Examining How Baseline Characteristics Influence Smoking Cessation in Adults with Major Depressive Disorder

Dingxuan Zhang

November 2024

Abstract

In this study we are going to explore how different baseline characteristics might influence smoking cessation success for adults with major depressive disorder (MDD). People with MDD usually face extra challenges when quitting as they tend to smoke more heavily, have a harder time with nicotine cravings, and go through tougher withdrawal symptoms. With data from a randomized, placebo-controlled, 2x2 factorial trial, this analysis will look at whether baseline characteristics affect how well behavioral treatments work for smoking cessation. We also check whether these baseline traits can predict who will be more likely to quit smoking, regardless of the type of treatment received. The trial included 300 adult smokers with current or past MDD, split between Behavioral Activation for Smoking Cessation (BASC) or Standard Treatment (ST), and varenicline or placebo. Findings from this analysis could point to which individual traits improve or reduce treatment success, providing clues to make cessation strategies more effective for people with MDD.

Introduction

In this project we aim to analyze data from a randomized trial to examine two main objectives: (1)identify baseline characteristics that predict smoking abstinence outcomes, controlling for the effects of behavioral treatment and pharmacotherapy (2) investigate the potential moderating effect of these baseline characteristics on the relationship between behavioral treatment and end-of-treatment (EOT) smoking abstinence.

The analysis will be structured as follows: 1. **Data Collection and Pre-processing**: We will detail the data sources, define the variables used in our analysis, and apply data cleaning steps. An exploratory data analysis (EDA) will be conducted to examine the distribution of key variables and prepare summary tables. 2. **Predictor Analysis**: We will first assess the predictive power of baseline characteristics using correlation analysis and LASSO regression to select key predictors of smoking abstinence. 3. **Moderation Analysis**: Using the predictors identified, we will construct a logistic regression model with interaction terms to test whether these variables moderate the effects of behavioral treatment on EOT smoking abstinence. 4. **Discussion**: We will summarize our findings, discuss the practical implications for personalized treatment strategies, and note any limitations in our methods. 5. **Conclusion**: We shall conclude our findings.

By analyzing both the predictive and moderating roles of baseline characteristics, we aim to enhance our understanding of factors that may influence the effectiveness of smoking cessation treatments for individuals with MDD. Notice here we switch the sequence of predictor and moderation analysis as there are too many baseline variables that will result in extremely large amount of interaction terms so we do the variable selection first and only consider the few important factors in interaction terms.

Data Collection

As we mentioned, the dataset for this project comes from a clinical trial involving 300 adult smokers with current or past MDD. Participants were randomly assigned to one of the two behavioral treatments BASC or ST and received either varenicline or a placebo. We shall firstly examine the dataset.

```
##   abst Var BA age_ps sex_ps NHW Black Hisp inc edu ftcd_score ftcd.5.mins
## 1    0   0  0     33      2   1     0    0   4   5          6           0
## 2    0   1  1     61      1   0     1    0   2   4          4           0
## 3    0   1  0     43      2   0     1    0   2   3          3           0
## 4    1   1  1     56      2   1     0    0   1   4          4           0
## 5    0   1  1     38      2   1     0    0   4   5          4           0
## 6    0   1  1     35      1   1     0    0   5   5          3           0
##   bdi_score_w00 cpd_ps crv_total_pq1 hedonsum_n_pq1 hedonsum_y_pq1
## 1            36     10            12             22             29
## 2            25     10             2             22             16
## 3            28     10            NA              9              4
## 4            28      8             1             38              2
## 5            22     17            15             25             13
## 6            15     10             8             25              9
##   shaps_score_pq1 otherdiag antidepmed mde_curr       NMR Only.Menthol
## 1               6         1          1        1 1.1896800           NA
## 2               0         0          0        1 0.7256241           NA
## 3               3         1          0        1 0.1520089            1
## 4               2         0          1        1 1.5936004            0
## 5               0         0          0        1 0.3379682            0
## 6               3         0          0        1 0.4909236            0
##   readiness
## 1        NA
## 2        NA
## 3        NA
## 4        NA
## 5        NA
## 6        NA
```

We can see that that there are 24 variables with some missing values and we can categorize the data as we are looking at the effect of baseline variables.Below is a list of all variables in the dataset, categorized to clarify which ones are baseline variables:

Outcome Variable

- **Smoking Abstinence (abst)**: This variable indicates whether participants achieved smoking abstinence by the end of the treatment.

Treatment Variables

- **Pharmacotherapy (Var)**: Specifies whether participants were given varenicline or a placebo.
- **Psychotherapy (BA)**: Indicates whether participants received BASC or ST as their behavioral treatment.

Baseline Variables

- **Demographic Variables**: Age at phone interview (age_ps), sex at phone interview (sex_ps), race/ethnicity indicators (Non-Hispanic White - NHW, Black, Hispanic), income level (inc), and education level (edu).
- **Smoking-Related Variables**: FTCD score at baseline (ftcd_score), daily cigarette count (cpd_ps), smoking within 5 minutes of waking (ftcd.5.mins), and cigarette reward value (crv_total_pq1).
- **Depression and Psychological Variables**: Baseline BDI score (bdi_score_w00), substitute reinforcers (hedonsum_n_pq1), complementary reinforcers (hedonsum_y_pq1), and anhedonia score (shaps_score_pq1).
- **Other Health and Psychological Variables**: Other DSM-5 diagnosis (otherdiag), antidepressant medication at baseline (antidepmed), current vs. past MDD status (mde_curr), nicotine metabolism ratio (NMR), exclusive menthol use (Only.Menthol), and readiness to quit smoking (readiness).

Data pre-processing

Since we can see that these baseline variables can be categorized further and we are therefore interested in the correlation between these variables which will also help us identifle the important variables where we shall conduct later. Now we firstly deal with the missing values by checking the missing values.

```
## # A tibble: 7 x 3
##   Variable        Missing_Count Missing_Percentage
##   <chr>                   <int>              <dbl>
## 1 inc                         3              1
## 2 ftcd_score                  1              0.333
## 3 crv_total_pq1              18              6
## 4 shaps_score_pq1             3              1
## 5 NMR                        21              7
## 6 Only.Menthol                2              0.667
## 7 readiness                  17              5.67
```

We can see that the amount of missing data is not substantive. Therefore, we conduct imputation using MICE package to ensure a complete dataset for analysis, reducing the risk of bias while maintaining the integrity of our data.

After dealing missing data, we then convert categorical variables into suitable formats and standardize numerical variables. For categorical variables like NHW, Black, and Hisp, we convert each into dummy variables (0 or 1), allowing these binary categories to be used effectively in our analyses. Additionally, we convert variables such as inc and edu into factors to reflect their ordinal nature. For numerical variables, including ftcd_score, bdi_score_w00, and cpd_ps, we apply standardization to place them on a comparable scale which ensures our dataset is ready for the next phases of analysis.

After doing these data preprocesssing, we can now do the data analysis. We can split the orginal data into 4 groups based on treatment to see the difference in eacch group by drawing a table.

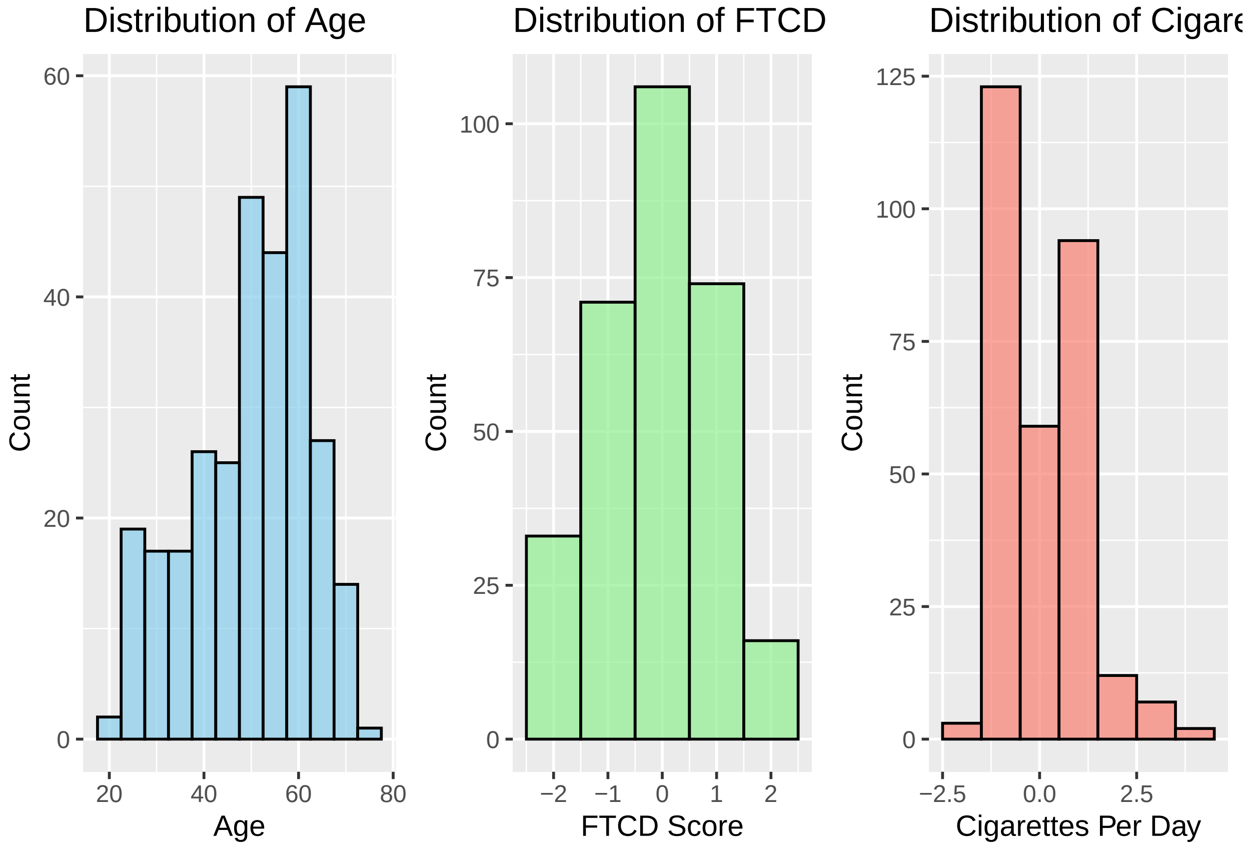
We can see in the table data was grouped into four treatment categories based on the combination of behavioral activation (BA) and varenicline therapy: BA Placebo, BA Varenicline, Control, and ST Varenicline. This table provides a summary of key characteristics across these groups, including age (age ps), smoking cessation readiness (readiness), and demographic variables such as gender (sex ps) and race indicators (NHW, Black, Hisp). Overall, the mean age across groups is around 50 years, with slight variations, and the readiness score remains relatively consistent, suggesting comparable baseline characteristics across treatment groups.

We then generates distribution plots for age_ps (Age), ftcd_score (Nicotine Dependence Score), and cpd_ps (Cigarettes Per Day) to understand the baseline characteristics of our sample. These

Characteristic	**Overall** N = 300	**BA_Placebo** N = 81	**BA_Varenicline** N = 83	**Control** N = 68	**ST_Varenicline** N = 68
age_ps	50 (13)	49 (13)	50 (13)	50 (11)	51 (14)
sex_ps					
1	135 / 300 (45%)	37 / 81 (46%)	39 / 83 (47%)	29 / 68 (43%)	30 / 68 (44%)
2	165 / 300 (55%)	44 / 81 (54%)	44 / 83 (53%)	39 / 68 (57%)	38 / 68 (56%)
NHW	105 / 300 (35%)	25 / 81 (31%)	34 / 83 (41%)	22 / 68 (32%)	24 / 68 (35%)
Black	157 / 300 (52%)	43 / 81 (53%)	37 / 83 (45%)	40 / 68 (59%)	37 / 68 (54%)
Hisp	18 / 300 (6.0%)	5 / 81 (6.2%)	4 / 83 (4.8%)	4 / 68 (5.9%)	5 / 68 (7.4%)
readiness	7 (1)	7 (1)	7 (1)	7 (1)	7 (1)

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

variables are critical in smoking cessation studies, as age, nicotine dependence, and smoking intensity can impact treatment success. The plots show a broad age range centered around 55-60, a normal distribution of nicotine dependence scores, and a skewed distribution of daily cigarette consumption, with most participants smoking at low to moderate levels. These insights help us identify potential patterns and tailor interventions based on participants' baseline characteristics.



Predictor Analysis

In this section, we aim to identify key baseline characteristics that serve as predictors of smoking abstinence at the end of treatment (EOT). Understanding which baseline variables are most predictive of abstinence can inform targeted intervention strategies and improve treatment efficacy. To accomplish this, we will use two primary methods.

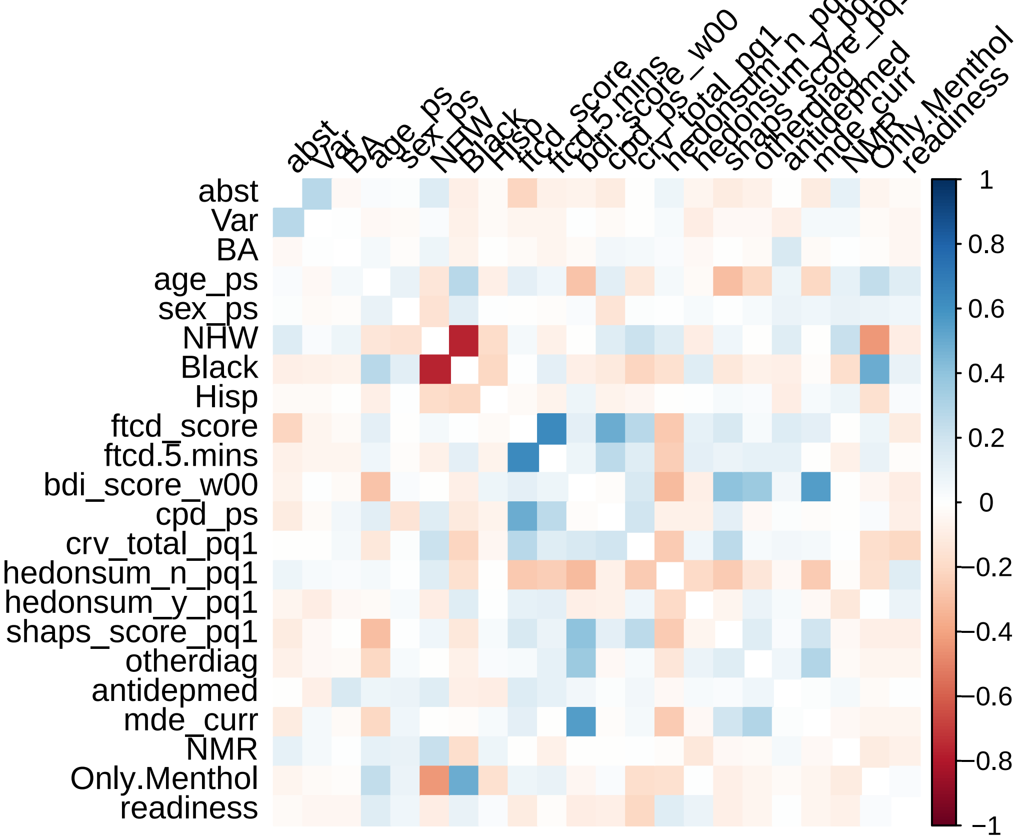
Firstly, we will compute a correlation matrix to examine the relationships between continuous baseline variables and the smoking abstinence outcome. This initial step will help us identify any strong linear associations between variables, providing insight into which baseline characteristics may have predictive power.

Following the correlation analysis, we will use Least Absolute Shrinkage and Selection Operator (LASSO) regression to select the most important predictors of smoking abstinence. LASSO is a regularization method that penalizes the absolute size of regression coefficients, effectively shrinking some coefficients to zero. This property makes LASSO particularly useful for variable selection, as it can reduce model complexity and identify the most influential predictors from a potentially large set of baseline variables. We will use cross-validation to select the optimal penalty parameter that minimizes prediction error.

By combining these methods, we aim to narrow down our set of baseline variables to a core group that best predicts smoking abstinence. These selected predictors will then be used in the subsequent Moderation Analysis to examine potential interactions with treatment effects.

Correlation Analysis

In the correlation matrix analysis, we included all numeric baseline variables, along with abst, which indicates smoking abstinence at the end of treatment. By examining correlations between abst and other numeric variables, we aim to identify baseline characteristics that may be predictive of smoking cessation success. Although abst is a binary variable, treating it as numeric allows us to observe approximate linear relationships with continuous variables, providing initial insights into potential predictors. This approach serves as a preliminary screening step, helping us to identify variables that might play a significant role in predicting abstinence, which we will further evaluate using LASSO regression.



From the matrix, we observe that abst has relatively weak correlations with most baseline variables, with no particularly strong linear relationships evident. Some baseline variables, such as ftcd_score and cpd_ps, show moderate correlations with each other, indicating a relationship between nicotine dependence and daily cigarette consumption. Additionally, NHW and Black exhibit a high negative correlation, likely due to racial categorization. Overall, this preliminary analysis suggests that individual baseline variables may not have strong standalone predictive power for abstinence. We will further investigate these variables using LASSO regression to select key predictors for abstinence in the next step.

Lasso Regression Analysis

Building on the insights gained from the correlation matrix, we now proceed with LASSO (Least Absolute Shrinkage and Selection Operator) regression to further refine our selection of key predictors for smoking abstinence (abst). While the correlation analysis provided a preliminary view of variable relationships, LASSO allows us to formally select the most influential baseline variables by applying a penalty to the regression coefficients, shrinking less important ones to zero. This step will help us identify the strongest predictors of abstinence, reducing model complexity and focusing on the most impactful variables.

```
## [1] "Best lambda for lasso: 0.02822"
```

```
## [1] "Selected variables with non-zero coefficients:"
```

```
##                           s0
## (Intercept)     -2.107925846
## Var              1.089308471
## NHW              0.409989502
## ftcd_score      -0.333795380
## shaps_score_pq1 -0.006668571
## mde_curr        -0.126868771
## NMR              0.029594387
```

The LASSO regression identified several key predictors of smoking abstinence. Using cross-validation, we determined the optimal lambda value to be 0.02822, minimizing the model's prediction error. The variables with non-zero coefficients included ftcd_score, shaps_score_pq1, mde_curr, and NHW (0.4099) and NMR (0.0296) are positively associated with abstinence, indicating that individuals with these characteristics may have higher probabilities of quitting smoking. Conversely, ftcd_score (-0.3338), shaps_score_pq1 (-0.0067), and mde_curr (-0.1269) exhibited negative coefficients, suggesting that higher scores on these measures may be associated with lower likelihoods of smoking cessation. Overall, the LASSO model successfully reduced the number of predictors by shrinking the coefficients of less relevant variables to zero, thus highlighting a subset of variables that are potentially most influential in predicting smoking abstinence in adults with major depressive disorder.

summary

In this section, we evaluated baseline variables as predictors of smoking abstinence, controlling for behavioral treatment and pharmacotherapy. The correlation matrix provided preliminary insights into the relationships between baseline variables and abstinence, while the LASSO regression allowed us to identify the most influential predictors by penalizing less important variables. Based on these analyses, we identified NHW, ftcd_score, shaps_score_pq1, mde_curr, and NMR as key predictors of abstinence. In the next section, we will examine the interaction terms between these selected predictors and treatment variables to explore whether these variables moderate the effects of behavioral treatment and pharmacotherapy on smoking cessation outcomes.

Moderation Analysis

The goal of this section is to examine baseline variables as potential moderators of the effects of behavioral treatment on end-of-treatment (EOT) smoking abstinence. Specifically, we aim to identify whether certain baseline characteristics can influence the effectiveness of two main treatment approaches: behavioral activation (`BA`) and pharmacotherapy (`Var`). By understanding how these baseline factors interact with treatment, we can potentially tailor smoking cessation strategies to individuals with major depressive disorder (MDD), enhancing treatment effectiveness.

To achieve this, we will employ a logistic regression model that includes main effects for both treatment variables (`BA` and `Var`) and all baseline variables (`Z`), as well as interaction terms between the treatment variables and a selected subset of baseline variables. This subset of baseline variables—`ftcd_score`, `shaps_score_pq1`, `mde_curr`, `NHW`, and `NMR`—was identified through LASSO regression as having the strongest association with smoking abstinence. By focusing on these variables for interaction terms, we aim to construct a model that is both informative and parsimonious.

The model can be represented as follows:

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 A_i + \beta_2 B_i + \sum_{k=1}^{K} \beta_{3k} Z_{ik} + \sum_{j=1}^{5} \gamma_{1j} A_i X_{ij} + \sum_{j=1}^{5} \gamma_{2j} B_i X_{ij}$$

where: - A_i represents pharmacotherapy (`Var`), - B_i represents behavioral treatment (`BA`), - Z_{ik} represents the main effects of all baseline variables in the dataset, - X_{ij} represents the selected baseline variables for interaction: `ftcd_score`, `shaps_score_pq1`, `mde_curr`, `NHW`, and `NMR`, - β coefficients capture the main effects of treatment and baseline variables, and - γ coefficients capture the moderation effects, representing how the selected baseline variables influence the effectiveness of each treatment.

In the next step, we will perform model selection to identify the significant main effects and interaction terms, ensuring the model is both interpretable and relevant to our research question.

Model selection

In this section, we continue to use LASSO regression for model selection to identify the most relevant predictors and interactions affecting smoking abstinence at the end of treatment (EOT). LASSO is particularly suited for high-dimensional data with many predictors, as it applies an L1 penalty that shrinks the coefficients of less relevant variables to zero, effectively selecting a subset of important predictors. By using LASSO, we aimed to simplify the model while retaining the most significant baseline variables and treatment interactions, thus achieving a balance between model interpretability and predictive power. The other method we learned, Ridge regression, is not ideal for variable selection in this context because it applies an L2 penalty, which shrinks coefficients towards zero but does not set any of them exactly to zero. As a result, ridge regression reduces the impact of less important predictors without fully eliminating them, meaning that all variables are retained in the model to some degree.Since our goal is not only to predict smoking abstinence outcomes but also to identify key baseline variables and treatment interactions, LASSO is preferable.

```
## [1] "Selected variables with non-zero coefficients:"
```

```
##                     s0
## (Intercept) -1.8915736
## Var          0.5211023
## NHW          0.2870637
## ftcd_score  -0.3206452
## Var:NMR      0.8992114
```

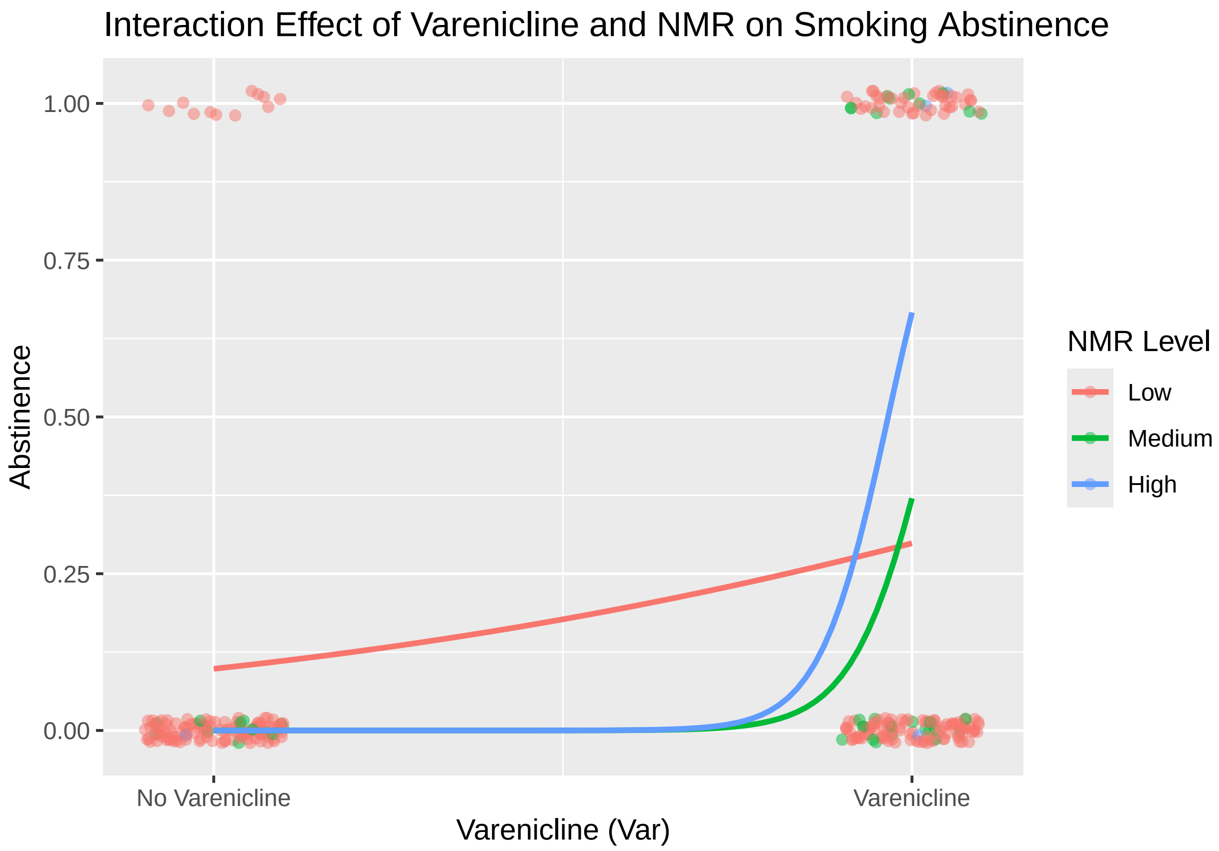
Interpretation

The LASSO model selected five variables with non-zero coefficients, indicating their significant influence on smoking abstinence outcomes. The intercept was -1.8916, serving as the baseline log-odds of abstinence when all predictors are at zero. The variable Var, representing Varenicline, had a positive coefficient (0.5211), suggesting that participants receiving Varenicline had a higher probability of refraining from smoking. NHW as expected also had a positive coefficient (0.2871), indicating that non-Hispanic White participants were more likely to achieve smoking abstinence.

ftcd_score, which measures nicotine dependence, had a negative coefficient (-0.3206), implying that higher nicotine dependence was associated with a lower probability of abstinence. Additionally, the interaction term Var:NMR (0.8992) was significant, indicating that the effect of Varenicline on smoking cessation was moderated by nicotine metabolism rate (NMR). This positive interaction suggests that participants with a higher NMR who received Varenicline were more likely to abstain, highlighting the potential for personalized treatment strategies based on nicotine metabolism.

We can use a plot to illustrate the effect, where we classify NMR into three levels: high, medium, and low.

```
## `geom_smooth()` using formula = 'y ~ x'
```



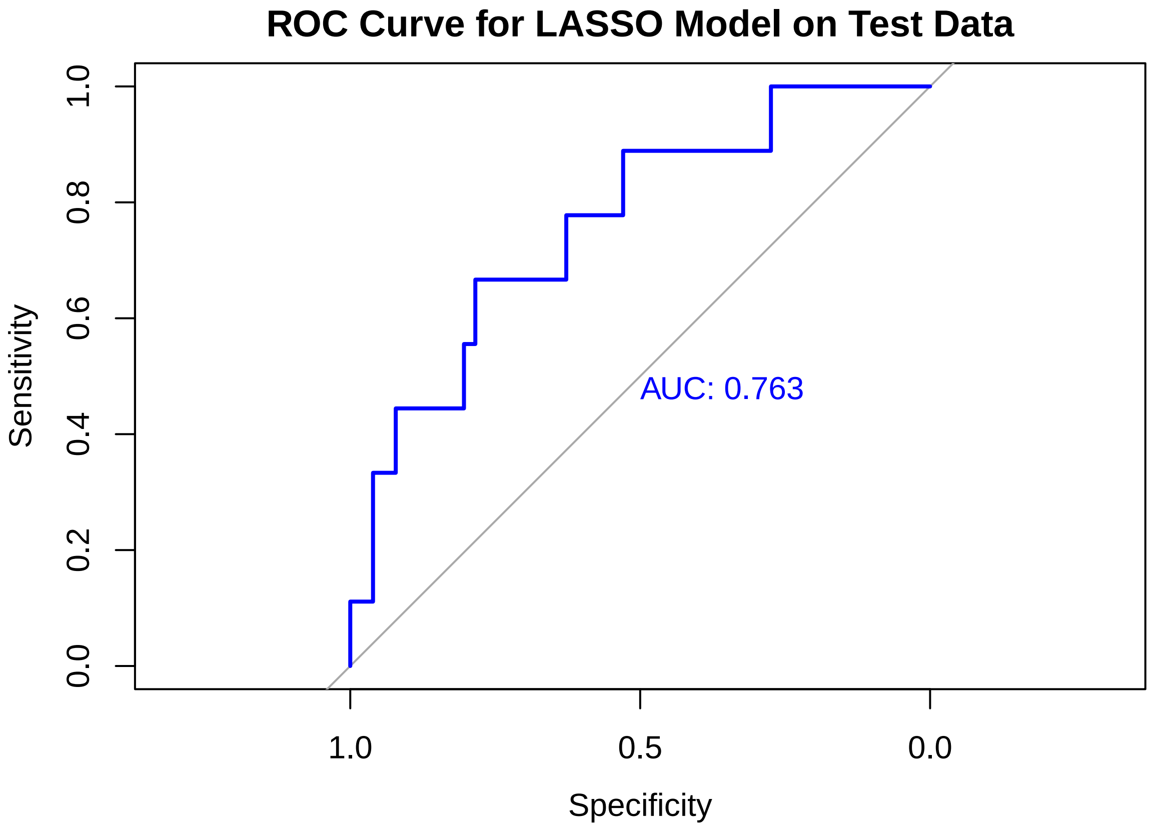
The plot illustrates the interaction effect between NMR levels and Varenicline treatment on the probability of abstinence. For participants who did not receive Varenicline, represented on the left side of the x-axis, the probability of abstinence remains consistently low across all NMR levels, suggesting that NMR alone does not significantly affect smoking cessation outcomes in the absence of Varenicline. However, for those who received Varenicline, the probability of abstinence increases with higher NMR levels. Participants with high NMR show a dramatic increase in abstinence probability, reaching nearly 100%, indicating that individuals with a high nicotine metabolism rate benefit substantially from Varenicline. Those with medium NMR also see an improvement in abstinence probability, though to a lesser extent, while individuals with low NMR experience only a slight increase in abstinence probability when taking Varenicline.

This analysis suggests that NMR moderates the effectiveness of Varenicline treatment, with higher NMR levels amplifying the positive effect of Varenicline on smoking abstinence. These findings support the potential for a personalized treatment approach, where NMR levels could be used to identify individuals who are more likely to benefit from Varenicline, thereby improving the overall success rate in smoking cessation programs.

Next step, we will examine the AUC and ROC for model fit.

Model fit

```
## [1] "AUC on test data: 0.763"
```



The model's performance was evaluated on test data using the area under the ROC curve (AUC), which was 0.763, as shown in the ROC curve plot. An AUC of 0.763 indicates a fair level of discrimination, meaning the model has a reasonable ability to distinguish between participants who achieve abstinence and those who do not. The ROC curve further demonstrates the model's predictive accuracy, with sensitivity and specificity balanced across various threshold levels.

sumamry

In summary, LASSO regression enabled effective variable selection, isolating key predictors and interactions that impact smoking abstinence outcomes. The selected model highlights the positive effect of Varenicline and identifies important baseline factors like NHW and ftcd_score, as well as the moderating role of NMR in the treatment effect. With an AUC of 0.763, the model demonstrates a fair predictive performance, supporting the potential for using these predictors in tailoring smoking cessation treatments for individuals with major depressive disorder.

Discussion

Findings

In this study, we examined how baseline variables influence the effectiveness of two treatments, Varenicline (Var) and Behavioral Activation (BA), on end-of-treatment (EOT) smoking abstinence in adults with major depressive disorder (MDD). Our analysis focused on identifying key baseline predictors of abstinence and exploring potential interaction effects between these baseline variables and the treatments. Using LASSO regression for variable selection, we identified important predictors such as ftcd_score (nicotine dependence score), NMR (nicotine metabolism rate), NHW (Non-Hispanic White status), and mde_curr (current major depressive episode). Additionally, our interaction analysis highlighted that NMR acts as a significant moderator of Varenicline's effect on abstinence, where higher nicotine metabolism rates lead to improved treatment outcomes with Var.

Implications

Our findings provide insights for personalized smoking cessation strategies in adults with MDD. The interaction between NMR and Varenicline suggests that individuals with high nicotine metabolism rates may benefit more from Varenicline treatment, potentially leading to higher abstinence rates in this subgroup, which means NMR levels could be assessed prior to treatment initiation to guide medication selection, thereby optimizing the chances of successful smoking cessation.

Limitations

While our analysis provides valuable insights, there are several limitations to consider:

1.Sample Size and Generalizability: The study sample may be limited in size, and findings might not generalize to broader populations outside the study group (e.g., individuals without MDD).

2.Missing Data and Imputation: Although we used multiple imputation to handle missing data, this approach can introduce uncertainty, particularly if the missingness is not completely at random.

3.Simplified Modeling of Interactions: We focused on a subset of interactions based on selected baseline variables, which may limit our ability to detect other potentially meaningful interactions. Future studies could explore a broader set of interaction terms to capture additional nuances.

4.Cross-Sectional Nature of Analysis: This analysis was based on EOT abstinence, a cross-sectional measure, rather than long-term abstinence outcomes. Longitudinal studies could provide a more comprehensive understanding of sustained smoking cessation over time.

Conclusion

In summary, these results indicate that baseline variables such as ftcd_score, NHW, mde_curr, and NMR play crucial roles in predicting abstinence outcomes and in moderating treatment effects. By incorporating these baseline factors, clinicians can better personalize smoking cessation interventions for adults with MDD, potentially improving success rates. These findings underscore the value of evaluating baseline characteristics not only as predictors of abstinence but also as modifiers of treatment efficacy, advancing our understanding of personalized approaches in smoking cessation.

References

1. Roufosse F, Kahn JE, Rothenberg ME, Wardlaw AJ, Klion AD, Kirby SY, Gilson MJ, Bentley JH, Bradford ES, Yancey SW, Steinfeld J, Gleich GJ; HES Mepolizumab study group. Efficacy and safety of mepolizumab in hypereosinophilic syndrome: A phase III, randomized, placebo-controlled trial. J Allergy Clin Immunol. 2020 Dec;146(6):1397-1405. doi: 10.1016/j.jaci.2020.08.037. Epub 2020 Sep 18. PMID: 32956756; PMCID: PMC9579892.

Code Appendix:

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(summarytools)
library(kableExtra)
library(RColorBrewer)
library(ggplot2)
library(dplyr)
library(randomForest)
library(tidyr)
library(mice)


# Load necessary libraries
library(dplyr)
library(tidyr)

# Load the dataset
data <- read.csv("project2.csv")
data$id <- NULL
# Inspect data structure
head(data)
# Calculate the number and percentage of missing values for each column, and filter to show only column
missing_data_summary <- data %>%
  summarise_all(~sum(is.na(.))) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Missing_Count") %>%
  mutate(Missing_Percentage = (Missing_Count / nrow(data)) * 100) %>%
  filter(Missing_Count > 0)  # Only keep variables with missing values

# Display the missing data summary
missing_data_summary

data <- mice(data, m = 5, method = "pmm", maxit = 50, seed = 500, printFlag=FALSE)
data <- complete(data)
data <- data %>%
  mutate(
    NHW = as.numeric(NHW == 1),
    Black = as.numeric(Black == 1),
    Hisp = as.numeric(Hisp == 1),
    inc = as.factor(inc),
    edu = as.factor(edu)
  )

# Standardize numerical variables
data <- data %>%
  mutate(
    ftcd_score = scale(ftcd_score),
    bdi_score_w00 = scale(bdi_score_w00),
    cpd_ps = scale(cpd_ps)
  )
# Load necessary libraries
library(dplyr)
```

```r
library(gtsummary)
library(kableExtra)

# We backup one to create table without changing original data
data_copy <- data

# Create a new grouping variable in the copied dataset based on Var and BA columns
data_copy <- data_copy %>%
  mutate(
    treatment_group = case_when(
      Var == 0 & BA == 0 ~ "Control",
      Var == 1 & BA == 0 ~ "BA_Placebo",
      Var == 0 & BA == 1 ~ "ST_Varenicline",
      Var == 1 & BA == 1 ~ "BA_Varenicline"
    )
  )

# Select a subset of important variables for the table
selected_vars <- c("age_ps", "sex_ps", "NHW", "Black", "Hisp", "readiness")

# Specify which variables are categorical
factor_vars <- c("sex_ps", "NHW", "Black", "Hisp")

# Generate a summary table grouped by the new treatment_group variable
summary_table <- tbl_summary(
  data_copy %>% select(all_of(selected_vars), treatment_group),  # Select only important variables and
  by = treatment_group,  # Group by treatment group
  type = list(readiness ~ "continuous"),  # Specify readiness as continuous
  statistic = list(
    all_continuous() ~ "{mean} ({sd})",  # Mean and SD for continuous variables
    all_categorical() ~ "{n} / {N} ({p}%)"  # Count and percentage for categorical variables
  ),
  missing = "no"  # Ignore missing data counts for simplicity
) %>%
  add_overall() %>%  # Add an overall summary column
  as_kable_extra(booktabs = TRUE) %>%
  kable_styling(latex_options = "scale_down")

# Display the summary table
summary_table

library(ggplot2)
library(gridExtra)

plot_age <- ggplot(data, aes(x = age_ps)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Age", x = "Age", y = "Count")

plot_ftcd <- ggplot(data, aes(x = ftcd_score)) +
  geom_histogram(binwidth = 1, fill = "lightgreen", color = "black", alpha = 0.7) +
  labs(title = "Distribution of FTCD Score", x = "FTCD Score", y = "Count")

plot_cpd <- ggplot(data, aes(x = cpd_ps)) +
  geom_histogram(binwidth = 1, fill = "salmon", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Cigarettes Per Day", x = "Cigarettes Per Day", y = "Count")

grid.arrange(plot_age, plot_ftcd, plot_cpd, ncol = 3)
# Load package
library(corrplot)

# Calculate correlation matrix for numeric variables
numeric_vars <- data %>% select(where(is.numeric))
cor_matrix <- cor(numeric_vars, use = "complete.obs")

# Plot the correlation matrix with only colors and no numbers
corrplot(cor_matrix, method = "color",
         tl.col = "black",       # Set text label color
         tl.srt = 45,            # Rotate text labels for better readability
         addCoef.col = NULL,     # Disable displaying correlation coefficients
         diag = FALSE)           # Hide the diagonal

library(glmnet)
set.seed(2001)
data1 <- data[, -1]
y <- data$abst
X <- model.matrix(~ BA + Var + . - 1, data = data1)  # "-1" removes the intercept for glmnet

# glmnet requires specifying family = "binomial" for logistic regression
lasso_model <- glmnet(X, y, family = "binomial", alpha = 1)

cv_lasso <- cv.glmnet(X, y, family = "binomial", alpha = 1)

best_lambda <- cv_lasso$lambda.min
print(paste("Best lambda for lasso:", round(best_lambda, 5)))

final_model <- glmnet(X, y, family = "binomial", alpha = 1, lambda = best_lambda)

lasso_coefficients <- as.matrix(coef(final_model))
print("Selected variables with non-zero coefficients:")
print(lasso_coefficients[lasso_coefficients != 0, , drop = FALSE])
# Load necessary packages
library(glmnet)
library(mice)
library(pROC)

set.seed(123)  # for reproducibility
train_index <- sample(1:nrow(data), 0.8 * nrow(data))
train_data <- data[train_index, ]
test_data <- data[-train_index, ]

# Construct the design matrix for the training data with interaction terms
X <- model.matrix(abst ~ Var + BA + . + Var:(ftcd_score + shaps_score_pq1 + mde_curr + NHW + NMR) +
                  BA:(ftcd_score + shaps_score_pq1 + mde_curr + NHW + NMR), data = train_data)
X <- X[,-1]  # Remove intercept
Y <- factor(train_data$abst)
```

```r
# Fit LASSO model with cross-validation
lasso_fit <- cv.glmnet(X, Y, family = "binomial", alpha = 1, type.measure = "auc", nfolds = 10)
best_lambda <- lasso_fit$lambda.min

# Final LASSO model with selected lambda
final_lasso_model <- glmnet(X, Y, family = "binomial", alpha = 1, lambda = best_lambda)

# Construct the design matrix for the test data using the same formula
test_X <- model.matrix(abst ~ Var + BA + . + Var:(ftcd_score + shaps_score_pq1 + mde_curr + NHW + NMR) +
                       BA:(ftcd_score + shaps_score_pq1 + mde_curr + NHW + NMR), data = test_data)
test_X <- test_X[,-1]  # Remove intercept

# Ensure column names of test_X match those of X
test_X <- test_X[, colnames(X), drop = FALSE]  # Reorder or subset columns to match training matrix

# Predict on test data
pred_probs <- predict(final_lasso_model, type = "response", newx = test_X)

# Extract selected variables and their coefficients
selected_vars <- as.matrix(coef(final_lasso_model))
selected_vars <- selected_vars[selected_vars != 0, , drop = FALSE]

# Display selected variables and coefficients
print("Selected variables with non-zero coefficients:")
print(selected_vars)

library(ggplot2)
library(dplyr)

data %>%
  mutate(NMR_group = cut(NMR, breaks = 3, labels = c("Low", "Medium", "High"))) %>%
  ggplot(aes(x = Var, y = abst, color = NMR_group)) +
  geom_jitter(width = 0.1, height = 0.02, alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  labs(
    x = "Varenicline (Var)",
    y = "Abstinence",
    color = "NMR Level",
    title = "Interaction Effect of Varenicline and NMR on Smoking Abstinence"
  ) +
  scale_x_continuous(breaks = c(0, 1), labels = c("No Varenicline", "Varenicline"))

library(pROC)

# Evaluate model performance on test data using AUC
auc_test <- roc(test_data$abst, as.numeric(pred_probs))
print(paste("AUC on test data:", round(auc(auc_test), 3)))

# Plot ROC curve
plot.roc(auc_test, main = "ROC Curve for LASSO Model on Test Data", col = "blue", print.auc = TRUE)
```