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[1]. 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[2]. However, most smoking cessation clinical trials have excluded this important group from trial enrollment[3], limiting information and suggestion for this group who wants to quit smoking.

In a recent 2 × 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 x 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. 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 generate a new treatment variables to capture the four distinct intervention groups formed by the 2x2 factorial design, including `ST + placebo`, `ST + varenicline`, `BASC + placebo`, and `BASC + varenicline.` This new treatment variable was set to reference `ST + placebo` for comparison among groups. Additionally, 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

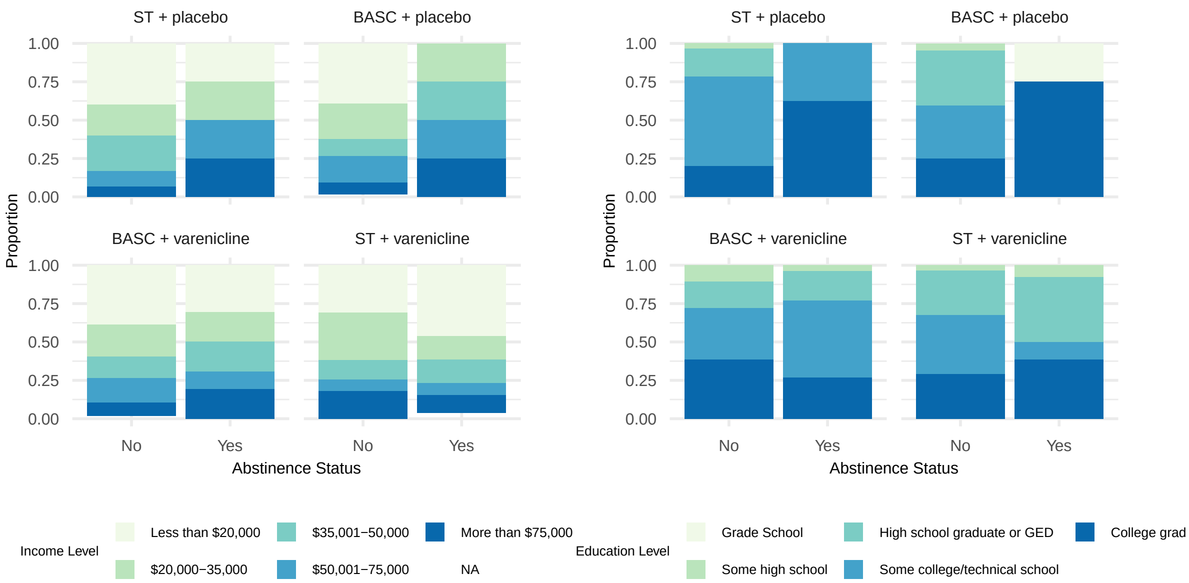
Variable	Missing Count	Missing Percentage
NMR	21	7 %
crv_total_pq1	18	6 %
readiness	17	5.67 %
inc	3	1 %
shaps_score_pq1	3	1 %
Only.Menthol	2	0.67 %
ftcd_score	1	0.33 %

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 differences among groups and abstinence outcomes. For example, participants with income less than $20,000 are less likely to stop smoking in both the `ST + placebo` and `BASC + varenicline` groups that more people would still continue to smoke at the week 27 follow-up. Even in the `BASC + placebo` group, no patients stop smoking at week 27. However, this pattern reverses in the `ST + varenicline` group that lower-income individuals show a relatively higher likelihood of smoking abstinence. That is, the combination of standard treatment with varenicline may have a greater impact on smoking cessation for lower-income participants. This reversal pattern suggests that income level might be a potential moderator of the behavioral treatment effectiveness on the EOT abstinence among people with MDD.

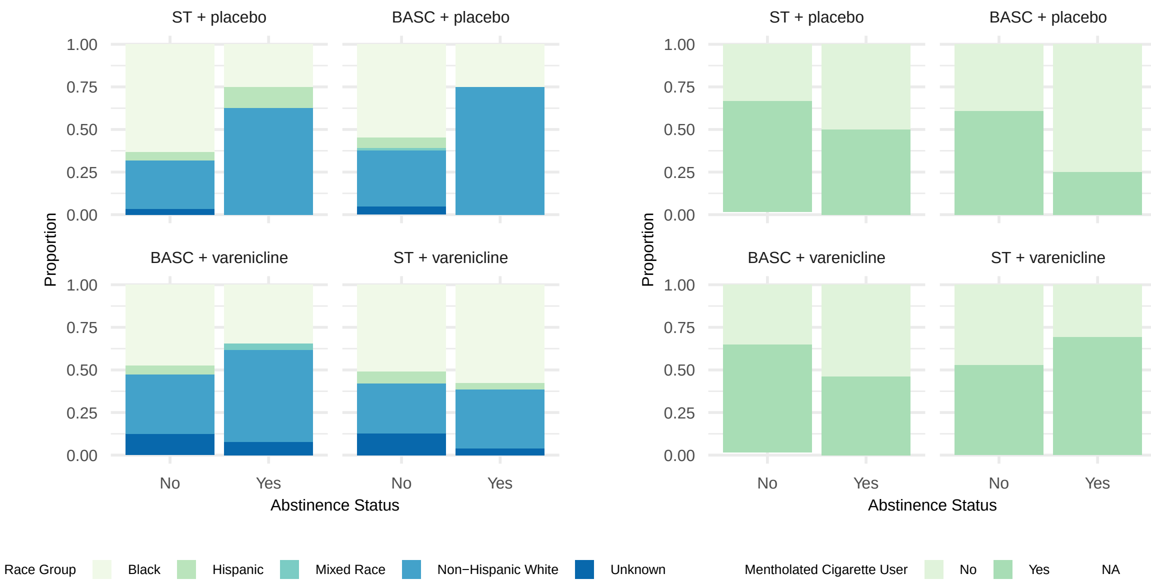
Figure 1: Baseline Characteristics by Abstinence Status and Treatment Group (Categorical 1)



Similar for education level presented in `Figure 2`, people with lower education level (grade school, some high school, or high school graduate) shows less probability of smoking cessation, especially in the `ST + placebo`, `BASC + placebo`, and `BASC + varenicline` group, suggesting the potential association between education level and smoking cessation after treatment.However, participants in the `ST + varenicline` groups show relatively higher likelihood of abstinence, suggesting potential interacting relationship between education level and treatment assignment on the abstinence. In addition, college graduated participants in the two placebo groups show higher probability of abstinence while those participants in the two varenicline groups show reversed pattern as well, 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 instance, black people generally face greater challenges of smoking cessation as they continuously exhibit lower abstinence rate across different treatment groups, particularly in the `ST + placebo` and `BASC + placebo` groups. However, in the `BASC + varenicline` group, the difference between proportion of black participants who continue or stop smoking becomes less pronounced. Even in the `ST + varenicline` group, black participants exhibit higher abstinence rates, suggesting that varenicline may be particularly effective to mitigate challenges in cessation among black individuals. Also for `Only.Menthol`, among the first three treatment groups, mentholated cigarette users show lower likelihood to achieve smoking cessation and non-mentholated cigarette users exhibit higher probability to stop smoking. This pattern is reversed in the `ST + varenicline` group again. 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.

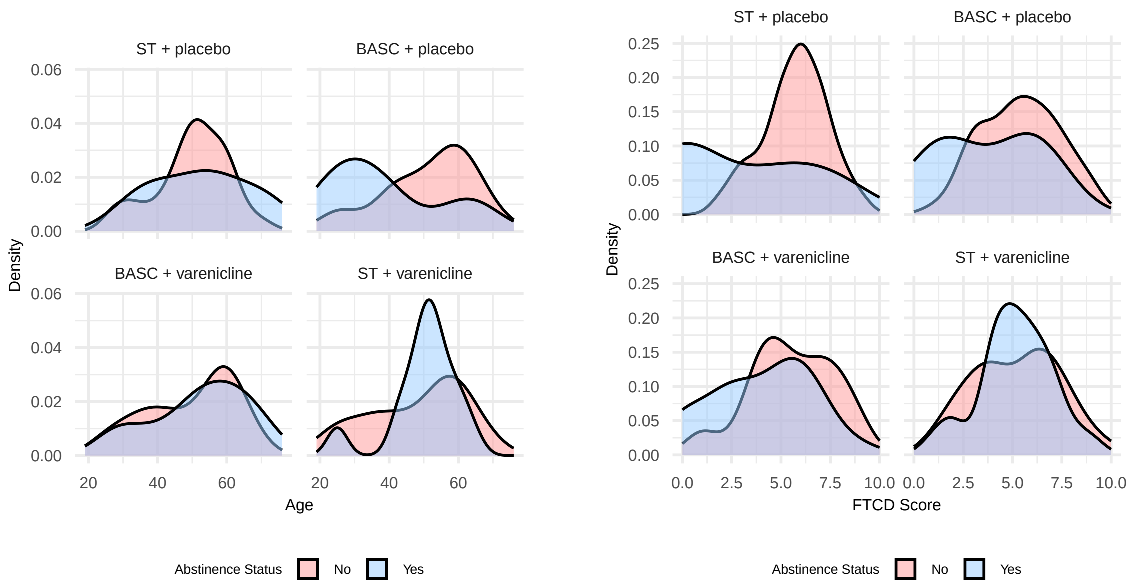
Figure 2: Baseline Characteristics by Abstinence Status and Treatment Group (Categorical 2)



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.

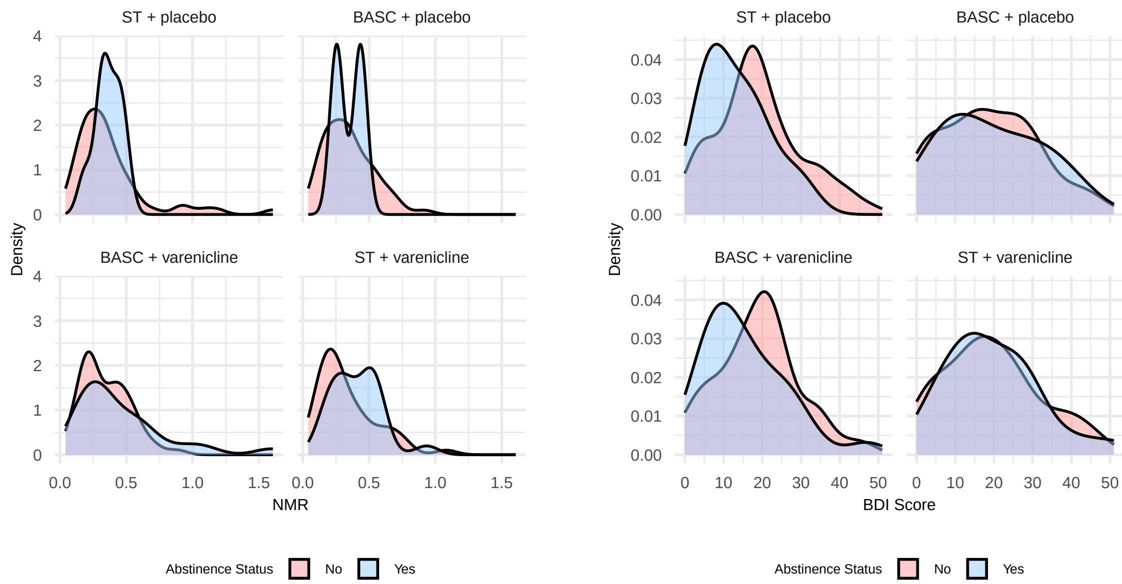
Seeing `Figure 3`, higher-middle age people (50-60 years) show lower abstinence rate in the `ST + placebo`, `BASC + placebo`, and `BASC + varenicline` groups while they exhibit higher abstinence rate in the `ST + varenicline` treatment. This indicates that the combination of varenicline and ST might be relatively more effective for helping older individuals quit smoking. Moreover, younger participants present higher abstinence rate in the `BASC + placebo` group while show more challenges of stop smoking with the other treatment. In addition, participants with more nicotine dependence (higher FTCD score) exhibit a significantly higher non-abstinence rate in the `ST + placebo` group while this pattern is mitigated in the other three treatment groups. Even in the `ST + varenicline` group, they show much larger abstinence rate at week 27 follow-up. Moreover, participants with relatively low FTCD score present lower abstinence rate in the `ST + varenicline` group while show much higher likelihood of achieving smoking cessation in the other three treatment groups. These findings suggest a potential age-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)



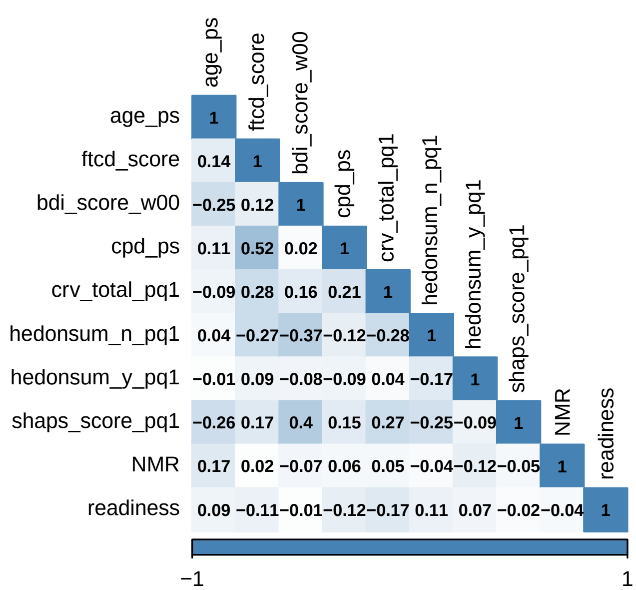
Seeing `Figure 4`, the distribution of NMR is skewed towards participants with higher nicotine metabolism ratio (NMR) across all groups. In the two placebo groups, participants with lower NMR are more likely to quit smoking at week 27 follow-up while this gap is less pronounced in the two varenicline groups. Moreover, participants with lower BDI scores (less severe depressive symptoms) are more likely to stop smoking in the `ST + placebo` and `BASC + varenicline` groups while people in these two groups with middle level BDI scores are less likely to stop smoking. However, in the `BASC + placebo` and `ST + varenicline` groups, 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. These findings suggest that both NMR and BDI scores may interact with 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



	age_ps	ftcd_score	bdi_score_w00	cpd_ps	crv_total_pq1	hedonsum_n_pq1	hedonsum_y_pq1	shaps_score_pq1	NMR	readiness
age_ps	1									
ftcd_score	0.14	1								
bdi_score_w00	−0.25	0.12	1							
cpd_ps	0.11	0.52	0.02	1						
crv_total_pq1	−0.09	0.28	0.16	0.21	1					
hedonsum_n_pq1	0.04	−0.27	−0.37	−0.12	−0.28	1				
hedonsum_y_pq1	−0.01	0.09	−0.08	−0.09	0.04	−0.17	1			
shaps_score_pq1	−0.26	0.17	0.4	0.15	0.27	−0.25	−0.09	1		
NMR	0.17	0.02	−0.07	0.06	0.05	−0.04	−0.12	−0.05	1	
readiness	0.09	−0.11	−0.01	−0.12	−0.17	0.11	0.07	−0.02	−0.04	1

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
hedonsum_n_pq1	Square Root Transformation	1.338843	-0.0591728
hedonsum_y_pq1	Square Root Transformation	1.391398	0.0620129
shaps_score_pq1	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 2×2, 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 obervations 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

	Behavioral and Pharmacological Treatment Assignment				
Characteristic	**ST + placebo, N = 68**	**BASC + placebo, N = 68**	**BASC + varenicline, N = 83**	**ST + varenicline, N = 81**	**Overall, N = 300**
Smoking abstinence	8 (12%)	4 (5.9%)	26 (31%)	26 (32%)	64 (21%)
Age	50 (11)	51 (14)	50 (13)	49 (13)	50 (13)
Sex					
Male	29 (43%)	30 (44%)	39 (47%)	37 (46%)	135 (45%)
Female	39 (57%)	38 (56%)	44 (53%)	44 (54%)	165 (55%)
Income					
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
Education					
Grade School	0 (0%)	1 (1.5%)	0 (0%)	0 (0%)	1 (0.3%)
Some high school	2 (2.9%)	3 (4.4%)	7 (8.4%)	4 (4.9%)	16 (5.3%)
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%)
FTCD score	5 (2)	5 (2)	5 (2)	5 (2)	5 (2)
Missing	1	0	0	0	1
Smoking within 5 mins of waking up	35 (51%)	32 (47%)	33 (40%)	38 (47%)	138 (46%)
BDI score	18 (11)	19 (12)	18 (11)	20 (12)	19 (11)
Cigarettes smoked per day	15 (7)	16 (9)	16 (9)	14 (7)	15 (8)
Cigarette reward value	7 (4)	7 (4)	7 (4)	7 (3)	7 (4)
Missing	8	1	3	6	18
Pleasurable events (substitute reinforcers)	21 (20)	23 (20)	23 (19)	23 (19)	23 (20)
Pleasurable events (complementary reinforcers)	27 (20)	28 (22)	22 (17)	25 (19)	25 (19)
Anhedonia	3 (3)	2 (3)	2 (3)	2 (3)	2 (3)
Missing	1	2	0	0	3
Other lifetime DSM-5 diagnosis	28 (41%)	35 (51%)	30 (36%)	40 (49%)	133 (44%)
Taking antidepressant	15 (22%)	28 (41%)	24 (29%)	15 (19%)	82 (27%)
Current vs. past MDD					
Past MDD	37 (54%)	36 (53%)	43 (52%)	37 (46%)	153 (51%)
Current MDD	31 (46%)	32 (47%)	40 (48%)	44 (54%)	147 (49%)
Nicotine metabolism ratio	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
Exclusive mentholated cigarette user	43 (64%)	40 (59%)	48 (59%)	47 (58%)	178 (60%)
Missing	1	0	1	0	2
Readiness to quit smoking	7 (1)	7 (1)	7 (1)	7 (1)	7 (1)
Missing	4	4	5	4	17
Race					
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%)

[1] 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 treatment. 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. 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` and `Table 5`.

Table 4: Main Effect Estimates

Variable	Estimate
(Intercept)	-0.4327118
crv_total_pq1	0.0004343
eduSome college/technical school	-0.0224811
ftcd_score	-0.1479283
mde_curr1	-0.2546973
NMR	0.3533395
raceNon-Hispanic White	0.3370103
readiness	-0.0205558

Table 5: Interaction Estimates

Treatment	Variable	Estimate
trtST + varenicline	Only.Menthol1	0.4910082
trtBASC + placebo	age_ps	-0.0046514
trtBASC + varenicline	age_ps	0.0024278
trtBASC + varenicline	antidepmed1	0.1200809
trtST + varenicline	bdi_score_w00	0.0002893
trtST + varenicline	crv_total_pq1	0.0707485
trtBASC + placebo	eduHigh school graduate or GED	-0.1269962
trtST + varenicline	eduHigh school graduate or GED	0.1735452
trtBASC + placebo	eduSome college/technical school	-0.2178426
trtBASC + varenicline	eduSome college/technical school	0.6996237
trtST + varenicline	eduSome college/technical school	-0.4873256
trtST + varenicline	ftcd.5.mins1	0.2790025
trtST + varenicline	ftcd_score	0.0044188
trtBASC + varenicline	inc$35,001-50,000	0.0026236
trtBASC + varenicline	incMore than $75,000	0.1163245
trtST + varenicline	incMore than $75,000	-0.4187670
trtBASC + varenicline	raceMixed Race	0.9585245
trtBASC + varenicline	raceNon-Hispanic White	0.4503014
trtBASC + varenicline	raceUnknown	0.0828347
trtST + varenicline	raceUnknown	-0.1966441

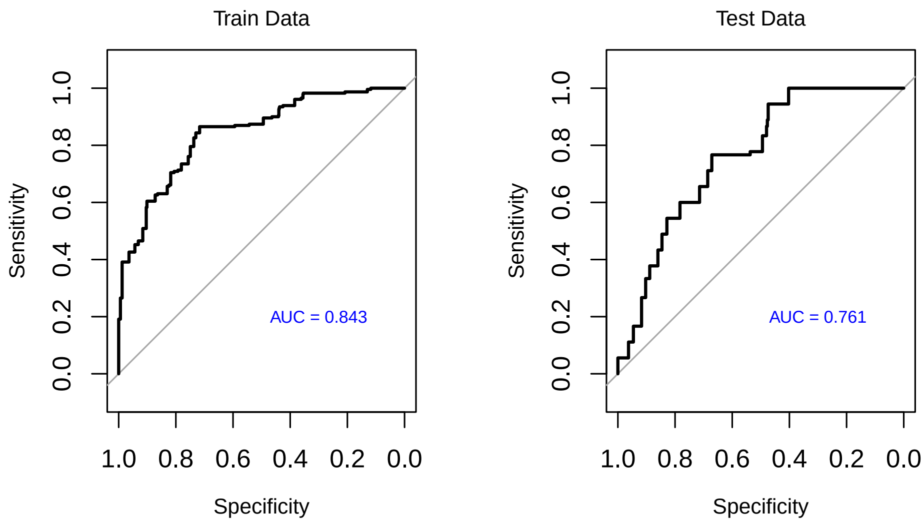
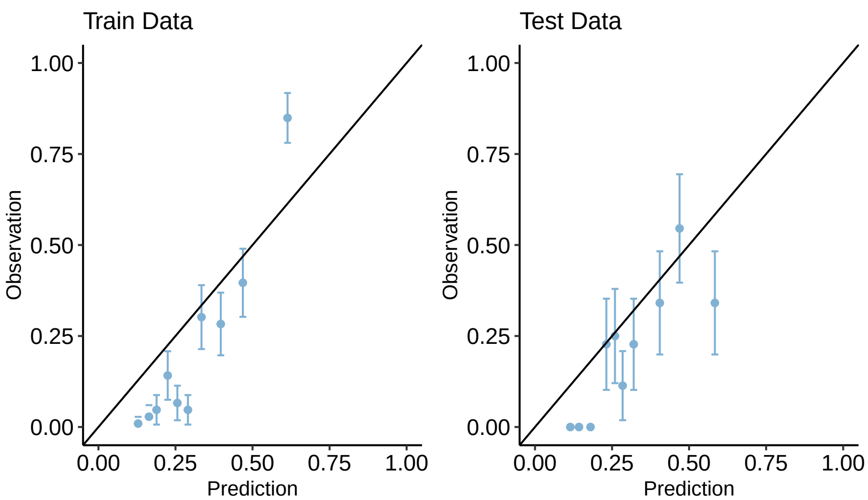


Figure 4: Calibration Plot Comparison



Discussion

Appendix

```r
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",
```

```r
                          "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"))
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),
```

```r
        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 = "bott
  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() +
```

```r
  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 = "botto
  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$shaps_score_pq1, na.rm = TRU
                                              skewness(averaged_data_factor$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")) %>%
```

```r
  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",
                                     antidepmed ~ "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 Assignme
  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",
```

```r
          "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)

# generate and recode necessary columns
new_data <- data %>%
  mutate(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_)),
         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_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"))

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

# relevel inc and edu to make them ordinal with correct level
new_data <- new_data %>%
  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"))

new_data$edu <- recode(new_data$edu, "Grade School" = "Some high school & Grade School")
new_data$edu <- recode(new_data$edu, "Some high school" = "Some high school & Grade School")
# multiple imputation with m = 5
imputed_data <- mice(new_data, m = 5, method = 'pmm', maxit = 50, seed = 2550, printFlag = FALSE)

# extract the five imputed datasets to a data list
completed_datasets <- list()
for (i in 1:5) {
```

```r
  completed_datasets[[i]] <- complete(imputed_data, i)
}

for (i in 1:length(completed_datasets)) {
  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
lasso_model_function <- function(data_list) {
  lasso_coef <- list()

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

    # split train and test sets
    set.seed(2550)
    train_index <- createDataPartition(new_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 ~ trt * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
                                    bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    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, foldid = train_data$foldid)
    best_lambda <- cv_model$lambda.min

    # fit the final lasso model using the 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"
    coefficients <- coefficients[, c("Variable", "Estimates", setdiff(names(coefficients), c("Estimates

    # store coef results in list
```

```
    lasso_coef[[index]] <- coefficients
  }

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

# run the lasso model function on the list of imputed datasets
lasso_coef_results <- lasso_model_function(completed_datasets)
# generate a coefficient data frame extracting from five lasso models
imputed_coefs_list <- list()

for (i in seq_along(lasso_coef_results)) {
  coefs <- lasso_coef_results[[i]]
  colnames(coefs)[colnames(coefs) == "Estimates"] <- paste0("Estimates_", i)
  imputed_coefs_list[[i]] <- coefs[, c("Variable", paste0("Estimates_", i))]
}

# combine all imputed datasets' coefficients by column and calculate pooled estimates
wide_format_coefficients <- Reduce(function(x, y) merge(x, y, by = "Variable", all = TRUE), imputed_coe
wide_format_coefficients$Pooled_Estimate <- rowMeans(
  wide_format_coefficients[ , grep("Estimates_", names(wide_format_coefficients))],
  na.rm = TRUE)

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

colnames(coef_table)[2] <- "Estimate"

coef_table_maineffect <- coef_table[1:8, ]
coef_table_interaction <- coef_table[9:28, ]

coef_table_interaction <- coef_table_interaction %>%
  separate(Variable, into = c("Treatment", "Variable"), sep = ":", remove = FALSE) %>%
  arrange(Variable, Treatment) %>%
  dplyr::select(Treatment, Variable, Estimate)

coef_table_maineffect %>%
  kable(booktabs = TRUE, caption = "Main Effect Estimates") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down"))

coef_table_interaction %>%
  kable(booktabs = TRUE, caption = "Interaction Estimates") %>%
  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(new_data$trt, p = 0.7, list = FALSE)

# generate long format of train and test dataframe from the five imputed datasets
```

```r
for (i in seq_len(imputed_data$m)) {
  imputed_dataset <- complete(imputed_data, 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 ~ trt * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.m
                                                  bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq
                                                  hedonsum_y_pq1 + shaps_score_pq1 + otherdiag + antidepr
                                                  mde_curr + NMR + Only.Menthol + 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 %>%
  filter(Variable == "(Intercept)") %>%
  pull(Pooled_Estimate)

# extract only non-intercept pooled coefficients
pooled_coefs <- wide_format_coefficients %>%
  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_

# 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 ~ trt * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mi
                                                  bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq
                                                  hedonsum_y_pq1 + shaps_score_pq1 + otherdiag + antidepr
                                                  mde_curr + NMR + Only.Menthol + 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_Es

# convert log-odds to probabilities
```

```r
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.