Optimizing Smoking Cessation Strategies for Adults with Major Depressive Disorder: A Predictive Analysis of Behavioral and Pharmacological Interventions

Morgan Cunningham

2024-11-07

Abstract

Smoking cessation remains particularly challenging for individuals with Major Depressive Disorder (MDD) due to increased nicotine dependence and unique psychological barriers. This study analyzed data from a 2x2 randomized, placebo-controlled trial of 300 participants with past or present MDD, evaluating the efficacy of Behavioral Activation for Smoking Cessation (BASC) and varenicline, a pharmacological intervention. Using multiple imputation to address missing data, cross-validated LASSO regression identified predictors of abstinence at the end of treatment (EOT).

Key baseline predictors included nicotine dependence (FTCD score), complementary reinforcers associated with smoking, and the Nicotine Metabolism Ratio (NMR). Interaction terms revealed that varenicline combined with higher nicotine dependence positively influenced abstinence, while behavioral activation demonstrated increased effectiveness when paired with higher readiness to quit smoking. Despite a moderate decline in validation performance (AUC = 0.703), the model retained its ability to distinguish abstinent and non-abstinent participants.

The findings suggest that personalizing smoking cessation strategies, particularly by tailoring behavioral and pharmacological treatments to individual readiness and dependence levels, can enhance outcomes for individuals with MDD. However, limitations such as the modest sample size, reliance on self-reported data, and a focus on short-term abstinence highlight the need for larger, longitudinal studies to validate and extend these results.

Introduction

Smoking cessation poses significant challenges for individuals with Major Depressive Disorder (MDD). These individuals are more likely to smoke heavily, have stronger nicotine dependence, and experience severe withdrawal symptoms. Psychological factors associated with MDD can increase the appeal of nicotine and reduce motivation to quit. Despite the importance of addressing smoking in this group, individuals with MDD are often excluded from clinical trials. This exclusion limits the evidence available to guide treatment strategies for this vulnerable population.

To address this gap, a study led by Dr. George Papandonatos explores the efficacy of combining behavioral and pharmacological treatments tailored for smokers with MDD. Behavioral Activation for Smoking Cessation (BASC), a strategy designed to enhance engagement in rewarding, health-promoting activities, has been evaluated alongside varenicline, a pharmacotherapy known to mitigate nicotine cravings. A 2x2 randomized, placebo-controlled trial assessed these approaches, comparing BASC and standard treatment (ST), with or without adjunctive varenicline, among smokers with current or past MDD. While varenicline consistently improved abstinence rates relative to placebo, BASC did not outperform ST, suggesting the need for further refinement of behavioral interventions.

This project builds on prior research by examining baseline characteristics that may influence treatment outcomes. It investigates how these factors predict end-of-treatment (EOT) abstinence, while accounting for both behavioral and pharmacological interventions. By identifying key moderators, this study aims to improve smoking cessation strategies for individuals with MDD.

Methods

In our initial EDA, we created a Table 1 (see figure appendix) showcasing baseline characteristics across the four treatment groups involved in the smoking cessation trial: BASC + placebo, BASC + varenicline, ST + placebo, and ST + varenicline. Key demographic and clinical variables are presented, alongside p-values to assess the statistical differences across groups. Age, sex, and race distribution are similar across groups with no significant differences, suggesting that the groups are relatively balanced regarding these baseline demographics.

"Taking Antidepressants" shows a statistically significant difference ($p = 0.013$), with a higher proportion in the BASC + placebo and BASC + varenicline groups. This could potentially influence treatment outcomes related to MDD symptoms and smoking cessation. Indicators like the Fagerstrom Test for Cigarette Dependence score, cigarettes per day, and smoking within five minutes of waking show no significant differences, indicating that baseline nicotine dependence is comparable among groups.

Factors such as income, education level, readiness to quit, and previous DSM-5 diagnoses do not differ significantly across groups. This helps in attributing any treatment effect observed during the study to the interventions rather than to pre-existing differences in these socioeconomic or psychiatric factors.

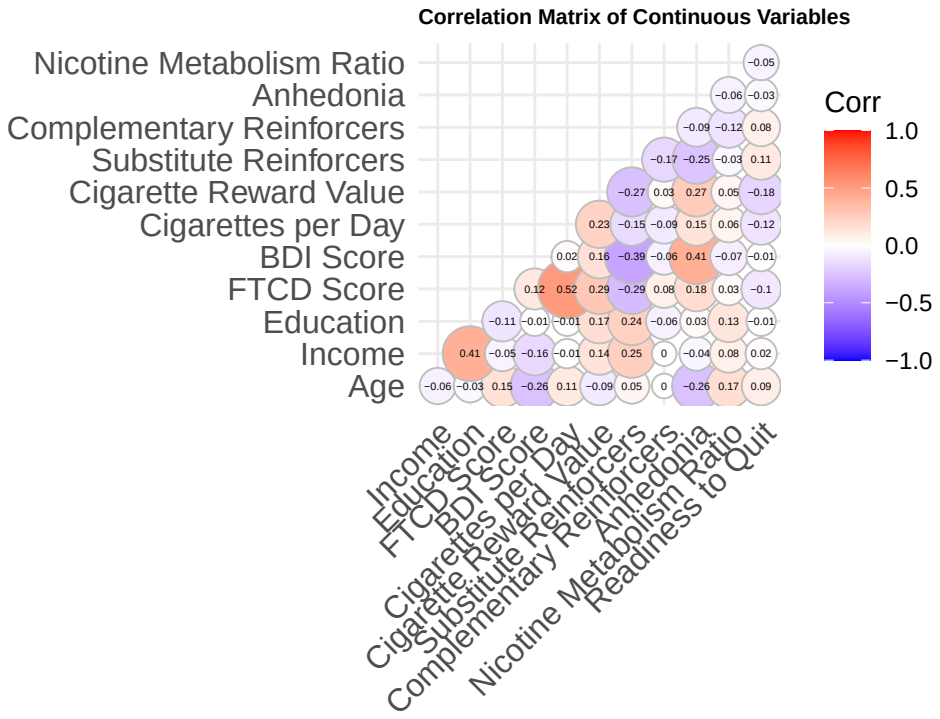
Overall, our four treatment groups are fairly well-matched for most variables. However, we do see there is a considerable amount of missing data. We look more into that next.

Table 1: Missing Data Pattern

	Number of Missing	Percentage of Missing (%)
Income	3	1
FTCD score at baseline	1	0.33
Cigarette reward value at baseline	18	6
Anhedonia	3	1
Nicotine Metabolism Ratio	21	7
Exclusive Mentholated Cigarette User	2	0.67
Baseline readiness to quit smoking	17	5.67
Total Missing Entries Overall	65	21.67

The missing data percentages are relatively low across most variables, with the exceptions of "Nicotine Metabolism Ratio" at 7% and "Baseline readiness to quit smoking" at 5.67%. No specific missing data pattern was identified so we conclude that the data are missing completely at random. The total percentage of missing entries across all variables is 21.67%. Given the small sample size of 300, nearly omitting these data values would significantly lessen statistical power, accuracy, and potentially bias the results. To handle missing data in the baseline characteristics, we used the Multiple Imputation by Chained Equations (MICE) method. This approach is particularly suited for datasets with missing values, as it estimates missing data multiple times to account for the uncertainty introduced by imputation.

Before performing imputation, we split the dataset into training and test sets to facilitate model validation. A random sample without replacement was drawn, with 70% of the data allocated to the training set for model derivation, and the remaining 30% reserved for the test set for validation. Since the split occurred before addressing missing data, multiple imputation was applied separately to the training and test sets, ensuring no data leakage between them.



In the correlation matrix of continuous variables, there is a moderate positive correlation of 0.52 between the number of cigarettes smoked per day at baseline (`cpd_ps`) and the nicotine dependence score (`ftcd_score`). This suggests that individuals who smoke more cigarettes tend to have higher dependence levels. Additionally, income (`inc`) and education (`edu`) are moderately correlated at 0.41, indicating a link between socioeconomic status and education level. Depression, as measured by the BDI score at baseline (`bdi_score_w00`), is also negatively correlated with anhedonia (`shaps_score_pq1`), with a correlation of -0.39. These negative correlations indicate that as depressive symptoms or nicotine dependence increase, the likelihood or frequency of substituting with pleasurable activities decreases. Next we will more closely look at the correlation between our possible predictor variables and dependent variable of smoking abstinence.

From Table 2, we see that there is a positive correlation of 0.16 between standard treatment with varenicline and abstinence. Non-Hispanic White ethnicity, nicotine metabolism rate, and behavioral activation combined with varenicline positively correlated and the most influential predictors for predicting abstinence. On the other hand, baseline cigarette dependence and receiving a placebo rather than active treatment are negatively associated with abstinence. Similarly, being currently diagnosed with MDD or being a heavier smoker reduces the chances of abstinence.

We will look to include 3 interactions in our model. First, since varenicline and FTCD score have opposite effects on abstinence, an interaction between these two might reveal if the effectiveness of varenicline treatment is moderated by nicotine dependence level. Second, since varenicline is positively correlated with abstinence and current MDD is negatively correlated, testing an interaction could help determine if the presence of MDD affects the efficacy of the treatment. Lastly, behavioral activation combined with a participant's readiness to quit might help determine if motivation enhances the effect of behavioral intervention.

Table 2: Correlation Between Predictors and Smoking Abstinence

	Description	Correlation
Var	Pharmacotherapy (Varenicline)	0.28
BA	Behavioral Activation	-0.04
age_ps	Age at Phone Interview	0.03
sex_ps	Sex at Phone Interview	0.01
NHW	Non-Hispanic White Indicator	0.15
Black	Black Indicator	-0.09
Hisp	Hispanic Indicator	-0.03
inc	Income (Low to High)	0.07
edu	Education (Low to High)	0.04
ftcd_score	FTCD Score at Baseline	-0.22
ftcd.5.mins	Smoking with 5 mins of Waking Up	-0.07
bdi_score_w00	BDI Score at Baseline	-0.07
cpd_ps	Cigarettes per Day at Baseline	-0.10
crv_total_pq1	Cigarette Reward Value at Baseline	-0.01
hedonsum_n_pq1	Pleasurable Events – Substitute Reinforcers	0.07
hedonsum_y_pq1	Pleasurable Events – Complementary Reinforcers	-0.05
shaps_score_pq1	Anhedonia (SHAPS Score)	-0.10
otherdiag	Other Lifetime DSM-5 Diagnosis	-0.07
antidepmed	Taking Antidepressant Medication	-0.01
mde_curr	Current vs Past MDD	-0.10
NMR	Nicotine Metabolism Ratio	0.13
Only.Menthol	Exclusive Mentholated Cigarette User	-0.05
readiness	Baseline Readiness to Quit Smoking	-0.04
st_placebo	Pharmacotherapy (Placebo)	-0.13
basc_placebo	Behavioral Activation + Placebo	-0.20
st_varenicline	Pharmacotherapy (Varenicline)	0.16
basc_varenicline	Behavioral Activation + Varenicline	0.15

Model Building

For the imputation process, we generated five imputed datasets for both the training and test sets. Missing values were estimated using predictive models tailored to each variable type. Each imputed dataset was analyzed individually, and results were aggregated to produce pooled coefficients. This pooling process accounts for both within-imputation and between-imputation variability, providing robust and unbiased parameter estimates.

Next, we applied LASSO regression on the training data set to refine predictor selection. LASSO was chosen as the selection method over other techniques due to its ability to handle multicollinearity and shrink coefficients of less important variables to zero, effectively selecting a subset of relevant predictors. In a study with numerous baseline variables, multicollinearity can complicate model interpretation and lead to inflated standard errors. By applying a penalty to the absolute size of coefficients, LASSO mitigates these issues and simplifies the model while maintaining predictive accuracy. Using cross-validation, we identified the optimal penalty parameter (`lambda.min`) that minimizes the cross-validation error.

To incorporate all imputed data sets, LASSO was applied separately to each imputed dataset. The coefficients from all imputations were pooled to identify predictors consistently selected across imputations. Variables with non-zero pooled coefficients were retained for inclusion in the final model.

Several interaction terms were included in the model to capture complex relationships between treatment effects and participant characteristics that might influence smoking cessation outcomes. Specifically, the following interactions were examined:

1. `st_varenicline:sqrt(ftcd_score)` - whether an individual received standard varenicline and the square root-transformed FTCD score.
2. `basc_varenicline:sqrt(ftcd_score)` - whether an individual received a basic version of the varenicline treatment and the square root-transformed FTCD score.
3. `st_varenicline:mde_curr` - standard varenicline treatment group and current major depressive episode.
4. `basc_varenicline:mde_curr` - basic varenicline treatment group and current major depressive episode.
5. `basc_placebo:readiness` - whether an individual received a basic placebo treatment and readiness to quit smoking.
6. `basc_varenicline:readiness` - basic varenicline treatment group and readiness to quit smoking.

By including these interaction terms, the model aims to identify specific conditions under which treatments are more or less effective.

Finally, the reduced set of predictors was used to refit the LASSO model and evaluate its performance on the test set. This approach ensures that the selected predictors are not only interpretable but also generalizable to unseen data.

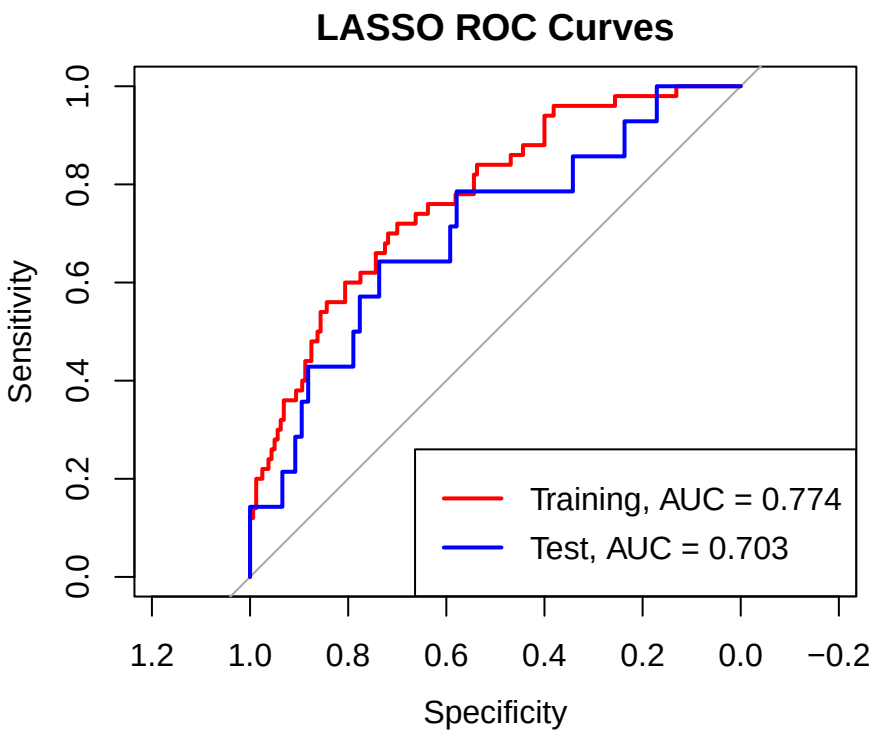
Results

Below, we list each predictor variable included in the final model. The table also displays the coefficient value for each predictor, reflecting its impact on smoking abstinence.

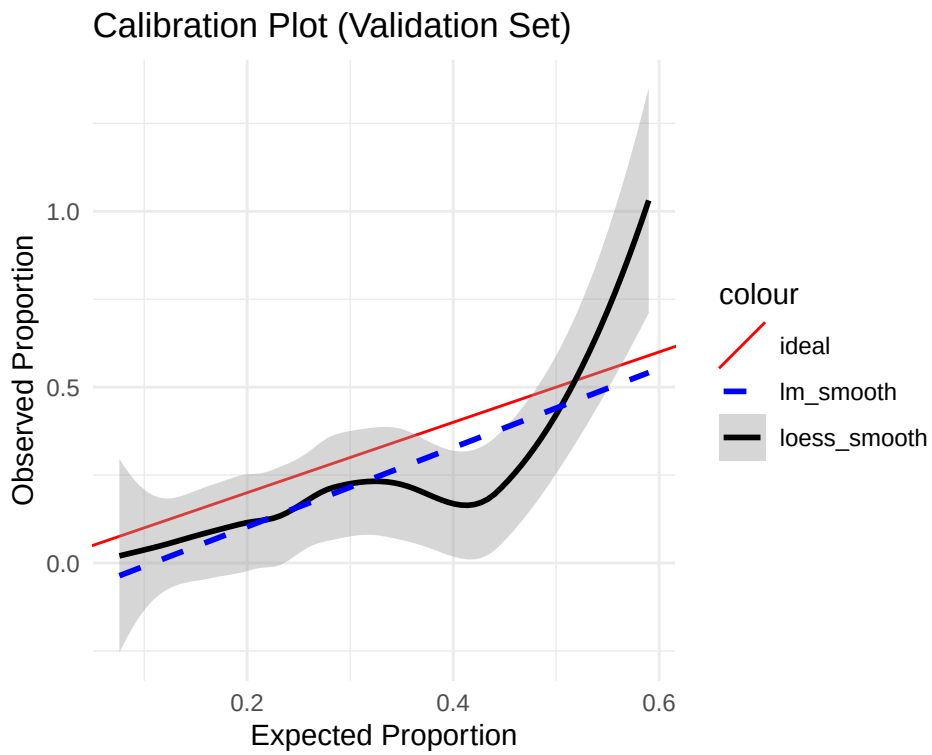
Table 3: LASSO Coefficients for Relevant Predictors

Coefficient	Variable
0.2141530700	(Intercept)
-0.1772637027	basc_placebo
-0.9197469789	sqrt(ftcd_score)
0.0005748088	age_ps
0.3506629782	NHW
-0.0003683357	hedonsum_y_pq1
-0.1556134279	otherdiag
0.6135192118	NMR
0.3089487092	st_varenicline:sqrt(ftcd_score)
0.0844195678	st_varenicline:mde_curr
0.0277146247	basc_varenicline:readiness

The LASSO regression coefficients for the selected predictors highlight their relative influence on the outcome of smoking abstinence. The intercept suggests the baseline level of the response variable when all predictors are at their reference levels. Among the predictors, `basc_placebo` has a negative coefficient (-0.177), indicating a reduced likelihood of smoking abstinence for individuals in this group. Similarly, `sqrt(ftcd_score)`, which reflects dependence on nicotine, has a strong negative association (-0.919), suggesting that higher dependence significantly reduces the probability of abstinence. Positive coefficients like `NMR` (0.613) indicate that a higher nicotine metabolism ratio improves abstinence likelihood. Interaction terms, such as `st_varenicline:sqrt(ftcd_score)` (0.309), show that specific treatment effects can be more impactful in conjunction with dependence levels. The inclusion of variables like `age_ps`, `otherdiag`, and `hedonsum_y_pq1` reflects the model's ability to capture nuanced relationships, but their small coefficients suggest a modest direct impact. Overall, the coefficients provide insight into the most influential predictors and their direction of effect, which can guide treatment strategies for the future.



This ROC curve compares the performance of the LASSO model on the training and validation sets. The AUC values are 0.774 for the training set and 0.703 for the validation set, demonstrating moderate model performance on unseen data. The training ROC curve consistently has a slightly higher sensitivity than the validation curve for a given specificity, indicating better discrimination ability on the training data. However, the decrease in AUC from training to validation suggests some overfitting, where the model performs better on the data it was trained on compared to new data. Nonetheless, the AUC of 0.703 reflects the model's ability to reasonably distinguish between positive and negative outcomes in unseen data. This balance between performance on the training and test sets demonstrates that the LASSO regression approach, with the selected predictors, captures relevant patterns in the data while avoiding severe overfitting.



The calibration plot evaluates how well the predicted probabilities align with the observed proportions in the validation set. The red diagonal line represents perfect calibration, indicating an ideal match between predicted probabilities and observed outcomes. The black LOESS smooth curve generally follows the red line but shows deviations, especially at the extremes. For lower predicted probabilities (below 0.2), the model slightly underpredicts the observed proportions, as the black line falls below the red line. In the mid-range probabilities (0.3–0.4), the model predictions are closer to ideal. However, for higher predicted probabilities (above 0.4), the model begins to overpredict, with the black line rising above the red. The dashed blue linear fit highlights these trends, showing systematic underprediction at the lower range and overprediction at the higher range. These deviations suggest the model performs reasonably well in predicting mid-range probabilities but may require further tuning to improve calibration at the extremes.

Limitations

This study has several limitations related to both the methods and the data. First, the multiple imputation approach used for handling missing data assumes that data are missing at random. If missingness depends on unobserved factors, this assumption may lead to bias in the imputations. Moreover, performing imputation separately for the training and test datasets could introduce inconsistencies in data representation. While LASSO regression is effective for variable selection, it imposes the assumption that some coefficients should shrink to zero, potentially excluding predictors with minor contributions that could impact outcomes in combination. Alternative methods, such as Elastic Net, could offer a balance between variable selection and flexibility. Although cross-validation was employed to tune the LASSO model, its reliance on random splits can introduce variability in selecting the penalty parameter and performance metrics, particularly given the relatively small sample size. Additionally, the binary outcome of smoking abstinence at the end of treatment (EOT) simplifies the evaluation of intervention success, as it does not account for relapse rates or long-term cessation outcomes. The inclusion of only a few pre-specified interaction terms may have overlooked other meaningful relationships among variables, further limiting the model's explanatory power. Finally, despite the LASSO penalty, the slight overperformance on the training data suggests a risk of overfitting, which could limit the model's generalizability.

The data itself also presents several limitations. The study population, consisting of individuals with MDD participating in a randomized trial, may not represent the broader population of smokers with different levels of depressive symptoms, demographics, or access to interventions. Imbalances in baseline characteristics, such as antidepressant usage, may confound the effects of the interventions despite efforts to randomize participants. With only 300 participants and notable missing data in key variables like nicotine metabolism ratio (7%) and readiness to quit smoking (5.67%), the small sample size reduces the diversity and statistical power needed to draw robust conclusions. Additionally, the assumption of randomness in the missing data patterns may not hold, introducing potential bias. Many variables, such as readiness to quit smoking and cigarettes per day, rely on self-reported measures, which are prone to recall bias and social desirability bias, affecting data accuracy. Simplistic metrics, such as summarizing nicotine dependence into a single FTCD score, may oversimplify complex dependencies, reducing the nuance in the dataset. Lastly, focusing solely on short-term EOT abstinence does not account for long-term outcomes, such as relapse or sustained behavioral change, limiting the scope of the findings.

Conclusion

This study sought to optimize smoking cessation strategies for individuals with Major Depressive Disorder (MDD) by identifying key baseline predictors and treatment moderators through a combination of multiple imputation and LASSO regression. The analysis revealed that baseline nicotine dependence (FTCD score), nicotine metabolism rates (NMR), and complementary reinforcers associated with smoking significantly predicted abstinence at the end of treatment (EOT). Interaction effects highlighted the importance of tailoring pharmacological and behavioral treatments to individual readiness to quit smoking and dependence levels, with varenicline proving particularly effective for participants with higher dependence.

While the model demonstrated reasonable discriminatory power (AUC = 0.703) and calibration on validation data, limitations—including the modest sample size, reliance on self-reported measures, and a focus on short-term outcomes—restrict the generalizability of these findings. Moreover, potential biases in handling missing data through imputation and class imbalance in abstinence outcomes should be addressed in future research.

This study contributes to the field by elucidating nuanced relationships between treatment effects and individual characteristics, offering valuable insights for personalizing smoking cessation interventions for individuals with MDD. To build on these findings, future research should incorporate larger, more diverse samples and extend follow-up periods to evaluate long-term abstinence and relapse rates. Addressing these limitations will further refine treatment strategies, improving outcomes for this vulnerable population.

Figure Appendix

Variable	**BASC + placebo** N = 68[1]	**BASC + varenicline** N = 83[1]	**ST + placebo** N = 68[1]	**ST + varenicline** N = 81[1]	**p-value**[2]
Age	51 (14)	50 (13)	50 (11)	49 (13)	0.7
Sex					>0.9
Female	30 (44%)	39 (47%)	29 (43%)	37 (46%)	
Male	38 (56%)	44 (53%)	39 (57%)	44 (54%)	
Non-Hispanic White					0.5
NHW	24 (35%)	34 (41%)	22 (32%)	25 (31%)	
Not NHW	44 (65%)	49 (59%)	46 (68%)	56 (69%)	

Variable	BASC + placebo N = 68[1]	BASC + varenicline N = 83[1]	ST + placebo N = 68[1]	ST + varenicline N = 81[1]	p-value[2]
Black					0.3
Black	37 (54%)	37 (45%)	40 (59%)	43 (53%)	
Not Black	31 (46%)	46 (55%)	28 (41%)	38 (47%)	
Hispanic					>0.9
Hispanic	5 (7.4%)	4 (4.8%)	4 (5.9%)	5 (6.2%)	
Not Hispanic	63 (93%)	79 (95%)	64 (94%)	76 (94%)	
Income (low to high)					0.8
1	25 (37%)	30 (37%)	26 (38%)	29 (36%)	
2	16 (24%)	17 (21%)	14 (21%)	21 (26%)	
3	8 (12%)	13 (16%)	14 (21%)	11 (14%)	
4	12 (18%)	12 (15%)	8 (12%)	6 (7.5%)	
5	6 (9.0%)	10 (12%)	6 (8.8%)	13 (16%)	
Missing	1	1	0	1	
Education (low to high)					
1	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)	
2	3 (4.4%)	7 (8.4%)	2 (2.9%)	4 (4.9%)	
3	23 (34%)	15 (18%)	11 (16%)	27 (33%)	
4	22 (32%)	32 (39%)	38 (56%)	24 (30%)	
5	19 (28%)	29 (35%)	17 (25%)	26 (32%)	
FTCD Score	5.31 (2.02)	5.07 (2.34)	5.39 (2.09)	5.17 (2.08)	0.7
Missing	0	0	1	0	
Smoking within 5 mins	32 (47%)	33 (40%)	35 (51%)	38 (47%)	0.5
BDI Score	19 (12)	18 (11)	18 (11)	20 (12)	>0.9
Cigarettes per Day	16 (9)	16 (9)	15 (7)	14 (7)	>0.9
Cigarette Reward Value	7.4 (3.8)	7.2 (3.9)	7.0 (3.7)	7.1 (3.5)	>0.9
Missing	1	3	8	6	
Pleasurable Events Scale (Substitute)	23 (20)	23 (19)	21 (20)	23 (19)	0.6

Variable	BASC + placebo N = 68[1]	BASC + varenicline N = 83[1]	ST + placebo N = 68[1]	ST + varenicline N = 81[1]	p-value[2]
Pleasurable Events Scale (Complementary)	28 (22)	22 (17)	27 (20)	25 (19)	0.3
Anhedonia Score	2.15 (3.23)	2.25 (3.12)	2.51 (3.38)	2.11 (3.00)	0.8
Missing	2	0	1	0	
Other DSM-5 Diagnosis	35 (51%)	30 (36%)	28 (41%)	40 (49%)	0.2
Taking Antidepressants	28 (41%)	24 (29%)	15 (22%)	15 (19%)	0.013
Current MDD	32 (47%)	40 (48%)	31 (46%)	44 (54%)	0.7
Nicotine Metabolism Ratio	0.34 (0.18)	0.38 (0.25)	0.37 (0.27)	0.36 (0.21)	>0.9
Missing	7	3	2	9	
Mentholated Cigarette User	40 (59%)	48 (59%)	43 (64%)	47 (58%)	0.9
Missing	0	1	1	0	
Readiness to Quit					
3	1 (1.6%)	0 (0%)	0 (0%)	0 (0%)	
4	2 (3.1%)	2 (2.6%)	1 (1.6%)	0 (0%)	
5	6 (9.4%)	11 (14%)	9 (14%)	9 (12%)	
6	18 (28%)	22 (28%)	14 (22%)	29 (38%)	
7	16 (25%)	21 (27%)	16 (25%)	18 (23%)	
8	17 (27%)	20 (26%)	19 (30%)	18 (23%)	
9	2 (3.1%)	1 (1.3%)	2 (3.1%)	2 (2.6%)	
10	2 (3.1%)	1 (1.3%)	3 (4.7%)	1 (1.3%)	
Missing	4	5	4	4	

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

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

Code Appendix

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

```
                      error=FALSE,
                      echo = FALSE,
                      fig.pos = "H",
                      fig.align = 'center')
setwd("/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 2")
project2 <- read.csv("project2.csv")
# libraries
library(dplyr)
library(tidyverse)
library(gtsummary)
library(kableExtra)
library(mice)
library(readr)
library(cardx)
library(ggcorrplot)
library(glmnet)
library(mice)
library(caret)
library(pROC)
library(MASS)
library(knitr)
library(kableExtra)
library(ggcorrplot)
# treatment columns
project2 <- project2 %>%
  mutate(
    st_placebo = case_when(BA == 0 & Var == 0 ~ 1, TRUE ~ 0),
    basc_placebo = case_when(BA == 1 & Var == 0 ~ 1, TRUE ~ 0),
    st_varenicline = case_when(BA == 0 & Var == 1 ~ 1, TRUE ~ 0),
    basc_varenicline = case_when(BA == 1 & Var == 1 ~ 1, TRUE ~ 0)
  )
write_csv(project2, "/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 2/project
# rows with at least one NA
rows_with_na <- apply(is.na(project2), 1, any)
# select cols w/ at least one NA
cols_with_na <- apply(is.na(project2), 2, any)
data_with_na <- project2[rows_with_na, cols_with_na]
# number and % of NA in each column and display table
num_na <- lapply (data_with_na,function(x) sum(is.na(x)))
pct_na <- round(as.numeric (num_na) /dim (project2) [1]*100,2)
missing_pattern <- cbind (num_na,pct_na)

# total missing values and percentage
total_missing <- sum(unlist(num_na))
total_pct_missing <- round((total_missing / (300) * 100), 2)
missing_pattern <- rbind(missing_pattern, Total = c(total_missing, total_pct_missing))

rownames (missing_pattern) <- c("Income", "FTCD score at baseline",
                                "Cigarette reward value at baseline", "Anhedonia",
                                "Nicotine Metabolism Ratio",
                                "Exclusive Mentholated Cigarette User",
                                "Baseline readiness to quit smoking",
                                "Total Missing Entries Overall")
```

```
kable (missing_pattern,caption = "Missing Data Pattern", booktabs=T, escape = T,
       col.names = c ("Number of Missing", "Percentage of Missing (%)"),
       align = "c" )
# train-test split (70% training, 30% testing)
set.seed(123)
train_index <- createDataPartition(project2$abst, p = 0.7, list = FALSE)
train_data <- project2[train_index, ]
test_data <- project2[-train_index, ]

# apply MICE to the training set
train_mice <- mice(train_data, m = 5, seed = 123, maxit = 5, print = FALSE)

# apply MICE to the test set
test_mice <- mice(test_data, m = 5, seed = 123, maxit = 5, print = FALSE)
labels <- c(
  age_ps = "Age",
  inc = "Income",
  edu = "Education",
  ftcd_score = "FTCD Score",
  bdi_score_w00 = "BDI Score",
  cpd_ps = "Cigarettes per Day",
  crv_total_pq1 = "Cigarette Reward Value",
  hedonsum_n_pq1 = "Substitute Reinforcers",
  hedonsum_y_pq1 = "Complementary Reinforcers",
  shaps_score_pq1 = "Anhedonia",
  NMR = "Nicotine Metabolism Ratio",
  readiness = "Readiness to Quit"
)

continuous_vars <- project2[, names(labels)]

cor_matrix <- cor(continuous_vars, use = "complete.obs")

rownames(cor_matrix) <- labels[names(labels)]
colnames(cor_matrix) <- labels[names(labels)]

ggcorrplot(cor_matrix,
           method = "circle",
           type = "lower",
           lab = TRUE,
           lab_size = 1.5,
           colors = c("blue", "white", "red"),
           title = "Correlation Matrix of Continuous Variables",
           ggtheme = theme_minimal() +
             theme(
               axis.text.x = element_text(size = 2, angle = 45, hjust = 1),
               axis.text.y = element_text(size = 2),
               plot.title = element_text(size = 8, face = "bold"),
               plot.margin = margin(1, 1, 1, 1, "cm")
               ))

# correlations between predictors and abst
# pairwise correlations between predictors and abst
```

```r
correlations <- sapply(project2[, !names(project2) %in% c("id", "abst")],
                       function(x) cor(x, project2$abst, use = "complete.obs"))

variable_descriptions <- c(
  Var = "Pharmacotherapy (Varenicline)",
  BA = "Behavioral Activation",
  age_ps = "Age at Phone Interview",
  sex_ps = "Sex at Phone Interview",
  NHW = "Non-Hispanic White Indicator",
  Black = "Black Indicator",
  Hisp = "Hispanic Indicator",
  inc = "Income (Low to High)",
  edu = "Education (Low to High)",
  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",
  crv_total_pq1 = "Cigarette Reward Value at Baseline",
  hedonsum_n_pq1 = "Pleasurable Events - Substitute Reinforcers",
  hedonsum_y_pq1 = "Pleasurable Events - Complementary Reinforcers",
  shaps_score_pq1 = "Anhedonia (SHAPS Score)",
  otherdiag = "Other Lifetime DSM-5 Diagnosis",
  antidepmed = "Taking Antidepressant Medication",
  mde_curr = "Current vs Past MDD",
  NMR = "Nicotine Metabolism Ratio",
  Only.Menthol = "Exclusive Mentholated Cigarette User",
  readiness = "Baseline Readiness to Quit Smoking",
  st_placebo = "Pharmacotherapy (Placebo)",
  basc_placebo = "Behavioral Activation + Placebo",
  st_varenicline = "Pharmacotherapy (Varenicline)",
  basc_varenicline = "Behavioral Activation + Varenicline"
)

correlations_table <- data.frame(
  Predictor = names(correlations),
  Correlation = round(correlations, 2)
)

correlations_table <- correlations_table %>%
  mutate(Description = variable_descriptions[as.character(Predictor)]) %>%
  dplyr::select(Description, Correlation)


correlations_table %>%
  kbl(caption = "Correlation Between Predictors and Smoking Abstinence") %>%
  kable_styling(
    bootstrap_options = c("striped", "hover", "condensed"),
    latex_options = c("hold_position")
  ) %>%
  column_spec(1, bold = TRUE) %>%
  row_spec(0, bold = TRUE, color = "white", background = "blue")
m <- 5
```

```r
# list to store Lasso coefficients
lasso_coefficients_list <- list()

# loop through all imputed datasets
for (i in 1:m) {
  train_complete <- complete(train_mice, action = i)

  # training data for Lasso with the specified model
  x_train <- model.matrix(
    abst ~ st_placebo + basc_placebo + st_varenicline + basc_varenicline +
      sqrt(ftcd_score) + sqrt(bdi_score_w00) + age_ps + sex_ps + NHW + Black +
      Hisp + inc + edu + ftcd.5.mins + cpd_ps + crv_total_pq1 +
      hedonsum_n_pq1 + hedonsum_y_pq1 + shaps_score_pq1 + otherdiag +
      antidepmed + mde_curr + NMR + Only.Menthol + readiness +
      st_varenicline:sqrt(ftcd_score) + basc_varenicline:sqrt(ftcd_score) +
      st_varenicline:mde_curr + basc_varenicline:mde_curr +
      basc_placebo:readiness + basc_varenicline:readiness,
    data = train_complete
  )[, -1]

  y_train <- train_complete$abst

  # cross-validated Lasso regression
  set.seed(123)
  lasso_model <- cv.glmnet(
    x = x_train,
    y = y_train,
    family = "binomial",
    alpha = 1,
    type.measure = "deviance"
  )

  # fit final Lasso model using  best lambda
  best_lambda <- lasso_model$lambda.min
  final_model <- glmnet(
    x = x_train,
    y = y_train,
    family = "binomial",
    alpha = 1,
    lambda = best_lambda
  )

  lasso_coefficients_list[[i]] <- coef(final_model)
}

# combine coefficients from all imputations
pooled_coefficients <- Reduce("+", lasso_coefficients_list) / m

# convert to a data frame for readability
pooled_coefficients_df <- as.data.frame(as.matrix(pooled_coefficients))
pooled_coefficients_df$Variable <- rownames(pooled_coefficients_df)
rownames(pooled_coefficients_df) <- NULL
colnames(pooled_coefficients_df)[1] <- "Coefficient"
```

```r
# consistently selected variables (non-zero coefficients)
selected_variables <- pooled_coefficients_df %>%
  filter(Coefficient != 0)
# reduced formula
reduced_formula <- as.formula(paste("abst ~", paste(selected_variables$Variable[-1], collapse = "+")))

# reduced matrices for training and test sets
train_complete <- complete(train_mice, action = 1)
test_complete <- complete(test_mice, action = 1)

x_train_reduced <- model.matrix(reduced_formula, data = train_complete)[, -1]
y_train <- train_complete$abst

x_test_reduced <- model.matrix(reduced_formula, data = test_complete)[, -1]
y_test <- test_complete$abst

# LASSO model using the reduced predictors
final_lasso_model <- glmnet(
  x = x_train_reduced,
  y = y_train,
  family = "binomial",
  alpha = 1,
  lambda = best_lambda
)

# predict on test set
test_predictions <- predict(final_lasso_model, newx = x_test_reduced, type = "response")
test_predicted_classes <- ifelse(test_predictions >= 0.5, 1, 0)

# # Evaluate performance
# conf_matrix <- confusionMatrix(factor(test_predicted_classes), factor(y_test))
# print(conf_matrix)
#
# # Additional metrics (optional)
# roc_curve <- roc(y_test, test_predictions)
# auc <- auc(roc_curve)
# print(paste("AUC:", auc))
# lasso coefficents
lasso_coefs_df <- as.data.frame(as.matrix(selected_variables))

lasso_coefs_df %>% kable(col.names = c("Coefficient", "Variable"),
                         caption = "LASSO Coefficients for Relevant Predictors",
                         align = "c") %>%
  kable_styling(full_width = F, bootstrap_options = c("striped", "hover", "condensed")) %>%
column_spec(1, bold = T) %>%
row_spec(0, background = "#D3D3D3")

# predictions for ROC curves
train_preds <- predict(final_lasso_model, newx = x_train_reduced, type = "response")
train_roc <- roc(y_train, train_preds)

test_preds <- predict(final_lasso_model, newx = x_test_reduced, type = "response")
test_roc <- roc(y_test, test_preds)
```

```r
# plot ROC curves
plot(train_roc, col = "red", main = "LASSO ROC Curves", lwd = 2)
lines(test_roc, col = "blue", lwd = 2)

# calculate AUC
auc_train <- auc(train_roc)
auc_test <- auc(test_roc)

legend(
  "bottomright",
  legend = c(
    paste("Training, AUC =", round(auc_train, 3)),
    paste("Test, AUC =", round(auc_test, 3))
  ),
  col = c("red", "blue"),
  lwd = 2
)

val_preds_prob <- predict(final_lasso_model, s = "lambda.min", newx = x_test_reduced, type = "response")

test_calib <- data.frame(
  prob = val_preds_prob,  # predicted probabilities on validation set
  bin = cut(val_preds_prob, breaks = 50),  # split predicted probabilities into 50 bins
  class = y_test  # actual outcome (assumed binary: 1 or 0)
)

colnames(test_calib)[colnames(test_calib) == "s1"] <- "prob"

test_calib_summary <- test_calib %>%
  group_by(bin) %>%
  summarize(
    observed = sum(class) / n(),  # positive cases in each bin
    expected = mean(prob),  # predicted probability in each bin
    se = sqrt(observed * (1 - observed) / n())  # standard error for plotting
  ) %>%
  ungroup()

cols <- c("ideal" = "red", "loess_smooth" = "black", "lm_smooth" = "blue")

ggplot(test_calib_summary) +
  geom_abline(aes(intercept = 0, slope = 1, color = "ideal")) +
  geom_smooth(aes(x = expected, y = observed, color = "loess_smooth"), se = TRUE) +
  geom_smooth(aes(x = expected, y = observed, color = "lm_smooth"), se = FALSE,
              method = "lm", linetype = "dashed") +
  scale_color_manual(values = cols) +
  labs(x = "Expected Proportion", y = "Observed Proportion",
       title = "Calibration Plot (Validation Set)") +
theme_minimal()
data <- read_csv("project2.csv")

# treatment_arm variable based on BA and Var
data <- data %>%
  mutate(
```

```
    treatment_arm = case_when(
      BA == "Standard Treatment" & Var == "Placebo" ~ "ST + placebo",
      BA == "Behavioral Activation" & Var == "Placebo" ~ "BASC + placebo",
      BA == "Standard Treatment" & Var == "Varenicline" ~ "ST + varenicline",
      BA == "Behavioral Activation" & Var == "Varenicline" ~ "BASC + varenicline",

      BA == 0 & Var == 0 ~ "ST + placebo",
      BA == 1 & Var == 0 ~ "BASC + placebo",
      BA == 0 & Var == 1 ~ "ST + varenicline",
      BA == 1 & Var == 1 ~ "BASC + varenicline"
    ),
    sex_ps = recode(sex_ps, `0` = "Male", `1` = "Female", .default = "Male"),
    NHW = factor(recode(NHW, `0` = "Not NHW", `1` = "NHW", .default = "Missing")),
    Black = factor(recode(Black, `0` = "Not Black", `1` = "Black", .default = "Missing")),
    Hisp = factor(recode(Hisp, `0` = "Not Hispanic", `1` = "Hispanic", .default = "Missing")),
    Only.Menthol = factor(recode(Only.Menthol, `0` = "No", `1` = "Yes", .default = "Missing"))
  )

# summary table stratified by treatment arm
table1 <- data %>%
  tbl_summary(
    by = treatment_arm,
    include = c(age_ps, sex_ps, NHW, Black, Hisp, inc, edu, ftcd_score, ftcd.5.mins,
                bdi_score_w00, cpd_ps, crv_total_pq1, hedonsum_n_pq1,
                hedonsum_y_pq1, shaps_score_pq1, otherdiag, antidepmed,
                mde_curr, NMR, Only.Menthol, readiness),
    label = list(
      age_ps ~ "Age",
      sex_ps ~ "Sex",
      NHW ~ "Non-Hispanic White",
      Black ~ "Black",
      Hisp ~ "Hispanic",
      inc ~ "Income (low to high)",
      edu ~ "Education (low to high)",
      ftcd_score ~ "FTCD Score",
      ftcd.5.mins ~ "Smoking within 5 mins",
      bdi_score_w00 ~ "BDI Score",
      cpd_ps ~ "Cigarettes per Day",
      crv_total_pq1 ~ "Cigarette Reward Value",
      hedonsum_n_pq1 ~ "Pleasurable Events Scale (Substitute)",
      hedonsum_y_pq1 ~ "Pleasurable Events Scale (Complementary)",
      shaps_score_pq1 ~ "Anhedonia Score",
      otherdiag ~ "Other DSM-5 Diagnosis",
      antidepmed ~ "Taking Antidepressants",
      mde_curr ~ "Current MDD",
      NMR ~ "Nicotine Metabolism Ratio",
      Only.Menthol ~ "Mentholated Cigarette User",
      readiness ~ "Readiness to Quit"
    ),
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
```

```
    missing = "ifany",
    missing_text = "Missing"
  ) %>%
  add_p() %>%
  modify_header(label ~ "**Variable**") %>%
  bold_labels()

table1 %>%
  as_flex_table() %>%
  flextable::width(width = 1.1)
```