

# Function Draft

Yingxi Kong

## Abstract

**Background:**

**Methods:**

**Results:**

**Conclusion:**

## Introduction

MDD has been one of the most prevalent mental health disorders in the world, with rates that have continued to rise, particularly during the COVID-19 pandemic<sup>1</sup>. Individuals with MDD are not only at risk for a range of adverse health outcomes but are also more likely to engage in harmful health behaviors, including tobacco use which is one of the most prevalent health-compromising behaviors among people with MDD, with the rate of smoking among individuals with major depressive disorder (MDD) is 2–3 higher than in the general population<sup>2</sup>. However, most smoking cessation clinical trials have excluded this important group from trial enrollment<sup>3</sup>, limiting information and suggestion for this group who wants to quit smoking.

In a recent  $2 \times 2$  factorial, randomized, placebo-controlled trial led by Dr. George Papandonatos, the efficacy and safety of combining behavioral and pharmacological treatment were evaluated among individuals with current or past MDD. Behavioral activation for smoking cessation (BASC), a behavioral treatment designed to enhance engagement in rewarding activities and reduce avoidance behavior, was paired with varenicline, a pharmacotherapy shown to reduce cravings and mitigate nicotine’s rewarding effects. The trial included 300 participants and compared BASC to standard treatment (ST) and varenicline to placebo. The results indicated that while varenicline significantly improved abstinence rates compared to placebo, BASC did not outperform ST, suggesting that while pharmacotherapy may provide substantial benefit for smokers with MDD, the behavioral component of cessation treatment may require further refinement.

This study is a collaboration with Dr. George Papandonatos, aiming to investigate the potential effect of baseline characteristics on the effectiveness of behavioral treatment on end-of-treatment (EOT) abstinence outcomes. Furthermore, we aim to assess these baseline characteristics as predictors of abstinence, while controlling for both behavioral treatment and pharmacotherapy. By identifying factors that may influence the efficacy of cessation interventions, this analysis seeks to inform targeted treatment strategies to enhance smoking cessation outcomes among individuals with MDD.

## Methods

The data in this analysis is a collaboration with Dr. George Papandonatos from a  $2 \times 2$  factorial, randomized, placebo-controlled study examining the efficacy and safety of behavioral activation for smoking cessation (BASC) and varenicline in treating tobacco dependence among adults with current or past major depressive disorder (MDD). Our sample population consists of 300 adults smokers with or previously with MDD. Patients were randomly assigned to either behavioral activation for smoking cessation (BASC) or standard behavioral treatment (ST) and either varenicline or placebo groups. That is, participants were assigned to four

distinct intervention groups, including **ST + placebo**, **ST + varenicline**, **BASC + placebo**, and **BASC + varenicline**. Randomization was stratified by clinical site, sex, and level of depressive symptoms to ensure balanced representation across these factors. The data also records patients’ smoking cessation outcomes at week 27 follow-up and relevant baseline characteristics. Key variables include smoking abstinence status (outcome), demographic characteristics (sex, age, income, education), their smoking behaviors (cigarettes per day, time to first cigarette after getting up, nicotine dependence score), and their psychiatric measures (MDD status, anhedonia score, other diagnoses, and antidepressant usage).

Using this data, our analysis aims to identify baseline variables as moderators of the behavioral treatment effects on end-of-treatment (EOT) abstinence and as predictors of smoking cessation, controlling for behavioral treatment and pharmacotherapy. Lasso regression will be applied to identify significant baseline characteristics and their interaction terms with the treatment, enable us to identify significant predictors and potential moderators on the EOT abstinence for people with MDD.

## Data Preprocessing

To prepare the data for analysis, we firstly convert all categorical variables to factor and for socioeconomic factors (income and education) with ordinal levels, we recoded levels in order to improve readability and interpretability. In addition, we combined race and ethnicity indicators into a single race variable with categories including Black, Hispanic, Non-Hispanic White, Mixed Race, and Unknown.

The data also contains various levels of missingness across several variables presented in **Table 1**. Nicotine Metabolism Ratio (NMR) has the highest missingness rate, with 7% of observations missing. The FTCD score at baseline (**ftcd\_score**) has the lowest missing rate, 0.33%, with only one patient missing information on this variable. Given the limited sample size of this data, we prefer to maintain as many observations as possible in our analysis. Thus, to address the missingness, we applied a multiple imputation approach using the `mice()` function from the `mice` package in R before taking data into the primary analysis which provides plausible values for all missing entries across five imputed datasets.

Table 1: Summary of Missing Data Patterns Across Variables

Variable	Missing Count	Missing Percentage
NMR	21	7 %
crv_total_pq1	18	6 %
readiness	17	5.67 %
inc	3	1 %
shaps_score_pq1	3	1 %
Only.Menthol	2	0.67 %
ftcd_score	1	0.33 %

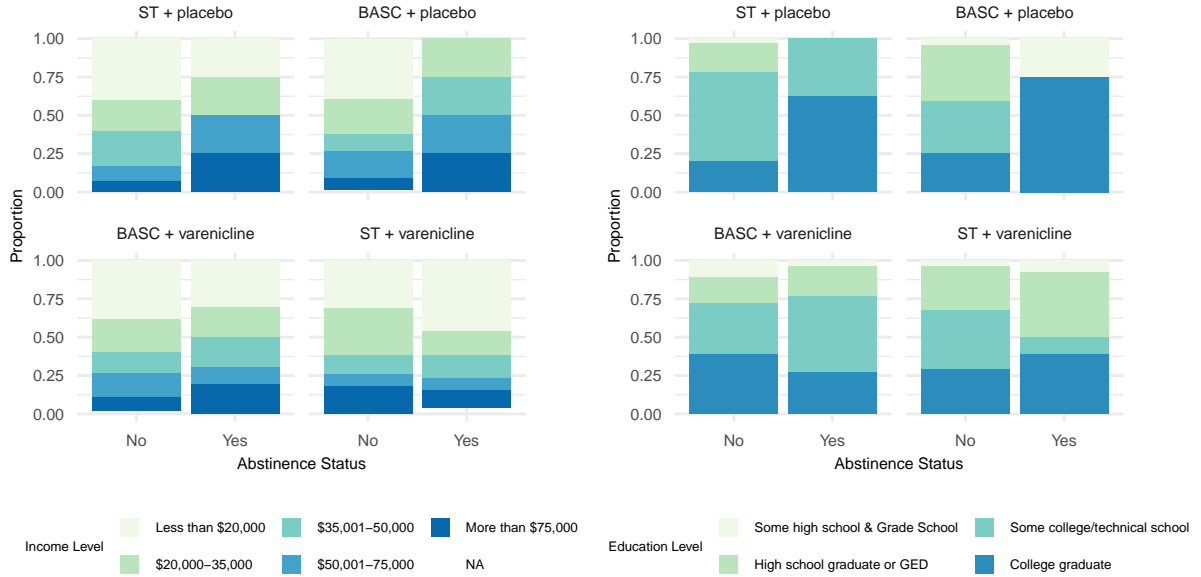
## Data Exploration and Transformation

To investigate potential interactions between baseline characteristics and treatment assignment on end-of-treatment (EOT) abstinence, we examined the distribution of each baseline variable across treatment groups and abstinence outcomes.

For categorical variables, we plot bar charts to show patterns across treatment groups and abstinence groups shown in **Figure 1** and **Figure 2**. Income in **Figure 1** exhibits different distribution among groups and abstinence outcomes. Within each treatment group, different income levels show various proportions of abstinent rate, suggesting that income level may be a potential predictor of treatment effect on abstinence. Most people with less than \$20,000 income **BASC + placebo** did not achieve abstinence at week 27 follow-up, while in the **ST + placebo** group, around 25% participants achieved abstinence. In addition, people with income ranging from \$35,001 to \$50,000 are more likely to stop smoking at week 27 follow-up in the **BASC + placebo** group while most participants with this income level who were assigned to the **ST + placebo** group did not achieve abstinence. Similar patterns observed when we compared the two behavioral treatments with varenicline that people with the same income level exhibit different outcomes. These findings suggest that

income level might be a potential moderator of the behavioral treatment effectiveness on the EOT abstinence among people with MDD.

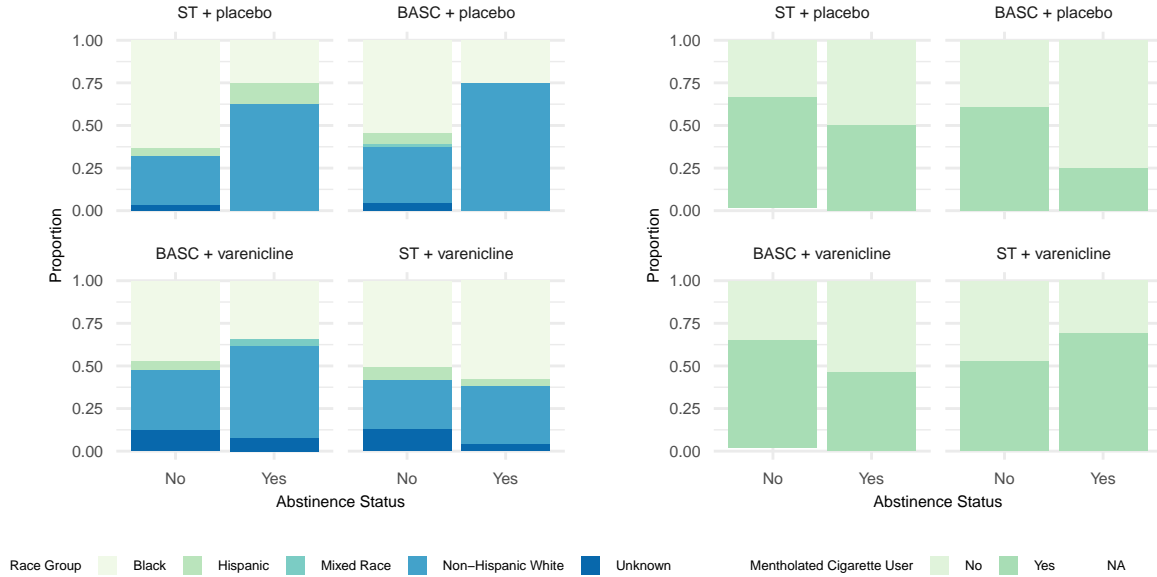
Figure 1: Baseline Characteristics by Abstinence Status and Treatment Group (Categorical 1)



Similar for education level presented in Figure 2, within each treatment group, people with difference education level exhibit various abstinence rate at follow up. For example, observing the BASC + placebo group, college graduated participants are more likely to achieve abstinence at follow up compared to those with lower education levels, suggesting that income level may be a potential predictor of treatment effect on abstinence. Additionally, comparing the two behavioral treatment with varenicline, college graduated participants are less likely to achieve abstinence with BASC while they present higher abstinence rate at week 27 follow-up with ST. Also, patients with high school graduate or GED exhibit higher likelihood of smoking cessation with BASC while their abstinence rate becomes much lower with ST, further suggesting that education level could be a potential moderator of the treatment effects on the EOT abstinence among people with MDD.

Race and the indicator of exclusive mentholated cigarette users (`Only.Menthol`) also exhibit difference distribution across treatment groups and outcome values shown in Figure 2. For example, in the ST + placebo, BASC + placebo, and BASC + varenicline groups, non-Hispanic White participants are more likely to stop smoking compared to Hispanic and Black participants. This indicates that race might be a potential predictor of the treatment effect on EOT abstinence for people with MDD. Additionally, comparing the two behavioral treatments with placebo, Hispanic participants with ST are more likely to achieve smoking cessation while this pattern reverses when they were assigned to BASC. Moreover, comparing the two varenicline groups, black people with BASC show less likelihood of stop smoking while they show larger abstinence rate in the ST group. Similar pattern observed for the indicator of exclusive mentholated cigarette users (`Only.Menthol`). In the two varenicline groups, mentholated cigarette users (`Only.Menthol` = 1) with ST are more likely to achieve abstinence while these users with BASC exhibit much lower abstinence rate at week 27 follow-up. These findings suggest that race and the indicator of exclusive mentholated cigarette users could be potential predictors or moderators of the treatment effects on the EOT abstinence for people with MDD.

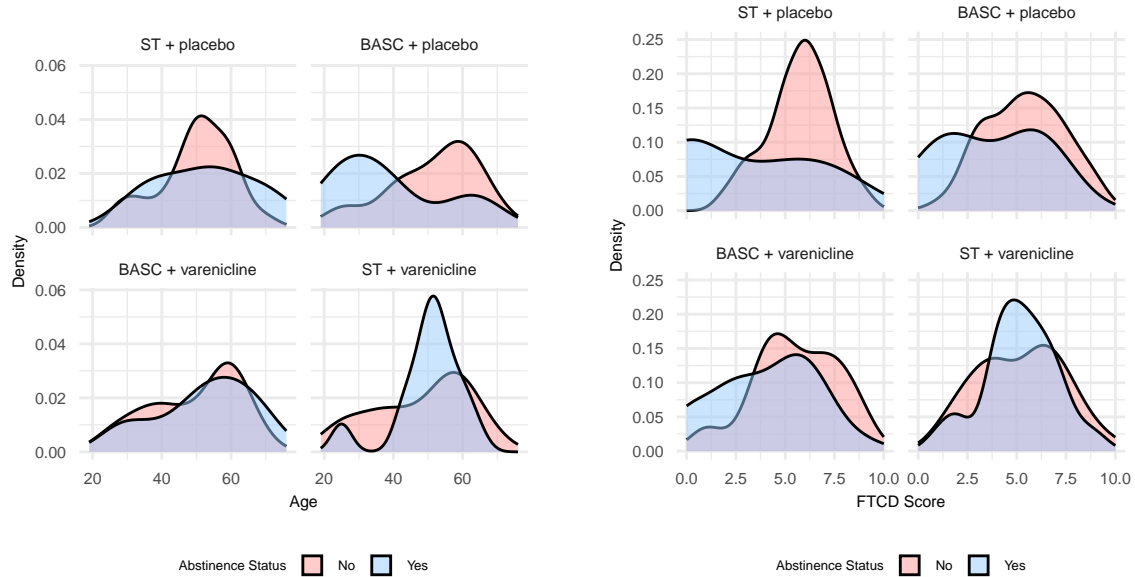
Figure 2: Baseline Characteristics by Abstinence Status and Treatment Group (Categorical 2)



We also examine the distribution of continuous variables by their treatment group and outcome values shown in **Figure 3** and **Figure 4**. Among all continuous variables, age, FTCD score, NMR, and BDI score exhibit different distribution among treatment groups and abstinence status at week 27 follow-up.

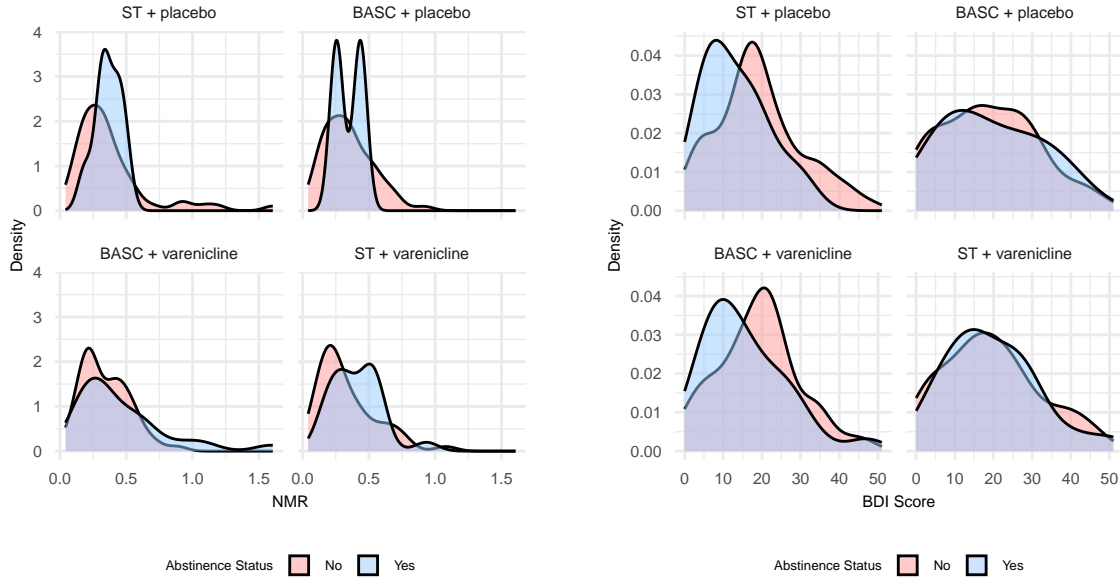
Seeing **Figure 3**, the abstinence rate changes as age varies within the same treatment group, suggesting age as a predictor of treatment effect on EOT abstinence. Moreover, comparing the two placebo groups, younger people with BASC shows higher abstinence rate while this group exhibits low abstinence rate with ST. In the two varenicline groups, although the general pattern of distributions are similar, middle-age people with ST show significantly higher abstinence rate compared to middle-age people with BASC. Additionally, the distribution of abstinence rate changes as FTCD score varies, suggesting the predicting role of FTCD score on the treatment effect. Moreover, participants with higher FTCD score in the ST + placebo group show significantly higher likelihood to continue smoking compared to whom in the BASC + placebo group. Participants with FTCD score around 5 show higher abstinence rate in the ST + varenicline group compared to those with BASC + varenicline. These findings suggest a potential age-behavioral treatment and FTCD score-treatment interaction terms that age and FTCD score might be predictors and moderators of the treatment effect on the EOT abstinence.

Figure 3: Baseline Characteristics by Abstinence Status and Treatment Group (Continuous 1)



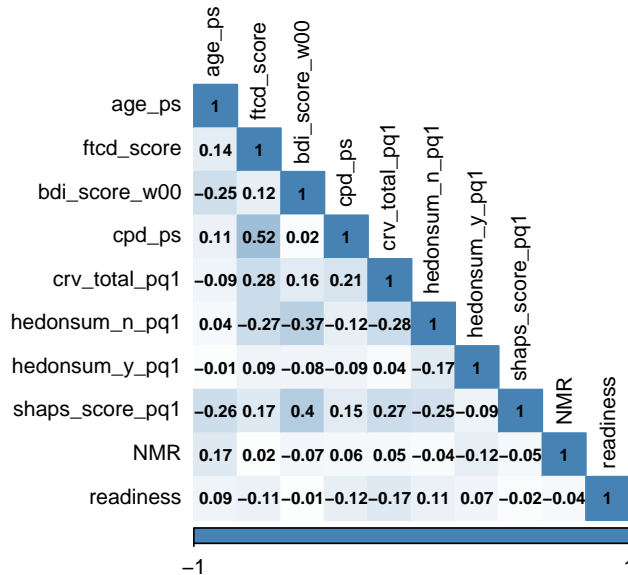
Seeing Figure 4, the distribution of NMR is skewed towards participants with higher nicotine metabolism ratio (NMR) across all groups. In the two placebo groups, participants with lower NMR are more likely to quit smoking at week 27 follow-up while this gap is less pronounced in the two varenicline groups. However, in the BASC + placebo group, there is a huge gap on abstinence rate between participants with NMR ranging from 0.3 to 0.5. Moreover, comparing the two varenicline groups, people with NMR ranging from 0.3 to 0.7 who were assigned to ST show higher abstinence rate while they show much lower abstinence rate with BASC. Moreover, comparing the two placebo groups, participants in the ST group with lower BDI scores (less severe depressive symptoms) are more likely to stop smoking and those with middle level BDI scores are less likely to stop smoking. However, in the BASC group, participants show similar distribution regardless of their depressive severity, suggesting the complex interacting relationship between BDI score and the combination of behavioral and pharmacological treatments. The same pattern observed when we compared the two varenicline groups. These findings suggest that both NMR and BDI scores may interact with behavioral treatment type, influencing smoking cessation outcomes.

Figure 4: Baseline Characteristics by Abstinence Status and Treatment Group (Continuous 2)



Additionally, examining the correlation among continuous variables, we observed that most variables show low to moderate correlations with each other, with both positive and negative relationship present.

Figure 3: Correlation Plot among Environmental Condition Characteristics



Next, to address skewness in several variables, we applied specific transformations based on the distributional characteristics of each variable (transformations were performed after the imputation step). Among all continuous variables, *hedonsum\_n\_pq1*, *hedonsum\_y\_pq1*, *shaps\_score\_pq1*, and *NMR* exhibit right skewness over distribution. Table 2 summarizes the skewness value of these variables before and after transformation.

For the two pleasurable events scale variables, `hedonsum_n_pq1` and `hedonsum_y_pq1`, we applied a square root transformation to reduce their high positive skewness values (1.34 and 1.39, respectively). This transformation brought their skewness close to zero (-0.06 and 0.06, respectively), resulting a more symmetric distribution.

Additionally, since `shaps_score_pq1` contains nearly 50% zero entries and it exhibits high positive skewness value 1.71, we explored several transformations, including log, square root, and inverse hyperbolic sine (`asinh()`). The inverse hyperbolic sine transformation produced the lowest skewness (0.52), making it the most suitable choice for this variable.

Finally, we applied a log transformation on NMR which presents the highest skewness value before transformation (1.92). The log transformation successfully reduced the skewness to a nearly symmetric value.

Table 2: Variable Transformation on Skewness

Variable	Transformation	Skewness before Transformation	Skewness after Transformation
<code>hedonsum_n_pq1</code>	Square Root Transformation	1.338843	-0.0591728
<code>hedonsum_y_pq1</code>	Square Root Transformation	1.391398	0.0620129
<code>shaps_score_pq1</code>	Inverse Hyperbolic Sine Transformation	1.707230	0.5217093
NMR	Log Transformation	1.915358	-0.2241582

## Results

Before conducting the primary analysis, we performed exploratory data analysis (EDA) to examine baseline characteristics, assess data distributions, and identify potential relationships within the dataset.

Table 1 presents an overall summary statistics of patients' baseline characteristics by their behavioral and pharmacological treatment assignment. Since our study is a  $2 \times 2$ , factorial, randomized, placebo-controlled trial, patients are randomly assigned to either behavioral activation for smoking cessation group (BASC) or standard behavioral treatment group (ST) and either varenicline or placebo blister pack. Patients can be categorized into four treatment arm groups: BASC + placebo, BASC + varenicline, ST + placebo, and ST + varenicline. Seeing from Table 1, the two placebo groups both have 68 observations while the two varenicline groups have 83 and 81 observations, respectively.

Most variables are evenly distributed across the four treatment arms, which reflects successful randomization in this factorial trial. However, a few key factors, such as socioeconomic indicators (income and education) and specific mental health variables (MDD status, DSM-5 diagnoses), exhibit slight variations that may influence outcomes. Notably, treatment arms with varenicline show higher abstinence rates than placebo groups, suggesting the potential efficacy of this pharmacotherapy in combination with behavioral interventions. While many baseline characteristics are evenly distributed across groups, some may still function as moderators, potentially interacting with treatment assignment to affect abstinence success. In addition, only one observation falls into the Grade School level in the education variable. To ensure the appropriate representation of categories, we combined the grade school level with the next level, some high school, during the regression analysis. This adjustment ensures sufficient sample sizes across categories when we split the data.

Table 3: Participant Characteristics by Treatment Arm

Characteristic	Behavioral and Pharmacological Treatment Assignment				Overall, N = 300
	ST + placebo, N = 68	BASC + placebo, N = 68	BASC + varenicline, N = 83	ST + varenicline, N = 81	
<b>Smoking abstinence</b>	8 (12%)	4 (5.9%)	26 (31%)	26 (32%)	64 (21%)
<b>Age</b>	50 (11)	51 (14)	50 (13)	49 (13)	50 (13)
<b>Sex</b>					
Male	29 (43%)	30 (44%)	39 (47%)	37 (46%)	135 (45%)
Female	39 (57%)	38 (56%)	44 (53%)	44 (54%)	165 (55%)
<b>Income</b>					
Less than \$20,000	26 (38%)	25 (37%)	30 (37%)	29 (36%)	110 (37%)

Table 3: Participant Characteristics by Treatment Arm (*continued*)

Characteristic	Behavioral and Pharmacological Treatment Assignment				
	ST + placebo, N = 68	BASC + placebo, N = 68	BASC + varenicline, N = 83	ST + varenicline, N = 81	Overall, N = 300
\$20,000-35,000	14 (21%)	16 (24%)	17 (21%)	21 (26%)	68 (23%)
\$35,001-50,000	14 (21%)	8 (12%)	13 (16%)	11 (14%)	46 (15%)
\$50,001-75,000	8 (12%)	12 (18%)	12 (15%)	6 (7.5%)	38 (13%)
More than \$75,000	6 (8.8%)	6 (9.0%)	10 (12%)	13 (16%)	35 (12%)
Missing	0	1	1	1	3
<b>Education</b>					
Some high school & Grade School	2 (2.9%)	4 (5.9%)	7 (8.4%)	4 (4.9%)	17 (5.7%)
High school graduate or GED	11 (16%)	23 (34%)	15 (18%)	27 (33%)	76 (25%)
Some college/technical school	38 (56%)	22 (32%)	32 (39%)	24 (30%)	116 (39%)
College graduate	17 (25%)	19 (28%)	29 (35%)	26 (32%)	91 (30%)
<b>FTCD score</b>	5 (2)	5 (2)	5 (2)	5 (2)	5 (2)
Missing	1	0	0	0	1
<b>Smoking within 5 mins of waking up</b>	35 (51%)	32 (47%)	33 (40%)	38 (47%)	138 (46%)
<b>BDI score</b>	18 (11)	19 (12)	18 (11)	20 (12)	19 (11)
<b>Cigarettes smoked per day</b>	15 (7)	16 (9)	16 (9)	14 (7)	15 (8)
<b>Cigarette reward value</b>	7 (4)	7 (4)	7 (4)	7 (3)	7 (4)
Missing	8	1	3	6	18
<b>Pleasurable events (substitute reinforcers)</b>	21 (20)	23 (20)	23 (19)	23 (19)	23 (20)
<b>Pleasurable events (complementary reinforcers)</b>	27 (20)	28 (22)	22 (17)	25 (19)	25 (19)
<b>Anhedonia</b>	3 (3)	2 (3)	2 (3)	2 (3)	2 (3)
Missing	1	2	0	0	3
<b>Other lifetime DSM-5 diagnosis</b>	28 (41%)	35 (51%)	30 (36%)	40 (49%)	133 (44%)
<b>Taking antidepressant Current vs. past MDD</b>	15 (22%)	28 (41%)	24 (29%)	15 (19%)	82 (27%)
Past MDD	37 (54%)	36 (53%)	43 (52%)	37 (46%)	153 (51%)
Current MDD	31 (46%)	32 (47%)	40 (48%)	44 (54%)	147 (49%)
<b>Nicotine metabolism ratio</b>	0.37 (0.27)	0.34 (0.18)	0.38 (0.25)	0.36 (0.21)	0.36 (0.23)
Missing	2	7	3	9	21
<b>Exclusive mentholated cigarette user</b>	43 (64%)	40 (59%)	48 (59%)	47 (58%)	178 (60%)
Missing	1	0	1	0	2
<b>Readiness to quit smoking</b>	7 (1)	7 (1)	7 (1)	7 (1)	7 (1)
Missing	4	4	5	4	17
<b>Race</b>					
Black	40 (59%)	36 (53%)	36 (43%)	43 (53%)	155 (52%)
Hispanic	4 (5.9%)	4 (5.9%)	3 (3.6%)	5 (6.2%)	16 (5.3%)
Mixed Race	0 (0%)	1 (1.5%)	1 (1.2%)	0 (0%)	2 (0.7%)
Non-Hispanic White	22 (32%)	24 (35%)	34 (41%)	25 (31%)	105 (35%)
Unknown	2 (2.9%)	3 (4.4%)	9 (11%)	8 (9.9%)	22 (7.3%)

<sup>1</sup> Mean (SD) for continuous; n (%) for categorical

To analyze the impact of behavioral treatment on end-of-treatment abstinence and examine the moderating role of baseline characteristics, we selected Lasso regression as our primary model. Lasso was chosen for its ability to perform both variable selection and regularization, making it particularly suited for our study, which involves numerous baseline predictors and interaction terms. By applying an L1 penalty, Lasso shrinks less relevant coefficients to zero, effectively selecting a subset of the most influential predictors and interactions.

As mentioned earlier, we performed multiple imputation using `mice()` function to generate five different imputed data to address missingness. To account for skewness in the data, we then applied the corresponding transformations shown in Table 2 to the skewed variables in each of the five imputed datasets. Each



imputed dataset was split into a 70% training set and a 30% test set, stratified by treatment group using the `createDataPartition()` function in the `caret` package. Lasso regression was then applied to each training set using `cv.glmnet()` with a design matrix that included all baseline characteristics and their interactions with the behavioral treatment and pharmacological treatment, respectively. To maintain the same distribution of treatment group across each cross-validation folds in lasso regression, we created custom fold assignments by treatment level and specified these assignments in the `foldid` argument in `cv.glmnet()`. During cross-validation, we identified the optimal regularization parameter, `lambda.min`, which minimized the cross-validated error and extract the coefficient estimates for each lasso model at this optimal lambda value. Finally, we averaged the coefficient estimates across all five Lasso models to obtain the final pooled estimates present in Table 4.

Table 4: Lasso Model Coefficient Estimate

Variable	Estimate	Exponential Estimate
(Intercept)	-0.9857992	0.3731409
ftcd_score	-0.1216751	0.8854360
hedonsum_n_pq1	-0.0018667	0.9981351
mde_curr1	-0.2054373	0.8142912
NMR	0.2363417	1.2666071
raceNon-Hispanic White	0.3216338	1.3793796
BA1:inc\$35,001-50,000	0.0124385	1.0125162
BA1:Only.Menthol1	-0.3842225	0.6809799
BA1:raceNon-Hispanic White	0.0060259	1.0060441
crv_total_pq1:Var1	0.0244015	1.0247016
eduHigh school graduate or GED:Var1	0.0456560	1.0467143
raceMixed Race:Var1	0.8554848	2.3525146
sex_ps2:Var1	0.0786102	1.0817826
age_ps:Var1	0.0170570	1.0172033

Observing the main effects variables in Table 4, the `ftcd_score` has a coefficient estimate of -0.1217, suggesting that for each one-unit increase in FTCD score, the odds of abstinence for people with MDD would be approximately 0.8854. That is, more severe nicotine dependence might lead to a lower probability of abstinence after treatment for people with MDD, holding other factors constant. `mde_curr1` which is an indicator of the participants' MDD status, has an estimate of -0.2054. This indicates that current MDD participants might have lower likelihood of abstinence after treatment adjusting for their treatment groups with the odds of smoking cessation decreases by around  $1 - 0.8142 = 18.58$ . The log-transformed Nicotine Metabolism Ratio (NMR) has an estimate of 0.2363, suggesting that for each unit increase in  $\log(\text{NMR})$ , the odds of abstinence rise by about 26.7%, holding other factors constant. That is, faster nicotine metabolism would be associated with higher odds of abstinence, adjusting for their treatment groups and other covariates.

The non-Hispanic White level in the race variable show an estimate of 0.3216, indicating that the odds of abstinence for Non-Hispanic White people with MDD is 37.93% higher than that for other racial groups. Notably, there is also an interaction term between BASC and the racial group. This interaction suggests that Non-Hispanic White people using the BASC treatment might experience a small additional increase in the odds of abstinence compared to people in other racial groups. Thus, race variable not only functions as predictor but also as a moderator since the treatment effect of BASC on smoking cessation varies among racial groups.

This is same for the indicator of exclusive mentholated cigarette user, `Only.Menthol1`, and the income characteristics, `inc`. The interaction term between BASC treatment and `Only.Menthol1` displays an estimate of -0.3842, suggesting that for MDD people receiving the BASC treatment, those who are mentholated cigarette user would have a 31.81% of decrease in the odds of abstinence compared to those who are not mentholated cigarette user. Additionally, the interaction term between BASC treatment and income (\$35,001-50,000) has an estimate of 0.0124. This suggests that for MDD people receiving the BASC treatment, those who have income ranging from \$35,000 to \$50,000 would experience approximately a 1.25% increase in the odds of abstinence compared to individuals with different income levels.

The model also has several interaction term with the indicator of varenicline treatment (Var) which identify potential moderators for the pharmacological treatment on the EOT abstinence for people with MDD. Specifically, cigarette reward value (`crv_total_pq1`), education level (level: high school graduate or GED), race (level: mixed race), sex, and age serve as potential moderators. These interactions suggest that the effectiveness of varenicline may vary depending on these baseline characteristics, highlighting the nuanced impact of pharmacological treatment across different demographic and behavioral profiles.

We also examine the model’s performance using ROC and calibration plots shown in **Figure 6** and **Figure 7**. The prediction values for the train and test sets are calculated using the long format of train and test sets from the five imputed datasets, respectively. Observing **Figure 6**, the ROC curves indicate that our model effectively distinguishes between positive and negative classes. The train set has an AUC of 0.809 while the test set has an AUC of 0.75. Although there is a slight drop in AUC on the test set, the model still shows great discriminative power on unseen data.

Moreover, observing **Figure 7**, the calibration plot for the training set shows that the model is well-calibrated with lower probability predictions, with points distributed around the diagonal line which indicates the alignment between the predictions and observed values. However, as probabilities increase, slight deviations appear. The test set calibration plot shows more pronounced deviations from the diagonal line. This suggests that while the model’s predictions are reasonably aligned with actual outcomes for the training data, calibration issues are more substantial in the test set, which needs further investigation.

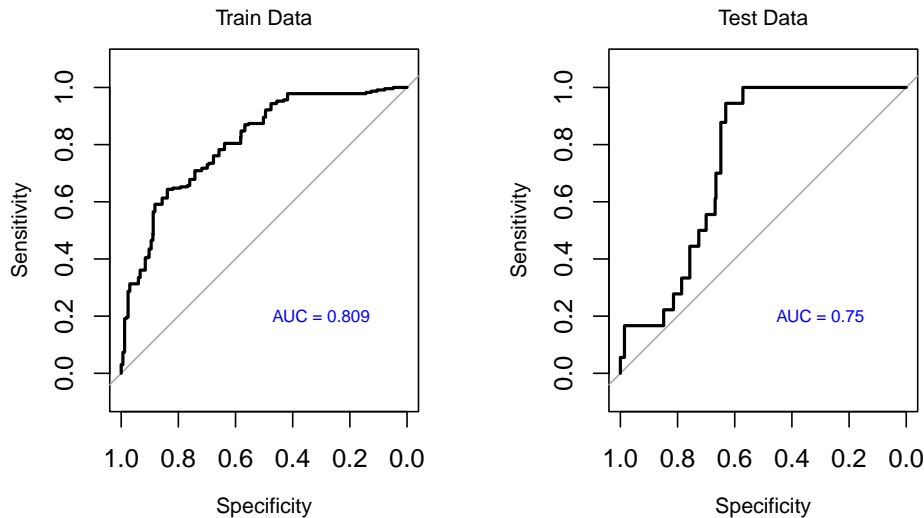
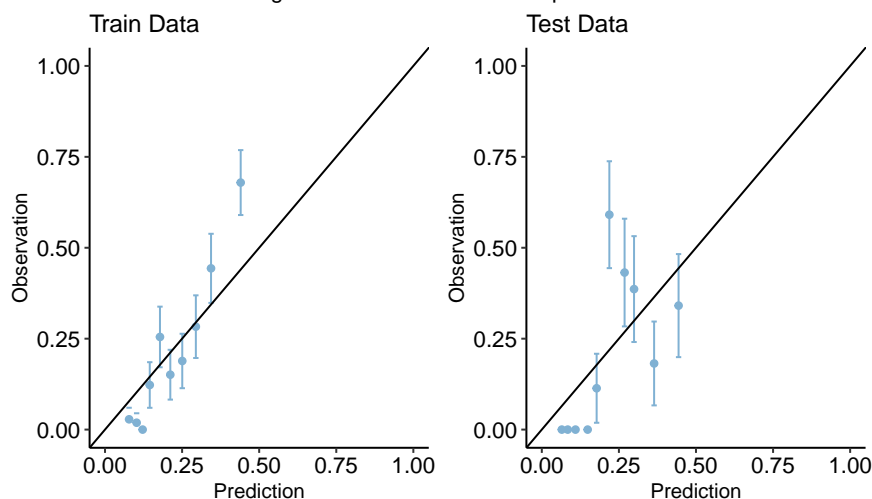


Figure 4: Calibration Plot Comparison



## Discussion

Collaborating with the study of smoking cessation for people with MDD led by Dr. George Papandonatos, this study aims to further investigate potential moderators of behavioral treatment (BASC vs. ST) on the EOT abstinence for this group of people and dedicates to identify baseline characteristics as predictors of abstinence, controlling for behavioral treatment and pharmacotherapy. Lasso regression is chosen to perform variable selection that helps to focus on the most influential baseline characteristics and their interaction terms to balance predictivity and interpretability. As predictors, control for treatment and other factors, higher nicotine dependence (higher FTCD score), and current MDD status both associate with lower likelihood of abstinence. Conversely, faster nicotine metabolism (higher NMR in log scale) associates with higher odds of abstinence, adjusting for treatment and other factors.

The race variable, particularly Non-Hispanic White level, serves as both predictor and moderator of behavioral treatment. Non-Hispanic White people had higher odds of abstinence and they experience more benefits from the BASC treatment. The indicator of menthol cigarette users (`Only.Menthol`) and income also serve as the moderators of behavioral treatment on the EOT abstinence for people with MDD. Specifically, menthol cigarette users treat with BASC face lower abstinence odds compared to non-menthol cigarette users treat with BASC, while in the BASC treatment group, people with income ranging from \$35,001 to \$50,000 benefits more from the BASC treatment compared to people with other levels of income. Additionally, the significant interaction terms with varenicline (`Var`) also suggest that the effectiveness of pharmacological treatment would vary by cigarette reward value, education level, race, sex, and age.

The model evaluation using ROC and calibration plots reveals our model’s strong discriminative power but raises slight concern regarding calibration on the test set which needs further investigation.

As the original paper states, a limitation of this study is that adherence to both behavioral treatment and pharmacotherapy was low, particular in the BASC-only group. The use of ITT approach under this scenario might underestimate treatment effects due to low adherence. An additional limitation of our analysis is that, although our sample provides a diverse phase of patients with MDD, certain levels within some categorical variables have limited sample sizes. Specifically, as shown in **Table 3**, only one observation falls within the “Grade school” category for the education variable. This limited representation may reduce the statistical power to detect effects for these categories, limiting the generalizability of our study. Further researches could explore strategies to enhance engagement and adherence and expand data with more balanced underrepresented groups to improve the accuracy and generalizability of our model. A future direction for this study would be to apply a multilevel modeling approach to better understand the influences of both individual and contextual factors on smoking cessation outcomes in individuals with MDD. This approach would allow for the analysis of time-varying and group-level predictors, such as provider expertise or site-specific support,

to assess how these factors interact with individual characteristics to impact abstinence.

## Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

# load necessary packages
library(tidyverse)
library(mice)
library(gt)
library(gtsummary)
library(kableExtra)
library(RColorBrewer)
library(scico)
library(caret)
library(glmnet)
library(pROC)
library(predtools)
library(gridExtra)
library(ggpubr)
library(patchwork)
library(e1071)
library(corrplot)
library(L0Learn)
library(MASS)
# set working directory
# Windows
setwd("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Data/")

# Mac
# setwd("~/Desktop/Fall 2024/PHP 2550/Data/")

# read in data
data <- read.csv("project2.csv")
# factor categorical variables
data[, c("abst", "Var", "BA", "sex_ps", "NHW",
        "Black", "Hisp", "inc", "edu", "ftcd.5.mins",
        "otherdiag", "antidepmed", "mde_curr",
        "Only.Menthol")] <- lapply(data[, c("abst", "Var", "BA", "sex_ps", "NHW",
        "Black", "Hisp", "inc", "edu",
        "ftcd.5.mins", "otherdiag", "antidepmed",
        "mde_curr", "Only.Menthol")], as.factor)

# Recode factor levels in the dataset
averaged_data_factor <- data %>%
  mutate(abst = fct_recode(as.factor(abst), "Yes" = "1", "No" = "0"),
         inc = fct_recode(as.factor(inc),
                        "Less than $20,000" = "1",
                        "$20,000-35,000" = "2",
                        "$35,001-50,000" = "3",
                        "$50,001-75,000" = "4",
                        "More than $75,000" = "5"),
         sex_ps = fct_recode(as.factor(sex_ps), "Male" = "1", "Female" = "2"),
         edu = fct_recode(as.factor(edu),
                        "Grade School" = "1",
```

```

      "Some high school" = "2",
      "High school graduate or GED" = "3",
      "Some college/technical school" = "4",
      "College graduate" = "5"),
ftcd.5.mins = fct_recode(as.factor(ftcd.5.mins), "Yes" = "1", "No" = "0"),
otherdiag = fct_recode(as.factor(otherdiag), "Yes" = "1", "No" = "0"),
antidepmed = fct_recode(as.factor(antidepmed), "Yes" = "1", "No" = "0"),
mde_curr = fct_recode(as.factor(mde_curr), "Current MDD" = "1", "Past MDD" = "0"),
Only.Menthol = fct_recode(as.factor(Only.Menthol), "Yes" = "1", "No" = "0"),
race = as.factor(case_when(Black == 0 & Hisp == 0 & NHW == 0 ~ "Unknown",
                           Black == 1 & Hisp == 1 & NHW == 1 ~ "Mixed Race",
                           Black == 1 & Hisp == 1 ~ "Mixed Race",
                           Black == 1 & NHW == 1 ~ "Mixed Race",
                           NHW == 1 & Hisp == 1 ~ "Mixed Race",
                           Black == 1 ~ "Black",
                           Hisp == 1 ~ "Hispanic",
                           NHW == 1 ~ "Non-Hispanic White",
                           TRUE ~ "Other")),
trt = as.factor(case_when(Var == 1 & BA == 1 ~ "BASC + varenicline",
                           Var == 0 & BA == 1 ~ "BASC + placebo",
                           Var == 1 & BA == 0 ~ "ST + varenicline",
                           Var == 0 & BA == 0 ~ "ST + placebo",
                           TRUE ~ NA_character_)))

averaged_data_factor$trt <- relevel(factor(averaged_data_factor$trt), ref = "ST + placebo")

averaged_data_factor <- averaged_data_factor %>%
  mutate(inc = fct_relevel(inc, "Less than $20,000", "$20,000-35,000",
                           "$35,001-50,000", "$50,001-75,000", "More than $75,000"),
         edu = fct_relevel(edu, "Grade School", "Some high school", "High school graduate or GED",
                           "Some college/technical school", "College graduate"))

averaged_data_factor$edu <- recode(averaged_data_factor$edu, "Grade School" = "Some high school & Grade
averaged_data_factor$edu <- recode(averaged_data_factor$edu, "Some high school" = "Some high school & G
missingness_df <- averaged_data_factor %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Missing_Count") %>%
  mutate(Total_Count = nrow(averaged_data_factor),
         Missing_Percentage = paste(round((Missing_Count / Total_Count) * 100, 2), "%")) %>%
  arrange(desc(Missing_Percentage)) %>%
  filter(Missing_Count != 0) %>%
  dplyr::select(-Total_Count)

colnames(missingness_df) <- c("Variable", "Missing Count", "Missing Percentage")
missingness_df %>%
  kable(booktabs = TRUE, caption = "Summary of Missing Data Patterns Across Variables ") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down"))
income_stackplot <- ggplot(averaged_data_factor, aes(x = abst, fill = inc)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(x = "Abstinence Status",
       y = "Proportion",
       fill = "Income Level") +

```

```

theme_minimal() +
scale_fill_brewer(palette = "GnBu") +
theme(axis.title = element_text(size = 6),
      title = element_text(size = 6),
      axis.text = element_text(size = 6),
      legend.title = element_text(size = 5),
      legend.text = element_text(size = 5),
      legend.key.size = unit(0.3, "cm"),
      legend.position = "bottom",
      strip.text = element_text(size = 6)) +
guides(fill = guide_legend(nrow = 2))

edu_stackplot <- ggplot(averaged_data_factor, aes(x = abst, fill = edu)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(x = "Abstinence Status",
       y = "Proportion",
       fill = "Education Level") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6)) +
  guides(fill = guide_legend(nrow = 2))

combined_plot_eduinc <- (wrap_elements(panel = income_stackplot + theme(legend.position = "bottom")) /
  wrap_elements(panel = edu_stackplot + theme(legend.position = "bottom"))) +
  plot_layout(ncol = 2, guides = 'collect') +
  plot_annotation(title = "Figure 1: Baseline Characteristics by Abstinence Status and Treatment Group",
                  theme = theme(plot.title = element_text(size = 8, hjust = 0.5)))

combined_plot_eduinc <- combined_plot_eduinc & theme(plot.margin = margin(10, 10, 10, 10),
  legend.position = c(0.5, 0.1))

combined_plot_eduinc
race_stackplot <- ggplot(averaged_data_factor, aes(x = abst, fill = race)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(x = "Abstinence Status",
       y = "Proportion",
       fill = "Race Group") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),

```

```

    legend.key.size = unit(0.3, "cm"),
    legend.position = "bottom",
    strip.text = element_text(size = 6))

only.menthol_stackplot <- ggplot(averaged_data_factor, aes(x = abst, fill = Only.Menthol)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(x = "Abstinence Status",
       y = "Proportion",
       fill = "Mentholated Cigarette User") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

combined_plot_racementhol <- (wrap_elements(panel = race_stackplot + theme(legend.position = "bottom")) +
  wrap_elements(panel = only.menthol_stackplot + theme(legend.position = "bottom")))
plot_layout(ncol = 2, guides = 'collect') +
plot_annotation(title = "Figure 2: Baseline Characteristics by Abstinence Status and Treatment Group",
  theme = theme(plot.title = element_text(size = 8, hjust = 0.5)))

combined_plot_racementhol <- combined_plot_racementhol & theme(plot.margin = margin(10, 10, 10, 10),
  legend.position = c(0.5, 0.1))

combined_plot_racementhol

ftcd_score_stackplot <- ggplot(averaged_data_factor, aes(x = ftcd_score, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(x = "FTCD Score",
       y = "Density",
       fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF")) +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

age_stackplot <- ggplot(averaged_data_factor, aes(x = age_ps, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(title = "",

```



```

    x = "Age",
    y = "Density",
    fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF")) +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

combined_plot_ftcdage <- (wrap_elements(panel = age_stackplot + theme(legend.position = "bottom")) /
  wrap_elements(panel = ftcd_score_stackplot + theme(legend.position = "bottom"))
  plot_layout(ncol = 2, guides = 'collect') +
  plot_annotation(title = "Figure 3: Baseline Characteristics by Abstinence Status and Treatment Group",
    theme = theme(plot.title = element_text(size = 8, hjust = 0.5)))

combined_plot_ftcdage <- combined_plot_ftcdage & theme(plot.margin = margin(10, 10, 10, 10),
  legend.position = c(0.5, 0.1))

combined_plot_ftcdage
NMR_stackplot <- ggplot(averaged_data_factor, aes(x = NMR, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(x = "NMR",
    y = "Density",
    fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF")) +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

bdi_stackplot <- ggplot(averaged_data_factor, aes(x = bdi_score_w00, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(x = "BDI Score",
    y = "Density",
    fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF")) +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),

```

```

    legend.text = element_text(size = 5),
    legend.key.size = unit(0.3, "cm"),
    legend.position = "bottom",
    strip.text = element_text(size = 6))

combined_plot_NMRbdi <- (wrap_elements(panel = NMR_stackplot + theme(legend.position = "bottom")) /
    wrap_elements(panel = bdi_stackplot + theme(legend.position = "bottom"))) +
    plot_layout(ncol = 2, guides = 'collect') +
    plot_annotation(title = "Figure 4: Baseline Characteristics by Abstinence Status and Treatment Group",
        theme = theme(plot.title = element_text(size = 8, hjust = 0.5)))

combined_plot_NMRbdi <- combined_plot_NMRbdi & theme(plot.margin = margin(10, 10, 10, 10),
    legend.position = c(0.5, 0.1))

combined_plot_NMRbdi
# create a correlation matrix among environmental condition factors
cor_matrix <- cor(averaged_data_factor[, c(5, 12, 14, 15, 16, 17, 18, 19, 23, 25)], use = "complete.obs")

# correlation plot of environmental condition factors
corrplot(cor_matrix, method = "color", type = "lower",
    tl.col = "black", tl.cex = 0.8, addCoef.col = "black",
    number.cex = 0.7, col = colorRampPalette(c("steelblue", "white", "steelblue"))(200))
title("Figure 3: Correlation Plot among Environmental Condition Characteristics",
    cex.main = 0.9, line = 3)

# Take transformation
averaged_data_factor_transformed <- averaged_data_factor
averaged_data_factor_transformed$shaps_score_pq1 <- asinh(averaged_data_factor$shaps_score_pq1)
averaged_data_factor_transformed$hedonsum_n_pq1 <- sqrt(averaged_data_factor$hedonsum_n_pq1)
averaged_data_factor_transformed$hedonsum_y_pq1 <- sqrt(averaged_data_factor$hedonsum_y_pq1)
averaged_data_factor_transformed$NMR <- log(averaged_data_factor$NMR)

skewness_df <- data.frame(Variable = c("hedonsum_n_pq1", "hedonsum_y_pq1", "shaps_score_pq1", "NMR"),
    transformation = c("Square Root Transformation",
        "Square Root Transformation",
        "Inverse Hyperbolic Sine Transformation",
        "Log Transformation"),
    skewness_before = c(skewness(averaged_data_factor$hedonsum_n_pq1),
        skewness(averaged_data_factor$hedonsum_y_pq1),
        skewness(averaged_data_factor$shaps_score_pq1, na.rm = TRUE),
        skewness(averaged_data_factor$NMR, na.rm = TRUE)),
    skewness_after = c(skewness(averaged_data_factor_transformed$hedonsum_n_pq1),
        skewness(averaged_data_factor_transformed$hedonsum_y_pq1),
        skewness(averaged_data_factor_transformed$shaps_score_pq1),
        skewness(averaged_data_factor_transformed$NMR, na.rm = TRUE)))

colnames(skewness_df) <- c("Variable", "Transformation",
    "Skewness before Transformation", "Skewness after Transformation")

skewness_df %>%
    kable(booktabs = TRUE, caption = "Variable Transformation on Skewness") %>%
    kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
    column_spec(1, width = "2cm") %>%
    column_spec(2, width = "4cm") %>%

```

```

column_spec(3, width = "3.5cm") %>%
column_spec(4, width = "3.5cm")
# create the summary table
summary_table <- averaged_data_factor %>%
  dplyr::select(-c("id", "Var", "BA", "Black", "Hisp", "NHW")) %>%
  tbl_summary(by = trt, label = list(abst ~ "Smoking abstinence",
                                     race ~ "Race",
                                     age_ps ~ "Age",
                                     sex_ps ~ "Sex",
                                     inc ~ "Income",
                                     edu ~ "Education",
                                     ftcd_score ~ "FTCD score",
                                     ftcd.5.mins ~ "Smoking within 5 mins of waking up",
                                     bdi_score_w00 ~ "BDI score",
                                     cpd_ps ~ "Cigarettes smoked per day",
                                     crv_total_pq1 ~ "Cigarette reward value",
                                     hedonsum_n_pq1 ~ "Pleasurable events (substitute reinforcers)",
                                     hedonsum_y_pq1 ~ "Pleasurable events (complementary reinforcers)",
                                     shaps_score_pq1 ~ "Anhedonia",
                                     otherdiag ~ "Other lifetime DSM-5 diagnosis",
                                     antidepressant ~ "Taking antidepressant",
                                     mde_curr ~ "Current vs. past MDD",
                                     NMR ~ "Nicotine metabolism ratio",
                                     Only.Menthol ~ "Exclusive mentholated cigarette user",
                                     readiness ~ "Readiness to quit smoking"),
              type = list(readiness ~ "continuous"),
              statistic = all_continuous() ~ "{mean} ({sd})",
              missing = "ifany",
              missing_text = "Missing") %>%
  add_overall(last = TRUE) %>%
  modify_spanning_header(update = all_stat_cols() ~ "**Behavioral and Pharmacological Treatment Assignment") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categorical") %>%
  bold_labels()

summary_table %>%
  as_kable_extra(booktabs = TRUE, caption = "Participant Characteristics by Treatment Arm",
                longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 7,
                            latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3.5cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  column_spec(6, width = "2cm") %>%
  row_spec(0, bold = TRUE, font_size = 7)
# Set working directory
# Windows
setwd("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Data/")

# Mac
# setwd("~/Desktop/Fall 2024/PHP 2550/Data/")

```

```

# Read in data
data <- read.csv("project2.csv")

# Factor categorical variables
data[, c("abst", "Var", "BA", "sex_ps", "NHW",
        "Black", "Hisp", "inc", "edu", "ftcd.5.mins",
        "otherdiag", "antidepmed", "mde_curr",
        "Only.Menthol")] <- lapply(data[, c("abst", "Var", "BA", "sex_ps", "NHW",
        "Black", "Hisp", "inc", "edu",
        "ftcd.5.mins", "otherdiag", "antidepmed",
        "mde_curr", "Only.Menthol")], as.factor)

# Multiple imputation with m = 5
new_data <- data[, -1]
imputed_data <- mice(new_data, m = 5, method = 'pmm', seed = 2550, printFlag = FALSE)

# Extract the five imputed datasets into a list
completed_datasets <- list()
for (i in 1:5) {
  completed_datasets[[i]] <- complete(imputed_data, i)
}

for (i in 1:length(completed_datasets)) {
  completed_datasets[[i]] <- completed_datasets[[i]] %>%
    mutate(trt = as.factor(case_when(Var == 1 & BA == 1 ~ "BASC + varenicline",
    Var == 0 & BA == 1 ~ "BASC + placebo",
    Var == 1 & BA == 0 ~ "ST + varenicline",
    Var == 0 & BA == 0 ~ "ST + placebo",
    TRUE ~ NA_character_)),
    race = as.factor(case_when(Black == 0 & Hisp == 0 & NHW == 0 ~ "Unknown",
    Black == 1 & Hisp == 1 & NHW == 1 ~ "Mixed Race",
    Black == 1 & Hisp == 1 ~ "Mixed Race",
    Black == 1 & NHW == 1 ~ "Mixed Race",
    NHW == 1 & Hisp == 1 ~ "Mixed Race",
    Black == 1 ~ "Black",
    Hisp == 1 ~ "Hispanic",
    NHW == 1 ~ "Non-Hispanic White",
    TRUE ~ "Other")),
    inc = fct_recode(as.factor(inc),
    "Less than $20,000" = "1",
    "$20,000-35,000" = "2",
    "$35,001-50,000" = "3",
    "$50,001-75,000" = "4",
    "More than $75,000" = "5"),
    edu = fct_collapse(as.factor(edu),
    "Some high school & Grade School" = c("1", "2"),
    "High school graduate or GED" = "3",
    "Some college/technical school" = "4",
    "College graduate" = "5")) %>%
  mutate(inc = fct_relevel(inc, "Less than $20,000", "$20,000-35,000",
    "$35,001-50,000", "$50,001-75,000", "More than $75,000"),
    edu = fct_relevel(edu, "Some high school & Grade School",
    "High school graduate or GED",

```

```

      "Some college/technical school", "College graduate")) %>%
mutate(trt = relevel(factor(trt), ref = "ST + placebo"))

# Apply transformations
completed_datasets[[i]]$shaps_score_pq1 <- asinh(completed_datasets[[i]]$shaps_score_pq1)
completed_datasets[[i]]$hedonsum_n_pq1 <- sqrt(completed_datasets[[i]]$hedonsum_n_pq1)
completed_datasets[[i]]$hedonsum_y_pq1 <- sqrt(completed_datasets[[i]]$hedonsum_y_pq1)
completed_datasets[[i]]$NMR <- log(completed_datasets[[i]]$NMR)
}
lasso_model_function_moderator <- function(data_list) {
  lasso_coef <- list()

  for (index in seq_along(data_list)) {
    data <- data_list[[index]]

    # Split train and test sets
    set.seed(2550)
    train_index <- createDataPartition(data$trt, p = 0.7, list = FALSE)
    train_data <- data[train_index, ]
    test_data <- data[-train_index, ]

    # Create fold IDs for cross-validation
    train_data$foldid <- NA
    for (trt_level in unique(train_data$trt)) {
      treatment_data <- train_data[train_data$trt == trt_level, ]
      fold_ids <- sample(rep(1:10, length.out = nrow(treatment_data)))
      train_data$foldid[train_data$trt == trt_level] <- fold_ids
    }

    # Define model matrix
    X <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
      bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_y_pq1 +
      shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Merit +
      readiness + race) +
      Var * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
      bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_y_pq1 +
      shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Merit +
      readiness + race), data = train_data)[, -1]

    y <- train_data$abst

    # Fit lasso with cross-validation using custom foldid
    cv_model <- cv.glmnet(X, y, family = "binomial", alpha = 1, nfolds = 10, foldid = train_data$foldid)
    best_lambda <- cv_model$lambda.min

    # Fit final lasso model using best lambda
    lasso_model <- glmnet(X, y, family = "binomial", alpha = 1, lambda = best_lambda)

    # Extract coefficients and store in a data frame
    coefficients <- as.data.frame(as.matrix(coef(lasso_model)))
    coefficients$Variable <- rownames(coefficients)
    rownames(coefficients) <- NULL
    colnames(coefficients)[1] <- "Estimates"
  }
}

```

```

    # Store coefficients in list
    lasso_coef[[index]] <- coefficients
  }

  # Return list of coefficients for all imputed datasets
  return(lasso_coef)
}

# Run the lasso model function
lasso_coef_results_moderator <- lasso_model_function_moderator(completed_datasets)
# Generate combined coefficient data frame
imputed_coefs_list_moderator <- list()
for (i in seq_along(lasso_coef_results_moderator)) {
  coefs <- lasso_coef_results_moderator[[i]]
  colnames(coefs)[colnames(coefs) == "Estimates"] <- paste0("Estimates_", i)
  imputed_coefs_list_moderator[[i]] <- coefs[, c("Variable", paste0("Estimates_", i))]
}

# Combine and pool estimates
wide_format_coefficients_moderator <- Reduce(function(x, y) merge(x, y, by = "Variable", all = TRUE), imputed_coefs_list_moderator)
wide_format_coefficients_moderator$Pooled_Estimate <- rowMeans(wide_format_coefficients_moderator[, colnames(wide_format_coefficients_moderator) != "Variable"])

coef_table_moderator <- wide_format_coefficients_moderator %>%
  filter(Pooled_Estimate != 0) %>%
  dplyr::select(c("Variable", "Pooled_Estimate"))

colnames(coef_table_moderator)[2] <- "Estimate"
coef_table_moderator$exp_estimate <- exp(coef_table_moderator$Estimate)
# Order main effects and interactions
main_effects <- coef_table_moderator[!grepl(":", coef_table_moderator$Variable), ]
interaction_terms <- coef_table_moderator[grepl(":", coef_table_moderator$Variable), ]
age_interaction <- coef_table_moderator[coef_table_moderator$Variable == "age_ps:Var1", ]

# Combine for final ordered result
ordered_model_results <- rbind(main_effects, interaction_terms)[-7,]
ordered_model_results <- rbind(ordered_model_results, age_interaction)
rownames(ordered_model_results) <- NULL
colnames(ordered_model_results)[3] <- "Exponential Estimate"

# Display results
ordered_model_results %>%
  kable(booktabs = TRUE, caption = "Lasso Model Coefficient Estimate") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down"))
long_data_train <- data.frame()
long_data_test <- data.frame()

# get stratified training index based on treatment group
set.seed(2550)
train_index <- createDataPartition(completed_datasets[[1]]$trt, p = 0.7, list = FALSE)

# generate long format of train and test dataframe from the five imputed datasets
for (i in 1:length(completed_datasets)) {
  imputed_dataset <- completed_datasets[[i]]

```

```

train_set <- imputed_dataset[train_index, ]
test_set <- imputed_dataset[-train_index, ]

long_data_train <- rbind(long_data_train, train_set)
long_data_test <- rbind(long_data_test, test_set)
}
# create the design matrix with interaction terms
long_data_matrix_train <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
  bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_y
  shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Mer
  readiness + race) +
  Var * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
  bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_y
  shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Mer
  readiness + race),
  data = long_data_train)

# convert the design matrix to a data frame
long_data_trainset <- as.data.frame(long_data_matrix_train)

# extract the intercept from pooled coefficients
pooled_intercept <- wide_format_coefficients_moderator %>%
  filter(Variable == "(Intercept)") %>%
  pull(Pooled_Estimate)

# extract only non-intercept pooled coefficients
pooled_coefs <- wide_format_coefficients_moderator %>%
  filter(Variable != "(Intercept)")

# ensure the predictor variables in the data match those in pooled coefficients
predictor_vars <- pooled_coefs$Variable
long_data_trainset <- long_data_trainset[, predictor_vars, drop = FALSE]

# calculate log-odds using matrix multiplication with pooled coefficients
long_data_trainset$log_odds <- pooled_intercept + as.matrix(long_data_trainset) %*% pooled_coefs$Pooled_Estimate

# convert log-odds to probabilities
long_data_trainset$predicted_prob <- 1 / (1 + exp(-long_data_trainset$log_odds))
# create the design matrix with interaction terms
long_data_matrix_test <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
  bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_y
  shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Mer
  readiness + race) +
  Var * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
  bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_y
  shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Mer
  readiness + race),
  data = long_data_test)

# convert the design matrix to a data frame
long_data_testset <- as.data.frame(long_data_matrix_test)

# ensure the predictor variables in the data match those in pooled coefficients

```



```

long_data_testset <- long_data_testset[, predictor_vars, drop = FALSE]

# calculate log-odds using matrix multiplication with pooled coefficients
long_data_testset$log_odds <- pooled_intercept + as.matrix(long_data_testset) %*% pooled_coefs$Pooled_E

# convert log-odds to probabilities
long_data_testset$predicted_prob <- 1 / (1 + exp(-long_data_testset$log_odds))
# do roc on train and test sets
auc_result <- roc(long_data_train$abst, long_data_trainset$predicted_prob)
auc_result_test <- roc(long_data_test$abst, long_data_testset$predicted_prob)

# plot roc for both sets
par(mfrow= c(1,2), oma = c(0, 0, 2, 0))
plot(auc_result, main = "Train Data", font.main = 1, cex.main = 0.8, cex.lab = 0.8)
text(0.3, 0.2, paste("AUC =", round(auc(auc_result), 3)), col = "blue", cex = 0.7)

plot(auc_result_test, main = "Test Data", font.main = 1, cex.main = 0.8, cex.lab = 0.8)
text(0.3, 0.2, paste("AUC =", round(auc(auc_result_test), 3)), col = "blue", cex = 0.7)
long_data_trainset <- long_data_trainset %>%
  mutate(abst_num = as.numeric(as.character(long_data_train$abst)))
long_data_testset <- long_data_testset %>%
  mutate(abst_num = as.numeric(as.character(long_data_test$abst)))

cal_plot_train <- calibration_plot(data = long_data_trainset, obs = "abst_num", pred = "predicted_prob",
cal_plot_test <- calibration_plot(data = long_data_testset, obs = "abst_num", pred = "predicted_prob",

grid.arrange(cal_plot_train$calibration_plot,
              cal_plot_test$calibration_plot, ncol = 2,
              top = text_grob("Figure 4: Calibration Plot Comparison"))

```

## Reference

1. Santomauro, D. F., Herrera, A. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Vos, T., Whiteford, H., Ferrari, A. J., Charlson, F. J., et al. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, 398(10312), 1700–1712.
2. Weinberger, A. H., Chaiton, M. O., Zhu, J., Wall, M. M., Hasin, D. S., & Goodwin, R. D. (2020). Trends in the prevalence of current, daily, and nondaily cigarette smoking and quit ratios by depression status in the u.s.: 2005–2017. *American Journal of Preventive Medicine*, 58(5), 691–698.
3. Hitsman, B., Papandonatos, G. D., McChargue, D. E., & al., et. (2013). Past major depression and smoking cessation outcome: A systematic review and meta-analysis update. *Addiction*, 108(2), 294–306.