Moderation Analysis and Prediction of Smoking Abstinence in Behavioral Therapy and Pharmacotherapy

Miaoyan Chen

2024-12-13

Aim: This is a follow-up analysis for the study to examine baseline variables as potential moderators of the effects of behaviorial treatment on abstinence and evaluate baseline variable as predictors of abstinence, controlling for behavioral treatment and pharmacotherapy.

Method: Multiple imputation is used to address missing data. We performed our analysis with Lasso regression and Best Subset regression, and the models were evaluated and compared for their performance metrics and coefficients.

Conclusions: Baseline characteristics such as Non-hispanic white, income, and FTCD score are selected as significant predictors of smoking abstinence. Significant moderators are observed for both behavioral and pharmacological treatments. However, there are some uncertainty regarding the significance of the coefficients. Larger and more diverse samples are needed to validate these findings and enhance their reliability.

1. Introduction

Smokers with depression are more likely to perceive smoking as a pleasurable activity and showing greater dependence than smokers without depression (Breslau, Kilbey, and Andreski 1992). Behavioral activation is a behavioral treatment that may improve smoking cessation for people with major depressive disorder (MDD). Previous study intended to examine the effect of treatment combination of BA and varenicline on smoking cessation for individuals with MDD. Hitsman et al. evaluated the efficacy of the novel treatment combination in a 2x2 randomized, placebo-controlled design across two U.S universities. Four treatment arms were considered in the trial, BASC + varenicline, BASC + placebo, standard treatment (ST) + varenicline, and ST + placebo. Results show that varenicline effectively promotes smoking cessation without increased safety risks in individuals with MDD, but no significant superiority

of BASC was found over standard treatment in improving smoking cessation rates. Considering that individual characteristics could potentially impact the abstinence rate, it is worthy to investigate how the baseline characteristics impacts the efficacy of the treatment arms. This paper is an extension of the trial (Hitsman et al. 2023) to examine the relationship between the baseline variables and abstinence, controlling for behavioral therapy and pharmacotherapy, and to examine the baseline variables as potential moderators of the effects of behavioral treatment on end-of-treatment abstinence.

2. Data

Data source

The data for this analysis is derived from a manipulated subset of the cohort from Hitsman et al.'s research. The subset only includes the primary outcome (smoking abstinence) and the 23 baseline characteristics. 300 smokers with confirmed diagnosis of major depressive disorder (MDD), without psychotic features according to the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) (Association 2013) with interest in quitting smoking were recruited in the trial. Initial eligibility screening was conducted via telephone, followed by final eligibility screening, informed consent, treatment randomization, and baseline assessment at week 0. Missing not at random (MNAR) assumption was applied to treat missingness of smoking abstinence data (missing = smoking) for Intent to Treat (ITT) analysis.

Baseline characteristics were carefully examined during the initial data preprocessing stage. Income and education were considered as ordinal variables, with some categories collapsed due to low observations/sample sizes. 10 score and count based measurements in the data set were analyzed as continuous variables, and 11 baseline binary variables and outcome (abstinence) were factored. The following Table 1 provides a overview of the baseline characteristics for the study participants stratified by treatment arm. Although the sample sizes for several variables are relatively small across the treatment arms, the randomization process seems to have balanced the distributions across treatment arms, making the groups more comparable. Some of the percentages do not sum up to 100% due to missing data, and details in handling missing data will be discussed in section 3.

Train-test split

We implemented a train-test split as part of our model validation procedure, stratifying by the smoking abstinence outcome variable to ensure balanced representation. We randomly sampled without replacement, allocating 70% of the data ($n = 211$) to the training set for model derivation and remaining 30% of the data ($n = 89$) to the test set for model validation. Training set and test set were generated prior to addressing missing data issues to prevent data leakage. Consequently, we performed multiple imputation separately on each set.

Table 1: Participant characteristics by treatment arm

Characteristic	**ST + Placebo** N = 68	**BASC + Placebo** N = 68	**BASC + Varenicline** N = 83	**ST + Varenicline** N = 81
Age	50.3 (10.8)	50.7 (13.5)	50.3 (13.2)	48.7 (12.7)
Sex				
Male	29 (42.6%)	30 (44.1%)	39 (47.0%)	37 (45.7%)
Female	39 (57.4%)	38 (55.9%)	44 (53.0%)	44 (54.3%)
Income				
Less than $20,000	26 (38.2%)	25 (37.3%)	30 (36.6%)	29 (36.3%)
$20,000–50,000	28 (41.2%)	24 (35.8%)	30 (36.6%)	32 (40.0%)
More than $50,000	14 (20.6%)	18 (26.9%)	22 (26.8%)	19 (23.8%)
Education				
Some high school or below	2 (2.9%)	4 (5.9%)	7 (8.4%)	4 (4.9%)
High school graduate or GED	11 (16.2%)	23 (33.8%)	15 (18.1%)	27 (33.3%)
Some college/technical school	38 (55.9%)	22 (32.4%)	32 (38.6%)	24 (29.6%)
College graduate	17 (25.0%)	19 (27.9%)	29 (34.9%)	26 (32.1%)
Race				
Black	40 (60.6%)	37 (56.9%)	37 (50.0%)	43 (58.9%)
Hispanic	4 (6.1%)	4 (6.2%)	3 (4.1%)	5 (6.8%)
Non-Hispanic White	22 (33.3%)	24 (36.9%)	34 (45.9%)	25 (34.2%)
FTCD score	5.4 (2.1)	5.3 (2.0)	5.1 (2.3)	5.2 (2.1)
Smoking with 5 mins after waking				
More than 5 minutes	33 (48.5%)	36 (52.9%)	50 (60.2%)	43 (53.1%)
5 minutes or less	35 (51.5%)	32 (47.1%)	33 (39.8%)	38 (46.9%)
BDI score	18.5 (10.8)	19.0 (12.3)	18.0 (10.6)	19.5 (12.2)
Cigarettes per day	15.0 (7.2)	15.6 (9.1)	15.5 (8.5)	14.4 (6.6)
Cigarette reward value	7.0 (3.7)	7.4 (3.8)	7.2 (3.9)	7.1 (3.5)
Substitute reinforcers	20.8 (20.1)	23.2 (20.3)	22.9 (19.0)	23.4 (19.5)
Complementary reinforcers	27.4 (19.9)	27.7 (21.5)	22.4 (17.0)	25.0 (19.4)
Anhedonia	2.5 (3.4)	2.2 (3.2)	2.3 (3.1)	2.1 (3.0)
Other lifetime DSM-5 diagnosis	28 (41.2%)	35 (51.5%)	30 (36.1%)	40 (49.4%)
Antidepressant medication	15 (22.1%)	28 (41.2%)	24 (28.9%)	15 (18.5%)
Major depressive disorder status				
Past MDD only	37 (54.4%)	36 (52.9%)	43 (51.8%)	37 (45.7%)
Current and Past MDD	31 (45.6%)	32 (47.1%)	40 (48.2%)	44 (54.3%)
Nicotine Metabolism Ratio	0.4 (0.3)	0.3 (0.2)	0.4 (0.2)	0.4 (0.2)
Cigarette type				
Regular cigarettes (or both)	24 (35.8%)	28 (41.2%)	34 (41.5%)	34 (42.0%)
Methol cigarettes only	43 (64.2%)	40 (58.8%)	48 (58.5%)	47 (58.0%)

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

Table 2: Missing data

Variable	Missing	Total	Missing Percentage
Income	3	300	1.00
FTCD score	1	300	0.33
Cigarette reward value	18	300	6.00
Anhedonia	3	300	1.00
Nicotine Metabolism Ratio	21	300	7.00
Cigarette type	2	300	0.67
Readiness to quit	17	300	5.67

3. Missing data and multiple imputation

Table 2 summarizes the percentage of missing data for each variable. A total of 7 baseline characteristics ($n = 59$ participants) contain missing values. Given the absence of missing data patterns, we can assume that the data are missing completely at random (MCAR).Given the absence of discernible missing data patterns, we assume that the data are missing completely at random (MCAR). Considering the small sample size ($N = 300$), limiting the analysis to complete cases only would exclude 19.67% of the participants. Such an omission would significantly reduce statistical power and potentially introduce bias into the results. To address this, we implemented Multiple Imputation by Chained Equations (MICE) to handle missing data in the baseline characteristics. MICE consists of three stages: imputation, analysis, and pooling. We generated $m = 5$ data sets with missing values imputed using a predictive model. Each imputed data is analyzed individually but identically to obtain a set of parameter estimates. To derive the pooled estimate coefficients, the estimates from each data set are aggregated to obtain overall parameter estimates, accounting for both within imputation and between imputation variability to ensure robust and unbiased inference. For continuous variables, predictive mean matching ("pmm") was used, while binary variables were imputed using logistic regression ("logreg"), and categorical variables were handled using polytomous logistic regression ("polyreg"). The pooled estimates from the five imputed datasets were averaged to provide aggregated coefficients.

4. Exploratory data analysis

No strong correlations were identified between the continuous variables in the baseline characteristics. Figure 1 presents boxplots grouped by behavioral activation and abstinence status, providing insight into potential moderators. Each boxplot compares smoking abstinence status across treatment groups (ST vs. BA). No obvious differences were observed between abstinent

and non-abstinent groups across treatment conditions, as the distributions largely overlapped. Therefore, these variables may not serve as strong moderators for behavioral activation.

Table 3 summarizes the categorical baseline characteristics stratified by treatment (ST vs. BA) and end-of-treatment outcome (abstinence status). Among participants who did not achieve abstinence ($n = 236$), the use of antidepressant medication was significantly higher in the BA group (34%) compared to the ST group (21%). However, among participants who achieved abstinence ($n = 64$), the difference between the ST and BA groups was not statistically significant, though it approached significance ($p = 0.086$). This suggests that antidepressant medication might act as a potential moderator, with participants on antidepressants potentially experiencing different outcomes depending on treatment assignment. Other distributions, such as sex, race, income, education, and smoking behavior, appeared similar across behavioral activation treatment groups and abstinence outcomes.

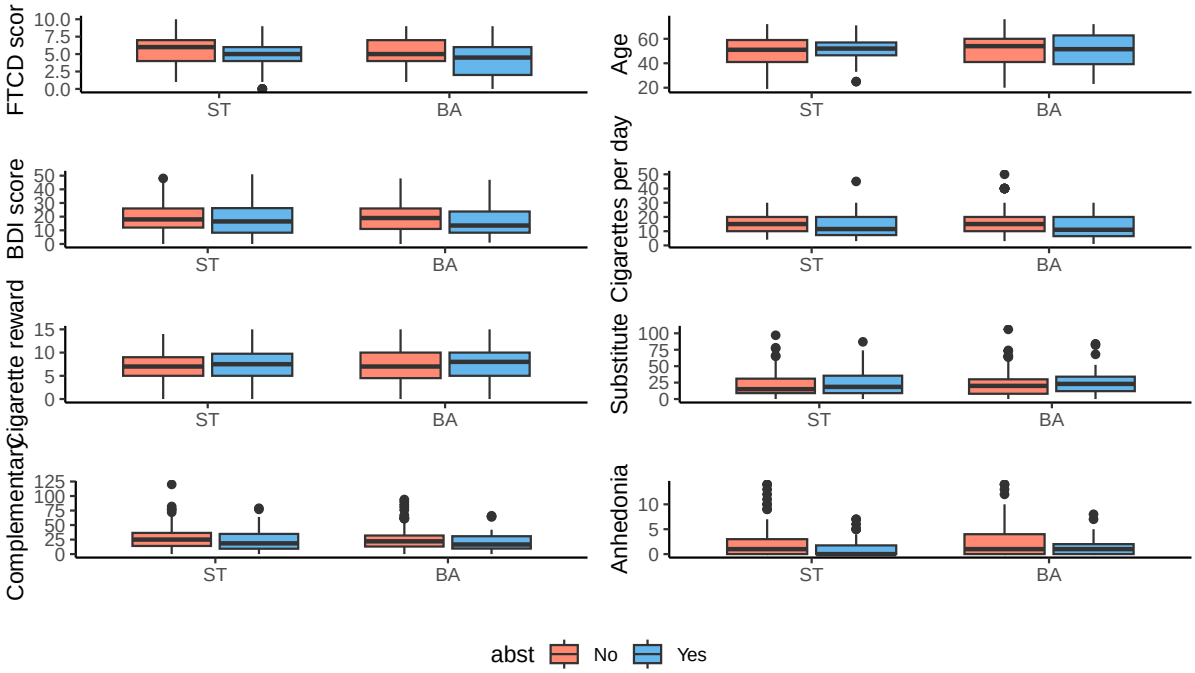

Figure 1: Distribution of baseline variable by bahavioral therapy and outcome

5. Models

We performed Lasso regression and Best Subset regression (Hazimeh and Mazumder 2020) to examine baseline variables as potential moderators of the effects of behavioral treatment on end-of-treatment (EOT) abstinence. The initial model included interaction terms between baseline variables and behavioral activation, as well as between baseline variables and varenicline. Lasso

Table 3: Smoking habit and depression by abstinence and BA treatment arms

Characteristic	**Abstinence = No**, N = 236			**Abstinence = Yes**, N = 64		
	ST N = 115	**BA** N = 121	**p-value**	**ST** N = 34	**BA** N = 30	**p-value**
Smoking with 5 mins after waking			0.6			0.2
More than 5 minutes	58 (50%)	65 (54%)		18 (53%)	21 (70%)	
5 minutes or less	57 (50%)	56 (46%)		16 (47%)	9 (30%)	
Other lifetime DSM-5 diagnosis	55 (48%)	54 (45%)	0.6	13 (38%)	11 (37%)	0.9
Antidepressant medication	24 (21%)	41 (34%)	0.025	6 (18%)	11 (37%)	0.086
Major depressive disorder status			0.7			0.7
Past MDD only	54 (47%)	60 (50%)		20 (59%)	19 (63%)	
Current and Past MDD	61 (53%)	61 (50%)		14 (41%)	11 (37%)	
Cigarette type			0.7			0.087
Regular cigarettes (or both)	46 (40%)	45 (38%)		12 (35%)	17 (57%)	
Methol cigarettes only	68 (60%)	75 (63%)		22 (65%)	13 (43%)	

[1] n (%)

[2] Pearson's Chi-squared test

(Least Absolute Shrinkage and Selection Operator) and Best Subset regression are especially helpful for this analysis, as it performs both variable selection and regularization, and enhances predictive performance in high-dimensional data.

In Lasso regression, the ℓ_1 penalty shrinks the coefficients of less influential predictors towards zero, performing variable selection from numerous predictors and improving interpretability by typically selecting one predictor from a group of highly correlated variables. In Best Subset regression, the optimal subset of predictors is selected to minimize the test MSE. The ℓ_0 penalty enhances sparsity by reducing the number of predictors with non-zero coefficients, while the squared ℓ_2 norm adds a shrinkage penalty to control the magnitude of the coefficients. By tuning the ℓ_0 regularization parameter (λ), we can manage the number of non-zero coefficients. We employed a logistic loss function and set maxSuppSize = 20 to allow for a maximum of 20 non-zero coefficients in the model.

We implemented 10-fold cross-validation (CV) to identify the optimal tuning parameter, λ, that minimizes prediction error and produces the most predictive model. The CV folds were stratified to ensure equivalent proportions of treatment combinations across folds in Lasso regression, maintaining balance between behavioral and pharmacological interventions to prevent bias results due to information leakage. For each imputed dataset, a cross-validated model was fitted using the optimal λ. The coefficients from these models were then averaged across all imputed datasets to produce an aggregated model that accounts for both within- and between-imputation variability.

Model performance evaluation

To evaluate the model's fit and predictive accuracy, we assessed the Brier score, calibration, and discrimination. The final aggregated Lasso regression and Best Subset model were validated using the test dataset ($n = 91$). Table 5 summarizes key performance metrics, including accuracy, specificity, sensitivity, AUC, and Brier score for both models. Calibration plots and ROC curves were plotted to visualize the model's predictive performance and discrimination ability.

Brier score measures the accuracy of probabilistic predictions, with lower values indicating better predictive performance. For Lasso regression, the Brier score was slightly lower in the training set (0.143) compared to the test set (0.180). The Area Under the Curve (AUC) is a measure of discrimination ability, which was higher in the training set (0.835) than in the test set (0.638), suggesting that the model performed worse in distinguishing between abstinent and non-abstinent cases in the test set, likely due to a small test sample size ($n = 91$). The ROC curves (Figure 2) demonstrated excellent performance on the training data, indicating strong discriminatory ability. Overall, the training set showed better performance metrics: accuracy (0.836), specificity (0.864), and sensitivity (0.733), compared to the test set, which had an accuracy of 0.562, specificity of 0.486, and sensitivity of 0.842. For Best Subset regression, we observed a drop in accuracy in both training data (0.790) and test data (0.438) compared to Lasso. Other metrics followed a similar pattern as those observed with Lasso regression.

Calibration plots (Figure 3 and Figure 4) were used to further evaluate the models. These plots revealed discrepancies between predicted and observed probabilities for both the Lasso and Best Subset regression models. While both models exhibited calibration issues, the Best Subset regression with $L_0 + L_2$ regularization appeared to be better calibrated than Lasso. Specifically, the predicted probabilities (blue line) deviated from the ideal diagonal line, indicating a tendency to under- or overestimate probabilities within certain ranges.This suggests potential calibration issues that may require further refinement of our models.

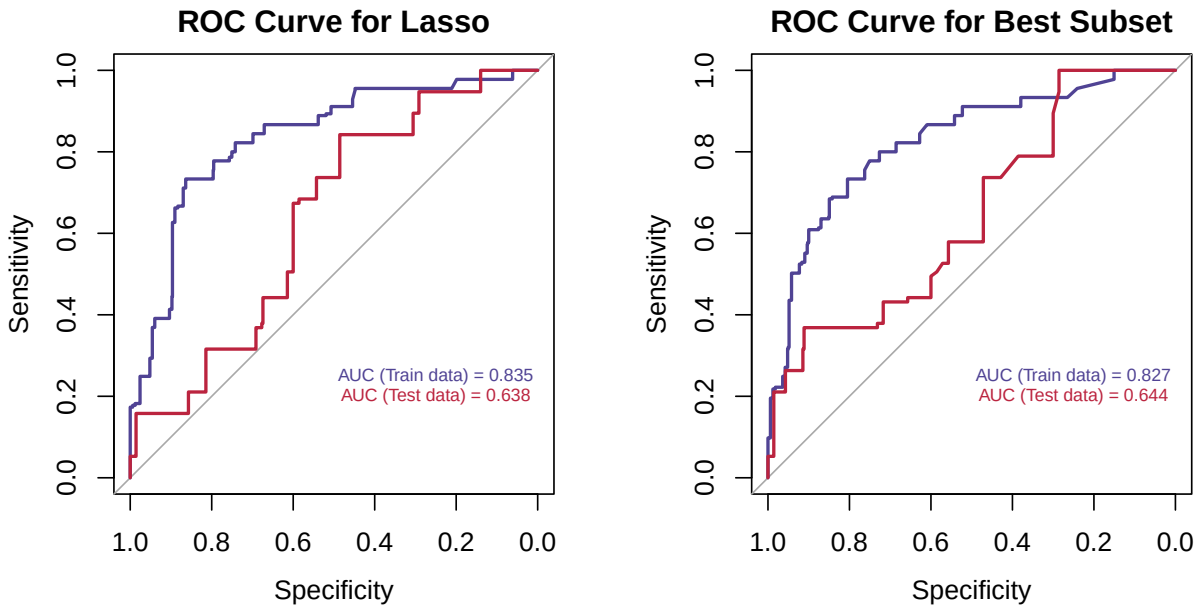

Figure 2: ROC curve for Lasso and Best Subset regression

Table 4: Model results from Lasso and Best Subset regression

	Lasso		Best Subset	
	OR	Estimate	OR	Estimate
Non-hispanic white	1.1982	0.1809	2.9983	1.0980
Income 20,000 - 75,000	0.8843	-0.1229		
FTCD score	0.7814	-0.2466	0.5851	-0.5360
Behavior activation x Income more than 75,000	1.0012	0.0012		
Behavior activation x Major depressive disorder	0.9458	-0.0557		
Behavior activation x Cigarette type	0.8904	-0.1161		
Varenicline x Age	1.0074	0.0074		
Varenicline x Non-hispanic white	1.4763	0.3895		
Varenicline x Income 35,001 - 50,000	1.0113	0.0112		
Varenicline x High school graduate or GED	1.1371	0.1285		
Varenicline x Smoking with 5 mins of waking up	1.4755	0.3890	6.6356	1.8924
Varenicline x Cigarette reward value	1.0016	0.0016		
Varenicline x Antidepressant medication	1.3048	0.2661		
Varenicline x Nicotine Metabolism Ratio	1.9481	0.6669	6.1040	1.8089

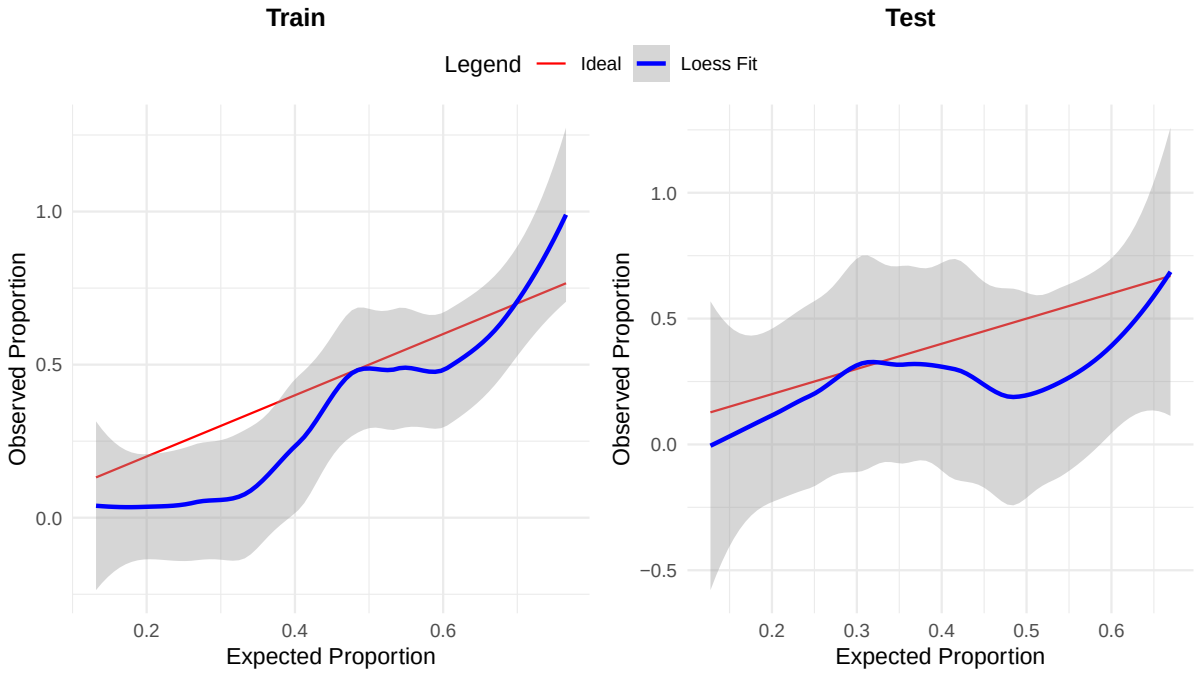

Figure 3: Calibration plots for Lasso (train vs. test)

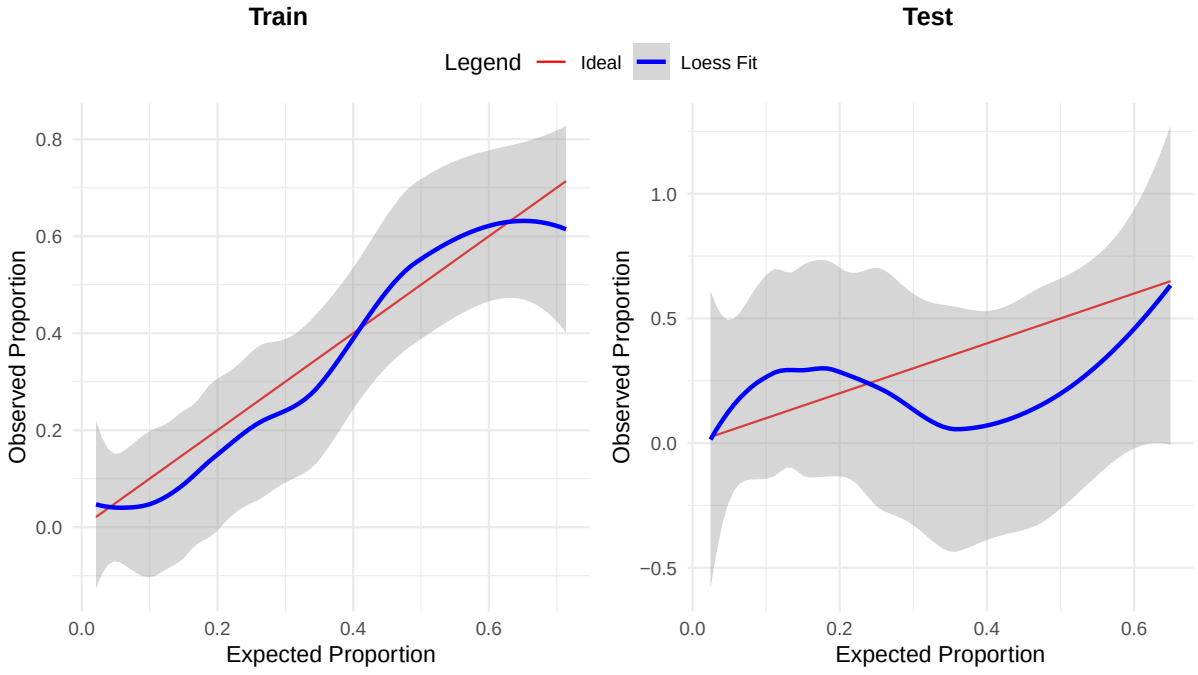

Figure 4: Calibration plots for Best Subset (train vs. test)

Table 5: Model Evaluation

	Lasso		Best Subset	
	Train	Test	Train	Test
Accuracy	0.836	0.562	0.790	0.438
Specificity	0.864	0.486	0.805	0.286
Sensitivity	0.733	0.842	0.733	1.000
AUC	0.835	0.638	0.827	0.644
Brier score	0.143	0.180	0.122	0.171

Model interpretation

The moderation analysis using Lasso regression identified several baseline characteristics and interaction terms that influence the efficacy of behavioral activation and pharmacological treatment. Table 4 presents the selected variables and aggregated model coefficients across five imputed datasets. Notably, the coefficients for behavioral treatment (BA) and pharmacotherapy (Var) are reduced to zero by the Lasso penalty.

From Lasso regression baseline characteristics such as Non-hispanic white, income, and FTCD score were selected as significant predictors of smoking abstinence, independent of the effects of behavioral treatment or pharmacotherapy. Being non-hispanic white (OR: 1.1982) increases the odds of smoking abstinence. Conversely, having an income range from $ 20,000 to $ 75,000 (OR: 0.8843) decreases the odds of smoking abstinence compared to individuals with income less than $20,000. Higher FTCD score (OR: 0.781) is also associated with lower odds of abstinence. The Best Subset regression further emphasizes key predictors influencing smoking abstinence as it identified similar terms as Lasso regression. Non-hispanic white (OR: 2.9983) seems to be a stronger predictor of abstinence in Best Subset model. In this case, being a Non-Hispanic White increases the odds of smoking abstinence by 2.9983, which is a 150.23% increase in odds ratio compared to the coefficient from Lasso. FTCD score (OR: 0.5851) remains as a negative predictor of the smoking abstinence.

Additionally, our moderation analysis using Lasso identified several significant moderators for both behavioral and pharmacological treatments. Individuals with an income above $75,000 experienced a slightly enhanced effect of behavioral activation (OR: 1.0012) compared to those earning below $20,000. Diagnoses of major depressive disorder (OR: 0.9458) and exclusive smoking of methanol cigarettes (OR: 0.890) negatively influenced smoking abstinence as moderators for behavioral treatment. The effectiveness of varenicline varied slightly by age (OR: 1.007). Furthermore, the impact of pharmacological treatment on smoking abstinence differed by race; non-Hispanic whites (OR: 1.4763) had higher odds of abstinence compared to Black individuals under varenicline treatment. Varenicline was also more effective for individuals with incomes between $35,001 and $50,000 (OR: 1.0113) and for high school graduates or GED

holders (OR: 1.1371). Other moderators, such as the perceived reward value of cigarettes and antidepressant medication usage, were also significant for pharmacological treatment.

Notably, both Lasso and Best Subset regression indicated that the effect of varenicline on smoking abstinence could be moderated by morning smoking habits and the nicotine metabolism ratio, with Best Subset yielding larger effects. Varenicline was more effective for individuals who smoked within 5 minutes of waking (Lasso OR: 1.4755; Best Subset OR: 6.6356) compared to those who smoked later. Additionally, the interaction between varenicline and nicotine metabolism ratio showed that varenicline was more effective for individuals with faster nicotine metabolism (Lasso OR: 1.9481; Best Subset OR: 6.1040).

7. Discussion

There were some agreement and disagreement between the model outputs. Non-Hispanic White and FTCD score are common variables across the models, suggesting they are strong predictors of abstinence. Morning smoking habits and nicotine metabolism were identified by the models, particularly by Best Subset, as very influential moderators of the effect of pharmacological treatments on smoking abstinence. The interaction between varenicline and age also appears in both models, though it might not be a significant moderator given the consistently positive association with age.

A key limitation of this study is the absence of an external validation dataset. Relying solely on an internal dataset for validation, especially with a small sample size, may limit the generalizability and transportability of the findings. External validation is essential to validate the model's robustness and performance in a diverse populations.

While the model demonstrated strong discrimination ability and good overall performance on the test set, the calibration plots reveal discrepancies between predicted and observed probabilities. This indicates that the model's predicted probabilities are not perfectly aligned with actual outcomes, suggesting potential areas for improvement in calibration.

Additionally, the process of performing multiple imputation after the train-test split may result in different imputed values for similar missing observations in the dataset. This approach can introduce variability in the analysis, potentially affecting the consistency and stability of the model's results.

8. Conclusion

This analysis used Lasso regression and Best Subset regression to investigate baseline characteristics and their interactions with behavioral therapy and pharmacotherapy. Both models identified key factors associated with smoking abstinence. Specifically, Non-Hispanic White ethnicity and FTCD score consistently emerged as robust predictors across both methods.

Lasso regression additionally highlighted income, major depressive disorder, and cigarette type as moderators influencing the effectivenss of behavioral treatment. Morning smoking habits and nicotine metabolism ratio were identified as strong moderators for pharmacotherapy. Other significant moderators revealed by Lasso includes age, race, income, education, cigarette reward value, and antidepressant medication usage.

While the model demonstrated strong discriminatory power and good overall performance in the test set, calibration plots revealed notable discrepancies between predicted and observed probabilities, indicating certain limitations. The absence of an external validation dataset and the relatively small sample size further constrain the generalizability of these findings. Future research involving larger, more diverse samples is essential to validate these results and enhance model robustness.

References

Association, American Psychiatric. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. Doi:10.1176/appi.books.9780890425596. DSM Library. American Psychiatric Association. https://doi.org/doi:10.1176/appi.books.9780890425596.

Breslau, N, M M Kilbey, and P Andreski. 1992. "Nicotine Withdrawal Symptoms and Psychiatric Disorders: Findings from an Epidemiologic Study of Young Adults." *Am J Psychiatry* 149 (4): 464–69. https://doi.org/10.1176/ajp.149.4.464.

Hazimeh, Hussein, and Rahul Mazumder. 2020. "Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms." *Operations Research* 68 (5): 1517–37. https://doi.org/10.1287/opre.2019.1919.

Hitsman, Brian, George D Papandonatos, Jacqueline K Gollan, Mark D Huffman, Raymond Niaura, David C Mohr, Anna K Veluz-Wilkins, et al. 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 $\times$ 2 Factorial, Randomized, Placebo-Controlled Trial." *Addiction* 118 (9): 1710–25. https://doi.org/10.1111/add.16209.

Appendix

```
knitr::opts_chunk$set(echo = F,
                  #     fig.align = "center",
                      fig.height = 4.5,
                      fig.width = 8,
                      message = FALSE,
                      warning = FALSE,
                      tidy = TRUE,
                      kable = TRUE
)
# Load libraries
library(tidyverse) # Data manipulation
library(gtsummary) # Summary table
library(dplyr) # Data manipulation
library(pROC) # ROC curve
library(latex2exp) # Writing symbols in latex
library(kableExtra) # Create nice table output
library(psych) # Exploratory data
library(glmnet) # For Lasso regression
library(caret) # For creating folds in CV
library(ggplot2) # Plotting
library(LOLearn) # Best subset
library(ggpubr) # arrange plots
setwd("~/Desktop/PHP 2550 Pratical Data Analysis/Project 2")
smoke <- read.csv("~/Desktop/PHP 2550 Pratical Data Analysis/Project 2/project2.csv")

# Remove id
smoke <- subset(smoke, select = c(-id))

# Factor categorical data
col_names <- names(smoke)[-c(4,11,13:18,22,24)]
smoke[,col_names] <- lapply(smoke[,col_names] , factor)

# Create a race category for all the races
smoke <- smoke %>%
  mutate(Race = case_when(NHW == 1 ~ "Non-Hispanic White",
                          Black == 1 ~ "Black",
                          Hisp == 1 ~ "Hispanic"))

# Create treatment variable for 2x2 factorial design
```

```r
smoke <- smoke %>%
  mutate(treatment_arm = case_when(BA == 0 & Var == 0 ~ "ST + Placebo",
                                   BA == 1 & Var == 0 ~ "BASC + Placebo",
                                   BA == 0 & Var == 1 ~ "ST + Varenicline",
                                   BA == 1 & Var == 1 ~ "BASC + Varenicline"))

smoke$treatment_arm <- factor(smoke$treatment_arm)
smoke$treatment_arm <- relevel(smoke$treatment_arm, ref = "ST + Placebo")

# Regroup and make income and education into ordinal variables
## Income
smoke <- smoke %>%
  mutate(inc = case_when(inc == 1 ~ 1,
                         inc == 2|  inc == 3 ~ 2,
                         inc == 4|  inc == 5 ~ 4))
smoke$inc <- ordered(smoke$inc, levels = c(1,2,4),
                        labels = c("Less than $20,000", "$20,000-50,000", "More than $50,000
## Education
smoke <- smoke %>%
  mutate(edu = ifelse(edu == 1| edu == 2, 1, edu))
smoke$edu <- ordered(smoke$edu, levels = c(1,3,4,5),
                           labels = c("Some high school or below", "High school graduate or
                                      "Some college/technical school", "College graduate"))

# Relabel binary variables
smoke$sex_ps <- factor(smoke$sex_ps, levels = c(1,2), labels = c("Male","Female"))
smoke$ftcd.5.mins <- factor(smoke$ftcd.5.mins, levels = c(0,1), labels = c("More than 5 minu
smoke$otherdiag <- factor(smoke$otherdiag, levels = c(0,1), labels = c("No", "Yes"))
smoke$antidepmed <- factor(smoke$antidepmed, levels = c(0,1), labels = c("No", "Yes"))
smoke$mde_curr <- factor(smoke$mde_curr, levels = c(0,1), labels = c("Past MDD only","Curren
smoke$Only.Menthol <- factor(smoke$Only.Menthol, levels = c(0,1), labels = c("Regular cigare

smoke %>%
  tbl_summary(include = c(age_ps, sex_ps, inc, edu, Race, ftcd_score, ftcd.5.mins,
                          bdi_score_w00,cpd_ps,crv_total_pq1,hedonsum_n_pq1,
                          hedonsum_y_pq1,shaps_score_pq1,otherdiag,antidepmed,
                          mde_curr,NMR,Only.Menthol, treatment_arm),
              by = treatment_arm,
              missing = "no",
              statistic = list(
                all_continuous() ~ "{mean} ({sd})",
                all_categorical() ~ "{n} ({p}%)"
```

```
            ),
            digits = list(all_continuous() ~ 1, all_categorical() ~ c(0, 1)),
            label = list(
                        age_ps = "Age",
                        sex_ps = "Sex",
                        inc = "Income",
                        edu = "Education",
                        ftcd_score = "FTCD score",
                        bdi_score_w00 = "BDI score",
                        cpd_ps = "Cigarettes per day",
                        ftcd.5.mins ~ "Smoking with 5 mins after waking",
                        crv_total_pq1 ~ "Cigarette reward value",
                        hedonsum_n_pq1 ~ "Substitute reinforcers",
                        hedonsum_y_pq1 ~ "Complementary reinforcers",
                        shaps_score_pq1 ~ "Anhedonia",
                        otherdiag ~ "Other lifetime DSM-5 diagnosis",
                        antidepmed ~ "Antidepressant medication",
                        NMR ~ "Nicotine Metabolism Ratio",
                        mde_curr ~ "Major depressive disorder status",
                        Only.Menthol ~ "Cigarette type"
                        )
  ) %>%
  as_kable_extra(booktabs = T, escape = F, caption = "Participant characteristics by treatmen
  kable_styling(font_size = 9) %>%
  landscape()


# Rename variables -- rename(New_Name = Old_Name)
smoke <- smoke %>%
   rename(
         Age = age_ps,
         Sex = sex_ps,
         "Income" = inc,
         "Education" = edu,
         "FTCD score" = ftcd_score,
         "BDI score" = bdi_score_w00,
         "Cigarettes per day" = cpd_ps,
         "Cigarette reward value" = crv_total_pq1,
         "Anhedonia"= shaps_score_pq1,
         "Nicotine Metabolism Ratio" = NMR,
         "Cigarette type" = Only.Menthol,
         "Readiness to quit" = readiness,
```

```
         "Substitute reinforcers" = hedonsum_n_pq1,
         "Complementary reinforcers" = hedonsum_y_pq1
         )
smoke <- subset(smoke, select = c(-treatment_arm,-Race))

# missing data summarized by columns
missing_tb1 <- smoke %>%
  summarise_all(~ sum(is.na(.))) %>%
  pivot_longer(everything(), names_to = "Variable", values_to = "Missing") %>%
  mutate(Total = nrow(smoke),
         "Missing Percentage" = round((Missing / Total) * 100,2)) %>%
  filter(Missing >0)

missing_tb1 %>%
   kbl(caption = "Missing data",
      booktabs = T,
      escape = T,
      align = "c") %>%
  kable_styling(full_width = FALSE, latex_options = c('HOLD_position'))
# Histogram after log transformation
smoke_log <- smoke %>%
  mutate(`log(Substitute reinforcers)` = ifelse(`Substitute reinforcers` == 0, log(`Substitut
         `log(Complementary reinforcers)` = ifelse(`Complementary reinforcers` == 0, log(`Co
         `log(Nictoine Metabolism Ratio)` = ifelse(`Nicotine Metabolism Ratio` == 0, log(`Ni
  )

# Histograms of non-normal continuous data and its log transformation
# multi.hist(smoke_log[,c(16,17,22,25:27)], ncol = 3, density=TRUE, freq=FALSE, global = F, l

smoke$NHW <- factor(smoke$NHW, levels = c(0, 1), labels = c("No", "Yes"))
colnames(smoke)[colnames(smoke) == "NHW"] <- "Non-Hispanic White"
smoke$Black <- factor(smoke$Black, levels = c(0, 1), labels = c("No", "Yes"))
smoke$Hisp <- factor(smoke$Hisp, levels = c(0, 1), labels = c("No", "Yes"))
colnames(smoke)[colnames(smoke) == "Hisp"] <- "Hispanic"
smoke$abst <- factor(smoke$abst, levels = c(0, 1), labels = c("No", "Yes"))
smoke$BA <- factor(smoke$BA, levels = c(0, 1), labels = c("ST", "BA"))

# Potential interaction with BA or Var
plot_int <- function(yval, trt1, ylab){
  ggplot(data = smoke) +
    geom_boxplot(aes(x = trt1, y = yval, fill = abst)) +
    theme_classic() +
```

```
    labs(x = "", y = ylab) +
    scale_fill_manual(values = c("Yes" = "#64B6EC", "No" = "#FF8972"))
}

p1 <- plot_int(yval = smoke$`FTCD score`, trt1 = smoke$BA, ylab = "FTCD score")
p2 <- plot_int(yval = smoke$Age, trt1 = smoke$BA, ylab = "Age")
p3 <- plot_int(yval = smoke$`BDI score`, trt1 = smoke$BA, ylab = "BDI score")
p4 <- plot_int(yval = smoke$`Cigarettes per day`, trt1 = smoke$BA, ylab = "Cigarettes per day
p5 <- plot_int(yval = smoke$`Cigarette reward value`, trt1 = smoke$BA, ylab = "Cigarette rewa
p6 <- plot_int(yval = smoke$`Substitute reinforcers`, trt1 = smoke$BA, ylab = "Substitute")
p7 <- plot_int(yval = smoke$`Complementary reinforcers`, trt1 = smoke$BA, ylab = "Complementa
p8 <- plot_int(yval = smoke$`Anhedonia`, trt1 = smoke$BA, ylab = "Anhedonia")

ggarrange(p1, p2, p3, p4, p5, p6, p7,p8, ncol = 2, nrow = 4,
                                  common.legend = TRUE, legend = "bottom")
smoke[-c(2,4,11,13:18,22,24)] %>%
   mutate(abst = paste("Abstinence = ", abst)) |>
  tbl_strata(
    strata = abst,
    .tbl_fun =
      ~ .x |>
        tbl_summary(include = c(ftcd.5.mins,otherdiag,antidepmed,mde_curr,`Cigarette type`),
                    by = BA, missing = "no",
                    label = list(
                           ftcd.5.mins ~ "Smoking with 5 mins after waking",
                           otherdiag ~ "Other lifetime DSM-5 diagnosis",
                           antidepmed ~ "Antidepressant medication",
                           mde_curr ~ "Major depressive disorder status")) |>
        add_p(),
    .header = "**{strata}**, N = {n}"
  ) %>%
  as_kable_extra(booktabs = T, escape = F, caption = "Smoking habit and depression by abstine
  kable_styling(font_size = 9)
# Test-train split
setwd("~/Desktop/PHP 2550 Pratical Data Analysis/Project 2")
load("smoke_train.Rdata")
load("smoke_test.Rdata")
load("imp_train_long.RData")
load("imp_test_long.RData")
#////////////////////////////////////////////////////////////////////////////
#---- Lasso regression
#////////////////////////////////////////////////////////////////////////////
```

```r
Lasso <- function(imp_data, idx) {
  #' Perform Lasso regression with 10-fold CV
  #' @param imp_data, list of imputed data
  #' @param idx, index for the imputed data
  #' @return Lasso_coef, Lasso coefficients for minimum lambda

  df <- imp_data[[idx]]

  # Set contrasts to treatment coding
  options(contrasts = c("contr.treatment", "contr.treatment"))
  f <- as.formula(abst ~ (. - BA - Var) * BA + (. - BA - Var) * Var)
  y <- df$abst
  xvar <- model.matrix(f, df)[,-1]

  # Generate stratified folds

  ## Define grouping variable by taking the interaction of BA and Var
  grouping <- interaction(imp_train[[1]]$BA, imp_train[[1]]$Var)

  ## Create folds to ensure proportions of BA+Var are similar in each fold
  set.seed(123) # for reproducibility
  folds <- createFolds(grouping, k = 10)

  foldid <- rep(NA, length(grouping))
  for (i in seq_along(folds)) {
    foldid[folds[[i]]] <- i
  }

  # Lasso regression
  set.seed(123)
  lambda_values <- 10^seq(4, -4, length = 100)
  Lasso_reg <- cv.glmnet(xvar, y, nfolds = 10, foldid = foldid,
                         alpha = 1, family = "binomial", lambda = lambda_values)
  best_Lasso <- glmnet(xvar, df$abst, nfolds = 10, alpha = 1,
                       family = "binomial", lambda = Lasso_reg$lambda.min)
  Lasso_coef <- coef(best_Lasso)[-1,]

  # Return estimates
  return(Lasso_coef)
}

# Compute the pooled Lasso coefficients from MI dataset
```

```r
average_Lasso_coefs <- function(imp_long, n = 5) {
  #' Average across coefficients from the MI data sets
  #' Calculate the predicted probabilities, ROC and AUC from
  #' @param imp_long, imputed data set in long format
  #' @param n, number of iterations/imputed data

  # Iterate through each set of coefficient from Lasso
  Lasso_coefs_list <- vector("list", n)

  for (i in 1:n) {
    Lasso_coefs_list[[i]] <- Lasso(imp_train, i)
  }

  # Combine the coefficients into a matrix
  Lasso_coefs_matrix <- do.call(cbind, Lasso_coefs_list)

  # Compute the average for these coefficients
  avg_coefs <- apply(Lasso_coefs_matrix, 1, mean)

  coef_matrix <- as.matrix(avg_coefs)
  coef_estimates <- as.matrix(coef_matrix[coef_matrix[,1] != 0, ])

  df_long <- imp_long[,-c(1,2)]

  # Set contrasts to treatment coding
  options(contrasts = c("contr.treatment", "contr.treatment"))
  f_long <- as.formula(abst ~ (. - BA - Var) * BA + (. - BA - Var) * Var)
  y <- df_long$abst
  xvar <- model.matrix(f_long, df_long)[,-1]

  # Use long data for predicted probabilities
  pred_probs <- xvar %*% coef_matrix
  pred_probs <- 1 / (1 + exp(-pred_probs)) #convert log_odds to probabilities

   # Compute AUC, accuracy, specificity, and sensitivity
  roc_value <- roc(y, pred_probs, quiet = TRUE)
  auc_value <- auc(roc_value)
  accuracy <- unlist(unname(coords(roc_value, "best", ret = c("accuracy", "specificity", "se

  # Compute Brier score
  actual <- as.numeric(as.character(df_long$abst))
  brier_score <- mean((pred_probs-actual)^2)
```

```
  # Calibration
  num_cuts <- 10

  # Create a dataset for plotting calibration
  calib_data <-  data.frame(prob = pred_probs,
                            bin = cut(pred_probs, breaks = num_cuts),
                            class = as.numeric(as.character(df_long$abst)))

  calib_data <- calib_data %>%
    group_by(bin) %>%
    summarize(observed = sum(class)/n(),
              expected = sum(prob)/n(),
              se = sqrt(observed*(1-observed)/n()))

  # plot calibration
  calib_plot <- ggplot(calib_data, aes(x = expected, y = observed)) +
    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", color = "Legend") +
    theme_minimal() +
    scale_color_manual(values = c("Ideal" = "red", "Loess Fit" = "blue"))
  # geom_abline(linetype = "dashed", color = "red") +
    # geom_point(aes(x = expected, y = observed), size = 2, color = "blue") +
    # geom_errorbar(aes(x = expected,
    #                   ymin = observed - 1.96*se,
    #                   ymax = observed + 1.96*se),
    #               colour="black", width=.01)+
    # theme_bw() +
    # labs(x = "Predicted Probability", y = "Observed Probabilities")

  return(list(
    avg_coefs = coef_estimates,
    avg_auc = auc_value,
    df_long_abst = df_long$abst,
    pred_probs = pred_probs,
  #  roc_value = roc_value,
    accuracy = accuracy,
    brier_score = brier_score,
    calib_plot = calib_plot
  ))
}
```

```r
avg_coefs_Lasso_train <- average_Lasso_coefs(imp_train_long, n = 5)
avg_coefs_Lasso_test <- average_Lasso_coefs(imp_test_long, n = 5)
#////////////////////////////////////////////////////////////////////////////
#---- Best Subset regression
#////////////////////////////////////////////////////////////////////////////
bestsubset <- function(imp_data, idx)  {
  #' Runs 10-fold CV for Best Subset
  #' @param imp_data, data set
  #' @return bestsubset_coef, coefficients for minimum CV error

  # Matrix form for ordered variables
  df <- imp_train[[idx]]
  x.ord <- model.matrix(abst~ (. - BA - Var) * BA + (. - BA - Var) * Var, data = df)[, -1]
  y.ord <- df$abst

  # Best Subset model
  bestsubset_mod <- L0Learn.cvfit(x.ord, y.ord, penalty="L0L2", loss="Logistic", maxSuppSize=

  # Get optimal lambda
  optimalGammaIndex <- which.min(sapply(bestsubset_mod$cvMeans, min))
  optimalLambdaIndex =which.min(bestsubset_mod$cvMeans[[optimalGammaIndex]])
  optimalLambda = bestsubset_mod$fit$lambda[[optimalGammaIndex]][optimalLambdaIndex]

  # Get coefficients
  best_subset_coef <- coef(bestsubset_mod, lambda=optimalLambda, gamma=bestsubset_mod$fit$ga

  names(best_subset_coef) <- colnames(x.ord)
  return(best_subset_coef)
}

average_best_subset <- function(imp_long, n = 5) {
  #' Runs 10-fold CV for Best Subset and average across coefficients from the MI data sets
  #' Calculate the predicted probabilities, ROC, AUC, accuracy, Brier score, and calibration
  #' @param imp_long, imputed data set in long format
  #' @param n, number of iterations/imputed data

  best_subset_coefs_list <- vector("list", n)

  for (i in 1:n) {
    best_subset_coefs_list[[i]] <- bestsubset(imp_train, i)
  }
```

```r
# Combine the coefficients into a matrix
best_subset_coefs_matrix <- do.call(cbind, best_subset_coefs_list)

# Compute the average for these coefficients
avg_coefs <- apply(best_subset_coefs_matrix, 1, mean)

coef_matrix <- as.matrix(avg_coefs)
coef_estimates <- as.matrix(coef_matrix[coef_matrix[,1] != 0, ])

df_long <- imp_long[,-c(1,2)]

# Set contrasts to treatment coding
options(contrasts = c("contr.treatment", "contr.treatment"))
f_long <- as.formula(abst ~ (. - BA - Var) * BA + (. - BA - Var) * Var)
y <- df_long$abst
xvar <- model.matrix(f_long, df_long)[,-1]

# Use long data for predicted probabilities
pred_probs <- xvar %*% coef_matrix
pred_probs <- 1 / (1 + exp(-pred_probs)) #convert log_odds to probabilities

# Compute AUC, accuracy, specificity, and sensitivity
roc_value <- roc(y, pred_probs)
auc_value <- auc(roc_value)
accuracy <- unlist(unname(coords(roc_value, "best", ret = c("accuracy", "specificity", "se

# ROC Curve
# roc_curve <- plot.roc(roc_value, col = "blue", main = paste("ROC Curve (AUC =", round(au

# Compute Brier Score
actual <- as.numeric(as.character(df_long$abst))
brier_score <- mean((pred_probs - actual)^2)

# Calibration
num_cuts <- 10

calib_data <-  data.frame(prob = pred_probs,
                          bin = Hmisc::cut2(pred_probs, g = num_cuts),
                          class = as.numeric(as.character(df_long$abst)))

calib_data <- calib_data %>%
  group_by(bin) %>%
```

```
    summarize(observed = sum(class)/n(),
              expected = sum(prob)/n(),
              se = sqrt(observed*(1-observed)/n()))

  # Plot calibration
  calib_plot <- ggplot(calib_data, aes(x = expected, y = observed)) +
    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", color = "Legend") +
    theme_minimal() +
    scale_color_manual(values = c("Ideal" = "red", "Loess Fit" = "blue"))
    # geom_abline(linetype = "dashed", color = "red") +
    # geom_errorbar(aes(ymin = observed - 1.96 * se, ymax = observed + 1.96 * se),
    #               colour = "black", width = 0.02) +
    # geom_point(size = 2, color = "blue") +
    # theme_bw() +
    # labs(x = "Predicted Probability", y = "Observed Probability")

  return(list(
    avg_coefs = coef_estimates,
    df_long_abst = df_long$abst,
    pred_probs = pred_probs,
    avg_auc = auc_value,
   # roc_curve = ROC_curve,
    accuracy = accuracy,
    brier_score = brier_score,
    calib_plot = calib_plot
  ))
}

avg_coefs_bestsubset_train <- average_best_subset(imp_train_long, 5)
avg_coefs_bestsubset_test <- average_best_subset(imp_test_long, 5)
par(mfrow = c(1, 2))
# Calculate the pooled ROC curve for Lasso
pooled_roc_train <- roc(avg_coefs_Lasso_train$df_long_abst, avg_coefs_Lasso_train$pred_probs)
pooled_roc_test <- roc(avg_coefs_Lasso_test$df_long_abst, avg_coefs_Lasso_test$pred_probs)

# Plot the pooled ROC curve
plot(pooled_roc_train, col = "#504394", lwd = 2, main = "ROC Curve for Lasso")
lines(pooled_roc_test, col = "#BB243F")
text(x = 0.25, y = 0.25, labels = paste("AUC (Train data) =", round(avg_coefs_Lasso_train$avg
text(x = 0.25, y = 0.2, labels = paste("AUC (Test data) =", round(avg_coefs_Lasso_test$avg_au
```

```r
# Calculate the pooled ROC curve for Best Subset
pooled_roc_train2 <- roc(avg_coefs_bestsubset_train$df_long_abst, avg_coefs_bestsubset_train$
pooled_roc_test2 <- roc(avg_coefs_bestsubset_test$df_long_abst, avg_coefs_bestsubset_test$pre

# Plot the pooled ROC curve
plot(pooled_roc_train2, col = "#504394", lwd = 2, main = "ROC Curve for Best Subset")
lines(pooled_roc_test2, col = "#BB243F")
text(x = 0.25, y = 0.25, labels = paste("AUC (Train data) =", round(avg_coefs_bestsubset_trai
text(x = 0.25, y = 0.2, labels = paste("AUC (Test data) =", round(avg_coefs_bestsubset_test$a
train_coef_estimates <- as.data.frame(as.matrix(avg_coefs_Lasso_train$avg_coefs))
train_coef_estimates <- train_coef_estimates %>%
  mutate(OR = exp(V1)) %>%
  select(OR, V1)

rownames(train_coef_estimates) <- c("Non-hispanic white", "Income 20,000 - 75,000", "FTCD sco

train_coef_estimates_bestsubset <- as.data.frame(as.matrix(avg_coefs_bestsubset_train$avg_coe
train_coef_estimates_bestsubset <- train_coef_estimates_bestsubset %>%
  mutate(OR =exp(V1)) %>%
  select(OR, V1)

rownames(train_coef_estimates_bestsubset) <- c("Non-hispanic white", "FTCD score", "Varenicli

# Convert row names to a column for joining
train_coef_estimates <- train_coef_estimates %>%
  rownames_to_column(var = "Variable")

train_coef_estimates_bestsubset <- train_coef_estimates_bestsubset %>%
  rownames_to_column(var = "Variable")

# Full join by first column
combined_coef_estimates <- full_join(train_coef_estimates, train_coef_estimates_bestsubset, b

options(knitr.kable.NA = '')
combined_coef_estimates %>%
  kbl(col.names =c("","OR", "Estimate", "OR", "Estimate"),
      caption ="Model results from Lasso and Best Subset regression", booktabs =T,escape = F
  add_header_above(c(" "=1, "Lasso" = 2, "Best Subset" = 2)) %>%
  kable_styling(full_width =FALSE,latex_options =c("hold_position"))
# Calibration plot for training set
train_calib <- avg_coefs_Lasso_train$calib_plot
```

```r
# Calibration plot for test set
test_calib <- avg_coefs_Lasso_test$calib_plot
combine_plot <- ggarrange(train_calib, test_calib, ncol = 2, nrow = 1, common.legend = T)
annotate_figure(combine_plot,
                top = text_grob("Train
# Calibration plot for training set
train_calib_bs <- avg_coefs_bestsubset_train$calib_plot
test_calib_bs <-avg_coefs_bestsubset_test$calib_plot
combine_plot2 <- ggarrange(train_calib_bs, test_calib_bs, ncol =2, nrow =1, common.legend = T
annotate_figure(combine_plot2,
                top = text_grob("Train
#////////////////////////////////////////////////////////////////////////////
#---- Model evaluation
#////////////////////////////////////////////////////////////////////////////

tbl <- data.frame(Lasso_train = c(avg_coefs_Lasso_train$accuracy, avg_coefs_Lasso_train$avg_a
                  Lasso_test = c(avg_coefs_Lasso_test$accuracy, avg_coefs_Lasso_test$avg_auc
                  bestsubset_train = c(avg_coefs_bestsubset_train$accuracy, avg_coefs_bestsub
                  bestsubset_test= c(avg_coefs_bestsubset_test$accuracy, avg_coefs_bestsubset
  round(3)
rownames(tbl) <- c("Accuracy", "Specificity", "Sensitivity", "AUC", "Brier score")

kable(tbl, booktabs = T, escape = T,
      caption = "Model Evaluation",
      row.names = TRUE,
      col.names = c("Train", "Test", "Train", "Test")) %>%
  add_header_above(c(" "=1, "Lasso" = 2, "Best Subset" = 2)) %>%
  kable_styling(full_width = F,
                latex_options = c("HOLD_position"))
```