Moderators of Behavioral Activation for Smoking Cessation in Individuals with Major Depressive Disorder

Yingqiu Huang

Abstract

Background: There have been many calls for greater attention to the treatment of tobacco dependence in individuals with major depressive disorder (MDD). Addressing depression-related psychological aspects linked to smoking behavior in individuals with MDD may enhance their chances of quitting smoking. This project aims to examine baseline variables as potential moderators of the effects of Behavioral Activation for Smoking Cessation (BABC) among individuals with current or past MDD, as well as to identify baseline predictors of abstinence while accounting for behavioral treatment and pharmacotherapy.

Methods: This project utilizes data from a randomized, placebo-controlled trial that was conducted on individuals with current or past MDD who received 12 weeks of either Behavioral Activation for Smoking Cessation (BASC) or standard treatment (ST) and either varenicline or placebo. The trial included 300 daily smokers with a diagnosis of MDD. Multiple imputation was applied to address missing data, and a Lasso algorithm was used for variable selection, with 30% of the observations set aside for validation. Pooled estimates were used to calculate predictions and evaluate the model's performance in terms of discrimination and calibration.

Results: FTCD Score and Current Major Depressive Disorder (MDD) were found to be important moderators of the effects of BA for smoking cessation. Age, sex, cigarette reward value, income, education, exclusive mentholated cigarette use, pleasurable events scale of complementary reinforcers, and non-hispanic white indicator are valuable predictors of abstinence.

Conclusion: The Lasso regression model successfully identified moderators and predictors with moderately good discrimination power. However, given the small sample size and imbalanced class, future work needs to be conducted for enhanced generalizability.

Introduction

Individuals with Major Depressive Disorder (MDD) have psychological factors that increase their dependence on smoking and make quitting more challenging, such as reward impairment and greater withdrawal severity. Varenicline is an effective drug that helps people quit smoking by blocking the effect of nicotine on the brain. However, other treatment of depression-related psychological factors linked with smoking behavior might also improve the rate of smoking cessation among adults with MDD.

A previous study investigated the efficacy and safety of combining Behavioral Activation for Smoking Cessation (BASC) with the pharmacotherapy varenicline among individuals with MDD, using a randomized, placebo-controlled, 2 × 2 factorial design. Participants who were daily smokers with current or past MDD were assigned to one of the four treatment groups: BASC with varenicline, BASC with placebo, standard treatment (ST) with varenicline, or ST with placebo. The primary outcome of this study is rate of smoking abstinence at 27 weeks. The results indicated that varenicline significantly improved smoking cessation rates compared to placebo, but there is no significant differences between BASC and ST.

Given the surprising results of the study, which found no significant difference between Behavioral Activation (BA) and standard treatment (ST) for smoking cessation, there may be underlying factors worth exploring. This finding suggests the possibility that specific baseline characteristics or psychological factors may moderate the effectiveness of BA. Thus, this project aims to investigate these potential moderators to better understand the effect of BASC and to identify baseline factors that predicts abstinence, while accounting for the use of pharmacotherapy.

Methods

Data Description

The dataset for this project consists of 300 observations across 25 variables. The primary outcome is a binary indicator of abstinence (`abst`), coded as 0 or 1. Treatment assignments are represented by two binary indicators: `BA` for Behavioral Activation and `Var` for Varenicline. Socioeconomic and demographic variables include: Age at phone interview (`age_ps`), Sex at phone interview (`sex_ps`), Non-Hispanic White indicator (`NHW`), Black indicator (`Black`), Hispanic indicator (`Hisp`), Income (`inc` = ordinal categorical, low to high) and Education (`edu` = ordinal categorical, low to high). Additional baseline characteristics include measures like FTCD score at baseline (`ftcd_score`), cigarette reward value at baseline (`crv_total_pq1`), and indicator for other lifetime DSM-5 diagnosis (`otherdiag`). More infomation on the variables can be found in Table 2.

Data Pre-Processing

For the purpose of Exploratory Data Analysis (EDA), we created a new variable `treatment` to categorize participants into the four treatment groups: BASC with varenicline, BASC with placebo, standard treatment (ST) with varenicline, or ST with placebo. This allows us to generate plots stratified by treatment group and explore potential interaction terms for inclusion in the Lasso regression model. We also consolidated race information by creating a new `race` variable that combines the existing race indicators (which were previously separate columns) and adds a category for mixed race while also accounting for missing race information. The new race variable includes the categories: Non-Hispanic White (NHW), Black, Hispanic, Mixed race and Unknown. The original race indicator columns were removed after creating this consolidated column. For model fitting purposes, categorical and binary columns were converted to factors, with income and education levels set as ordinal factors given their five levels.

Given that regression models will be used in subsequent analyses, we examined the distribution of continuous variables to assess normality and ensure they meet regression assumptions. During this process, we identified that the Nicotine Metabolism Ratio (NMR) variable was skewed. To address this, we applied a log transformation to NMR. The histograms below show the distribution of NMR before and after transformation.

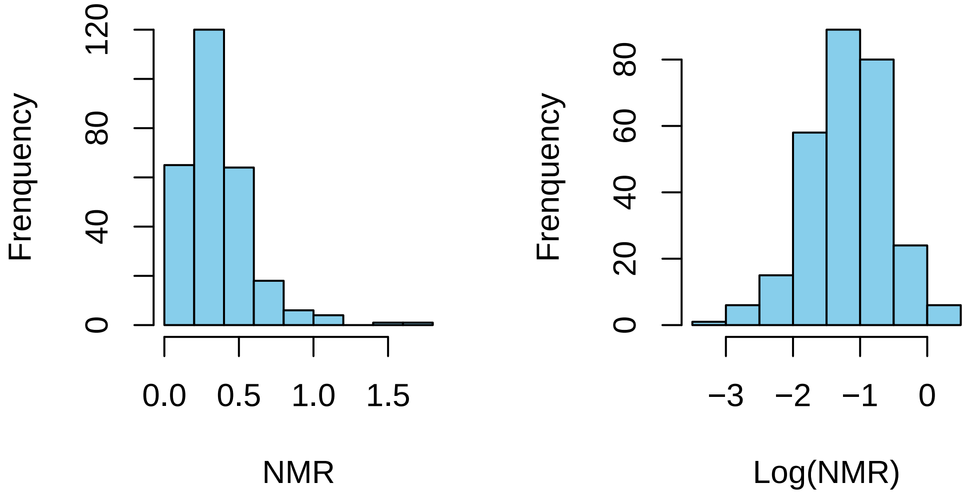

Figure 1: Histogram of Log Transformation of Nicotine Metabolism Ratio

Variable Selection

The rationale for conducting a variable selection process prior to applying Lasso includes several key considerations. First, Lasso can select from a pool of interaction terms, but it does not automatically generate these terms. Therefore, it is necessary to specify the interaction terms to include in the model matrix before running Lasso. This ensures that Lasso has the option to choose relevant interactions, rather than being restricted to main effects alone. Second, variable selection helps to avoid multicollinearity by identifying and eliminating highly correlated variables, ensuring that only one variable from a correlated pair is retained. Third, including all possible main effects and interaction terms in the Lasso model can lead to overfitting and a computationally intensive process. By pre-selecting interactions based on observed relationships, we limit the model to those terms that have preliminary support, thus improving model interpretability and prevents overfitting.

The correlation heatmap below illustrates the pairwise correlations among variables in the dataset. The strongest associations are observed between `ftcd_score` (FTCD score at baseline), `cpd_ps` (cigarettes per day at baseline phone survey), and `ftcd.5.mins` (smoking within 5 minutes of waking up). The FTCD score serves as a measure of nicotine dependence and is calculated using metrics that include both `ftcd.5.mins` and `cpd_ps`. Given the theoretical basis of their interrelationships and the observed high correlations, we opted to retain only `ftcd_score` and exclude `cpd_ps` and `ftcd.5.mins` to minimize multicollinearity.

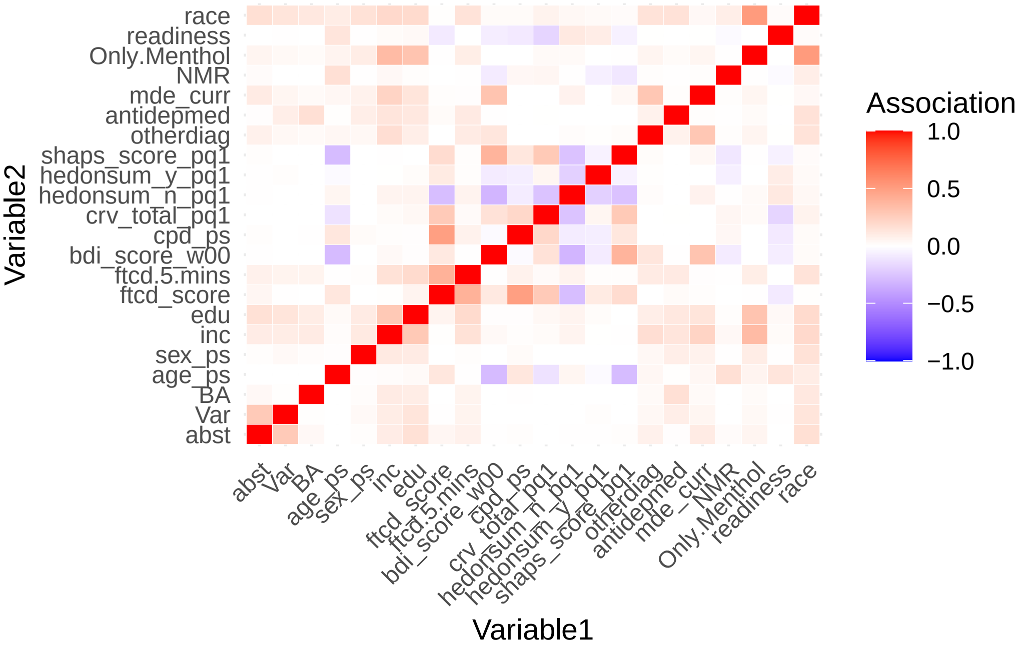

Figure 2: Correlation Heatmap

To select appropriate interaction terms for Lasso, we considered the primary aim of this project: to investigate baseline variables as potential moderators of the effects of behavioral treatment on abstinence. This means including interaction terms between `BA` (Behavioral Activation) and all other covariates, encompassing both socioeconomic/demographic and baseline health-related variables. This need is further supported by the following exploratory data analysis (EDA) plot, which illustrates a potential interaction effect between current vs. past MDD status and treatment on abstinence. The plot suggests that abstinence proportions vary across treatment groups depending on MDD status, indicating a potential moderating effect of MDD status on treatment outcomes.

In addition to interactions with `BA`, we decided to include interaction terms between `Var` (Varenicline) and all covariates, as pharmacotherapy serves as a secondary treatment that could enhance the model's predictive power.

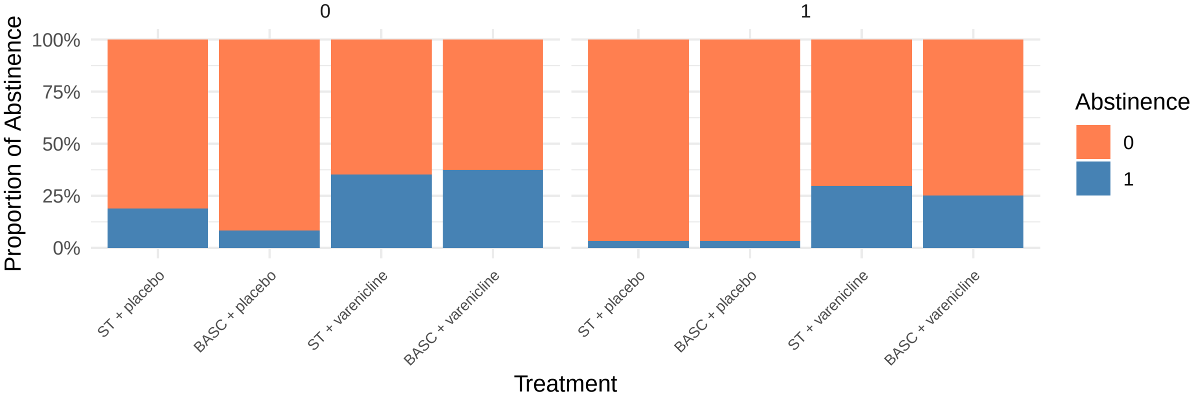

Figure 3: Interaction Between MDD Status and Treatment on Abstinence

Other than these two sets of interaction terms, we also included interactions between socioeconomic/demographic factors and other health related baseline characteristics in the Lasso model as well as main effects. Theoretically, variables such as age, income, and education often interact with health-related factors, influencing access to resources, coping mechanisms, and health outcomes. Empirically, the EDA plot shows that abstinence rates vary with education levels more evidently among individuals without MDD than those with MDD, indicating that education's impact on abstinence may differ by MDD status.

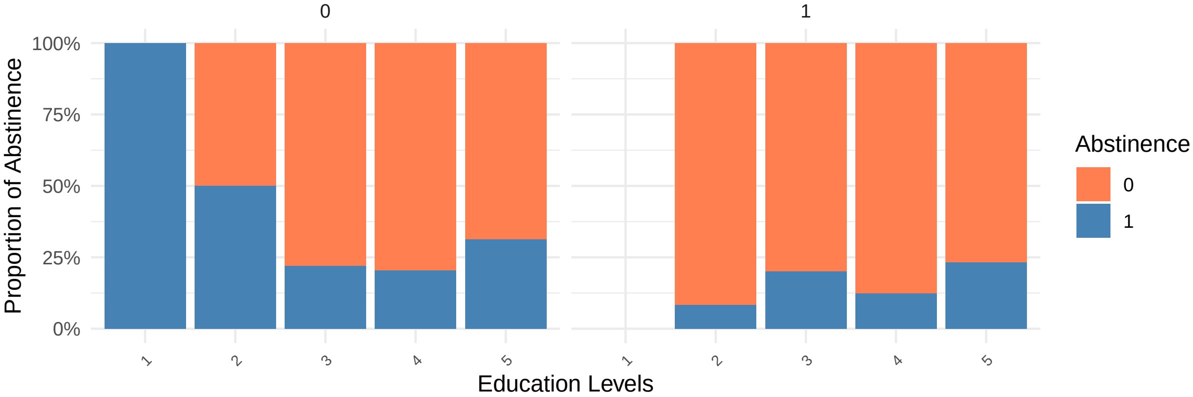

Figure 4: Interaction Between Education and MDD Status on Abstinence

Missing Data

The table below presents the patterns of missingness in the dataset. Excluding observations with missing data would result in a 20% reduction in the dataset. To address this, we applied multiple imputation using the `mice` package in R. This algorithm iteratively imputes missing values for each variable based on the observed distributions of other variables. We generated five imputed datasets using the predictive mean matching method. The imputation model did not include the `treatment` variable created for EDA purposes; instead, it utilized the original `BA` and `Var` variables, the newly created `race` variable, and other baseline covariates.

Table 1: Percentage of Missing Data

Variable	Percentage Missing
Nicotine Metabolism Ratio (NMR)	7.00
Cigarette reward value at baseline (crv_total_pq1)	6.00
Baseline readiness to quit smoking (readiness)	5.67
Income	1.00
Anhedonia (shaps_score_pq1)	1.00
Exclusive Mentholated Cigarette User (Only.Menthol)	0.67

Lasso Regression Model

Lasso regression is a technique that combines variable selection and regularization to prevent overfitting and improve both the predictive accuracy and interpretability of the model. The primary rationale for using a Lasso model in this project is to identify important baseline covariates that may act as moderators of the effect of Behavioral Activation (BA) on abstinence. Lasso achieves this by penalizing the coefficients of less relevant variables to zero, while keeping coefficients for the most influential variables. This approach is appropriate for handling the large number of main effects and interaction terms we have, creating a more interpretable model that highlights the most important factors.

Based on preliminary variable selection, we identified three sets of interaction terms to include in the Lasso model: interactions between BA and all other covariates, interactions between Varenicline and all other covariates, and interactions between demographic/socioeconomic factors (sex, income, education, race, age) and other health-related variables (excluding `BA` and `Var`).

The dataset was split into a training set (70%) and a validation set (30%) using stratified sampling based on the `treatment` variable to ensure proportional representation of each treatment group in both sets. We applied the Lasso model to each of the five imputed training datasets separately using the `glmnet` package in R. To address the imbalance between outcome groups (`abst = 1` and `abst = 0`), we applied more weight on the `abst = 1` group. The weight is calculated as the `abst = 0` population size divided by the `abst = 0` population size.

For each Lasso run on the imputed datasets, we used 10-fold cross-validation to identify the optimal λ value, then refitted the model with the best λ. We recorded the coefficient estimates from each of the five Lasso runs.

To obtain pooled estimates, we retained only those variables with non-zero coefficients in at least 3 out of the 5 imputations, providing a threshold for consistent variable selection. We then averaged the coefficients of these selected variables across the five models to derive the pooled estimates.

Finally, we used these pooled coefficients to calculate predictions by performing matrix multiplication with the long format of the dataset (concatenating the five imputed datasets). This allowed us to generate predictions on both the training and validation sets in their long formats, effectively averaging the results of the five Lasso models to obtain a single prediction outcome across the five imputed datasets.

Evaluation Metric

The performance of the developed model, based on the pooled estimates, was assessed using discrimiation and calibration. Discrimination was evaluated through AUC and ROC, which measures the model's ability to correctly distinguish between abstinent and non-abstinent individuals. Calibration was assessed using calibration plots, which compare the predicted probabilities of abstinence to the observed outcomes. These plots provide insights into how well the model's predicted probabilities align with actual probabilities.

Results

Summary of Baseline Characteristics

The table below provides a summary of baseline characteristics stratified by the outcome (abstinence). As shown in the table, the non-abstinence group (n = 236) is significantly larger than the abstinence group (n = 64). P-values are included for each variable to indicate potential associations with abstinence. Based on these p-values, FTCD score at baseline, Nicotine Metabolism Ratio (NMR), and treatment variables (BA and Var) emerge as potentially important factors associated with abstinence. Although other covariates did not show statistical significance, differences can still be observed between abstinent and non-abstinent individuals in terms of education level, MDD status, race, and Pleasurable Events Scale scores at baseline for complementary reinforcers.

Table 2: Candidate Variable Summary by Abstinence Status

Characteristic	**No Abstinence (N = 236)**	**Abstinence (N = 64)**	**p-value**
Age at phone interview	50 (13)	51 (13)	0.8
Sex at phone interview			0.8
1	107 (45%)	28 (44%)	
2	129 (55%)	36 (56%)	
Income			0.6
1	88 (38%)	22 (35%)	
2	56 (24%)	12 (19%)	
3	36 (15%)	10 (16%)	
4	30 (13%)	8 (13%)	
5	24 (10%)	11 (17%)	
Education			0.13
1	0 (0%)	1 (1.6%)	
2	13 (5.5%)	3 (4.7%)	
3	60 (25%)	16 (25%)	
4	97 (41%)	19 (30%)	
5	66 (28%)	25 (39%)	
FTCD score at baseline	5 (2)	4 (2)	0.002
Smoking with 5 mins of waking up	113 (48%)	25 (39%)	0.2
BDI score at baseline	19 (12)	17 (11)	0.2
Cigarettes per day at baseline phone survey	16 (8)	14 (8)	0.052
Cigarette reward value at baseline	7 (4)	7 (4)	>0.9
Pleasurable Events Scale at baseline:substitute reinforcers	22 (19)	25 (22)	0.4
Pleasurable Events Scale at baseline:complementary reinforcers	26 (19)	23 (20)	0.15
Anhedonia	2 (3)	2 (2)	0.3
Other lifetime DSM-5 diagnosis			0.2
0	127 (54%)	40 (63%)	
1	109 (46%)	24 (38%)	
Taking antidepressant medication at baseline			0.9
0	171 (72%)	47 (73%)	
1	65 (28%)	17 (27%)	
Current vs past MDD			0.073
0	114 (48%)	39 (61%)	
1	122 (52%)	25 (39%)	
Nicotine Metabolism Ratio (NMR)	-1.23 (0.61)	-1.02 (0.55)	0.023
Exclusive Mentholated Cigarette User			0.4
0	91 (39%)	29 (45%)	
1	143 (61%)	35 (55%)	
Baseline readiness to quit smoking			0.6
3	1 (0.4%)	0 (0%)	
4	5 (2.2%)	0 (0%)	
5	24 (11%)	11 (19%)	
6	68 (30%)	15 (25%)	
7	53 (24%)	18 (31%)	
8	61 (27%)	13 (22%)	
9	6 (2.7%)	1 (1.7%)	
10	6 (2.7%)	1 (1.7%)	
Treatment Groups			<0.001
ST + placebo	60 (25%)	8 (13%)	
BASC + placebo	64 (27%)	4 (6.3%)	
ST + varenicline	55 (23%)	26 (41%)	

Table 2: Candidate Variable Summary by Abstinence Status *(continued)*

Characteristic	**No Abstinence (N = 236)**	**Abstinence (N = 64)**	**p-value**
BASC + varenicline	57 (24%)	26 (41%)	
Race Groups			0.079
Unknown	19 (8.1%)	3 (4.7%)	
NHW	74 (31%)	31 (48%)	
Black	128 (54%)	27 (42%)	
Hisp	14 (5.9%)	2 (3.1%)	
Mixed Race	1 (0.4%)	1 (1.6%)	

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

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

Lasso Model Results

We used the pooled coefficients to generate predictions by applying matrix multiplication to the long format of the dataset, which combined the five imputed datasets. The following table represents the lasso selected variables and their corresponding coefficients for each impuation, as well as the averaged coefficients.

Table 3: Lasso Selected Variables Coefficients

	Imp 1	Imp 2	Imp 3	Imp 4	Imp 5	Pooled
Intercept	0.5439	0.5463	0.5470	0.5373	0.5471	0.5443
FTCD score at baseline	-0.0295	-0.0288	-0.0293	-0.0278	-0.0282	-0.0287
BA : FTCD score at baseline	-0.0044	-0.0036	-0.0042	-0.0056	-0.0036	-0.0043
BA : Current MDD	-0.0607	-0.0570	-0.0604	-0.0508	-0.0559	-0.0570
Age : Varenicline	0.0034	0.0032	0.0033	0.0032	0.0032	0.0033
Sex : Cigarette reward value at baseline	-0.0043	-0.0042	-0.0049	-0.0042	-0.0050	-0.0045
Income Group 4: Pleasurable Events Scale at baseline complementary reinforcers	-0.0004	-0.0002	-0.0003	-0.0003	-0.0001	-0.0003
Education Group 4: Current MDD	-0.0456	-0.0394	-0.0463	-0.0524	-0.0416	-0.0451
Education Group 4 : Exclusive Mentholated Cigarette User	-0.0508	-0.0503	-0.0532	-0.0535	-0.0528	-0.0521
Non-Hispanic White: Cigarette reward value at baseline	0.0050	0.0031	0.0044	0.0046	0.0026	0.0039

Model performance was initially assessed through discrimination. The plot below displays the ROC curves and AUC values for both the training and validation sets. The AUC for the training set is approximately 0.8, while the validation set has an AUC of around 0.75. These AUC values indicate that the model has a moderately good ability to distinguish between abstinent and non-abstinent individuals. The slight decrease in AUC for the validation set suggests some overfitting in the training set, but it remains within an acceptable range.

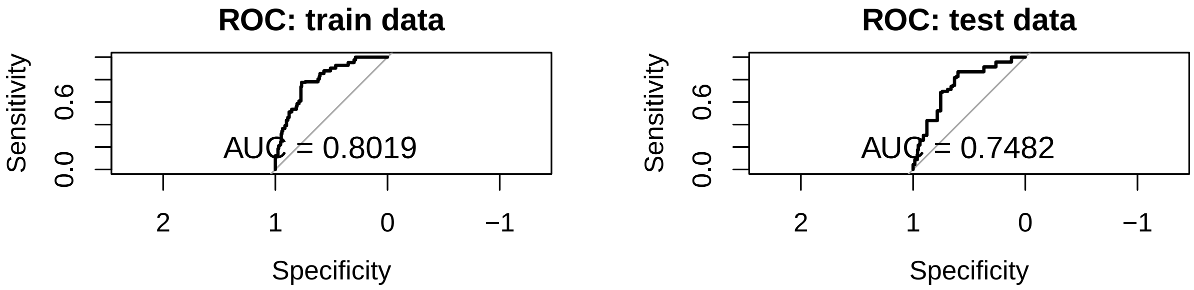

Figure 5: ROC Curves with AUC Values

The calibration plots displayed here compare the predicted probabilities with the actual observed proportions of abstinence in both the training set and testing set. The red line represents the ideal calibration line (predicted probabilities match observed proportion). The blue line shows a loess-smoothed fit of the observed data, and the shaded gray area represents the confidence interval around this fit. In both plots, the blue line is consistently below the red line, indicating that the model tends to overestimate the likelihood of the abstinence. This overestimation is less prominent at higher probability ranges, as the blue line approaches the red line. The red line spans only from 0.5 to 0.7. This conservative prediction range may result from lasso's shrinkage effect, which pulls probabilities toward the center and limits extreme predictions. In the testing set, there is more fluctuation and worse calibration, indicating some overfitting in the training set.

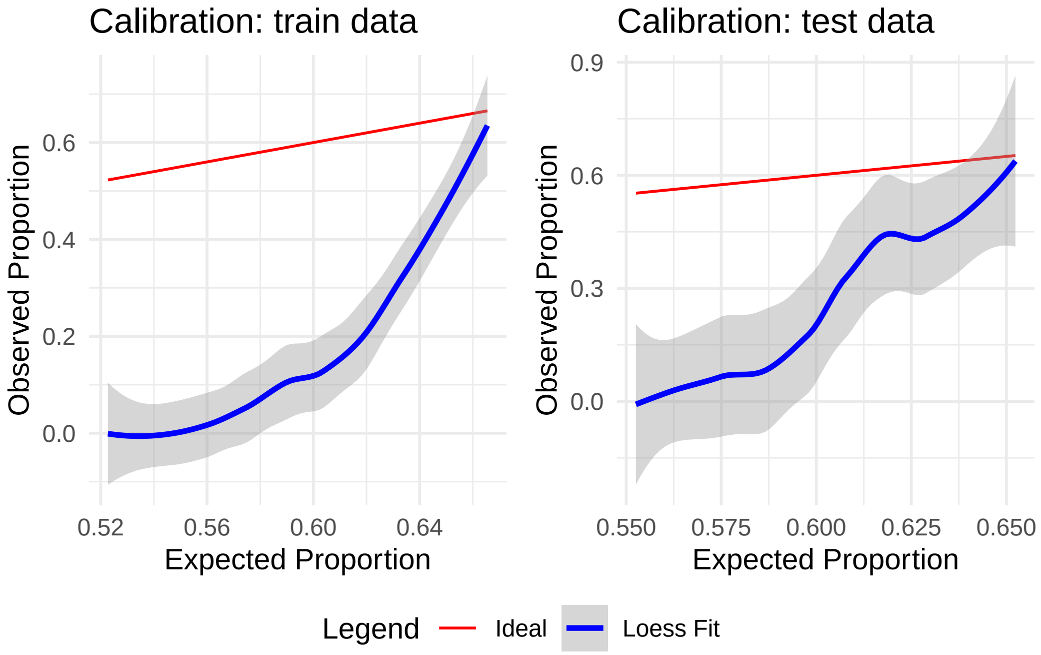

Figure 6: Calibration plots

Model Coefficients and Interpretation

This table provides the coefficients and odds ratios of the variables included in the Lasso model. **FTCD score at baseline** and **MDD status** appear to be potential moderators. Specifically, a higher FTCD score (odds ratio = 0.9957) is associated with slightly lower odds of abstinence, indicating that individuals with greater nicotine dependence may have a harder time achieving abstinence through BA. Similarly, individuals with **current MDD** (odds ratio = 0.9446) show slightly lower odds of abstinence compared to those with past MDD, suggesting that MDD status may influence the effectiveness of BA treatment. **Age** also shows a potential moderating effect when interacting with **Varenicline** (odds ratio = 1.0033), although the impact is minimal. The remaining interaction terms likely contribute to the predictive power of the model. Variables such as age, sex, cigarette reward value, income group 4, education group 4, exclusive mentholated cigarette use, pleasurable events scale complementary reinforcers, and non-hispanic white are included primarily for their predictive value, helping improve the model's accuracy in predicting abstinence outcomes. It's worth noting that all the odds ratio are close to 1, suggesting that even in the presence of moderators, the moderating effect is relatively small.

Table 4: Coefficient and Odds Ratio of Model

Variable	Coefficient	Odds Ratio
Intercept	0.5443	1.7234
FTCD score at baseline	-0.0287	0.9717
BA : FTCD score at baseline	-0.0043	0.9957
BA : Current MDD	-0.0570	0.9446
Age : Varenicline	0.0033	1.0033
Sex : Cigarette reward value at baseline	-0.0045	0.9955
Income Group 4: Pleasurable Events Scale at baseline complementary reinforcers	-0.0003	0.9997
Education Group 4: Current MDD	-0.0451	0.9559
Education Group 4 : Exclusive Mentholated Cigarette User	-0.0521	0.9492
Non-Hispanic White: Cigarette reward value at baseline	0.0039	1.0039

Discussion

The project aimed to examine baseline variables as potential moderators of the effects of Behavioral Activation (BA) on smoking cessation, while accounting for Varenicline use. Initially, we identified key main effects and interaction terms through exploratory data analysis (EDA) for inclusion in the Lasso model. To handle missing data, we performed multiple imputation and then applied Lasso regression for variable selection and regularization. We ran the Lasso model independently on each of the five imputed datasets and then combined the estimates, using the Lasso coefficients from each run. The model demonstrated a moderate ability to discriminate between outcomes, though it showed limited calibration, tending to overestimate probabilities.

The findings indicate that **FTCD Score** and **MDD Status** may act as modest moderators of BA's effects on smoking cessation. Specifically, individuals with higher FTCD scores or current MDD status tend to have lower odds of quitting smoking. This aligns with theoretical expectations, as a higher FTCD score reflects greater nicotine dependence, and current MDD status suggests a depressive disorder that could moderate BA's impact.

Additionally, the analysis identifies potential predictors of smoking cessation, such as age, sex, cigarette reward value, income group 4, education group 4, exclusive menthol cigarette use, pleasurable events scale complementary reinforcers, and non-Hispanic white indicator. While these factors may not serve as moderators, their associations highlight important predictors of abstinence.

While this project provides valuable insights into moderators of Behavioral Activation (BA) effects and potential predictive factors for abstinence, several limitations warrant consideration. First, the sample size is relatively small, with only 300 observations, which limits the model's ability to learn complex patterns effectively. Additionally, class imbalance worsen this issue, potentially leading to biased estimates and limiting the generalizability of the findings. The reliability of these results could be strengthened with more data. Second, the analysis is influenced by randomness inherent in multiple steps, such as multiple imputation, Lasso cross-validation, data splitting. Each contributing to an additional layer of uncertainty in model performance and feature selection. A future improvement could be to implement bootstrap Lasso, which would potentially produce more stable estimates. Third, the calibration plot suggests potential over-penalization by Lasso. The predicted values tend to cluster around 0.5, reflecting Lasso's conservative tendency to shrink coefficients, which may suppress important signals and result in under-confident predictions. Exploring alternative regularization methods or adjusting penalty strength could better balance prediction accuracy and model interpretability in future work.

In conclusion, this project provides a general framework to investigate moderators of effects of BA on smoking cessation. Future research can produce more reliable and nuanced results if limitations mentioned here are addressed.

Reference

Hitsman, B., Papandonatos, G. D., Gollan, J. K., Huffman, M. D., Niaura, R., Mohr, D. C., Veluz-Wilkins, A. K., Lubitz, S. F., Hole, A., Leone, F. T., Khan, S. S., Fox, E. N., Bauer, A.-M., Wileyto, E. P., Bastian, J., & Schnoll, R. A. (2023). Efficacy and safety of combination behavioral activation for smoking cessation and varenicline for treating tobacco dependence among individuals with current or past major depressive disorder: A 2 × 2 factorial, randomized, placebo-controlled trial. Addiction, 118(9), 1710–1725. https://doi.org/10.1111/add.16209

Appendix: Code

```r
# Load in dataset
project2 <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/Desktop/PHP2550/Proje

# Library used
library(dplyr)
library(knitr)
library(kableExtra)
library(ggplot2)
library(gtsummary)
library(reshape2)
library(mice)
library(glmnet)
library(caret)
library(ggpubr)
library(tibble)
library(pROC)
library(vcd)
library(bestglm)
library(gt)
# Add a new treatment column for EDA purpose
project2 <- project2 %>%
  mutate(treatment = case_when(Var == 1 & BA == 1 ~ "BASC + varenicline",
                               Var == 1 & BA == 0 ~ "ST + varenicline",
                               Var == 0 & BA == 1 ~ "BASC + placebo",
                               Var == 0 & BA == 0 ~ "ST + placebo"))

# Add a race column to indicate different race groups
project2<- project2 %>%
  mutate(race = case_when(NHW == 1 & Black == 0 & Hisp == 0 ~ "NHW",
                          NHW == 0 & Black == 1 & Hisp == 0 ~ "Black",
                          NHW == 0 & Black == 0 & Hisp == 1 ~ "Hisp",
                          NHW == 1 | Black == 1 | Hisp == 1 ~ "Mixed Race",
                          TRUE ~ "Unknown"))

# Factor and reset levels of covariates
project2$abst <- factor(project2$abst)
project2$Var <- factor(project2$Var)
project2$BA <- factor(project2$BA)
project2$sex_ps <- factor(project2$sex_ps)
project2$otherdiag <- factor(project2$otherdiag)
project2$antidepmed <- factor(project2$antidepmed)
project2$mde_curr <- factor(project2$mde_curr)
project2$Only.Menthol <- factor(project2$Only.Menthol)
project2$inc <- factor(project2$inc, levels = c(1,2,3,4,5))
project2$edu <- factor(project2$edu, levels = c(1,2,3,4,5))
project2$treatment <- factor(project2$treatment, levels = c("ST + placebo",
                                                            "BASC + placebo",
                                                            "ST + varenicline",
                                                            "BASC + varenicline"))
project2$race <- factor(project2$race, levels = c("Unknown","NHW",
                                                  "Black",
                                                  "Hisp",
                                                  "Mixed Race"))
project2$readiness <- as.numeric(project2$readiness)

# Delete original treatment and race columns for eda data
eda_data <- project2 %>%
  dplyr::select(-Var,-BA, -NHW, -Black, -Hisp)

# Prepare data for modeling
model_data <- project2 %>%
  dplyr::select(-NHW, -Black, -Hisp)

par(mfrow = c(1,2))
# Plot before transform distribution
hist(eda_data$NMR, main = NULL,
     xlab = "NMR", ylab = "Frenquency", col = "skyblue")

# Log transfor of NMR
```

```r
model_data$NMR <- log(model_data$NMR)
eda_data$NMR <- log(eda_data$NMR)

# Plot post transform distribution
hist(eda_data$NMR, main = NULL,
     xlab = "Log(NMR)", ylab = "Frenquency", col = "skyblue")
# Sample dataset for correlation heatmap
cor_data <- model_data %>%
  select(-treatment,-id)

# Identify categorical and continuous variables
categorical_vars <- c("abst", "Var", "BA", "sex_ps", "inc", "edu", "ftcd.5.mins",
                      "otherdiag", "antidepmed", "mde_curr", "Only.Menthol", "race")
continuous_vars <- setdiff(names(cor_data), categorical_vars)

# Create an empty matrix to store correlations
association_matrix <- matrix(NA, nrow = length(names(cor_data)),
                             ncol = length(names(cor_data)),
                             dimnames = list(names(cor_data), names(cor_data)))

# Calculate Cramér's V for categorical-categorical pairs
for (i in categorical_vars) {
  for (j in categorical_vars) {
    if (i != j) {
      tbl <- table(cor_data[[i]], cor_data[[j]])
      association_matrix[i, j] <- assocstats(tbl)$cramer
    } else {
      association_matrix[i, j] <- 1
    }
  }
}

# Calculate Pearson correlation for continuous-continuous pairs
for (i in continuous_vars) {
  for (j in continuous_vars) {
    association_matrix[i, j] <- cor(cor_data[[i]], cor_data[[j]],
                                    method = "pearson",
                                    use = "complete.obs")
  }
}
```

```
# Calculate Eta-Squared for categorical-continuous pairs
for (i in categorical_vars) {
  for (j in continuous_vars) {
    # Fit an ANOVA model to calculate eta-squared
    model <- aov(cor_data[[j]] ~ as.factor(cor_data[[i]]))
    eta_squared <- summary(model)[[1]][["Sum Sq"]][1] /
      sum(summary(model)[[1]][["Sum Sq"]])
    association_matrix[i, j] <- eta_squared
    association_matrix[j, i] <- eta_squared
  }
}

# Convert matrix to a dataframe for plotting
assoc_df <- melt(association_matrix,
                 varnames = c("Variable1", "Variable2"),
                 value.name = "Association")

# Plot the heatmap
ggplot(assoc_df, aes(x = Variable1, y = Variable2, fill = Association)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0,
                       limit = c(-1, 1), space = "Lab", name="Association") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# exclude `ftcd.5.mins` and `cpd_ps` to avoid multicolinearity
model_data <- model_data %>%
  select(-ftcd.5.mins, -cpd_ps)
# Plot interaction plot between MDD status and treatment
ggplot(eda_data, aes(x = treatment, fill = as.factor(abst))) +
  geom_bar(position = "fill") +
  facet_wrap(~ mde_curr) +
  labs(x = "Treatment", y = "Proportion of Abstinence",
       fill = "Abstinence") +
  scale_y_continuous(labels = scales::percent_format()) +
  scale_fill_manual(values = c("0" = "coral", "1" = "steelblue")) +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45,hjust=1, size = 7))


# Plot interaction plot between education and mde_curr
```

```
ggplot(eda_data, aes(x = edu, fill = as.factor(abst))) +
  geom_bar(position = "fill") +
  facet_wrap(~ mde_curr) +
  labs(x = "Education Levels", y = "Proportion of Abstinence",
       fill = "Abstinence") +
  scale_y_continuous(labels = scales::percent_format()) +
  scale_fill_manual(values = c("0" = "coral", "1" = "steelblue")) +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45,hjust=1, size = 7))


# Missingness dataframe
missing_df <- project2 %>%
  select(NMR, crv_total_pq1, readiness, inc, shaps_score_pq1, Only.Menthol)
missing_prop_df <- as.data.frame(sapply(missing_df,
                                        function(x)
                                          round(mean(is.na(x))*100, 2)))

# Convert row names to a column
missing_prop_df <- rownames_to_column(missing_prop_df, var = "Variable")
missing_prop_df$Variable <- c("Nicotine Metabolism Ratio (NMR)",
                               "Cigarette reward value at baseline (crv_total_pq1)",
                               "Baseline readiness to quit smoking (readiness)",
                               "Income", "Anhedonia (shaps_score_pq1)",
                               "Exclusive Mentholated Cigarette User (Only.Menthol)")

# Table output of % missing results
knitr::kable(missing_prop_df,
             col.names = c("Variable",
                           "Percentage Missing"),
      caption = "Percentage of Missing Data") %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
# Run multiple imputation
set.seed(1234)
imputed_data <- mice(model_data %>% select(-treatment), m = 5, method = 'pmm',
                     printFlag = FALSE)
# Split data into train and test
set.seed(1234)
train_index <- createDataPartition(model_data$treatment, p = 0.7, list = FALSE)

# Delete treatment from the model dataset
```

```
model_data <- model_data %>% select(-treatment)
# Run lasso on 5 imputed dataset and store the results
set.seed(1234)
lasso_result_list <- list()
for (i in 1:5){

  # Get imputed dataset for i
  data <- mice::complete(imputed_data,i)

  # Subset to training data
  train_data <- data[train_index, ]

  # Create matrix with selected interaction terms and main effects
  X_train <- model.matrix(abst ~
                    BA * (age_ps + sex_ps + inc + edu + ftcd_score +
                                 bdi_score_w00  + crv_total_pq1 +
                            hedonsum_n_pq1 +hedonsum_y_pq1 + shaps_score_pq1 +
                            otherdiag + antidepmed+ mde_curr +
                            NMR + Only.Menthol + readiness + race)
                    + Var * (age_ps + sex_ps + inc + edu + ftcd_score
                             + bdi_score_w00  + crv_total_pq1 + hedonsum_n_pq1 +
                               hedonsum_y_pq1 + shaps_score_pq1 + otherdiag +
                               antidepmed + mde_curr + NMR + Only.Menthol +
                               readiness + race)
                    + (sex_ps + inc + edu + race + age_ps) * (ftcd_score +
                              bdi_score_w00 + crv_total_pq1 + hedonsum_n_pq1 +
                                hedonsum_y_pq1 + shaps_score_pq1 +
                                otherdiag + antidepmed +
                              mde_curr + NMR + Only.Menthol + readiness)
                    ,
                  data = train_data)[,-c(1,2)]

  # Create outcome of abst status
  y <- as.numeric(as.character(train_data$abst))

  # Fit the model on just the train data with cv to find best lambda
  set.seed(1234)
  weights <- ifelse(train_data$abst == 1, 236 / 64, 1)
  fit_cv <- cv.glmnet(X_train,y, alpha=1, weights = weights, nfolds = 10)

  # Extract best lambda value
  lambda_min <- fit_cv$lambda.min

  # Fit again with the best lambda
  fit <- glmnet(X_train,y, alpha=1, lambda = lambda_min, weights = weights)

  # Extract the coefficients from the model
  res <- as.data.frame(as.matrix(coef(fit)))

  # Append the results
  lasso_result_list[[i]] <- res
}

# Combine the results and rename columns
lasso_result <- do.call(cbind, lasso_result_list)
colnames(lasso_result) <- paste0("Imputation_", 1:5)
# Count the number of zeros in each row
row_zero_counts <- apply(lasso_result, 1, function(x) sum(x == 0))

# Filter out rows with 2 or more zeros
filtered_lasso_result <- lasso_result[row_zero_counts <= 2, ]
# Create summary table stratified by outcome and add p-values
summary_table <- eda_data[,-1] %>%
  tbl_summary(
    by = abst,
    label = c(age_ps ~ "Age at phone interview",
              sex_ps ~ "Sex at phone interview",
              inc ~ "Income",
              edu ~ "Education",
              ftcd_score ~ "FTCD score at baseline",
              ftcd.5.mins ~ "Smoking with 5 mins of waking up",
              bdi_score_w00 ~ "BDI score at baseline",
              cpd_ps ~ "Cigarettes per day at baseline phone survey",
              crv_total_pq1 ~"Cigarette reward value at baseline",
              hedonsum_n_pq1 ~ "Pleasurable Events Scale at baseline:substitute reinforcer
              hedonsum_y_pq1 ~ "Pleasurable Events Scale at baseline:complementary reinfor
              shaps_score_pq1 ~ "Anhedonia",
              otherdiag ~ "Other lifetime DSM-5 diagnosis",
              antidepmed ~ "Taking antidepressant medication at baseline",
              mde_curr ~ "Current vs past MDD",
              NMR ~ "Nicotine Metabolism Ratio (NMR)",
              Only.Menthol ~ "Exclusive Mentholated Cigarette User",
              readiness ~ "Baseline readiness to quit smoking",
              treatment ~ "Treatment Groups",
              race ~ "Race Groups"),
    statistic = all_continuous() ~ "{mean} ({sd})",
    missing = "no"
  ) %>%
  modify_header(list(stat_1 ~ "No Abstinence (N = 236)",
                     stat_2 ~ "Abstinence (N = 64)")) %>%
  add_p() %>%
  gtsummary::as_kable_extra(
    booktabs = TRUE,
    caption = "Candidate Variable Summary by Abstinence Status",
    longtable = TRUE,
    linesep = ""
  ) %>%
  kableExtra::kable_styling(
    font_size = 8,
    latex_options = c("repeat_header", "HOLD_position")
  ) %>%
  kableExtra::column_spec(1, width = "5cm") %>%
  kableExtra::column_spec(2, width = "4cm") %>%
  kableExtra::column_spec(3, width = "4cm") %>%
  kableExtra::row_spec(0, bold = TRUE, font_size = 8)

summary_table
# Prepare test and train data matrices and datasets in the long format for prediction
X_train_long <- c()
X_test_long <- c()
train_data_long <- c()
test_data_long <- c()
for(i in 1:5){
  # Get train data in long format and train matrix in long format
  data <- mice::complete(imputed_data,i)
  train_data <- data[train_index, ]
  train_data_long <- rbind(train_data, train_data_long)
  X_train <- model.matrix(abst ~
                    BA * (age_ps + sex_ps + inc + edu + ftcd_score +
                                 bdi_score_w00  + crv_total_pq1 + hedonsum_n_pq1 +
                                 hedonsum_y_pq1 + shaps_score_pq1 +
                            otherdiag + antidepmed
                                 + mde_curr + NMR + Only.Menthol + readiness + race)
                    + Var * (age_ps + sex_ps + inc + edu + ftcd_score +
                                 bdi_score_w00  + crv_total_pq1 + hedonsum_n_pq1 +
                                 hedonsum_y_pq1 + shaps_score_pq1 +
                               otherdiag + antidepmed
                                 + mde_curr + NMR + Only.Menthol + readiness + race)
                    + (sex_ps + inc + edu + race + age_ps) * (ftcd_score +
                              bdi_score_w00 + crv_total_pq1 + hedonsum_n_pq1 +
                              hedonsum_y_pq1 + shaps_score_pq1 +
                                otherdiag + antidepmed +
                              mde_curr + NMR + Only.Menthol + readiness)
                    ,
                  data = train_data)[,-c(1,2)]
  X_train_long <- rbind(X_train, X_train_long)

  # Get test data in long format and test matrix in long format
  test_data <- data[-train_index, ]
  test_data_long <- rbind(test_data, test_data_long)
  X_test <- model.matrix(abst ~
                    BA * (age_ps + sex_ps + inc + edu + ftcd_score +
                                 bdi_score_w00  + crv_total_pq1 + hedonsum_n_pq1 +
                                 hedonsum_y_pq1 + shaps_score_pq1 + otherdiag +
                            antidepmed
                                 + mde_curr + NMR + Only.Menthol + readiness + race)
                    + Var * (age_ps + sex_ps + inc + edu + ftcd_score +
                                 bdi_score_w00  + crv_total_pq1 + hedonsum_n_pq1 +
                                 hedonsum_y_pq1 + shaps_score_pq1 + otherdiag
                             + antidepmed
                                 + mde_curr + NMR + Only.Menthol + readiness + race)
                    + (sex_ps + inc + edu + race + age_ps) * (ftcd_score +
                              bdi_score_w00 + crv_total_pq1 + hedonsum_n_pq1 +
                              hedonsum_y_pq1 + shaps_score_pq1 + otherdiag +
                                antidepmed +
                              mde_curr + NMR + Only.Menthol + readiness)
                    ,
                  data = test_data)[,-c(1,2)]
  X_test_long <- rbind(X_test, X_test_long)
}
# Calculate pooled estimate and intercept
filtered_lasso_result$average_estimate <- apply(filtered_lasso_result, 1,
                                                function(x) mean(x))
```

```
# Round to 4 decimal places
filtered_lasso_result_output <- filtered_lasso_result %>%
  mutate(across(where(is.numeric), ~ round(., 4)))

# Rename the rows
rownames(filtered_lasso_result_output) <-  c("Intercept", "FTCD score at baseline",
                      "BA : FTCD score at baseline",
                      "BA : Current MDD",
                      "Age : Varenicline",
                      "Sex : Cigarette reward value at baseline",
                      "Income Group 4: Pleasurable Events Scale at baseline complementary
                      "Education Group 4: Current MDD",
                      "Education Group 4 : Exclusive Mentholated Cigarette User",
                      "Non-Hispanic White: Cigarette reward value at baseline")


# Output the filtered_lasso_result table
knitr::kable(filtered_lasso_result_output,
             col.names = c("Imp 1",
                           "Imp 2",
                           "Imp 3",
                           "Imp 4",
                           "Imp 5",
                           "Pooled"),
      caption = "Lasso Selected Variables Coefficients") %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down","resizebox"),
                font_size = 6)

# Extract intercept and coefficient estimate
pooled_intercept <- filtered_lasso_result$average_estimate[1]
pooled_coefs <- filtered_lasso_result$average_estimate[-1]

# Calculate log-odds with pooled coefficients on train data
selected_columns <- rownames(filtered_lasso_result[-1,])
X_train_long_selected <- X_train_long[, selected_columns, drop = FALSE]

# Calculate log-odds and convert to probabilities on train data
log_odds <- as.data.frame(pooled_intercept + X_train_long_selected %*% pooled_coefs)
train_data_long$log_odds <- log_odds$V1
train_data_long$predicted_prob <- 1 / (1 + exp(-train_data_long$log_odds))
```

```
# Calculate log-odds with pooled coefficients on test data
selected_columns <- rownames(filtered_lasso_result[-1,])
X_test_long_selected <- X_test_long[, selected_columns, drop = FALSE]

# Calculate log-odds and convert to probabilities on test data
log_odds <- as.data.frame(pooled_intercept + X_test_long_selected %*% pooled_coefs)
test_data_long$log_odds <- log_odds$V1
test_data_long$predicted_prob <- 1 / (1 + exp(-test_data_long$log_odds))

# Calculate ROC and AUC for both test and train
roc_result_train <- roc(response = train_data_long$abst,
                        predictor = train_data_long$predicted_prob)
auc_result_train <- auc(roc_result_train)
roc_result_test <- roc(response = test_data_long$abst,
                       predictor = test_data_long$predicted_prob)
auc_result_test <- auc(roc_result_test)
# Plot the ROC curve and AUC values for both test and train
par(mfrow=c(1,2))
plot(roc_result_train, main = "ROC: train data")
text(x = 0.6, y = 0.2, labels = paste("AUC =", round(auc_result_train, 4)), cex = 1.2)
plot(roc_result_test, main = "ROC: test data")
text(x = 0.6, y = 0.2, labels = paste("AUC =", round(auc_result_test, 4)), cex = 1.2)

# Calibration for Train Set
num_cuts <- 10
calib_data_train <-  data.frame(prob = train_data_long$predicted_prob,
                          bin = cut(train_data_long$predicted_prob, breaks = num_cuts),
                          class = as.numeric(as.character(train_data_long$abst)))
calib_data_train <- calib_data_train %>%
             group_by(bin) %>%
             summarise(observed = sum(class)/n(),
                       expected = sum(prob)/n(),
                       se = sqrt(observed * (1-observed) / n()))
calib_plot_train <- ggplot(calib_data_train) +
  geom_line(aes(x = expected, y = expected, color = "Ideal")) +
  geom_smooth(aes(x = expected, y = observed, color = "Loess Fit"), method = "loess") +
  labs(x = "Expected Proportion", y = "Observed Proportion",
       title = "Calibration: train data", color = "Legend") +
  theme_minimal() +
  scale_color_manual(values = c("Ideal" = "red", "Loess Fit" = "blue"))
```

```r
# Calibration for Test set
calib_data_test <-  data.frame(prob = test_data_long$predicted_prob,
                          bin = cut(test_data_long$predicted_prob, breaks = num_cuts),
                          class = as.numeric(as.character(test_data_long$abst)))
calib_data_test <- calib_data_test %>%
             group_by(bin) %>%
             summarise(observed = sum(class)/n(),
                       expected = sum(prob)/n(),
                       se = sqrt(observed * (1-observed) / n()))
calib_plot_test <- ggplot(calib_data_test) +
  geom_line(aes(x = expected, y = expected, color = "Ideal")) +
  geom_smooth(aes(x = expected, y = observed, color = "Loess Fit"), method = "loess") +
  labs(x = "Expected Proportion", y = "Observed Proportion",
       title = "Calibration: test data", color = "Legend") +
  theme_minimal() +
  scale_color_manual(values = c("Ideal" = "red", "Loess Fit" = "blue"))

calib_plot_combined <- ggarrange(calib_plot_train, calib_plot_test, ncol = 2, nrow = 1,
                                 common.legend = TRUE, legend = "bottom")
calib_plot_combined
# Create a data frame for pooled coefficients and odds ratio
coef_df <- data.frame(Variable = rownames(filtered_lasso_result),
                      Coefficient = round(filtered_lasso_result$average_estimate,4))
coef_df$Odds_ratio <- round(exp(coef_df$Coefficient),4)
coef_df$Variable <- c("Intercept", "FTCD score at baseline",
                      "BA : FTCD score at baseline",
                      "BA : Current MDD",
                      "Age : Varenicline",
                      "Sex : Cigarette reward value at baseline",
                      "Income Group 4: Pleasurable Events Scale at baseline complementary
                      "Education Group 4: Current MDD",
                      "Education Group 4 : Exclusive Mentholated Cigarette User",
                      "Non-Hispanic White: Cigarette reward value at baseline")

# Table of Model Coefficients and Odds Ratio
knitr::kable(coef_df,
             col.names = c("Variable",
                           "Coefficient",
                           "Odds Ratio"),
      caption = "Coefficient and Odds Ratio of Model") %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"),
                    font_size = 8)
```