Predicting smoking abstinence and moderation analysis for behavioral therapy and pharmacotherapy

Project 2: Regression and moderation Analysis

Miaoyan Chen

2024-11-11

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 employed LASSO regression and Best Subset regression in our analysis. Models are evaluated and compared for their performance metrics and coefficients.

Conclusions: Baseline characteristics such as Non-Hispanic White, education level, FTCD score, and MDD are selected as significant predictors of smoking abstinence. Significant moderators are observed for both behavioral and pharmacological treatments. However, there is 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 employ a train-test split as our model validation procedure. A random sample without replacement is drawn, with 70% of the rows (n = 210) allocated to the training set for model derivation, while the remaining 30% of the rows (n = 90) are retained as the test set for model validation. Training set and test set are generated prior to addressing missing data issues. As a result, multiple imputation is performed separately for the training and test sets.

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). Considering the small sample size (N = 300), if the analysis were limited to complete cases only, 19.67% of the participants would be excluded. Therefore, omitting this proportion of the data would significantly reduce statistical power and potentially bias the results. To address this, we will use the Multiple Imputation by Chained Equations (MICE) to handle missing data in the baseline characteristics. There are three stages in MICE: 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 get 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, we will use Predictive Mean Matching (PMM), binary variable will be imputed using logistic regression ("logreg"), and categorical variable will be imputed using polytomous logistic regression ("polyreg"). The pooled estimates from the five imputed datasets will be averaged to provide aggregated coefficients.

4. Exploratory data analysis

No strong correlations are identified between the continuous variables in the baseline characteristics. Figure 1 shows boxplots grouped by behavioral activation and abstinence status to provide insight about potential moderators. Each boxplot compares between abstinent and non-abstinent across treatent type (ST vs. BA). There's no obvious difference between and non-abstinent groups across treatment for these characteristics as the distributions largely overlap. Therefore, these may not be strong moderators for behavioral activation.

Table 3 represents the categorical baseline characteristics stratified by treatment (ST vs. BA) and end of treatment outcome (abstinence status). Among those who didn't achieve abstinence (N = 236), the use of antidepressant medication was significantly higher than in BA group (34%) compared to the ST group (21%). However, for participants who achieve abstinence (N = 64), the difference between ST and BA group is not statistically significant, though is near signficance (p = 0.086). This could indicate that antidepressant medication could be a potential moderator and participants on antidepressant might experience different outcomes depending on the treatment assignment. Other distributions, such as sex, race, income, education, and smoking behavior appears similar across behavioral activation treatment groups and abstinence.

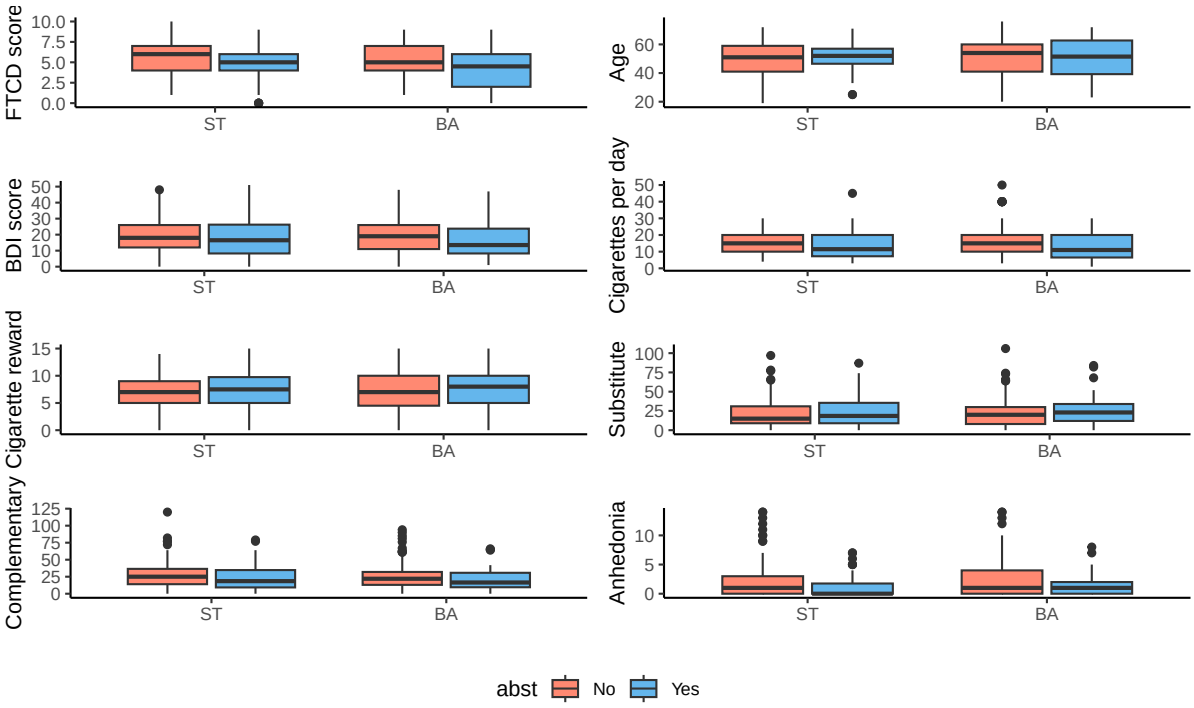

Figure 1: Distribution of baseline variable by bahavioral therapy and outcome

5. Models

We employed 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, resulting in a total of 84 terms. LASSO (Least Absolute Shrinkage and Selection Operator) and Best Subset regression are especially helpful for this analysis, as it performs both

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

variable selection and regularization, and enhances predictive performance in high-dimensional data.

In LASSO regression, the penalty term (ℓ_1 regularization) shrinks the coefficients of less influential predictors towards zero to prevent overfitting. By eliminating predictors with coefficients reduced to zero, LASSO is effectively performing variable selection when working with numerous predictors, including interaction terms. Additionally, when predictors are highly correlated, LASSO tends to select one variable from the correlated group, further improving model interpretability.

LASSO regression:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In Best Subset regression, the best subset of predictors is identified to minimize test MSE. The ℓ_0 penalty in Best Subset regression increases sparsity by directly controlling the number of predictors with non-zero coefficients. Therefore, we can control the number of non-zero coefficients by tuning the ℓ_0 regularization parameter (λ) in our model. We used logistic loss function and set maxSuppSize = 20 to allow a maximum of 20 non-zero coefficients in the model.

Best Subset Regression:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{i=1}^{p} 1(\beta_j \neq 0)$$

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 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=90). Table 5 summarizes key performance metrics, including accuracy, specificity, sensitivity, AUC, and Brier scores for both models.

The Brier score, which measures the accuracy of probabilistic predictions, with lower values indicating better predictive performance. For LASSO regression, the Brier score was slightly lower in the test set (0.176) compared to the training set (0.192). The Area Under the Curve (AUC) is a measure of discrimination ability, which was higher in the test set (0.840) than in the training set (0.761), suggesting that the model performed better in distinguishing between abstinent and non-abstinent cases in the test set. Overall, the test set showed better performance metrics: accuracy (0.849), specificity (0.873), and sensitivity (0.750), compared to the training set, which had an accuracy of 0.678, specificity of 0.640, and sensitivity of 0.812. This improved performance in the test set could be attributed to random variation and the small sample size of the test set. For Best Subset regression, we observe a drop in accuracy in both training data (0.596) and test data (0.683) compared to LASSO. Other metrics are behaving in similar pattern as the metrics of LASSO regression.

Further model evaluation was conducted using calibration plots. The calibration plots (Figure 2 and Figure 3) reveal discrepancies between predicted and observed probabilities in both LASSO and Best Subset. Specifically, the predicted probabilities (black dots) deviate significantly from the ideal diagonal line, indicating that the model tends to under or overestimate probabilities in certain probability ranges. This suggests potential calibration issues that may require further refinement in our models.

Table 4: Model results from LASSO and Best Subset regression

	LASSO		Best Subset	
	OR	Estimates	OR	Estimates
Non-Hispanic White	1.2099055	0.1905423	2.5147157	0.9221597
Education (Some college/technical school)	0.9491899	-0.0521464		
FTCD score	0.8626622	-0.1477320	0.7481244	-0.2901860
Major depressive disorder	0.8958597	-0.1099715		
Behavior activation x Income more than 75,000	1.0027425	0.0027388		
Varenicline x Age	1.0128308	0.0127491	1.0313394	0.0308583
Varenicline x Smoking with 5 mins of waking up	1.1743103	0.1606810		
Varenicline x Nicotine Metabolism Ratio	1.5761713	0.4549986		

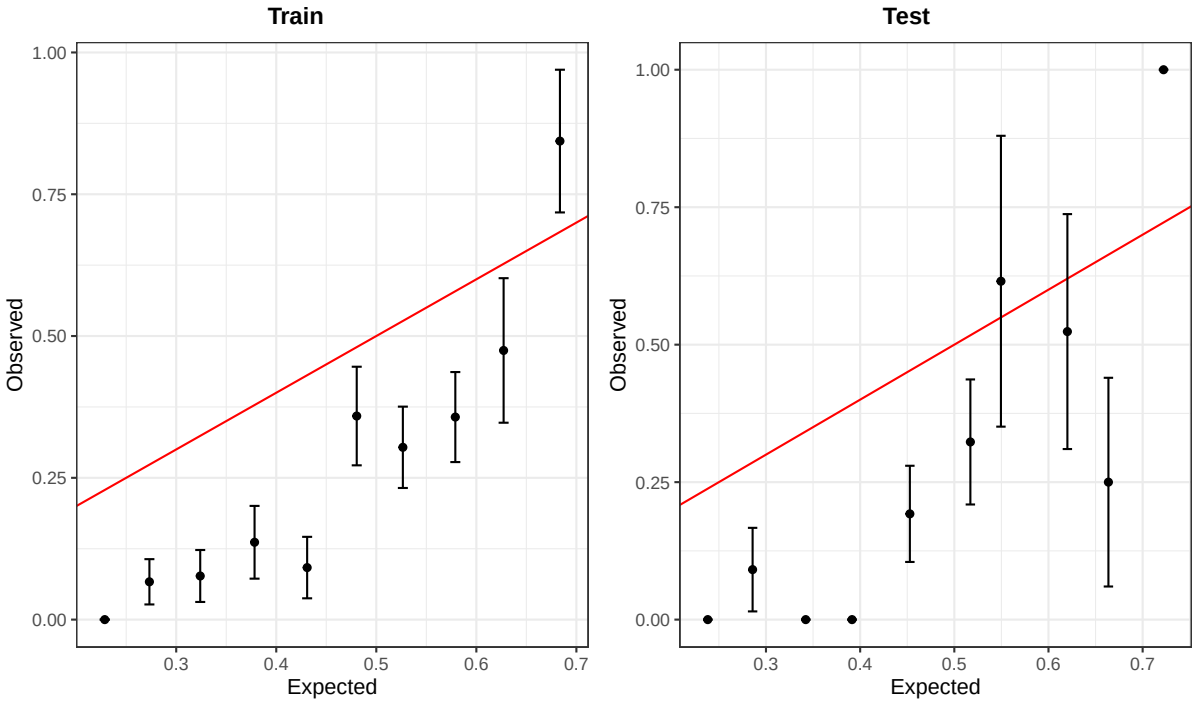

Figure 2: Calibration plots for LASSO (train vs. test)

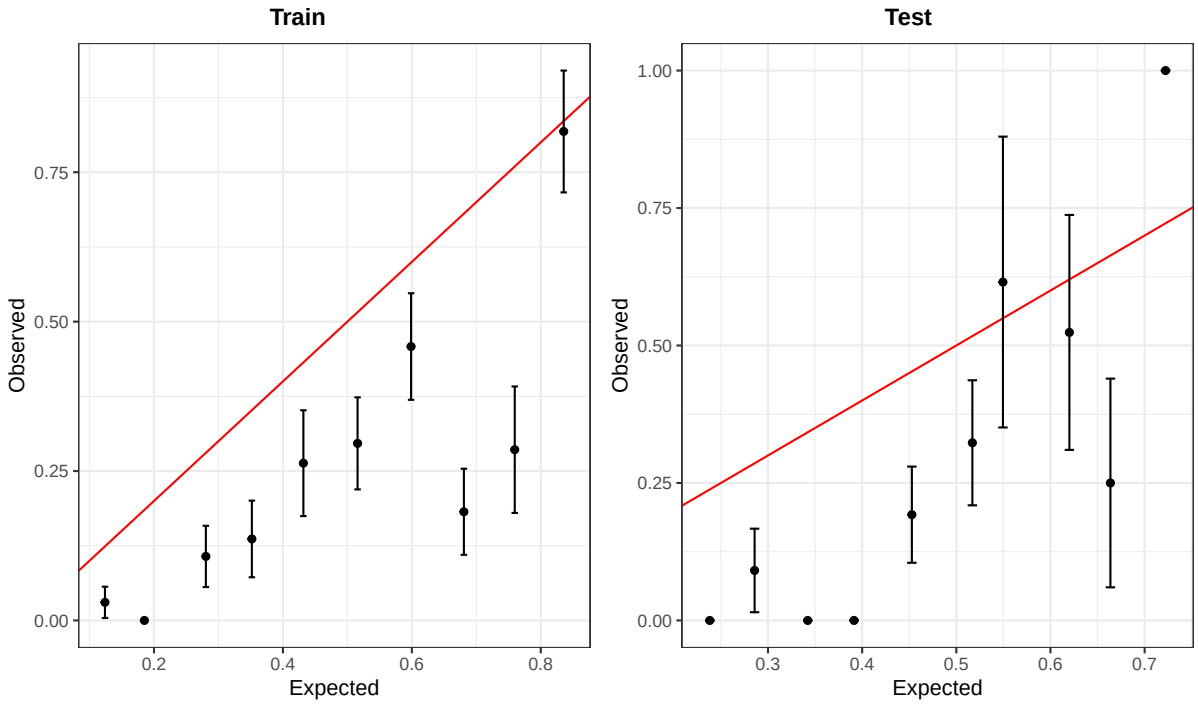

Figure 3: Calibration plots for Best Subset (train vs. test)

Table 5: Model Evaluation

	LASSO		Best Subset	
	Train	Test	Train	Test
Accuracy	0.678	0.849	0.596	0.683
Specificity	0.640	0.873	0.512	0.636
Sensitivity	0.812	0.750	0.896	0.875
AUC	0.761	0.840	0.751	0.812
Brier score	0.192	0.176	0.198	0.178

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. However, baseline characteristics

such as Non-Hispanic White, education level, FTCD score, and MDD are selected as significant predictors of smoking abstinence, independent of the effects of behavioral treatment or pharmacotherapy. Being Non-Hispanic White (OR: 1.2099) increases the odds of smoking abstinence. Conversely, having some college or technical education (OR: 0.9492) decreases the odds of smoking abstinence compared to individuals with some high school education or below. Higher FTCD score (OR: 0.8627) and a diagnosis of major depressive disorder (OR: 0.8959) are also associated with lower odds of abstinence.

Additionally, several significant moderators are observed for both behavioral and pharmacological treatments. Having an income above $75,000 slightly enhanced the effect of behavioral activation (OR: 1.0027) compared to those with an income below $20,000. The effect of varenicline varied slightly by age (OR: 1.0128). Furthermore, the impact of pharmacological treatment on smoking abstinence differed for individuals who smoked within 5 minutes of waking (OR: 1.1743) compared to those who smoked later. Notably, the interaction between varenicline and nicotine metabolism ratio (OR: 1.5762) suggests that varenicline is more effective for individuals with faster nicotine metabolism.

The Best Subset regression further emphasizes key predictors influencing smoking abstinence as it identified similar, but fewer terms that LASSO regression. Non-Hispanic White (OR: 2.5147) 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 2.5147, which is a 107.844% increase in odds ratio compared to LASSO coefficient. FTCD score (OR: 0.7481) remains as a negative predictor of the outcome.

7. Discussion

There's some agreement and disagreement between the model outputs. Non-Hispanic White and FTCD score are common variables in the models, suggesting these are strong predictors of abstinence. The interaction between varenicline and age also appeared both models, however it might not be a strong moderator given that the age is consistently positive.

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 demonstrates 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 employed LASSO regression and Best Subset regression to investigate baseline characteristics and their interactions with behavioral therapy and pharmacotherapy. Both models identified factors influencing smoking cessation. Specifically, Non-Hispanic White and FTCD score emerged as strong predictors of smoking abstinence, as they were consistently selected by both LASSO and Best Subset regression. Additionally, LASSO regression highlighted education and major depressive disorder as predictors of abstinence, while income was identified as a moderator for behavioral treatment. Morning smoking habits and nicotine metabolism ratio were found to be moderators of pharmacotherapy.

Although the model demonstrated strong discrimination ability and overall good performance on the test set, the calibration plots revealed discrepancies between predicted and observed probabilities, suggesting some limitations exist in our model. Furthermore, the absence of an external validation dataset and the small sample size limit the generalizability of these findings to other populations. Future studies with larger and more diverse samples are needed to validate these results and improve model reliability.

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 = 5,
                      fig.width = 8.5,
                      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(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
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 treatmer
  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(`Substitu
         `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
load("smoke_train.Rdata")
load("smoke_test.Rdata")
load("imp_train_long.RData")
load("imp_test_long.RData")
#////////////////////////////////////////////////////////////////////////////
#---- Lasso regression
#////////////////////////////////////////////////////////////////////////////

lasso <- function(imp_data, idx) {
```

```r
  #' 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)
  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
average_lasso_coefs <- function(imp_long, n = 5) {
```

```r
#' 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) +
    geom_abline(intercept = 0, slope = 1, color="red") +
    geom_errorbar(aes(x = expected,
                      ymin = observed - 1.96*se,
                      ymax = observed + 1.96*se),
                  colour="black", width=.01)+
    geom_point(aes(x = expected, y = observed)) +
    theme_bw() +
    labs(x = "Expected", y = "Observed")

  return(list(
    avg_coefs = coef_estimates,
    avg_auc = auc_value,
  #  roc_value = roc_value,
    accuracy = accuracy,
    brier_score = brier_score,
    calib_plot = calib_plot
  ))
}

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
```

```r
  #' @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="L0", loss="Logistic", maxSuppSize=20

  # 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$fit$gam

  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)
  }

  # 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)
```

```r
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 = 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) +
  geom_abline(intercept = 0, slope = 1, color="red") +
```

```
    geom_errorbar(aes(x = expected,
                      ymin = observed - 1.96*se,
                      ymax = observed + 1.96*se),
                  colour="black", width=.01) +
    geom_point(aes(x = expected, y = observed)) +
    theme_bw() +
    labs(x = "Expected", y = "Observed")

  return(list(
    avg_coefs = coef_estimates,
    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)
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", "Education (Some college/technical

train_coef_estimates_bestsubset <- as.data.frame(as.matrix(avg_coefs_bestsubset_train$avg_co
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", "Varenicl

# 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
```

```r
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", "Estimates", "OR", "Estimates"),
      caption ="Model results from LASSO and Best Subset regression", booktabs =T,escape =F,a
  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

# Calibration plot for test set
test_calib <- avg_coefs_lasso_test$calib_plot
combine_plot <- ggarrange(train_calib, test_calib, ncol = 2, nrow = 1)
annotate_figure(combine_plot,
                top = text_grob("Train
# Calibration plot for training set
train_calib_bs <- avg_coefs_bestsubset_train$calib_plot
test_calib_bs <-avg_coefs_lasso_test$calib_plot
combine_plot2 <- ggarrange(train_calib_bs, test_calib_bs, ncol =2,nrow =1)
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("striped", "HOLD_position"))
```