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, logistic regression models were fitted on each imputed dataset, and consistent predictors across imputations were identified. Lasso regression was also applied for robust variable selection across imputed datasets, focusing on main effects and treatment interactions. Key predictors were identified based on the Lasso variable selection.

Results: Moderator analysis found that menthol cigarette use negatively moderated the effectiveness of behavioral activation. Predictor analysis showed that the Nicotine Metabolism Ratio, interaction between Varenicline and Antidepressant Medication Use, FTCD Score, and its interaction between Behavioral Activation (BA) are strong 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 aimed to assess whether BASC would improve smoking cessation outcomes compared to ST, both with and without the addition of varenicline. However, the 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. To elaborate, the baseline characteristics includes demographic factors such as age, sex, race, and socioeconomic indicators like income and education. Additionally, clinical and behavioral factors are accounted for, 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 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 results sections for data summary). 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 the baseline depression score, daily cigarette consumption, and both substitute and complementary reinforcers, which are aspects of pleasurable activities. The Nicotine Metabolism Ratio 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, in turns improving the suitability of these variables for statistical modeling and the robustness of the analysis.

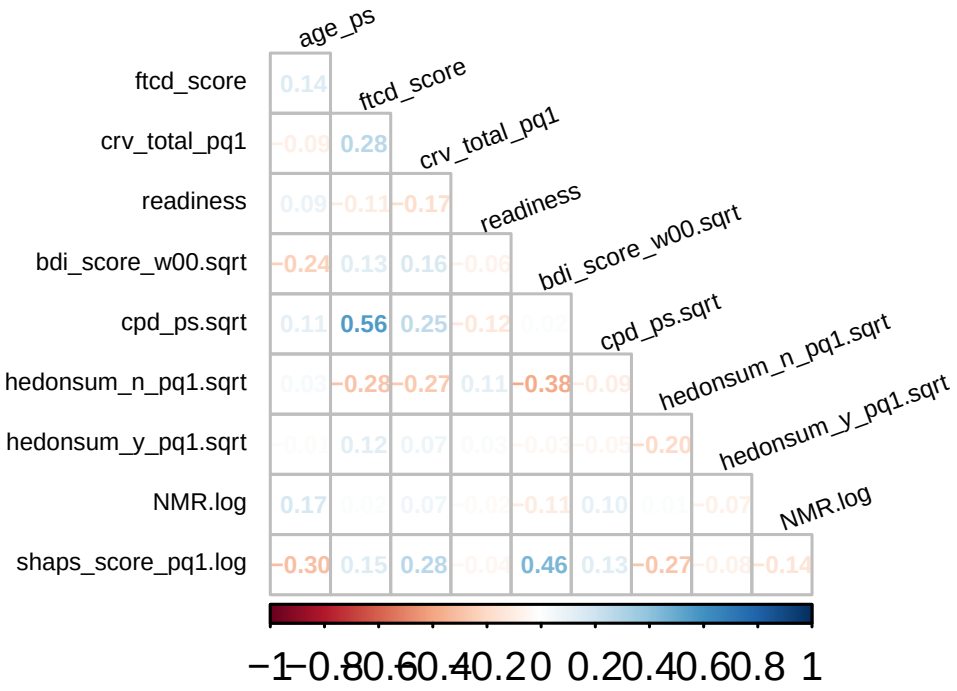

Figure 1: Correlation plot

The correlation plot above 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 (0.56) between the nicotine dependence (FTCD score) and cigarettes per day (CPD). 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 (0.46). 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.

Missing Data

Most baseline variables have 0% missing values, so no imputation is required for those. The column with highest missingness is the Nicotine Metabolism Ratio (7%). Then, we have 6% missing data in both baseline readiness to quit smoking and cigarette reward value. The columns for income, anhedonia, and only menthol preferences have 1% of missing data. 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.

Regression Models and Model Evaluation

To examine the potential moderators, we will fit a logistic model to each imputed dataset using `glm()` and then performed stepwise selection in both directions using `step()` via the *Stats* and *MASS* package in R respectively. With a goal of examining whether baseline variables moderate the effects of behavioral treatment on EOT abstinence, the main effects of all baseline variables and all their interactions with BA was considered in the initial model. After the model selection, a frequency table of the predictors selected across the models is created, where we extracted final predictors that appear in at least three of the models. This is choosing the most consistent predictors across the imputed datasets. With these consistent predictors, we fitted a final model to all the imputed datasets, pooled the results, and obtained the final coefficients and standard errors.

The model will be evaluated by using model diagnostics that looks at multicollinearity, residuals, and influential points. Since the model was selected using the step function, it is selected based on the minimum AIC. A lower AIC means a better AIC fit. Therefore, by applying the automatic model selection function, we obtained the best fitting model that has the smallest AIC value.

To evaluate the 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 60/40 train test split is used. We will apply Lasso regression with 10 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. 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 score for each participant, which we will later regress abstinence against the score in a logistic regression model and predict the probability of smoking abstinence for each participant.

The model evaluation will consists of using 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

Data Summary

Using the pre-transformed data, **Table 1** 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. **Tables 2** and **3** show 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 1: 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 2: 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%)
BL FTCD	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%)
BL BDI	19 (11)	18 (11)	20 (12)	19 (12)	18 (11)
BL Cigarettes/day	15 (8)	15 (7)	14 (7)	16 (9)	16 (9)
BL 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
BL Substitute Reinforcers	23 (20)	21 (20)	23 (19)	23 (20)	23 (19)
BL 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%)
BL 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
BL 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 (%)

Baseline variables as potential moderators of the effects of behavioral treatment on end-of-treatment (EOT) abstinence

After performing model selection with the step function, the final logistic regression model is as follows:

$$
\begin{aligned}
logit(E[\text{abstinence}]) &= ln(\frac{p}{1-p}) \\
&= \beta_0 + \beta_1 \cdot \text{Var} + \beta_2 \cdot \text{BA} + \beta_3 \cdot \text{Age} + \beta_4 \cdot \text{Edu2} + \beta_5 \cdot \text{Edu3} + \\
&+ \beta_6 \cdot \text{Edu4} + \beta_7 \cdot \text{OnlyMenthol} + \beta_8 \cdot \text{FTCD Score} \\
&+ \beta_9 \cdot \text{Smoke5mins} + \beta_{10} \cdot \text{NHW} + \beta_{11} \cdot \text{NMR.log} \\
&+ \beta_{12} \cdot (\text{BA} \times \text{edu2}) + \beta_{13} \cdot (\text{BA} \times \text{edu3}) + \beta_{14} \cdot (\text{BA} \times \text{edu4}) + \\
&+ \beta_{15} \cdot (\text{BA} \times \text{Only.Menthol})
\end{aligned}
$$

The following table provides a summary for each variable.

Table 3: Linear Regression Model Summary

Variable	**Estimate**	**Std. Error**	**Statistic**	**df**	**Significance**
(Intercept)	-1.733	0.997	-1.738	280.025	0.083
Var1	1.644	0.383	4.288	281.949	0.000
BA1	0.374	0.576	0.649	281.849	0.517
age_ps	0.021	0.014	1.507	281.787	0.133
edu.L	-0.435	0.686	-0.635	281.836	0.526
edu.Q	1.061	0.598	1.776	281.923	0.077
edu.C	0.713	0.472	1.510	281.698	0.132
Only.MentholYes	0.825	0.528	1.563	281.590	0.119
ftcd_score	-0.418	0.103	-4.051	281.988	0.000
ftcd.5.minsYes	0.796	0.463	1.719	281.882	0.087
NHWYes	1.057	0.400	2.640	280.872	0.009
NMR.log	0.535	0.301	1.777	264.329	0.077
BA1:edu.L	-0.035	0.937	-0.037	281.657	0.971
BA1:edu.Q	-1.108	0.834	-1.328	281.367	0.185
BA1:edu.C	-1.272	0.670	-1.897	281.815	0.059
BA1:Only.MentholYes	-1.377	0.697	-1.976	281.993	0.049

According to the selected model above, we observe a significant effect of Varenicline on EOT abstinence (p-value < 0.001). This suggests that the odds of participants receiving Varenicline to have EOT abstinence is $exp(1.644) = 5.18$ times the odds of those receiving placebo, adjusting for all other covariates in this model. However, the main effect for behavioral treatment is not significant, which means that the behavioral treatment alone were not associated with the effects of EOT abstinence. This finding aligns with the prior study. Among the predictors, the significance for FTCD score (p-value < 0.001) suggests that the odds of EOT abstinence for individuals with higher nicotine dependence at baseline are $exp(-0.418) = 0.66$ times than those with lower nicotine dependence. The significance for Non-Hispanic White (NHW) (p-value $= 0.009$) means that the odds of EOT abstinence for NHW individuals are $exp(1.057) = 2.88$ times than those with non-NHW individuals.

Regarding the potential moderators, only exclusively **menthol cigarette users** significantly moderated the effects of behavioral treatment (p-value = 0.049). Specifically, the estimate of -1.377 indicates that the menthol users may have lower odds of abstinence when receiving behavioral treatment, compared to the non-menthol users. **Education**, on the other hand, shows a marginal moderating effects on behavioral treatment (p-value = 0.059 for group 3, some college/technical school). However, this is not significant in other groups. This suggests that individuals with behavioral treatment that had some college/technical school education may have lower odds of abstinence compared to those who had grade school/some high school education. Note that this conclusion is not statistically significant.

Overall, this concludes that Varenicline is an effective treatment for EOT abstinence, while behavioral treatment effectiveness depends. Exclusively menthol cigarette use significantly moderates the effect of behavioral treatment, and the education of some college or technical school may marginally influence that as well. These findings suggests that it could be helpful in designing a targeted approach to certain groups such as exclusively menthol smokers, to improve the EOT abstinence outcome.

Model Evaluation

In validating the model, we further performed some model diagnostics. Multicollinearity is assessed by calculating VIF of the main effect part of the model, which we picked the first complete imputed data and obtained the calculations below. The low VIF values indicates that there is no multicollinearity in this model.

Table 4: Model diagnostics part 1

	GVIF	Df	GVIF^(1/(2*Df))
Var	1.064965	1	1.031972
BA	1.038550	1	1.019093
NMR.log	1.116099	1	1.056456
NHW	1.264532	1	1.124514
edu	1.309381	3	1.045950
ftcd_score	1.982987	1	1.408186
ftcd.5.mins	1.934739	1	1.390949
mde_curr	1.186739	1	1.089375
hedonsum_n_pq1.sqrt	1.554527	1	1.246807
shaps_score_pq1.log	1.206782	1	1.098536

More diagnostic plots shown below provides insights into the fit and potential issues for the model. The QQ plot on the left shows that the residuals mostly follow the line, but with some slight deviations at the upper tail. This suggests some moderate deviations from normality, but this will not be a major concern for our model. The Cook's distance plot in the middle

shows that points 186, 226, and 283 have the highest Cook's distances, which means that they could be influential points for the model. However, their distance is still relatively small and below the threshold 1, so they are not too concerning. The Residuals vs. Leverage plot on the right again shows these points as potential outliers. Given these diagnostics, the is a relatively robust model.

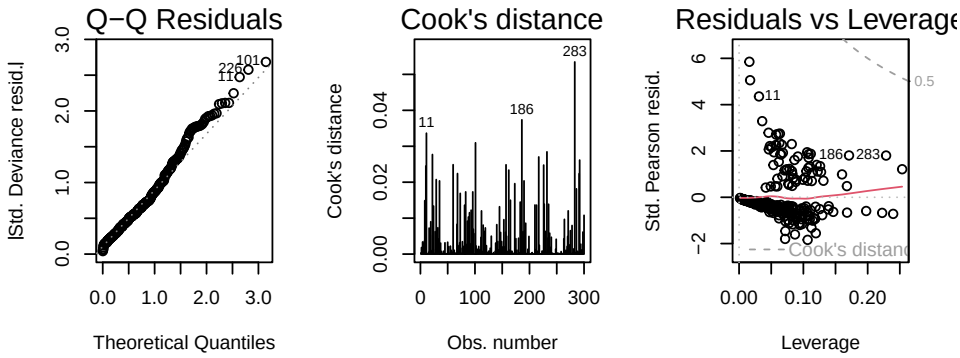

Figure 2: Model diagnostics part 2

Baseline variables as predictors of abstinence, controlling for behavioral treatment and pharmacotherapy

For the next part, we will evaluate which baseline variables could be predictors for EOT abstinence, controlling for their behavioral treatment and pharmacotherapy. The final Lasso Regression Model coefficients is displayed in the table below.

Table 5: Lasso Model Summary

Predictor	**Mean**	**SD**
BA1:ftcd_score	-0.055	0.007
BA1:inc.Q	-0.003	0.007
BA1:mde_currYes	-0.005	0.012
NHWYes	0.042	0.068
NMR.log	0.247	0.092
Var1:NHWYes	0.026	0.023
Var1:age_ps	0.013	0.001
Var1:antidepmedYes	0.235	0.052
ftcd_score	-0.176	0.009
otherdiagYes	-0.002	0.005

The 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.

The slight negative mean effect (-0.055) between the interaction of **BA** and **FTCD score** shows that higher nicotine dependence (FTCD score) may reduce the effectiveness of BA. The positive mean effect (0.042) for **NMR** (nicotine metabolism ratio) indicates that individuals with a higher metabolism of nicotine have better odds of abstinence. Another significant positive effect is seen between the interaction of **Varenicline** and **antidepressant medication use** (0.235), suggests that those taking antidepressant may increase the effectiveness of varenicline. **FTCD score** alone also has a notable negative effect (-0.176), which means that higher nicotine dependence is associated with loewr odds of abstinence. Other effects are smaller but they all play parts in predicting smoking cessation for participants with MDD.

With these estimates, we will calculate a scores for each participant and create a score model by fitting a logistic regression model with only the composite score as a predictor. Once the model is fitted, we can use this to generate predicted probabilities for abstinence based on the composite score. Table below shows the score model summary.

Table 6: Score Model Summary

Variable	**Estimate**	**Std. Error**	**z value**	**Significance**
(Intercept)	0.425	0.141	3.013	0.003
score	2.156	0.182	11.862	0.000

From the summary above, the score estimate 1.232 (p-value < 0.001) indicates that the composite score is a strong predictor of smoking abstinence. For every one-unit increase in the score, the log-odds of abstinence increase by 1.232. In other words, the odds of abstinence increase by $exp(1.232) = 3.43$. At score $= 0$, the odds of abstinence is $exp(-0.469) = 0.63$.

Model Evaluation

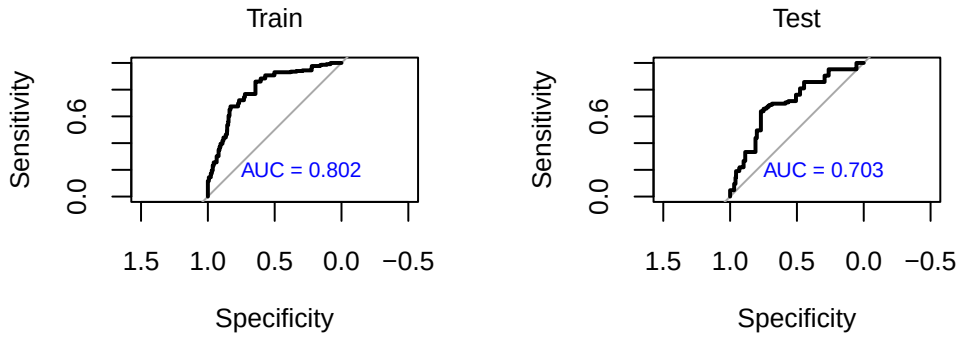

Figure 3: ROC/AUC Plots

We used the same scoring and fitted the same score model to the test data. The AUC for the training data is 0.802, and the AUC for the test data is slightly lower at 0.703, but it is still acceptable. These value shows the generally good discriminatory power of the model. The slight decline in predicted performance in the test data indicates that there may still be some characteristics the model was not able to fully capture.

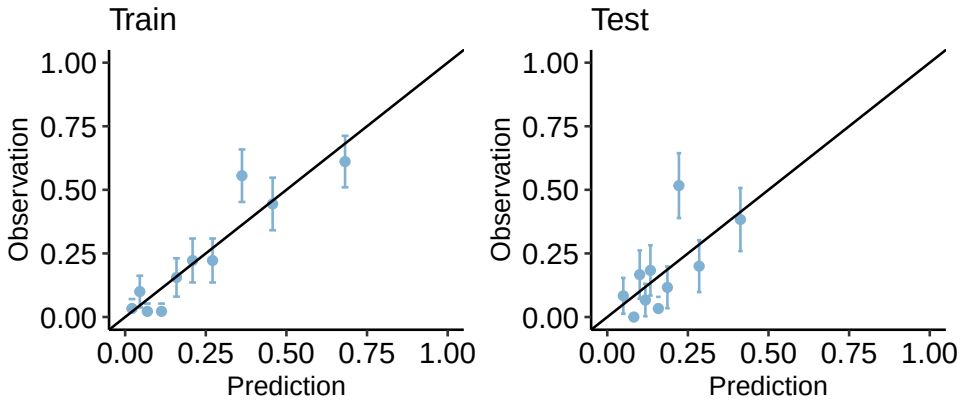

Figure 4: Calibration Plot Comparison - LASSO

The left panel shows the calibartion plot for the training data, our points generally follow the diagonal line, indicating good calibration. There are some deviation at around x = 0.3, and we do not have high values. In the test data case on the right panel, we observe more deviated points at around 0.25, with wider confidence interval. The lines also gather around the smaller values. This indicates that our model might still need further refinement to better capture the characteristics of the popultation.

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 logistic and lasso regression. The model diagnostics and evaluation of both of them showed reasonable discrimination, for both train and test set in the lasso. The composite scoring 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. First, this relatively small sample limits the generalizability of our results, particularly in the Lasso model, we could only do a 60/40 train test split to ensure enough data for both sets. There could also be biases in the self-reported variables, which 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 throwing away some less important but slightly informative variables that may improve the model predictability. Another limitation was the low treatment adherence, mentioned in the 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, our 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, menthol cigarette use could moderate the effects of behavioral treatment - the menthol users may have lower odds of abstinence when receiving behavioral treatment, compared to the non-menthol users. Nicotine Metabolism Ratio, interaction between Varenicline and Antidepressant Medication Use, FTCD Score, and its interaction between Behavioral Activation (BA) are strong 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(DataExplorer)
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
# Import data sets
data <- read_csv("project2.csv")

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

## 1- Convert Variable Types
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"))
```

```r
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"))

## 2- Variable Transformations

# Histograms of all numeric data
# plot_histogram(select(data, -id), nrow = 5L)

# Transform bdi_score_w00 with sqrt
data_transformed$bdi_score_w00.sqrt <- sqrt(data_transformed$bdi_score_w00)

# Transform cpd_ps with sqrt
data_transformed$cpd_ps.sqrt <- sqrt(data_transformed$cpd_ps)
```

```r
# Transform hedonsum_n_pq1, hedonsum_y_pq1 with sqrt
data_transformed$hedonsum_n_pq1.sqrt <- sqrt(data_transformed$hedonsum_n_pq1)
data_transformed$hedonsum_y_pq1.sqrt <- sqrt(data_transformed$hedonsum_y_pq1)

# Transform NMR with log
data_transformed$NMR.log <- log(data_transformed$NMR)

# Transform shaps_score_pq1 with log(x+1) to handle zero values
data_transformed$shaps_score_pq1.log <- log(data_transformed$shaps_score_pq1
 ↪  + 1)


data_transformed <- data_transformed %>% select(-c(bdi_score_w00, cpd_ps,
 ↪  hedonsum_n_pq1,
                                        hedonsum_y_pq1, NMR, shaps_score_pq1))

data_numeric <- select_if(data_transformed,is.numeric) %>% na.omit()

M = cor(data_numeric[,-1])
corrplot(M, method = 'number', type = "lower", diag = FALSE,
         number.cex = 0.5,
         tl.cex = 0.5, tl.col = "black", tl.srt = 20)
## 3- Missing data
# data[,-(1:9)] %>% abbreviate_vars() %>% vis_miss()

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",)

# Baseline Demographics
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",
          "BL FTCD" = "ftcd_score",
          "Smoking with 5 mins of waking up" = "ftcd.5.mins",
          "BL BDI" = "bdi_score_w00",
          "BL Cigarettes/day " = "cpd_ps",
          "BL Cigarette reward value" = "crv_total_pq1",
          "BL Substitute Reinforcers" = "hedonsum_n_pq1",
          "BL Complementary Reinforcers" = "hedonsum_y_pq1",
          "Anhedonia" = "shaps_score_pq1",
          "Other lifetime DSM-5 diagnosis" = "otherdiag",
          "BL Taking antidepressant medication" = "antidepmed",
          "Current MDD" = "mde_curr",
          "Nicotine Metabolism Ratio" = "NMR",
          "Exclusive Mentholated Cigarette User " = "Only.Menthol",
          "BL 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")
                            )

###### 1: Moderator Analysis - Logistic Regression Model ######

selected_models <- list()

# Select model with step() for each imputed data
for (i in 1:5) {
  imputed_data <- complete(data.mice, i)
  fit <- glm(abst ~ Var * BA +
                    BA * age_ps + # Age impact on BASC
                    BA * sex_ps + # Sex impact on BASC
                    BA * NHW + BA * Black + BA * Hisp + # Race impact on BASC
                    BA * inc + BA * edu + # Income & Education impact on BASC
```

```
                    BA * ftcd_score + # FTCD Score impact on BASC
                    BA * ftcd.5.mins + # Smoking within 5 mins of waking up
 ↪  impact on BASC
                    BA * bdi_score_w00.sqrt + # BDI Score impact on BASC
                    BA * cpd_ps.sqrt + # Cigarettes per day impact on BASC
                    BA * crv_total_pq1 + # Cigarette reward value impact on
 ↪  BASC
                    BA * hedonsum_n_pq1.sqrt + BA * hedonsum_y_pq1.sqrt + #
 ↪  Pleasurable Events Scale impact on BASC
                    BA * shaps_score_pq1.log + # Anhedonia impact on BASC
                    BA * otherdiag + # Other lifetime DSM-5 diagnosis impact
 ↪  on BASC
                    BA * antidepmed + # Antidepressant medication impact on
 ↪  BASC
                    BA * mde_curr + # Current MDD impact on BASC
                    BA * NMR.log + # Nicotine Metabolism Ratio impact on BASC
                    BA * Only.Menthol + # Exclusive Menthol user impact on
 ↪  BASC
                    BA * readiness, # Readiness to quit impact on BASC
                    family = binomial, data = imputed_data)

  selected_models[[i]] <- step(fit, direction = "both", trace = 0) # Select
 ↪  using AIC
}

selected_predictors <- lapply(selected_models, function(model)
 ↪  rownames(summary(model)$coefficients)[-1])

freq_table <- table(unlist(selected_predictors))

# Select those appeared at least 3 models
# names(freq_table[freq_table>=3])

# Final logistic model
fit_fin <- with(data.mice,
                glm(abst ~ Var + BA + age_ps  + edu + Only.Menthol +
                      ftcd_score + ftcd.5.mins + NHW + NMR.log +
                      # Interaction terms
                      BA * edu + BA * Only.Menthol,
                    family = binomial))

pooled_results_final <- pool(fit_fin)
```

```r
# Table
summary <- summary(pooled_results_final)
coefficients <- as.data.frame(summary)
kable(coefficients, digits = 3,
      caption = "Linear Regression Model Summary",
      col.names = c("**Variable**", "**Estimate**", "**Std. Error**",
       ↪  "**Statistic**", "**df**", "**Significance**"))

# Use first imputed data set, and model without interaction
completed_data <- complete(data.mice, 1)

model <- glm(abst ~ Var + BA + NMR.log + NHW + edu +
                      ftcd_score + ftcd.5.mins + mde_curr +
                      hedonsum_n_pq1.sqrt + shaps_score_pq1.log,# +
                      # Interaction terms
                      #BA * edu + BA * hedonsum_n_pq1.sqrt +
                      #BA * shaps_score_pq1.log,
                    family = binomial, data = completed_data)

vif(model, type = "terms", digits=2) %>%
 kable(col.names = c("GVIF","Df","GVIF^(1/(2*Df))")
   )

model <- glm(abst ~ Var + BA + NMR.log + NHW + edu +
                      ftcd_score + ftcd.5.mins + mde_curr +
                      hedonsum_n_pq1.sqrt + shaps_score_pq1.log +
                      # Interaction terms
                      BA * edu + BA * hedonsum_n_pq1.sqrt +
                      BA * shaps_score_pq1.log,
                    family = binomial, data = completed_data)
par(mfrow = c(1,3))
plot(model,2) # QQ Plot
plot(model,4) # Cook's
plot(model,5) # Residuals vs Leverage

###### 2: Predictor Analysis - Lasso Regression Model ######
# Train test split
set.seed(123)

# Sample indices in train data (60/40 train test split)
train_idx <- createDataPartition(data$abst, p = 0.6, list = FALSE)
```

```r
lasso_models <- list()
lasso_lambda <- numeric(5)

for (i in 1:5) {
  imputed_data <- complete(data.mice, i)[train_idx, ]

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

  # Cross-validated Lasso to find optimal lambda
  lasso_cv <- cv.glmnet(X, y, 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
}

# Matrix of all resulted coefficients
all_coefs <- lapply(lasso_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
```

```r
pooled_coef_summary <- coef_data %>%
  group_by(Predictor) %>%
  summarise(Mean = mean(Estimate),
            SD = sd(Estimate)) %>%
  as.data.frame()

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

tbl %>%
  kable(digits = 3,
      caption = "Lasso Model Summary",
      col.names = c("**Predictor**", "**Mean**", "**SD**"))

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

X_final <- model.matrix(abst ~ . + . * Var + . * BA + Var * BA,
                   data = composite_complete_data[,-1])[,-1]

X_final <- X_final[,pooled_coef_summary$Predictor]

y_final <- composite_complete_data$abst

# Scores for each ppt
composite_complete_data$score <- X_final %*% pooled_coef_summary$Mean
# Score model
score_model <- glm(abst ~ score, data = composite_complete_data, family =
 ↪  "binomial")

# Score model summary
summary <- summary(score_model)
coefficients <- as.data.frame(summary$coefficients)
kable(coefficients, digits = 3,
      caption = "Score Model Summary",
      col.names = c("**Variable**", "**Estimate**", "**Std. Error**", "**z
       ↪  value**", "**Significance**"))
```

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

# AUC/ROC
roc <- roc(composite_complete_data$abst, composite_complete_data$pred)
auc <- auc(roc)

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

cal_plot <- calibration_plot(data = cal_plot_data,
                             obs = "abst", pred = "pred",
                             title = "Train", 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),]
  imputed_data
}))


X_final <- model.matrix(abst ~ . + . * Var + . * BA + Var * BA,
                  data = test_data[,-1])[,-1]

X_final <- X_final[,pooled_coef_summary$Predictor]

y_final <- test_data$abst

test_data$score <- X_final %*% pooled_coef_summary$Mean
score_model <- glm(abst ~ score, data = test_data, family = "binomial")

test_data$pred <- predict(score_model, test_data, type = "response")

roc_test <- roc(test_data$abst, test_data$pred)
auc_test <- auc(roc_test)

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

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

cal_plot_test <- calibration_plot(data = cal_plot_data,
                                  obs = "abst", pred = "pred",
                                  title = "Test", y_lim = c(0, 1), x_lim=c(0,
                                   ↪  1))
# ROC/AUC Plots
par(mfrow= c(1,2))
plot(roc, main = "Train", font.main = 1,
     cex.main = 0.8, cex.axis = 0.8, cex.lab = 0.8)
text(0.3, 0.2, paste("AUC =", round(auc, 3)), col = "blue", cex = 0.7)

plot(roc_test, main = "Test", font.main = 1,
     cex.main = 0.8, cex.axis = 0.8, cex.lab = 0.8)
text(0.3, 0.2, paste("AUC =", round(auc_test, 3)), col = "blue", cex = 0.7)

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