PHP 2550: Project 2

Peirong Hao

Introduction

Since adults with major depressive disorder (MDD) often exhibit higher nicotine dependence, this project, in collaboration with Dr. George Papandonatos from Brown University's Department of Biostatistics, investigates smoking cessation in adults with MDD. The prior study used a 2x2 factorial design to assess the effectiveness of `Behavioral Activation for Smoking Cessation (BASC)` versus `Standard Behavioral Treatment (ST)`, alongside `Varenicline` versus `Placebo`. Results showed that `Varenicline` significantly improved cessation rates compared to `Placebo`, though no substantial differences were found between `BASC` and `ST` in increasing abstinence rates. Building on these findings, this study analyzes baseline variables to identify potential predictors of smoking abstinence, while adjusting for treatment effects.

Exploratory Data Analysis (EDA)

There were 300 observations with unique IDs (no duplicated observations) and 24 variables. Except for one variable serving as the ID, the others are variables of interest, including the outcome, treatments, and baseline covariates. Among all baseline characteristics, only `Antidepressant medication` has a p-value $< \alpha = 0.05$, indicating it is statistically significantly different across the four treatment groups (see Table 1).

According to the assumptions in the paper, the full sample was randomized at baseline and employed a missing-not-at-random (MNAR) assumption, meaning all missing outcomes were considered as smoking. There were no systematic differences in terms of missingness across the four treatment groups (see Table 2). The `ST + varenicline` group had the most missing data, with 18 observations, while the `BASC + varenicline` group had the least, with 12 missing observations.

As shown in Table 3, the top three variables with the highest missingness were `Nicotine Metabolism Ratio`, `Cigarettes reward value`, and `Readiness to quit`. Examining the pattern of missingness across multiple variables revealed that some variables tended to

Table 1: Participant characteristics by treatment group

Characteristic	**BASC + placebo** N = 68	**BASC + varenicline** N = 83	**ST + placebo** N = 68	**ST + varenicline** N = 81	**p-value**
Age	51 (14)	50 (13)	50 (11)	49 (13)	0.7
Sex					>0.9
Female	38 (56%)	44 (53%)	39 (57%)	44 (54%)	
Male	30 (44%)	39 (47%)	29 (43%)	37 (46%)	
Race					0.5
Black/African American	37 (54%)	37 (45%)	40 (59%)	43 (53%)	
Others	7 (10%)	12 (14%)	6 (8.8%)	13 (16%)	
White	24 (35%)	34 (41%)	22 (32%)	25 (31%)	
Hispanic					>0.9
False	63 (93%)	79 (95%)	64 (94%)	76 (94%)	
True	5 (7.4%)	4 (4.8%)	4 (5.9%)	5 (6.2%)	
Education					
College graduate	19 (28%)	29 (35%)	17 (25%)	26 (32%)	
Grade school	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)	
High school graduate or GED	23 (34%)	15 (18%)	11 (16%)	27 (33%)	
Some college/technical school	22 (32%)	32 (39%)	38 (56%)	24 (30%)	
Some high school	3 (4.4%)	7 (8.4%)	2 (2.9%)	4 (4.9%)	
Income per year					0.8
$20,000 - 35,000	16 (24%)	17 (21%)	14 (21%)	21 (26%)	
$35,001 - 50,000	8 (12%)	13 (16%)	14 (21%)	11 (14%)	
$50,001 - 75,000	12 (18%)	12 (15%)	8 (12%)	6 (7.5%)	
Less than $20,000	25 (37%)	30 (37%)	26 (38%)	29 (36%)	
More than $75,000	6 (9.0%)	10 (12%)	6 (8.8%)	13 (16%)	
Cigarettes smoked per day	16 (9)	16 (9)	15 (7)	14 (7)	>0.9
Cigarettes reward value	7 (4)	7 (4)	7 (4)	7 (3)	>0.9
FTCD	5 (2)	5 (2)	5 (2)	5 (2)	0.7
Readiness to quit					
3	1 (1.6%)	0 (0%)	0 (0%)	0 (0%)	
4	2 (3.1%)	2 (2.6%)	1 (1.6%)	0 (0%)	
5	6 (9.4%)	11 (14%)	9 (14%)	9 (12%)	
6	18 (28%)	22 (28%)	14 (22%)	29 (38%)	
7	16 (25%)	21 (27%)	16 (25%)	18 (23%)	
8	17 (27%)	20 (26%)	19 (30%)	18 (23%)	
9	2 (3.1%)	1 (1.3%)	2 (3.1%)	2 (2.6%)	
10	2 (3.1%)	1 (1.3%)	3 (4.7%)	1 (1.3%)	
Time to first cigarette after waking					0.5
5 minutes or less	32 (47%)	33 (40%)	35 (51%)	38 (47%)	
More than 5 minutes	36 (53%)	50 (60%)	33 (49%)	43 (53%)	
Cigarette type					0.9
Menthol cigarettes only	40 (59%)	48 (59%)	43 (64%)	47 (58%)	
Regular cigareettes (or both)	28 (41%)	34 (41%)	24 (36%)	34 (42%)	
Major depressive disorder status					0.7
Current MDD only/Current and past MDD	32 (47%)	40 (48%)	31 (46%)	44 (54%)	
Past MDD only	36 (53%)	43 (52%)	37 (54%)	37 (46%)	
Antidepressant medication					**0.013**
False	40 (59%)	59 (71%)	53 (78%)	66 (81%)	
True	28 (41%)	24 (29%)	15 (22%)	15 (19%)	
Other psychiatric diagnosis					0.2
False	33 (49%)	53 (64%)	40 (59%)	41 (51%)	
True	35 (51%)	30 (36%)	28 (41%)	40 (49%)	
Depressive symptoms (BDI-II)	19 (12)	18 (11)	18 (11)	20 (12)	>0.9
Pleasurable Events – substitute reinforcers	23 (20)	23 (19)	21 (20)	23 (19)	0.6
Pleasurable Events – complementary reinforcers	28 (22)	22 (17)	27 (20)	25 (19)	0.3
Anhedonia	2 (3)	2 (3)	3 (3)	2 (3)	0.8
Nicotine Metabolism Ratio	0.34 (0.18)	0.38 (0.25)	0.37 (0.27)	0.36 (0.21)	>0.9

[1] Mean (SD); n (%)

[2] Kruskal-Wallis rank sum test; Pearson's Chi-squared test; Fisher's exact test

Table 2: Missingness by treatment group

Group	Number of missing observations
BASC + placebo	15
BASC + varenicline	12
ST + placebo	14
ST + varenicline	18

Table 3: Single variable missingness table

variable	n_miss	pct_miss
Nicotine Metabolism Ratio	21	7
Cigarettes reward value	18	6
Readiness to quit	17	5.67
Income per year	3	1
Anhedonia	3	1
Cigarette type	2	0.667
FTCD	1	0.333

be missing together. Specifically, two observations had missing values in both `Cigarette reward value at baseline` and `Nicotine Metabolism Ratio`; one observation had missing values in both `Baseline readiness to quit smoking` and `Cigarette reward value at baseline`; two observations had missingness in both `Exclusive Mentholated Cigarette User` and `Baseline readiness to quit smoking`; and one observation had missing values in both `FTCD score at baseline` and `Anhedonia`.

To further determine the type of missingness, I created binary indicators (1 for missing, 0 for not missing) for each variable and generated a correlation plot for missingness (refer to Figure 2). Since some correlations had absolute values greater than 0.3, with a maximum of 0.58, I considered the missingness to plausibly be Missing at Random (MAR), meaning the probability of an observation being missing depends only on the observed variables.

I then checked the boxplots of all quantitative variables against the outcome. Most of the boxplots showed a similar range for both outcome groups. However, the range of `Anhedonia` appeared to differ significantly between the two outcome groups (see Figure 3), suggesting that `Anhedonia` could be an important predictor.

The next step was to create a correlation plot (Figure 4) for all pairs of quantitative variables. Overall, the correlations were not strong, with only three pairs showing an absolute correlation value greater than 0.5. The strongest correlation, with an absolute value of 0.625, was between the `FTCD score at baseline` and `Smoking within 5 minutes of waking up`. The correlation between `Current vs. Past MDD` and `BDI score` at baseline was 0.577, and the

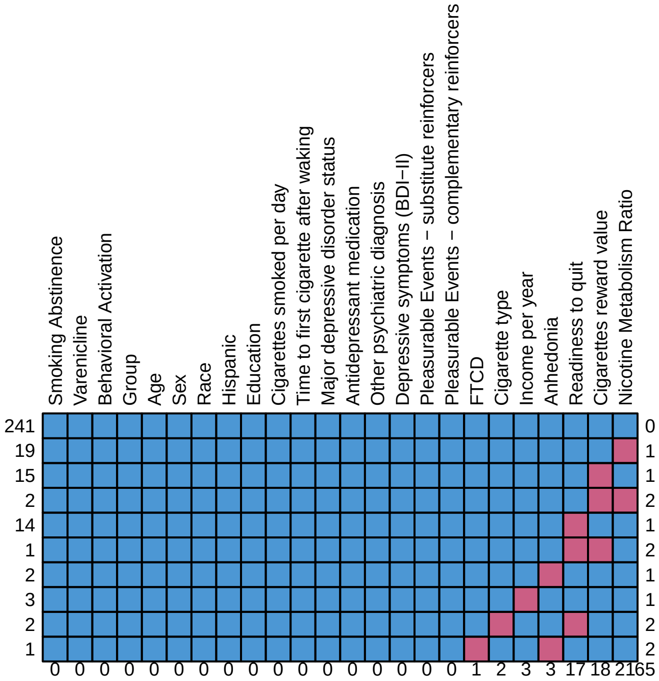

Figure 1: Multi-variable missingness pattern

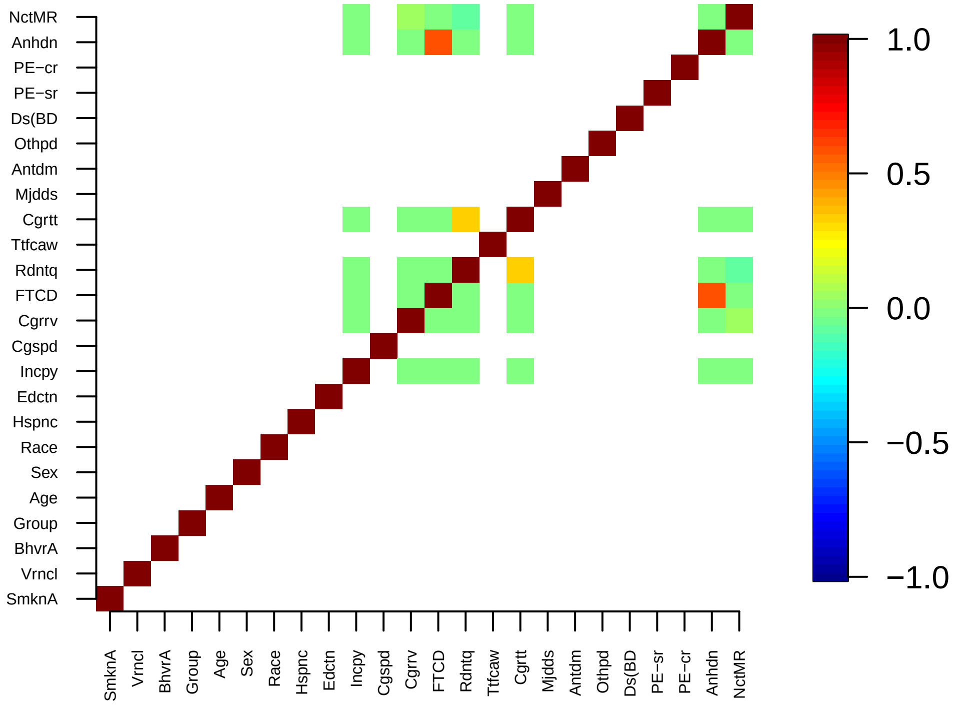

Figure 2: Missing Data Correlation Plot

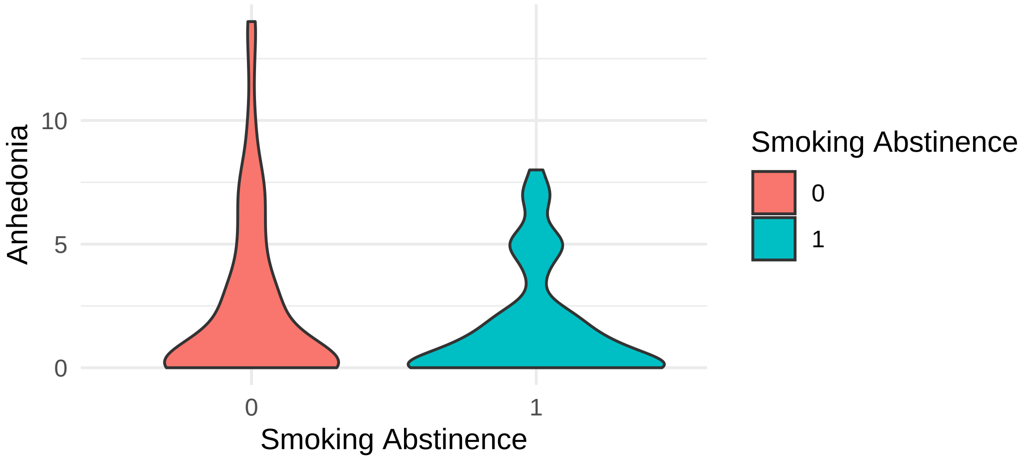

Figure 3: Violin plot of Anhedonia by Smoking Abstinence

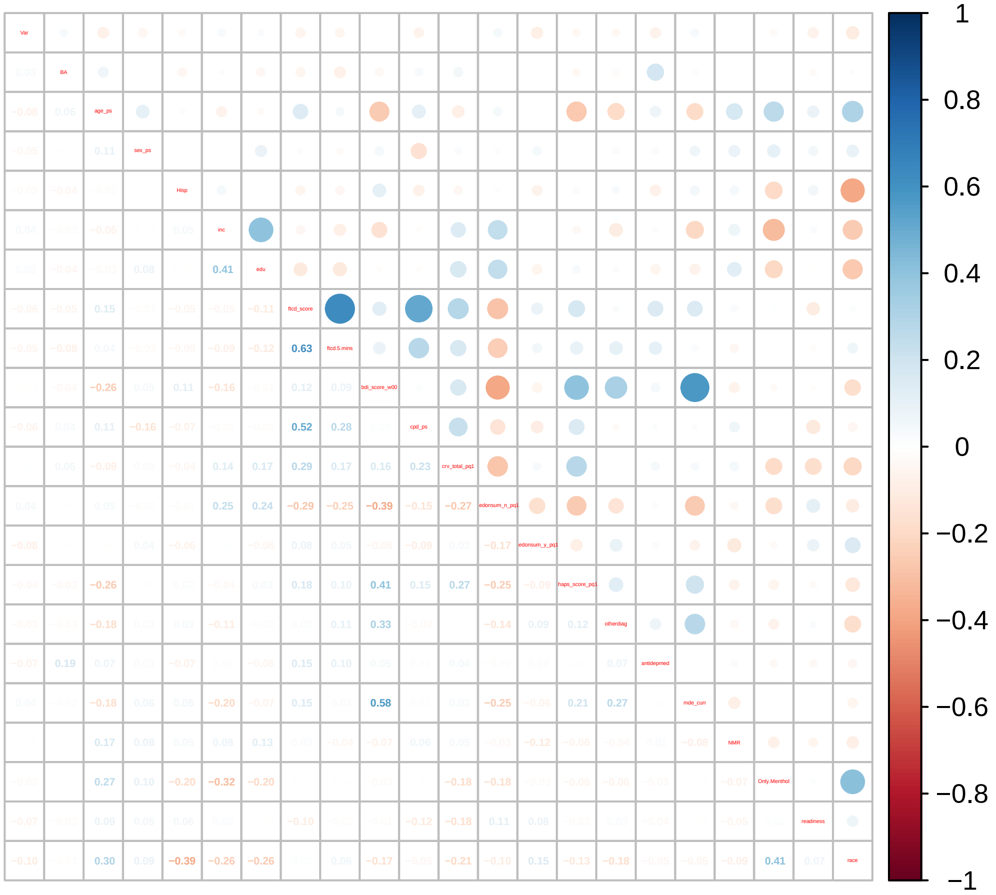

Figure 4: Multi-variable correlation

correlation between `FTCD score at baseline` and `Cigarettes per day at the baseline phone survey` was 0.517.

The data was unbalanced, with 79% of observations having an outcome value of 0 and 21% having an outcome value of 1. From Table 4, four variables were significantly different between groups: `Varenicline`, `Race`, `FTCD score at baseline`, and `Nicotine Metabolism Ratio`, each with p-values $< \alpha = 0.05$, indicating they could be important predictors. For each categorical variable (except the outcome), I created a two-way proportion table with the outcome. Notably, individuals with a `Baseline readiness to quit smoking` score of 3 or 4 did not achieve smoking abstinence (outcome = 0). Among those with a `readiness` score of 5, 31% achieved smoking abstinence (outcome = 1). For individuals with higher `readiness` scores, 14% to 25% achieved smoking abstinence.

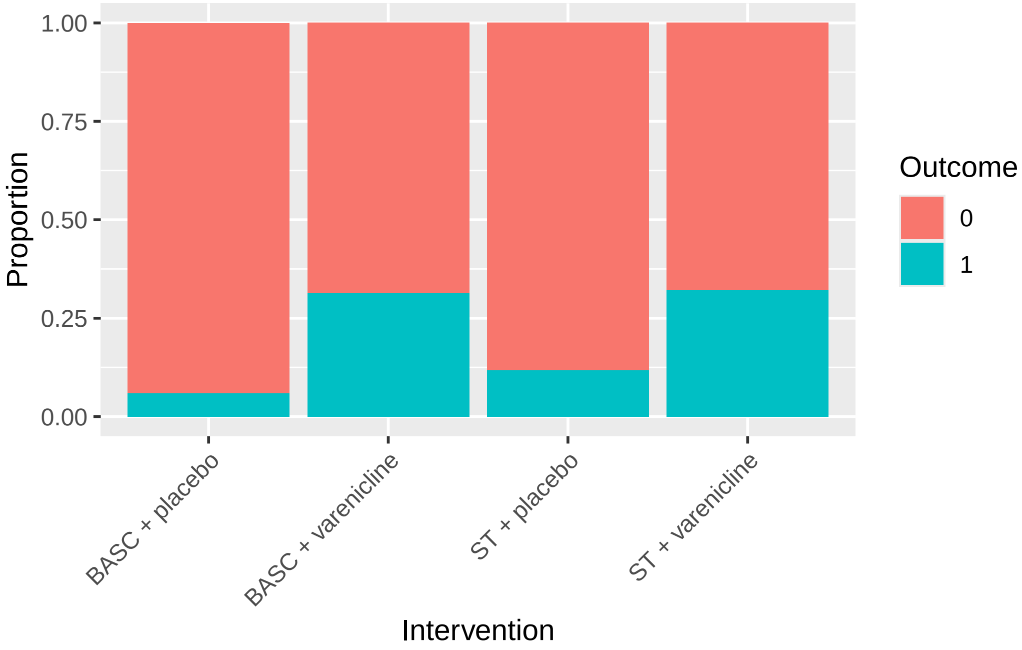

Figure 5: Stacked barplot of outcome by treatment group

From Figure 5, participants in the `Varenicline` group had a higher proportion of outcomes equal to 1 (Smoking Abstinence). Among the `Placebo` groups, `ST` appeared to be more effective than `BASC`, as `ST` was associated with a higher proportion of outcomes equal to 1.

Because `Antidepressant medication` is statistically significantly different between four treatment groups, and `Nicotine Metabolism Ratio(NMR)`, `FTCD`, `Race` are statistically significantly different between two outcome groups, I made Table 5 for them. White participants have higher `NMR` and `FTCD` compared to other races. However, there is not a clear trend of whether taking antidepressant medication affect `NMR` and `FTCD`.

Table 4: Participant characteristics by outcome group

	Smoking Abstinence		
Characteristic	**0** N = 236	**1** N = 64	**p-value**
Varenicline			<0.001
False	124 (53%)	12 (19%)	
True	112 (47%)	52 (81%)	
Behavioral Activation			0.5
False	115 (49%)	34 (53%)	
True	121 (51%)	30 (47%)	
Age	50 (13)	51 (13)	0.8
Sex			0.8
Female	129 (55%)	36 (56%)	
Male	107 (45%)	28 (44%)	
Race			0.032
Black/African American	129 (55%)	28 (44%)	
Others	33 (14%)	5 (7.8%)	
White	74 (31%)	31 (48%)	
Hispanic			0.8
False	221 (94%)	61 (95%)	
True	15 (6.4%)	3 (4.7%)	
Education			0.13
College graduate	66 (28%)	25 (39%)	
Grade school	0 (0%)	1 (1.6%)	
High school graduate or GED	60 (25%)	16 (25%)	
Some college/technical school	97 (41%)	19 (30%)	
Some high school	13 (5.5%)	3 (4.7%)	
Income per year			0.6
$20,000 - 35,000	56 (24%)	12 (19%)	
$35,001 - 50,000	36 (15%)	10 (16%)	
$50,001 - 75,000	30 (13%)	8 (13%)	
Less than $20,000	88 (38%)	22 (35%)	
More than $75,000	24 (10%)	11 (17%)	
Cigarettes smoked per day	16 (8)	14 (8)	0.052
Cigarettes reward value	7 (4)	7 (4)	>0.9
FTCD	5 (2)	4 (2)	0.002
Readiness to quit			0.6
3	1 (0.4%)	0 (0%)	
4	5 (2.2%)	0 (0%)	
5	24 (11%)	11 (19%)	
6	68 (30%)	15 (25%)	
7	53 (24%)	18 (31%)	
8	61 (27%)	13 (22%)	
9	6 (2.7%)	1 (1.7%)	
10	6 (2.7%)	1 (1.7%)	
Time to first cigarette after waking			0.2
5 minutes or less	113 (48%)	25 (39%)	
More than 5 minutes	123 (52%)	39 (61%)	
Cigarette type			0.4
Menthol cigarettes only	143 (61%)	35 (55%)	
Regular cigareettes (or both)	91 (39%)	29 (45%)	
Major depressive disorder status			0.073
Current MDD only/Current and past MDD	122 (52%)	25 (39%)	
Past MDD only	114 (48%)	39 (61%)	
Antidepressant medication			0.9
False	171 (72%)	47 (73%)	
True	65 (28%)	17 (27%)	
Other psychiatric diagnosis			0.2
False	127 (54%)	40 (63%)	
True	109 (46%)	24 (38%)	
Depressive symptoms (BDI-II)	19 (12)	17 (11)	0.2
Pleasurable Events – substitute reinforcers	22 (19)	25 (22)	0.4
Pleasurable Events – complementary reinforcers	26 (19)	23 (20)	0.15
Anhedonia	2 (3)	2 (2)	0.3
Nicotine Metabolism Ratio	0.35 (0.22)	0.42 (0.26)	0.023

[1] n (%); Mean (SD)

[2] Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's exact test

Table 5: Summary statistics for four variables

Antidepressant medication	NMR_mean	NMR_sd	FTCD_mean	FTCD_sd
Black/African American				
Not take med	0.33	0.22	4.87	2.11
Take med	0.33	0.15	6.32	1.49
Others				
Not take med	0.33	0.24	4.93	1.74
Take med	0.23	0.11	4.62	2.72
White				
Not take med	0.42	0.20	5.35	2.34
Take med	0.47	0.32	5.38	2.27

Statistical Modeling

As discussed previously, the missingness in this case is likely Missing at Random (MAR). Since MICE is designed for MAR data — imputing missing values based on observed data in other variables — I used MICE to impute five datasets before splitting them into testing and training sets. It is preferable to perform multiple imputation before splitting, as this ensures imputations are consistent across the entire dataset, leading to comparable results in both training and test sets. For each imputed dataset, there are four treatment groups, and within each group, I applied an 80%-20% train-test split. In other words, for each imputed dataset, I obtained a final training and testing set by combining the training and testing subsets from each treatment group.

In addition to main effects, I included several interaction terms for modeling. One interaction was between `Varenicline` and `Behavioral Activation`, while the others were all possible interactions between a quantitative predictor and a categorical predictor (except the outcome and two treatment variables). I selected the L0+L1 regularization model from the L0Learn library and used 5-fold cross-validation to identify the optimal gamma and lambda values. Each imputed dataset produced one set of estimated coefficients, and the final pooled estimates were obtained by averaging the results across the five imputed datasets.

The ROC curves in Figure 6 did not approach the upper left corner, indicating no points with both high sensitivity and high specificity. With an area under the curve (AUC) of approximately 0.7, the model demonstrated limited predictive capability, suggesting it may not be highly effective. Surprisingly, the model fit the derivation data worse than the validation data, with a slightly lower AUC for the derivation data (0.740) compared to the validation data (0.753).

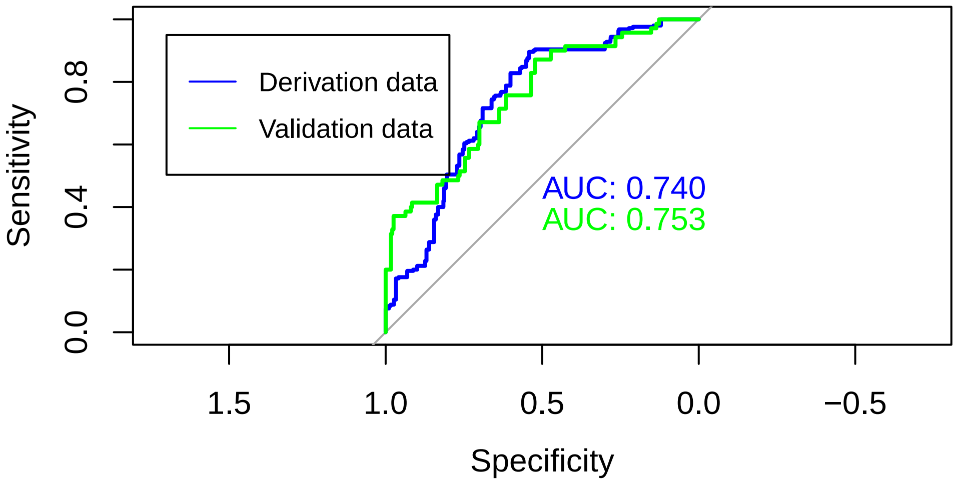

Figure 6: Discrimination plot

Table 6: Discrimination table

Measures	Derivation	Validation
Sens	0.896	0.871
Spec	0.542	0.523
PPV	0.341	0.351
NPV	0.952	0.932
Acc	0.616	0.603

Table 6 shows an accuracy of approximately 60% (61.6% for the derivation set and 60.3% for the validation set), indicating that the model correctly predicts outcomes 60% of the time, which is not particularly strong. For the derivation data, sensitivity is greater than specificity, suggesting that the model performs well in identifying individuals with smoking abstinence but has lower specificity, reflecting weaker performance in correctly predicting outcomes for those without smoking abstinence. The validation set shows a similar pattern, with sensitivity higher than specificity. The Positive Predictive Value (PPV) indicates that 34.1% of those with a positive test in the derivation set actually reach smoking abstinence, slightly lower than the 35.1% in the validation set. The Negative Predictive Value (NPV) shows that 95.2% of those with a negative test in the derivation set do not achieve smoking abstinence, which is slightly higher than the 93.2% observed in the validation set.

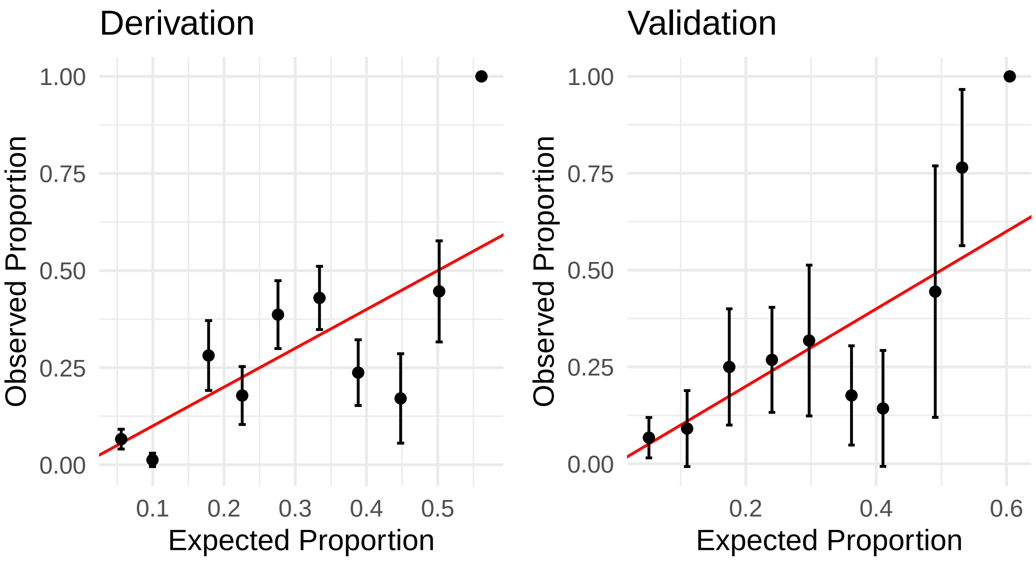

Figure 7: Calibration plot

Figure 7 compares observed versus expected proportions, with the red line representing a perfect fit where the estimated and true distributions align. In the derivation set's calibration plot, the estimated and true distributions do not align well, as 6 out of 9 confidence intervals fail to capture the red line. The calibration in the validation set looks better, with most 95% confidence intervals capturing the red line. However, the validation plot's confidence intervals are much wider, indicating substantial uncertainty. Notably, a single dot appears in the upper right corner of both calibration plots, which may be due to the limited proportions of observations with smoking abstinence — only 21% in the derivation set and 23% in the validation set.

Table 7: Coefficients of the Logistic Model

	Coefficients
Behavioral Activation	1.384
Age * Education	0.066
Age * Current vs past MDD	0.253
BDI score * Readiness to quit smoking	-0.024

Results

Interestingly, the "best" model does not include `Varenicline` (which was considered important in the previous study) but instead includes `Behavioral Activation` (which was considered unimportant previously), `Age` and its interactions with both `Education` and `Current vs Past`

MDD, along with the interaction between BDI score and Readiness to Quit Smoking. This model is relatively simple, likely due to the effects of L1 (lasso) regularization, which penalizes the absolute sum of the coefficients, reducing complexity by setting less influential variable coefficients to zero. Additionally, incorporating L0 regularization, which minimizes the count of non-zero coefficients, further helps prevent overfitting by selecting an optimal subset of features.

The interpretation of the coefficients is as follows: For a participant who received Behavioral Activation, the odds of achieving Smoking Abstinence are $e^{1.384} = 3.99$ times higher than for a participant who received Standard Behavioral Treatment, holding all other variables constant. A positive coefficient of 0.066 indicates a slight positive interaction effect between Age and Education on Smoking Abstinence, meaning that the combined effect of Age and Education marginally increases the odds of Smoking Abstinence.The effect size ($e^{0.066} = 1.07$) is relatively modest.

Similarly, the positive coefficient of 0.253 suggests an interaction between Age and Having Current vs. Past Major Depressive Disorder (MDD) that affects Smoking Abstinence. This implies that as Age increases, individuals with current MDD have approximately 1.29 times ($e^{0.253} = 1.29$) higher odds of achieving Smoking Abstinence than those with past MDD. A negative coefficient of -0.024 indicates an interaction effect between the BDI score (a measure of depression severity) and Readiness to Quit Smoking on Smoking Abstinence likelihood. This suggests that for individuals with higher BDI scores, even with readiness to quit, their likelihood of achieving Smoking Abstinence might be slightly lower. Although this effect size is small ($e^{-0.024} = 0.98$), it highlights a potential challenge for individuals with more severe depression symptoms.

I also explored various transformations — squared, square root, and log — for quantitative variables including Age, BDI score, Cigarettes per day, Pleasurable Events Scale (both substitute and complementary reinforcers), Anhedonia, and Nicotine Metabolism Ratio, focusing on those with non-normal histogram distributions. The selected model was updated to include Behavioral Activation, Exclusive Mentholated Cigarette User, an interaction between Pleasurable Events Scale (complementary reinforcers) and Current vs. Past MDD, and an interaction between BDI score and Readiness to Quit Smoking. Despite these adjustments, the coefficients remained small, the AUC increased only slightly to 0.76, 95% confidence intervals in the calibration plot moved a little closer toward the red line, indicating minimal improvement. Since Behavioral Activation, Current vs. Past MDD, and the interaction between BDI score and Readiness to Quit Smoking were selected in both models (with or without variable transformation), they appear to be essential predictors of the outcome, Smoking Abstinence.

Since the absolute value of the mean of pearson residuals is less than 2 for both outcome groups, the model appears to fit well. Comparing the means, the model seems to overestimate the probability of smoking abstinence for non-abstinent individuals (indicated by a negative mean) and underestimate the likelihood of abstinence for abstinent individuals (positive mean).

Table 8: Pearson Residuals Summary Table

Smoking Abstinence	Mean	SD
Abstinent Group	1.71	0.60
Non-abstinent Group	-0.47	0.19

Discussion

This study focuses on individuals with mental health disorders, specifically MDD, and examines the effectiveness of different treatments, psychological and demographic factors that influence smoking cessation. The model, constructed using L1 and L0 regularization techniques, is simplified by reducing complexity and selecting only the most predictive features. Key predictors and interaction effects influencing smoking abstinence are identified. However, certain limitations should be noted: the model's simplicity may exclude complex relationships, and missing data could introduce potential bias. Despite these limitations, the findings of this study highlight the value of `Behavioral Activation` and the intricate role of psychological factors such as `Current vs. Past MDD`, `BDI score`, and `Readiness to Quit Smoking` in influencing smoking cessation outcomes. This information is vital for future research and policy decisions in the field of mental health and smoking cessation.

Code Appendix

```r
set.seed(1)
library(tidyverse)
library(knitr)
library(tidyr)
library(dplyr)
library(kableExtra)
library(readr) #read csv
library(visdat) #missing data pattern
library(gridExtra)
library(corrplot)

library(fields)
library(glue)
library(mice)
library(gtsummary)
library(naniar)
library(gt)
library(L0Learn)
library(pROC)

#read in data file
df.proj2<-read.csv("project2.csv")
#length(unique(df.proj2$id))#no duplicate obs
df.proj2<-df.proj2[,-1]#remove id column
#str(df.proj2)#int, num, convert some to factors
#rename variables and their categories
df.proj2.revise <- df.proj2 %>% mutate(
  `Smoking Abstinence` = abst,
  `Behavioral Activation` = case_when(
      BA == 1 ~ "True",
      BA == 0 ~ "False"),
  Varenicline = case_when(
      Var == 1 ~ "True",
      Var == 0 ~ "False"),
  Group = case_when(
          Var == 0 & BA == 0 ~ "ST + placebo",
          Var == 0 & BA == 1 ~ "BASC + placebo",
          Var == 1 & BA == 0 ~ "ST + varenicline",
          Var == 1 & BA == 1 ~ "BASC + varenicline"),
  Age = age_ps,
  Sex = case_when(
          sex_ps == 2 ~ "Female",
          sex_ps == 1 ~ "Male"),
  Race = case_when(
          NHW == 1 & Black == 0 ~ "White",
          NHW == 0 & Black == 1 ~ "Black/African American",
          .default = "Others"),
  Hispanic = case_when(
          Hisp == 1 ~ "True",
          Hisp == 0 ~ "False"),
  Education = case_when(
          edu == 1 ~ "Grade school",
          edu == 2 ~ "Some high school",
          edu == 3 ~ "High school graduate or GED",
          edu == 4 ~ "Some college/technical school",
          edu == 5 ~ "College graduate"),
  `Income per year` = case_when(
          inc == 1 ~ "Less than $20,000",
          inc == 2 ~ "$20,000 - 35,000",
          inc == 3 ~ "$35,001 - 50,000",
          inc == 4 ~ "$50,001 - 75,000",
          inc == 5 ~ "More than $75,000"),
  `Cigarettes smoked per day` = cpd_ps,
  `Cigarettes reward value` = crv_total_pq1,
  FTCD = ftcd_score,
  `Readiness to quit` = readiness,
  `Time to first cigarette after waking` = case_when(
          ftcd.5.mins == 1 ~ "5 minutes or less",
          ftcd.5.mins == 0 ~ "More than 5 minutes"),
  `Cigarette type` = case_when(
          Only.Menthol == 1 ~ "Menthol cigarettes only",
          Only.Menthol == 0 ~ "Regular cigareettes (or both)"),
  `Major depressive disorder status` = case_when(
          mde_curr == 1 ~ "Current MDD only/Current and past MDD",
          mde_curr == 0 ~ "Past MDD only"),
  `Antidepressant medication` = case_when(
          antidepmed == 1 ~ "True",
          antidepmed == 0 ~ "False"),
  `Other psychiatric diagnosis` = case_when(
          otherdiag == 1 ~ "True",
          otherdiag == 0 ~ "False"),
  `Depressive symptoms (BDI-II)` = bdi_score_w00,
  `Pleasurable Events - substitute reinforcers` = hedonsum_n_pq1,
  `Pleasurable Events - complementary reinforcers` = hedonsum_y_pq1,
  Anhedonia = shaps_score_pq1,
  `Nicotine Metabolism Ratio` = NMR,
  across(c(`Smoking Abstinence`, `Behavioral Activation`, Varenicline,
   ↪  Group, Sex, Race,
           Hispanic, `Income per year`, Education, `Time to first
 ↪  cigarette after waking`,
          `Other psychiatric diagnosis`, `Antidepressant medication`,
          `Major depressive disorder status`, `Cigarette
           ↪  type`,`Readiness to quit`), as.factor)) %>%
  dplyr::select(`Smoking Abstinence`, Varenicline, `Behavioral
 ↪  Activation`, Group,
                Age, Sex, Race, Hispanic, Education, `Income per year`,
 ↪  `Cigarettes smoked per day`,
                `Cigarettes reward value`, FTCD, `Readiness to quit`,
                 ↪  `Time to first cigarette after waking`,
                `Cigarette type`, `Major depressive disorder status`,
                 ↪  `Antidepressant medication`,
                `Other psychiatric diagnosis`, `Depressive symptoms
                 ↪  (BDI-II)`,
                `Pleasurable Events - substitute reinforcers`,
                `Pleasurable Events - complementary reinforcers`,
                Anhedonia, `Nicotine Metabolism Ratio`)

#summary table
tbl.summary<-df.proj2.revise %>%
  dplyr::select(-c(`Smoking Abstinence`, Varenicline, `Behavioral
 ↪  Activation`))%>%
  tbl_summary(by = Group,
              type = list(where(is.numeric) ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({sd})"),
              missing="no") %>% add_p() %>%
  bold_labels() %>%
  #italicize_levels() %>%
  bold_p(t = 0.05) %>%
  as_kable_extra(format = "latex", booktabs = TRUE) %>%
  kable_styling(latex_options = "scale_down")
```

```
tbl.summary
#modify race variable
df.proj2 <- df.proj2 %>% mutate(
  race = case_when(
          NHW == 1 & Black == 0 ~ 1,
          NHW == 0 & Black == 1 ~ 2,
          .default = 0))%>%
  select(-c(NHW, Black))

# Visualize missing data pattern

#df.proj2%>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")

# Calculate the number of missing values for each column
missing_counts <- colSums(is.na(df.proj2.revise))
miss_tbl<-df.proj2.revise %>% filter(if_any(everything(), is.na))%>%
  group_by(Group)%>%
  summarise(n = n(),.groups = 'drop')%>%
  rename(`Number of missing observations` = n)

kable(miss_tbl, format = "latex", booktabs = TRUE, digits = 2)%>%
  kable_styling(latex_options = "hold_position")
missingness_table <- miss_var_summary(df.proj2.revise)
missingness_table <- missingness_table[c(1:7),]

kable(missingness_table, format = "latex", booktabs = TRUE, digits =
 ↪  2)%>%
  kable_styling(latex_options = "hold_position")
par(cex = 0.6)
md.plot<-md.pattern(df.proj2.revise, plot = TRUE, rotate.names = TRUE)
#check MAR

# Create a missingness indicator matrix and correlation matrix
missing_data <- as.data.frame(is.na(df.proj2.revise) * 1)
cor_matrix <- cor(missing_data)

# Use short labels on the axes
short_labels <- abbreviate(colnames(cor_matrix), minlength = 5)

# Plot and add a legend
```

```
image.plot(cor_matrix, zlim = c(-1, 1), axes = FALSE)
axis(1, at = seq(0, 1, length = ncol(cor_matrix)), labels =
 ↪  short_labels, las = 2, cex.axis = 0.5)
axis(2, at = seq(0, 1, length = nrow(cor_matrix)), labels =
 ↪  short_labels, las = 1, cex.axis = 0.5)
#single variable eda

#might not need to do any variable transformation

#quantitative variables
par(mfrow=c(1,2))
hist((df.proj2$age_ps))#left skew
hist((df.proj2$age_ps)^2)#left skew

hist((df.proj2$ftcd_score))#slight left skew

par(mfrow=c(1,2))
hist((df.proj2$bdi_score_w00))#right skew
hist(sqrt(df.proj2$bdi_score_w00))#right skew

par(mfrow=c(1,2))
hist((df.proj2$cpd_ps))#right skew
hist(log(df.proj2$cpd_ps))#right skew

hist((df.proj2$crv_total_pq1))#slight right skew

par(mfrow=c(1,2))
hist((df.proj2$hedonsum_n_pq1))#right skew
hist(sqrt(df.proj2$hedonsum_n_pq1))#right skew

par(mfrow=c(1,2))
hist((df.proj2$hedonsum_y_pq1))#right skew
hist(sqrt(df.proj2$hedonsum_y_pq1))#right skew

par(mfrow=c(1,2))
hist((df.proj2$shaps_score_pq1))
hist(log(df.proj2$shaps_score_pq1))#right skew

par(mfrow=c(1,2))
hist((df.proj2$NMR))
hist(log(df.proj2$NMR))#right skew->slight left skew
```

```
#multivariable eda, C+Q

par(mfrow=c(1,2))
boxplot(df.proj2$age_ps~df.proj2$abst)
boxplot(df.proj2$ftcd_score~df.proj2$abst)

par(mfrow=c(1,2))
boxplot(df.proj2$bdi_score_w00~df.proj2$abst)
boxplot(df.proj2$cpd_ps~df.proj2$abst)

par(mfrow=c(1,2))
boxplot(df.proj2$crv_total_pq1~df.proj2$abst)
boxplot(df.proj2$hedonsum_n_pq1~df.proj2$abst)

par(mfrow=c(1,2))
boxplot(df.proj2$hedonsum_y_pq1~df.proj2$abst)
boxplot(df.proj2$shaps_score_pq1~df.proj2$abst) #other ranges of
 ↪  boxplots are similar

boxplot(df.proj2$NMR~df.proj2$abst)
ggplot(df.proj2.revise, aes(x = `Smoking Abstinence`, y = Anhedonia,
 ↪  fill = `Smoking Abstinence`)) +
  geom_violin() +
  theme_minimal()
non_factor_df <- df.proj2[, sapply(df.proj2, function(col)
 ↪  !is.factor(col))]
non_factor_df <- non_factor_df[,-1]#remove id column
corrplot.mixed(cor(non_factor_df, use = "complete.obs"), tl.cex = 0.2,
 ↪  number.cex = 0.3)
#cor(non_factor_df, use = "complete.obs")

#ftcd_score and ftcd.5.mins have cor=0.6254919659
#FTCD score at baseline, Smoking with 5 mins of waking up

#mde_curr and bdi_score_w00 have cor=0.57701349
#Current vs past MDD, BDI score at baseline

#ftcd_score and cpd_ps have cor=0.516755685
#FTCD score at baseline, Cigarettes per day at baseline phone survey
df.proj2.revise %>%
  dplyr::select(-c(Group))%>%
  tbl_summary(by = `Smoking Abstinence`,
              type = list(where(is.numeric) ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({sd})"),
              missing="no") %>%
  bold_labels() %>%
  #italicize_levels() %>%
  add_p()%>%
  modify_spanning_header(all_stat_cols() ~ "**Smoking Abstinence**") %>%
  as_kable_extra(format = "latex", booktabs = TRUE) %>%
  kable_styling(latex_options = "scale_down", font_size = 7)
#categorical variables
prop.table(table(df.proj2$abst))#outcome of interest: 79% of 0s, 21% of
 ↪  1s, so unbalanced data

tab<-prop.table(table(df.proj2$Var, df.proj2$abst),1)
barplot(t(tab))

prop.table(table(df.proj2$BA, df.proj2$abst),1)

prop.table(table(df.proj2$sex_ps, df.proj2$abst),1)

prop.table(table(df.proj2$race, df.proj2$abst),1)

prop.table(table(df.proj2$Hisp, df.proj2$abst),1)

prop.table(table(df.proj2$inc, df.proj2$abst),1)

prop.table(table(df.proj2$edu, df.proj2$abst),1)

prop.table(table(df.proj2$ftcd.5.mins, df.proj2$abst),1)

prop.table(table(df.proj2$otherdiag, df.proj2$abst),1)

prop.table(table(df.proj2$antidepmed))#73% no, 27% yes
prop.tab<-prop.table(table(df.proj2$antidepmed, df.proj2$abst),1)
barplot(t(prop.tab),xlab="Antidepressant Medication",main="Stacked
 ↪  barplot of smoking abstinence",legend.text = c("Not taking
 ↪  it","Taking medication"),col=c("lightskyblue1","lightblue3"))

prop.table(table(df.proj2$mde_curr, df.proj2$abst),1)
prop.table(table(df.proj2$Only.Menthol, df.proj2$abst),1)

prop.table(table(df.proj2$readiness, df.proj2$abst),1)#explain this

prop.table(table(df.proj2.revise$Group, df.proj2.revise$`Smoking
 ↪  Abstinence`),1)
prop_table<-prop.table(table(df.proj2.revise$Group,
 ↪  df.proj2.revise$`Smoking Abstinence`),1)
prop_df <- as.data.frame(prop_table)
colnames(prop_df) <- c("Intervention", "Outcome", "Proportion")
ggplot(prop_df, aes(x = Intervention, y = Proportion, fill = Outcome)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(y = "Proportion", x = "Intervention", fill = "Outcome") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#sig different between four trt groups:
#Antidepressant medication: C

#sig different between two outcome groups:
#Nicotine Metabolism Ratio(NMR): Q

#FTCD: Q

#Race: C

#Varenicline: C


ggplot(df.proj2, aes(x=NMR, y=ftcd_score, fill=factor(race))) +
  geom_violin()

df.proj2.revise %>% ggplot(aes(y=`Nicotine Metabolism Ratio`, x=FTCD,
 ↪  color=`Antidepressant medication`)) +
  geom_smooth()+
  theme_minimal()+
  facet_grid(.~Race)#+
  #theme(legend.position="none")+
 # labs(x = "Wind speed in km/hr", y = "Completion Time in hours")


ggplot(df.proj2.revise,
       aes(x = Race, y = `Nicotine Metabolism Ratio`, fill = Race)) +
  geom_boxplot() #+
  #labs(x = "Group", y = "Value") +
  #theme_minimal()

ggplot(df.proj2.revise,
       aes(x = Race, y = `Nicotine Metabolism Ratio`,fill = Race)) +
  geom_violin() +
  theme_minimal()+
  facet_grid(.~`Antidepressant medication`)+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggplot(df.proj2.revise,
       aes(x = Race, y = FTCD, fill = Race)) +
  geom_violin() +
  theme_minimal()+
  facet_grid(.~`Antidepressant medication`)+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


ggplot(df.proj2.revise, aes(y = `Nicotine Metabolism Ratio`, x = FTCD,
                 color = Race, shape = `Antidepressant medication`)) +
  geom_point(size = 3) +
  #labs(x = "Var1", y = "Var2", color = "Category 1", shape = "Category
   ↪  2") +
  theme_minimal()
df.proj2.revise %>%
  mutate(`Antidepressant medication` = case_when(
          `Antidepressant medication` == "True" ~ "Take med",
          `Antidepressant medication` == "False" ~ "Not take med"))%>%
  group_by(Race, `Antidepressant medication`) %>%
  summarize(
    NMR_mean = mean(`Nicotine Metabolism Ratio`,na.rm=T),
    NMR_sd = sd(`Nicotine Metabolism Ratio`,na.rm=T),
    FTCD_mean = mean(FTCD,na.rm=T),
    FTCD_sd = sd(FTCD,na.rm=T)) %>%
  gt() %>%
  tab_options(table.font.size = px(13))%>%
  fmt_number(decimals = 2)
# possible variable transformation to normzalize quantitative variables
# df.proj2$age_ps<-(df.proj2$age_ps)^2
# df.proj2$bdi_score_w00<-sqrt(df.proj2$bdi_score_w00)
# df.proj2$cpd_ps<-log(df.proj2$cpd_ps+1)
# df.proj2$hedonsum_n_pq1<-sqrt(df.proj2$hedonsum_n_pq1)
# df.proj2$hedonsum_y_pq1<-sqrt(df.proj2$hedonsum_y_pq1)
# df.proj2$shaps_score_pq1<-log(df.proj2$shaps_score_pq1+1)
# df.proj2$NMR<-log(df.proj2$NMR+1)

#mice imputation and then test train split
imp <- mice(df.proj2, meth='pmm', maxit = 10, seed=500, print=F)
df.imp <- complete(imp, action="long")

numeric_vars <- c("age_ps", "ftcd_score", "bdi_score_w00", "cpd_ps",
                          "crv_total_pq1",
                           ↪  "hedonsum_n_pq1","hedonsum_y_pq1",
                          "shaps_score_pq1","NMR")#9
categorical_vars <- c("sex_ps", "race", "Hisp", "inc", "edu",
 ↪  "ftcd.5.mins",
                      "otherdiag", "antidepmed", "mde_curr",
                       ↪  "Only.Menthol",
                        "readiness")#11

interaction_formula <- as.formula(
  paste("~", paste(outer(numeric_vars, categorical_vars,
                         function(x, y) paste(x, y, sep = "*")),
                          ↪  collapse = " + ")))

interaction_matrix <- model.matrix(interaction_formula,
 ↪  df.imp)[,-c(1:21)]#first 21 cols are main effects
interaction_df <- as.data.frame(interaction_matrix)
df.imp <- cbind(df.imp, interaction_df)
df.imp$Var_BA <- df.imp$Var * df.imp$BA

ncol = ncol(df.imp)-2 #main effect and all possible interaction

train.all <- data.frame(matrix(ncol = ncol, nrow = 0))
colnames(train.all) <- colnames(df.imp)[-c(1:2)]

test.all <- data.frame(matrix(ncol = ncol, nrow = 0))
colnames(test.all) <- colnames(df.imp)[-c(1:2)]

coef.all <- matrix(ncol = 5, nrow = ncol) #5 imputation
```

```
for(m in 1:5){
  data.subset <- df.imp[df.imp$.imp == m,-c(1,2)]

  #Var == 0 & BA == 0
  #Var == 0 & BA == 1
  #Var == 1 & BA == 0
  #Var == 1 & BA == 1
  trt1 <- data.subset[data.subset$Var == 0 & data.subset$BA == 0,]
  trt2 <- data.subset[data.subset$Var == 0 & data.subset$BA == 1,]
  trt3 <- data.subset[data.subset$Var == 1 & data.subset$BA == 0,]
  trt4 <- data.subset[data.subset$Var == 1 & data.subset$BA == 1,]

  trt.list <- list(trt1, trt2, trt3, trt4)

  #test train split, stratify by treatment

  train.final <- data.frame(matrix(ncol = ncol, nrow = 0))
  colnames(train.final) <- colnames(data.subset)

  test.final <- data.frame(matrix(ncol = ncol, nrow = 0))
  colnames(test.final) <- colnames(data.subset)

  for(df in trt.list){
    idx <- sample(c(TRUE, FALSE), nrow(df), replace=TRUE,
 ↪  prob=c(0.8,0.2))
    train  <- df[idx, ]
    test   <- df[!idx, ]

    train.final <- rbind(train.final, train)
    test.final <- rbind(test.final, test)
    }

  train.all <- rbind(train.all, train.final)
  test.all <- rbind(test.all, test.final)

  x1 <- model.matrix(abst ~ ., data=train.final)[, -1]
  y1 <- train.final$abst

  cvfit = L0Learn.cvfit(x1, y1, nFolds=5, seed=1, loss = "Logistic",
 ↪  penalty="L0L1", nGamma=5, gammaMin=0.0001, maxSuppSize=8)
  optimalGammaIndex = which.min(lapply(cvfit$cvMeans, min))
  optimalLambdaIndex = which.min(cvfit$cvMeans[[optimalGammaIndex]])
  optimalLambda =
 ↪  cvfit$fit$lambda[[optimalGammaIndex]][optimalLambdaIndex]
  coef.iter <- coef(cvfit, lambda=optimalLambda,
 ↪  gamma=cvfit$fit$gamma[optimalGammaIndex])

  coef.all[,m]<-as.vector(coef.iter)

}
#final coef = avg/pool coef
coef.avg <- rowMeans(coef.all, na.rm=T)
#discrimination

#training
x.train <- model.matrix(abst ~ ., data=train.all)[, -1]
y.train <- train.all$abst

scores <- coef.avg[1] + x.train %*% coef.avg[-1]
mod1<-glm(y.train~scores, family= binomial())
pred1 <- predict(mod1, train.all, type = "response")#convert to
 ↪  probabilities

roc1 <- roc(predictor = pred1,
               response = as.factor(mod1$y),
               levels = c(0,1), direction = "<")
plot(roc1, col = "blue", print.auc = TRUE)#print.thres = TRUE

#testing
x2 <- model.matrix(abst ~ ., data=test.all)[, -1]
y2 <- test.all$abst
scores <- coef.avg[1] + x2 %*% coef.avg[-1]
mod2<-glm(y2~scores, family= binomial())
pred2 <- predict(mod2, test.all, type = "response")

roc2 <- roc(predictor = pred2,
               response = as.factor(mod2$y),
               levels = c(0,1), direction = "<")
plot(roc2, col = "green", print.auc = TRUE, print.auc.y = .4,
 ↪  add=TRUE)#print.thres = TRUE

legend(1.7, 0.95, legend=c("Derivation data", "Validation data"),
       col=c("blue", "green"),lty=1, cex=0.8)
pred_ys <- ifelse(pred1 > 0.157, 1, 0)
tab_outcome <- table(mod1$y, pred_ys)

sens1 <- tab_outcome[2, 2]/(tab_outcome[2, 1]+tab_outcome[2, 2])
spec1 <- tab_outcome[1, 1]/(tab_outcome[1, 1]+tab_outcome[1, 2])
ppv1 <- tab_outcome[2, 2]/(tab_outcome[1, 2]+tab_outcome[2, 2])
npv1 <- tab_outcome[1, 1]/(tab_outcome[1, 1]+tab_outcome[2, 1])
acc1 <- (tab_outcome[1, 1]+tab_outcome[2, 2])/sum(tab_outcome)

pred_ys <- ifelse(pred2 > 0.165, 1, 0)
tab_outcome <- table(mod2$y, pred_ys)

sens2 <- tab_outcome[2, 2]/(tab_outcome[2, 1]+tab_outcome[2, 2])
spec2 <- tab_outcome[1, 1]/(tab_outcome[1, 1]+tab_outcome[1, 2])
ppv2 <- tab_outcome[2, 2]/(tab_outcome[1, 2]+tab_outcome[2, 2])
npv2 <- tab_outcome[1, 1]/(tab_outcome[1, 1]+tab_outcome[2, 1])
acc2 <- (tab_outcome[1, 1]+tab_outcome[2, 2])/sum(tab_outcome)

data.frame(Measures = c("Sens", "Spec", "PPV", "NPV", "Acc"),
          Derivation = round(c(sens1, spec1, ppv1, npv1, acc1),3),
          Validation = round(c(sens2, spec2, ppv2, npv2, acc2),3)) %>%
  kable()
#calibration plot1
num_cuts <- 10
calib_data <-  data.frame(prob = pred1,
                          bin = cut(pred1, breaks = num_cuts),
                          class = mod1$y)
calib_data <- calib_data %>%
             group_by(bin) %>%
             summarize(observed = sum(class)/n(),
                       expected = sum(prob)/n(),
                       se = sqrt(observed * (1-observed) / n()))

p1<-ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  geom_errorbar(aes(x = expected, ymin = observed - 1.96 * se,
                    ymax = observed + 1.96 * se),
                colour="black", width=.01)+
  geom_point(aes(x = expected, y = observed)) +
  labs(x = "Expected Proportion", y = "Observed Proportion") +
  theme_minimal()+
  ggtitle("Derivation")
#calibration plot2
num_cuts <- 10
calib_data <-  data.frame(prob = pred2,
                          bin = cut(pred2, breaks = num_cuts),
                          class = mod2$y)
calib_data <- calib_data %>%
             group_by(bin) %>%
             summarize(observed = sum(class)/n(),
                       expected = sum(prob)/n(),
                       se = sqrt(observed * (1-observed) / n()))

p2<-ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  geom_errorbar(aes(x = expected, ymin = observed - 1.96 * se,
                    ymax = observed + 1.96 * se),
                colour="black", width=.01)+
  geom_point(aes(x = expected, y = observed)) +
  labs(x = "Expected Proportion", y = "Observed Proportion") +
  theme_minimal()+
  ggtitle("Validation")

grid.arrange(p1, p2, nrow = 1)
#find out the selected model predictors
nonzero.param.idx <- which(coef.avg != 0)[-1]
# length(coef.avg)-1
# length(colnames(df.imp)[-c(1:3)])
all.param<-colnames(df.imp)[-c(1:3)]
nonzero.param <- all.param[nonzero.param.idx]
nonzero.param.value <- coef.avg[nonzero.param.idx]
nonzero.param.value <- as.data.frame(nonzero.param.value)
row.names(nonzero.param.value) <- c("Behavioral Activation", "Age *
 ↪  Education",
                                    "Age * Current vs past MDD",
                                    "BDI score * Readiness to quit
                                     ↪  smoking")

#no transformation: "BA"  "age_ps:edu"  "age_ps:mde_curr"
 ↪  "bdi_score_w00:readiness"
```

```
kable(as.data.frame(nonzero.param.value), col.names = c("Coefficients"),
 ↪
      booktabs = TRUE, digits = 3)

#variable transformation: "BA"   "Only.Menthol"
 ↪  "hedonsum_y_pq1:mde_curr"  "bdi_score_w00:readiness"
#model assumption check
#deviance, perason residuals to find outliers

train.all$logit<-coef.avg[1] + x.train %*% coef.avg[-1]
train.all$p_hat <- 1 / (1 + exp(-train.all$logit))

# Calculate deviance residuals
train.all$deviance_residuals <- with(train.all,
  sign(abst - p_hat) * sqrt(2 * ((abst * log(abst / p_hat)) + ((1 -
   ↪  abst) * log((1 - abst) / (1 - p_hat)))))
)

# log(0) is undefined (set 0*log(0) to 0)
train.all$deviance_residuals[is.na(train.all$deviance_residuals)] <- 0

train.all$pearson_residuals <- with(train.all, (abst - p_hat) /
 ↪  sqrt(p_hat * (1 - p_hat)))


#pearson_residuals table
summary_table <- train.all %>%
  mutate(`Smoking Abstinence` = case_when(
      abst == 1 ~ "Abstinent Group",
      abst == 0 ~ "Non-abstinent Group"),)%>%
  group_by(`Smoking Abstinence`) %>%
  summarize(
    Mean = mean(pearson_residuals),
    SD = sd(pearson_residuals)
  )

kable(summary_table, format = "latex", booktabs = TRUE, digits = 2)%>%
  kable_styling(latex_options = "hold_position")
#plots
ggplot(train.all, aes(x = p_hat, y = deviance_residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Fitted values (Predicted Probability)", y = "Deviance
   ↪  Residuals", title = "Deviance Residuals vs Fitted Values")

ggplot(train.all, aes(x = p_hat, y = pearson_residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Fitted values (Predicted Probability)", y = "Pearson
   ↪  Residuals", title = "Pearson Residuals vs Fitted Values")

# Loop through each predictor variable and plot deviance residuals
for (var in numeric_vars) {
  p <- ggplot(train.all, aes_string(x = var, y = "deviance_residuals"))
 ↪  +
    geom_point() +
    geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
    labs(x = paste("Predictor", var),
         y = "Deviance Residuals",
         title = paste("Deviance Residuals vs Predictor", var))

  print(p)  # Print each plot to the graphics device
}

# Q-Q plot for deviance residuals
qqnorm(train.all$deviance_residuals, main = "Q-Q Plot of Deviance
 ↪  Residuals")
qqline(train.all$deviance_residuals, col = "red")

# Repeat similarly for Pearson residuals
qqnorm(train.all$pearson_residuals, main = "Q-Q Plot of Pearson
 ↪  Residuals")
qqline(train.all$pearson_residuals, col = "blue")
```