censored (relapse unknown) and uncensored (relapsed) patients.



(a) Percent not relapsed vs. Time.

(b) Percent relapsed vs. Time.

Figure 3: Ground truth survival for uncensored (relapsed) patients.

4 First Empirical Results

The results of the project can be grouped into two distinct tasks based on the project outline. On one hand, we will perform clustering tasks over the data which we will visualize and explain. On the other, we will use the data to train several survival models.

4.1 Clustering

To get a complete picture of sub-groups within the dataset, clusters based on the patients features are obtained. These exclude the censoring variable, the days elapsed until the first alcohol intake after entering the treatment as well as the number of days a patient drank alcohol in each period between surveys. Geographical indicators are also excluded.

Table 3: Descriptive statistics on target variables. **Days to relapse for censored patients does not indicate a relapse event** rather it indicates the day they completed the program (1) or the day they did not followup (2).

Variable Type	Avg. Days to Relapse	Std. Dev.	Median Days to Relapse	Count
Uncensored	**42.618**	**46.300**	**30**	**1759**
Censored	**244.271**	**105.964**	**197**	**1531**
(1) censored by end of study period	355.419	7.841	357	313
(2) lost to followup	215.709	100.532	184	1218
Censored, Uncensored	**136.457**	**128.411**	**91**	**3290**

Initially, we planned to obtain results for time-series clustering using only the patient features representing the number of days they consumed alcohol in between surveys. However, the time series were too noisy and clustering results were not significant, so results will not be shown.

As the feature space we are working in exceeds the dimensions humans are able to interpret, we will be applying dimensionality reduction by using Principal Component Analysis (PCA). An detailed explanation on the meaning of the PCA features will be discussed later in this section.

4.1.1 Dimensionality reduction with PCA

PCA is a statistical technique widely used to obtain better analyses of datasets that have a large number of features per sample. This technique reduces the number of dimensions used while maximizing the information contained. This is achieved by linearly transforming the data into a new reduced space where all the features are orthogonal to each other and obtained by making linear combinations of the original set of features that best explain the variability of data.

For the visualization of the data in the subsequent sections, PCA has been applied to obtain a reduction to a 2-dimensional space. To better explain the clusters obtained, we first need to interpret the meaning of the values for each dimension in the new space.

For the clustering problems using all of the patient features, we focus our attention on the top 6 original features in each dimension. To determine these top 6 features we order them in decreasing order of absolute value of their weights in the linear combination that represents each new dimension. We will refer to the dimension plotted in the x axis of subsequent plots as PCA1 and to the y axis as PCA2. The most important features for each dimension are:

- **PCA1**: ADHD (adhdg_0_cd), VictimizationEvents (gvsg_cd), Depression (dssg_0_cd), Delinquency (cdsg_0_cd), GlobalEmotionalProblems (epsg_0_cd) and CriminalActivity (gcsg_0_cd). All of the variables are categorical, meaning that they take values 0, 1 or 2 depending on how high the original value was. 5 out of the 6 variables are mental health (MH) indicators. The other, gcsg_0_cd, is a social factor (SF). The weights of these features in the linear combination can be found in Table 4. From them we cannot see any negative value. All the negative values in the linear combination have absolute values lower than 0.2, thus we can conclude that $PCA1 > 0$ is correlated with patients having high severity in mental health indicators while $PCA1 < 0$ is correlated with patients having low severity in the same indicators.

- **PCA2**: AgeGroup (B2a_0g), CriminalJusticeEngagement (cjsig_0_cd), PopulationDensity (pop_deng), VictimizationEvents (gvsg_cd), CriminalActivity (gcsg_0_cd) and CurrentlyUnemployed (unemplmt_cd). All of the variables are categorical, meaning that they take values 0, 1 or 2 depending on how high the original value was. 3 out of the 6 variables are demographic indicators (DI). 2 out of 6 are social factors (SF) and the other, gvsg_cd, is a mental health (MH) indicator. The weights of these features in the linear combination can be found in Table 5. The rest of the values are, in absolute value, lower than 0.2. This is a mix of several types of features, thus the explanation is more difficult to explain. However,

Table 4: PCA1 linear combination weights for the top 6 features in it.

Feature name	English variable name	Indicator type	Weight
adhdg_0_cd	ADHD	MH	0.4479
gvsg_cd	VictimizationEvents	MH	0.4238
dssg_0_cd	Depression	MH	0.3823
cdsg_0_cd	Delinquency	MH	0.3324
epsg_0_cd	GlobalEmotionalProblems	MH	0.3249
gcsg_0_cd	CriminalActivity	SF	0.3098

Table 5: PCA2 linear combination weights for the top 6 features in it.

Feature name	English variable name	Indicator type	Weight
B2a_0g	AgeGroup	DI	0.6090
cjsig_0_cd	CriminalJusticeEngagement	SF	0.3700
pop_deng	PopulationDensity	DI	-0.2942
gvsg_cd	VictimizationEvents	MH	0.2801
gcsg_0_cd	CriminalActivity	SF	-0.2099
unemplmt_cd	CurrentlyUnemployed	DI	0.2052

we see that the negative values are correlated with the population density and the criminal cases the patient is involved. When those are high and the Demographic indicators low, $PCA2 < 0$. If the opposite happens, then $PCA2 > 0$. So PCA2 positive values are correlated with the severity of the victimization events, the group age (the older, the greater the value), the intensity of involvement with the criminal justice system, and the unemployment rate near the treatment center.

4.1.2 K-Means clustering

To decide the best values for the number of clusters k hyperparameter, we run K-Means over the data with different values for k. For each clustering setup we compute the silhouette score. Plotting the silhouette score versus the value of k allows us to spot what values of k are best. Silhouette score versus k for K-means is shown in Figure 4. From there, we select the two best values for k which are $k = 2, 3$.

A visual representation of the clustering (performing the PCA dimensionality reduction) can be seen in Figure 5a and Figure 5b for $k = 2$ and $k = 3$, respectively.

When selecting $k = 2$ we can see that the two clusters obtained are samples with $PCA1 < 0$ and samples with $PCA1 > 0$, which is equivalent to saying that one cluster (Cluster 0) contains patients with milder or null mental health indicators while the other (Cluster 1) contains patients with high mental heath indicators.

With that, it may seem reasonable to hypothesize that patients in Cluster 0 will have less difficulties to follow the treatment than patients in Cluster 1, and this will potentially lead to longer relapse times in Cluster 0 than in Cluster 1.

Samples in this problem can be divided into two different groups, uncensored and censored, as explained previously. Inside the censored group we can distinguish between two reasons for censoring the data: (a) Censored 1: ending the 1-year follow-up period without any relapse or (b) Censored 2: failing to follow up.

Table 6 shows this breakup per cluster. We can see that Cluster 0 has a majority of censored samples while Cluster 1 has a majority of uncensored samples. We can conclude that people with greater mental health problems may be less likely to follow up, and for those who follow up consistently, it can be expected that relapse times are shorter.

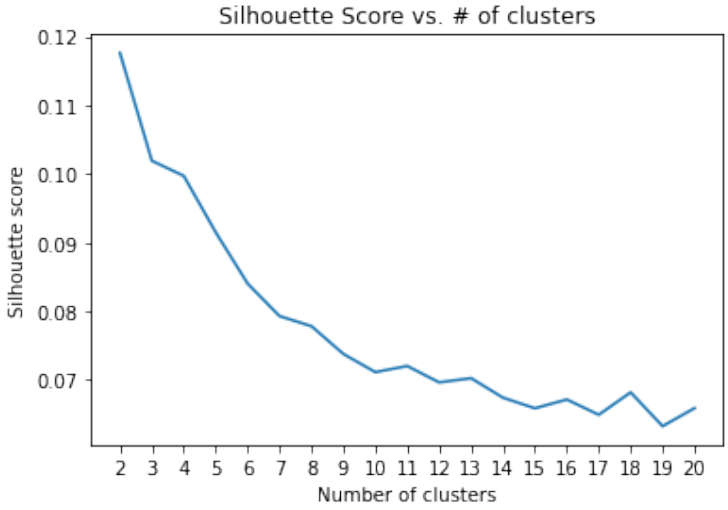
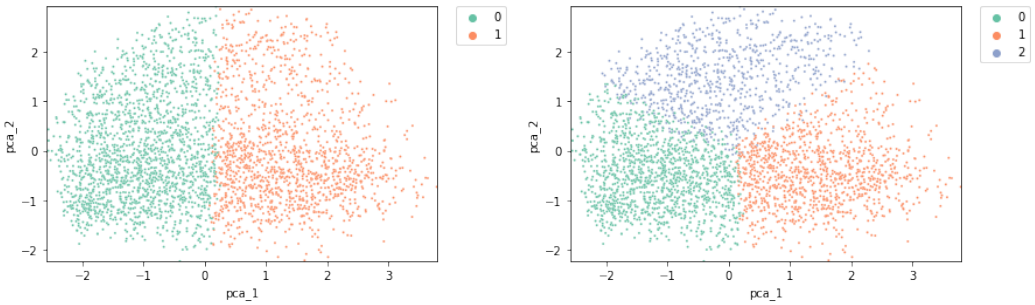


Figure 4: Silhouette score versus k value for K-means clustering.



(a) $k = 2$ (b) $k = 3$

Figure 5: K-Means clustering visualization for different values of k

Table 6: Metrics per cluster for K-Means clustering ($k = 2$). Median and Average show the median and average Alcohol Days value for all the uncensored samples in the cluster. Pct. relapsed, Pct. Censored 1 and Pct. Censored 2 make reference to the percentage of uncensored samples, censored samples due to end of the 1-year follow up period and censored samples due to the patient missing to follow up, respectively.

	Cluster	
Metric	0	1
Median	30	23
Average	46.21	38.83
% Relapsed	48.10	60.56
% Censored 1 (Not relapsed before the end of study)	11.52	6.85
% Censored 2 (Failed to follow up)	40.37	32.57

Table 7: Metrics per cluster for K-Means clustering ($k = 3$). Median and Average show the median and average Alcohol Days value for all the uncensored samples in the cluster. Pct. relapsed, Pct. Censored 1 and Pct. Censored 2 make reference to the percentage of uncensored samples, censored samples due to end of the 1-year follow up period and censored samples due to the patient missing to follow up, respectively.

	Cluster		
Metric	0	1	2
Median	30	21	45
Average	41.19	37.71	57.59
% Relapsed	51.88	66.01	38.53
% Censored 1 (Not relapsed before the end of study)	10.71	6.10	12.26
% Censored 2 (Failed to follow up)	37.39	27.87	49.19

All that would explain the low percentage of Censored 1 samples (at least one year without a relapse) compared with Censored 2 samples (failed to follow up) and also the shorter Median and Average values for Alcohol Days of the uncensored samples in Cluster 1.

When selecting $k = 3$ we can see a similar but more refined clustering. All the samples with positive PCA2 are in one cluster. For the samples with negative PCA2, samples with negative PCA1 are in one cluster and the rest are in another. This means that patients living in not highly populated zones and with low involvement in the criminal justice system are in one cluster (Cluster 2).For the ones living in highly populated zones and with high involvement in the criminal justice, the ones with low mental health indicators are in one cluster (Cluster 0) and the ones with high mental health indicators in another (Cluster 1).

In Table 7 we can see that Clusters 0 and 1 have similar values for all the metrics to Clusters 0 and 1 obtained when setting $k = 2$. Then the difference here will reside in how to explain Cluster 2 samples. Given that the patients in that cluster live in zones with potentially better social conditions (again, low unemployment and not crowded) it will make sense that the cluster had greater Censored 1 percentage than the other 2 (more patients not relapsing within the 1-year period) and among the uncensored patients, the largest Mean and Average Alcohol Days values. Data in the table confirms the trends. Something still to be explained is the high percentage of Censored 2 patients in that cluster (almost half of the patients failed to follow up in Cluster 2).

4.1.3 Agglomerative clustering

In this case we also obtained the Silhouette Score for each value of k. The plot can be seen in Figure 6. In it we can see that the best values for k are again $k = 2, 4$.

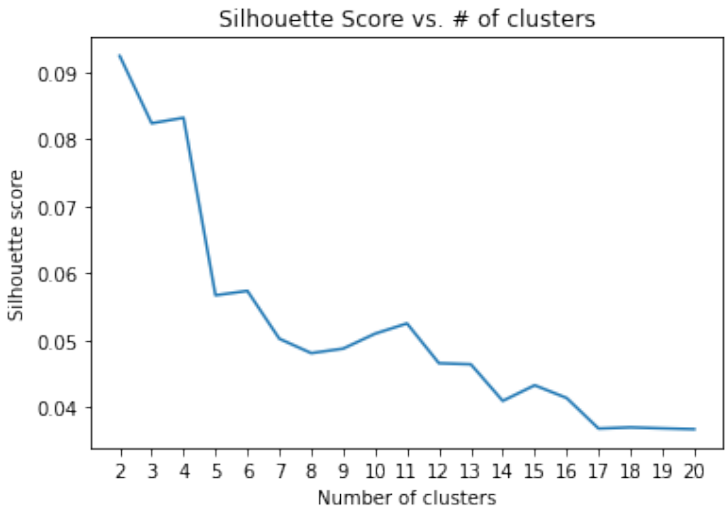
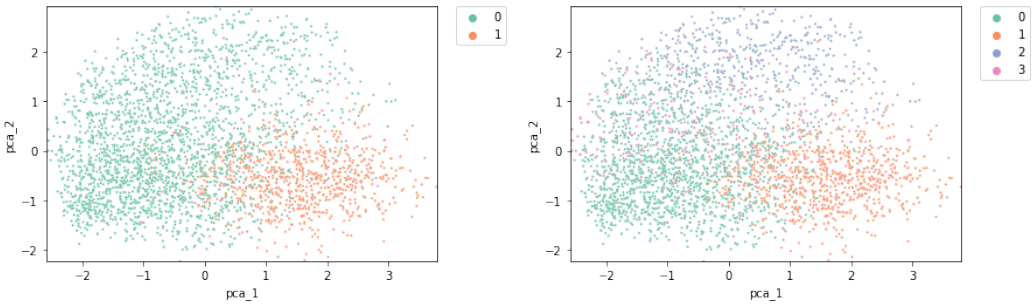


Figure 6: Silhouette score versus k value for Agglomerative clustering.

Visual representations of the clusters obtained can be seen in Figure 7a and Figure 7b for $k = 2$ and $k = 4$, respectively.



(a) $k = 2$ (b) $k = 4$

Figure 7: Agglomerative clustering visualization for different values of k

Visualizations in this case are worse. Separation between clusters is less defined in the case of $k = 2$ and is almost non-existent between clusters 2 and 3 in the case of $k = 4$.

If we have a look at the results for Agglomerative Clustering and $k = 2$ shown in Table 8, we can check that the values for each cluster are really similar to the clusters obtained with K-Means and $k = 2$.

Table 9 shows metrics for Agglomerative Clustering and $k = 4$ we can see that Clusters 0, 1 and 3 have majority of uncensored samples. Among them Median and Average Alcohol Days values are quite close between them and the only remarkable difference between them is that Cluster 1 has way less Censored 1 samples than Cluster 0. In Cluster 2, the majority of samples are censored. Among the censored samples, the majority of them (54.59%) are related to patients who failed to follow up.

Without having more visually separable clusters, it is hard to find a reason behind the clusters obtained by the Agglomerative Clustering algorithm. A possible drawback that these two clustering algorithms have is that they are using distances to compute the similarity between samples. However, it is worth mentioning again that our samples are vectors of categorical features, therefore, distance may not be the best similarity metric given the data.

Table 8: Metrics per cluster for Agglomerative clustering ($k = 2$). Median and Average show the median and average Alcohol Days value for all the uncensored samples in the cluster. Pct. relapsed, Pct. Censored 1 and Pct. Censored 2 make reference to the percentage of uncensored samples, censored samples due to end of the 1-year follow up period and censored samples due to the patient missing to follow up, respectively. .

	Cluster	
Metric	0	1
Median	30	21
Average	45.51	36.92
% Relapsed	48.22	67.96
% Censored 1 (Not relapsed before the end of study)	11.09	5.14
% Censored 2 (Failed to follow up)	40.68	26.88

Table 9: Metrics per cluster for Agglomerative clustering ($k = 4$). Median and Average show the median and average Alcohol Days value for all the uncensored samples in the cluster. Pct. relapsed, Pct. Censored 1 and Pct. Censored 2 make reference to the percentage of uncensored samples, censored samples due to end of the 1-year follow up period and censored samples due to the patient missing to follow up, respectively.

	Cluster			
Metric	0	1	2	3
Median	25	21	45	87
Average	37.49	36.92	57.36	70.13
% Relapsed	50.28	67.96	35.81	58.33
% Censored 1 (Not relapsed before the end of study)	10.84	5.14	9.58	15.00
% Censored 2 (Failed to follow up)	38.87	26.88	54.59	26.67

4.2 Survival Analysis Overview

The goal of survival analysis is to predict the time to an event of interest. The event of interest is a patient's first drink after treatment, otherwise known as the time to relapse. There are three main advantages to survival models over linear regression models for clinical data. First, survival models can use right censored data. Table 3 shows that there are 1,531 right censored patients in the longitudinal study (46% of the data). In other types of models 46% of the data would need to be discarded. Second, survival models output survival and hazard functions rather than point estimates. A survival function $S(t)$ is the probability of surviving –not relapsing in our problem– up to time t. This probability makes the model easier to interpret as instead of trusting one point estimate of time to survival, a probability over time is returned. Finally, survival models estimate the hazard function for individuals, which indicates the probability that a patient relapses at time t, *given* that they have not yet relapsed.

This section will introduce 7 survival models, the evaluation metrics used to evaluate the accuracy of each model, as well as discuss model results.

4.2.1 Evaluation Methods

Two common evaluation metrics for survival modeling are C-index and Brier Score. The Brier score is used to mainly assess calibration: how close our model is to the actual prediction. The C-index is the most common measure of discrimination: whether our model is able to predict risk scores that allow us to correctly determine the order of the events of interest. Since the C-index disregards the actual values of predicted risk scores – it is a ranking metric – it is unable to tell us anything about calibration. In our experiments, we focus more on Brier score, because it is able to both calibrate

and discriminate, and we care about being more accurate with predicting relapse for individuals, rather than ordering relapses of individuals correctly.

- **Concordance index (C-index):** this metric cares about the rank of the predictions rather than their actual values. For every pair of two individuals in our study, we compare the rank of their predictions against their actual time-to-event, and if their order is correct, we add this pair to our list of concordant pairs. The number of concordant pairs divided by the total pairs gives us the C-index score. A score of 1 is considered the best possible score, and a score of 0 is the worst possible score. 0.5 is considered equivalent to random guessing. This method supports right censoring, because we know for sure that patients who are right censored have survival time greater than or equal to the patient they're being compared against.
- **Integrated Brier score (IBS):** this metric estimates the time-dependent Brier score for right censored data. Brier score is the mean squared error of the estimated probability at every time step in the survival function. For survival analysis, to account for right censored data, it is weighted by the inverse probability of censoring weight, estimated by the Kaplan-Meier estimator. We have used Integrated Brier Score (IBS) as our evaluation metric, as it provides the cumulative Brier Score over the entire time period we are evaluating for.

4.2.2 Kaplan-Meier Estimator

The Kaplan-Meier estimator is a non-parametric model. It does not consider any features: it instead estimates a survival function only from the censoring variable and time-to-event. The estimated curve is a step function, with steps occurring at time points where one or more individuals relapsed. Using the Kaplan-Meier approach, it is possible to estimate survival curves for smaller populations by dividing the dataset into smaller sub-groups according to a variable. If we want to consider more than 1 or 2 variables however, this approach becomes infeasible, because subgroups will get very small.

4.2.3 Cox Proportional Hazard Model

The Cox Proportional Hazard Model (Cox PH) is a semi-parametric model that can be used to predict time to an event and feature importance. Cox PH specifies a linear-like model for the log hazard [10]. The hazard ratio is the ratio of the actual (all feature values set) hazard versus the baseline group hazard (with no feature values set). An assumption of proportional hazards regression is that the hazard ratio is constant over time. Cox PH is named proportional hazard because it makes a strong assumption that the hazard ratio of two individuals is proportional or does not vary with time [10]. While Cox PH is a popular model, it has notable drawbacks in practice. The model fails when features are highly correlated or when the number of features approaches the number of observations [11]. Training a model with all features fails using the vanilla Cox PH, so no results are reported.

4.2.4 Cox PH with Penalty

To overcome the shortcoming of Cox PH, a penalty can be added to feature coefficients to remove or greatly reduce the effect of the feature. The following three Cox PH penalty models were evaluated:

- **Cox PH with Ridge Regression:** Adding ridge regression to the Cox PH model reduces the effect of some of the features by multiplying the feature by a very small number. Ridge Regression does not reduce the number of features.
- **Cox PH with Lasso Penalty:** The lasso penalty will reduce the number of features, by making some of the coefficients zero and choosing a subset of features.
- **Coxnet:** Coxnet combines the penalties from both ridge regression and lasso. A mixing parameter is selected as a ratio, where 0 means the ridge regression penalty is used and 1 means the lasso penalty is used.

Table 10: The table shows Brier Score and Concordance Index for each one of the models. The random survival forest model performs the best as it has both the lowest brier score and the highest concordance index. Cox PH with Lasso Penalty performs well too.

Model Name	Integrated Brier Score	Concordance Index
Kaplan-Meier Estimator	0.243	-
Cox PH with Ridge Regression Penalty	0.217	0.609
Cox PH with Lasso Penalty	0.216	0.612
Coxnet	0.216	0.612
Survival Trees	0.310	0.579
Random Survival Forests	**0.206**	**0.642**
Neural Multi-Task Logistic Regression	0.223	0.621

4.2.5 Survival Trees

Survival trees is a simple, non-parametric method that holds less restrictive assumptions than Cox proportional hazard model. It originated from the idea of regression trees, and is created by putting the data points into groups. The goal is to maximize the difference in response between groups, and this can be done using any survival metrics discussed above.

4.2.6 Random Survival Forests

Random survival forests is a non-parametric, ensemble method, that uses multiple survival trees. It is based on the concept of random forests, but it also incorporates right censored data. For prediction, a sample traverses down each tree in the forest until it reaches a terminal node. Data in each terminal is used to estimate the survival and cumulative hazard function. The ensemble model uses the average of all decision trees to predict a value.

4.2.7 Neural Multi-Task Logistic Regression

Multi-Task Logistic Regression (MTLR) is a series of logistic regression models built on different time intervals with the goal of estimating the probability that an event of interest happened within each of those intervals. MTLR also does away with the Cox proportional hazards assumption and additionally, can be better for predicting survival graphs, since it is time dependent modeling. Neural MTLR combines neural networks (for non-linear modeling) with the MTLR technique. Since this method is a classification technique, while training this model, we specified 365 bins (1 for each day in the year of the study) to achieve better granularity. The hyperparameters set for the model are 1e-3 as learning rate, 100 epochs and ReLU activation function.

4.2.8 Results

To obtain preliminary baseline results, we trained each model using all features except the number of days a patient drank alcohol in each period between surveys. The data was split 80-20% into train and test set. Right censored data (CensorCatergory in Appendix A =1,2) was set to False in the dataset to indicate the event of interest (relapse) did not occur, and the remaining data (CensorCatergory in Appendix A =0) was set to True in the dataset to indicate our event of interest did occur.

Each model was evaluated using both the Integrated Brier Score (IBS) and C-index. Kaplan-Meier is the simplest model estimating survival for the entire population without using any features. The estimated curve is a step function, with steps occurring at time points where one or more individuals relapsed. Kaplan-Meier can thus be considered a baseline for all the other models to improve upon. Note that this curve in Figure 8 is very similar to Figure 3a, which justifies how the population's survival changes over time. The survival curve for the full population is higher because intuitively,

the uncensored patients all relapsed. Thus, the survival curve for the uncensored population will fall faster than the whole population.

Table 10 shows the model evaluation results. Here we can see that Random Survival Forests (RSF) are performing best in terms of IBS and C-index. Cox PH with Lasso Regression Penalty is also performing well. Survival Trees perform worse than the baseline, because it is a fairly basic decision tree and likely cannot model more complex behavior. This issue is solved by using the aggregation of many such trees, which is exactly why we believe RSF performs much better. Neural MTLR also shows good promise in terms of the metrics, so we plan to explore this model and what features it deems important in further detail.

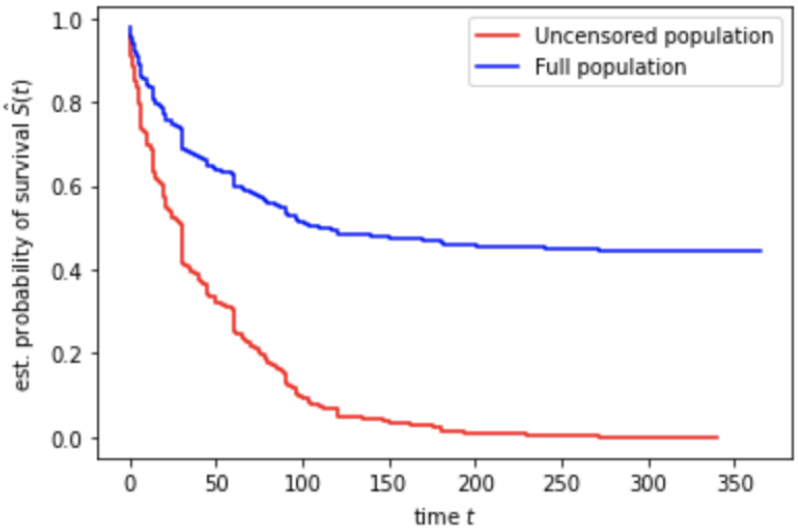
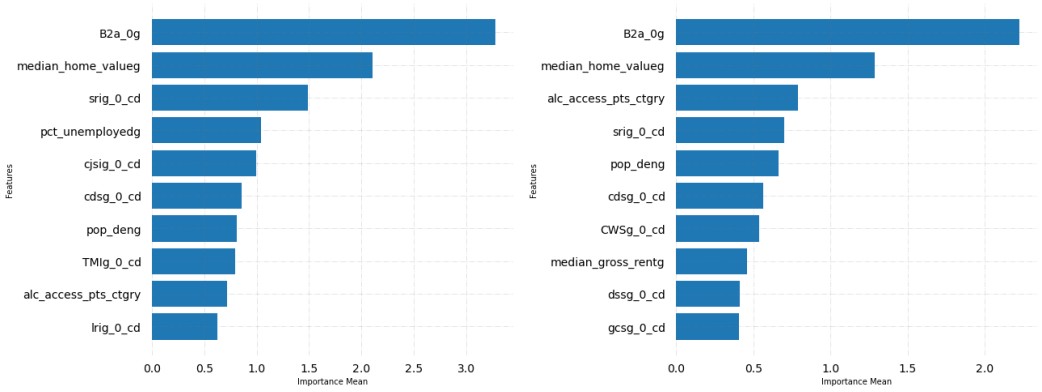


Figure 8: This figure shows the Kaplan Meier (KM) estimator survival curve for the entire population versus the uncensored patients only.

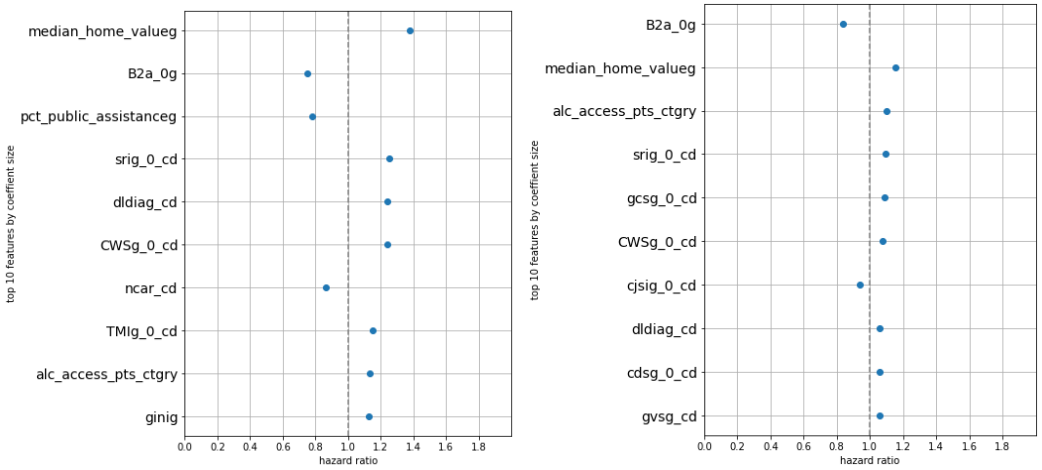


(a) Survival Tree Feature Importance

(b) RSF Feature Importance

Figure 9: Top 10 features selected by Survival Trees and Random Survival Forests

We performed feature importance using the Permutation Feature Importance algorithm for non-parametric models (RSF and Survival Tree). The top 10 features across both models are shown in Figures 9a and 9b. The survival tree model identified that the most important features as AgeGroup, MedianHomeValue and FriendsWithOthersWithAUD. While in RSF, the most important features are the AgeGroup, MedianHomeValue and AlcoholAccessPoints. An interesting finding is that



(a) Ridge Regression Hazard Ratios

(b) Lasso Penalty Hazard Ratios

Figure 10: Top 10 features selected by different Cox PH models

across these models, we see 6 overlapping most important features: AgeGroup, MedianHomeValue, FriendsWithOthersWithAUD, PopulationDensity, AlcoholAccessPoints and Delinquency.

For Cox PH models, we obtain the 10 most important features using the Hazard Ratio values. For Ridge Regression, we can see in Figure 10a, the most important positive features having HR $>$ 1.0 Witkiewitz et al.(increased risk of relapse) are MedianHomeValue and FriendsWithOthersWithAUD, while the negative features having HR $<$ 1.0 (decreased risk of relapse) are AgeGroup and PctPublicAssistance. For Lasso Regression, we can see in Figure 10b, the most important positive features are MedianHomeValue and AlcoholAccessPoints, while the negative features are AgeGroup and CriminalJusticeEngagement. With categorical features, it isn't easy to understand how hazard ratios of the features affect relapse risk. To better explain this, we plan to one-hot encode our categorical features, as shown in Davis et.al. [7], and rerun this experiment.

Comparing the non-parametric models with the parametric Cox models, MedianHomeValue AgeGroup, FriendsWithOthersWithAUD and AlcoholAccessPoints contribute significantly across all models.

5 Remaining parts

As stated in the project proposal, the team has completed, Task 0 and most of Task 1.

In Task 0 data cleaning and data visualization was performed. As part of Task 1, survival model baselines have been obtained using the whole dataset. In addition, clustering has been performed over the data to see if interesting clusters were obtained.

Time-series clustering has been proven to not be useful for this task after obtaining the results. As for the interpret-ability of the cluster results obtained with K-Means and Agglomerative clustering, more exploration needs to be done. As we are dealing with vectors of categorical features, it is highly probable that the selected clustering algorithms are not the best choice as they compute distances between the vectors to find the closest samples. In the coming weeks, we will explore a new clustering algorithm specifically designed to deal with categorical values: K-Modes, which is based on K-Means. However, K-Modes uses dissimilarities (total mismatches of vector positions with respect to the same positions of another vector) to update the centroids. The lower the number of mismatches, the more similar two vectors (or samples) are. With that, we aim to achieve (a) meaningful categorical centroids and, (b) potentially more explainable clusters and visualizations.

We also plan to rerun our experiments with a smaller feature set, selected by analyzing feature correlation (for example, Region and PctPublicHealthcare are highly correlated and one of them could be removed) and feature importance. We also plan to understand whether the features are positively or negatively correlated to the survival by combining results from the non-parametric methods (RSF) with the parametric methods (Cox). Additional work will also be done to interpret the Cox Hazard Ratios reported in Figures 10a and 10b. Finally, we plan to try some Deep Learning models to generate non-linear interactions.

Remaining tasks are Task 2 and, if possible, Task 3. For Task 2, we will create expert defined clusters to train a survival model for each cluster to see if this increases model performance. We would also use good clusters obtained in Task 1 to see what are the differences they have with respect to expert-defined clusters. Additionally, Kaplan-Meier estimator can be used to divide the population based on features of interest (for example people with ADHD versus people who do not have ADHD) and see the survival curves for both of these subgroups without considering any other features.

For task 3, literature research will be performed to see if deep neural models can be applied for survival tasks like this one.

References

[1] WHO, "Alcohol," May 2022.

[2] N. I. on Alcohol Abuse and Alcoholism, "Understanding alcohol use disorder," Apr 2021.

[3] U. Nations, "Goal 3 — department of economic and social affairs," 2021.

[4] K. Witkiewitz, A. D. Wilson, M. R. Pearson, K. S. Montes, M. Kirouac, C. R. Roos, K. A. Hallgren, and S. A. Maisto, "Profiles of recovery from alcohol use disorder at three years following treatment: can the definition of recovery be extended to include high functioning heavy drinkers?," *Addiction*, vol. 114, no. 1, pp. 69–80, 2019.

[5] Y. Wen, M. F. Rahman, Y. Zhuang, M. Pokojovy, H. Xu, P. McCaffrey, A. Vo, E. Walser, S. Moen, and T.-L. B. Tseng, "Time-to-event modeling for hospital length of stay prediction for covid-19 patients," *Machine Learning with Applications*, vol. 9, p. 100365, 2022.

[6] X. Dai, J. H. Park, S. Yoo, N. D'Imperio, B. H. McMahon, C. T. Rentsch, J. P. Tate, and A. C. Justice, "Survival analysis of localized prostate cancer with deep learning," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[7] J. P. Davis, P. Rao, B. Dilkina, J. Prindle, D. Eddie, N. C. Christie, G. DiGuiseppi, S. Saba, C. Ring, and M. Dennis, "Identifying individual and environmental predictors of opioid and psychostimulant use among adolescents and young adults following outpatient treatment," *Drug and alcohol dependence*, vol. 233, p. 109359, 2022.

[8] M. L. Dennis, J. C. Titus, M. K. White, J. I. Unsicker, and D. Hodgkins, "Global appraisal of individual needs: Administration guide for the gain and related measures," *Bloomington, IL: Chestnut Health Systems*, 2003.

[9] M. Ives, R. Funk, P. Ihnes, T. Feeney, M. Dennis, and G. C. Center, "Global appraisal of individual needs evaluation manual," *Bloomington, IL: Chestnut Health Systems.[Google Scholar]*, 2010.

[10] W. Fox, "Cox proportional-hazards regression for survival data in r," *An R Companion to Applied Regression, Second Edition*, 2011.

[11] H. T. T. R. Simon N, Friedman J, "Regularization paths for cox's proportional hazard model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

A Appendix: Data Dictionary

Table 11: Data dictionary of the final dataset used in analysis.

Category	No.	Data Type	English Variable Name	Variable Name	Short Description
demographic	1	boolean	Northeast	region_ne	Northeast region: T/F
	2	boolean	Southeast	region_se	Southeast region: T/F
	3	boolean	Midwest	region_mw	Midwest region: T/F
	4	boolean	Southwest	region_sw	Southwest region: T/F
	5	boolean	West	region_w	West region: T/F
	6	boolean	CurrentlyUnemployed	unemplmt_cd	Is patient currently unemployed?
	7	categorical	Homeless	homeless_0_cd	Is patient homeless?
	8	categorical	AgeGroup	B2a_0g	0: adolescent (12-17), 1 - young adult (18-29), 2 - adult (30+)
	9	boolean	White	Raceg4_gr_1	White/Caucasion
	10	boolean	Black	Raceg4_gr_2	Black/African American
	11	boolean	Hispanic	Raceg4_gr_3	Hispanic
	12	boolean	OtherRace	Raceg4_gr_4	Other Race
	13	categorical	PopulationDensity	pop_deng	Population Density: Low, med, high
	14	categorical	PctUnemployed	pct_unemployedg	Percent unemployed in area: Low, med, high
	15	categorical	PctPublicAssistance	pct_public_assistanceg	Percent receiveing public asssistance: Low, med, high
	16	categorical	PctPoverty	pct_povertyg	Percent below poverty line: Low, med, high
	17	categorical	MedianFamilyIncome	median_family_incomeg	Median family income: Low, med, high
	18	categorical	MedianHomeValue	median_home_valueg	Median home value: Low, med, high
	19	categorical	MedianGrossRent	median_gross_rentg	Median gross monthly rent: Low, med, high
	20	categorical	GiniIndex	ginig	Census Gini Index: Low, med, high
	21	categorical	PctPublicHealthcare	pct_on_public_healthcareg	Percent on public healthcare: Low, med, high
environment	22	categorical	AlcoholAccessPoints	alc_access_pts_ctgry	Number of alcohol access points in area (bars, liquor stores, etc.): Low, med, high
mental health	23	categorical	TreatmentReisistance	TRIg_0_cd	Resistance to treatment: Low, med, high
	24	categorical	VictimizationEvents	gvsg_cd	Experienced victimization events: Low, med, high
	25	boolean	SUDAndMentalHealthDisorder	dldiag_cd	SUD and co-occuring mental health disorder: T/F
	26	categorical	Depression	dssg_0_cd	Depression symptoms: Low, med, high
	27	categorical	GlobalEmotionalProblems	epsg_0_cd	Global emotional problems (wars, unrest, etc.): Low, med, high
	28	categorical	ADHD	adhdg_0_cd	Likelyhood of ADHD: Low, med, high
	29	categorical	Delinquincy	cdsg_0_cd	Delinquency or conduct disorder: Low, med, high
	30	boolean	Suicidal	suicprbs_0_cd	Reported problems of suicide: T/F
social factor	31	categorical	CriminalJusticeEngagement	cjsig_0_cd	Engagement with criminal justice system: Low, med, high
	32	categorical	LivingWithOthersWithAUD	lrig_0_cd	Living with others with AUD: Low, med, high
	33	categorical	FriendsWithOthersWithAUD	srig_0_cd	Friends with others with AUD: Low, med, high
	34	categorical	CriminalActivity	gcsg_0_cd	Involvement in criminal activity in last 90 days: Low, med, high
	35	categorical	NotCloseToSomeoneInRecoveryAUD	ncar_cd	Not close to someone in active recovery: T/F
substance use	36	boolean	SUDUnder15	und15_cd	Age of first substance use under 15: T/F
	37	categorical	WithdrawlSymptoms	CWSg_0_cd	Reported withdrawl symptoms: Low, med, high
	38	numeric	DaysConsumedIntake	S2a1_0	The # days of alcohol use @ intake
	39	numeric	DaysConsumedThreeMonths	S2a1_3	The # days of alcohol use @ month 3 survey
	40	numeric	DaysConsumedSixMonths	S2a1_6	The # days of alcohol use @ month 6 survey
	41	numeric	DaysConsumedTwelveMonths	S2a1_12	The # days of alcohol use @ month 12 survey
target	42	boolean	CensorVariable	Alcohol_Cens	Censored/Uncensored
	43	categorical	CensorCategory	Alcohol_Cens_Ctgry	0: uncensored, 1: censored by end of study period, 2: lost to followup
	44	numeric	DaysToRelapseAUD	Alcohol_Days	Time in days until first alcohol consumption after beginning treatment
treatment	45	boolean	EngagmentLevel	engage30	Engagement in treatment for 30+ days and 3+ sessions: T/F
	46	boolean	PriorSUDTreatment	prsatx_cd	Prior substance abuse treatment: T/F
	47	categorical	TreatmentMotivation	TMIg_0_cd	Motivation for successful treatment: Low, med, high
	48	categorical	ConfidenceToAvoidSubstances	SESg_0_cd	Self confidence to avoid substance use during treatment: Low, med, high

B Appendix: Descriptive Statistics

Table 12: Count of categorical variables with levels low, medium, high.

Variable	LOW	MED	HIGH
pct_public_assistanceg	0	2727	563
ginig	0	1027	2263
CWSg_0_cd	2925	300	65
SESg_0_cd	2119	910	261
cdsg_0_cd	1987	1008	295
B2a_0g	1912	741	637
Alcohol_Cens_Ctgry	1759	313	1218
adhdg_0_cd	1692	849	749
gcsg_0_cd	1678	980	632
TRIg_0_cd	1475	1662	153
dssg_0_cd	1473	1150	667
epsg_0_cd	1284	1633	373
alc_access_pts_ctgry	1268	961	1061
gvsg_cd	1176	590	1524
cjsig_0_cd	1127	630	1533
TMIg_0_cd	520	2371	399
median_family_incomeg	478	2557	255
pct_on_public_healthcareg	470	2724	96
pct_unemployedg	396	2500	394
median_gross_rentg	378	2535	377
pop_deng	350	1269	1671
srig_0_cd	41	1198	2051
lrig_0_cd	34	1925	1331
pct_povertyg	31	2702	557
median_home_valueg	15	2589	686

Table 13: Count of categorical variables with levels low, medium, high.

Variable	Type	FALSE	TRUE
engage30	boolean	2255	1035
region_ne	boolean	3113	177
region_se	boolean	2611	679
region_mw	boolean	2849	441
region_sw	boolean	2508	782
region_w	boolean	2079	1211
unemplmt_cd	boolean	2447	843
prsatx_cd	boolean	1996	1294
und15_cd	boolean	1054	2236
dldiag_cd	boolean	602	2688
suicprbs_0_cd	boolean	2995	295
homeless_0_cd	boolean	2741	549
ncar_cd	boolean	895	2395
Raceg4_gr_1	boolean	2168	1122
Raceg4_gr_2	boolean	2921	369
Raceg4_gr_3	boolean	2091	1199
Raceg4_gr_4	boolean	2690	600
Alcohol_Cens	boolean	1759	1531

Table 14: Count of categorical variables with levels low, medium, high.

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Null Count
Alcohol_Days	0	25	91	136.5	195	365	0
S2a1_0	0	0	3	10.61	12	90	6
S2a1_3	0	0	0	4.932	4	90	474
S2a1_6	0	0	0	4.944	4	90	640
S2a1_12	0	0	0	4.836	4	90	1715