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

Diahmin Hawkins

11/6/2024

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). Dr. George Papandonatos, from Brown University's highly regarded Biostatistics Department (ranked #14 in the country by U.S. News & World Report's Best Graduate Schools), has investigated smoking cessation outcomes in adults diagnosed with MDD.

This study was conducted in research clinics at Northwestern University (Chicago, IL) and the University of Pennsylvania (Philadelphia, PA). Findings from this research indicate that individuals with MDD tend to smoke more heavily, exhibit greater nicotine dependence, and endure more severe withdrawal symptoms than individuals without MDD. According to `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`, "In conclusion, we found strong evidence that varenicline improved both short- and long-term abstinence rates relative to placebo among racially and socio-economically diverse adults with varied motivation to quit and varied 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.

The prior 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.

This current study 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 will significantly interact with treatment type, influencing the likelihood of smoking cessation. interact with treatment type, influencing the likelihood of smoking cessation.

Methods

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

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 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) In order to obtain the these treatments, we mutated the treatment groups to to conform to the paper's treatment groups. In the case here, we are are able to to able to obtain treatment groups by `Var == 1 & BA == 1 ~ "BA_VA"`, `Var == 0 & BA == 0 ~ "ST_Placebo"`, `Var == 0 & BA == 1 ~ "BA_Placebo"`, `Var == 1 & BA == 0 ~ "ST_VA"` to equate to `BASC+ Varenicline`,`ST+Placebo`, `BASC + Placebo`, and `ST + Varenicline`.

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 imputate 5 different datasets 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 analyzed independently using the same statistical model, 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 pool Rubin's Rules to find the best variables for model selection.

EDA (Exploratory Data Analysis)

To explore the relationships of the categorical variables, a `Chi-Squared` test is used to determine whether there is a significant association . Through this process, it compares the observed frequencies in different

categories with the expected frequencies to see if any differences are likely due to chance.This test assesses whether two categorical variables are independent of each other, identify patterns, relationships, and potential associations of the risk factors and the outcome of abstinence.

After exploring the categorical relationship, we will observe the correlations of all the variables. Based on this plot, we we will observe potential risk factors to explore in our logistic model.

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 based on the given predictors by maximizing the likelihood (or minimizing the negative log-likelihood).In our logistic model, we will implement a test, train, split process with 70% on the trained data and 30% on the test data. After filtering the statistical significant variables using $\alpha <= .05$, we will observe the variables that are left after pooling the results.

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 β 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 β coefficients, we observe the non-zero coeffiecients in both our main effects model and interaction term model.In our lass model, we will implement a test, train, split process with 70% on the trained data and 30% on the test data.Following this process is crossvalidation and using the respective lasso mechanics. We will pool the results using `Rubin's Rules` find the standard errors, means, 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 will compare the models based on performance metrics AUC (Area Under the Curve) both training and testing datasets. These metrics will help evaluate how well each model generalizes to unseen data and its ability to distinguish between abstinent and non-abstinent outcomes.The Area Under the Curve (AUC) is a metric used to evaluate the performance of binary classification models. Specifically, it refers to the area under the Receiver Operating Characteristic (ROC) Curve, which plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold levels. The AUC provides an aggregate measure of a model's ability to distinguish between positive and negative classes across all possible classification thresholds. An AUC value of 1 indicates a perfect model that can correctly classify all cases, while an AUC of 0.5 suggests the model performs no better than random guessing. Higher AUC values reflect stronger model performance and greater discriminatory power. AUC is particularly useful because it is independent of any specific threshold, offering a more comprehensive understanding of the model's effectiveness in distinguishing between classes.

Table 1: Demographics Table of the Smoking Cessation Participants

Demograhic Characteristics Variable	BA+Placebo	BA+VA	ST+Placebo	ST+VA
N	68	83	68	81
Mean Ages	50.7	50.3	50.3	48.7
Standard Deviation Ages	13.5	13.2	10.8	12.7
Sex				

Demograhic Characteristics Variable	BA+Placebo	BA+VA	ST+Placebo	ST+VA
Blacks	37	37	40	43
Hispanics	5	4	4	5
Black Percentage	54.4	44.6	58.8	53.1
Hispanic Percentage	7.4	4.8	5.9	6.2
Non-Hispanic Whites	24	34	22	25
Non-Hispanic White Percentage	35.3	41.0	32.4	30.9
Education Levels				
Grade School	1	0	0	0
Some High School	3	7	2	4
High School Graduate or GED	23	15	11	27
Some College	22	32	38	24
College Graduate	19	29	17	26
Income Levels				
Less than $20,000	25	30	26	29
$20,000–35,000	16	17	14	21
$35,001–50,000	8	13	14	11
$50,001–75,000	12	12	8	6
More than $75,000	6	10	6	13
Unknown	1	1	0	1
Major Depressive Disorder				
Cigarette Type				
Menthol Only	40	48	43	47
Antidepressant Medication (%)	41.2	28.9	22.1	18.5
Other Lifetime Diagnosis	35	30	28	40
Smoking				
Cigarettes Per Day at Baseline	15.6	15.5	15.0	14.4
Cigarette Reward Value at Baseline	7.4	7.2	7.0	7.1
FTCD at Baseline	5.31	5.07	5.39	5.17
Readiness to Quit	6.8	6.7	7.0	6.7
Time to smoking upon waking up				
Smoking 5 minutes into waking up (%)	47.1	39.8	51.5	46.9
Pleasurable Events at Baseline				
Substitute Reinforcers	23.2	22.9	20.8	23.4
Complimentary Reinforcers	27.7	22.4	27.4	25.0
0	36	43	37	37
1	32	40	31	44
Male	30	39	29	37
Female	38	44	39	44

Chi-Squared Test

The `Significant Chi-Square Test Results` This table presents chi-square test results assessing the relationships between various categorical variables related to smoking cessation and associated factors, such as demographic, psychological, and behavioral measures. Each row shows the chi-square statistic, degrees of freedom, and p-value for a specific variable pair, indicating whether their association is statistically significant. Notably, variables like abstinence (`abst`), nicotine dependence (as measured by `ftcd_score`), and demographic factors (e.g., race/ethnicity variables like `NHW` and `Black`, socioeconomic status indicators such as `inc` and `edu`) exhibit strong associations ($p < 0.05$) with various other factors, suggesting significant relationships that could impact smoking cessation outcomes. For example, `ftcd_score` shows a highly significant relationship with `cpd_ps` (cigarettes per day) and `Only.Menthol`, highlighting the link between nicotine dependence and smoking habits. The significant associations between demographic factors (e.g., `NHW` with `edu` and `inc`) may reflect socioeconomic disparities influencing smoking behaviors. Additionally, the

relationship between mental health indicators (`antidepmed` for antidepressant medication) and readiness to quit (`readiness`) underscores the importance of psychological support in cessation programs. Overall, these results suggest that demographic, socioeconomic, and psychological factors play a role in smoking cessation and should be considered when designing interventions.

Table 2: Significant Chi-Square Test Results

	Variable 1	Variable 2	Chi-Square	Degrees of Freedom	P-Value
X-squared	abst	Var	21.8554	1	0.0000
X-squared3	abst	NHW	5.7282	1	0.0167
X-squared8	abst	ftcd_score	40.4872	10	0.0000
X-squared42	BA	antidepmed	7.0210	1	0.0081
X-squared45	sex_ps	NHW	6.2197	1	0.0126
X-squared46	sex_ps	Black	4.5201	1	0.0335
X-squared51	sex_ps	readiness	19.3378	7	0.0072
X-squared58	NHW	Black	174.1406	1	0.0000
X-squared59	NHW	Hisp	8.7393	1	0.0031
X-squared60	NHW	inc	24.4229	4	0.0001
X-squared61	NHW	edu	22.6856	4	0.0001
X-squared63	NHW	readiness	15.9639	7	0.0254
X-squared67	NHW	antidepmed	4.4881	1	0.0341
X-squared69	NHW	Only.Menthol	53.8773	1	0.0000
X-squared70	Black	Hisp	11.3454	1	0.0008
X-squared71	Black	inc	35.5929	4	0.0000
X-squared72	Black	edu	39.1821	4	0.0000
X-squared74	Black	readiness	14.1423	7	0.0487
X-squared80	Black	Only.Menthol	69.7664	1	0.0000
X-squared90	Hisp	Only.Menthol	6.7797	1	0.0092
X-squared91	inc	edu	95.3405	16	0.0000
X-squared92	inc	ftcd_score	64.1709	40	0.0090
X-squared98	inc	mde_curr	15.2423	4	0.0042
X-squared99	inc	Only.Menthol	36.6583	4	0.0000
X-squared100	edu	ftcd_score	72.4146	40	0.0013
X-squared103	edu	ftcd.5.mins	10.6205	4	0.0312
X-squared107	edu	Only.Menthol	28.6503	4	0.0000
X-squared109	ftcd_score	cpd_ps	382.4318	230	0.0000
X-squared110	ftcd_score	ftcd.5.mins	129.9010	10	0.0000
X-squared113	ftcd_score	mde_curr	19.7527	10	0.0317
X-squared120	readiness	Only.Menthol	23.3752	7	0.0015
X-squared121	cpd_ps	ftcd.5.mins	35.2673	23	0.0489
X-squared131	otherdiag	mde_curr	24.5906	1	0.0000

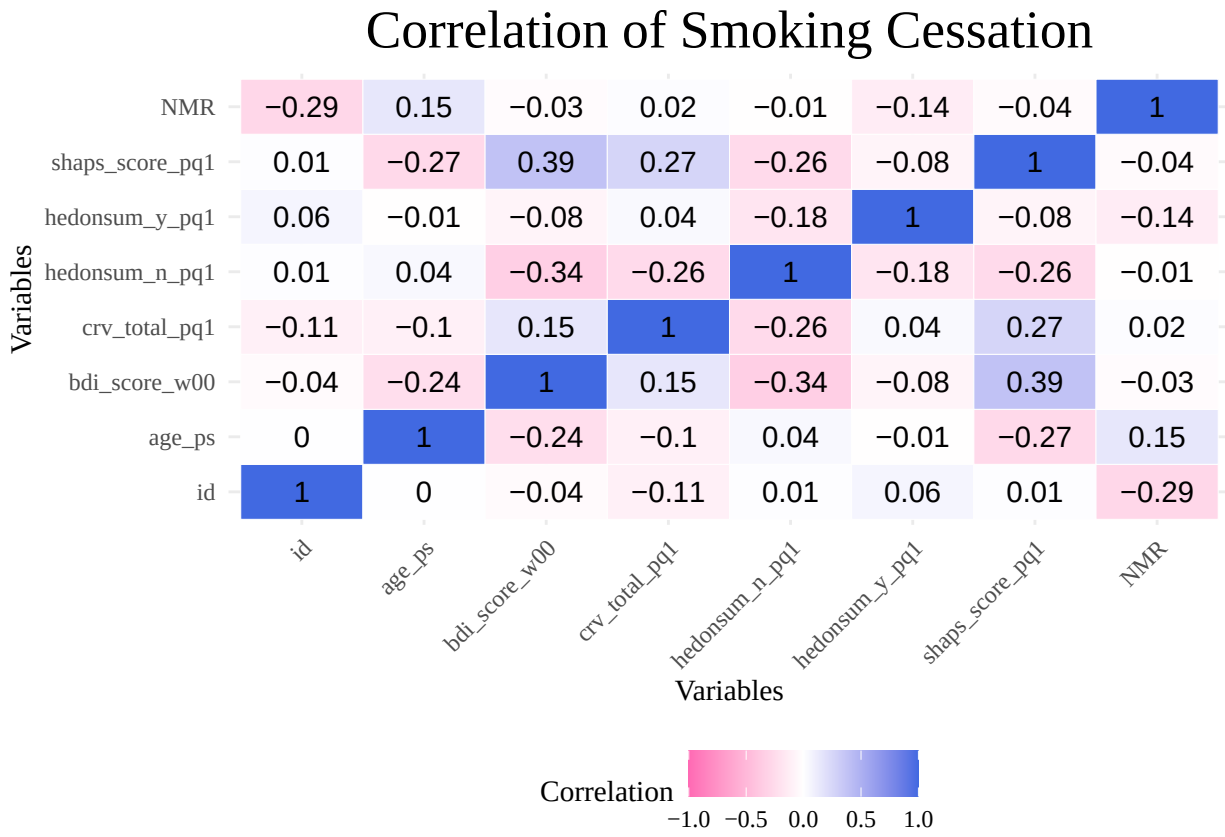
Correlation Plot

According to the `Correlation to Smoking Cessation Plot`, I noticed 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.

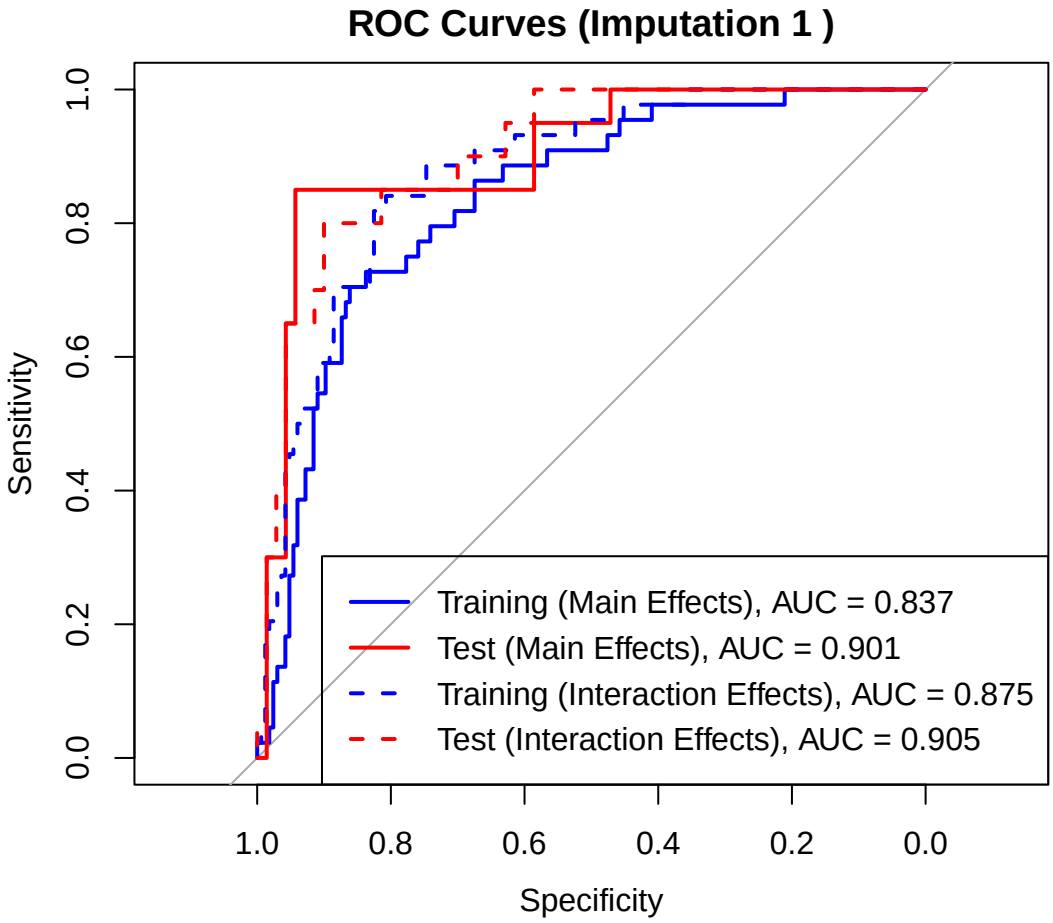


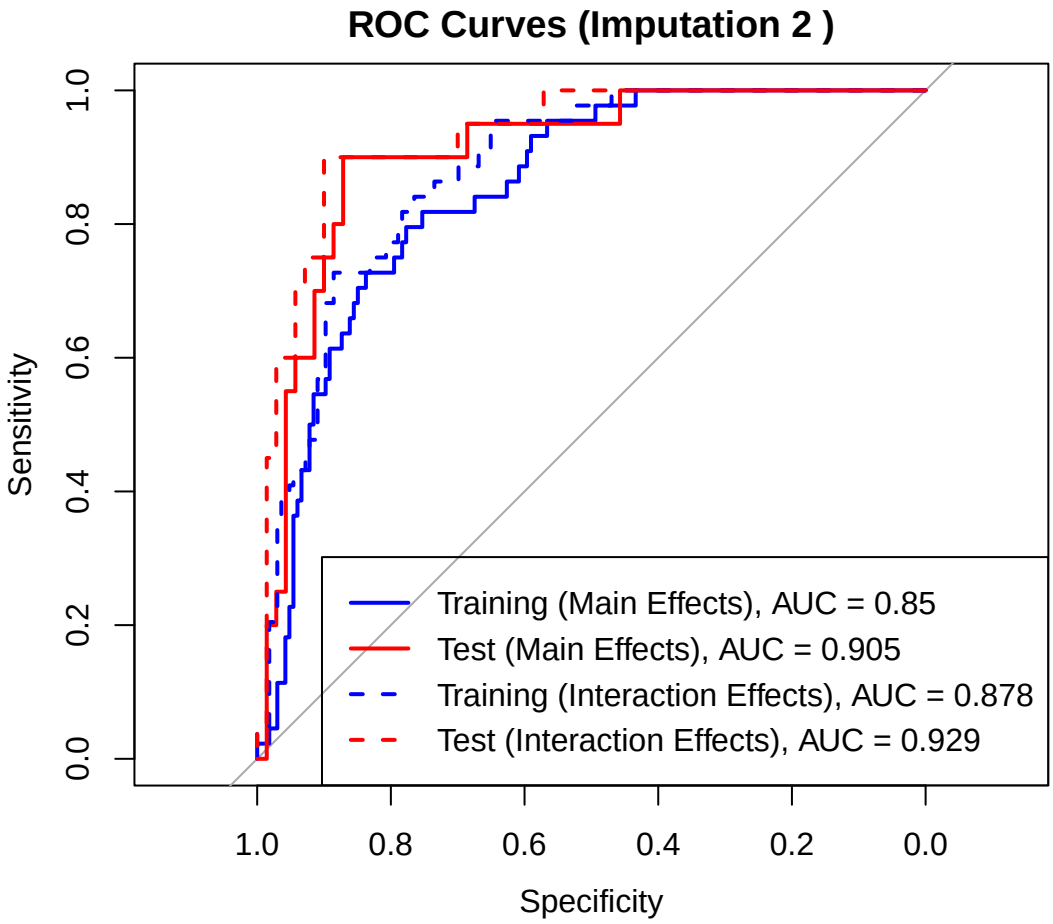
Results

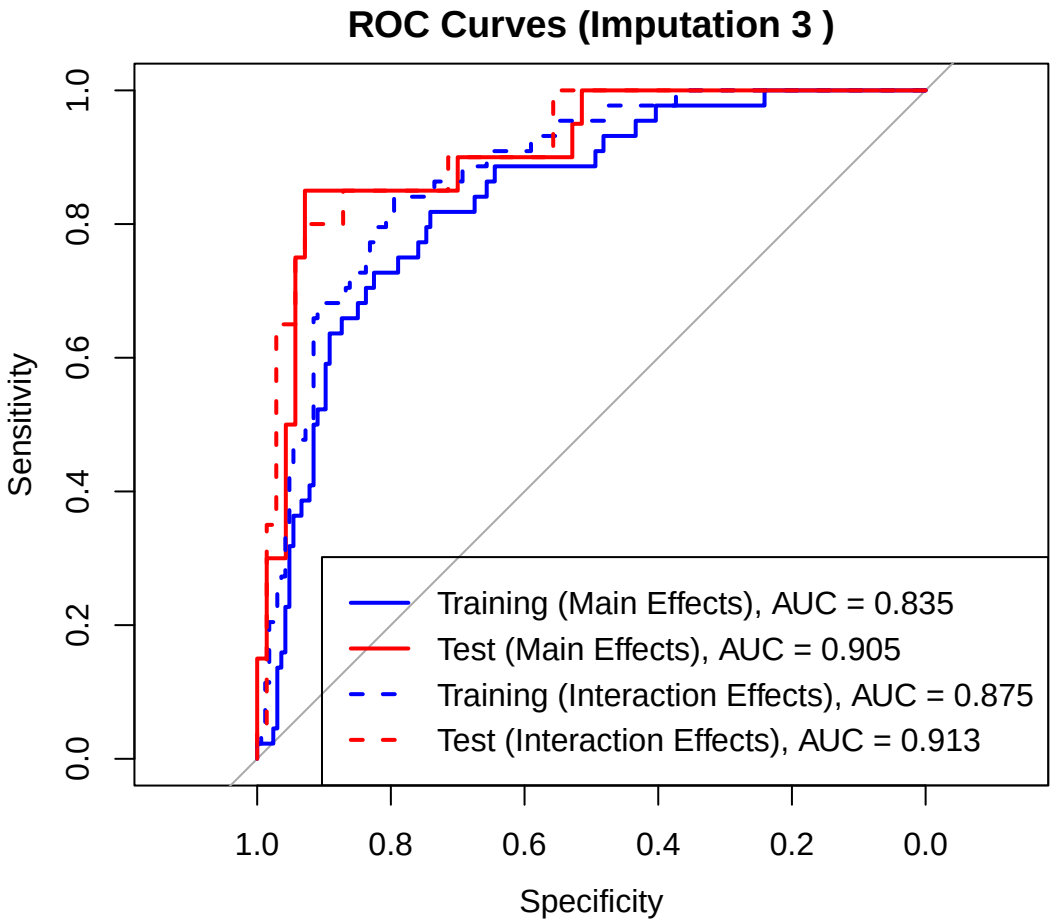
Logistic Modeling

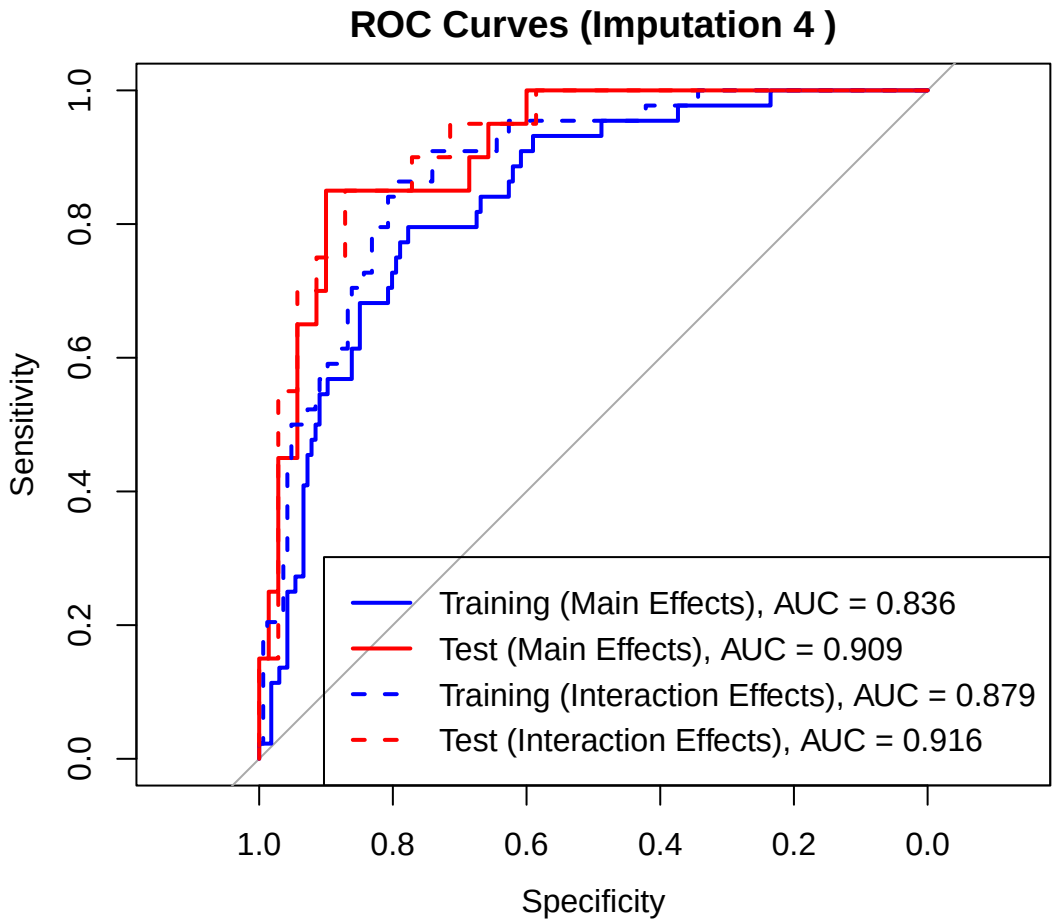
To begin our analysis, we will observe the logistic regression model with all of the main effects on the test and trained data. The trained data represented in this logistic mode is 70% and test data is 30%. In the first table, we notice all the significant predictors that attributes to abstinence amongst the participants in study. So far, Pharmacotherapy (Var), Non-Hispanic Whites, FTCD_score (ftcd_score), Psychotherapy (BA), and Current vs past Major Depressive Disorder (mde_curr) are statistical significant predictors in this model across various imputations.Use of pharmacotherapy, such as nicotine replacement therapy or other medications, is known to improve abstinence rates by alleviating withdrawal symptoms and reducing cravings, making it easier for participants to maintain abstinence.Research indicates that smoking cessation success rates can vary by race/ethnicity due to differences in socioeconomic factors, access to healthcare, and cultural attitudes

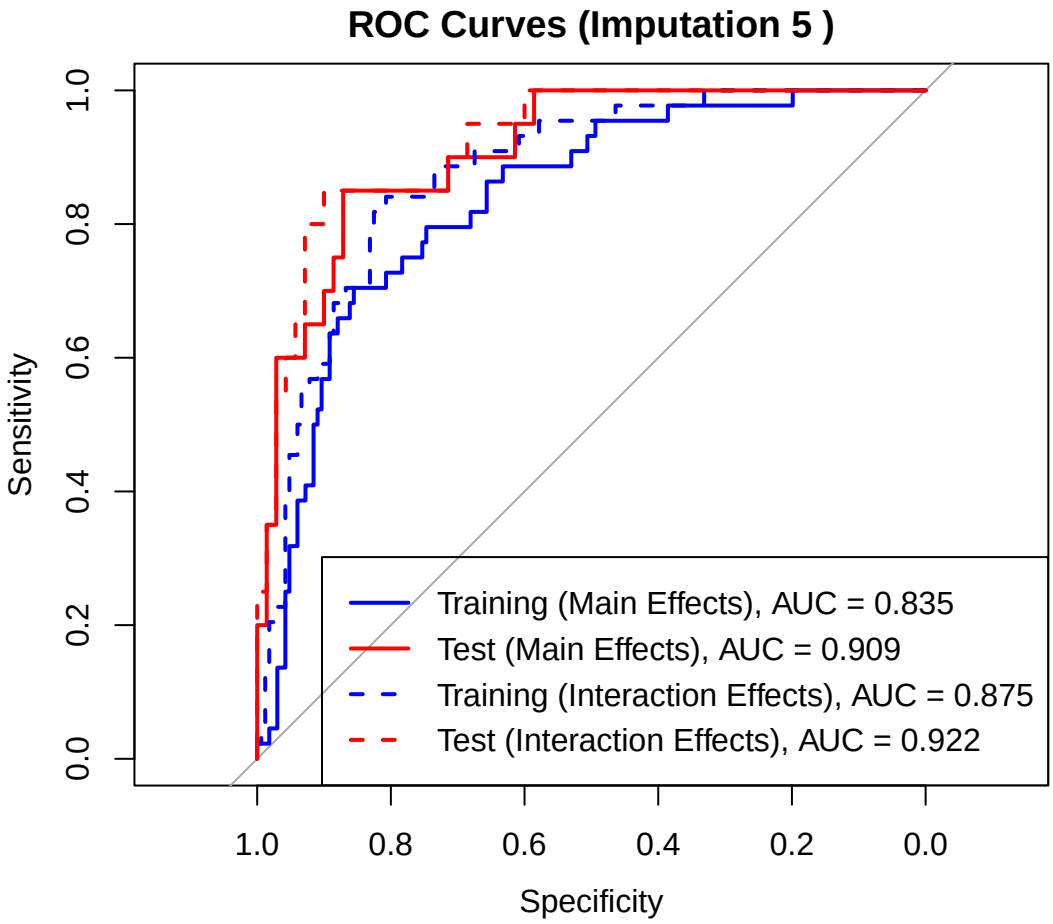
towards smoking. In this study, Non-Hispanic White participants appear to have higher odds of achieving abstinence, potentially reflecting broader access to cessation resources or varying social support structures. This score measures tobacco dependence, with higher scores indicating greater dependence. Lower dependence is generally associated with better cessation outcomes, as individuals with higher dependence often face greater challenges with withdrawal and cravings, reducing their likelihood of achieving abstinence.Behavioral therapy is a common intervention to support smoking cessation, helping participants develop coping mechanisms to resist cravings and manage triggers. Its significance in this model suggests that psychotherapy is an effective tool in aiding participants to achieve and maintain abstinence.Depression is a critical factor in smoking cessation, as those with current depressive symptoms may struggle more with abstinence due to nicotine's role as a mood stabilizer. Participants with active depressive symptoms often face additional challenges in quitting, which may explain the negative association with abstinence.




ROC Curves (Imputation 2)
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
Training (Main Effects), AUC = 0.85
Test (Main Effects), AUC = 0.905
Training (Interaction Effects), AUC = 0.878
Test (Interaction Effects), AUC = 0.929


ROC Curves (Imputation 3)
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
Training (Main Effects), AUC = 0.835
Test (Main Effects), AUC = 0.905
Training (Interaction Effects), AUC = 0.875
Test (Interaction Effects), AUC = 0.913


ROC Curves (Imputation 4)
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Training (Main Effects), AUC = 0.836
Test (Main Effects), AUC = 0.909
Training (Interaction Effects), AUC = 0.879
Test (Interaction Effects), AUC = 0.916



Area Under Curve

These ROC (Receiver Operating Characteristic) curves evaluate the performance of logistic models across five imputations, comparing training and test datasets for models with main effects and interaction effects. The ROC curves illustrate the trade-off between sensitivity (true positive rate) and specificity (false positive rate) for each model.

Across all imputations, the test data generally achieves higher AUC (Area Under the Curve) values than the training data, indicating better predictive accuracy on the test set. The AUC values for the main effects model on the test data range from 0.901 to 0.935, while the interaction effects model on the test data achieves AUCs between 0.905 and 0.929. These high AUC values (close to 1) suggest excellent model performance, with both main and interaction effects contributing to accurate predictions.

The consistent performance across imputations and the higher AUC on test data imply that the model is not overfitted to the training data and generalizes well to new data. The inclusion of interaction effects appears to slightly enhance model performance, as seen in the slightly higher AUCs for test data with interactions. Overall, these ROC curves demonstrate robust model predictability and the importance of considering both main and interaction effects for a comprehensive understanding of smoking abstinence predictors.

Logistic Regression Significant Predictor Variables for Training Data

Table 3: Logistic Model Results (Statistically Significant Coefficients Across Imputations)

term	estimate	std.error	statistic	p.value	imputation

Var	1.6433532	0.4557069	3.606163	0.0003108	1
NHW	2.0258497	0.8933806	2.267622	0.0233523	1
ftcd_score	-0.4068411	0.1443170	-2.819078	0.0048162	1
mde_curr	-1.0667848	0.5233301	-2.038455	0.0415045	1
Var	1.6619218	0.4650306	3.573790	0.0003519	2
NHW	1.9738076	0.9041043	2.183164	0.0290238	2
ftcd_score	-0.4623460	0.1495188	-3.092227	0.0019866	2
mde_curr	-1.0435068	0.5300319	-1.968762	0.0489804	2
Var	1.6439457	0.4584891	3.585572	0.0003363	3
NHW	2.0672436	0.8936462	2.313268	0.0207079	3
ftcd_score	-0.4168512	0.1449833	-2.875166	0.0040381	3
mde_curr	-1.0692759	0.5251498	-2.036135	0.0417368	3
Var	1.6798725	0.4562666	3.681778	0.0002316	4
NHW	1.9632767	0.8981891	2.185817	0.0288290	4
ftcd_score	-0.4074906	0.1443221	-2.823480	0.0047505	4
mde_curr	-1.0521853	0.5254606	-2.002406	0.0452411	4
Var	1.6643650	0.4558300	3.651285	0.0002609	5
NHW	2.0108050	0.8959982	2.244207	0.0248191	5
ftcd_score	-0.4135940	0.1445738	-2.860781	0.0042260	5
mde_curr	-1.0706589	0.5233630	-2.045729	0.0407830	5

Pooled Results Logistic Regression Model on Training Data with Main Effects

In the case of this `Pooled Results from Logistic Regression Model on Training Data with Main Effects`, we noticed that Non-Hispanic Whites and FTCD_Score are sgnificant indicators in this pooled data.Non-Hispanic White (NHW) has positive coefficient (Mean = 2.096) with a CI that does not include zero suggests that being Non-Hispanic White is significantly associated with higher odds of achieving smoking abstinence. This may reflect differential access to resources or social factors that facilitate smoking cessation. FTCD Score (ftcd_score) has a negative coefficient (Mean = -0.332) which implies that higher tobacco dependence (as measured by the FTCD score) is associated with lower odds of achieving abstinence. This aligns with the understanding that greater nicotine dependence creates additional barriers to quitting due to stronger withdrawal symptoms and cravings.

Table 4: Pooled Logistic Model Results (Statistically Significant Coefficients Across Imputations)

term	Mean	SE_within	SE_between	Count	Pooled_SE	Lower_CI	Upper_CI
NHW	2.0081965	0.0374310	0.0017513	5	0.1988280	1.6184936	2.3978994
Var	1.6586916	0.0137402	0.0002360	5	0.1184202	1.4265881	1.8907952
ftcd_score	-0.4214246	0.0208021	0.0005409	5	0.1464622	-0.7084904	-0.1343588
mde_curr	-1.0604823	0.0107482	0.0001444	5	0.1045059	-1.2653139	-0.8556508

Logistic with Testing Data and Main Effects

From observing the data, there isn't any significant predictors but Var the pharmacotherapy with a mean of 3.46 and pooled standard error 0.68.

Table 5: Pooled Logistic Model Results with Test Data and Main Effects (Statistically Significant Coefficients Across Imputations)

term	Mean	SE_within	SE_between	Count	Pooled_SE	Lower_CI	Upper_CI
Var	3.455093	0.3119449	0.121637	5	0.67669	2.12878	4.781405

Logistic Regression Using Training Data and Interactions

Pharmacotherapy is positive, with a mean estimate of 1.721 and a confidence interval (CI) from 1.497 to 1.945. This positive association suggests that pharmacotherapy significantly increases the odds of smoking abstinence. The low pooled standard error (0.114) and consistent significance across all imputations (Count = 5) reinforce the robustness of this predictor.The mean coefficient for ftcd_score is -0.471, with a CI ranging from -0.726 to -0.215. This negative association indicates that higher tobacco dependence, as measured by FTCD score, is linked to lower odds of achieving abstinence. A higher FTCD score reflects greater dependence, which typically makes it harder for individuals to quit smoking due to increased withdrawal symptoms and cravings. The pooled standard error (0.130) is relatively low, and this variable is significant across all imputations.Current Major Depressive Disorder has a mean coefficient is -1.365, with a CI from -1.833 to -0.898. This negative coefficient suggests that participants currently experiencing major depressive disorder (MDD) have lower odds of achieving smoking abstinence. The presence of current MDD may make quitting more challenging due to the potential use of nicotine as a mood regulator. The larger pooled standard error (0.239) still indicates consistent significance across imputations.

Table 6: Pooled Logistic Model Results on Training Data (Statistically Significant Coefficients Across Imputations)

term	Mean	SE_within	SE_between	Count	Pooled_SE	Lower_CI	Upper_CI
Var	1.7209110	0.0128427	0.0002062	5	0.1144121	1.4966632	1.9451588
ftcd_score	-0.4705946	0.0166064	0.0003447	5	0.1304610	-0.7262981	-0.2148910
mde_curr	-1.3653827	0.0526549	0.0034657	5	0.2383563	-1.8325609	-0.8982044

Logistic Regression Using Test Data and Interactions

This table presents pooled logistic model results on the test data, showing that the variable Var is a statistically significant predictor of the outcome across imputations. The mean coefficient for Var is 3.804, indicating a strong positive association with the likelihood of smoking abstinence. The pooled standard error (SE) is relatively low at 0.617, suggesting that the estimate is precise.

The confidence interval (CI) for Var ranges from 2.595 to 5.013, which does not include zero, reinforcing the statistical significance of this predictor. This wide yet positive range further indicates a substantial effect size. Given that Var is representes as pharmacotherapy in this study, these results imply that pharmacotherapy is highly effective in increasing the odds of abstinence in the test data.

Overall, this strong and consistent positive association highlights the critical role of pharmacotherapy in smoking cessation. It suggests that, even when tested on new data, pharmacotherapy remains a powerful tool for helping participants achieve and sustain abstinence. This reinforces the importance of including pharmacotherapy in smoking cessation interventions, as it significantly boosts participants' chances of quitting successfully.

Table 7: Pooled Logistic Model Results on Test Data (Statistically Significant Coefficients Across Imputations)

term	Mean	SE_within	SE_between	Count	Pooled_SE	Lower_CI	Upper_CI
Var	3.803727	0.2706728	0.0915797	5	0.6169023	2.594599	5.012856

Lasso Modeling

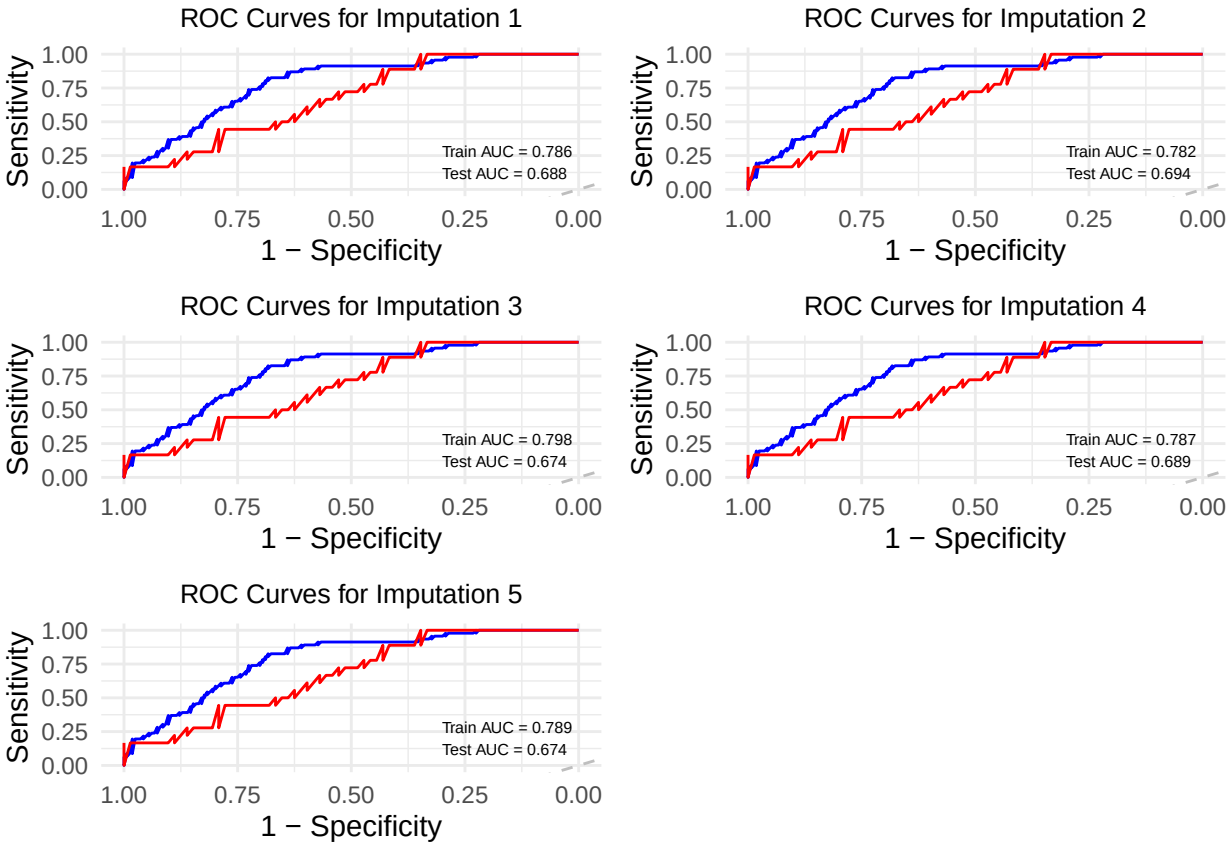


Table 8: Lasso Model Selected Variables (Non-Zero Coefficients)

	Variable	Mean	SE
3	Var	1.0578609	0.2992835
23	NMR	0.6604634	0.7569528
12	ftcd_score	-0.1760150	0.1188627
7	NHW	0.1646441	0.3305748
10	inc	0.0204397	0.1006115
19	shaps_score_pq1	-0.0157695	0.1170541
18	hedonsum_y_pq1	-0.0054638	0.0499208
16	crv_total_pq1	0.0052087	0.1028597
2	id	-0.0007159	0.0217926
15	cpd_ps	-0.0002115	0.0205747

The ROC curves for the five imputations illustrate the performance of the Lasso model on training (blue) and testing (red) sets, with AUC values provided for each. Across all imputations, training AUC values are consistently higher (around 0.78–0.79), indicating strong model performance on the training data. However, the test AUC values are lower (ranging from 0.674 to 0.689), suggesting a decrease in model performance when applied to new data, potentially indicating slight overfitting. In Table 8, the Lasso model selected specific variables with non-zero coefficients, such as NMR, ftcd_score, shaps_score_pq1, and crv_total_pq1, indicating these variables are significant predictors in the model. The positive and negative coefficients represent the direction of association with the outcome. For instance, NMR (nicotine metabolism rate) has a positive coefficient, indicating a higher metabolism rate might increase the likelihood of the outcome, while ftcd_score (a nicotine dependence measure) has a negative coefficient, suggesting higher dependence might be associated with a lower probability of the desired outcome. Overall, the selected variables highlight the key factors affecting smoking cessation, with the model capturing significant relationships despite some generalization issues indicated by the test AUC values.

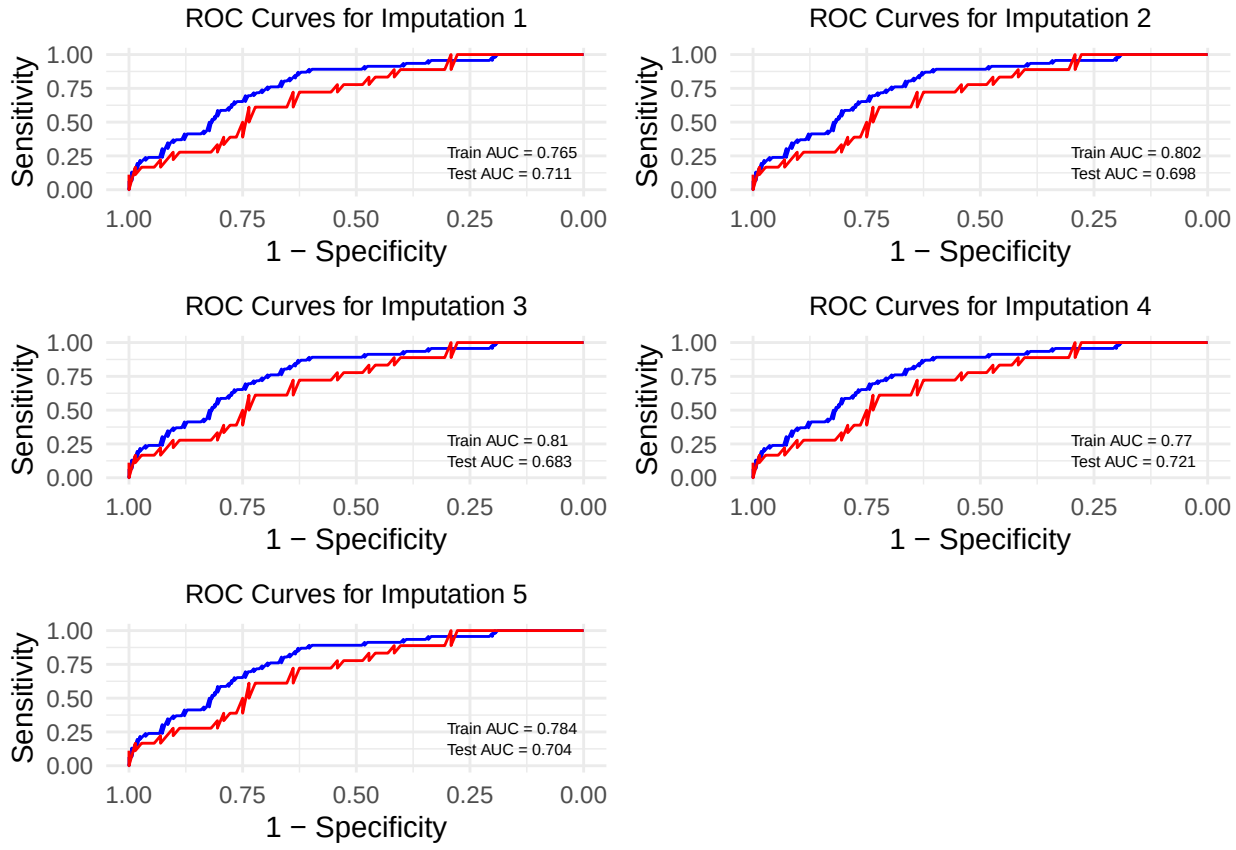


Table 9: Lasso Model Selected Variables with Interaction Terms (Non-Zero Coefficients)

	Variable	Mean	SE
68	Var:NMR	1.2435251	0.8212252
61	Var:crv_total_pq1	0.0283203	0.1645366
12	ftcd_score	-0.0215332	0.1721554
3	Var	0.0154619	0.1396994
223	ftcd_score:readiness	-0.0146495	0.0615760
271	hedonsum_n_pq1:NMR	0.0021274	0.0654365
158	Black:hedonsum_y_pq1	-0.0013556	0.0409328
50	Var:age_ps	0.0011500	0.0387292

188	inc:hedonsum_n_pq1	0.0002054	0.0143640
38	id:cpd_ps	-0.0000217	0.0041243
37	id:bdi_score_w00	-0.0000079	0.0032275

Interaction Terms Lasso Model

The Lasso model presented in this table highlights selected variables with interaction terms that have non-zero coefficients, indicating their potential relevance in predicting outcomes related to smoking cessation. Interaction terms, such as ftcd_score:readiness and hedonsum_n_pq1:NMR, capture the combined effects of these variables, providing deeper insights into how certain factors jointly influence smoking cessation. For instance, the interaction between ftcd_score (a nicotine dependence measure) and readiness to quit may suggest that the impact of nicotine dependence varies depending on an individual's motivation level. Similarly, the hedonsum_n_pq1:NMR interaction indicates that pleasure-seeking behaviors may relate differently to nicotine metabolism rates, potentially affecting cessation success. Including these interactions allows for a nuanced understanding of complex relationships, emphasizing that the effect of one variable can depend on the levels of another.

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 β coefficents to and identify predictor variables for model selection. In the case of our best model, Pharmacotherapy (Var), Nicotine Metabolism Ratio (NMR), Current vs past current (mde_curr), Non-Hispanic Whites (NHW), Sex at phone interview (sex_ps), FTCD_score(ftcd_score),baseline readiness to quit smoking (readiness), and cigarette reward value at baseline(crv_total_pq1) are good predictors of abstinence amongst tobacco dependent individuals.

Conclusion

In conclusion, the logistic performed better according to the area under the curve models due to it's target estimates of AUC being closer to 1. 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, the Chi Squared demonstrated highly correlated interaction terms which impacted the way the logistic reacted. 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. The significant variables for logistic models are Non-Hispanic White (NHW), Pharmacotherapy (Var), FTCD_score, and Current vs Past MDD.

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.

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)




# 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)
```

```r
# 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)

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"))

# 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
  )



#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"
```

```
  ))

# Create demographics table

demographics_table <- project2_table %>%
  group_by(treatment_groups) %>%
  summarise(
     N= n(),
    `Mean Ages` = round(mean(age_ps, na.rm = TRUE), 1),
    `Standard Deviation Ages` = round(sd(age_ps, na.rm = TRUE), 1),
    Blacks = round(sum(as.numeric(as.character(Black)), na.rm = TRUE), 1),
    Hispanics = round(sum(as.numeric(as.character(Hisp)), na.rm = TRUE), 1),
    `Non-Hispanic Whites` = round(sum(as.numeric(as.character(NHW)), na.rm = TRUE), 1),
    `Black Percentage` = round(mean(as.numeric(as.character(Black)), na.rm = TRUE) * 100, 1),
    `Hispanic Percentage` = round(mean(as.numeric(as.character(Hisp)), na.rm = TRUE) * 100, 1),
    `Non-Hispanic White Percentage` = round(mean(as.numeric(as.character(NHW)), na.rm = TRUE) * 100, 1)
    `Education Levels`= c(""),
   `Income Levels`= c("") ,
   `Major Depressive Disorder`= c(""),
    `Antidepressant Medication (%)`=round(mean(as.numeric(as.character(antidepmed)), na.rm = TRUE)*100,
    `Other Lifetime Diagnosis`= round(sum(as.numeric(as.character(otherdiag)), na.rm = TRUE), 1),
    `Cigarette Type`= c(""),
   ` Menthol Only` = round(sum(as.numeric(as.character(Only.Menthol)), na.rm = TRUE), 1),
   `Smoking`= c(""),
   `Cigarettes Per Day at Baseline` = round(mean(as.numeric(as.character(cpd_ps)), na.rm = TRUE), 1),
   `Cigarette Reward Value at Baseline`=round(mean(as.numeric(as.character(crv_total_pq1)), na.rm = TRU
    `FTCD at Baseline`=round(mean(as.numeric(as.character(ftcd_score)), na.rm = TRUE), 2),
  `Readiness to Quit`=round(mean(as.numeric(as.character(readiness)), na.rm = TRUE), 1),
  `Time to smoking upon waking up`= c(""),
  `Smoking 5 minutes into waking up (%)`=round(mean(as.numeric(as.character(ftcd.5.mins)), na.rm = TRUE

  `Pleasurable Events at Baseline`= c(""),
  `Substitute Reinforcers`= round(mean(as.numeric(hedonsum_n_pq1), na.rm = TRUE), 1),
  `Complimentary Reinforcers`= round(mean(as.numeric(hedonsum_y_pq1), na.rm = TRUE), 1),
   Sex= c("")
  )

education_counts <- project2_table %>%
  group_by(treatment_groups, edu) %>%
  summarize(Education_Count = n()) %>%
  pivot_wider(names_from = edu, values_from = Education_Count, values_fill = 0)



inc_counts <- project2_table%>%
  group_by(treatment_groups, inc) %>%
  summarize(Income_Count = n()) %>%
  pivot_wider(names_from = inc, values_from = Income_Count, values_fill = 0)



mdd_counts <- project2_table%>%
  group_by(treatment_groups, mde_curr) %>%
```

```r
  summarize(MDE_Count = n()) %>%
  pivot_wider(names_from = mde_curr, values_from = MDE_Count, values_fill = 0)

sex_counts <- project2_table %>%
  group_by(treatment_groups, sex_ps) %>%
  summarize(Sex_Count = n()) %>%
  pivot_wider(names_from = sex_ps, values_from = Sex_Count, values_fill = 0)



demographics_table <- demographics_table %>%
  left_join(education_counts, by = "treatment_groups") %>%
  left_join(inc_counts, by = "treatment_groups") %>%
  left_join(mdd_counts, by = "treatment_groups")

demographics_table <- demographics_table %>%
  select(treatment_groups, `N`, `Mean Ages`, `Standard Deviation Ages`, `Sex`,Blacks, Hispanics, `Black
         `Cigarette Type`, ` Menthol Only`,  everything())




# Transpose the table
transposed_table <- as.data.frame(t(demographics_table))

# Update column names to reflect the new format
colnames(transposed_table) <- transposed_table[1, ]  # Set the first row as column names
transposed_table <- transposed_table[-1, ]           # Remove the first row (now redundant)
#rownames(transposed_table) <- c("Blacks")            # Rename rows if needed

# Transpose the sex_counts table
sex_counts_transposed <- as.data.frame(t(sex_counts))
colnames(sex_counts_transposed) <- sex_counts$treatment_groups  # Set treatment groups as column names
sex_counts_transposed <- sex_counts_transposed[-1, ]  # Remove the first row (which was treatment group
rownames(sex_counts_transposed) <- c("Male", "Female")  # Set appropriate row names

# Rename specific row names
rownames(transposed_table)[rownames(transposed_table) == "1.y"] <- "Less than $20,000"
rownames(transposed_table)[rownames(transposed_table) == "2.y"] <- "$20,000-35,000"
rownames(transposed_table)[rownames(transposed_table) == "3.y"] <- "$35,001-50,000"
rownames(transposed_table)[rownames(transposed_table) == "4.y"] <- "$50,001-75,000"
rownames(transposed_table)[rownames(transposed_table) == "5.y"] <- "More than $75,000"
rownames(transposed_table)[rownames(transposed_table) == "NA"] <- "Unknown"
rownames(transposed_table)[rownames(transposed_table) == "1.x"] <- "Grade School"
rownames(transposed_table)[rownames(transposed_table) == "2.x"] <- "Some High School"
rownames(transposed_table)[rownames(transposed_table) == "3.x"] <- "High School Graduate or GED"
rownames(transposed_table)[rownames(transposed_table) == "4.x"] <- "Some College"
rownames(transposed_table)[rownames(transposed_table) == "5.x"] <- "College Graduate"



# Combine the transposed sex counts with the main transposed demographic table
transposed_table <- rbind(transposed_table, sex_counts_transposed)
```

```r
#Create Kable Extra Table
kable(transposed_table,
      caption = "Demographics Table of the Smoking Cessation Participants",
      col.names = c("Demograhic Characteristics Variable", "BA+Placebo", "BA+VA", "ST+Placebo", "ST+VA"),
      digits =2) # Set digits for p-values and chi-square



#Change variables into factor and continous variables
factor_vars <- c("abst","Var","BA","sex_ps", "NHW",
                 "Black", "Hisp", "inc", "edu","ftcd_score",
                 "readiness", "cpd_ps",
                 "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))


chi_square_results <- data.frame(
  Variable1 = character(),
  Variable2 = character(),
  Chi_Square = numeric(),
  Degrees_of_Freedom = numeric(),
  P_Value = numeric(),
  stringsAsFactors = FALSE
)

# Loop through each unique pair of variables
for (i in 1:(length(factor_vars) - 1)) {
  for (j in (i + 1):length(factor_vars)) {
    var1 <- factor_vars[i]
    var2 <- factor_vars[j]

    # Check if both columns exist in the data frame
    if (all(c(var1, var2) %in% colnames(project2))) {
      # Create a contingency table
      table_data <- table(project2[[var1]], project2[[var2]])

      # Perform the chi-square test
      chi_test <- chisq.test(table_data)

       # Add results to the data frame
      chi_square_results <- rbind(chi_square_results, data.frame(
        Variable1 = var1,
        Variable2 = var2,
        Chi_Square = chi_test$statistic,
        Degrees_of_Freedom = chi_test$parameter,
        P_Value = chi_test$p.value
      ))
    }
```

```
  }
}


significance_level <- 0.05

# Filter for significant associations only
significant_results <- chi_square_results %>%
  filter(P_Value < significance_level)

# Display the significant results as a kable table
kable(significant_results,
      caption = "Significant Chi-Square Test Results",
      col.names = c("Variable 1", "Variable 2", "Chi-Square", "Degrees of Freedom", "P-Value"),
      digits = 4) # Set digits for p-values and chi-square


# 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))

# 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
```

```r
# Data Imputation for Logistic Modeling
imputed_data1 <- mice(project2, m = 5, method = 'pmm', maxit = 10, seed = 222, print=FALSE)

# Initialize lists to store model results, significant coefficients, and AUC values
model_results <- list()
model_results2 <- list()
significant_coef_estimates<- list()
significant_coef_estimates2 <- list()
significant_coef_estimates3 <- list()
significant_coef_estimates4 <- list()
auc_train_main <- list()
auc_test_main <- list()
auc_train_interaction <- list()
auc_test_interaction <- list()

# Loop over each imputed dataset
for (i in 1:5) {
  # Get a complete dataset from the imputed data
  complete_data_ <- complete(imputed_data1, i)

  # Split into training and testing sets
  set.seed(222)
  trainIndex1 <- createDataPartition(complete_data_$abst, p = 0.7, list = FALSE)
  train_data1 <- complete_data_[trainIndex1, ]
  test_data1 <- complete_data_[-trainIndex1, ]

  # Fit the logistic model on Training and Test Data
  main_effects <- glm(abst ~ ., family = binomial(link = "logit"), data = train_data1)
  main_effects2 <- glm(abst ~ ., family = binomial(link = "logit"), data = test_data1)

  # 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 ,
                               family =binomial(link = "logit"), data=train_data1)

  # Fit the logistic model with interaction terms on Test Data
  interaction_effects2 <- 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 ,
                               family =binomial(link = "logit"), data =test_data1)
```

```r
# Tidy the model summary and add an identifier column for the imputed dataset
model_summary <- broom::tidy(main_effects) %>%
  mutate(imputation = i)

  model_summary2 <- broom::tidy(main_effects2) %>%
  mutate(imputation = i)

# Store the tidy data frame
model_results[[i]] <- model_summary
model_results2[[i]] <- model_summary2


# Store only statistically significant coefficients (p-value < 0.05) Training Data
significant_coef_estimates[[i]] <- broom::tidy(main_effects) %>%
  filter(p.value < 0.05) %>%
  select(term, estimate) %>%
  mutate(imputation = i)


# Store only statistically significant coefficients (p-value < 0.05) with Test Data
significant_coef_estimates2[[i]] <- broom::tidy(main_effects2) %>%
  filter(p.value < 0.05) %>%
  select(term, estimate) %>%
  mutate(imputation = i)


# Store only statistically significant coefficients (p-value < 0.05) Train Data
significant_coef_estimates3[[i]] <- broom::tidy(interaction_effects) %>%
  filter(p.value < 0.05) %>%
  select(term, estimate) %>%
  mutate(imputation = i)

 # Store only statistically significant coefficients (p-value < 0.05) Test Data
significant_coef_estimates4[[i]] <- broom::tidy(interaction_effects2) %>%
  filter(p.value < 0.05) %>%
  select(term, estimate) %>%
  mutate(imputation = i)

# ROC Curve and AUC calculation for Main Effects Model
train_data1$predicted_probs_main_train <- predict(main_effects, newdata = train_data1, type = "respons
test_data1$predicted_probs_main_test <- predict(main_effects2, newdata = test_data1, type = "response
roc_train_main <- roc(train_data1$abst, train_data1$predicted_probs_main_train)
roc_test_main <- roc(test_data1$abst, test_data1$predicted_probs_main_test)

# ROC Curve and AUC calculation for Interaction Model
train_data1$predicted_probs_interaction_train <- predict(interaction_effects, newdata = train_data1,
test_data1$predicted_probs_interaction_test <- predict(interaction_effects2, newdata = test_data1, ty
roc_train_interaction <- roc(train_data1$abst, train_data1$predicted_probs_interaction_train)
roc_test_interaction <- roc(test_data1$abst, test_data1$predicted_probs_interaction_test)


# # Calculate AUC values for legend
auc_train_main_val <- auc(roc_train_main)
```

```
  auc_test_main_val <- auc(roc_test_main)
  auc_train_interaction_val <- auc(roc_train_interaction)
  auc_test_interaction_val <- auc(roc_test_interaction)

   # Plot ROC curves for main and interaction effects on the same plot
  plot(roc_train_main, col = "blue", lty = 1, main = paste("ROC Curves (Imputation", i, ")"))
  lines(roc_test_main, col = "red", lty = 1)
  lines(roc_train_interaction, col = "blue", lty = 2)
  lines(roc_test_interaction, col = "red", lty = 2)

  # Add a legend with AUC values
  legend("bottomright",
         legend = c(paste("Training (Main Effects), AUC =", round(auc_train_main_val, 3)),
                    paste("Test (Main Effects), AUC =", round(auc_test_main_val, 3)),
                    paste("Training (Interaction Effects), AUC =", round(auc_train_interaction_val, 3))
                    paste("Test (Interaction Effects), AUC =", round(auc_test_interaction_val, 3))),
         col = c("blue", "red", "blue", "red"),
         lty = c(1, 1, 2, 2),
         lwd = 2)


}




# Combine all data frames into a single data frame
combined_results <- bind_rows(model_results)


# Filter only the significant coefficients (p-value < 0.05)
significant_logistic_results <- combined_results %>%
  filter(p.value < 0.05)

# Display results in a kable table
significant_res <- kable(significant_logistic_results,
                                caption = "Logistic Model Results (Statistically Significant Coefficie
  kable_styling(full_width = F, font_size = 12)

significant_res






# Combine all significant coefficients into a single data frame
combined_significant_results <- bind_rows(significant_coef_estimates)

# Calculate pooled estimates only for terms that are consistently significant across imputations
pooled_significant_results <- combined_significant_results %>%
  group_by(term) %>%
  summarize(
```

```
    Mean = mean(estimate, na.rm = TRUE),
    SE_within = sqrt(mean((estimate - mean(estimate, na.rm = TRUE))^2, na.rm = TRUE)),
    SE_between = var(estimate, na.rm = TRUE),
    Count = n(),  # Count of imputations where the term was significant
    .groups = 'drop'
  ) %>%
  filter(Count == 5) %>%  # Only keep terms significant in all imputations
  mutate(
    Pooled_SE = sqrt(SE_within + (1 + 1/5) * SE_between),  # Rubin's Rules for SE
    Lower_CI = Mean - 1.96 * Pooled_SE,
    Upper_CI = Mean + 1.96 * Pooled_SE
  )

# Display results in a kable table
kable_pooled_significant <- kable(pooled_significant_results,
                                  caption = "Pooled Logistic Model Results (Statistically Significant C
  kable_styling(full_width = F, font_size = 12)

kable_pooled_significant
```

```
# Combine all significant coefficients into a single data frame
combined_significant_results2 <- bind_rows(significant_coef_estimates2)

# Calculate pooled estimates only for terms that are consistently significant across imputations
pooled_significant_results2 <- combined_significant_results2 %>%
  group_by(term) %>%
  summarize(
    Mean = mean(estimate, na.rm = TRUE),
    SE_within = sqrt(mean((estimate - mean(estimate, na.rm = TRUE))^2, na.rm = TRUE)),
    SE_between = var(estimate, na.rm = TRUE),
    Count = n(),  # Count of imputations where the term was significant
    .groups = 'drop'
  ) %>%
  filter(Count == 5) %>%  # Only keep terms significant in all imputations
  mutate(
    Pooled_SE = sqrt(SE_within + (1 + 1/5) * SE_between),  # Rubin's Rules for SE
    Lower_CI = Mean - 1.96 * Pooled_SE,
    Upper_CI = Mean + 1.96 * Pooled_SE
  )

# Display results in a kable table
kable_pooled_significant2 <- kable(pooled_significant_results2,
                                  caption = "Pooled Logistic Model Results with Test Data and Main Effe
  kable_styling(full_width = F, font_size = 12)

kable_pooled_significant2
```

```r
# Combine all significant coefficients into a single data frame
combined_significant_results3 <- bind_rows(significant_coef_estimates3)

# Calculate pooled estimates only for terms that are consistently significant across imputations
pooled_significant_results3 <- combined_significant_results3 %>%
  group_by(term) %>%
  summarize(
    Mean = mean(estimate, na.rm = TRUE),
    SE_within = sqrt(mean((estimate - mean(estimate, na.rm = TRUE))^2, na.rm = TRUE)),
    SE_between = var(estimate, na.rm = TRUE),
    Count = n(),  # Count of imputations where the term was significant
    .groups = 'drop'
  ) %>%
  filter(Count == 5) %>%  # Only keep terms significant in all imputations
  mutate(
    Pooled_SE = sqrt(SE_within + (1 + 1/5) * SE_between),  # Rubin's Rules for SE
    Lower_CI = Mean - 1.96 * Pooled_SE,
    Upper_CI = Mean + 1.96 * Pooled_SE
  )

# Display results in a kable table
kable_pooled_significant3 <- kable(pooled_significant_results3,
                                   caption = "Pooled Logistic Model Results on Training Data (Statistical
  kable_styling(full_width = F, font_size = 12)

kable_pooled_significant3

# Combine all significant coefficients into a single data frame
combined_significant_results4 <- bind_rows(significant_coef_estimates4)

# Calculate pooled estimates only for terms that are consistently significant across imputations
pooled_significant_results4<- combined_significant_results4 %>%
  group_by(term) %>%
  summarize(
    Mean = mean(estimate, na.rm = TRUE),
    SE_within = sqrt(mean((estimate - mean(estimate, na.rm = TRUE))^2, na.rm = TRUE)),
    SE_between = var(estimate, na.rm = TRUE),
    Count = n(),  # Count of imputations where the term was significant
    .groups = 'drop'
  ) %>%
  filter(Count == 5) %>%  # Only keep terms significant in all imputations
  mutate(
    Pooled_SE = sqrt(SE_within + (1 + 1/5) * SE_between),  # Rubin's Rules for SE
    Lower_CI = Mean - 1.96 * Pooled_SE,
    Upper_CI = Mean + 1.96 * Pooled_SE
  )

# Display results in a kable table
```

```r
kable_pooled_significant4 <- kable(pooled_significant_results4,
                                   caption = "Pooled Logistic Model Results on Test Data (Statistically S
  kable_styling(full_width = F, font_size = 12)

kable_pooled_significant4

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

# Set up parameters
set.seed(1)
m <- 5
lasso_coef_estimates <- list()
lasso_optimal_lambdas <- list()
lasso_train_predictions <- list()
lasso_test_predictions <- list()
roc_list <- list()  # To store ROC plots


# Impute missing data
imputed_data1 <- mice(project2, m = m, method = 'pmm', maxit = 10, seed = 222, print= FALSE)

# Set up train-test split
trainIndex <- createDataPartition(complete(imputed_data1, 1)$abst, p = 0.7, list = FALSE)

# Loop over each imputed dataset
for (i in 1:m) {
  completed_data <- complete(imputed_data1, i)

  # Split data
  train_data <- completed_data[trainIndex, ]
  test_data <- completed_data[-trainIndex, ]

  # Model matrix for training and test sets
  X_train <- model.matrix(formula, data = train_data)[, -1]  # Remove intercept column
  Y_train <- train_data$abst
  X_test <- model.matrix(formula, data = test_data)[, -1]  # Remove intercept column
  Y_test <- test_data$abst

  # Fit Lasso model with cross-validation on training data
  cv_fit <- cv.glmnet(X_train, Y_train, alpha = 1, family = "binomial")

  # Store coefficients
  lasso_coef_estimates[[i]] <- coef(cv_fit, s = "lambda.min")

  # Store optimal lambda for reference
  lasso_optimal_lambdas[[i]] <- cv_fit$lambda.min

  # Predict on both train and test sets using the optimal lambda
  lasso_train_predictions[[i]] <- predict(cv_fit, newx = X_train, s = "lambda.min", type = "response")
  lasso_test_predictions[[i]] <- predict(cv_fit, newx = X_test, s = "lambda.min", type = "response")

  # Calculate ROC and AUC for training and test data
```

```
  roc_train <- roc(Y_train, as.numeric(lasso_train_predictions[[i]]))
  auc_train <- auc(roc_train)

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

 # Create ROC plot for both training and test data
  p <- ggplot() +
    geom_line(aes(x = roc_train$specificities, y = roc_train$sensitivities), color = "blue") +
    geom_line(aes(x = roc_test$specificities, y = roc_test$sensitivities), color = "red") +
    geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "gray") +  # Add diagonal refere
    labs(title = paste("ROC Curves for Imputation", i),
         x = "1 - Specificity",
         y = "Sensitivity") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5, size = 10)) +  # Smaller title size
    scale_x_reverse() +
    # Add AUC to the legend
    annotate("text", x = 0.3, y = 0.2, label = paste(
      "Train AUC =", round(auc_train, 3), "\nTest AUC =", round(auc_test, 3)),
      color = "black", size = 2, hjust = 0)

  # Add plot to list
  roc_list[[i]] <- p
}

# Arrange all ROC plots in a grid layout
grid.arrange(grobs = roc_list, ncol = 2)

# Create a data frame to store pooled results for Lasso
lasso_pooled_results <- data.frame(Variable = rownames(lasso_coef_estimates[[1]]),
                                   Mean = NA, SE = NA)

# Calculate the mean and standard error for each coefficient
for (var in lasso_pooled_results$Variable) {
  coefs <- sapply(lasso_coef_estimates, function(x) as.numeric(x[var, 1]))

  # Mean of the coefficients across imputations
  lasso_pooled_results[lasso_pooled_results$Variable == var, "Mean"] <- mean(coefs, na.rm = TRUE)

  # Rubin's Rules for standard error calculation
  se_within <- sqrt(mean((coefs - mean(coefs, na.rm = TRUE))^2))
  se_between <- var(coefs, na.rm = TRUE)
  pooled_se <- sqrt(se_within + (1 + 1/m) * se_between)

  lasso_pooled_results[lasso_pooled_results$Variable == var, "SE"] <- pooled_se
}

# Filter for non-zero mean coefficients
lasso_selected_vars <- lasso_pooled_results[lasso_pooled_results$Mean != 0 &
                                              lasso_pooled_results$Variable != "(Intercept)", ]
lasso_selected_sorted <- lasso_selected_vars[order(-abs(lasso_selected_vars$Mean)), ]
```

```r
# Create a table using kable for Lasso results
kable_lasso_model_table<- kable(lasso_selected_sorted, caption = "Lasso Model Selected Variables (Non-Z
  kable_styling(full_width = F, font_size = 12)

kable_lasso_model_table
```

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

# Set up parameters
set.seed(222)
m <- 5
lasso_coef_estimates2 <- list()
lasso_optimal_lambdas2 <- list()
lasso_train_predictions2 <- list()
lasso_test_predictions2 <- list()
roc_list2 <- list()  # To store ROC plots
```

```r
# Impute missing data
imputed_data2 <- mice(project2, m = m, method = 'pmm', maxit = 10, seed = 222, print= FALSE)

# Set up train-test split (80% train, 20% test) using the first imputed dataset
trainIndex2 <- createDataPartition(complete(imputed_data2, 1)$abst, p = 0.7, list = FALSE)

# Loop over each imputed dataset
for (i in 1:m) {
  completed_data2<- complete(imputed_data2, i)

  # Split data
  train_data2 <- completed_data2[trainIndex2, ]
  test_data2 <- completed_data2[-trainIndex2, ]

  # Model matrix for training and test sets
  X_train2 <- model.matrix(formula2, data = train_data2)[, -1]  # Remove intercept column
  Y_train2 <- train_data2$abst
  X_test2 <- model.matrix(formula2, data = test_data2)[, -1]  # Remove intercept column
  Y_test2 <- test_data2$abst

  # Fit Lasso model with cross-validation on training data
  cv_fit2 <- cv.glmnet(X_train2, Y_train2, alpha = 1, family = "binomial")

  # Store coefficients
  lasso_coef_estimates2[[i]] <- coef(cv_fit2, s = "lambda.min")

  # Store optimal lambda for reference
  lasso_optimal_lambdas2[[i]] <- cv_fit2$lambda.min
```

```r
  # Predict on both train and test sets using the optimal lambda
  lasso_train_predictions2[[i]] <- predict(cv_fit2, newx = X_train2, s = "lambda.min", type = "response
  lasso_test_predictions2[[i]] <- predict(cv_fit2, newx = X_test2, s = "lambda.min", type = "response")

  # Calculate ROC and AUC for training and test data
  roc_train2 <- roc(Y_train2, as.numeric(lasso_train_predictions2[[i]]))
  auc_train2 <- auc(roc_train2)

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

 # Create ROC plot for both training and test data
  p <- ggplot() +
    geom_line(aes(x = roc_train2$specificities, y = roc_train2$sensitivities), color = "blue") +
    geom_line(aes(x = roc_test2$specificities, y = roc_test2$sensitivities), color = "red") +
    geom_abline(intercept = -2, slope = 1, linetype = "dashed", color = "gray") +  # Add diagonal refer
    labs(title = paste("ROC Curves for Imputation", i),
         x = "1 - Specificity",
         y = "Sensitivity") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5, size = 10)) +  # Smaller title size
    scale_x_reverse() +
    # Add AUC to the legend
    annotate("text", x = 0.3, y = 0.2, label = paste(
      "Train AUC =", round(auc_train2, 3), "\nTest AUC =", round(auc_test2, 3)),
      color = "black", size = 2, hjust = 0)

  # Add plot to list
  roc_list2[[i]] <- p
}

# Arrange all ROC plots in a grid layout
grid.arrange(grobs = roc_list2, ncol = 2)

# Create a data frame to store pooled results for Lasso
lasso_pooled_results2 <- data.frame(Variable = rownames(lasso_coef_estimates2[[1]]),
                                    Mean = NA, SE = NA)

# Calculate the mean and standard error for each coefficient
for (var in lasso_pooled_results2$Variable) {
  coefs2 <- sapply(lasso_coef_estimates2, function(x) as.numeric(x[var, 1]))

  # Mean of the coefficients across imputations
  lasso_pooled_results2[lasso_pooled_results2$Variable == var, "Mean"] <- mean(coefs2, na.rm = TRUE)

  # Rubin's Rules for standard error calculation
  se_within2 <- sqrt(mean((coefs2 - mean(coefs2, na.rm = TRUE))^2))
  se_between2 <- var(coefs2, na.rm = TRUE)
  pooled_se2 <- sqrt(se_within2 + (1 + 1/m) * se_between2)

  lasso_pooled_results2[lasso_pooled_results2$Variable == var, "SE"] <- pooled_se2
}
```

```r
# Filter for non-zero mean coefficients
lasso_selected_vars2 <- lasso_pooled_results2[lasso_pooled_results2$Mean != 0 &
                                                lasso_pooled_results2$Variable != "(Intercept)", ]
lasso_selected_sorted2 <- lasso_selected_vars2[order(-abs(lasso_selected_vars2$Mean)), ]

# Create a table using kable for Lasso results
kable_lasso_model_table2<- kable(lasso_selected_sorted2, caption = "Lasso Model Selected Variables with
  kable_styling(full_width = F, font_size = 12)

kable_lasso_model_table2
```