# Assessing Baseline Variables as Potential Moderators of the Behavioral Treatment Effects on End-of-treatment (EOT) Abstinence

Diahmin Hawkins

11/6/2024

## Abstract

**Background:** Smoking cessation among individuals with Major Depressive Disorder (MDD) presents unique challenges, including heightened nicotine dependence and withdrawal severity. Prior research has shown that pharmacotherapy (varenicline) and behavioral activation can improve abstinence rates; however, the role of baseline demographic, psychological, and clinical variables as moderators of treatment outcomes remains underexplored.

**Objective:** This study aims to assess baseline characteristics as potential moderators of behavioral and pharmacological treatment effects on end-of-treatment (EOT) smoking abstinence. Specifically, it evaluates how these variables influence treatment efficacy, including interaction effects between treatment type and participant characteristics.

**Methods:** Using data from a 2×2 factorial, randomized, placebo-controlled trial with 300 adult smokers (current or past MDD), multiple imputation was applied to address missing data (21.67% missing). Logistic and Lasso regression models were used to identify significant predictors and their interactions, followed by a comparative analysis of model performance based on Area Under the Curve (AUC) metrics. Key predictors included FTCD scores, readiness to quit, and demographic factors such as race/ethnicity.

**Results:** Several key baseline variables emerged as significant moderators of smoking abstinence. Higher Nicotine Metabolite Ratio (NMR) was associated with increased odds of abstinence (OR = 2.453, 95% CI: 1.187–5.046), potentially moderating the effectiveness of pharmacological treatments. Lower readiness to quit drastically reduced abstinence odds (OR = 0.038, 95% CI: 0.001–0.878), highlighting the need for motivational enhancement strategies for less-ready individuals. Race/ethnicity was a significant factor, with Non-Hispanic White individuals showing substantially higher odds of abstinence (OR = 5.82, 95% CI: 2.497–14.597), possibly due to socioeconomic and healthcare access disparities. Other moderators included menthol use (OR = 2.169, 95% CI: 1.335–3.565), cigarettes per day (cpd_ps) (OR = 0.921, 95% CI: 0.884–0.959), income level (OR = 0.491, 95% CI: 0.274–0.861), and age (OR = 1.034, 95% CI: 1.013–1.056, p = 0.0011). Logistic regression outperformed Lasso in predictive accuracy (AUC: training = 0.964; testing = 1), identifying robust predictors and interactions.

**Conclusion:** Baseline characteristics significantly moderated treatment effects on smoking cessation, with pharmacotherapy and demographic variables playing pivotal roles. Logistic regression outperformed Lasso in predictive accuracy, emphasizing the value of including interaction terms in modeling efforts. These findings underscore the importance of personalized interventions for enhancing smoking cessation outcomes among individuals with MDD.

# Introduction

Mental health disorders are among the most common health conditions associated with tobacco dependence. Studies have shown that smokers with depression find smoking more pleasurable and are more dependent on nicotine, leading to more severe withdrawal symptoms than smokers without major depressive disorder (MDD). "Smokers with depression are more likely to smoke heavily, perceive smoking as more pleasureable than other traditionally rewarding activities, show greater dependence and experience more severe withdrawal than smokers without MDD (Hitsman, Papandonatos, et al., 1711). Dr. George Papandonatos, from Brown University's highly regarded Biostatistics Department, investigated smoking cessation outcomes in adults diagnosed with MDD.

The motivation behind this study stems from the need to address tobacco dependence in individuals with major depressive disorder(MDD). Due to high prevalence of smoking among people with MDD, more than 30% of individuals with depression are daily smokers. Smokers that has depression, tend to smoke more heavily and frequently, which causes a greater dependence. To evaluate the impact of tobacco dependence on individuals with MDD, this study was conducted in research clinics at Northwestern University (Chicago, IL) and the University of Pennsylvania (Philadelphia, PA). Recruitment for this study was conducted between 06/01/2015 and 03/13/2020. This study employed a randomized, placebo-controlled, 2x2 factorial design to compare behavioral activation for smoking cessation (BASC) against standard behavioral treatment (ST), with an additional comparison of varenicline versus placebo. This study included 300 adult smokers with either current or past MDD.

Findings indicated that BASC did not surpass standard behavioral treatment in effectiveness, regardless of concurrent varenicline therapy.It also indicate that individuals with MDD tend to smoke more heavily, exhibit greater nicotine dependence, and endure more severe withdrawal symptoms than individuals without MDD. Studies have found strong evidence that the varenicline treatment improved short and long term abstinence rates compared to placebo among racially and socioeconomically diverse groups with varied motivations and psychiatric presentations (Hitsman, Papandonatos, et al., 1722). While varenicline has proven effective in supporting smoking cessation, addressing the psychological aspects of smoking behavior, particularly those linked to depression, may further enhance cessation rates among adults with MDD.

In this analysis, we will focus on three primary aims. The first aim is to assess baseline characteristics as potential moderators of the effects of behavioral treatment on end-of-treatment (EOT) abstinence. The second aim is to investigate whether baseline predictors of abstinence vary depending on the type of behavioral intervention and pharmacotherapy administered. The third aim is to identify which baseline variables (e.g., demographic, psychological, and clinical factors) have the strongest influence on abstinence outcomes. We hypothesize that certain baseline characteristics like (race, income, educ) will significantly interact with treatment type, influencing the likelihood of smoking cessation. interact with treatment type, influencing the likelihood of smoking cessation.

# Methods

## Missingness

The raw data used for this analysis consisted of 300 rows and 25 columns. To begin this analysis, I got the sum of missing values from the dataset by columns. From this analysis, it was observed that the variables `Income`, `FTCD Score at Baseline`, `Cigarette Reward Value at  Baseline`, `Anhedonia`, `Nicotine Metabolism Ratio`, `Exclusive Mentholated Cigarette User`, and`Baseline Readiness to Quit Smoking` contained missing data. The missing data were examined using the `naniar` package in R to determine the percentage missing and available in the data. Following this procedure, there is a percentage of 21.67% of missing data.The missing percentage goes as follows: Income(1%), FTCD Score at Baseline (.33%), Cigarette Reward Value at Baseline (6%), Anhedonia (1%), Nicotine Metabolism Ratio (7%), Exclusive Mentholated Cigarette User(.67%), and Baseline Readiness to Quit Smoking (5.67%). To further quantify the extent of missingness, the `naniar package` in R was employed to calculate the percentage of missing and available data.The missing data represent only .9% of the dataset, while 99.1% of the data remains well-represented. Therefore,these

Missing Data Summary for Smoking Sessation

| Variables | Missing Values | Percentage Missing (%) |
|---|---|---|
| Income | 3 | 1.00 |
| FTCD Score at Baseline | 1 | 0.33 |
| Cigarette Reward Value at Baseline | 18 | 6.00 |
| Anhedonia | 3 | 1.00 |
| Nicotine Metabolism Ratio | 21 | 7.00 |
| Exclusive Mentholated Cigarette User | 2 | 0.67 |
| Baseline Readiness to Quit Smoking | 17 | 5.67 |
| Total | 65 | 21.67 |

missing data properties led to the implementation imputation.

## Pre-processing

The initial exploratory analysis was conducted to identify patterns and relationships among key variables. During preprocessing, it was observed that there are treatment groups in the dataset, but the paper discussed a 2 by 2 factorial design to compare behavioral activation for smoking cessation (BASC) against standard behavioral treatment (ST). The four treatment groups that we will be exploring consists of `BASC + Varenicline` `ST + Placebo`, `BASC + Placebo` and `ST + Varenicline`.Due to limited number of participants in education levels, `Grade School, Some High School, and High School` were combined to `High School or Less` to make it more balance to other education levels in the data.

## Multiple Imputations

`Multiple Imputation` is a statistical process in the `mice` package that handles missing data by creating several datasets that fills in missing values.In the case here, we will implement 5 imputations for more accurate analyses by accounting for the uncertainty around the missing data, compared to single imputation methods that replace missing values with a single estimate.Each imputed dataset is then arrange into long format using the complete long function treating each as if it were the real, complete data. The statistical model we will be anticipating throughout this process is `Logistic` and `Lasso` regression. Following these results, we will examine summary statitiscs like the p-values, odd ratios, and beta estimates to examine the best potential moderators for abstinence for smokers that experience MDD.

## EDA (Exploratory Data Analysis)

To investigate the relationships between categorical variables, we utilized barplots to examine significant associations between baseline characteristics, such as income and education levels, and abstinence outcomes. This approach involves comparing observed frequencies across different categories to determine if the differences are statistically significant or likely due to chance. By assessing the independence of two categorical variables, we aim to identify patterns, relationships, and potential associations between risk factors and abstinence.

Barplots were used to visually evaluate these relationships, highlighting interactions between two categorical variables. Patterns in bar heights or proportions provided insights into how one variable influences or interacts with another. Additionally, density plots were included to observe data concentration, with peaks indicating areas of high frequency. Overlaying these density plots allowed for a comparative analysis of differences in distributions, further revealing potential trends or differences in the data.

### Test Train Splits on models

In our logistic and lasso models, we will implement a test, train, split process with 70% on the trained data and 30% on the test data.

### Logistic

The logistic regression model doesn't use a penalty term and it's objective is to model the probability of a binary outcome Pr(Y=1|X) OR Pr(Y=0|X). Logistic regression uses the MLE (Maximum Likelihood Estimate) of the observed data without imposing any regularization on the coefficients. Logistic regression finds the coefficients that best predict the outcome (abstinence in this case) based on the given predictors by maximizing the likelihood (or minimizing the negative log-likelihood).

To refine the model, variables with statistical significance ($\alpha <= .05$) will be selected and observed for their summary statistics of p-values, beta coefficients, confidence intervals, and odd ratios. This step enables the identification of potential moderating factors that may influence abstinence outcomes among participants which provide insights into key predictors.

### Lasso

In the lasso regression model, we use the $l_1$ penalty rather than the $l_2$, where we take the absolute value of the $\beta$ rather than the squaring them. The $l_1$ penalty has the effect of forcing some of the coefficients to be exactly equal to zero. Lasso performs variable selection and models are easier to interpret that produces sparse models due to all the zeroes represented inside of the model.After observing our $\beta$ coefficients, we observe the non-zero coeffiecients in both our main effects model and interaction term model.Following the test, train , split, we will implement the process of crossvalidation and using the respective lasso mechanics. We will pool the results using `Rubin's Rules` find the odd ratios, means,confidence intervals and other statistical attributes.Then we observe the non-zero coefficients to represent the predictor variables that have the most impact on our model and on the prediction of abstinence.

## Comparison Analysis (Area Under the Curve (AUC))

Following the regression model analysis, we will evaluate and compare the performance of the logistic and LASSO regression models. This comparison aims to assess each model's predictive accuracy, interpretability, and ability to identify significant predictors of abstinence.We'll compare models using AUC on both training and testing data to assess generalization and discrimination between abstinent and non-abstinent outcomes. Higher AUC indicates better performance.

## Summary Statistics Table

This table provides a detailed breakdown of baseline characteristics for participants in four treatment groups: BASC + Placebo, BASC + Varenicline, ST + Placebo, and ST + Varenicline. Demographically, there is a high representation of Black participants, ranging from 45% in the BASC + Varenicline group to 59% in the ST + Placebo group. Non-Hispanic White participants make up a smaller proportion, ranging from 31% in the ST + Varenicline group to 41% in the BASC + Varenicline group. Hispanic participants account for less than 8% in all groups, with the lowest representation (4%) in BASC + Varenicline. Regarding sex, females are slightly more represented across all groups, with percentages ranging from 53% to 57%. These demographic trends highlight a study population where Black participants and females are highly represented, potentially reflecting disparities in smoking prevalence or recruitment efforts targeting underserved populations.

Socioeconomic factors also provide important context for understanding the challenges these participants may face in smoking cessation. A significant proportion of participants report annual incomes below $20,000, ranging from 36% in BASC + Varenicline and ST + Varenicline to 38% in ST + Placebo. In contrast, those with incomes above $75,000 are minimally represented, accounting for only 9–16% across groups. Educational attainment reveals that a large percentage of participants (30–39%) have attended some college or technical

school, while a smaller subset (25–35%) are college graduates. This socioeconomic profile indicates that the study population includes a high proportion of individuals from lower-income and mid-level educational backgrounds, groups often at higher risk for smoking persistence and less likely to access cessation resources. These demographic and socioeconomic factors are critical in interpreting treatment effects, as they may influence both baseline smoking behaviors and responsiveness to cessation interventions.

| Characteristic | BASC + Placebo N = 68[1] | BASC + Varenicline N = 83[1] | ST + Placebo N = 68[1] | ST + Varenicline N = 81[1] |
|---|---|---|---|---|
| Age | 54 (42, 61) | 53 (40, 60) | 51 (45, 58) | 52 (41, 59) |
| Sex | | | | |
| Female | 38 (56%) | 44 (53%) | 39 (57%) | 44 (54%) |
| Male | 30 (44%) | 39 (47%) | 29 (43%) | 37 (46%) |
| Non-Hispanic White | 24 (35%) | 34 (41%) | 22 (32%) | 25 (31%) |
| Black | 37 (54%) | 37 (45%) | 40 (59%) | 43 (53%) |
| Hispanic | 5 (7.4%) | 4 (4.8%) | 4 (5.9%) | 5 (6.2%) |
| Income | | | | |
| $20,000–35,000 | 16 (24%) | 17 (20%) | 14 (21%) | 21 (26%) |
| $35,001–50,000 | 8 (12%) | 13 (16%) | 14 (21%) | 11 (14%) |
| $50,001–75,000 | 12 (18%) | 12 (14%) | 8 (12%) | 6 (7.4%) |
| Less than $20,000 | 25 (37%) | 30 (36%) | 26 (38%) | 29 (36%) |
| More than $75,000 | 6 (8.8%) | 10 (12%) | 6 (8.8%) | 13 (16%) |
| Unknown | 1 (1.5%) | 1 (1.2%) | 0 (0%) | 1 (1.2%) |
| Education Level | | | | |
| College graduate | 19 (28%) | 29 (35%) | 17 (25%) | 26 (32%) |
| Some college/technical school | 22 (32%) | 32 (39%) | 38 (56%) | 24 (30%) |
| Some High School or Less | 27 (40%) | 22 (27%) | 13 (19%) | 31 (38%) |
| FTCD Score | 5.00 (4.00, 7.00) | 5.00 (4.00, 7.00) | 6.00 (4.00, 7.00) | 5.00 (4.00, 7.00) |
| Unknown | 0 | 0 | 1 | 0 |
| FTCD Score (5 mins) | 32 (47%) | 33 (40%) | 35 (51%) | 38 (47%) |
| BDI Score | 18 (9, 27) | 18 (10, 25) | 18 (12, 25) | 18 (11, 27) |
| Cigarettes per day | 15 (10, 20) | 15 (10, 20) | 13 (10, 20) | 15 (10, 20) |
| Craving Total | 7.0 (5.0, 10.0) | 8.0 (4.5, 10.0) | 7.0 (4.5, 9.0) | 7.0 (5.0, 9.0) |
| Unknown | 1 | 3 | 8 | 6 |
| Hedonic Sum (Negative) | 21 (10, 31) | 20 (9, 32) | 14 (9, 27) | 20 (9, 35) |
| Hedonic Sum (Positive) | 23 (14, 34) | 17 (11, 31) | 25 (12, 38) | 21 (13, 34) |
| Shaps Score | 0.00 (0.00, 3.00) | 1.00 (0.00, 4.00) | 1.00 (0.00, 5.00) | 1.00 (0.00, 3.00) |
| Unknown | 2 | 0 | 1 | 0 |
| Other Diagnoses | 35 (51%) | 30 (36%) | 28 (41%) | 40 (49%) |
| Antidepressant Medication | 28 (41%) | 24 (29%) | 15 (22%) | 15 (19%) |
| Current Major Depression Episode | 32 (47%) | 40 (48%) | 31 (46%) | 44 (54%) |
| Nicotine Metabolism Ratio | 0.32 (0.23, 0.46) | 0.33 (0.22, 0.50) | 0.32 (0.20, 0.43) | 0.29 (0.20, 0.51) |
| Unknown | 7 | 3 | 2 | 9 |
| Only Menthol | 40 (59%) | 48 (59%) | 43 (64%) | 47 (58%) |
| Unknown | 0 | 1 | 1 | 0 |
| Readiness to Quit | | | | |
| 3 | 1 (1.6%) | 0 (0%) | 0 (0%) | 0 (0%) |
| 4 | 2 (3.1%) | 2 (2.6%) | 1 (1.6%) | 0 (0%) |
| 5 | 6 (9.4%) | 11 (14%) | 9 (14%) | 9 (12%) |
| 6 | 18 (28%) | 22 (28%) | 14 (22%) | 29 (38%) |
| 7 | 16 (25%) | 21 (27%) | 16 (25%) | 18 (23%) |
| 8 | 17 (27%) | 20 (26%) | 19 (30%) | 18 (23%) |
| 9 | 2 (3.1%) | 1 (1.3%) | 2 (3.1%) | 2 (2.6%) |
| 10 | 2 (3.1%) | 1 (1.3%) | 3 (4.7%) | 1 (1.3%) |
| Unknown | 4 | 5 | 4 | 4 |

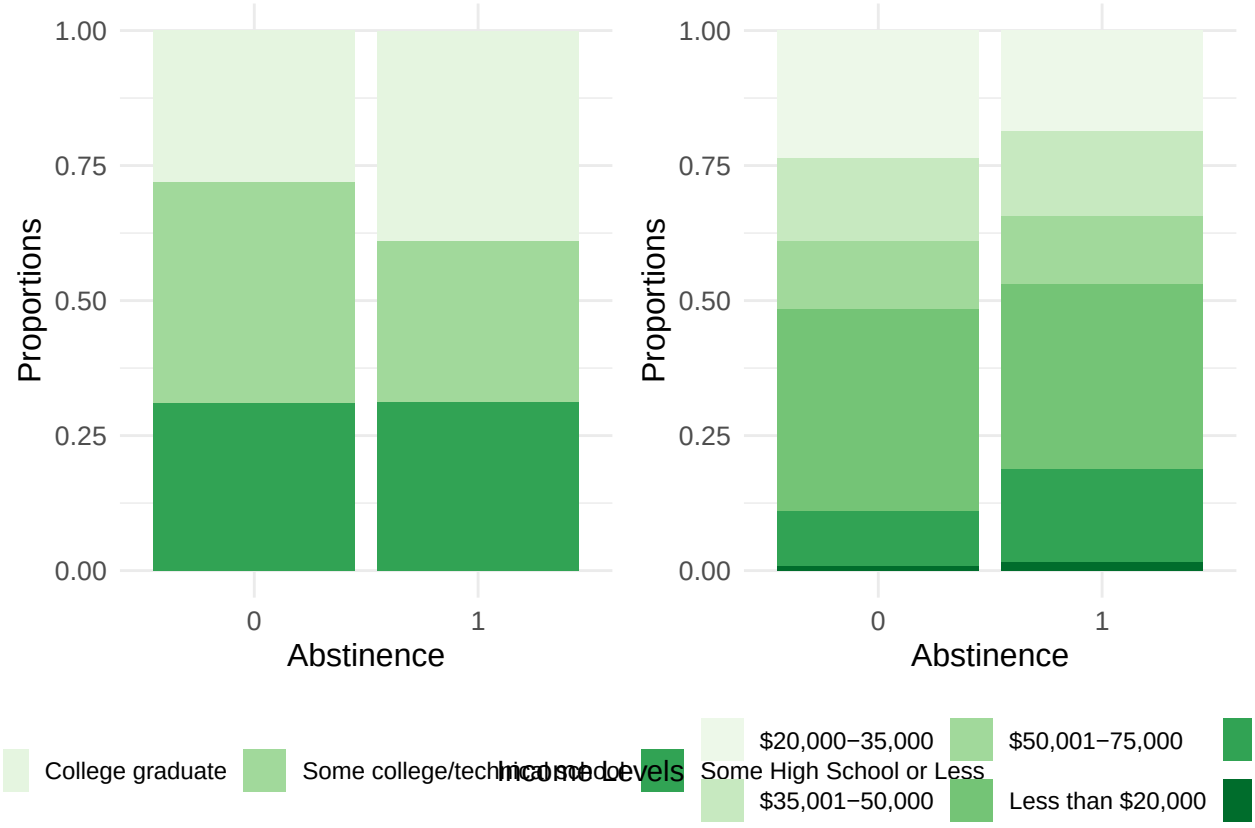
[1] Median (Q1, Q3); n (%)

## Exploratory Data Analysis

In the **Baseline Characteristics of Abstinence Status by Socioecomonic Levels** ,the stacked bar plot offers a clear visual depiction of the association between education and income levels based on abstinence status.The plot indicates that individuals with higher education levels, such as "College Graduate," appear more prevalent among the abstinent group compared to the non-abstinent group, while those with "Some High School or Less" education are more represented in the non-abstinent group. This suggests a potential positive relationship between higher education levels and the likelihood of abstinence.

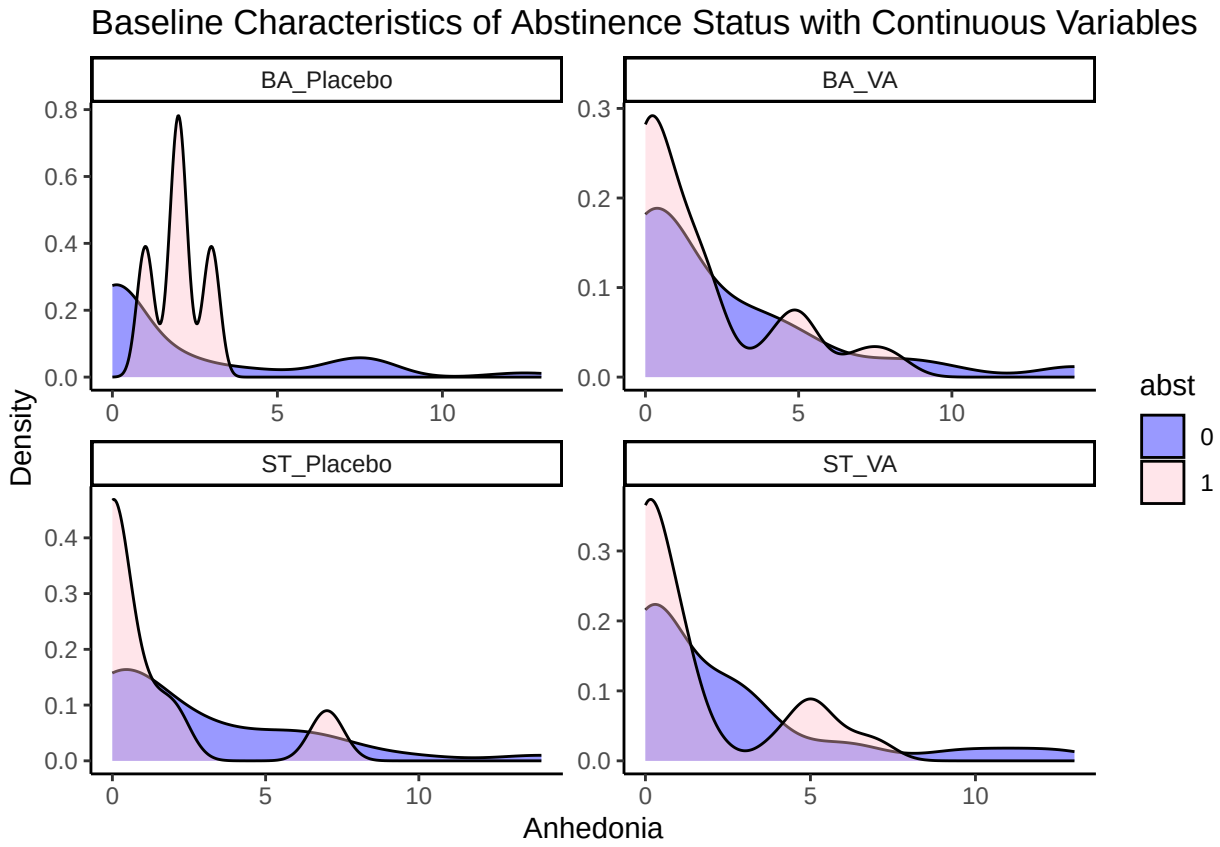
The proportions of individuals in each income bracket appear relatively consistent between the abstinent and non-abstinent groups, with no clear dominance of any specific income level in either group. However, a slight trend may suggest higher proportions of abstinent individuals in middle to higher income brackets (e.g., "$35,001–$75,000"), whereas lower income categories (e.g., "Less than $20,000") appear more prominent in the non-abstinent group. Through further analysis, it would be interesting to see if income and education levels may be potential moderators for abstinence.

**Baseline Characteristics of Abstinence Status by Socioecomonic Levels**




`Anhedonia` is the inability to feel pleasure, experience joy, or pleasure. Anhedonia is a common symtom of depression and mental health disorders. Based off the observations **Baseline Characteristics of Abstinence Status with Continous Variables** plot, the BA_Placebo group exhibit lower anhedonia scores in the abstinent group (blue) compared to non-abstinent individuals (red), with the density peaks concentrated at lower values. This is similarly reflected in the BA_VA group, where the abstinent participants also tend to have lower anhedonia scores, though the distributions for the two groups overlap more. The ST_Placebo group, a similar trend is observed, where abstinent individuals have a higher density at lower anhedonia scores, suggesting less severe anhedonia among those who are abstinent. The ST_VA group, demonstrated a consisted distribution towards lower anhedonia scores in the abstinent group compared to their non-abstinent counterparts. These plot offers interesting insights due to the overlapping different treatment groups which contributes to lower anhedonia scores are associated with higher likelihoods of abstinence, regardless of the

treatment group.

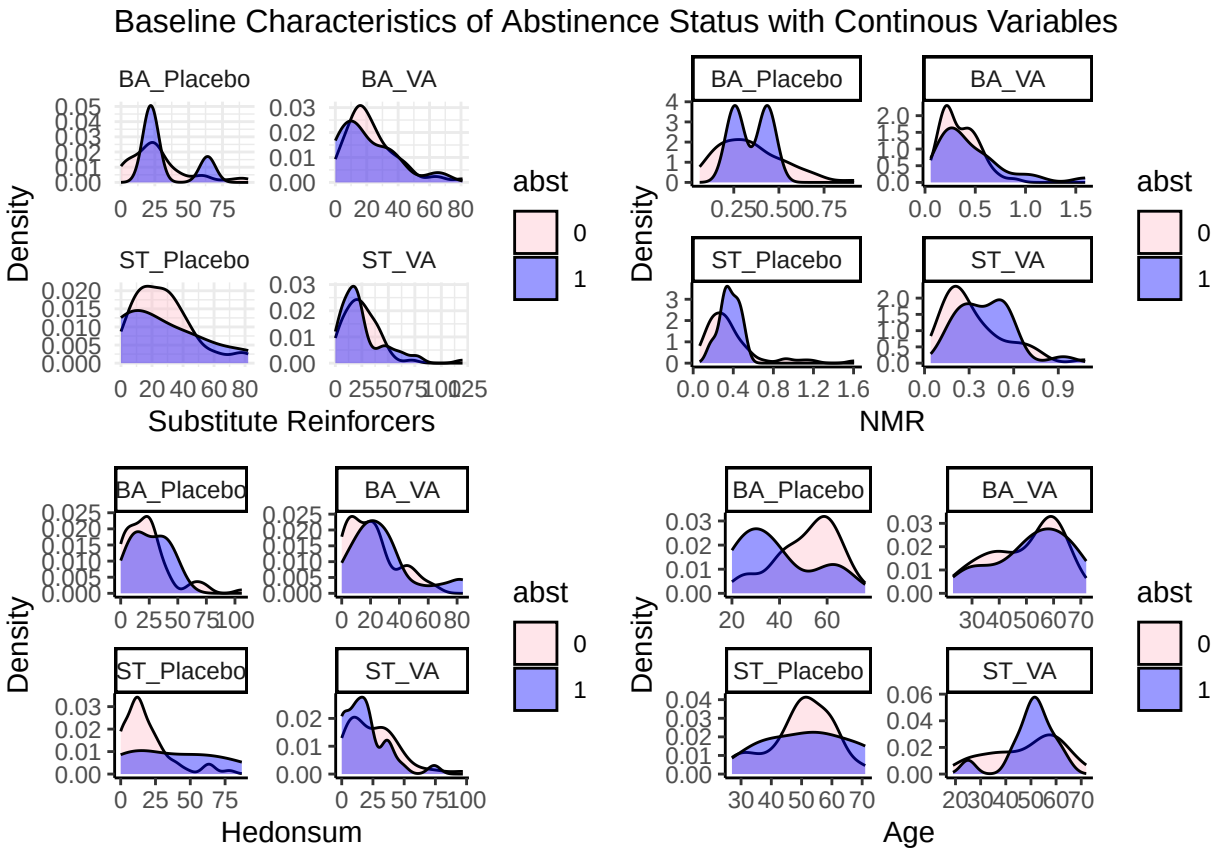



For substitute reinforcers, abstinent individuals (blue) consistently exhibit a higher density at lower scores compared to non-abstinent individuals, particularly in the BA_VA and ST_VA groups. This suggests that individuals who rely less on substitute reinforcers may have a greater likelihood of abstinence, potentially due to a higher capacity for intrinsic motivation in these groups.

Regarding `NMR`, there is noticeable variation across the treatment groups. In the BA_Placebo and ST_Placebo groups, abstinent individuals tend to cluster around moderate to high NMR values, indicating a potential metabolic or physiological influence on treatment efficacy. However, in the BA_VA and ST_VA groups, the distributions are more similar, suggesting that NMR may have a weaker predictive value in these settings.

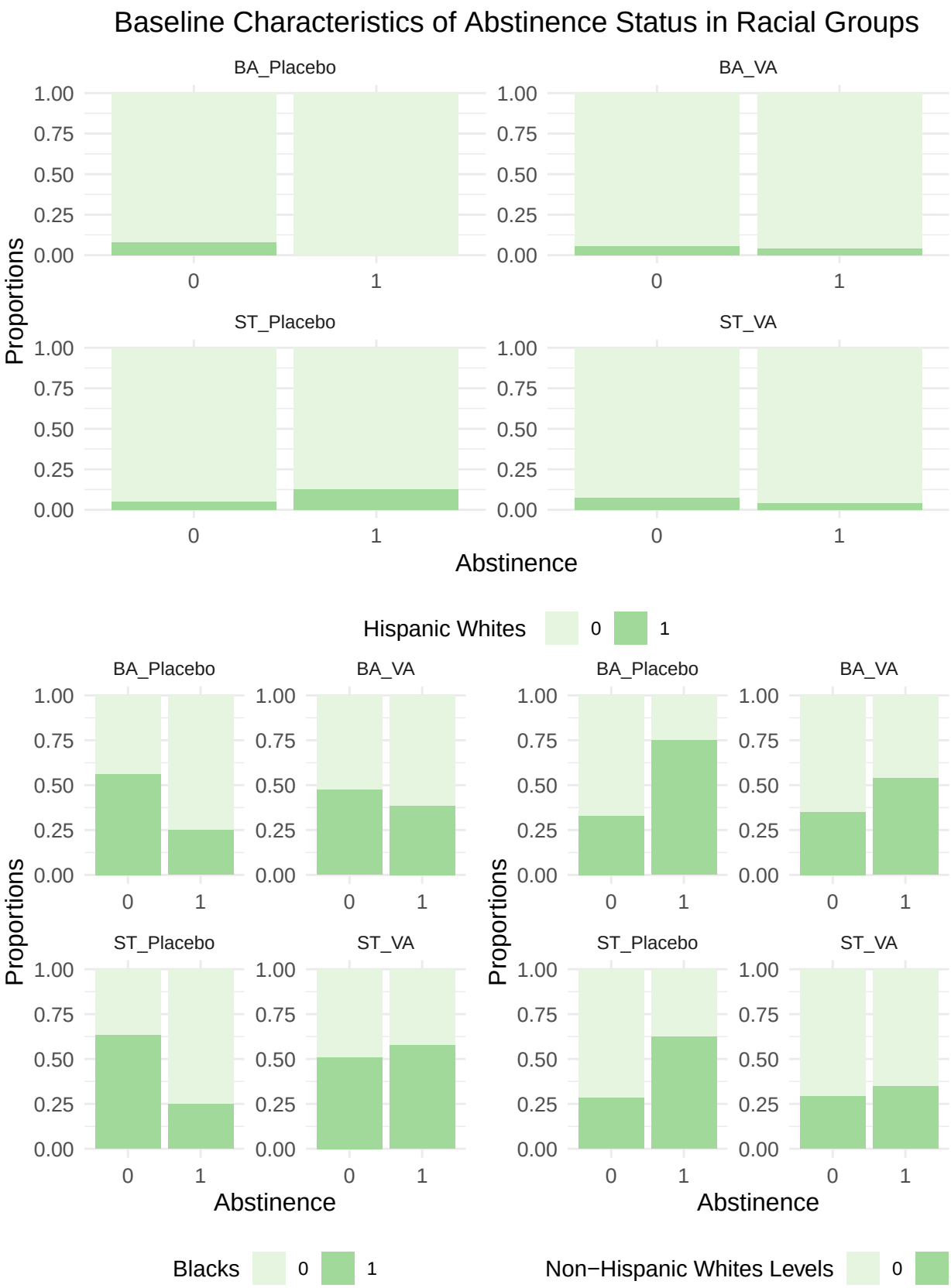
For `Hedonic` capacity, the BA_Placebo and ST_Placebo groups show higher densities for abstinent individuals at moderate hedonic scores, while non-abstinent participants are more evenly distributed across the range. This indicates that hedonic capacity may be a key moderator for predicting abstinence, as individuals with moderate to high capacity for experiencing pleasure may respond better to treatment.

In terms of `Age`, the relationship with abstinence varies by treatment group. In the BA_Placebo group, abstinent individuals are skewed toward younger ages, while the ST_Placebo group shows a more balanced distribution. In the BA_VA and ST_VA groups, older participants appear more likely to be abstinent, suggesting age may interact with the type of intervention to influence outcomes.




According to **Baseline Characteristics of Abstinence Status in Racial Groups**, the proportion of abstinent individuals appears slightly lower compared to non-abstinent individuals in the `Black` group, suggesting a potential disparity in abstinence rates. For `Non-Hispanic Whites`, the proportions of abstinent and non-abstinent individuals appear relatively balanced, indicating no strong visual difference in abstinence outcomes for this group. In contrast, the Hispanic White group shows a much smaller representation of abstinent individuals compared to non-abstinent individuals, suggesting that this group may face unique barriers to achieving abstinence.

These patterns highlight potential racial disparities in abstinence outcomes, with Black and Hispanic White individuals appearing less likely to achieve abstinence compared to Non-Hispanic Whites. This may be bias,because the racial groups were highly sampled in the data compared to `Non-Hispanic Whites`.

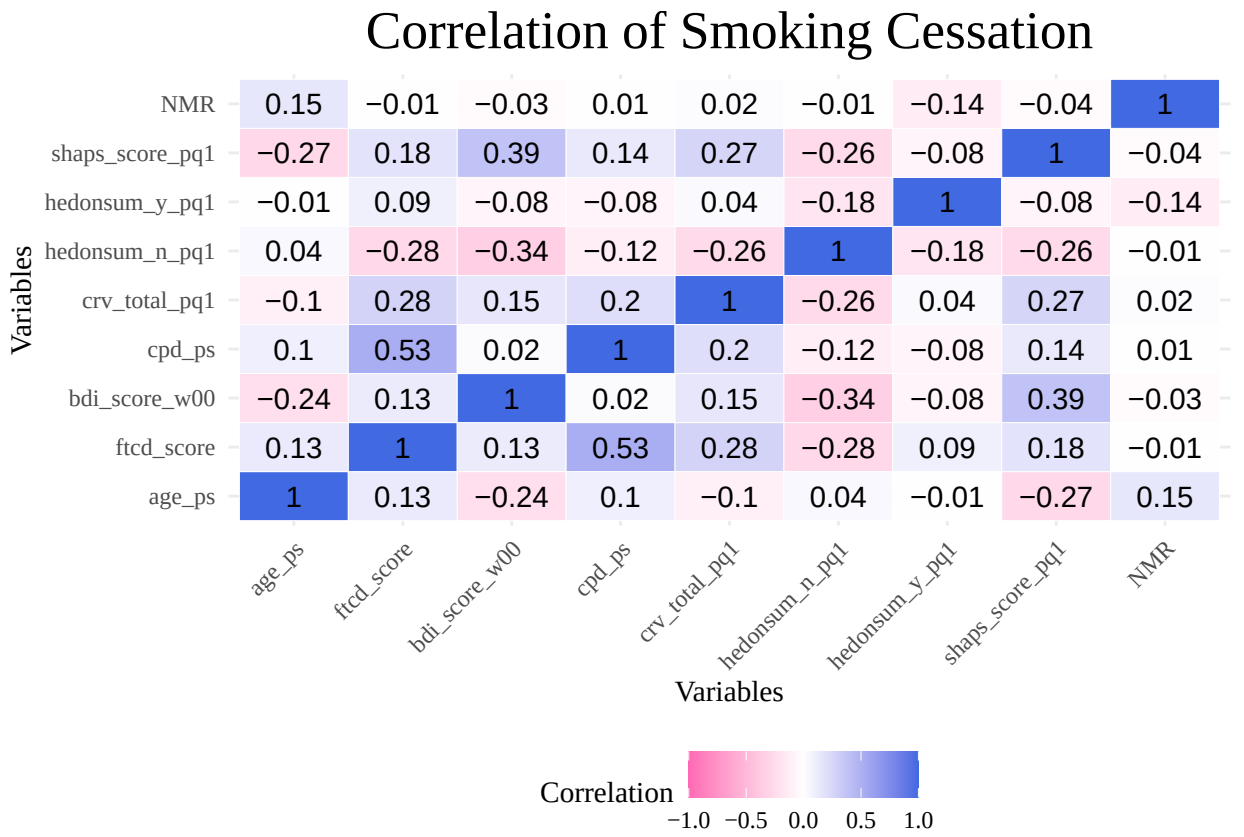

Baseline Characteristics of Abstinence Status in Racial Groups
BA_Placebo
BA_VA
ST_Placebo
ST_VA
Proportions
1.00
0.75
0.50
0.25
0.00
0
1
Abstinence
Hispanic Whites
0
1
Blacks
Non−Hispanic Whites Levels

## Correlation Plot

According to the `Correlation to Smoking Cessation Plot`, the highest postive correlation presented is the ftcd_score with cpd_ps of (.53) which may introduce multicollinearity in our regression models. The rest of the plot indicates weaker associations which are more ideal for our regression. Depression Score (bdi_score_w00) demonstrates a notable positive correlation with shaps_score_pq1 (0.39), which might indicate that higher depression levels are associated with anhedonia or reduced pleasure response (as measured by the SHAPS score).Negative correlations with several variables like age_ps and hedonsum_n_pq1 (-0.34), indicating that higher depression scores might be associated with specific clinical or psychological profiles. Anhedonia (shaps_score_pq1) offers a positive correlation with bdi_score_w00 (depression score, 0.39), highlighting an association between depressive symptoms and anhedonia. Negative correlations with abstinence-related measures (hedonsum_n_pq1, -0.26), which could indicate that anhedonia is negatively associated with behaviors linked to smoking cessation. Craving (crv_total_pq1) illustrates a moderate positive correlation with shaps_score_pq1 (0.27), potentially indicating that craving and anhedonia are related. High craving scores might represent a barrier to smoking cessation, as they indicate higher dependence and difficulty abstaining.NMR (Nicotinic Metabolism Rate) indicate aminor correlation with other variables, suggesting it may not be as directly related to the psychological measures shown here but could independently influence cessation outcomes by affecting nicotine processing and addiction levels.

Abstinence from smoking may be more challenging for individuals with higher craving levels, depressive symptoms, and anhedonia. The strong relationships between these psychological variables indicate a potential cumulative effect, where individuals facing multiple psychological challenges might experience a higher barrier to achieving and maintaining abstinence.

This matrix provides valuable insights into the psychological and demographic factors that may need to be addressed to improve smoking cessation success, highlighting the importance of targeting depressive symptoms, managing craving, and enhancing motivation in cessation interventions.



## Logistic Main Effects

Several key variables emerge as potential moderators for abstinence, influencing how other factors or interventions impact cessation outcomes. One significant moderator is the **Nicotine Metabolite Ratio (NMR)**, which reflects an individual's metabolic capacity to process nicotine. With an odds ratio of 2.453 (95% CI: 1.187–5.046), higher NMR values are associated with increased odds of abstinence. This suggests that NMR could moderate the effectiveness of pharmacological treatments, such as nicotine replacement therapy, as faster metabolizers may require higher doses or alternative strategies to achieve success.

Another critical factor is **readiness to quit**. Individuals with lower readiness to quit (readiness9) have drastically reduced odds of abstinence, with an odds ratio of 0.038 (95% CI: 0.001–0.878). Readiness may moderate the relationship between treatment intensity and outcomes, where higher readiness levels may enhance the effectiveness of behavioral interventions, while lower readiness levels might necessitate motivational enhancement strategies. Similarly, **race/ethnicity** plays a significant role, with Non-Hispanic White (NHW1) individuals showing an odds ratio of 5.82 (95% CI: 2.497–14.597), indicating much higher odds of abstinence compared to other racial/ethnic groups. Race could moderate the impact of socioeconomic factors, such as income and access to culturally tailored healthcare resources, which may affect treatment outcomes.

Other moderators include **menthol use** and **cigarettes per day (cpd_ps)**. Individuals who exclusively smoke menthol products have 116.9% higher odds of abstinence (odds ratio: 2.169, 95% CI: 1.335–3.565), potentially reflecting unique behavioral or physiological patterns in menthol smokers. This could moderate the effectiveness of treatment modalities, as menthol smokers may respond differently to pharmacological or behavioral interventions. On the other hand, higher cigarette consumption significantly reduces abstinence odds, with an odds ratio of 0.921 (95% CI: 0.884–0.959). Cigarette consumption may moderate the relationship between treatment type and cessation outcomes, where heavier smokers likely require more intensive or combined therapy approaches.

Socioeconomic factors, such as **income level (inc2)**, also show a significant moderating effect. Individuals in lower income brackets are 50.9% less likely to achieve abstinence (odds ratio: 0.491, 95% CI: 0.274–0.861), indicating that affordability and accessibility of resources could moderate cessation success. **Age (age_ps)** is another important moderator, with older individuals showing higher odds of abstinence (odds ratio: 1.034, 95% CI: 1.013–1.056, p-value:0.00111). Age may moderate the effectiveness of interventions, as younger individuals might require different or more intensive strategies compared to older adults, who may be more motivated by health concerns or long-term smoking behavior patterns.

In summary, variables such as **NMR**, **readiness to quit**, **race/ethnicity**, **menthol use**, **cigarettes per day**, **income level**, and **age** are critical moderators that influence the effectiveness of treatments and the likelihood of abstinence. Understanding and addressing these moderators can guide tailored, individualized interventions, ultimately improving cessation outcomes for diverse populations.

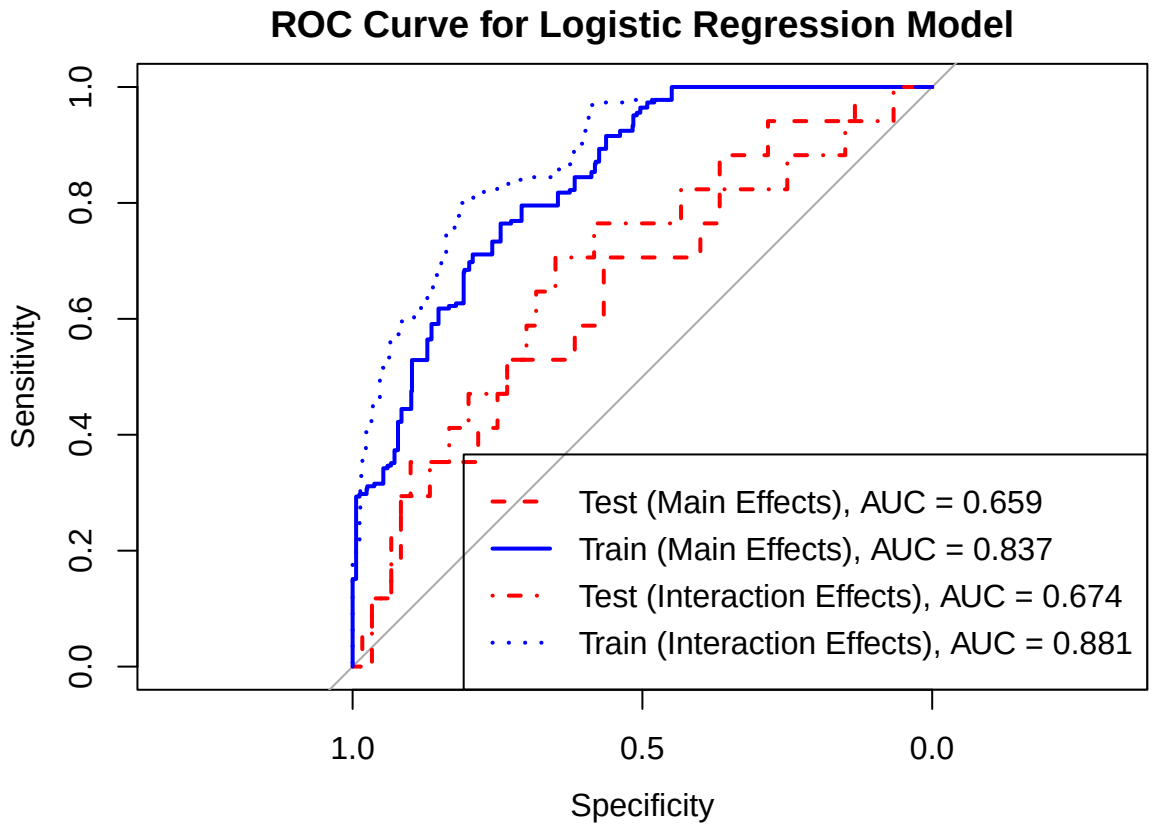
#Logistic Interaction Results Positive predictors include **age** (OR: 1.03), **Non-Hispanic White (NHW)** race (OR: 5.82), and **menthol use** (OR: 2.17), all significantly increasing the odds of abstinence. In contrast, negative predictors such as **cigarettes per day** (OR: 0.92), **low income levels** (e.g., inc2: OR: 0.49), and **current major depression** (OR: 0.66) significantly reduce the odds. Notably, **readiness to quit** is a critical factor, where lower readiness levels (readiness9: OR: 0.04) drastically decrease abstinence likelihood.

| Variable | Estimate | Std. Error | z value | Pvalue | Odds Ratios | Lower Conf. | Upper Conf. |
|---|---|---|---|---|---|---|---|
| age_ps | 0.03372 | 0.01034 | 3.26154 | 0.00111 | 1.03429919 | 1.013784799 | 1.0557966 |
| BA1 | -1.44057 | 0.43135 | -3.33966 | 0.00084 | 0.23679232 | 0.097605331 | 0.5365057 |
| cpd_ps | -0.08184 | 0.02063 | -3.96780 | 0.00007 | 0.92141492 | 0.884188515 | 0.9587546 |
| inc2 | -0.71296 | 0.29151 | -2.44572 | 0.01446 | 0.49019013 | 0.273563876 | 0.8600788 |
| NHW1 | 1.76130 | 0.44778 | 3.93340 | 0.00008 | 5.82000635 | 2.496757092 | 14.5965742 |
| NMR | 0.89738 | 0.36773 | 2.44034 | 0.01467 | 2.45317212 | 1.186540044 | 5.0457181 |
| Only.Menthol1 | 0.77433 | 0.25023 | 3.09447 | 0.00197 | 2.16913248 | 1.334872735 | 3.5653320 |
| readiness9 | -3.27570 | 1.64530 | -1.99094 | 0.04649 | 0.03779048 | 0.001021493 | 0.8781230 |
| sex_ps2 | -0.57481 | 0.21921 | -2.62222 | 0.00874 | 0.56281421 | 0.364560624 | 0.8620887 |
| Var1 | 1.73726 | 0.29346 | 5.91997 | 0.00000 | 5.68172908 | 3.245743547 | 10.2838379 |

| Variable | Estimate | Std. Error | z value | Pvalue | Odds Ratios | Lower Conf. | Upper Conf. |
|---|---|---|---|---|---|---|---|
| age_ps | 0.05721 | 0.02578 | 2.21956 | 0.02645 | 1.03429919 | 1.0073187162 | 1.1146117 |
| BA1 | -1.39716 | 0.42812 | -3.26346 | 0.00110 | 0.23679232 | 0.1048287398 | 0.5645902 |
| cpd_ps | -0.11507 | 0.05048 | -2.27949 | 0.02264 | 0.92141492 | 0.8069043096 | 0.9839622 |
| crv_total_pq1 | -0.07802 | 0.03590 | -2.17343 | 0.02975 | 0.99741471 | 0.8613717262 | 0.9918186 |
| inc2 | -1.03869 | 0.31014 | -3.34916 | 0.00081 | 0.49019013 | 0.1900225325 | 0.6429069 |
| inc4 | -0.97457 | 0.42630 | -2.28611 | 0.02225 | 0.52150065 | 0.1600182669 | 0.8547352 |
| mde_curr1 | -0.73720 | 0.29216 | -2.52326 | 0.01163 | 0.65973224 | 0.2678539287 | 0.8436322 |
| NHW1 | 1.75800 | 0.65575 | 2.68089 | 0.00734 | 5.82000635 | 1.6796287603 | 22.6375553 |
| Only.Menthol1 | 0.85021 | 0.33973 | 2.50260 | 0.01233 | 2.16913248 | 1.2069551958 | 4.5882301 |
| readiness9 | -3.51694 | 1.70093 | -2.06766 | 0.03867 | 0.03779048 | 0.0007221262 | 0.7141533 |
| Var1 | 2.23885 | 0.26203 | 8.54442 | 0.00000 | 5.68172908 | 5.7062888765 | 15.9763840 |

## Area Under Curve

**ROC Curve for Logistic Regression Model** plot compares the predictive performance of logistic regression models with main effects and interaction effects on both training and test datasets. The interaction effects model demonstrates slightly better performance than the main effects model, with higher AUC values for both training (0.881 vs. 0.837) and test datasets (0.674 vs. 0.659). However, the gap between training and test AUC values for both models highlights a decline in predictive accuracy when applied to unseen data, indicating potential overfitting, particularly for the interaction effects model. The training ROC curves for both models (solid blue for main effects and dotted blue for interaction effects) fit the data well, but the sharper drop-off in performance on the test curves (red dashed and red dot-dash) suggests that these models generalize less effectively. This disparity reinforces the need to balance model complexity, as adding interaction terms may capture relationships in the training data at the expense of broader applicability. While the interaction effects model slightly improves predictions, the main effects model may offer comparable performance with greater simplicity.




## Lasso Main Effects Model

## Interpretation of the Lasso Model

The Lasso regression model identified significant predictor variables influencing abstinence status with both main effects and odds ratios that provide meaningful insights into the relationships. Variables such as **Nicotine Metabolite Ratio (NMR)** (OR = 1.09), which reflects an individual's ability to metabolize nicotine, were associated with higher odds of abstinence, suggesting that individuals with faster nicotine metabolism might benefit more from pharmacological treatments. Similarly, **menthol smoking status (Only.Menthol1)** was associated with increased odds of abstinence (OR = 1.06), possibly due to behavioral or physiological characteristics specific to menthol smokers that make them more responsive to certain interventions. Sociodemographic factors, including **income levels** (e.g., inc2: OR = 0.91, inc5: OR =

1.00), had varied effects, indicating that financial stability and access to resources are critical moderators of treatment success. Moreover, **race/ethnicity (e.g., NHW1: OR = 1.20, Hisp1: OR = 1.03)**, reflecting the role of cultural, socioeconomic, and systemic disparities in shaping cessation outcomes.

Behavioral and psychological variables further underscore the complexity of smoking cessation interventions. **Readiness to quit** emerged as a significant predictor, with lower readiness levels (e.g., readiness4: OR = 0.74, readiness9: OR = 0.82) associated with reduced odds of abstinence, highlighting the need for motivational enhancement strategies for individuals in these stages. Variables such as **current depressive episodes (mde_curr1: OR = 0.95)** and **use of antidepressant medication (antidepmed1: OR = 1.03)** further emphasized the interaction of mental health with smoking cessation outcomes. Lastly, education levels (e.g., edu4: OR = 0.48) and cigarette consumption (cpd_ps: OR = 0.99) were inversely related to abstinence, suggesting that tailored interventions targeting heavy smokers and those with limited educational opportunities could be essential for improving cessation rates. These findings collectively underscore the need for individualized approaches that integrate sociodemographic, behavioral, and mental health factors for effective smoking cessation interventions.

| Variable | Coefficient | Odds_Ratio |
|---|---|---|
| Var1 | 0.1963241115 | 1.2169213 |
| BA1 | -0.0617443960 | 0.9401232 |
| age_ps | 0.0030070519 | 1.0030116 |
| sex_ps2 | -0.0437059135 | 0.9572354 |
| NHW1 | 0.1795789498 | 1.1967134 |
| Black1 | 0.0370573431 | 1.0377525 |
| Hisp1 | 0.0253552300 | 1.0256794 |
| inc2 | -0.0952689647 | 0.9091284 |
| inc3 | -0.0492222622 | 0.9519695 |
| inc4 | -0.0834304770 | 0.9199550 |
| edu2 | -0.6699831563 | 0.5117172 |
| edu3 | -0.7116073213 | 0.4908546 |
| edu4 | -0.7250695338 | 0.4842909 |
| edu5 | -0.6737517135 | 0.5097924 |
| ftcd_score | -0.0243890796 | 0.9759059 |
| ftcd.5.mins1 | 0.0365625459 | 1.0372392 |
| bdi_score_w00 | 0.0009785729 | 1.0009791 |
| cpd_ps | -0.0043346447 | 0.9956747 |
| crv_total_pq1 | -0.0005169362 | 0.9994832 |
| hedonsum_n_pq1 | -0.0003202853 | 0.9996798 |
| hedonsum_y_pq1 | 0.0006423462 | 1.0006426 |
| shaps_score_pq1 | -0.0035341547 | 0.9964721 |
| otherdiag1 | -0.0418391279 | 0.9590240 |
| antidepmed1 | 0.0282303314 | 1.0286326 |
| mde_curr1 | -0.0526112558 | 0.9487488 |
| NMR | 0.0938947450 | 1.0984441 |
| Only.Menthol1 | 0.0617780267 | 1.0637262 |
| readiness4 | -0.3005134288 | 0.7404380 |
| readiness5 | 0.0129826545 | 1.0130673 |
| readiness6 | -0.1041613249 | 0.9010799 |
| readiness7 | 0.0369034199 | 1.0375928 |
| readiness8 | -0.1276852819 | 0.8801303 |
| readiness9 | -0.1932869156 | 0.8242455 |
| readiness10 | -0.1384063057 | 0.8707448 |
| treatment_groupsST_VA | 0.0338250561 | 1.0344036 |

## Interaction Lasso Model

The Lasso regression analysis identified several potential moderators of abstinence, reflecting diverse demographic, behavioral, and clinical factors that influence cessation outcomes. Notably, interaction terms such as BA1:inc2 (OR = 1.14) and antidepmed1:treatment_groupsBA_VA (OR = 1.14) suggest that socioeconomic status and antidepressant use may significantly interact with behavioral activation and treatment assignment to impact abstinence rates. Behavioral readiness, measured through variables like readiness5 interacting with depression status (mde_curr1:readiness5, OR = 1.13) and treatment group (readiness5:treatment_groupsBA_VA, OR = 1.13), emerged as critical moderators, highlighting the importance of psychological preparedness in determining the effectiveness of both pharmacological and behavioral interventions. Additionally, demographic factors such as race/ethnicity (e.g., NHW1:edu4, OR = 1.13) and menthol smoking status (e.g., NHW1:Only.Menthol1, OR = 1.12) appear to play a role

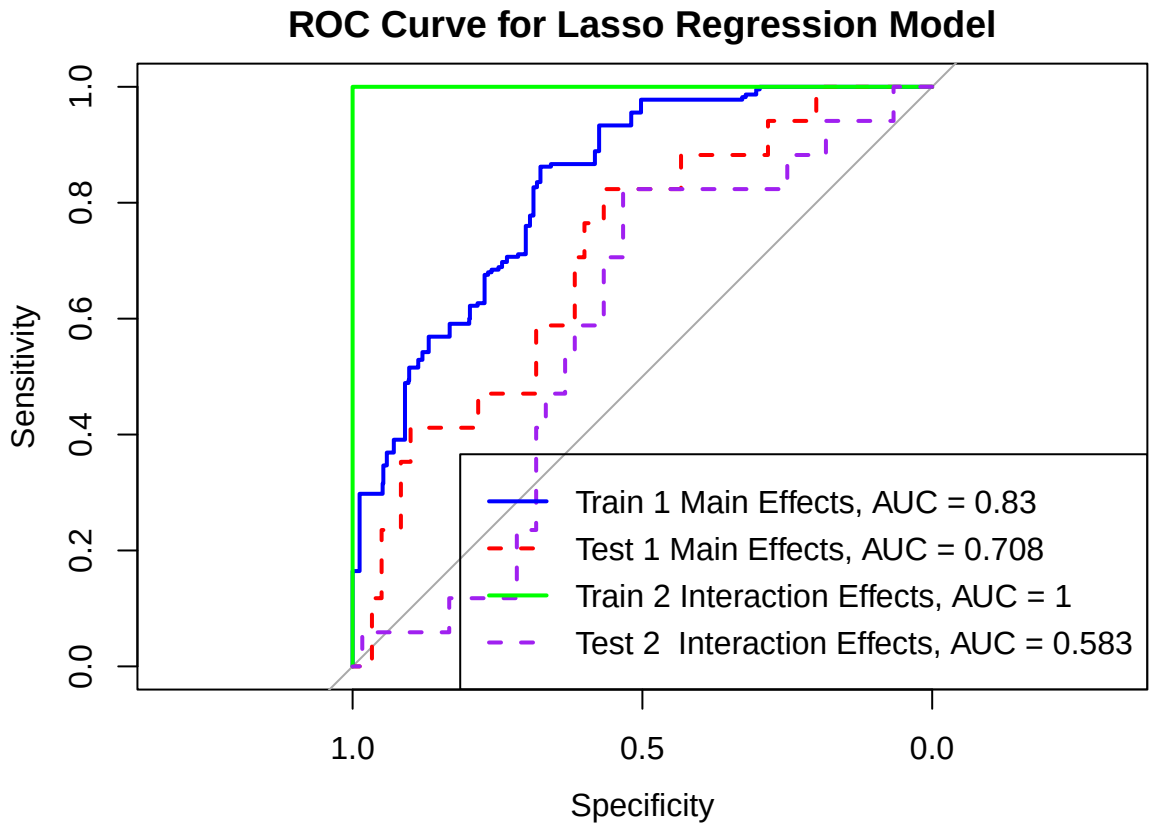
*Table is in Supplementary Materials*

## AUC CURVE for Lasso Models

The ROC curves provide a comparative evaluation of the Lasso regression model's predictive performance across different datasets and interaction effects. The Train 1 Main Effects curve (blue) shows strong predictive performance, with an AUC of 0.83, indicating a well-calibrated model for the training set. However, the corresponding Test 1 Main Effects curve (red, AUC = 0.708) demonstrates some degree of performance drop-off, suggesting potential overfitting to the training data or challenges in generalizing to the test set.

For the interaction effects, the Train 2 Interaction Effects curve (green, AUC = 1) indicates a perfect separation of classes in the training set, likely reflecting a high degree of model complexity or overfitting to the specific interaction terms. Conversely, the Test 2 Interaction Effects curve (purple, AUC = 0.583) shows diminished performance on the test set, suggesting that the interaction terms might not generalize well and could be overparameterized or poorly estimated.

Overall, while the main effects model demonstrates reasonable generalizability (train-test AUC difference of ~0.12), the interaction effects model likely introduces overfitting, as evidenced by the stark disparity in training and test performance.




### Comparative Analysis

In comparing logistic regression models to Lasso (Least Absolute Shrinkage and Selection Operator) models, we find distinct advantages in each approach regarding predictor selection and model performance. Logistic regression models, especially those with main and interaction effects, provide insights into significant predictors with interpretable coefficients, making it easier to understand the direct impact of variables on smoking cessation. However, logistic models can be sensitive to multicollinearity and may include non-contributing predictors.

Lasso models, on the other hand, are effective at handling multicollinearity and automatically selecting the most relevant predictors by applying a penalty to reduce less important coefficients to zero. This feature makes Lasso ideal for variable selection, often resulting in a more parsimonious model focused on predictors with the strongest associations to the outcome. If the Lasso model excludes non-contributing or weak predictors that logistic regression retained, this indicates Lasso's utility in identifying the most critical predictors for smoking cessation.

## Limitations

The study was conducted in a research clinic at Northwestern University (Chicago, Illinois) and University of Pennsylvania (Philadelphia, Pennsylvania). While the study mentions that the overal therapist competence was rated very good, but it may raise potential concerns because of the therapist level of education. The students from University of Pennsylvania has their Bachelor's degree while Northwestern University students has their Master's. This a potential limitation because the quality of education, knowledge, and expertise may be compromise which will influence results.

In logistic regression, models are sensitive to multicollinearity, which occurs when predictor variables are highly correlated. Introducing interaction terms to capture the combined effects of risk factors can amplify this issue, as seen in some of our results. The appearance of high multicollinearity in certain tables may indicate

overfitting, where the model captures noise instead of meaningful patterns. Overfitting not only complicates interpretation but also undermines the model's generalizability to new data. Addressing multicollinearity and considering techniques like regularization may help to mitigate these risks and improve the model's reliability.

For small sample sizes, logistic regression may outperform lasso because lasso can be prone to overfitting with limited data. Lasso's feature selection might end up removing too many predictors, reducing the model's ability to generalize well. In the case here, we have 300 participants in this dataset which is a indication of a small sample which causes the logistic to outperform lasso. logistic regression uses all available information, which can be an advantage when data is scarce.Lastly, predictors in logistic are highly correlated and need to be considered together which can impact the multicollinearity.

Logistic regression are sensitive to multicollimearity and most of the interact terms are highly correlated. This complicates things and causes overfitting.Lasso regressions uses a penalty term which implements more constraints. Because of it's constraints, we can observe the non-zero $\beta$ coefficents to and identify predictor variables for model selection.

One of the key limitations observed with the Lasso regression model is its susceptibility to **data leakage** and **overfitting**, particularly when the model complexity increases or interaction terms are introduced. Data leakage occurs when information from the test set or future data inadvertently influences the training process, leading to overly optimistic performance metrics during model development. This leakage can bias the estimated coefficients and inflate the apparent predictive accuracy of the model, as seen in the perfect separation (AUC = 1) for the **Train 2 Interaction Effects** curve. While this might indicate a highly predictive model in training, the corresponding test set performance (AUC = 0.583) reveals substantial degradation, underscoring the lack of generalizability and the potential for spurious relationships in the data.

Additionally, overfitting emerges as a prominent limitation of Lasso when the model captures noise or overly specific patterns in the training data. This is particularly evident in the stark discrepancy between the training and test set AUC values for the interaction effects model, suggesting that the additional complexity introduced by interaction terms does not translate to better real-world performance. Overfitting not only undermines the robustness of the model but can also bias the coefficient estimates, leading to misinterpretation of predictor importance and compromised inference. These issues emphasize the importance of rigorous validation procedures, careful feature selection, and mechanisms to mitigate data leakage, such as proper train-test splits, cross-validation, and ensuring that imputation or scaling processes are confined strictly to the training data. Without addressing these limitations, the reliability of Lasso model estimates can be significantly compromised.

# Conclusion

In conclusion, the logistic performed better according to the area under the curve models due to it's target estimates of AUC and not alot of data leakage in our models. In the lasso models, the AUC are more moderate compare.When predictors are highly correlated, Lasso may arbitrarily select one variable from a correlated set and set others to zero, potentially missing relevant information and leading to suboptimal performance. In the case here, t. Lasso remove those correlated variable and keep the one's that are more important as a nonzero. Logistic regression keeps all correlated variables, which can sometimes capture the overall signal better when correlations are essential to the model's interpretation. Factors like NMR, readiness to quit, race/ethnicity, menthol use, cigarettes per day, income level, and age are critical moderators that influence the effectiveness of treatments and the likelihood of abstinence.

# References

Hitsman B, Papandonatos GD, Gollan JK, Huffman MD, Niaura R, Mohr DC, Veluz-Wilkins AK, Lubitz SF, Hole A, Leone FT, Khan SS, Fox EN, Bauer AM, Wileyto EP, Bastian J, Schnoll RA. Efficacy and safety of combination behavioral activation for smoking cessation and varenicline for treating tobacco dependence among individuals with current or past major depressive disorder: A 2×2 factorial, randomized, placebo-controlled trial. Addiction. 2023 Sep;118(9):1710-1725. doi: 10.1111/add.16209. Epub 2023 May 3. Erratum in: Addiction. 2024 Sep;119(9):1669. doi: 10.1111/add.16609. PMID: 37069490.

# Supplementary Materials

| | | | |
|---|---|---|---|
| *10* | -4.144018e-02 | inc3 | 0.9594067 |
| *619* | 4.065996e-02 | shaps_score_pq1:readiness7 | 1.0414979 |
| *66* | -4.019463e-02 | Var1:Only.Menthol1 | 0.9606025 |
| *25* | -4.011894e-02 | otherdiag1 | 0.9606752 |
| *240* | 3.997340e-02 | Black1:readiness7 | 1.0407831 |
| *35* | -3.958669e-02 | readiness9 | 0.9611866 |
| *245* | 3.946074e-02 | Black1:treatment_groupsST_Placebo | 1.0402497 |
| *103* | -3.927609e-02 | BA1:readiness4 | 0.9614852 |
| *337* | 3.870995e-02 | inc5:mde_curr1 | 1.0394689 |
| *327* | -3.864365e-02 | inc3:otherdiag1 | 0.9620935 |
| *46* | -3.841813e-02 | Var1:inc2 | 0.9623105 |
| *11* | -3.821582e-02 | inc4 | 0.9625052 |
| *623* | -3.710541e-02 | shaps_score_pq1:treatment_groupsBA_VA | 0.9635746 |
| *205* | 3.703220e-02 | NHW1:readiness4 | 1.0377264 |
| *78* | -3.672510e-02 | BA1:sex_ps2 | 0.9639411 |
| *9* | -3.561019e-02 | inc2 | 0.9650164 |
| *52* | 3.537222e-02 | Var1:edu4 | 1.0360053 |
| *423* | -3.529737e-02 | edu3:antidepmed1 | 0.9653183 |
| *357* | 3.501472e-02 | inc5:readiness6 | 1.0356350 |
| *18* | 3.500110e-02 | ftcd.5.mins1 | 1.0356208 |
| *80* | 3.478548e-02 | BA1:Black1 | 1.0353976 |
| *43* | -3.469420e-02 | Var1:NHW1 | 0.9659007 |
| *514* | 3.451364e-02 | ftcd.5.mins1:readiness7 | 1.0351162 |
| *47* | -3.433152e-02 | Var1:inc3 | 0.9662511 |
| *300* | -3.236629e-02 | inc4:ftcd.5.mins1 | 0.9681519 |
| *50* | -3.225057e-02 | Var1:edu2 | 0.9682639 |
| *87* | -3.217310e-02 | BA1:edu3 | 0.9683389 |
| *392* | 3.119173e-02 | edu4:ftcd.5.mins1 | 1.0316833 |
| *626* | 3.117328e-02 | otherdiag1:antidepmed1 | 1.0316643 |
| *391* | 3.073254e-02 | edu3:ftcd.5.mins1 | 1.0312097 |
| *497* | -3.068088e-02 | ftcd_score:treatment_groupsBA_VA | 0.9697850 |
| *625* | -3.046779e-02 | shaps_score_pq1:treatment_groupsST_VA | 0.9699917 |
| *342* | -3.044256e-02 | inc2:Only.Menthol1 | 0.9700161 |
| *42* | -2.992877e-02 | Var1:sex_ps2 | 0.9705147 |
| *148* | 2.945693e-02 | sex_ps2:NHW1 | 1.0298951 |
| *31* | -2.898138e-02 | readiness5 | 0.9714346 |
| *346* | 2.890208e-02 | inc2:readiness4 | 1.0293238 |
| *171* | 2.804910e-02 | sex_ps2:Only.Menthol1 | 1.0284462 |
| *267* | 2.799559e-02 | Hisp1:Only.Menthol1 | 1.0283912 |
| *88* | -2.791671e-02 | BA1:edu4 | 0.9724694 |
| *489* | 2.761532e-02 | ftcd_score:Only.Menthol1 | 1.0280002 |
| *102* | -2.681021e-02 | BA1:Only.Menthol1 | 0.9735460 |
| *5* | -2.673049e-02 | sex_ps2 | 0.9736236 |
| *201* | -2.667583e-02 | NHW1:antidepmed1 | 0.9736768 |
| *105* | 2.666636e-02 | BA1:readiness6 | 1.0270251 |
| *466* | 2.592155e-02 | edu2:treatment_groupsBA_VA | 1.0262604 |
| *421* | -2.572611e-02 | edu5:otherdiag1 | 0.9746020 |
| *168* | 2.516614e-02 | sex_ps2:antidepmed1 | 1.0254855 |
| *12* | -2.484870e-02 | inc5 | 0.9754575 |
| *110* | 2.475563e-02 | BA1:treatment_groupsBA_VA | 1.0250646 |
| *76* | 2.434814e-02 | Var1:treatment_groupsST_VA | 1.0246470 |
| *439* | -2.424942e-02 | edu3:readiness4 | 0.9760422 |
| *55* | 2.379892e-02 | Var1:ftcd.5.mins1 | 1.0240844 |
| *415* | 2.358531e-02 | edu3:shaps_score_pq1 | 1.0238656 |
| *612* | 2.337514e-02 | shaps_score_pq1:antidepmed1 | 1.0236505 |
| *414* | 2.297496e-02 | edu2:shaps_score_pq1 | 1.0232409 |
| *186* | -2.296784e-02 | NHW1:inc4 | 0.9772939 |
| *634* | 2.285670e-02 | otherdiag1:readiness8 | 1.0231199 |
| *377* | -2.254351e-02 | inc5:treatment_groupsBA_VA | 0.9777087 |
| *61* | -2.250250e-02 | Var1:shaps_score_pq1 | 0.9777488 |
| *32* | 2.183303e-02 | readiness6 | 1.0220731 |
| *81* | 2.181754e-02 | BA1:Hisp1 | 1.0220573 |
| *72* | -2.174718e-02 | Var1:readiness9 | 0.9784876 |
| *97* | 2.168834e-02 | BA1:shaps_score_pq1 | 1.0219252 |
| *683* | -2.159195e-02 | Only.Menthol1:treatment_groupsBA_VA | 0.9786395 |
| *307* | -2.109819e-02 | inc3:cpd_ps | 0.9791228 |
| *323* | -2.067182e-02 | inc3:shaps_score_pq1 | 0.9795404 |
| *26* | 2.059626e-02 | antidepmed1 | 1.0208098 |
| *488* | 1.971221e-02 | ftcd_score:NMR | 1.0199078 |
| *324* | 1.883712e-02 | inc4:shaps_score_pq1 | 1.0190157 |
| *611* | -1.860867e-02 | shaps_score_pq1:otherdiag1 | 0.9815634 |
| *269* | 1.825807e-02 | Hisp1:readiness5 | 1.0184258 |
| *417* | -1.817693e-02 | edu5:shaps_score_pq1 | 0.9819873 |
| *486* | -1.803245e-02 | ftcd_score:antidepmed1 | 0.9821292 |
| *502* | 1.787130e-02 | ftcd.5.mins1:crv_total_pq1 | 1.0180319 |
| *70* | 1.628342e-02 | Var1:readiness7 | 1.0164167 |
| *478* | -1.613096e-02 | ftcd_score:ftcd.5.mins1 | 0.9839984 |
| *304* | 1.581552e-02 | inc4:bdi_score_w00 | 1.0159413 |
| *398* | 1.565681e-02 | edu2:cpd_ps | 1.0157800 |
| *192* | -1.552678e-02 | NHW1:ftcd_score | 0.9845931 |
| *309* | 1.549789e-02 | inc5:cpd_ps | 1.0156186 |
| *313* | 1.515760e-02 | inc5:crv_total_pq1 | 1.0152731 |
| *312* | -1.513291e-02 | inc4:crv_total_pq1 | 0.9849810 |
| *468* | -1.482428e-02 | edu4:treatment_groupsBA_VA | 0.9852851 |
| *427* | 1.449902e-02 | edu3:mde_curr1 | 1.0146046 |

## Code Appendix

```
knitr::opts_chunk$set(warning = FALSE,
                      message = FALSE,
                      echo = FALSE,
                      fig.align = "center")
library(readr)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(MASS)
library(tidyr)
library(kableExtra)
library(knitr)
library(GGally)
library(naniar)
library(visdat)
library(gtsummary)
library(gt)
library(mice)
library(corrplot)
library(reshape2)
library(ggwordcloud)
library(magick)
library(glmnet)
library(caret)
library(car)
library(broom)
library(gridExtra)
library(pROC)
library(broom)
 library(cowplot)



# Path to your image
fig_path2 <- "/Users/diahminhawkins/Documents/GitHub/Project2/SmokePicture.png"

# Load the image using magick
img2<- image_read(fig_path2)

# Convert image to raster for use in ggplot
img_raster2<- as.raster(img2)


# Example data
words <- c("Smoking Cessation", "Depression", "MDD", "Readiness", "FTCD",
           "Menthol", "Antidepressant", "Gender", "Carbon Monoxide", "BDI",
           "Duration", "Waking up Smoking", "Psychiatric Diagnosis", "Black", "Hispanic","Non-white His
         "Anhedonia","Complimentary Reinforcers", "Substitute Reinforcers",
         "Cigarette Reward", "Varenicline", "Behavioral Activations", "Pharmacotherapy","Psychotherapy"

frequencies <- c(1, 18, 1, 16, 14, 55, 5, 1, 4, 42, 3, 14, 14, 18, 16, 4,7,9,10,22, 55, 36, 40,55)
```

Table 1: Smoking Cessation Data Description

| Variables | Type | Description |
|---|---|---|
| id | Numeric | Participants id number |
| abst | Categorical | Smoking Abstinence |
| Var | Categorical | Varenicline (Pharmacotherapy) |
| BA | Categorical | Behavioral Activation (Pyschotherapy) |
| age_ps | Numeric | Age at phone interview |
| sex_ps | Categorical | Sex at interview |
| NHW | Categorical | Non-Hispanic White Indicator |
| Black | Categorical | Black Indicator |
| Hisp | Categorical | Hispanic Indicator |
| inc | Categorical | Income Levels (Low to High) |
| edu | Categorical | Education Levels (Low to High. |
| ftcd_score | Numeric | FTCD score at Baseline |
| ftcd.5.mins | Categorical | Smoking within 5 minutes of waking up |
| bdi_score_w00 | Numeric | BDI score at baseline |
| cpd_ps | Numeric | Cigarettes per day at baseline phone interview |
| crv_total_pq1 | Numeric | Cigarette reward value at baseline |
| hedonsum_n_pq1 | Numeric | Pleasurable events scale at baseline-substitute reinforcers |
| hedonsum_y_pq1 | Numeric | Pleasureable events scale at baseline-complementary reinforcers |

```r
new_frame <- data.frame(words, frequencies)

# Generate the word cloud on top of the image background
ggplot(new_frame, aes(label = words, size = frequencies)) +
  # Add the image background
  annotation_raster(img_raster2, xmin = -Inf, xmax = Inf, ymin = -Inf, ymax = Inf) +

  # Generate the word cloud
  geom_text_wordcloud(aes(color = frequencies)) +
  scale_size_area(max_size = 10) +

  # Customize the colors of the words
  scale_color_gradient(low = "white", high = "black") +

  # Remove axis titles and labels since we want the word cloud only
  theme_void()


#Load the data in
project2<- read_csv("project2.csv")


#Check for missing data
#project2%>% vis_dat()
#vis_miss(project2)
#project2%>% glimpse()
#Get all the missing data from each column
Missing_Data<- sapply(project2, function(x) sum(is.na(x)))
# Convert to dataframe
Missing_Data_df <- data.frame(ColumnName = names(Missing_Data), `Missing Data` = Missing_Data)

# Set names for the dataframe columns if necessary
names(Missing_Data_df) <- c("Variables", "Missing Data")

# Calculate the total number of rows in the dataset
total_rows <- nrow(project2)


#Create Missing Data Summary
missing_data_summary <-Missing_Data_df %>%
  filter(Variables %in% c('ftcd_score', 'inc', 'crv_total_pq1', 'shaps_score_pq1',
                          'NMR', 'Only.Menthol', 'readiness')) %>%
    mutate(Percent_Missing = (`Missing Data` / total_rows) * 100) %>%
  dplyr::select(Variables, `Missing Data`, Percent_Missing)%>%
  mutate(Variables = case_when(
   Variables == "ftcd_score" ~ "FTCD Score at Baseline",
     Variables == "inc" ~ "Income",
    Variables == "crv_total_pq1" ~ "Cigarette Reward Value at  Baseline",
   Variables == "shaps_score_pq1" ~ "Anhedonia",
    Variables == "NMR" ~ "Nicotine Metabolism Ratio",
    Variables == "Only.Menthol" ~ "Exclusive Mentholated Cigarette User",
    Variables == "readiness" ~ "Baseline Readiness to Quit Smoking"))
```

```r
# Obtain the total missing data and the percentage or missing data
total_missing <- sum(missing_data_summary$`Missing Data`)
percent_missing_total <- (total_missing / total_rows) * 100

#Create row to combine with the summart table
total_row <- data.frame(
  Variables = "Total",
  `Missing Data` = total_missing,
  Percent_Missing =percent_missing_total
)

#Both data frames have identical column names
names(total_row) <- names(missing_data_summary)

# Bind the summary table and the total row
missing_data_summary <- rbind(missing_data_summary, total_row)



# Convert to a gtsummary table
missing_data_summary %>%
  gt() %>%
  tab_header(
    title = "Missing Data Summary for Smoking Sessation"
  ) %>%
  cols_label(
    Variables = "Variables",
    `Missing Data` = "Missing Values",
    Percent_Missing = "Percentage Missing (%)"
  ) %>%
  fmt_number(
    columns = vars(Percent_Missing),
    decimals = 2  # Format percentage to two decimal places
  )



#Create variable table dataframe with description of the Marathon Dat
Variables_description<- data_frame(
  Variables= c("id", "abst", "Var", "BA", "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"),
  Type= c("Numeric", "Categorical", "Categorical", "Categorical","Numeric",
          "Categorical","Categorical", "Categorical", "Categorical","Categorical",
          "Categorical","Numeric","Categorical", "Numeric", "Numeric", "Numeric","Numeric","Numeric", "
          "Categorical", "Categorical","Categorical","Numeric","Categorical", "Numeric"),
  Description= c("Participants id number",
                 "Smoking Abstinence",
                 "Varenicline (Pharmacotherapy)",
                 "Behavioral Activation (Pyschotherapy)",
```

```
                  "Age at phone interview",
                  "Sex at interview",
                  "Non-Hispanic White Indicator",
                  "Black Indicator ",
                  "Hispanic Indicator",
                  "Income Levels (Low to High)",
                  "Education Levels (Low to High.",
                  "FTCD score at Baseline",
                  "Smoking within 5 minutes of waking up ",
                  "BDI score at baseline",
                  "Cigarettes per day at baseline phone interview",
                  "Cigarette reward value at baseline",
                  "Pleasurable events scale at baseline- substitute reinforcers",
                  "Pleasureable events scale at baseline- complementary reinforcers",
                  "Anhedonia",
                  "Other lifetime DSM-5 diagnosis",
                  "Taking antidepressant medication at baseline",
                  "Current vs. Past MDD",
                  "Nicotine  Metabolism Cigarette User",
                  "Exclsuive Mentholated Cigarette User",
                  "Baseline readiness to quit smoking"
                  )
)



# Create the table with kable and customize with kableExtra
table_summary<- kable(Variables_description, "latex", booktabs = TRUE, caption = "Smoking Cessation Data
  kable_styling(latex_options = c("striped", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "8cm")







# Load the PNG file
table_image <- image_read("adjustedtable2.png")



# Display the image in RStudio's Viewer or an external window
table_image


# define treatment categories
project2_treatments <- project2 %>%
  mutate(
    treatment_groups = case_when(
      Var == 1 & BA == 1 ~ "BASC + Varenicline",
```

```
      Var == 0 & BA == 1 ~ "BASC + Placebo",
      Var == 1 & BA == 0 ~ "ST + Varenicline",
      Var == 0 & BA == 0 ~ "ST + Placebo"
    )
  )

# Combine education levels
project2_treatments <- project2_treatments %>%
   mutate(edu = as.character(edu))%>%
  mutate(edu = case_when(
    edu %in% c( "2", "3") ~ "1",
    TRUE ~ edu  # Keep other values as they are
  ))

# recode income, education, and sex levels
project2_treatments <- project2_treatments  %>%
  mutate(
    inc= case_when(
      inc == 1 ~ "Less than $20,000",
      inc == 2 ~ "$20,000-35,000",
      inc == 3 ~ "$35,001-50,000",
      inc == 4 ~ "$50,001-75,000",
      inc == 5 ~ "More than $75,000",
      TRUE ~ "Unknown"
    ),
    edu = case_when(
      edu == 1 ~ "Some High School or Less",
      edu == 4 ~ "Some college/technical school",
      edu == 5 ~ "College graduate",
      TRUE ~ "Unknown"
    ),
    sex_ps = case_when(
      sex_ps == 1 ~ "Male",
      sex_ps == 2 ~ "Female",
      TRUE ~ "Unknown")
  )




#Define Treatment Groups
project2_table <- project2 %>%
  mutate(treatment_groups = case_when(
    Var == 1 & BA == 1 ~ "BA_VA",
    Var == 0 & BA == 0 ~ "ST_Placebo",
    Var == 0 & BA == 1 ~ "BA_Placebo",
    Var == 1 & BA == 0 ~ "ST_VA"
  ))
```

```r
#Change variables into factor and continous variables
factor_vars <- c("abst","Var","BA","sex_ps", "NHW",
                 "Black", "Hisp", "inc", "edu","readiness",
                 "ftcd.5.mins","otherdiag", "antidepmed","mde_curr",
                 "Only.Menthol", "treatment_groups")

#Mutate variables as factors
project2_table<- project2_table%>%
  mutate(across(all_of(factor_vars), as.factor))

#Mutate variables as factors
project2_treatments<- project2_treatments%>%
  mutate(across(all_of(factor_vars), as.factor))


# Create a stacked bar plot
income_plot<-ggplot(project2_treatments, aes(x = abst, fill = inc)) +
  geom_bar(position = "fill") + # Use position = "fill" for proportions
  labs(
    x = "Abstinence",
    y = "Proportions",
    fill = "Income Levels"
  ) +
  scale_fill_brewer(palette = "Set4") + # Apply a Brewer palette
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12),
    legend.position = "bottom"
  )

# Create a stacked bar plot
education_plot<-ggplot(project2_treatments, aes(x = abst, fill = edu)) +
  geom_bar(position = "fill") + # Use position = "fill" for proportions
  labs(
    x = "Abstinence",
    y = "Proportions",
    fill = "Education Levels"
  ) +
  scale_fill_brewer(palette = "Set4") + # Apply a Brewer palette
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12),
    legend.position = "bottom"
  )
```

```r
# Combine plots with cowplot
demo_plot <- plot_grid(
  education_plot, income_plot,
  ncol = 2,  # Two columns
  align = "hv",  # Align both horizontally and vertically
  label_size = 14  # Increase label size


)

demo_plot



# Hedonsum Complementary Rewards
hedonsum_y_pq_plot <- ggplot(project2_table, aes(x = hedonsum_y_pq1, fill = abst)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~treatment_groups, scales = "free") +
  theme_minimal() +
  labs(
    x = "Substitute Reinforcers",
    y = "Density"
  ) +
  scale_fill_manual(values = c("0" = "pink", "1" = "blue"), name = "abst")

 # Nicotine Metabolism Ratio
nmr_plot <- ggplot(project2_table, aes(x = NMR, fill = abst)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~treatment_groups, scales = "free") +
  theme_classic() +
  labs(
    x = "NMR",
    y = "Density"
  ) +
  scale_fill_manual(values = c("0" = "pink", "1" = "blue"), name = "abst")


# Hedonsum Substitute Reinforcers
hedonsum_N_plot<-ggplot(project2_table, aes(x = hedonsum_n_pq1 , fill = abst)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~treatment_groups, scales = "free")+
  theme_classic()+
  labs(
       x = "Hedonsum",
       y = "Density")+
  scale_fill_manual(values = c("0" = "pink", "1" = "blue"), name = "abst")


age_plot<-ggplot(project2_table, aes(x =age_ps , fill = abst)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~treatment_groups, scales = "free")+
  theme_classic()+
  labs(
```

```
       x = "Age",
       y = "Density")+
  scale_fill_manual(values = c("0" = "pink", "1" = "blue"), name = "abst")
```

```
anhedonia_plot <- ggplot(project2_table, aes(x = shaps_score_pq1, fill = abst)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~treatment_groups, scales = "free") +
  theme_classic() +
  labs(
    title = "Baseline Characteristics of Abstinence Status with Continuous Variables",
    x = "Anhedonia",
    y = "Density"
  ) +
  scale_fill_manual(values = c("1" = "pink", "0" = "blue"), name = "abst")

anhedonia_plot
```

```
grid.arrange(
  hedonsum_y_pq_plot, nmr_plot, hedonsum_N_plot, age_plot,
  ncol = 2,
  nrow = 2,
  top= "Baseline Characteristics of Abstinence Status with Continous Variables"
)
```

```
# Create a stacked bar plot
NHW_plot<-ggplot(project2_table, aes(x = abst, fill = NHW)) +
  geom_bar(position = "fill") + # Use position = "fill" for proportions
  facet_wrap(~treatment_groups, scales = "free")+
  labs(
    x = "Abstinence",
    y = "Proportions",
    fill = "Non-Hispanic Whites Levels"
  ) +
  scale_fill_brewer(palette = "Set4") + # Apply a Brewer palette
  theme_minimal() +
  theme(

    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12),
    legend.position = "bottom"
```

```
  )


# Create a stacked bar plot
Hisp_plot<-ggplot(project2_table, aes(x = abst, fill = Hisp)) +
  geom_bar(position = "fill") + # Use position = "fill" for proportions
  facet_wrap(~treatment_groups, scales = "free")+
  labs(title = "Baseline Characteristics of Abstinence Status in Racial Groups",
    x = "Abstinence",
    y = "Proportions",
    fill = "Hispanic Whites "
  ) +
  scale_fill_brewer(palette = "Set4") + # Apply a Brewer palette
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12),
    legend.position = "bottom"
  )


# Create a stacked bar plot
Black_plot<-ggplot(project2_table, aes(x = abst, fill = Black)) +
  geom_bar(position = "fill") + # Use position = "fill" for proportions
  facet_wrap(~treatment_groups, scales = "free")+
  labs(
    x = "Abstinence",
    y = "Proportions",
    fill = "Blacks"
  ) +
  scale_fill_brewer(palette = "Set4") + # Apply a Brewer palette
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12),
    legend.position = "bottom"
  )



 Hisp_plot



# Combine plots with cowplot
combined_plot <- plot_grid(
  Black_plot, NHW_plot,

  ncol = 2,  # Two columns
  align = "hv",  # Align both horizontally and vertically
  label_size = 14  # Increase label size
```

```
)

combined_plot
# Find the variables in project2 that are not in factor_vars
continous_vars <- setdiff(names(project2), factor_vars)

project2 <- project2 %>%
  mutate(across(all_of(continous_vars), as.numeric))

# Select only numeric columns for the correlation plot
numeric_data <- project2 %>% select(all_of(continous_vars))%>%
dplyr::select(-c(id))


# Calculate the correlation matrix
cor_matrix <- cor(numeric_data, use = "complete.obs")


# Melt the correlation matrix for ggplot2
cor_data2 <- melt(cor_matrix)


#Cessation Smoking correlation plot
cessation_smoking_plot2<-ggplot(data = cor_data2, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(value, 2)), color = "black", size = 4) +
  scale_fill_gradient2(low = "hotpink", high = "royalblue", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  labs(title = "Correlation of Smoking Cessation",
       x= "Variables",
       y= "Variables") +
  theme_minimal(base_family = "Times") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        legend.position = "bottom",
        plot.title = element_text(hjust = 0.5, size= 20))

cessation_smoking_plot2




set.seed(222)

#project2_table<- project2_table%>%
  #dplyr::select(-id,)

trainIndex <- createDataPartition(project2_table$abst, p = 0.7, list = FALSE)
  train_data <- project2_table[trainIndex, ]
  test_data <- project2_table[-trainIndex, ]
```

```r
# Data Imputation for Logistic Modeling

imputed_data <- mice(train_data, m = 5, method = 'pmm', maxit = 10, seed = 222, print=FALSE)
stacked_data_primary <- complete(imputed_data, action = "long", include = FALSE)
stacked_data_primary<-stacked_data_primary%>%
  dplyr::select(-.id, -.imp)

# Fit the logistic regression model with
main_effects <- glm(abst ~ ., family = binomial(link = "logit"), data = stacked_data_primary)

# Fit the logistic model with interaction terms on Training Data
  interaction_effects <- glm(abst ~ Var + BA + 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 +
                               sex_ps * Black + sex_ps * NHW + Black * Only.Menthol+ Hisp *
                               Only.Menthol +NHW * edu  +  Black * edu+ ftcd_score*age_ps + ftcd_score * c
                                 BA*NMR,
                               family =binomial(link = "logit"), data=stacked_data_primary)


#Make table for Main Effects
data_frame2<- round(summary(main_effects)$coefficients,5)

# Make into dataframs
data_frame2<-as.data.frame(data_frame2)

# Convert row names to a column in the dataframe
data_frame2$Variable <- rownames(data_frame2)

# Rearrange the columns to make "Variable" the first column
data_frame2 <- data_frame2[, c("Variable", colnames(data_frame2)[-ncol(data_frame2)])]

# Compute odds ratios
odds_ratios <- exp(coef(main_effects, na.rm=TRUE))

# Make odd ratios into dataframe
yea<-as.data.frame(odds_ratios)

# Create row names for odd ratios
yea$Variable<- rownames(yea)

# Rearrange the columns to make "Variable" the first column
yea <- yea[, c("Variable", colnames(yea)[-ncol(yea)])]

#Removes NAs in odd ratios
yea<- yea[!is.na(yea$odds_ratios), ]


# Compute confidence intervals for the coefficients
conf_intervals <- exp(confint(main_effects))

# Make confidence intervals into a dataframe
```

```r
conf_intervals_<-as.data.frame(conf_intervals)

# A
conf_intervals_$Variable<- rownames(conf_intervals_)
# Rearrange the columns to make "Variable" the first column
 conf_intervals_<- conf_intervals_[, c("Variable", colnames(conf_intervals_)[-ncol(conf_intervals_)])]
 # Remove rows where treatment_groups are "ST_Placebo" or "ST_VA"
 conf_intervals_ <-  conf_intervals_[!(conf_intervals_$Variable %in% c("treatment_groupsST_Placebo", "t


combined_df <- merge(merge(data_frame2, yea, by = "Variable"), conf_intervals_, by = "Variable")

combined_df<-combined_df%>%
  group_by(Variable)%>%
  filter(`Pr(>|z|)` < 0.05) %>%
  rename(`Lower Conf.` = `2.5 %`,
         `Upper Conf.`= `97.5 %`,
         `Pvalue`= `Pr(>|z|)`,
    `Odds Ratios` = `odds_ratios`)
combined_df<- combined_df[!(combined_df$Variable %in% c("id")), ]

# Format numerical columns for LaTeX (e.g., Odds Ratios, Confidence Intervals)

table2<-combined_df %>%
  kbl(caption = "Logistic Model:Significant Predictor Variables for Main Effects Data",
      booktabs = TRUE, escape = FALSE, align = "c") %>%
  kable_styling(full_width = FALSE, latex_options = c('hold_position'))









# Path to your image
fig_path4<- "~/Documents/GitHub/PDA Final Project Portfolio/table2.png"

# Load the image using magick
img4<- image_read(fig_path4)

rotated_image <- image_rotate(img4, 90)
rotated_image







#Make table for Main Effects
data_frame3<- round(summary(interaction_effects)$coefficients,5)

# Make into dataframs
```

```r
data_frame3<-as.data.frame(data_frame3)

# Convert row names to a column in the dataframe
data_frame3$Variable <- rownames(data_frame3)

# Rearrange the columns to make "Variable" the first column
data_frame3<- data_frame3[, c("Variable", colnames(data_frame3)[-ncol(data_frame3)])]

# Compute odds ratios
odds_ratios2 <- exp(coef(interaction_effects, na.rm=TRUE))

# Make odd ratios into dataframe
yea1<-as.data.frame(odds_ratios2)

# Create row names for odd ratios
yea1$Variable<- rownames(yea1)

# Rearrange the columns to make "Variable" the first column
yea1 <- yea1[, c("Variable", colnames(yea1)[-ncol(yea1)])]

#Removes NAs in odd ratios
yea1<- yea1[!is.na(yea1$odds_ratios), ]


# Compute confidence intervals for the coefficients
conf_intervals2<- exp(confint(interaction_effects))

# Make confidence intervals into a dataframe
conf_intervals_2<-as.data.frame(conf_intervals2)


conf_intervals_2$Variable<- rownames(conf_intervals_2)
# Rearrange the columns to make "Variable" the first column
 conf_intervals_2<- conf_intervals_2[, c("Variable", colnames(conf_intervals_2)[-ncol(conf_intervals_2)]
 # Remove rows where treatment_groups are "ST_Placebo" or "ST_VA"
 conf_intervals_2 <-  conf_intervals_2[!(conf_intervals_2$Variable %in% c("NHW1:edu5", "Black1:edu2", "

combined_df2<- merge(merge(data_frame3, yea, by = "Variable"), conf_intervals_2, by = "Variable")

combined_df2 <- combined_df2 %>%
  group_by(Variable) %>%
  filter(`Pr(>|z|)` < 0.05) %>%
  rename(`Lower Conf.` = `2.5 %`,
         `Upper Conf.`= `97.5 %`,
         `Pvalue`= `Pr(>|z|)`,
          `Odds Ratios` = `odds_ratios`  # Rename column for readability
  )



combined_df2<- combined_df2[!(combined_df2$Variable %in% c("id")), ]
```

```r
table0<-combined_df2 %>%
  kbl(caption = "Logistic Model:Significant Predictor Variables for Training Data with Interaction Terms
      booktabs = TRUE, escape = FALSE, align = "c") %>%
  kable_styling(full_width = FALSE, latex_options = c('hold_position'))



# Path to your image
fig_path3 <- "~/Documents/GitHub/PDA Final Project Portfolio/Table0.png"

# Load the image using magick
img3<- image_read(fig_path3)

rotated_image2 <- image_rotate(img3, 90)

rotated_image2

# Predict probabilities on test data
predicted_probs <- predict(main_effects, newdata = test_data, type = "response")
predict_probs_2<- predict(main_effects, newdata = stacked_data_primary, type = "response")
predicted_probs3 <- predict(interaction_effects, newdata = test_data, type = "response")
predict_probs_4<- predict(interaction_effects, newdata = stacked_data_primary, type = "response")



# Calculate ROC
roc_curve <- roc(test_data$abst, predicted_probs)
roc_curve2<- roc(stacked_data_primary$abst, predict_probs_2)
roc_curve3 <- roc(test_data$abst, predicted_probs3)
roc_curve4<- roc(stacked_data_primary$abst, predict_probs_4)


#Calculate AUC
auc_value <- auc(roc_curve)
auc_value2<-auc(roc_curve2)
auc_value3 <- auc(roc_curve3)
auc_value4<-auc(roc_curve4)




plot(roc_curve, col = "red", lty = 2, main = "ROC Curve for Logistic Regression Model", lwd = 2)
lines(roc_curve2, col = "blue", lty = 1, lwd = 2)  # Main Effects Training
lines(roc_curve3, col = "red", lty = 4, lwd = 2)  # Interaction Effects Test
lines(roc_curve4, col = "blue", lty = 3, lwd = 2)  # Interaction Effects Training

# Add a legend
legend("bottomright",
       legend = c(
         paste("Test (Main Effects), AUC =", round(auc_value, 3)),
         paste("Train (Main Effects), AUC =", round(auc_value2, 3)),
         paste("Test (Interaction Effects), AUC =", round(auc_value3, 3)),
         paste("Train (Interaction Effects), AUC =", round(auc_value4, 3))
```

```
       ),
       col = c("red", "blue", "red", "blue"),
       lty = c(2, 1, 4, 3),
       lwd = 2)




# Define the formula for Lasso
formula <- abst ~ .  # This includes all variables in the dataset for prediction

# Set up parameters
set.seed(222)

stacked_data_primary<-stacked_data_primary%>%
  dplyr::select(-id)

test_data<-test_data%>%
   dplyr::select(-id)
stacked_data_primary <- stacked_data_primary[complete.cases(stacked_data_primary), ]
test_data <- test_data[complete.cases(test_data), ]



# Model matrix for training and test sets
  X_train <- model.matrix(formula, data = stacked_data_primary)[, -1]  # Remove intercept column
  Y_train <- stacked_data_primary$abst
  X_test <- model.matrix(formula, data = test_data)[, -1]  # Remove intercept column
  Y_test <- test_data$abst
  BA_position <- which(colnames(X_train) == "BA")
  penalty_factors <- rep(1, ncol(X_train))
  penalty_factors[BA_position] <- 0

  Y_train <- as.numeric(as.character(stacked_data_primary$abst))
  Y_test <- as.numeric(as.character(test_data$abst))

cv_lasso <- cv.glmnet(X_train, Y_train, alpha = 1, penalty.factor = penalty_factors, family= "binomial"
  # Use lambda.min for final model
best_lambda <- cv_lasso$lambda.min

 lasso_model <- glmnet(X_train, Y_train, alpha = 1, penalty.factor = penalty_factors,lambda = best_lambd


  lasso_train_predictions <- predict(lasso_model, newx = X_train, s=best_lambda,type = "response")
  lasso_test_predictions <- predict(lasso_model, newx = X_test,  s=best_lambda, type = "response")

  # Calculate ROC and AUC for training and test data
  roc_train <- roc(Y_train, lasso_train_predictions)
  auc_train <- auc(roc_train)


  roc_test <- roc(Y_test,as.numeric(lasso_test_predictions))
```

```
  auc_test <- auc(roc_test)




# Extract coefficients for the best lambda
lasso_coefficients <- coef(lasso_model, s = best_lambda)

# Convert to a readable data frame
lasso_coefficients_df <- as.data.frame(as.matrix(lasso_coefficients))
colnames(lasso_coefficients_df) <- "Coefficient"
lasso_coefficients_df$Variable <- rownames(lasso_coefficients_df)
rownames(lasso_coefficients_df) <- NULL

# Calculate Odds Ratios (OR)
lasso_coefficients_df$Odds_Ratio <- exp(lasso_coefficients_df$Coefficient)


# Reorder columns: Variable first, Coefficient second, then Odds_Ratio
lasso_coefficients_df <- lasso_coefficients_df[, c("Variable", "Coefficient", "Odds_Ratio")]

# Filter for nonzero coefficients
lasso_coefficients_df <- lasso_coefficients_df[lasso_coefficients_df$Coefficient != 0, ]


lasso_coefficients_df <- lasso_coefficients_df %>%
  filter(!(Variable %in% c("(Intercept)", "id")))

kable_lasso_table3<-lasso_coefficients_df %>%
  kbl(caption = "Lasso Model:Significant Predictor Variables for with Main Effects",
      booktabs = TRUE, escape = FALSE, align = "c") %>%
  kable_styling(full_width = FALSE, latex_options = c('hold_position'))



# Path to your image
fig_path_lasso <- "~/Documents/GitHub/PDA Final Project Portfolio/kable_lasso_table_3.png"

# Load the image using magick
img_lasso<- image_read(fig_path_lasso)

img_lasso

# Define the formula for Lasso
formula2 <- abst ~ .^2  # This includes all variables in the dataset for prediction

# Set up parameters
set.seed(222)


stacked_data_primary <- stacked_data_primary[complete.cases(stacked_data_primary), ]
```

```r
test_data <- test_data[complete.cases(test_data), ]
# Model matrix for training and test sets
  X_train2 <- model.matrix(formula2, data = stacked_data_primary)[, -1]  # Remove intercept column
  Y_train2 <- stacked_data_primary$abst
  X_test2 <- model.matrix(formula2, data = test_data)[, -1]  # Remove intercept column
  Y_test2 <- test_data$abst
  BA_position2 <- which(colnames(X_train2) == "BA")
  penalty_factors2 <- rep(1, ncol(X_train2))
  penalty_factors2[BA_position2] <- 0

  Y_train2 <- as.numeric(as.character(stacked_data_primary$abst))
  Y_test2 <- as.numeric(as.character(test_data$abst))

cv_lasso2 <- cv.glmnet(X_train2, Y_train2, alpha = 1, penalty.factor = penalty_factors2, family= "binom
  # Use lambda.min for final model
best_lambda2 <- cv_lasso2$lambda.min



 lasso_model2 <- glmnet(X_train2, Y_train2, alpha = 1, penalty.factor = penalty_factors2,lambda = best_


  lasso_train_predictions2 <- predict(lasso_model2, newx = X_train2, s=best_lambda2,type = "response")


  lasso_test_predictions2 <- predict(lasso_model2, newx = X_test2,  s=best_lambda2, type = "response")

  # Calculate ROC and AUC for training and test data
  roc_train2 <- roc(Y_train2, lasso_train_predictions2)
  auc_train2 <- auc(roc_train2)


  roc_test2 <- roc(Y_test2,as.numeric(lasso_test_predictions2))
  auc_test2 <- auc(roc_test2)



#stacked_data_primary <- stacked_data_primary[complete.cases(stacked_data_primary), ]
#test_data <- test_data[complete.cases(test_data), ]

#length(test_data)
#dim(stacked_data_primary)

#head(stacked_data_primary)
#names(stacked_data_primary)
#stacked_data_primary<-stacked_data_primary_%>%
 # select(-id)


# Extract non-zero coefficients from the Lasso model
lasso_coefficients2 <- coef(lasso_model2, s = best_lambda2)

# Convert to a data frame
```

```r
lasso_summary <- as.data.frame(as.matrix(lasso_coefficients2))
colnames(lasso_summary) <- "Coefficient"
lasso_summary$Variable <- rownames(lasso_summary)
rownames(lasso_summary) <- NULL

# Calculate odds ratios (OR)
lasso_summary$Odds_Ratio <- exp(lasso_summary$Coefficient)

# Filter for non-zero coefficients
lasso_summary_filtered <- lasso_summary[lasso_summary$Coefficient != 0 & lasso_summary$Variable != "(In

# Sort by the absolute value of the coefficient
lasso_summary_filtered <- lasso_summary_filtered[order(-abs(lasso_summary_filtered$Coefficient)), ]

# Create a summary statistics table with kable


kable_lasso_summary <- lasso_summary_filtered %>%
  kable(caption = "Summary Statistics: Lasso Model (Non-Zero Coefficients and Odds Ratios)",
        col.names = c("Variable", "Coefficient", "Odds Ratio")) %>%
  kable_styling(full_width = FALSE, font_size = 12)









# Calculate ROC and AUC for training and test data
roc_train <- roc(Y_train, lasso_train_predictions)
auc_train <- auc(roc_train)

roc_train2 <- roc(Y_train2, lasso_train_predictions2)
auc_train2 <- auc(roc_train2)

roc_test <- roc(Y_test, as.numeric(lasso_test_predictions))
auc_test <- auc(roc_test)

roc_test2 <- roc(Y_test2, as.numeric(lasso_test_predictions2))
auc_test2 <- auc(roc_test2)

# Plot ROC curve for the first training data
plot(roc_train, col = "blue", lty = 1, main = "ROC Curve for Lasso Regression Model", lwd = 2, xlim = c

# Add ROC curves for test data and second training/test sets
lines(roc_test, col = "red", lty = 2, lwd = 2)
lines(roc_train2, col = "green", lty = 1, lwd = 2)
lines(roc_test2, col = "purple", lty = 2, lwd = 2)

# Add a legend
legend(
  "bottomright",
```

```
  legend = c(
    paste("Train 1 Main Effects, AUC =", round(auc_train, 3)),
    paste("Test 1 Main Effects, AUC =", round(auc_test, 3)),
    paste("Train 2 Interaction Effects, AUC =", round(auc_train2, 3)),
    paste("Test 2  Interaction Effects, AUC =", round(auc_test2, 3))
  ),
  col = c("blue", "red", "green", "purple"),
  lty = c(1, 2, 1, 2),
  lwd = 2
)



#kable_lasso_summary
# Path to your image

fig_path5 <- "~/Documents/GitHub/PDA Final Project Portfolio/table_lasso_interractions.png"

# Load the image using magick
img5<- image_read(fig_path5)

img5
table_summary
```