Moderators and Predictors Analysis for Smoking Abstinence with Behavioral Treatment And Pharmacotherapy

Tianna Chan

2024-11-11

Abstract

Background: Smoking cessation is particularly challenging for individuals with major depressive disorder (MDD). Pharmacotherapy by using the drug Varenicline is effective for aiding smoking cessation, while behavior treatments may help depression and impact cessation success among MDD smokers. This project builds on prior research by examining the baseline characteristics as potential moderators and identify predictors of end-of-treatment (EOT) abstinence among adults with MDD.

Methods: Data were processed with transformations for normality and multiple imputation for missing values. To explore baseline moderators of behavior treatment effects on abstinence, lasso models focusing on main effects and treatment interactions were fitted on each imputed dataset for robust variable selection across imputed datasets. The fitted models were pooled to a final model with key predictors identified. Model results are compared between the data with and withour transformation.

Results: Exclusively menthol cigarette use negatively moderated the effectiveness of behavioral activation. FTCD Score, interaction between Varenicline and NMR, being Non-Hispanic White, and the interaction between Varenicline and Age are the strongest predictors that meaningfully influence smoking abstinence outcomes in individuals with MDD.

Conclusion: The reduced effectiveness of behavioral activation among menthol users highlights a need for targeted behavioral activation treatment for exclusive menthol smokers. Key predictors of abstinence shows the importance of biological, psychological, and behavioral factors in intervention design. These findings suggests that a tailored behavioral treatment for MDD smokers could be beneficial in improving smoking cessation success.

Introduction

Among the preventable cause of death reasons in the United States, smoking remains the top of the list [1], and individuals with major depressive disorder (MDD) are more likely to smoke heavily, exhibit greater nicotine dependence, and experience worse withdrawal symptoms than those who do not have MDD [2]. While using the widely-used Varenicline to help people stop smoking, depression-targeted behavioral treatments may also help improve the rates of MDD smokers quit smoking.

A previous randomized, placebo-controlled study used a 2x2 factorial design that compared two behavioral treatments - behavioral activation for smoking cessation (BASC) versus standard behavioral treatment (ST) and Varenicline versus placebo [3]. The study results indicated that BASC did not significantly outperform ST in promoting smoking cessation, regardless of varenicline usage.

Based on these findings, this project seeks to further examine whether any specific baseline characteristics may moderate the effects of behavioral treatments on end-of-treatment (EOT) abstinence. Additionally, we aim to identify the baseline predictors of abstinence while controlling for behavioral treatment and pharmacotherapy. With the identified moderators and predictors, this project aims to create a more personalized intervention for MDD smokers that could potentially enhance abstinence outcomes by targeting on certain individual demographics and/or characteristics.

Methods

Study Data

The data consists of 300 participants, with their characteristics at baseline, treatment group, and the end-of-treatment (EOT) smoking abstinence from the Hitsman study. The study took place in research clinics at two urban universities in the United States. The baseline characteristics includes demographic factors such as age, sex, race, and socioeconomic indicators like income and education. Additional clinical and behavioral factors are also accounted, including baseline nicotine dependence, depression symptoms, smoking habits, and reward valuation associated with smoking. Indicators of prior diagnoses of major depressive disorder (MDD), antidepressant medication use, and readiness to quit smoking are also included to provide a comprehensive view of psychological and behavioral readiness. The data also included biological markers such as nicotine metabolism ratio (NMR) and preference for exclusively menthol cigarettes help capture individual differences that may impact treatment outcomes.

Data Preprocessing and Summary

To process the data, we factored the categorical variables and created ordering for those ordinal variables. For example, income and education are ordered from low to high with 5 subgroups each. We will further combine the first two groups of education into one because of insufficient data for the first group (See ***Table 3***). Thus, only four education groups will be considered in the analyses.

Transformations were then done to ensure normality distribution for some specific variables. A square root transformation was applied to cigarette per day, substitute, and complementary reinforcers. The Nicotine Metabolism Ratio (NMR) was log-transformed to handle skewness in its distribution. Additionally, the Anhedonia measure was log-transformed with an adjustment of adding one to account for any zero values. With these transformations, variables are normalized and less skewed, which will be tested in whether it will improve the statistical modeling and the robustness of the analysis in later sections.

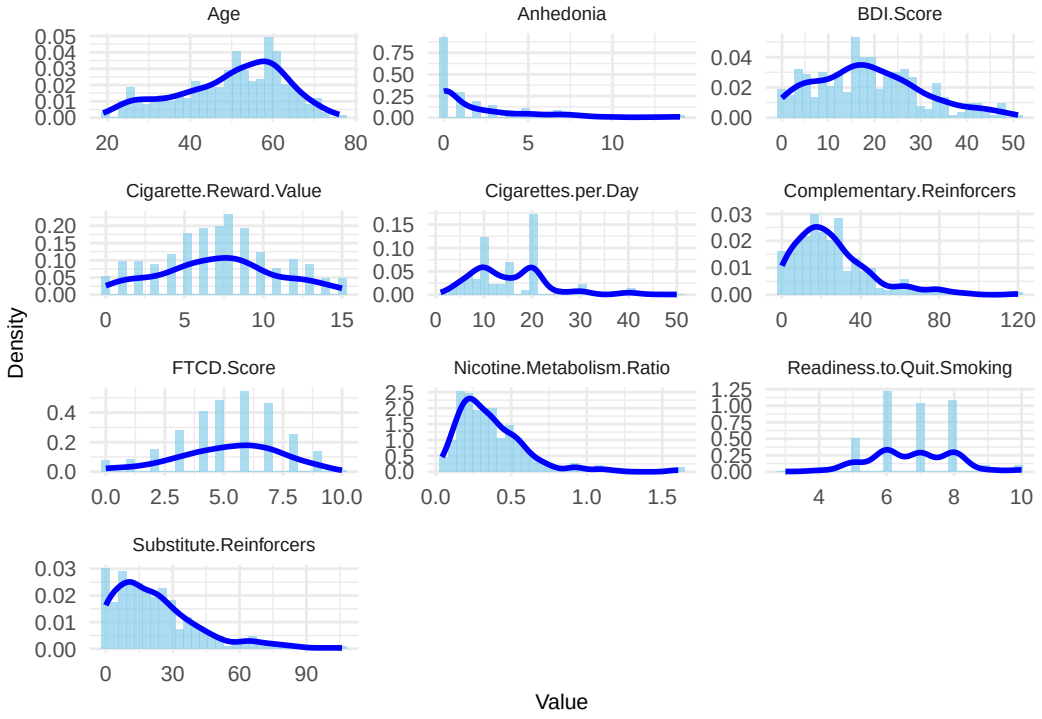

Figure 1: Distributions of Numeric Variables

The correlation plot below (***Figure 2***) provides a visual representation of all the numeric variables in the transformed data. The number represents the correlation coefficient between each pair, with color corresponding to the level of correlation, orange to blue from negative to positive correlation. Notably, there is a relatively higher correlation (0.52) between the

nicotine dependence (FTCD score) and cigarettes per day. This suggests that individuals who consume more cigarettes daily tend to have higher nicotine dependence scores. Another significant positive correlation is between Anhedonia and baseline depression score (0.40). This means that with higher levels of Anhedonia, they tend to have higher baseline depression scores. The plot shows that the relationships between variables do not have significant high correlations that would suggest multicollinearity issues, which means each variable is likely to provide unique information. Therefore, we will retain to consider all variables in the following analysis.

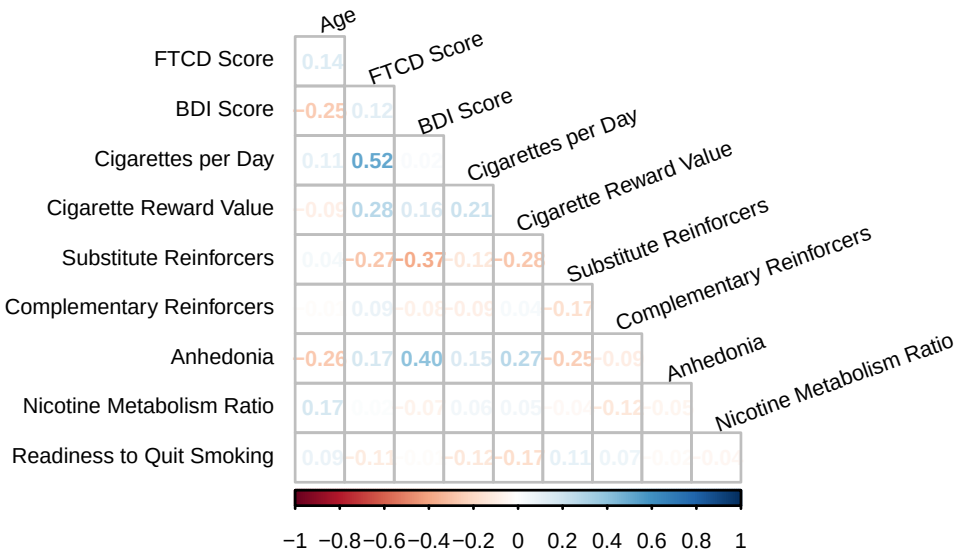

Figure 2: Correlation plot

Missing Data

Most baseline variables have 0% missing values, with only 7 columns having missing values (See ***Table 1***). The column with highest missingness is the Nicotine Metabolism Ratio (7%). Then, we have about 6% missing data in both baseline readiness to quit smoking and cigarette reward value. The columns for income and anhedonia have 1% of missing data, while only menthol preferences and the FTCD score has 0.667% and 0.333% respectively. The overall missing data in this data is 1.3%. It is plausible that these missingness in the aforementioned columns are due to participant characteristics. The low overall missingness also makes it seem unlikely that the missing values are due to systematic patterns on unobserved values. Therefore, we can assume that they are missing at random (MAR), and we proceed with multiple imputation using the *mice* package in R. We created 5 multiple imputed datasets with the predictive mean matching method. This accounts for the uncertainty in these missing values, which allows a more accurate estimates in subsequent analyses and avoid dropping data.

Table 1: Missing Data Table

Variable	Number of Missing	Percent of Missing
Nicotine.Metabolism.Ratio	21	7
Cigarette.Reward.Value	18	6
Readiness.to.Quit.Smoking	17	5.67
Income	3	1
Anhedonia	3	1
Exclusive.Mentholated.Cigarette.User	2	0.667
FTCD.Score	1	0.333

Data Summary

Using the pre-transformed data, ***Table 2*** below shows the end of treatment smoking apps to statistics for each treatment groups, we can find the biggest percentage of smoking abstinence among those groups that used Varenicline. Among the placebo groups, the group with standard behavioral treatment has a surprisingly higher percentage.

Table 2: EOT Abstinence Results

	ST+Placebo	ST+Varenicline	BA+Placebo	BA+Varenicline
n	68.00	81.0	68.00	83.00
Smoking Abstinence	8.00	26.0	4.00	26.00
Percentage (%)	11.76	32.1	5.88	31.33

Table 3 below shows the baseline demographics and characteristics of the participants. The proportions for most of them seems to be balanced between groups. However, for education, there is only 1 person categorized into 1- Grade school. Thus, we re-categorized education into only four groups: 1) Grade school/Some high school; 2) High school grad/GED; 3) Some college/technical school; and 4) College graduate in future analyses. Note that there is also no significant imbalances for baseline characteristics, except whether they are taking antidepressant mediaction - We see a significant higher percentage of those taking in the behavioral treatment and placebo group.

Table 3: Baseline Characteristics Summary

Characteristic	**Overall** N = 300	**ST+Placebo** N = 68	**ST+Varenicline** N = 81	**BA+Placebo** N = 68	**BA+Varenicline** N = 83
Age	50 (13)	50 (11)	49 (13)	51 (14)	50 (13)
Sex	1.55 (0.50)	1.57 (0.50)	1.54 (0.50)	1.56 (0.50)	1.53 (0.50)
Non-Hispanic White	0.35 (0.48)	0.32 (0.47)	0.31 (0.46)	0.35 (0.48)	0.41 (0.49)
Black	0.52 (0.50)	0.59 (0.50)	0.53 (0.50)	0.54 (0.50)	0.45 (0.50)
Hispanic	0.06 (0.24)	0.06 (0.24)	0.06 (0.24)	0.07 (0.26)	0.05 (0.22)
Income					
1-Less than $20,000	110 (37%)	26 (38%)	29 (36%)	25 (37%)	30 (37%)
2-$20,000–35,000	68 (23%)	14 (21%)	21 (26%)	16 (24%)	17 (21%)
3-$35,001–50,000	46 (15%)	14 (21%)	11 (14%)	8 (12%)	13 (16%)
4-$50,001–75,000	38 (13%)	8 (12%)	6 (7.5%)	12 (18%)	12 (15%)
5-More than $75,000	35 (12%)	6 (8.8%)	13 (16%)	6 (9.0%)	10 (12%)
Unknown	3	0	1	1	1
Education					
1-Grade school	1 (0.3%)	0 (0%)	0 (0%)	1 (1.5%)	0 (0%)
2-Some high school	16 (5.3%)	2 (2.9%)	4 (4.9%)	3 (4.4%)	7 (8.4%)
3-High school grad/GED	76 (25%)	11 (16%)	27 (33%)	23 (34%)	15 (18%)
4-Some college/tech school	116 (39%)	38 (56%)	24 (30%)	22 (32%)	32 (39%)
5-College graduate	91 (30%)	17 (25%)	26 (32%)	19 (28%)	29 (35%)
FTCD Score	5.22 (2.14)	5.39 (2.09)	5.17 (2.08)	5.31 (2.02)	5.07 (2.34)
Unknown	1	1	0	0	0
Smoking with 5 mins of waking up	138 (46%)	35 (51%)	38 (47%)	32 (47%)	33 (40%)
BDI Score	19 (11)	18 (11)	20 (12)	19 (12)	18 (11)
Cigarettes/day	15 (8)	15 (7)	14 (7)	16 (9)	16 (9)
Cigarette reward value	7.2 (3.7)	7.0 (3.7)	7.1 (3.5)	7.4 (3.8)	7.2 (3.9)
Unknown	18	8	6	1	3
Substitute Reinforcers	23 (20)	21 (20)	23 (19)	23 (20)	23 (19)
Complementary Reinforcers	25 (19)	27 (20)	25 (19)	28 (22)	22 (17)
Anhedonia	2.25 (3.16)	2.51 (3.38)	2.11 (3.00)	2.15 (3.23)	2.25 (3.12)
Unknown	3	1	0	2	0
Other lifetime DSM-5 diagnosis	133 (44%)	28 (41%)	40 (49%)	35 (51%)	30 (36%)
Taking antidepressant medication	82 (27%)	15 (22%)	15 (19%)	28 (41%)	24 (29%)
Current MDD	147 (49%)	31 (46%)	44 (54%)	32 (47%)	40 (48%)
Nicotine Metabolism Ratio	0.36 (0.23)	0.37 (0.27)	0.36 (0.21)	0.34 (0.18)	0.38 (0.25)
Unknown	21	2	9	7	3
Exclusive Mentholated Cigarette User	178 (60%)	43 (64%)	47 (58%)	40 (59%)	48 (59%)
Unknown	2	1	0	0	1
Readiness to quit smoking	6.78 (1.24)	6.95 (1.34)	6.71 (1.11)	6.80 (1.36)	6.68 (1.19)
Unknown	17	4	4	4	5

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

Selected categorical variables with observable differences on abstinence outcomes were plotted below ***(Figure 3)***. We can see a significant differences of end-of-treatment abstinence for using Varenicline. There is also slightly more abstinence among non-hispanic white, non-hispanic, non-black, higher income, higher education, not currently having an MDD, and not exclusive mentholated smokers. The most significant difference shown here is the use of varenicline, being non-hispanic white, higher income and education, which indicates that they could be possible predictors.

Figure 3: Abstinence vs. Categorical Bar Plots

Similarly, we plotted continuous variables by abstinence ***(Figure 4)***. People who had EOT abstinence have had a lower FTCD score and BDI score, higher cigarette reward value, slightly higher substitute reinforcer and NMR, and slightly lower complementary reinforcers. The biggest difference is in FTCD score, which indicates a possibility of strong predictor of EOT abstinence.

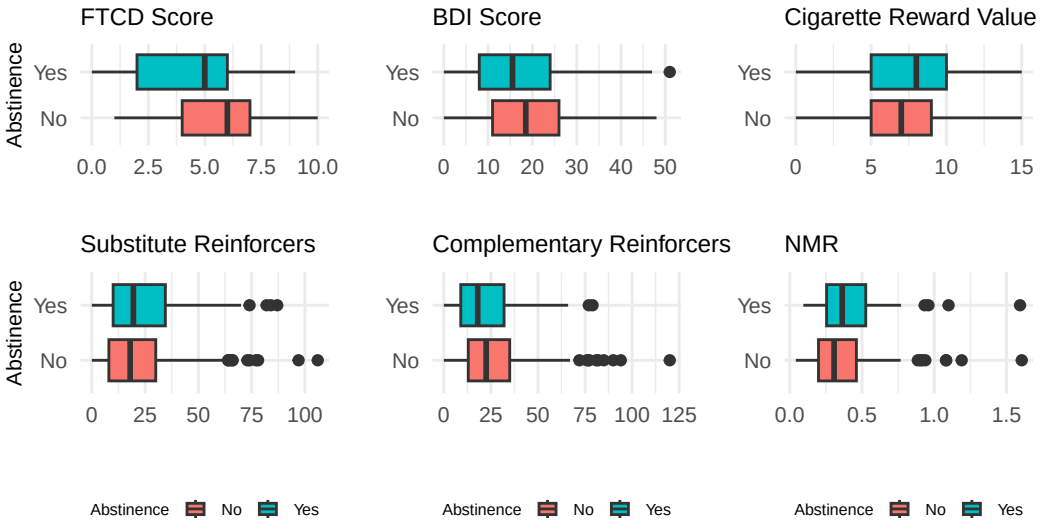

Figure 4: Abstinence vs. Continuous Box Plots

We further investigated some pairwise relationships ***(Figure 5)*** by the assumption that behavioral treatment and Varenicline has different impact to different demographics. Selected informative plots were included below. Higher Anhedonia scores seem to be associated with BA = 0 and Non-Abstinent (No) individuals, suggesting that Anhedonia may be a moderator for BA. Varenicline looks more effective for individuals with higher NMR, fewer cigarettes per day, less severe depression (lower BDI scores), and being Non-Hispanic White. Menthol Cigarette Use looks like a significant barrier to abstinence, even when BA is employed.

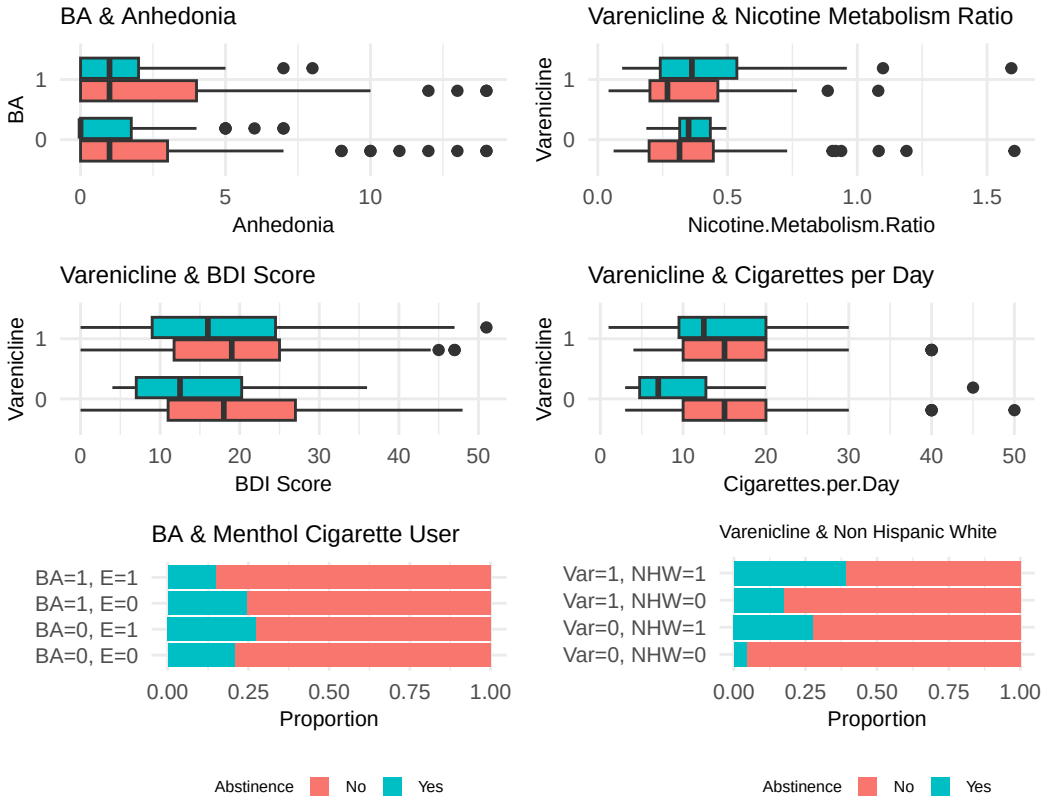

Figure 5: Pairwise Relationships for Relevant Variables

Regression Models and Model Evaluation

To evaluate the potential moderators and predictors, we will first use the *caret* package to split data into test and train sets using the function called `createDataPartition()`. With this function, it automatically stratifies sampling that ensures the proportion of EOT abstinence is maintained in both training and test sets. Because of the relatively small data set, a 70/30 train test split is used. We will apply Lasso regression with 5 fold cross-validation on each imputed dataset. The choice of Lasso regression will select the most predictive variables and shrink the others to 0. This effectively performs variable selection by excluding some less important

predictors out of the model, which could be helpful for easier calculation. In addition, Lasso regression regularize the high variance of a small data set, which is particularly suitable in our situation.

The lambda in each Lasso regression will be obtained from the cross-validated Lasso, providing the optimal value that minimizes prediction error. With a Lasso Model for each imputed data, we will then calculate the mean and standard deviation of each predictor's coefficient, creating the final Lasso estimates. These estimates will then be used to calculate a predicted probability of smoking abstinence for each participant.

We will repeat the process and compare models that has transformed variables and no transformed variables. The model evaluation for both models will consist of using Area Under the Curve (AUC) and calibration plots to assess model discrimination. A higher AUC indicates better discriminative ability of the model. The calibration plot visually compares the predictions and observations across different predicted values.

Results

Table 4: Linear Regression Models Summary

	Model Without Transformation			Model With Transformation		
Predictor	Pooled Mean	Pooled SD	Non Zero	Pooled Mean	Pooled SD	Non Zero
Anhedonia	-0.001	0.003	1(20%)	NA	NA	NA
BA:Exclusive.Menthol.Cigarette.User	-0.054	0.071	4(80%)	-0.141	0.072	5(100%)
FTCD.Score	-0.145	0.020	5(100%)	-0.176	0.018	5(100%)
Non.Hispanic.White	0.067	0.085	3(60%)	0.233	0.082	5(100%)
Varenicline:Age	0.011	0.004	5(100%)	0.019	0.002	5(100%)
Varenicline:Black	0.003	0.007	1(20%)	0.040	0.054	2(40%)
Varenicline:Education.C	0.030	0.044	3(60%)	0.073	0.057	4(80%)
Varenicline:Nicotine.Metabolism.Ratio	0.896	0.269	5(100%)	NA	NA	NA
BA:log.NMR	NA	NA	NA	0.004	0.009	1(20%)
Exclusive.Mentholated.Cigarette.User	NA	NA	NA	-0.010	0.017	2(40%)
Varenicline:BDI.Score	NA	NA	NA	0.000	0.000	2(40%)
Varenicline:Income.L	NA	NA	NA	-0.014	0.019	2(40%)
Varenicline:Smoking.5mins.of.waking.up	NA	NA	NA	0.001	0.003	1(20%)
log.Anhedonia	NA	NA	NA	-0.001	0.003	1(20%)
log.Nicotine.Metabolism.Ratio	NA	NA	NA	0.116	0.084	5(100%)

Model Evaluation

We fitted both models to the test data. The AUC for both train sets are higher than the test (See ***Figure 6***). The slight decline in predicted performance in the test data indicates that there are still some characteristics the model was not able to fully capture. For the model without transformations, the AUC for training data is 0.771, and that for the test data is slightly lower at 0.749. For the model with transformations, the AUC for training data is higher at 0.796, but the AUC for the test data is now at 0.695. These values are still an

acceptable discriminatory power of the model. Although the model with transformations has a higher AUC in train, it goes down in test. The model without transformations shows better consistency between train and test.

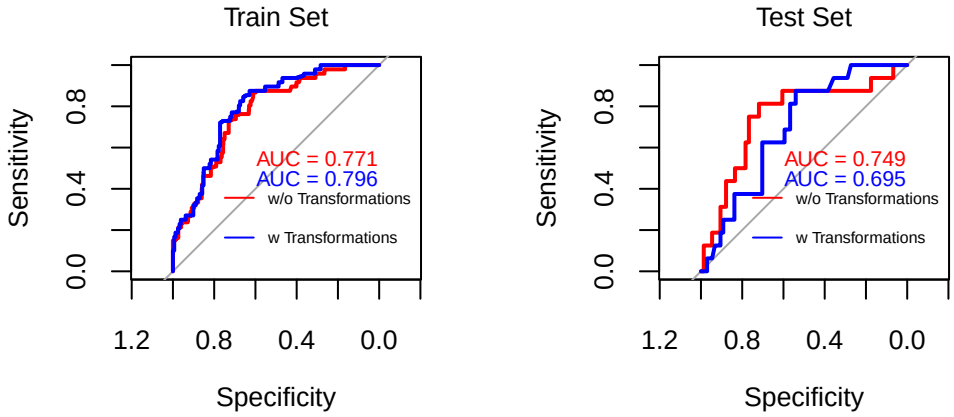

Figure 6: ROC/AUC Plots

A Calibration plot for each model is evaluated (***Figure 7***). The training data points generally follow the diagonal line for both models, with some deviation at small values, indicating a moderate calibration. For the test data, we observe more deviated points at around 0.25, with wider confidence interval. The lines are near the smaller values for the model without transformation, while the lines go to the bigger values for the model with transformation, and the deviation was significant. This indicates that the model without transformation might be a better fit.

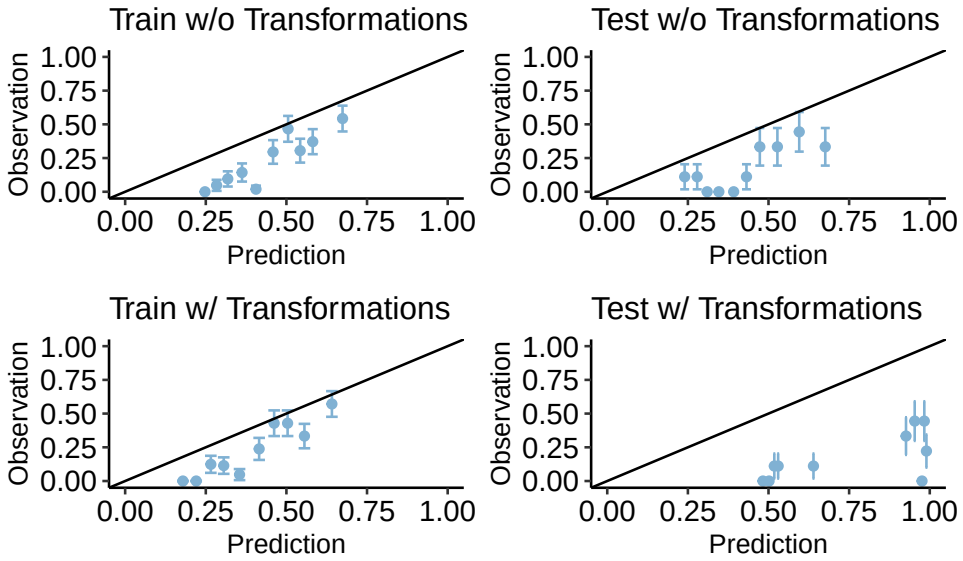

Figure 7: Calibration Plot Comparison - LASSO

With the above model evaluation results, we chose the Lasso model without transformations to be our final model. This final Lasso model highlights important baseline variables that influence end-of-treatment abstinence, both as main effects and as moderators of treatment response. The variables that remained to the model summary table above means that they have a certain level of predictive power in predicting smoking abstinence (See ***Table 4***).

The slight negative mean effect (-0.001) for **Anhedonia** means that individuals with Anhedonia have lower odds of abstinence, but this effect is small and only 1 out of 5 imputed models included this predictor.

The negative mean effect (-0.054) between the interaction of **BA** and **Exclusive Mentholated Cigarette User** shows that those that exclusively consume menthol cigarettes may reduce the effectiveness of BA. This also showed that being an exclusive mentholated cigarette user is a moderator that could impact the behavioral treatment.

The negative mean effect (-0.145) for **FTCD Score** means that individuals with higher nicotine independence have lower odds of abstinence. This is a strong predictor since all 5 imputed models included this predictor.

The positive mean effect (0.067) for **Non-Hispanic White** means that non-hispanic white individuals have higher odds of abstinence. This is a relatively strong predictor, being included from 3 out of 5 imputed models.

There are four positive mean effects between the interaction of **Varenicline** and **Age** (0.011), **Black** (0.003), **Education** (0.030), and **Nicotine Metabolism Ratio** (0.896). This suggests that those older, being black, being a college graduate, or having a higher NMR may increase the effectiveness of Varenicline. Being a college graduate seems counterintuitive here, but the

reason may be due to taking as prescribed and are more likely to adhere to treatment guidelines. This could be attributed to greater health literacy among college graduates, which enables them to understand the benefits of Varenicline and follow through with its recommended usage. The strongest interaction is between **Varenicline** and **Nicotine Metabolism Ratio**, with the highest pooled mean and included from all imputed models.

Discussion

This project investigated the baseline variables and their role in moderating and/or predicting the effects of behavioral and pharmacological treatments on end-of-treatment (EOT) smoking abstinence for individuals with major depressive disorder (MDD). Our results indicates the potential moderators and predictors in the use of lasso regression. The model evaluation showed reasonable discrimination for both train and test set. The model could potentially be used in clinical settings to identify individuals who might benefit most from specific cessation strategies.

While the findings provide some valuable insights, there are several limitations to this analyses. This relatively small sample limits the generalizability of our results, particularly in the Lasso model, we could only do a 70/30 train test split to ensure enough data for both sets. Data leakage may also occur if information from the test set accidentally leaks into the training process. This is because the train and test set were imputed together. Additionally, the outcome for EOT abstinence is rare, which may affect our results.

There could also be biases in the self-reported variables. Further research could be done to adjust for measurement error. Additionally, the use of Lasso regression shrinks some less predictive variables estimate to zero, so the result focus on selection rather than the precise effect sizes, which could be penalize some slightly informative variables that may improve the model predictability and interpretation. Another limitation was the low treatment adherence mentioned in prior research [3]. Future research should aim to further validate these findings in a larger population with a delivery of treatment that would ensure treatment adherence.

Conclusions

Overall, this project suggests that personalized treatment approaches based on baseline characteristics could potentially improve the smoking cessation outcome and some baseline characteristics could be used as predictors for smoking cessation. Particularly, exclusively menthol users may have lower odds of abstinence when receiving behavioral treatment, compared to the non-exclusively menthol users. FTCD Score, interaction between Varenicline and NMR, being Non-Hispanic White, and the interaction between Varenicline and Age are the strongest predictors that meaningfully influence smoking abstinence outcomes in individuals with MDD. These could be some important considerations when designing smoking cessation interventions or predicting smoking cessation outcomes based on participants' characteristics.

References

[1] American Lung Association. (n.d.). Tobacco Facts | State of tobacco control. Retrieved from https://www.lung.org/research/sotc/facts.

[2] Breslau, N., Kilbey, M. M., & Andreski, P. (1992). Nicotine withdrawal symptoms and psychiatric disorders: findings from an epidemiologic study of young adults. *The American Journal of Psychiatry*, 149(4), 464–469. https://doi.org/10.1176/ajp.149.4.464

[3] Hitsman, B., Papandonatos, G. D., Gollan, J. K., Huffman, M. D., Niaura, R., Mohr, D. C., Veluz-Wilkins, A. K., Lubitz, S. F., Hole, A., Leone, F. T., Khan, S. S., Fox, E. N., Bauer, M., Wileyto, E. P., Bastian, J., & Schnoll, R. A. (2023). Efficacy and safety of combination behavioral activation for smoking cessation and varenicline for treating tobacco dependence among individuals with current or past major depressive disorder: A 2x2 factorial, randomized, placebo-controlled trial. *Addiction*, 118(9), 1710–1725. https://doi.org/10.1111/add.16209

Code Appendix

```
# Load libraries
library(readr)
library(tidyverse)
set.seed(123456)
library(knitr)
library(tidyr)
library(dplyr)
library(kableExtra)
library(visdat)
library(gtsummary)
library(corrplot)
library(mice)
library(car)
library(glmnet) # For lasso
library(pROC) # For ROC
library(predtools) # For calibration plots
library(caret) # For stratified sampling
library(gridExtra) # For grid arrange
library(naniar) # For miss var summary

th <- theme(plot.title = element_text(size = 9),
            axis.title.x = element_text(size = 8),
            axis.title.y = element_text(size = 8),
            axis.text.x = element_text(size = 8),
            axis.text.y = element_text(size = 8),
            legend.position = "bottom",
            legend.justification = "center",
            legend.box = "horizontal",
            legend.title = element_text(size = 6),
            legend.text = element_text(size = 6),
            legend.key.size = unit(0.3, 'cm')
            )
# Import data sets
data <- read_csv("project2.csv")

###### DATA PREPROCESSING ######
#str(data)

## 1- Convert Variable Types
```

```r
data_transformed <- data

# Convert binary variables to factors
data_transformed$abst <- factor(data_transformed$abst, labels = c("No",
 ↪  "Yes"))
data_transformed$Var <- factor(data_transformed$Var)
data_transformed$BA <- factor(data_transformed$BA)
data_transformed$sex_ps <- factor(data_transformed$sex_ps, labels = c("Male",
 ↪  "Female"))
data_transformed$NHW <- factor(data_transformed$NHW, labels = c("No", "Yes"))
data_transformed$Black <- factor(data_transformed$Black, labels = c("No",
 ↪  "Yes"))
data_transformed$Hisp <- factor(data_transformed$Hisp, labels = c("No",
 ↪  "Yes"))
data_transformed$ftcd.5.mins <- factor(data_transformed$ftcd.5.mins, labels =
 ↪  c("No", "Yes"))
data_transformed$otherdiag <- factor(data_transformed$otherdiag, labels =
 ↪  c("No", "Yes"))
data_transformed$antidepmed <- factor(data_transformed$antidepmed, labels =
 ↪  c("No", "Yes"))
data_transformed$mde_curr <- factor(data_transformed$mde_curr, labels =
 ↪  c("No", "Yes"))
data_transformed$Only.Menthol <- factor(data_transformed$Only.Menthol, labels
 ↪  = c("No", "Yes"))

# Convert income and education to ordinal factors
data_transformed$inc <- ordered(data_transformed$inc, levels = 1:5,
                    labels = c("Less than $20,000", "$20,000-35,000",
                               "$35,001-50,000", "$50,001-75,000",
                               "More than $75,000"))

data_transformed$edu <- case_when(
  data_transformed$edu == 1 ~ 1,
  data_transformed$edu == 2 ~ 1,
  data_transformed$edu == 3 ~ 2,
  data_transformed$edu == 4 ~ 3,
  data_transformed$edu == 5 ~ 4,
  )

data_transformed$edu <- ordered(data_transformed$edu, levels = 1:4,
                labels = c("Grade school/Some high school",
                           "High school grad/GED",
```

```
                          "Some college/tech school",
                          "College graduate"))


data_transformed <- data_transformed %>%
  rename("Abstinence" = "abst",
         "Varenicline" = "Var",
         "Age" = "age_ps",
         "Sex" = "sex_ps",
         "Non.Hispanic.White" = "NHW",
         "Hispanic" = "Hisp",
         "Income" = "inc",
         "Education" = "edu",
         "FTCD.Score" = "ftcd_score",
         "Smoking.5mins.of.waking.up" = "ftcd.5.mins",
         "BDI.Score" = "bdi_score_w00",
         "Cigarettes.per.Day" = "cpd_ps",
         "Cigarette.Reward.Value" = "crv_total_pq1",
         "Substitute.Reinforcers" = "hedonsum_n_pq1",
         "Complementary.Reinforcers" = "hedonsum_y_pq1",
         "Anhedonia" = "shaps_score_pq1",
         "Other.Lifetime.DSM5.Diagnosis" = "otherdiag",
         "Taking.Antidepressant.Medication" = "antidepmed",
         "Current.MDD" = "mde_curr",
         "Nicotine.Metabolism.Ratio" = "NMR",
         "Exclusive.Mentholated.Cigarette.User" = "Only.Menthol",
         "Readiness.to.Quit.Smoking" = "readiness"
         )
## 2- Variable Transformations
numeric_columns <- data_transformed[, sapply(data_transformed, is.numeric)]
 ↪  %>% select(-id)

# Pivot data into long format
data_long <- numeric_columns %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to =
   ↪  "Value")

# Plot histograms with density overlay
ggplot(data_long, aes(x = Value)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", alpha =
   ↪  0.7) +
  geom_density(color = "blue", size = 1) +
```

```
  facet_wrap(~ Variable, scales = "free", ncol = 3) +
  labs(title = "",
       x = "Value", y = "Density") +
  theme_minimal() + th + theme(strip.text.x = element_text(size = 6.5))

# Transform Anhedonia with log(x+1) to handle zero values
data_transformed$log.Anhedonia <- log(data_transformed$Anhedonia + 1)

# Transform Cigarettes per day with sqrt
data_transformed$sqrt.Cigarettes.per.Day <-
 ↪  sqrt(data_transformed$Cigarettes.per.Day)

# Transform Substitute.Reinforcers, Complementary.Reinforcers with sqrt
data_transformed$sqrt.Substitute.Reinforcers <-
 ↪  sqrt(data_transformed$Substitute.Reinforcers)
data_transformed$sqrt.Complementary.Reinforcers <-
 ↪  sqrt(data_transformed$Complementary.Reinforcers)

# Transform NMR with log
data_transformed$log.Nicotine.Metabolism.Ratio <-
 ↪  log(data_transformed$Nicotine.Metabolism.Ratio)


transformed_vars <- c("log.Anhedonia",
 ↪  "sqrt.Cigarettes.per.Day","sqrt.Substitute.Reinforcers",
            "sqrt.Substitute.Reinforcers","sqrt.Complementary.Reinforcers",
             ↪  "log.Nicotine.Metabolism.Ratio")
original_vars <- c("Anhedonia", "Cigarettes.per.Day",
 ↪  'Substitute.Reinforcers',
                  "Substitute.Reinforcers", "Complementary.Reinforcers",
                   ↪  "Nicotine.Metabolism.Ratio")
data_numeric <- select_if(data_transformed,is.numeric) %>% na.omit() %>%
  select(-c(transformed_vars))

colnames(data_numeric) <- gsub("\\.", " ", colnames(data_numeric)) #replace
 ↪  period with space

M = cor(data_numeric[,-1])
corrplot(M, method = 'number', type = "lower", diag = FALSE,
         number.cex = 0.5, cl.cex = 0.5,
         tl.cex = 0.5, tl.col = "black", tl.srt = 20)
## 3- Missing data
```

```r
# data[,-(1:9)] %>% abbreviate_vars() %>% vis_miss()

miss_tbl <- data_transformed  %>% select(-c(transformed_vars)) %>%
 ↪  miss_var_summary()
colnames(miss_tbl) <- c("Variable", "Number of Missing", "Percent of
 ↪  Missing")

kable(miss_tbl[1:7,],booktabs = TRUE, digits = 2, caption = "Missing Data
 ↪  Table") %>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )
data.mice <- mice(data_transformed, m = 5, meth='pmm', seed=500)

# summary(data.mice)

###### DATA SUMMARY ######
# Create a new variable for the 4 groups
data$Group <- factor(paste(data$BA, data$Var, sep = "_"))
levels(data$Group) <- c("ST+Placebo", "ST+Varenicline", "BA+Placebo",
 ↪  "BA+Varenicline")

# Smoking Abstinence Rates Table
abst_tbl <- data %>%
  group_by(Group) %>%
  summarise(
    n = n(),
    "Smoking Abstinence" = sum(abst == 1)
  ) %>%
  mutate(
    "Percentage (%)" = round((`Smoking Abstinence`/n) *100,2)
  ) %>% as.data.frame()

rownames(abst_tbl) <- abst_tbl$Group

t(abst_tbl[-1]) %>% as.data.frame() %>%
  kable(booktabs=T, digits=2, caption = "EOT Abstinence Results")%>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )
```

```r
# Baseline Characteristics
data %>% dplyr::select(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, Group)) %>%
  mutate(
    inc = case_when(
      inc == 1 ~ "1-Less than $20,000",
      inc == 2 ~ "2-$20,000-35,000",
      inc == 3 ~ "3-$35,001-50,000",
      inc == 4 ~ "4-$50,001-75,000",
      inc == 5 ~ "5-More than $75,000"
    ),
    edu = case_when(
      edu == 1 ~ "1-Grade school",
      edu == 2 ~ "2-Some high school",
      edu == 3 ~ "3-High school grad/GED",
      edu == 4 ~ "4-Some college/tech school",
      edu == 5 ~ "5-College graduate"
    ),
    ftcd.5.mins = factor(ftcd.5.mins, labels = c("No", "Yes")),
    otherdiag = factor(otherdiag, labels = c("No", "Yes")),
    antidepmed = factor(antidepmed, labels = c("No", "Yes")),
    mde_curr = factor(mde_curr, labels = c("No", "Yes")),
    Only.Menthol = factor(Only.Menthol, labels = c("No", "Yes"))
  ) %>%
  rename("Age" = "age_ps",
         "Sex" = "sex_ps",
         "Non-Hispanic White" = "NHW",
         "Hispanic" = "Hisp",
         "Income" = "inc",
         "Education" = "edu",
         "FTCD Score" = "ftcd_score",
         "Smoking with 5 mins of waking up" = "ftcd.5.mins",
         "BDI Score" = "bdi_score_w00",
         "Cigarettes/day " = "cpd_ps",
         "Cigarette reward value" = "crv_total_pq1",
```

```
          "Substitute Reinforcers" = "hedonsum_n_pq1",
          "Complementary Reinforcers" = "hedonsum_y_pq1",
          "Anhedonia" = "shaps_score_pq1",
          "Other lifetime DSM-5 diagnosis" = "otherdiag",
          "Taking antidepressant medication" = "antidepmed",
          "Current MDD" = "mde_curr",
          "Nicotine Metabolism Ratio" = "NMR",
          "Exclusive Mentholated Cigarette User " = "Only.Menthol",
          "Readiness to quit smoking" = "readiness"
          ) %>%
  tbl_summary(by = Group,
              type = list(where(is.numeric) ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({sd})")) %>%
  add_overall() %>%
  # add_p %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Baseline Characteristics Summary") %>%
  kableExtra::kable_styling(font_size = 7.5,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )
# Abstinence vs. Categorical

library(ggplot2)
library(gridExtra)

a <- ggplot(data_transformed, aes(x = factor(Varenicline), fill =
 ↪  factor(Abstinence))) +
  geom_bar(position = "fill") +
  labs(title = "Varenicline", x = "", y = "", fill = "Abstinence") +
  coord_flip() + # Flip axes
  theme_minimal() + th +
  theme(legend.position = "none")

b <- ggplot(data_transformed, aes(x = factor(BA), fill = factor(Abstinence)))
 ↪  +
  geom_bar(position = "fill") +
  labs(title = "BA", x = "", y = "", fill = "Abstinence") +
  coord_flip() + # Flip axes
  theme_minimal() + th +
  theme(legend.position = "none")
```

```
c <- ggplot(data_transformed, aes(x = factor(Non.Hispanic.White), fill =
↪  factor(Abstinence))) +
  geom_bar(position = "fill") +
  labs(title = "Non Hispanic White", x = "", y = "", fill = "Abstinence") +
  coord_flip() + # Flip axes
  theme_minimal() + th +
  theme(legend.position = "none")

d <- ggplot(data_transformed, aes(x = factor(Hispanic), fill =
↪  factor(Abstinence))) +
  geom_bar(position = "fill") +
  labs(title = "Hispanic", x = "", y = "", fill = "Abstinence") +
  coord_flip() + # Flip axes
  theme_minimal() + th +
  theme(legend.position = "none")

e <- ggplot(data_transformed, aes(x = factor(Black), fill =
↪  factor(Abstinence))) +
  geom_bar(position = "fill") +
  labs(title = "Black", x = "", y = "", fill = "Abstinence") +
  coord_flip() + # Flip axes
  theme_minimal() + th +
  theme(legend.position = "none")

f <- ggplot(data_transformed[!is.na(data_transformed$Income),], aes(x =
↪  factor(Income), fill = factor(Abstinence))) +
  geom_bar(position = "fill") +
  labs(title = "Income", x = "", y = "", fill = "Abstinence") +
  scale_x_discrete(labels = c("1", "2", "3", "4", "5")) +
  coord_flip() + # Flip axes
  theme_minimal() + th +
  theme(legend.position = "none")

g <- ggplot(data_transformed, aes(x = factor(Education), fill =
↪  factor(Abstinence))) +
  geom_bar(position = "fill") +
  labs(title = "Education", x = "", y = "Proportion", fill = "Abstinence") +
  scale_x_discrete(labels = c("1", "2", "3", "4")) +
  coord_flip() + # Flip axes
  theme_minimal() + th

h <- ggplot(data_transformed, aes(x = factor(Current.MDD), fill =
↪  factor(Abstinence))) +
```

```
  geom_bar(position = "fill") +
  labs(title = "Current MDD", x = "", y = "Proportion", fill = "Abstinence")
   ↪  +
  coord_flip() + # Flip axes
  theme_minimal() + th

i <-
 ↪  ggplot(data_transformed[!is.na(data_transformed$Exclusive.Mentholated.Cigarette.User),],
           aes(x = factor(Exclusive.Mentholated.Cigarette.User), fill =
            ↪  factor(Abstinence))) +
  geom_bar(position = "fill") +
  labs(title = "Mentholated Cigarette User", x = "", y = "Proportion", fill =
   ↪  "Abstinence") +
  coord_flip() + # Flip axes
  theme_minimal() + th + theme(plot.title = element_text(size=7))

# Combine all plots into a grid
grid.arrange(a, b, c, d, e, f, g, h, i, ncol = 3, heights = c(0.9, 0.9, 1.3))
# Abstinence vs. Continuous

a <- ggplot(data_transformed, aes(x = factor(Abstinence), y = FTCD.Score,
 ↪  fill = factor(Abstinence))) +
    geom_boxplot() +
    labs(title = "FTCD Score", x = "Abstinence", y = "", fill = "Abstinence")
     ↪  +
    coord_flip() + # Flip axes
    theme_minimal() + th +
  theme(legend.position = "none")

b <- ggplot(data_transformed, aes(x = factor(Abstinence), y = BDI.Score, fill
 ↪  = factor(Abstinence))) +
    geom_boxplot() +
    labs(title = "BDI Score", x = "", y = "", fill = "Abstinence") +
    coord_flip() + # Flip axes
    theme_minimal() + th +
  theme(legend.position = "none")

c <- ggplot(data_transformed, aes(x = factor(Abstinence), y =
 ↪  Cigarette.Reward.Value, fill = factor(Abstinence))) +
    geom_boxplot() +
    labs(title = "Cigarette Reward Value", x = "", y = "", fill =
     ↪  "Abstinence") +
```

```
    coord_flip() + # Flip axes
    theme_minimal() + th +
  theme(legend.position = "none")

d <- ggplot(data_transformed, aes(x = factor(Abstinence), y =
 ↪  Substitute.Reinforcers, fill = factor(Abstinence))) +
    geom_boxplot() +
    labs(title = "Substitute Reinforcers", x = "Abstinence", y = "", fill =
     ↪  "Abstinence") +
    coord_flip() + # Flip axes
    theme_minimal() + th

e <- ggplot(data_transformed, aes(x = factor(Abstinence), y =
 ↪  Complementary.Reinforcers, fill = factor(Abstinence))) +
    geom_boxplot() +
    labs(title = "Complementary Reinforcers", x = "", y = "", fill =
     ↪  "Abstinence") +
    coord_flip() + # Flip axes
    theme_minimal() + th

f <- ggplot(data_transformed, aes(x = factor(Abstinence), y =
 ↪  Nicotine.Metabolism.Ratio, fill = factor(Abstinence))) +
    geom_boxplot() +
    labs(title = "NMR", x = "", y = "", fill = "Abstinence") +
    coord_flip() + # Flip axes
    theme_minimal() + th

grid.arrange(a, b, c, d, e, f, ncol = 3, bottom = legend, heights =
 ↪  c(0.9,1.3))
# Pairwise Relationships

a <- ggplot(data_transformed, aes(x = BA, y = Anhedonia, fill =
 ↪  factor(Abstinence))) +
    geom_boxplot()+
    labs(title = "BA & Anhedonia", x = "BA", y = "Anhedonia", fill =
     ↪  "Abstinence") +
    coord_flip() + # Flip axes
    theme_minimal() + th +
    theme(legend.position = "none")

b <- ggplot(data_transformed, aes(x = Varenicline, y =
 ↪  Nicotine.Metabolism.Ratio, fill = factor(Abstinence))) +
```

```
    geom_boxplot()+
    labs(title = "Varenicline & Nicotine Metabolism Ratio", x =
     ↪  "Varenicline", y = "Nicotine.Metabolism.Ratio", fill = "Abstinence")
     ↪  +
    coord_flip() + # Flip axes
    theme_minimal() + th +
    theme(legend.position = "none")

c <- ggplot(data_transformed, aes(x = Varenicline, y = BDI.Score, fill =
 ↪  factor(Abstinence))) +
    geom_boxplot()+
    labs(title = "Varenicline & BDI Score", x = "Varenicline", y = "BDI
     ↪  Score", fill = "Abstinence") +
    coord_flip() + # Flip axes
    theme_minimal() + th +
    theme(legend.position = "none")

d <- ggplot(data_transformed, aes(x = Varenicline, y = Cigarettes.per.Day,
 ↪  fill = factor(Abstinence))) +
    geom_boxplot()+
    labs(title = "Varenicline & Cigarettes per Day", x = "Varenicline", y =
     ↪  "Cigarettes.per.Day", fill = "Abstinence") +
    coord_flip() + # Flip axes
    theme_minimal() + th +
    theme(legend.position = "none")

e <-
 ↪  ggplot(data_transformed[!is.na(data_transformed$Exclusive.Mentholated.Cigarette.User),],
            aes(x = interaction(BA, Exclusive.Mentholated.Cigarette.User),
             ↪  fill = factor(Abstinence))) +
    geom_bar(position = "fill") +
    labs(title = "BA & Menthol Cigarette User",
         x = "",
         y = "Proportion",
         fill = "Abstinence") +
    theme_minimal() + th +
    coord_flip() + # Flip axes
    scale_x_discrete(labels = c("BA=0, E=0", "BA=0, E=1", "BA=1, E=0", "BA=1,
     ↪  E=1"))

f <- ggplot(data_transformed, aes(x = interaction(Varenicline,
 ↪  Non.Hispanic.White), fill = factor(Abstinence))) +
```

```
    geom_bar(position = "fill") +
    labs(title = "Varenicline & Non Hispanic White",
         x = "",
         y = "Proportion",
         fill = "Abstinence") +
    theme_minimal() + th + theme(plot.title = element_text(size=7)) +
    coord_flip() + # Flip axes
    scale_x_discrete(labels = c("Var=0, NHW=0", "Var=0, NHW=1", "Var=1,
     ↪  NHW=0", "Var=1, NHW=1"))

grid.arrange(a, b, c, d, e, f, ncol = 2, heights = c(0.9, 0.9, 1.1))

###### 2: Predictor Analysis - Lasso Regression Model ######
# Train test split
set.seed(1234)

# Sample indices in train data (70/30 train test split)
train_idx <- createDataPartition(data$abst, p = 0.7, list = FALSE)

### Without transformation
lasso_models <- list()
lasso_lambda <- numeric(5)

for (i in 1:5) {
  imputed_data <- complete(data.mice, i)[train_idx, ] %>%
 ↪  select(-c(transformed_vars))

  X <- model.matrix(Abstinence ~ . + . * Varenicline + . * BA + Varenicline *
 ↪  BA,
                    data = imputed_data[,-1])[, -1]
  y <- imputed_data$Abstinence

  # Cross-validated Lasso to find optimal lambda
  lasso_cv <- cv.glmnet(X, y,  nfolds = 5, alpha = 1, family = "binomial")
  lasso_lambda[i] <- lasso_cv$lambda.min

  # Fit Lasso model at optimal lambda (lambda.min)
  lasso_model <- glmnet(X, y, alpha = 1, family = "binomial", lambda =
 ↪  lasso_cv$lambda.min)
  lasso_models[[i]] <- lasso_model
}
generate_lasso_table <- function(list_of_models){
```

```r
  # Matrix of all resulted coefficients
  all_coefs <- lapply(list_of_models, function(model) {
    coefs <- coef(model)[-1,] %>% as.matrix()
    coefs
  })

  # Data frame combining the estimates
  coef_data <- data.frame(Predictor = character(),
                          Estimate = numeric())

  for (i in seq_along(all_coefs)) {

    coefs <- all_coefs[[i]]
    predictors <- rownames(coefs)
    estimates <- as.numeric(coefs)

    coef_data <- rbind(coef_data, data.frame(Predictor = predictors,
                                             Estimate = estimates))
  }

  # Pooled result
  pooled_coef_summary <- coef_data %>%
    group_by(Predictor) %>%
    summarise(Mean = mean(Estimate),
              SD = sd(Estimate),
              NonZeroCount = paste0(sum(Estimate != 0), "(", mean(Estimate !=
               ↪  0)*100, "%)")
              ) %>%
    as.data.frame()

  # Table for those not 0
  tbl <- pooled_coef_summary[pooled_coef_summary$Mean != 0,]
  rownames(tbl) <- seq(1:nrow(tbl))

  return(list(tbl, pooled_coef_summary))
}

tbl <- generate_lasso_table(lasso_models)[[1]]

# tbl %>%
#   kable(digits = 3,
#       caption = "Lasso Model Summary - Without Transformations",
```

```r
#       col.names = c("**Predictor**", "**Mean**", "**SD**",
 ↪  "**NonZeroCount**"))

# Composite complete data (Train set)
composite_complete_data <- bind_rows(lapply(1:5, function(i) {
  imputed_data <- complete(data.mice, i)[train_idx,] %>%
 ↪  select(-c(transformed_vars))
  imputed_data
}))


X_final <- model.matrix(Abstinence ~ . + . * Varenicline + . * BA +
 ↪  Varenicline * BA,
                  data = composite_complete_data[,-1])[, -1]

final_coef <- generate_lasso_table(lasso_models)[[2]]$Predictor

X_final <- X_final[,final_coef]


composite_complete_data$pred <- X_final %*%
 ↪  generate_lasso_table(lasso_models)[[2]]$Mean

# Convert to probabilities
composite_complete_data$prob <- 1 / (1 + exp(-composite_complete_data$pred))
### With transformation
lasso_models2 <- list()
lasso_lambda2 <- numeric(5)

for (i in 1:5) {
  imputed_data <- complete(data.mice, i)[train_idx, ] %>%
 ↪  select(-c(original_vars))

  X <- model.matrix(Abstinence ~ . + . * Varenicline + . * BA + Varenicline *
 ↪  BA,
                  data = imputed_data[,-1])[, -1]
  y <- imputed_data$Abstinence

  # Cross-validated Lasso to find optimal lambda
  lasso_cv <- cv.glmnet(X, y,  nfolds = 5, alpha = 1, family = "binomial")
  lasso_lambda2[i] <- lasso_cv$lambda.min
```

```r
  # Fit Lasso model at optimal lambda (lambda.min)
  lasso_model <- glmnet(X, y, alpha = 1, family = "binomial", lambda =
 ↪  lasso_cv$lambda.min)
  lasso_models2[[i]] <- lasso_model
}

tbl2 <- generate_lasso_table(lasso_models2)[[1]]

# tbl2 %>%
#   kable(digits = 3,
#       caption = "Lasso Model Summary - With Transformations",
#       col.names = c("**Predictor**", "**Mean**", "**SD**",
 ↪  "**NonZeroCount**"))


# Composite complete data (Train set)
composite_complete_data2 <- bind_rows(lapply(1:5, function(i) {
  imputed_data <- complete(data.mice, i)[train_idx,] %>%
 ↪  select(-c(original_vars))
  imputed_data
}))


X_final <- model.matrix(Abstinence ~ . + . * Varenicline + . * BA +
 ↪  Varenicline * BA,
                  data = composite_complete_data2[,-1])[, -1]

final_coef <- generate_lasso_table(lasso_models2)[[2]]$Predictor

X_final <- X_final[,final_coef]


composite_complete_data2$pred <- X_final %*%
 ↪  generate_lasso_table(lasso_models2)[[2]]$Mean

# Convert to probabilities
composite_complete_data2$prob <- 1 / (1 +
 ↪  exp(-composite_complete_data2$pred))

coefficients <- full_join(tbl, tbl2, by = "Predictor")

coefficients$Predictor <- c(
```

```
  "Anhedonia",
  "BA:Exclusive.Menthol.Cigarette.User",
  "FTCD.Score",
  "Non.Hispanic.White",
  "Varenicline:Age",
  "Varenicline:Black",
  "Varenicline:Education.C",
  "Varenicline:Nicotine.Metabolism.Ratio",
  "BA:log.NMR",
  "Exclusive.Mentholated.Cigarette.User",
  "Varenicline:BDI.Score",
  "Varenicline:Income.L",
  "Varenicline:Smoking.5mins.of.waking.up",
  "log.Anhedonia",
  "log.Nicotine.Metabolism.Ratio"
)

kable(coefficients, digits = 3,
      caption = "Linear Regression Models Summary",
      col.names = c("Predictor", "Pooled Mean", "Pooled SD", "Non Zero",
                    "Pooled Mean", "Pooled SD", "Non Zero")) %>%
  kableExtra::add_header_above(c(" " = 1, "Model Without Transformation" = 3,
 ↪  "Model With Transformation" = 3)) %>%
  kable_styling(font_size = 7, full_width = F, position = "center")

# Predict based on score model
# composite_complete_data$pred <- predict(score_model,
 ↪  composite_complete_data, type = "response")

# AUC/ROC
roc <- roc(composite_complete_data$Abstinence, composite_complete_data$prob)
auc <- auc(roc)


roc2 <- roc(composite_complete_data2$Abstinence,
 ↪  composite_complete_data2$prob)
auc2 <- auc(roc2)

# Evaluation: Calibration
cal_plot_data <- composite_complete_data %>%
  mutate(Abstinence = case_when(Abstinence == "No" ~ 0,
                           Abstinence == "Yes" ~ 1))
```

```r
cal_plot <- calibration_plot(data = cal_plot_data,
                             obs = "Abstinence", pred = "prob",
                             title = "Train w/o Transformations", y_lim =
                              ↪  c(0, 1), x_lim=c(0, 1))

cal_plot_data2 <- composite_complete_data2 %>%
  mutate(Abstinence = case_when(Abstinence == "No" ~ 0,
                          Abstinence == "Yes" ~ 1))

cal_plot2 <- calibration_plot(data = cal_plot_data2,
                             obs = "Abstinence", pred = "prob",
                             title = "Train w/ Transformations", y_lim =
                              ↪  c(0, 1), x_lim=c(0, 1))
# Validate with test data (Same process)
test_data <- bind_rows(lapply(1:5, function(i) {
  imputed_data <- complete(data.mice, i)
  imputed_data <- imputed_data[!(imputed_data$id %in% train_idx),] %>%
 ↪  select(-c(transformed_vars))
  imputed_data
}))


X_final <- model.matrix(Abstinence ~ . + . * Varenicline + . * BA +
 ↪  Varenicline * BA,
                  data = test_data[,-1])[, -1]

final_coef <- generate_lasso_table(lasso_models)[[2]]$Predictor

X_final <- X_final[,final_coef]


test_data$pred <- X_final %*% generate_lasso_table(lasso_models)[[2]]$Mean

# Convert to probabilities
test_data$prob <- 1 / (1 + exp(-test_data$pred))

roc_test <- roc(test_data$Abstinence, test_data$prob)
auc_test <- auc(roc_test)

# Evaluation: Calibration
cal_plot_data3 <- test_data %>%
```

```r
  mutate(Abstinence = case_when(Abstinence == "No" ~ 0,
                           Abstinence == "Yes" ~ 1))

cal_plot_test <- calibration_plot(data = cal_plot_data3,
                                  obs = "Abstinence", pred = "prob",
                                  title = "Test w/o Transformations", y_lim =
                                   ↪  c(0, 1), x_lim=c(0, 1))


# Validate with test data (With Transformations)
test_data2 <- bind_rows(lapply(1:5, function(i) {
  imputed_data <- complete(data.mice, i)
  imputed_data <- imputed_data[!(imputed_data$id %in% train_idx),] %>%
 ↪  select(-c(original_vars))
  imputed_data
}))


X_final <- model.matrix(Abstinence ~ . + . * Varenicline + . * BA +
 ↪  Varenicline * BA,
                  data = test_data2[,-1])[, -1]

final_coef <- generate_lasso_table(lasso_models2)[[2]]$Predictor

X_final <- X_final[,final_coef]


test_data2$pred <- X_final %*% generate_lasso_table(lasso_models)[[2]]$Mean

# Convert to probabilities
test_data2$prob <- 1 / (1 + exp(-test_data2$pred))

roc_test2 <- roc(test_data2$Abstinence, test_data2$prob)
auc_test2 <- auc(roc_test2)

# Evaluation: Calibration
cal_plot_data4 <- test_data2 %>%
  mutate(Abstinence = case_when(Abstinence == "No" ~ 0,
                           Abstinence == "Yes" ~ 1))

cal_plot_test2 <- calibration_plot(data = cal_plot_data4,
                                  obs = "Abstinence", pred = "prob",
```

```
                                  title = "Test w/ Transformations", y_lim =
                                   ↪  c(0, 1), x_lim=c(0, 1))
#|
# ROC/AUC Plots
par(mfrow = c(1, 2))

# Train Set
plot(roc, main = "Train Set", font.main = 1,
     cex.main = 0.8, cex.axis = 0.8, cex.lab = 0.8, col = "red")
lines(roc2, col = "blue")
legend("bottomright", legend = c("w/o Transformations", "w Transformations"),
       col = c("red", "blue"), lty = 1, cex = 0.5, bty= "n", inset= 0.01)
text(0.3, 0.55, paste("AUC =", round(auc, 3)), col = "red", cex = 0.7)
text(0.3, 0.45, paste("AUC =", round(auc2, 3)), col = "blue", cex = 0.7)

# Test Set
plot(roc_test, main = "Test Set", font.main = 1,
     cex.main = 0.8, cex.axis = 0.8, cex.lab = 0.8, col = "red")
lines(roc_test2, col = "blue")
legend("bottomright", legend = c("w/o Transformations", "w Transformations"),
       col = c("red", "blue"), lty = 1, cex = 0.5,  bty= "n", inset= 0.01)
text(0.3, 0.55, paste("AUC =", round(auc_test, 3)), col = "red", cex = 0.7)
text(0.3, 0.45, paste("AUC =", round(auc_test2, 3)), col = "blue", cex = 0.7)


#|
# Calibration Plots
grid.arrange(cal_plot$calibration_plot,
             cal_plot_test$calibration_plot,
             cal_plot2$calibration_plot,
             cal_plot_test2$calibration_plot,
             ncol = 2
            )
```