PHP 2550 Project 2: Predictors and Treatment Effects of Smoking Abstinence in Adults with Major Depressive Disorder

Peirong Hao

Abstract

Smoking cessation is an important public health challenge, particularly for individuals with major depressive disorder (MDD), who are more likely to smoke heavily and experience severe withdrawal symptoms. This project aimed to analyze predictors of smoking abstinence among adult smokers with MDD while evaluating the effectiveness of pharmacotherapy and psychotherapy interventions. Data from a 2×2 factorial randomized clinical trial involving 300 participants were used. The trial compared `varenicline` with `placebo` and `behavioral activation for smoking cessation (BASC)` with `standard behavioral treatment (ST)`. Using L1 and L0 regularization techniques, the project identified key predictors influencing smoking abstinence, including the `Fagerstrom Test for Nicotine Dependence (FTCD) score` and the interaction between `varenicline` and the `nicotine metabolism ratio`. The findings underscore the importance of `pharmacotherapy` with `varenicline` and highlight additional factors that could inform personalized smoking cessation strategies.

Introduction

According to [1], major depressive disorder (MDD) is the most common mental health disorder. Research has demonstrated that individuals with MDD not only tend to smoke more heavily but also face more severe withdrawal symptoms, making smoking cessation particularly challenging. Despite the high prevalence of smoking in this population, there have been limited clinical trials aimed at optimizing tobacco treatment strategies for smokers with MDD. In particular, the effectiveness of combining `psychotherapy` and `pharmacotherapy` requires further investigation. Therefore, [1] employed a 2×2 factorial design to evaluate the effectiveness of `psychotherapy` being `behavioral activation for smoking cessation (BASC)` versus `standard behavioral treatment (ST)` and `pharmacotherapy` being `varenicline` versus

`placebo`. It revealed that while `varenicline` significantly improved cessation rates compared to `placebo`, there were no substantial differences between `BASC` and `ST` in enhancing abstinence rates. These findings suggested that `pharmacotherapy` with `varenicline` is effective, but the benefits of additional behavioral interventions may depend on individual characteristics.

Building on the prior work, this project, conducted in collaboration with Dr. George Papandonatos from Brown University's Department of Biostatistics, focused on analyzing baseline variables to identify potential predictors of smoking abstinence, while accounting for treatment effects. The findings aimed to contribute to the development of personalized smoking cessation strategies. The data were collected at Northwestern University and the University of Pennsylvania research clinics. A total of 300 adult participants, daily smokers diagnosed with major depressive disorder (MDD) based on DSM-5 criteria, were enrolled in the study. Each participant expressed an interest in quitting smoking. The trial employed a stratified randomization strategy to ensure a balanced representation across treatment groups, accounting for the clinical site, sex, and severity of depressive symptoms.

The primary outcome of interest was smoking abstinence, a binary variable indicating whether a participant achieved smoking abstinence at the follow-up assessment.The dataset included two treatment variables: pharmacotherapy and psychotherapy. Pharmacotherapy was a binary variable that identified whether participants received varenicline or a placebo medication. Psychotherapy reflected whether participants underwent behavioral activation for smoking cessation (BASC) or standard behavioral treatment (ST). Demographic variables included age, sex, racial and ethnic identities, income, and educational attainment. Baseline smoking behavior and nicotine dependence were assessed through several measures, including the Fagerstrom Test for Nicotine Dependence (FTCD) score, nicotine metabolism ratio (NMR), and the number of cigarettes smoked per day. Additional variables captured whether participants smoked their first cigarette within five minutes of waking and whether they exclusively smoked menthol cigarettes.

Psychological and psychiatric variables provided insights into mental health and smoking-related behaviors. Depressive symptom severity at baseline was measured using the Beck Depression Inventory (BDI). Cigarette reward value quantified the perceived pleasure associated with smoking. The Pleasurable Events Scale measured two rewarding activities: substitute reinforcers and complementary reinforcers. Anhedonia, a condition characterized by a reduced ability to experience pleasure, was included as it was hypothesized that behavioral activation therapy could address this symptom to increase smoking abstinence. Additionally, baseline readiness to quit smoking captured participants' self-reported motivation to stop smoking. One variable indicated whether participants had any other lifetime psychiatric diagnoses based on DSM-5 criteria. Another variable reflected whether participants were taking antidepressant medication. There was also a variable distinguishing participants with current MDD or current and past MDD from those with only a past diagnosis of the disorder.

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 were variables of interest, including the outcome, treatments, and baseline covariates. Among all baseline characteristics, only `Antidepressant medication` had a p-value $< \alpha = 0.05$, indicating it was 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 `standard behavioral treatment + varenicline` group had the most missing data, with 18 observations, while the `behavioral activation for smoking cessation + 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`.

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 1). 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 depended only on the observed variables.

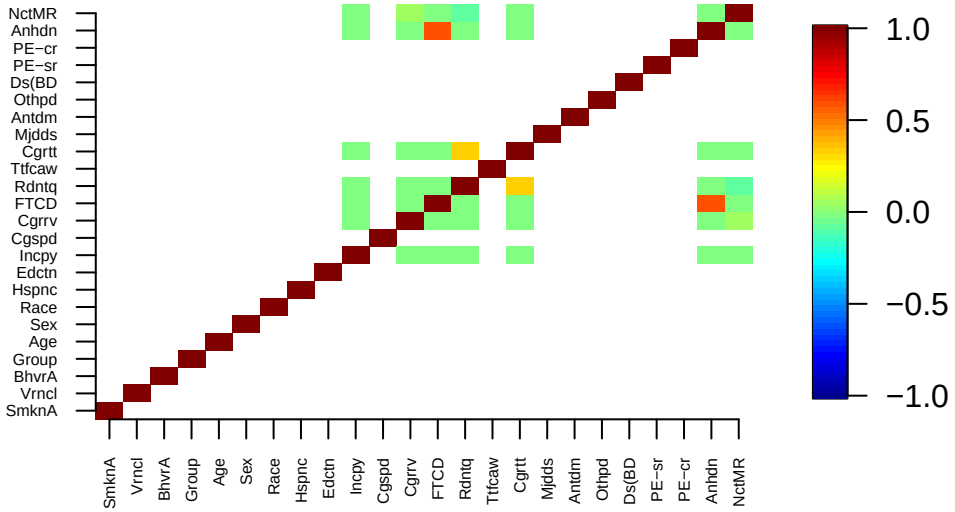

Figure 1: Missing Data Correlation Plot

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 2), suggesting that `anhedonia` could be an important predictor.

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

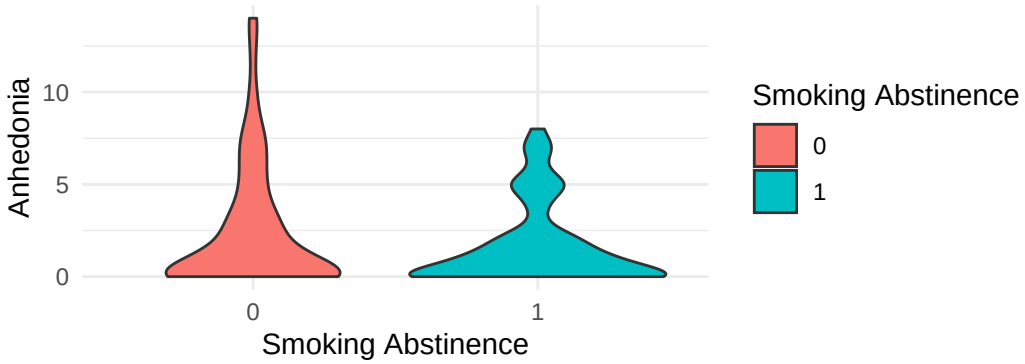

Figure 2: Violin plot of Anhedonia by Smoking Abstinence

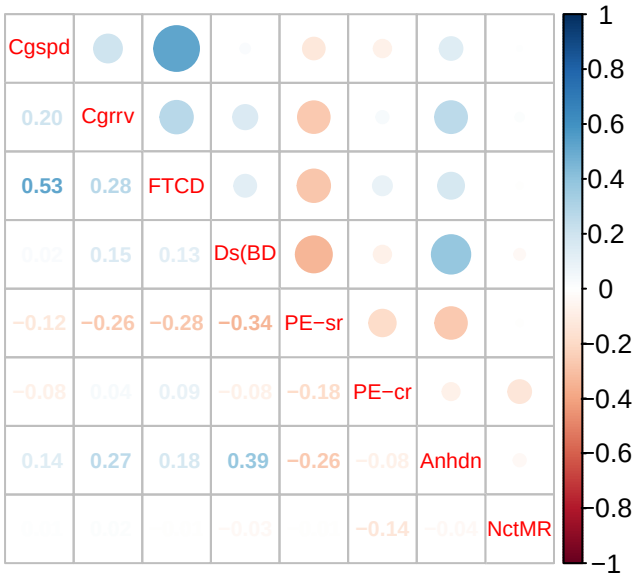

Figure 3: Multi-variable correlation

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

The next step was to create a correlation plot (Figure 3) for all pairs of quantitative variables. Overall, the correlations were not strong, with only one pair showing an absolute correlation value greater than 0.5. The strongest correlation, with an absolute value of 0.53, was between the `Fagerstrom Test for Nicotine Dependence (FTCD) score` and `cigarettes smoked per day`.

The data were 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`, `Fagerstrom Test for Nicotine Dependence (FTCD) score`, 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 `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.

From Figure 4, participants in the `varenicline` group had a higher proportion of outcomes equal to 1 (Smoking Abstinence). Among the `placebo` groups, `standard behavioral treatment (ST)` appeared to be more effective than `behavioral activation for smoking cessation (BASC)`, as `ST` was associated with a higher proportion of outcomes equal to 1.

Because `antidepressant medication` was statistically significantly different between four treatment groups, and `nicotine metabolism ratio (NMR)`, `Fagerstrom Test for`

Table 4: Participant characteristics by outcome group

Characteristic	Smoking Abstinence		p-value
	0 N = 236	1 N = 64	
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

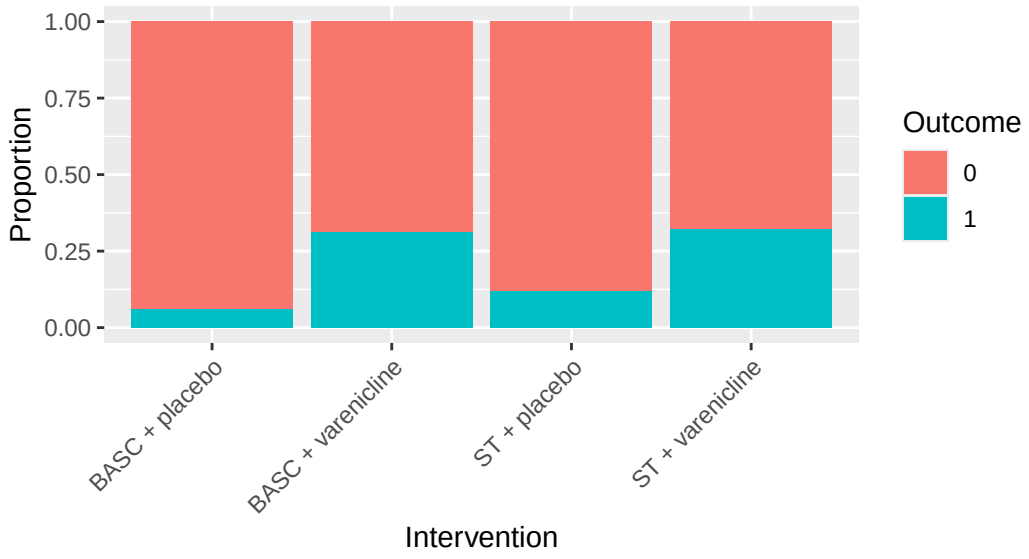

Figure 4: Stacked barplot of outcome by treatment group

`Nicotine Dependence (FTCD) score`, `race` were statistically significantly different between two outcome groups, I made a table for them. White participants have higher `NMR` and `FTCD score` compared to other races. However, there was not a clear trend of whether taking antidepressant medication affects `NMR` and `FTCD score`.

Statistical Modeling

As discussed previously, the missingness in this case was likely Missing at Random (MAR). Since the mice (Multivariate Imputation by Chained Equations) package was 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. Given the limited amount of available data, I performed multiple imputation prior to splitting the dataset, which could have resulted in data leakage. For each imputed dataset, there were four treatment groups, and within each treatment 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.

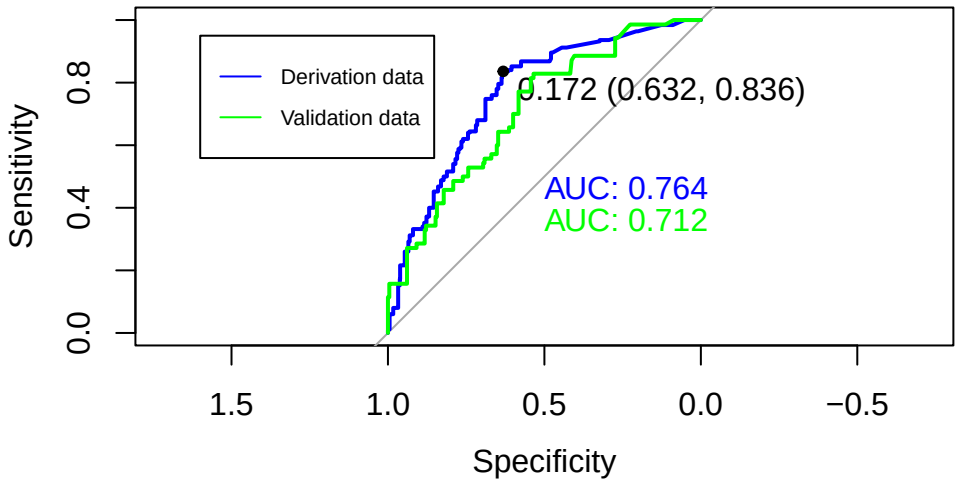

Figure 5: Discrimination plot

The ROC curves in Figure 5 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. Unsurprisingly, the model fit the derivation data better than the validation data, with a slightly higher AUC for the derivation data (0.764) compared to the validation data (0.712).

Table 5: Discrimination table

Measures	Derivation	Validation
Sens	0.836	0.771
Spec	0.632	0.543
PPV	0.374	0.340
NPV	0.936	0.887
Acc	0.674	0.597

Using 0.172 as the cut point for prediction, Table 5 showed an accuracy of approximately 67.4% for the derivation set and 59.7% for the validation set, indicating that the model correctly predicted outcomes in about 60 to 67% of cases, which was not particularly strong. In the derivation set, sensitivity was higher than specificity, suggesting that the model performed better in identifying individuals achieving smoking abstinence but had worse specificity, reflecting weaker performance in correctly predicting outcomes for those not achieving smoking abstinence. The validation set exhibited a similar pattern. The Positive Predictive Value

(PPV) indicated that 37.4% of those with a positive test in the derivation set actually achieved smoking abstinence, slightly higher than the 34% observed in the validation set. The Negative Predictive Value (NPV) showed that 93.6% of those with a negative test in the derivation set did not achieve smoking abstinence, slightly higher than the 88.7% observed in the validation set.

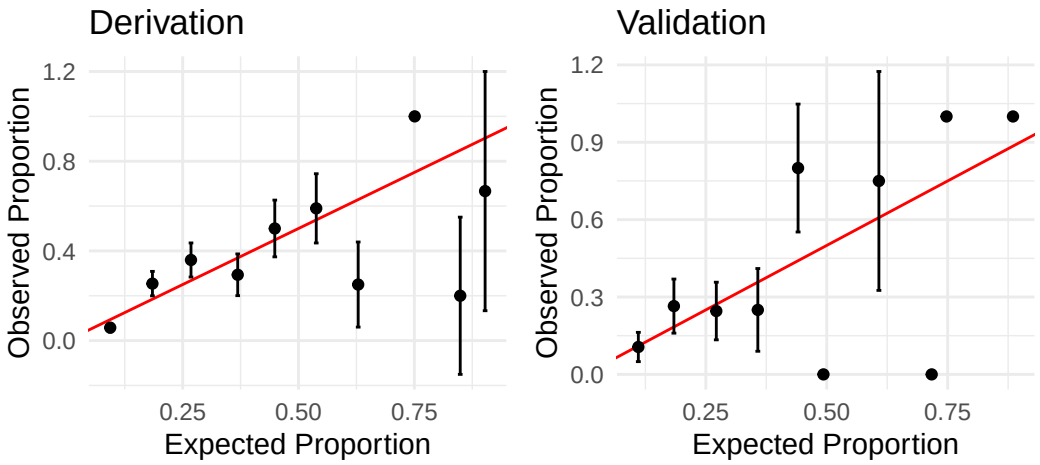

Figure 6: Calibration plot

Figure 6 displayed calibration plots comparing observed versus expected proportions, with the red line representing a perfect fit. In the derivation set's calibration plot, the estimated and true distributions showed poor alignment, with confidence intervals widening as the expected proportion increased. The calibration plot for the validation set demonstrated more narrower confidence intervals, they were reduced to just single points. Both plots showed concerning deviations from the red line for expected proportions greater than 0.5, which may have resulted from the limited proportion of observations achieving smoking abstinence—21% in the derivation set and 23% in the validation set.

Table 6: Coefficients of the Logistic Model

	Coefficients
FTCD score	-0.118
Varenicline * Nicotine Metabolism Ratio	2.720
Varenicline * Smoking with 5 mins of waking up	0.208
Current vs past MDD * Behavioral Activation	-0.210

Results

The model included Fagerstrom Test for Nicotine Dependence (FTCD) score, the interactions between varenicline with both nicotine metabolism ratio and smoking with 5 mins of waking up, along with the interaction between current vs past MDD and behavioral activation. This model was relatively simple, likely due to the effects of L1 (lasso) regularization, which penalized the absolute sum of the coefficients, reducing complexity by setting less influential variable coefficients to zero. Additionally, incorporating L0 regularization, which minimized the count of non-zero coefficients, further prevented overfitting by selecting an optimal subset of features.

The interpretation of the coefficients was as follows: For each additional point earned in FTCD score, we expected the odds of achieving smoking abstinence decreased multiplicately by $e^{-0.118} = 0.89$, holding all other variables constant. This finding aligned with expectations, as individuals with higher nicotine dependence typically faced greater challenges in quitting smoking. A positive coefficient of 2.720 indicated a strong positive interaction effect between varenicline and nicotine metabolism ratio on smoking abstinence. Specifically, as nicotine metabolism ratio increased, individuals who took varenicline have approximately 15.18 times ($e^{2.72} = 15.18$) higher odds of achieving smoking abstinence compared to those who used placebo. Similarly, the positive coefficient of 0.208 suggested an interaction between varenicline and smoking with 5 mins of waking up on smoking abstinence. This indicated that the combined effect of using varenicline and smoking shortly after waking slightly increased the odds of achieving smoking abstinence. While the effect size was modest, it underscored a potential benefit of varenicline for individuals with strong nicotine dependence habits. A negative coefficient of -0.21 indicated an interaction effect between the current vs past MDD and behavioral activation on the likelihood of smoking abstinence. This suggested that individuals with current MDD who received behavioral activation had slightly lower odds of achieving smoking abstinence compared to those with past only MDD who received the same intervention. This finding implied that behavioral activation was less effective for individuals with ongoing depressive 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 Fagerstrom Test for Nicotine Dependence (FTCD) score and an interaction between varenicline and nicotine metabolism ratio. The coefficients remained small, but the AUC decreased slightly (0.756 for derivation set and 0.658 for validation set). Since Fagerstrom Test for Nicotine Dependence (FTCD) score, and the interaction between varenicline and nicotine metabolism ratio were selected in both models (with or without variable transformation), they appeared to be essential predictors of the outcome, smoking abstinence.

Table 7: Pearson Residuals Summary Table

Smoking Abstinence	Mean	SD
Abstinent Group	1.72	0.58
Non-abstinent Group	-0.47	0.19

Since the absolute values of the mean Pearson residuals were less than 2 for both outcome groups, the model appeared to fit well. A comparison of the means suggested that the model overestimated the probability of smoking abstinence for non-abstinent individuals, as indicated by a negative mean, and underestimated the likelihood of smoking abstinence for abstinent individuals, as reflected by a positive mean.

Discussion

This project focused on individuals with mental health disorders, specifically major depressive disorder (MDD), and examined the effectiveness of various treatments as well as the psychological and demographic factors influencing smoking cessation. The model, developed using L1 and L0 regularization techniques, was simplified to reduce complexity and prioritize the most predictive features. Key predictors and interaction effects impacting smoking abstinence were identified. However, certain limitations should be acknowledged. The model's simplicity may have excluded complex relationships, and missing data could have introduced potential bias. Despite these challenges, the findings highlighted the importance of pharmacotherapy with `varenicline`, the `Fagerstrom Test for Nicotine Dependence (FTCD) score`, and the `nicotine metabolism ratio` in influencing smoking cessation outcomes. Additional variables such as `smoking within 5 minutes of waking up`, `current vs past MDD`, and `behavioral activation` were also selected as significant predictors. Interestingly, the "best" model included `varenicline`, which was highlighted as important in previous research, and `behavioral activation`, which was previously considered less significant. These insights could provide valuable directions for future research and policy decisions in the intersection of mental health and smoking cessation.

References

[1] B. Hitsman et al., "Efficacy and Safety of Combination Behavioral Activation for Smoking Cessation and Varenicline for Treating Tobacco Dependence among Individuals with Current or Past Major Depressive Disorder: A 2x2 Factorial, Randomized, Placebo-Controlled Trial," May 2023, doi: https://doi.org/10.1111/add.16209.

Code Appendix

```
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(LOLearn)
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")
```

```r
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),
  group = case_when(
          Var == 0 & BA == 0 ~ 1,
          Var == 0 & BA == 1 ~ 2,
          Var == 1 & BA == 0 ~ 3,
          Var == 1 & BA == 1 ~ 4))%>%
  select(-c(NHW, Black)) %>% mutate(
    across(c(abst, BA, Var, group, sex_ps, race,
           Hisp, inc, edu, ftcd.5.mins,
          otherdiag, antidepmed,
          mde_curr, Only.Menthol,readiness), as.factor))

# 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),]#first 7 variables have
 ↪  missing values

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)
```

```r
#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))
```

```r
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.revise[, sapply(df.proj2.revise, function(col)
 ↪  !is.factor(col))]
non_factor_df <- non_factor_df[,-1]#remove id column

short_colnames <- abbreviate(colnames(non_factor_df), minlength = 5)
colnames(non_factor_df) <- short_colnames
```

```
corrplot.mixed(cor(non_factor_df, use = "complete.obs"), tl.cex = 0.7,
 ↪  number.cex = 0.7)
#cor(non_factor_df, use = "complete.obs")
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)
```

```r
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 = "", y = "")


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)%>%
  cols_label(
    NMR_mean = "NMR\nMean",
    NMR_sd = "NMR\nSD",
    FTCD_mean = "FTCD\nMean",
    FTCD_sd = "FTCD\nSD"
  )
# 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")

base_vars <- c("age_ps", "ftcd_score", "bdi_score_w00", "cpd_ps",
                          "crv_total_pq1",
                           ↪  "hedonsum_n_pq1","hedonsum_y_pq1",
                          "shaps_score_pq1","NMR",
               "sex_ps", "race", "Hisp", "inc", "edu", "ftcd.5.mins",
                      "otherdiag", "antidepmed", "mde_curr",
                       ↪  "Only.Menthol",
                        "readiness")
trt_vars <- c("Var", "BA")


# select variables in df.imp starting with base_vars or trt_vars
selected_vars <- colnames(df.imp)[sapply(colnames(df.imp), function(var)
 ↪  {
  any(startsWith(var, base_vars)) || any(startsWith(var, trt_vars))
})]

interaction_formula <- as.formula(
```

```
  paste("~", paste(outer(base_vars, trt_vars, function(x, y) paste(x, y,
   ↪  sep = "*")),
                   collapse = " + "))
)

interaction_matrix <- model.matrix(interaction_formula, df.imp)
interaction_df <- as.data.frame(interaction_matrix)[,-c(1:36)]#remove
 ↪  main effects w/o main effect
df.imp <- cbind(df.imp, interaction_df)

ncol <- 90 #exclude #id and #imp, 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 = 105) #5 imputation
for(m in 1:5){
  data.subset <- df.imp[df.imp$.imp == m,-c(1,2)]

  trt1 <- data.subset[data.subset$group == 1,]
  trt2 <- data.subset[data.subset$group == 2,]
  trt3 <- data.subset[data.subset$group == 3,]
  trt4 <- data.subset[data.subset$group == 4,]

  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]#exclude intercept
  y1 <- train.final$abst

  rownames(coef.all) <- c("Intercept",colnames(x1))

  cvfit = L0Learn.cvfit(x1, y1, nFolds=5, seed=1, loss = "Logistic",
 ↪  penalty="L0L1",
                        nGamma=5, gammaMin=0.0001, gammaMax=0.1,
                         ↪  maxSuppSize=20)
  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)
#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(1.6, 0.95, legend=c("Derivation data", "Validation data"),
       col=c("blue", "green"),lty=1, cex=0.7)
#https://alicepaul.github.io/health-data-science-using-r/book/logistic_regression.html

pred_ys <- ifelse(pred1 > 0.172, 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 > 0.172, 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)


# summary(train.all$abst)/(950+250)
# summary(test.all$abst)/(230+70)
#find out the selected model predictors
nonzero.param.idx <- which(coef.avg != 0)[-1]
nonzero.param.value <- coef.avg[nonzero.param.idx]
nonzero.param.value <- as.data.frame(nonzero.param.value)
row.names(nonzero.param.value) <- c("FTCD score", "Varenicline *
 ↪  Nicotine Metabolism Ratio",
                                    "Varenicline * Smoking with 5 mins
                                     ↪  of waking up",
                                    "Current vs past MDD * Behavioral
                                     ↪  Activation")

#no transformation: ftcd_score   `Var1:NMR` `Var1:ftcd.5.mins1`
 ↪  `mde_curr1:BA1`
kable(as.data.frame(nonzero.param.value), col.names = c("Coefficients"),
 ↪
      booktabs = TRUE, digits = 3)

#variable transformation: ftcd_score -0.05414289, Var1:NMR 4.24669083
#model assumption check:deviance, perason residuals to find outliers

if (is.factor(train.all$abst)) {
  train.all$abst <- as.numeric(as.character(train.all$abst))
}

if (!is.numeric(x.train)) {
  x.train <- as.matrix(as.numeric(x.train))
}

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)))))
)

train.all$deviance_residuals <- with(train.all,
  sign(abst - p_hat) * sqrt(2 * (
    (abst * ifelse(abst == 0, 0, log(abst / p_hat))) +
    ((1 - abst) * ifelse((1 - abst) == 0, 0, 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")
```