Time-to-relapse predictions for alcohol use disorder patients

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

Bistra Dilkina, PhD
Associate Professor
Co-Director, USC Center for AI in Society
University of Southern California
dilkina@usc.edu

Jordan Davis, PhD
Assistant Professor
Associate Director, USC Center for AI in Society
University of Southern California
jordanpd@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 homogeneous subgroups within the heterogeneous 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 [1]. Moreover, it is a, "psychoactive substance with dependence-producing properties" and it is known to cause over 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. AUD treatment outcomes are often measured by abstinence, however another way to measure treatment outcomes is the level of function (low, medium, high) a patient has, post-treatment. For example, a patient who continues to engage in alcohol consumption post treatment may function better and have a higher quality of life than a person who completely abstains from alcohol consumption. In this project, we expand on this idea by developing a framework to understand the driving factors behind AUD relapse events. Our framework is divided into three tiers with each tier contributing to the overarching objective of using ML models to inform targeted AUD treatment programs: ***(1)*** *identify and understand homogeneity among various patient subgroups with respect to relapse events and other associative factors,* ***(2)*** *accurately predict time-to-relapse using survival analysis models,* ***(3)*** *determine and explain the key factors contributing to AUD relapse events across patient subgroups and the entire population.*

1.3 Project Outline

To successfully achieve the objectives of this project, we have broken the problem into two distinct stages which are summarized in Figure 1. In the first stage, we will cluster patients on their demographics and other relevant features. The objective of this clustering task is to determine which homogeneous subgroups are implicitly present within the data. In addition to clustering, we will also train survival models for the entire population and explain the driving factors behind these relapse predictions using model feature importance. In the second stage of the project, we will use the clustered patient subgroups as well as expert defined subgroups to train several survival models for each patient subgroup. The intent is to understand how these subgroup level survival models perform in comparison to the overall population and extract insights for each subgroup's relapse traits.

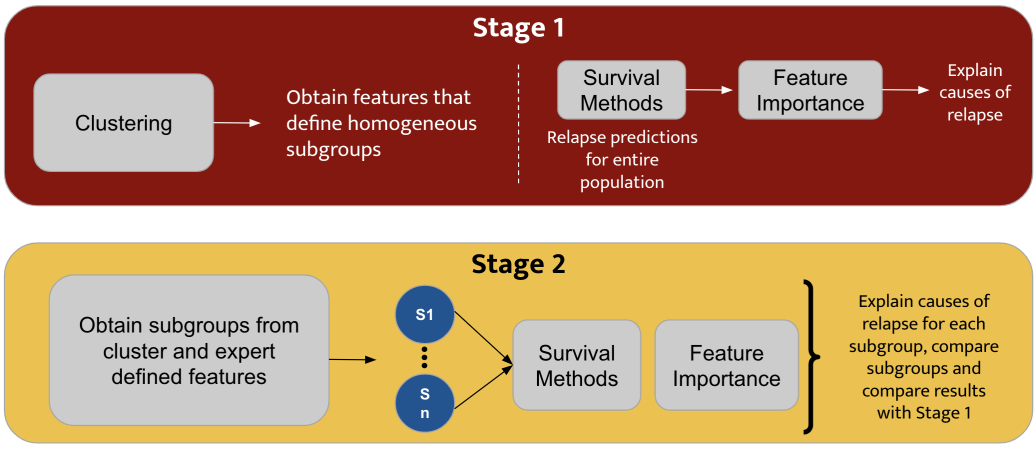


Figure 1: Two primary stages for determining the driving factors behind AUD relapse events. Stage 1 consists of clustering to determine patient subgroups and survival modeling across the entire patient population. Stage 2 builds upon stage 1 using clustered results and expert defined subgroups to determine the factors driving relapse events at a more granular level.

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 [4] authors used generic survival models such as the 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. The authors found that 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 non-linear relationships and that Random Survival Forests (RSF) provided a good C-index score, however, failed to perform when it came to other metrics like Brier score, due to the small dataset size. In [5], deep learning models were used in time-to-event predictions 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. The authors used the models to perform two different tasks: time-to-mortality and time-to-diagnosis. 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 understand the rank-order predictors (key factors) that lead to relapse. In a paper by Witkiewitz et al. [6] a latent class analysis was used to determine the level of function patients experienced following AUD treatment. Witkiewitz used 4 patient subgroups and concluded that treatment outcomes regarding a patient's level of function and quality of life can be determined using these subgroups. In Davis et. al. [7], the individual and environmental predictors of adolescent opioid use were examined using lasso regression and RSF. Based on their analysis, a mix of individual and environmental variables proved to be strong predictors of substance use for all severities (low, medium, high usage). Time-to-relapse models stratified by severity were influenced by a distinct set of features which indicates that SUD treatment should be tailored to the severity as well as the substance.

Building on some of the characteristics of these previous publications, we aim to develop a methodology to inform better AUD treatment programs. Our goal is identify (through cluster analysis and expert opinion) patient subgroups using methods from [6], predict time-to-relapse using ML and DL models outlined in [4] and [5], and utilize techniques performed in [7] to determine, quantify, and explain the driving factors of AUD relapse within each subgroup.

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.

Two data sources are used to create final dataset. The first data source, the Global Appraisal of Individual Needs (GAIN) encodes data gathered from patients aged 12+, before, during, and after undergoing treatment for a substance use disorder (SUD) [8] from 137 treatment centers in the United States [7]. Each participant completed a baseline assessment upon entering treatment and completed follow-up assessments at 3, 6, and 12 months [7]. It is important to note that the treatment program 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

Table 1: Details of censor variable. In the final analysis only censored/uncensored is 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. Patient was lost due to failure to followup on day d_2

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.

The primary dataset was provided by Davis et al., who previously did the work to process and join the two data sources in [7]. The dataset provided contains treatment data for various types of SUD treatments, therefore the dataset was filtered to contain only patients being treated for AUD leaving a total of (n = 3290) patient samples. Aside from filtering, 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 data contained few Null values with Nulls only present in 4 columns. The columns containing Null values were followup survey questions regarding the number of days where alcohol was consumed in each period. This result is expected since the null values corresponded to censored data. Therefore Null values were simply kept in the data. For each follow up period 0, 3, 6, 12 months, the count of Null values was 6 (0.18%), 474 (14.4%), 640 (19.45%), and 1715 (52.13%) respectively (Appendix B, Table 8).

3.2 Data Description

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 or refusal to participate in treatment. To account for these cases, an added variable –called a censoring variable– allows us to understand whether a patient has experienced a relapse event or has stopped participating in the study. This variable takes on three different values 0, 1, 2 which are described in Table 1. In the final dataset, the censor values have been modified to indicate the event that an individual has not relapsed (1, 2) with a value of True (censored) and those uncensored individuals who experienced a relapse (0) event are denoted with a value of False.

The variables *DaysToRelapseAUD* and *CensorVariable* are used as target variables in the analysis and features concerning an AUD patient's substance abuse, treatment, environmental factors, etc. are used as the independent variables to train the model. A breakdown of the variable types is shown in Table 2. A detailed description of each variable's meaning is shown in Appendix A, Table 5. Counts for categorical and boolean variable levels as well as descriptive statistics for continuous variables are shown in Appendix B, Tables 6, 7,8 respectively.

For each variable pair in the final dataset, a correlation value was calculated. As most of the variables are categorical or binary standard correlation metrics cannot be computed. To understand how these 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. After creating the correlation matrix, we saw that correlations between variables were mostly centered around zero. We set a threshold of 0.80 for high correlation where features with correlation values exceeding an absolute value of 0.80 were considered to be removed. Analyzing the pairwise correlations revealed no correlation pairs to be greater than an absolute value of 0.80 so as a result, all features were kept in the analysis. A few variables had relatively high correlations with the highest correlation being 0.67.

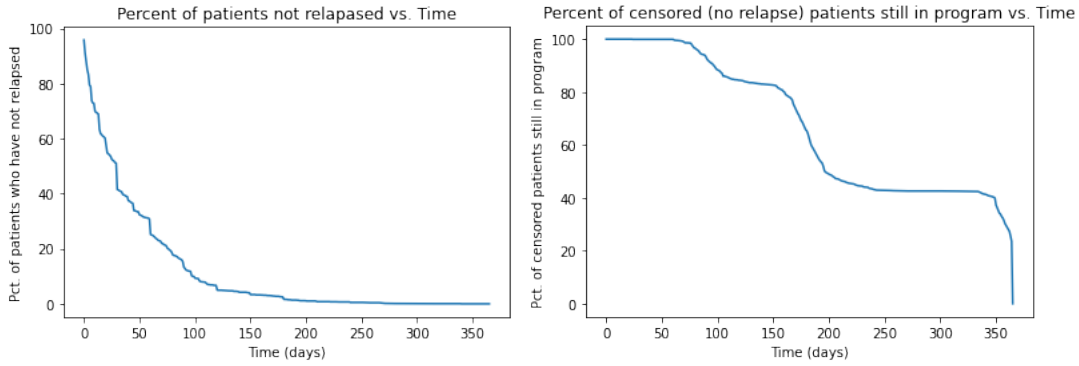
Table 2: Breakdown of variable types in the final dataset.

Variable Type	Boolean	Categorical	Continuous	Total	Example
Demographic	12	10	0	22	Race Group, Age Group, Gender
Environment	0	1	0	1	Alcohol Access Points
Mental Health	3	6	0	9	Depression, Suicidal Thoughts
Social Factors	1	4	0	5	Not Close to Anyone in Recovery
Substance Use	1	1	4	6	Withdrawal Symptoms
Treatment	2	2	0	4	Prior Substance Use, Treatment Motivation
Target	1	1	1	3	Days Until First Drink, Alcohol Censor
Total Count	20	25	5	50	

3.3 Ground Truth

In Figures 2a and 2b we can see the ground-truth time-to-event plotted as a cumulative distribution function for percentage of uncensored patients not relapsed and percentage of censored patients who are still in the study. In Figure 2a 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 bumps in the chart on 2b correspond to the followup periods at 3, 6, and 12 months when we would have information on whether or not the person attends their followup appointment. In 2b we see the steepest decline in percent of patients still in the study in between zero to 200 days meaning the majority of patients are lost due to followup which is confirmed in Table 3.

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).



(a) Uncensored percent not relapsed vs. Time.

(b) Censored percent still in program vs. Time.

Figure 2: Ground truth survival for uncensored (relapsed) patients and censored patients.

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	Description	Avg. Days to Relapse	Std. Dev.	Median Days to Relapse	Count	Pct. of Dataset
Uncensored	exact relapse known	**42.618**	**46.300**	**30**	**1759**	**53.47**
Censored	no relapse known	**244.271**	**105.964**	**197**	**1531**	**46.53**
	(1) no relapse in 12 month period (success)	355.419	7.841	357	313	9.51
	(2) lost to followup (relapse unknown)	215.709	100.532	184	1218	37.02
Censored, Uncensored	all patients	**136.457**	**128.411**	**91**	**3290**	**100**

4 Methods

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 longitudinal, 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 the metrics used to evaluate the accuracy of the survival models and all the survival models tried.

4.1 Evaluation Methods

Two common evaluation metrics for survival modeling are C-index and Integrated Brier Score (IBS). The IBS 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.

- **Concordance index (C-index):** this metric is concerned with 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.
- **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. An IBS score of 0 is considered the best possible score and 1 is the worst possible score.

4.2 Models

4.2.1 Kaplan-Meier Estimator

The Kaplan-Meier estimator is a continuous, non-parametric model used as a baseline in the study. It does not consider any features: it estimates a survival function only from the censoring variable and time-to-relapse. 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 subgroups. If we want to consider more than 1 or 2 variables however, this approach becomes infeasible, because subgroups will get very small.

4.2.2 Cox Proportional Hazard Model

The Cox Proportional Hazard Model (Cox PH) is a semi-parametric model that 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].

4.2.3 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.
- **Elastic net:** Elastic net combines the penalties from both ridge regression and lasso. A mixing parameter is selected (l1_ratio), where 0 means the ridge regression penalty is used and 1 means the lasso penalty is used.

4.2.4 Survival Trees

Survival trees is a simple, continuous, non-parametric method which 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 using evaluation metrics. We used 1000 estimators in our model training to obtain the best IBS and C-index score.

4.2.5 Random Survival Forests

Random survival forests is a continuous, non-parametric, ensemble method, that uses the average of multiple survival trees to predict a value. It is based on the concept of random forests, but it also incorporates right censored data. Data in each terminal is used to estimate the survival and cumulative hazard function. We used 1000 estimators in our model training to obtain the best IBS and C-index score.

4.2.6 Neural Multi-Task Logistic Regression

Neural MTLR is a neural, discrete model which combines neural networks (for non-linear modeling) with the Multi-Task Logistic Regression (MTLR) technique. 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. 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 as the activation function.

4.3 Population Subgroup Selection

To better understand relapse for homogeneous groups of patients (as explained in our motivation), there is a need to first define homogeneous subgroups in the overall heterogeneous population. We explored two ways of obtaining them: Clustering and Expert Definition.

For clustering, we used multiple clustering techniques and silhouette score as defined in Appendix D. We noticed that while the 2 clusters we obtained were able to define the population based on Post Traumatic Stress Disorder (PTSD) and Delinquency, the distribution of relapse among these

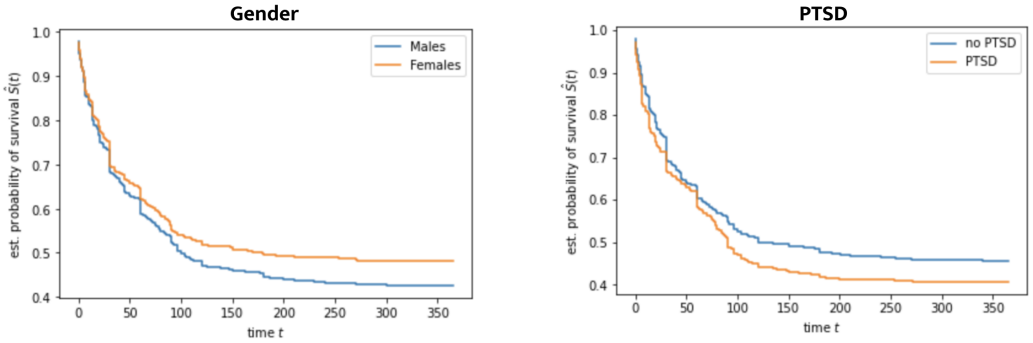


Figure 3: This figure shows the Kaplan Meier (KM) estimator survival curve for the 2 expert defined groups. The figure on the left shows Males versus Females, where Males relapse faster than Females. The figure on the right shows people with PTSD versus people without PTSD, where people with PTSD relapse faster.

clusters was more or less similar. Hence, we decided to not use these cluster definitions in further experiments.

In the Results section, we will only use the Expert defined subgroups. The subgroups have been obtained using a unique feature of interest for our experts and splitting samples by their distinct feature values. The two features are Gender and PTSD. Figure 3 shows the Kalpan-Meier curves for the sub-populations across those features.

5 Results

The dataset was split 80-20% into train and test set using stratified sampling to ensure equal distribution of censored and uncensored data. We dropped all variables related to the region where the treatment center was present, because Region was dominating all the models' feature importance. Thus, all our final results do not include the Region feature.

For evaluating the models, we used the Integrated Brier Score (IBS) for ranking our models and C-index for filtering out models which have a C-index score below 0.6. For the domain of addiction sciences, a C-index score above 0.6 is considered good because of the small overall dataset size and scarcity of time-to-event data; we only have approximately 1500 individuals who relapse during the course of the study.

5.1 Results & Relapse causes over entire population

Table 4: The table shows Integrated Brier Score (IBS) and Concordance Index (C-index) for each one of the models. The Random Survival Forest and Cox PH with Lasso Penalty models performs the best as they have both the lowest IBS and the highest C-index.

Model Name	Integrated Brier Score	Concordance Index
Kaplan-Meier Estimator	0.237	-
Cox PH with Ridge Regression Penalty	**0.195**	**0.665**
Cox PH with Lasso Penalty	0.200	0.660
Elastic net (l1_ratio=0.1)	0.197	0.662
Survival Trees	0.327	0.549
Random Survival Forests	**0.195**	**0.672**
Neural Multi-Task Logistic Regression	0.216	0.645

Table 4 shows the model evaluation results. The Random Survival Forests (RSF) model performs the best in both IBS and C-index for the non-parametric models. The Survival Trees model performs worse than the baseline (Kaplan-Meier), because it is a fairly basic decision tree and cannot model more complex behavior. RSF likely performed better because it is an aggregation of many survival trees and able to model complex behavior. Neural MTLR shows good promise, but could probably improve with a larger dataset to train the neural network.

For the semi-parametric models, we tried 3 different Cox PH model types. These models and their performance can be seen in Figure 4. A l1_ratio of 0 is a lasso penalty model. A l1_ratio of 1 is a ridge regression model. A l1_ratio of (0.1, 0.9) is an elastic net model. The best performing cox model was the ridge regression model, where the IBS was the lowest and the C-index was the highest.

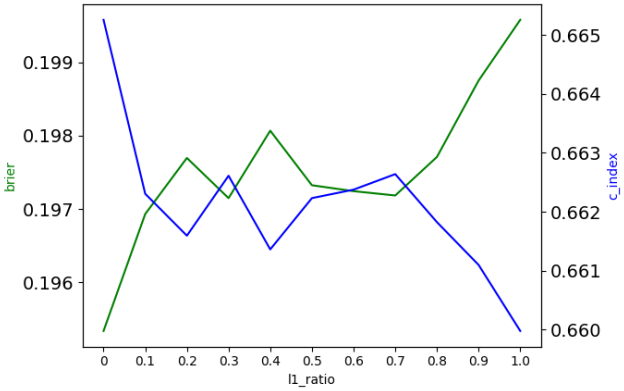
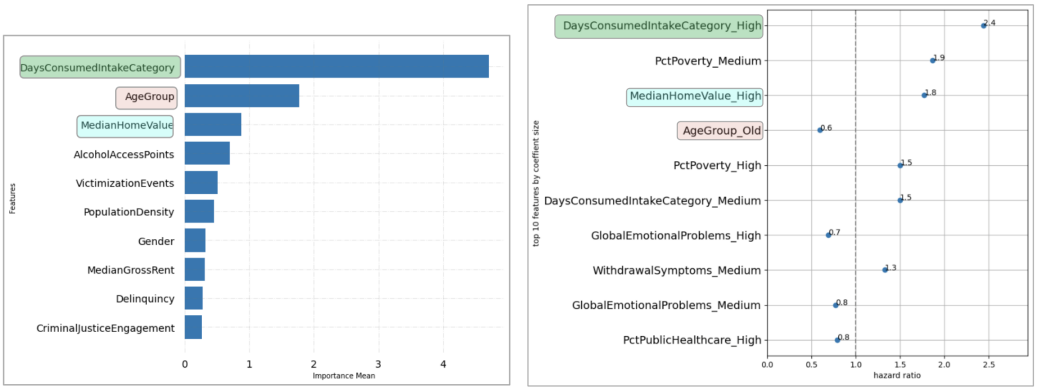


Figure 4: A comparison of the 11 Cox PH models. l1_ratio=0 (ridge regression) gives the best result.



(a) RSF Feature Importance

(b) Ridge Regression Hazard Ratios

Figure 5: Top 10 features selected by Random Survival Forests and top 10 values for hazard ratios from Ridge Regression Cox PH for the Overall Population.

Table 4 shows the two best performing models as RSF and Cox PH Ridge Regression model. The RSF model can give a rank order of feature importance, while the Cox model can measure the relative risk of having an attribute vs. not having an attribute.

We performed feature importance using the Permutation Feature Importance algorithm for non-parametric models. Since RSF is our top performing model, we compare the most important figures from the RSF model shown in Figure 5a. Here, we can see that the top features are the number of days that the person had alcohol before starting treatment (DaysConsumedIntakeCategory), their

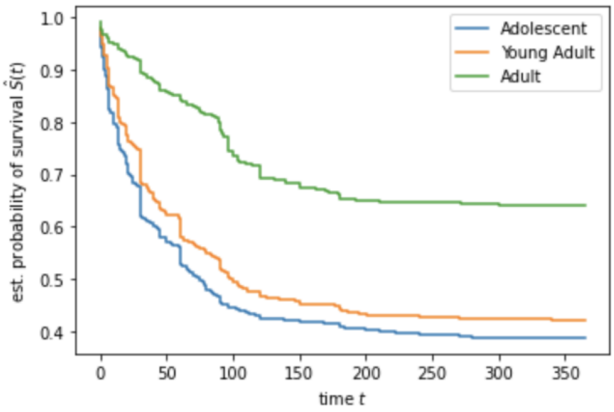


Figure 6: This figure shows the Kaplan Meier (KM) estimator survival curve for different age groups in our data. Adolescents and Young Adults relapse much faster than Adults (aged 30 years and up).

age group (AgeGroup) and median home value around the treatment centers the individuals went to (MedianHomeValue). Overall, we see a combination of personal factors like AgeGroup and VictimizationEvents as well as socio-ecocnomic factors like MedianHomeValue and PopulationDensity in the feature list from RSF.

The measurement of relative risk is done using a Hazard Ratio. Hazard Ratios for the best performing Cox model are shown in Figure 5b. A hazard ratio of 0.6 for AgeGroup_Old means a patient who is in AgeGroup_Old is 0.6x less likely to relapse than a patient in AgeGroup_Adolescent. A hazard ratio less than 1 is considered protective because is shows a decrease in risk vs. the control, while a hazard ratio greater than 1 is considered a risk factor. The largest risk factors identified by the Cox model for the entire population are high alcohol consumption prior to treatment (DaysCOnsumedIntakeCategory), high poverty (PctPoverty), and medium withdrawal symptoms (WithdrawalSymptoms).

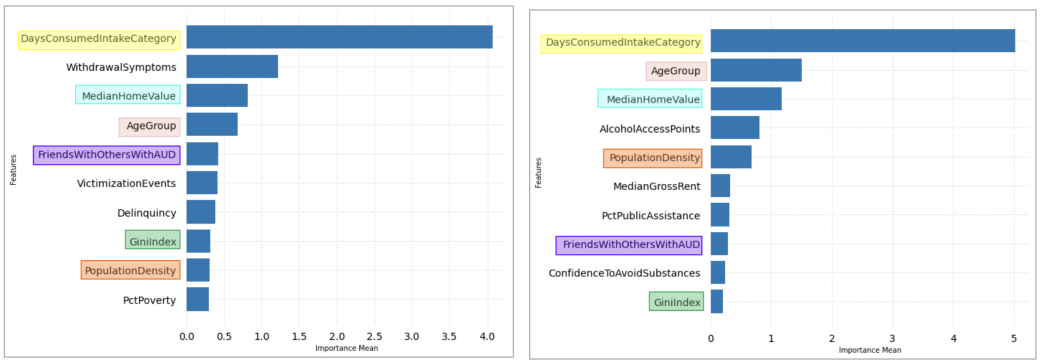
To understand and validate the results we get from these 2 models, we analyzed the Kaplan-Meier curve for some features of interest. Among them was the AgeGroup feature, which can be seen in Figure 6. The Cox model states that being an Adult (AgeGroup_Old) is a protective factor, which means Adults are at lower risk of relapsing. That makes sense when we compare against the Kaplan-Meier curve, which also states that Adults relapse slower than other age groups, since their curve falls the slowest.

5.2 Relapse causes over subgroups

5.2.1 Gender subgroup

Since RSF is the top performing non-parametric model for Gender subgroups as well as can be seen from Table 9 in the Appendix, we compare the most important figures from the RSF model shown in Figures 7a and 7b. The 2 figures highlight the common features between the subgroups. Comparing these two subgroups, we see common features across both genders like DaysConsumedIntakeCategory, AgeGroup and MedianHomeValue. This also lines up with the most important features for the overall population.

For females, we notice some new important features which do not show up in males, like whether they have any withdrawal symptoms (WithdrawalSymptoms) or delinquency (Delinquency). For males, features like the number of alcohol access points near them (AlcoholAccessPoints) and median gross rent near the treatment center (MedianGrossRent) are more important, which do not show up in females.

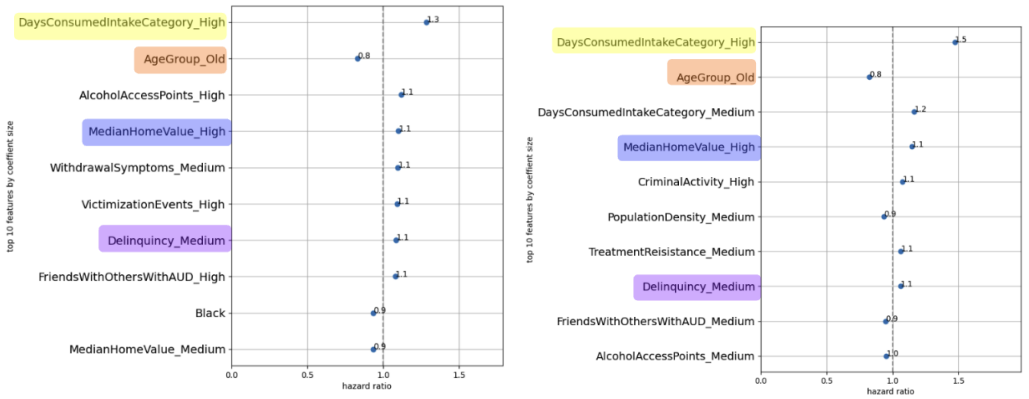


(a) RSF Feature Importance for Females

(b) RSF Feature Importance for Males

Figure 7: Top 10 features selected by Random Survival Forests for the population divided by Gender.

Another interesting finding is that the individual's involvement with criminal justice systems (CriminalJusticeInvolvement) does not show up in the top 10 most important features for either gender, even though it is important for the overall population, as can be seen in Figure 5a. This indicates that criminal justice involvement is a more global factor that influences recovery.



(a) Top 10 Cox Hazard Ratios for Females

(b) Top 10 Cox Hazard Ratios for Males

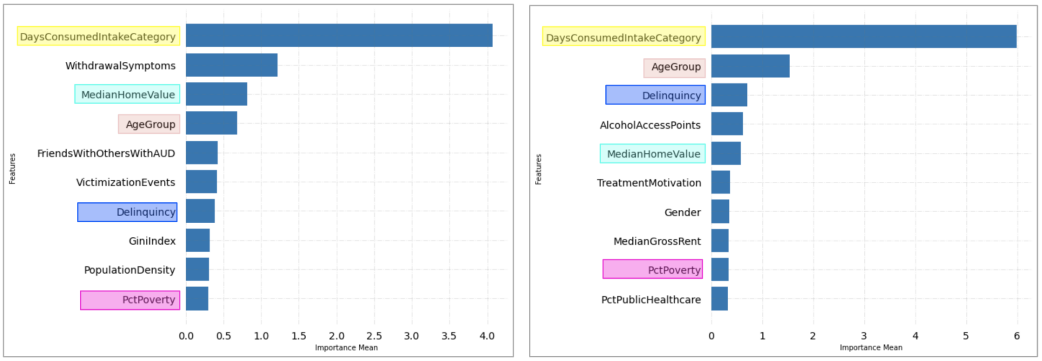
Figure 8: Top 10 hazard ratios selected by Cox model for the population divided by Gender.

The Cox model shows similar findings to the RSF model. Figures 8a and 8b show the largest hazard ratio for models trained on the female population vs. the male population. There are four features highlighted as being similar between the two genders. More interesting however, are the risk factors that differ between females and males. Females are at a higher risk of relapse based on personal features, while males are at a higher risk of relapse based on socio-economic or census tract features. Females show a higher risk when they experience withdrawal symptoms, and high victimization events. On the other hand, the features that add risk to males are census track features like criminal activity and population density.

In conclusion, for females, more personal factors play a greater role in their relapse and these factors may be more easily addressable in treatment plans. For males, socio-economic factors are more important. This can help domain experts tailor treatment plans according to the gender of the individual.

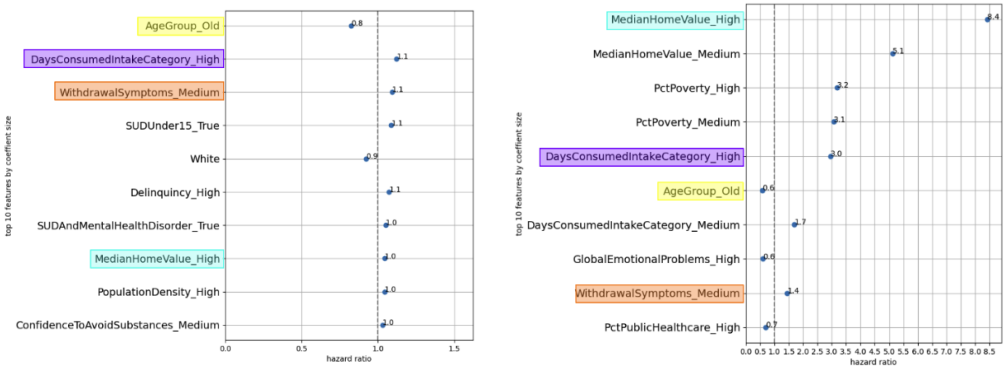
5.2.2 Post Traumatic Stress Disorder (PTSD) subgroup

For the PTSD feature, RSF is the top performing non-parametric model as can be seen from Table 9 in the Appendix, so we compare the most important figures from the RSF model shown in Figures 9a and 9b. Comparing these two subgroups, we see common features across both subgroups like DaysConsumedIntakeCategory, AgeGroup and MedianHomeValue. This also lines up with the most important features for the overall population as well as Gender. Additionally, Delinquency is an important feature across both subgroups. For people with PTSD, other important features are if they are friends with other people suffering from AUD (FriendsWithOthersWithAUD) or VictimizationEvents, which are not as important for people without PTSD.



(a) RSF Feature Importance for people with PTSD (b) RSF Feature Importance for people without PTSD

Figure 9: Top 10 features selected by Random Survival Forests for the population divided by PTSD.



(a) Hazard Ratios for people with PTSD (b) Hazard Ratios for people without PTSD

Figure 10: Top 10 hazard ratios selected by Cox model for the population divided by PTSD.

Figures 10a and 10b show the hazard ratios for patients who have PTSD vs. those who do not have PTSD. Again as in the female, male sub-population, there are four features that the two models agree on. However, the features that add risk to PTSD patients are largely personal factors, like substance abuse under the age of 15, and high delinquency. On the other hand, patients who present without PTSD have high risk factors that relate to socio-economic factors and census track. Interestingly, the highest hazard ratio of all model runs can be seen in Figure 10b. A person without PTSD is shown to be at a 8.4x higher risk when they have a high median home value, versus having a low medium home value. This is likely due to the Median home value being a highly unbalanced feature, having only 15 samples in the base case (low median income).

6 Conclusion

We evaluated 7 model types to determine if survival models could accurately predict the time to first drink for patients undergoing AUD treatment. Two models (RSF and Cox PH Ridge Regression) stood out showing promising IBS and C-index. These two models were then used to better understand the patient population risk factors to faster relapse.

Next, sub-population models were trained. We found that models trained on sub populations vs. the whole population performed similarly but with each subgroup having unique driving factors. These driving factors along with accurate relapse predictions indicate that subgroup models can and should be used to inform more targeted AUD treatment programs for patients within a designated subgroup.

6.1 Limitations

During the course of the research we have spotted two limitations. First, we have seen that there exists a set of features whose sample distribution across the different feature values is really unbalanced. An example of this is the Median Home Value (MHV) feature that has 2589 samples under the value 1 (Medium MHV), having only 15 samples under the value 0 (Low MHV). This is highly probable to be impacting the feature importance and relative risk analysis performed.

Secondly, the project aims to predict time to first drink (relapse) after entering treatment. However, absolute abstinence may not be sustainable where at least 53% of the population of this research relapses. Maybe solving other questions may be even more helpful to create better treatment plans.

6.2 Future Work

To address the limitation identified above, in the future we plan to remove highly unbalanced features like MHV. Secondly, to expand our work we think the same analysis could be done on all SUD (substance use disorder) instead of focusing the population only on AUD. Lastly, there are additional features that should be evaluated as potential homogeneous sub-populations. Figure 6 shows the Kaplan-Meier curves for Age. When comparing this Kaplan-Meier curve to the Gender and PTSD curves in Figure 3 and Figure 6, there is a much wider gap between Adolescents and Young Adults vs. Adults, than Female vs. Males or PTSD vs. No PTSD.

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] 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.

[5] 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.

[6] 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.

[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.

A Appendix: Data Dictionary

Table 5: 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	Gender	Female	Patient Gender Male/Female
	14	categorical	PopulationDensity	pop_deng	Population Density: Low, med, high
	15	categorical	PctUnemployed	pct_unemployedg	Percent unemployed in area: Low, med, high
	16	categorical	PctPublicAssistance	pct_public_assistanceg	Percent receiveing public asssistance: Low, med, high
	17	categorical	PctPoverty	pct_povertyg	Percent below poverty line: Low, med, high
	18	categorical	MedianFamilyIncome	median_family_incomeg	Median family income: Low, med, high
	19	categorical	MedianHomeValue	median_home_valueg	Median home value: Low, med, high
	20	categorical	MedianGrossRent	median_gross_rentg	Median gross monthly rent: Low, med, high
	21	categorical	GiniIndex	ginig	Census Gini Index: Low, med, high
	22	categorical	PctPublicHealthcare	pct_on_public_healthcareg	Percent on public healthcare: Low, med, high
environment	23	categorical	AlcoholAccessPoints	alc_access_pts_ctgry	Number of alcohol access points in area (bars, liquor stores, etc.): Low, med, high
mental health	24	categorical	TreatmentReisistance	TRIg_0_cd	Resistance to treatment: Low, med, high
	25	categorical	VictimizationEvents	gvsg_cd	Experienced victimization events: Low, med, high
	26	boolean	SUDAndMentalHealthDisorder	dldiag_cd	SUD and co-occuring mental health disorder: T/F
	27	categorical	Depression	dssg_0_cd	Depression symptoms: Low, med, high
	28	categorical	GlobalEmotionalProblems	epsg_0_cd	Global emotional problems (wars, unrest, etc.): Low, med, high
	29	categorical	ADHD	adhdg_0_cd	Likelyhood of ADHD: Low, med, high
	30	boolean	PTSD	tsd_0	Diagnosed with PTSD T/F
	31	categorical	Delinquincy	cdsg_0_cd	Delinquency or conduct disorder: Low, med, high
	32	boolean	Suicidal	suicprbs_0_cd	Reported problems of suicide: T/F
social factor	33	categorical	CriminalJusticeEngagement	cjsig_0_cd	Engagement with criminal justice system: Low, med, high
	34	categorical	LivingWithOthersWithAUD	lrig_0_cd	Living with others with AUD: Low, med, high
	35	categorical	FriendsWithOthersWithAUD	srig_0_cd	Friends with others with AUD: Low, med, high
	36	categorical	CriminalActivity	gcsg_0_cd	Involvement in criminal activity in last 90 days: Low, med, high
	37	categorical	NotCloseToSomeoneInRecoveryAUD	ncar_cd	Not close to someone in active recovery: T/F
substance use	38	boolean	SUDUnder15	und15_cd	Age of first substance use under 15: T/F
	39	categorical	WithdrawlSymptoms	CWSg_0_cd	Reported withdrawl symptoms: Low, med, high
	40	numeric	DaysConsumedIntake	S2a1_0	The # days of alcohol use @ intake
	41	numeric	DaysConsumedThreeMonths	S2a1_3	The # days of alcohol use @ month 3 survey
	42	numeric	DaysConsumedSixMonths	S2a1_6	The # days of alcohol use @ month 6 survey
	43	numeric	DaysConsumedTwelveMonths	S2a1_12	The # days of alcohol use @ month 12 survey
treatment	44	boolean	EngagmentLevel	engage30	Engagement in treatment for 30+ days and 3+ sessions: T/F
	45	boolean	PriorSUDTreatment	prsatx_cd	Prior substance abuse treatment: T/F
	46	categorical	TreatmentMotivation	TMIg_0_cd	Motivation for successful treatment: Low, med, high
	47	categorical	ConfidenceToAvoidSubstances	SESg_0_cd	Self confidence to avoid substance use during treatment: Low, med, high
target	48	boolean	CensorVariable	Alcohol_Cens	Censored/Uncensored
	49	categorical	CensorCategory	Alcohol_Cens_Ctgry	0: uncensored, 1: censored by end of study period, 2: lost to followup
	50	numeric	DaysToRelapseAUD	Alcohol_Days	Time in days until first alcohol consumption after beginning treatment

B Appendix: Descriptive Statistics

Table 6: Count of categorical variables with levels low, medium, high.

No.	Variable	LOW	MED	HIGH
1	pct_public_assistanceg	0	2727	563
2	ginig	0	1027	2263
3	CWSg_0_cd	2925	300	65
4	SESg_0_cd	2119	910	261
5	cdsg_0_cd	1987	1008	295
6	B2a_0g	1912	741	637
7	Alcohol_Cens_Ctgry	1759	313	1218
8	adhdg_0_cd	1692	849	749
9	gcsg_0_cd	1678	980	632
10	TRIg_0_cd	1475	1662	153
11	dssg_0_cd	1473	1150	667
12	epsg_0_cd	1284	1633	373
13	alc_access_pts_ctgry	1268	961	1061
14	gvsg_cd	1176	590	1524
15	cjsig_0_cd	1127	630	1533
16	TMIg_0_cd	520	2371	399
17	median_family_incomeg	478	2557	255
18	pct_on_public_healthcareg	470	2724	96
19	pct_unemployedg	396	2500	394
20	median_gross_rentg	378	2535	377
21	pop_deng	350	1269	1671
22	srig_0_cd	41	1198	2051
23	lrig_0_cd	34	1925	1331
24	pct_povertyg	31	2702	557
25	median_home_valueg	15	2589	686

Table 7: Count of boolean variables with levels T/F.

No.	Variable	Type	FALSE	TRUE
1	engage30	boolean	2255	1035
2	region_ne	boolean	3113	177
3	region_se	boolean	2611	679
4	region_mw	boolean	2849	441
5	region_sw	boolean	2508	782
6	region_w	boolean	2079	1211
7	unemplmt_cd	boolean	2447	843
8	prsatx_cd	boolean	1996	1294
9	und15_cd	boolean	1054	2236
10	dldiag_cd	boolean	602	2688
11	suicprbs_0_cd	boolean	2995	295
12	homeless_0_cd	boolean	2741	549
13	ncar_cd	boolean	895	2395
14	Raceg4_gr_1	boolean	2168	1122
15	Raceg4_gr_2	boolean	2921	369
16	Raceg4_gr_3	boolean	2091	1199
17	Raceg4_gr_4	boolean	2690	600
18	Alcohol_Cens	boolean	1759	1531
19	Female	boolean	2212	1078
20	tsd_0	boolean	2432	854

Table 8: Descriptive statistics for continuous variables.

No.	Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Null Count
1	Alcohol_Days	0	25	91	136.5	195	365	0
2	S2a1_0	0	0	3	10.61	12	90	6
3	S2a1_3	0	0	0	4.932	4	90	474
4	S2a1_6	0	0	0	4.944	4	90	640
5	S2a1_12	0	0	0	4.836	4	90	1715

C Appendix: Results

Table 9: Integrated Brier Scores for sub population model runs

	Integrated Brier Score				
Model	Female	Male	PTSD	No PTSD	Full Popluation
Kaplan-Meier Estimator	**0.234**	0.239	0.274	0.236	0.237
Survival Trees	0.363	0.329	0.337	0.338	**0.327**
Random Survival Forests	0.201	0.199	0.212	0.199	**0.195**
Neural MTLR	0.264	**0.214**	0.286	0.221	0.216
Best Performing Cox PH Model	0.205	0.206	0.223	0.204	**0.195**

D Appendix: 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.

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.

Regarding the clustering algorithms used , we started using the most common ones: K-Means and Agglomerative Clustering. However, the fact that these algorithm use distances (e.g. Euclidean distance) to compute the similarity between samples made us discard the results as we are dealing with categorical features in this problem. Instead we used another flavor of K-Means called K-Modes which will be explained in detail in the following section.

D.1 K-Modes clustering

K-Modes clustering uses dissimilarities between samples instead of distances to update centroids and cluster assignments. In this context, a dissimilarity is defined as a difference of value in a speccific feature between two different samples. Taking the example shown in Figure 11, bottom right sample is more similar to the top sample than bottom left sample as they have dissimilarity values of 2 and 5, respectively.

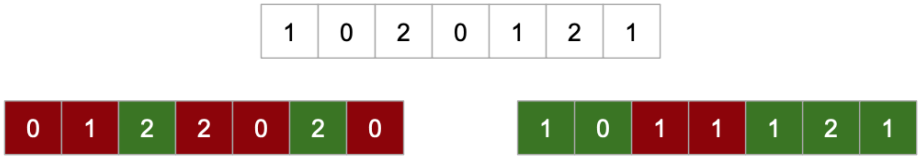
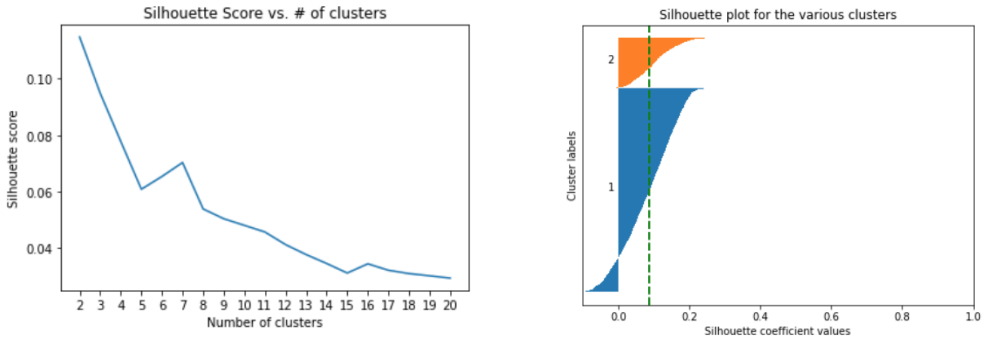


Figure 11: K-Modes sample comparison. Top sample is taken as reference. Bottom left sample has a dissimilarity value of 5. Bottom right sample has a dissimilarity of value of 2.

D.2 Clustering pipeline with K-Modes

To decide the best values for the number of clusters k hyperparameter, we run K-Modes 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 12a. From there, we select the best values for k which is



(a) Silhouette score versus number of clusters.

(b) Silhouette plot for $k = 2$

Figure 12: Silhouette related plots for K-Modes clustering.

$k = 2$. For this value, we obtain the silhouette plot shown in Figure 12b to check the clustering quality. In the plot we can see two things: the average silhouette score is extremelly low, and the silhouette score per sample is also low, meaning that the clusters are not good enough as the majority of the samples would be better placed in another cluster.

One of the causes that may be producing these results would be the number and/or importance of the features used. In total we used 44 features to perform clustering and possible not all of them are guiding the clustering algorithm to obtain meaningful clusters.

To see if that was the case, we used the clustering results to train a Random Forest Classifier receiving as inputs the 44-feature samples and predicting the assigned K-Modes cluster as output. After that, we performed Permutation Feature Importance to rank the 44 features from most to least importance in predicting the cluster assignment. From there we obtained the Top n most important features and rerun the K-Modes clustering using only the new set of features. A diagram of the described pipeline is shown in Figure 13.

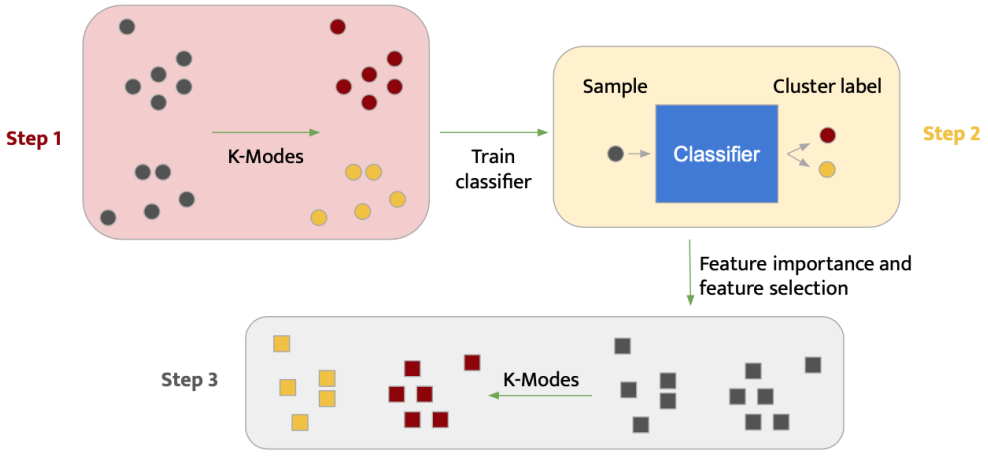
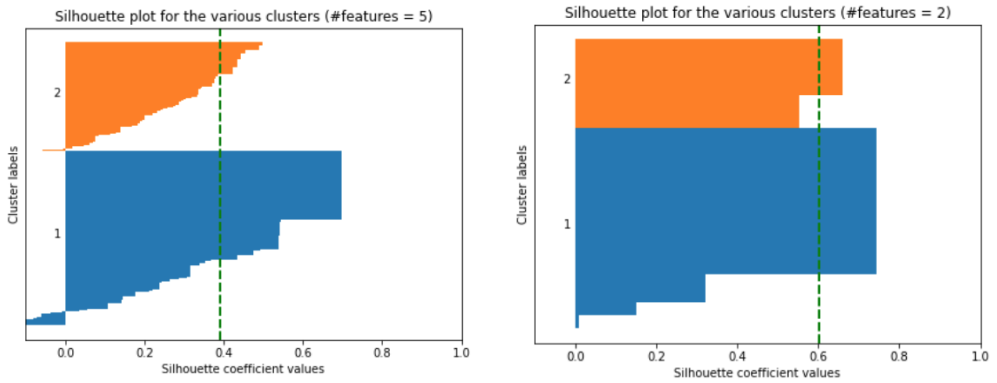


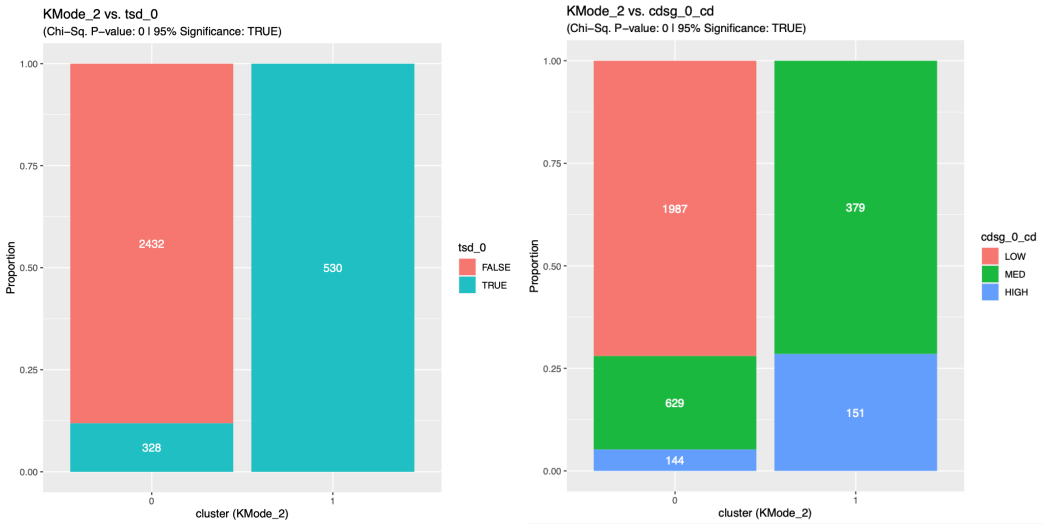
Figure 13: Pipeline diagram for K-Modes with feature reduction.

After rerunning the clustering algorithm with different number of features, we discovered that reducing the number of features, both, the silhouette score and the silhouette plot improved. We show in Figure 14 the silhouette plots for $k = 2$ using the Top-5 and Top-2 features.



(a) Silhouette plot for $k = 2$ using Top 5 features. (b) Silhouette plot for $k = 2$ using Top 2 features.

Figure 14: Silhouette plots for K-Modes clustering using 2 clusters and different number of features.



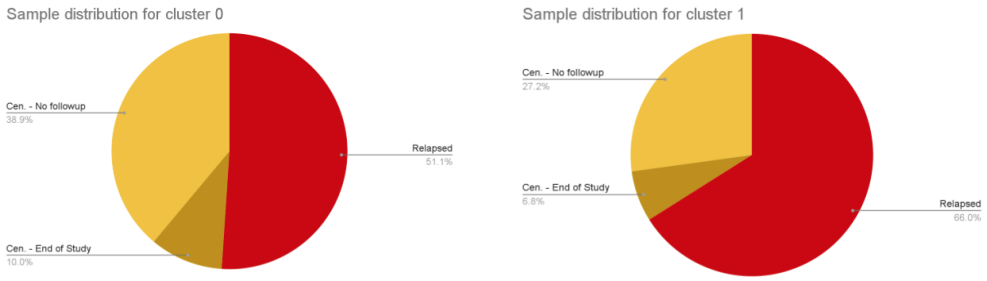
(a) PTSD (b) CDSG

Figure 15: Feature value sample distribution per cluster.

D.3 Results

After improving the clustering by increasing the silhouette score, we checked the quality and meaning of the clusters obtained. We performed all the checks with the best setup possible, 2 clusters ($k = 2$) obtained using the Top 2 features ($n = 2$). The selected features were Post Traumatic Stress Disorder (tsd_0) and Delinquincy/Conduct Disorder (cdsg_0_cd). The first check we performed is the distribution of samples per cluster for each one of the values of each variable. Distributions are shown in Figure 15. In the plots we can see that clusters are unbalanced in terms of number of samples. Cluster 1 has way less number of samples assigned than Cluster 0. In terms of feature value, Cluster 1 contains subjects with PTSD and Medium-High Delinquency indicators. Cluster 0, on the other hand, has majority of individuals without PTSD and Low Delinquency indicator. In that sense we would expect people in Cluster 1 to relapse before than people in Cluster 0, as the former has much more risk indicators than the latter.

Another check that we performed was in terms of distribution of samples per cluster with respect to their value in the censor variable. Figure 16 shows that distribution. Although there is a slight difference in terms of number of uncensored sample ratio in Cluster 1, this is not significant because



(a) Distribution in Cluster 0

(b) Distribution in Cluster 1

Figure 16: Sample distribution per cluster with respect to censor variable values.

Table 10: Alcohol Days Median and Average values of the uncensored samples in each cluster for K-Modes ($k = 2$).

Alcohol Days (Target Variable)	**Cluster 0**	**Cluster 1**	**All population**
Median	30	23	30
Average	46.21	38.83	42.62

that cluster contains only 530 samples versus the +2700 samples in Cluster 0. With these two results, we decided to not use the clusters obtained further in the experimental part.

However, one last check that we did and that may be interesting to the reader is the number of days until relapse of the uncensored variables in each cluster. Distribution is shown in Table 10. While Cluster 0 has equal median value as the whole population, Cluster 1 has a median value of 23, 7 days less than Cluster 0. This matches with the insight given above, where we pointed that, a priori, patients in Cluster 1 had more chances of relapsing before than patients in Cluster 0. In terms of average, it happens quite the same between Cluster 0 and 1.