Evaluating Predictors and Moderators of Smoking Abstinence in Individuals with Major Depressive Disorder: Insights from Behavioral and Pharmacological Interventions

Yunan Chen

2024-11-10

Abstract

This report examines data from a clinical trial exploring smoking abstinence treatments for individuals with major depressive disorder (MDD), who face unique challenges in quitting smoking. The study evaluates two behavioral treatments: Behavioral Activation for Smoking Cessation (BASC) and standard treatment (ST), paired with either varenicline or placebo. Results indicate that varenicline improves smoking abstinence rates compared to placebo, underscoring its effectiveness in this population. BASC, designed to counter depressive symptoms by encouraging engagement in meaningful activities, shows promise, especially when combined with varenicline. However, certain baseline factors, such as nicotine dependence and active depressive symptoms, appear to influence the behavioral treatment effectiveness. Higher nicotine dependence slightly reduces BASC's impact, while current MDD symptoms can act as a barrier to successful abstinence. Exploratory analyses highlight that demographic and behavioral factors like education, income, and age affect somking abstinence. Regression analysis reveals predictors of abstinence, suggesting that individuals with higher nicotine metabolism ratio, lower nicotine dependence, and certain income levels have greater odds of quitting. The model's classification accuracy indicates reliable performance, though calibration issues on test data suggest potential for improvement. These findings emphasize the need for tailored interventions for smokers with MDD, particularly combining BASC and varenicline, to enhance long-term cessation success.

Introduction

Major depressive disorder (MDD), or clinical depression, is a mood disorder characterized by a persistent feeling of sadness and loss of interest, impacting individuals' emotions, thoughts, and behaviors and potentially leading to various emotional and physical complications. Research indicates that smokers with a history of depression are less likely to quit successfully and are more susceptible to relapse than those without depression (Cook et al., 2010). Previous studies have shown that smokers with MDD tend to smoke more heavily, find smoking more pleasurable than other rewarding activities, exhibit higher nicotine dependence, and experience more severe withdrawal symptoms than smokers without MDD (Hitsman et al., 2023). Varenicline, a prescription drug specifically designed to help the general population quit smoking. However, unlike smokers without mental health disorders, those with mental health disorders, including MDD, are less likely to be prescribed varenicline than nicotine replacement therapy, despite the greater effectiveness of varenicline (Evins et al., 2019). This discrepancy has prompted researchers to explore targeted treatments for individuals with depression who wish to quit smoking. One promising approach is behavioral activation (BA), a therapeutic intervention aimed at enhancing motivation and engagement in rewarding and meaningful activities, which may address both anhedonia and depressive symptoms. However, the combined effect of behavioral activation for smoking cessation BASC and varenicline on smoking abstinence remains underexplored. In light of this, Hitsman et al. (2023) hypothesized that Behavioral Activation for Smoking

Cessation (BASC) would lead to higher long-term abstinence rates compared to standard treatment (ST), and that varenicline would increase long-term abstinence compared to placebo. To test these hypotheses, they conducted a clinical trial.

This report analyzes data from that clinical trial with the objectives of examining baseline variables as potential moderators of the effects of behavioral treatment on end-of-treatment abstinence. Additionally, this analysis assesses baseline variables as predictors of abstinence, while controlling for both behavioral treatment and pharmacotherapy, to identify factors that may enhance cessation outcomes for people with Major Depressive Disorder (MDD).

Data Collection and Data Preprocessing

The data used in this report was provided by Dr. George Papandonatos and derived from a randomized placebo-controlled trial that investigated smoking cessation interventions in individuals with a history of major depressive disorder (MDD) (Hitsman et al., 2023). In this trial, 300 adults who smoked daily (at least one cigarette per day) and had a lifetime diagnosis of MDD were recruited. Participants were randomized to receive one of two behavioral treatments, Behavioral Activation for Smoking Cessation (BASC) or Standard Treatment (ST), alongside either varenicline or placebo. The intervention period lasted 12 weeks. Medication blister packs were dispensed in two sets: at week 3 (for weeks 3–7) and at week 7 (for weeks 8–13). Both BA treatment arms involved eight 45-minute sessions, conducted weekly for the first 4 weeks and biweekly for the remaining 8 weeks, with a strong focus on stress reduction, loss of reward, and social-environmental strategies to support abstinence. Abstinence from smoking was assessed at a follow-up visit during week 27. Baseline demographic, smoking, and psychiatric history data were collected at the start of the study.

`Table 1` shows the baseline characteristics of participants by treatment group and overall sample. The treatment groups are combinations of pharmacotherapy (placebo vs varenicline) and behavioral treatment (ST vs BASC). Participants were approximately equally assigned to each of the four treatment groups. In the groups receiving varenicline (ST + varenicline group and BASC + varenicline group), a higher percentage of participants achieved smoking cessation (32% and 31%, respectively) compared to the placebo group (12% in the ST + placebo group and 5.9% in the BASC + placebo group). This suggests that varenicline may be more effective than placebo in helping participants achieve smoking cessation. Looking at the distribution of education levels across the four groups, only a very small number of participants had Grade school education, with only one participant in the BASC + placebo group. The high school level group was also underrepresented in all groups, ranging from 2.9% to 8.4% in each group. Given the low representation of the "Grade school" group and the "Some of the high school" group, it would be reasonable to combine these two groups into the "High school and below" group. This would reduce the sparsity within each treatment group. Similar to education levels, we could also consider combining the income levels "$50,001-$75,000" and "$75,000 and over" into "$50,000 and over".

`Table 1` also shows that missing data is minimal for most variables, with a few exceptions. For example, Race has 22 missing values (7.3%), Nicotine Metabolism Ratio has 21 missing values (7%), Cigarette Reward Value has 18 missing values (6%), and Readiness to Quit Smoking has 17 missing values (6%). While the overall proportion of missing values is small, these variables are critical for the analysis. To address this, multiple imputation under the assumption of missing at random (MAR) was implemented using the `mice()` function in R to create five imputed datasets, ensuring a robust analysis and minimizing potential bias.

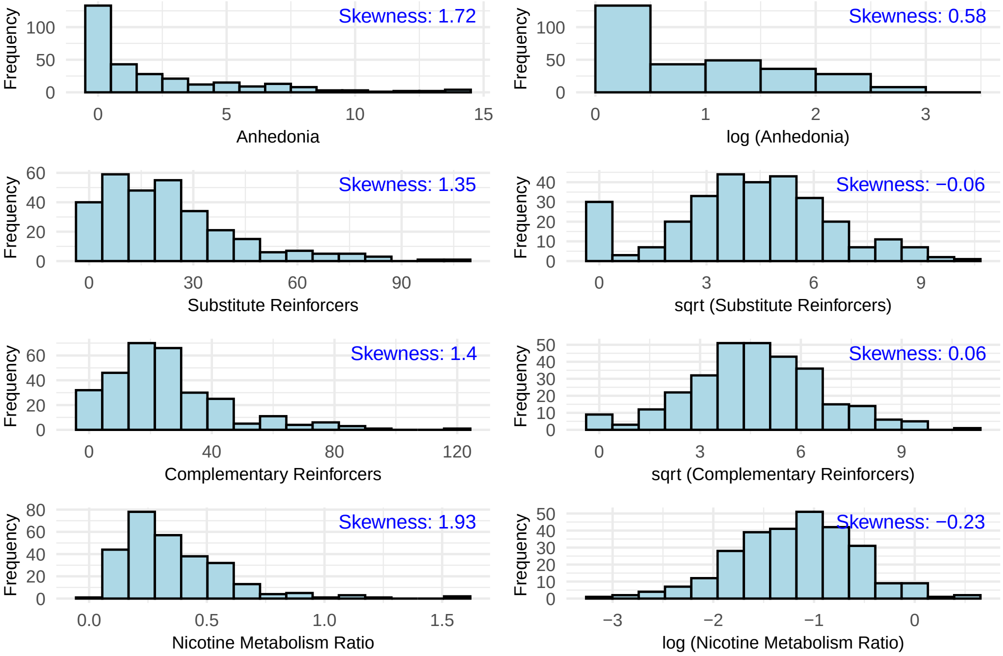
Table 1: Baseline Characteristics by Treatment Group

Characteristic	**ST + placebo** N = 68[1]	**BASC + placebo** N = 68[1]	**ST + varenicline** N = 81[1]	**BASC + varenicline** N = 83[1]	**Overall** N = 300[1]
Smoking Abstinence	8 (12%)	4 (5.9%)	26 (32%)	26 (31%)	64 (21%)
Age	50.3 (10.8)	50.7 (13.5)	48.7 (12.7)	50.3 (13.2)	50.0 (12.6)
Sex (% female)	39 (57%)	38 (56%)	44 (54%)	44 (53%)	165 (55%)
Income /yr					
Less than $20,000	26 (38%)	25 (37%)	29 (36%)	30 (37%)	110 (37%)
$20,000–35,000	14 (21%)	16 (24%)	21 (26%)	17 (21%)	68 (23%)
$35,001–50,000	14 (21%)	8 (12%)	11 (14%)	13 (16%)	46 (15%)
$50,001–75,000	8 (12%)	12 (18%)	6 (7.5%)	12 (15%)	38 (13%)
More than $75,000	6 (8.8%)	6 (9.0%)	13 (16%)	10 (12%)	35 (12%)
Missing	0	1	1	1	3
Education					
Grade school	0 (0%)	1 (1.5%)	0 (0%)	0 (0%)	1 (0.3%)
Some high school	2 (2.9%)	3 (4.4%)	4 (4.9%)	7 (8.4%)	16 (5.3%)
High school graduate or GED	11 (16%)	23 (34%)	27 (33%)	15 (18%)	76 (25%)
Some college/technical school	38 (56%)	22 (32%)	24 (30%)	32 (39%)	116 (39%)
College graduate	17 (25%)	19 (28%)	26 (32%)	29 (35%)	91 (30%)
FTCD score	5.4 (2.1)	5.3 (2.0)	5.2 (2.1)	5.1 (2.3)	5.2 (2.1)
Missing	1	0	0	0	1
Smoking with 5 mins of waking up	35 (51%)	32 (47%)	38 (47%)	33 (40%)	138 (46%)
BDI score	18.5 (10.8)	19.0 (12.3)	19.5 (12.2)	18.0 (10.6)	18.7 (11.5)
Cigarettes /day	15.0 (7.2)	15.6 (9.1)	14.4 (6.6)	15.5 (8.5)	15.1 (7.9)
Cigarette reward value	7.0 (3.7)	7.4 (3.8)	7.1 (3.5)	7.2 (3.9)	7.2 (3.7)
Missing	8	1	6	3	18
Substitute reinforcers	20.8 (20.1)	23.2 (20.3)	23.4 (19.5)	22.9 (19.0)	22.6 (19.6)
Complementary reinforcers	27.4 (19.9)	27.7 (21.5)	25.0 (19.4)	22.4 (17.0)	25.4 (19.4)
Anhedonia	2.5 (3.4)	2.2 (3.2)	2.1 (3.0)	2.3 (3.1)	2.2 (3.2)
Missing	1	2	0	0	3
Lifetime DSM-5 diagnosis	28 (41%)	35 (51%)	40 (49%)	30 (36%)	133 (44%)
Taking antidepressant medication	15 (22%)	28 (41%)	15 (19%)	24 (29%)	82 (27%)
Current vs past MDD	31 (46%)	32 (47%)	44 (54%)	40 (48%)	147 (49%)
Nicotine Metabolism Ratio	0.4 (0.3)	0.3 (0.2)	0.4 (0.2)	0.4 (0.2)	0.4 (0.2)
Missing	2	7	9	3	21
Exclusive Mentholated Cigarette User	43 (64%)	40 (59%)	47 (58%)	48 (59%)	178 (60%)
Missing	1	0	0	1	2
Readiness to quit smoking	7.0 (1.3)	6.8 (1.4)	6.7 (1.1)	6.7 (1.2)	6.8 (1.2)
Missing	4	4	4	5	17
Race					
Balck	40 (61%)	36 (55%)	43 (59%)	36 (49%)	155 (56%)
Hispanic	4 (6.1%)	4 (6.2%)	5 (6.8%)	3 (4.1%)	16 (5.8%)
White	22 (33%)	24 (37%)	25 (34%)	34 (46%)	105 (38%)
Balck and Hispanic	0 (0%)	1 (1.5%)	0 (0%)	1 (1.4%)	2 (0.7%)
Missing	2	3	8	9	22

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

Examining through the distribution of the continuous variables, four variables were found to have skewed distribution, therefore variable transformations were performed (`Figure 1`). Log transformation was performed on Anhedonia (`shaps_score_pq1_log`), and Nicotine Metabolism Ratio (`NMR`). Square root transformations were performed on Substitute Reinforcers (`hedonsum_n_pq1`), and Complementary Reinforcers (`hedonsum_y_pq1`). These transformations effectively reduce the skewness of each variable, making the distributions look more normally distributed. Such adjustments are beneficial for statistical analyses as they minimize the impact of extreme outliers; however, they may also reduce the interpretability of the transformed variables.

Figure 1: Variable Distribution Before and After Transformations



Exploratory Data Analysis

An exploratory analysis was conducted to identify patterns and relationships within the data prior to performing regression analysis. This preliminary investigation aimed to uncover potential trends and insights that could inform the subsequent modeling approach and enhance the interpretation of the regression results.

`Figure 2` illustrates the proportion and number of participants achieving smoking abstinence, stratified by Behavioral Activation (BASC) treatment groups, with lighter colors representing no abstinence and more solid colors indicating abstinence. Across clinically relevant variables, such as antidepressant use, mental health diagnoses, and smoking behavior, differences in abstinence rates are observed. For antidepressant medication users, abstinence was more frequent in the BASC group (38%) compared to the ST group (17%). However, menthol cigarette users receiving BASC had lower abstinence rates (35%) compared to those in the ST group (65%). Similarly, among individuals with high nicotine dependence (smoking within 5 minutes of waking), the BASC group achieved lower abstinence rates (29%) compared to the ST group (48%). These patterns suggest that while BASC supports smoking cessation for some subgroups, its effectiveness may vary, particularly when compared to standard treatment in certain populations.

Table 2: Number and proportion of smokers who have achieved cessation

Variables	ST + placebo	BASC + placebo	ST + vareni-cline	BASC + vareni-cline
Education Level				
Grade school	-	1 (100%)	-	-
Some high school	-	-	2 (50%)	1 (14%)
High school graduate or GED	-	-	11 (41%)	5 (33%)
Some college/technical school	3 (8%)	-	3 (12%)	13 (41%)
College graduate	5 (29%)	3 (16%)	10 (38%)	7 (24%)
Income Level				
Less than $20,000	2 (11%)	-	7 (35%)	6 (26%)
$20,000–35,000	2 (17%)	-	2 (13%)	5 (38%)
$35,001–50,000	-	-	2 (29%)	4 (40%)
$More than $50,000	4 (36%)	1 (7%)	4 (29%)	5 (29%)

Figure 2: Number of Smoking Abstinence by Group

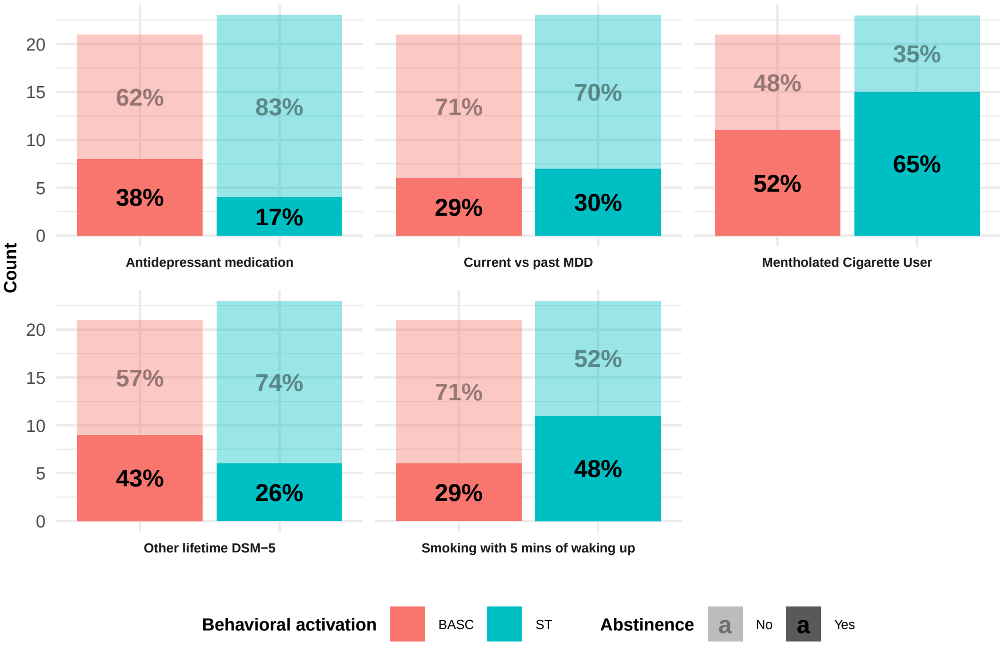


Table 2 summarizes the number and percentage of smokers who achieved cessation across various education and income levels within four treatment groups. Generally, the “ST + varenicline” and “BASC + varenicline” groups had higher cessation rates across both education and income categories, suggesting the efficacy of varenicline in aiding smoking cessation. Among education levels, “College graduate” and “Some college/technical school” participants showed particularly high cessation rates in the “ST + varenicline” and “BASC + varenicline” groups, while “High school graduate or GED” individuals also had notable cessation rates, especially within “ST + varenicline.” In terms of income, higher income brackets, especially those earning “$35,001–50,000” and “More than More than $50,000” had higher cessation rates within the “BASC + varenicline” group, while the “ST + varenicline” group showed consistent cessation rates across various income levels, including lower-income categories. These results suggest that varenicline may be effective across a diverse range of demographic groups.

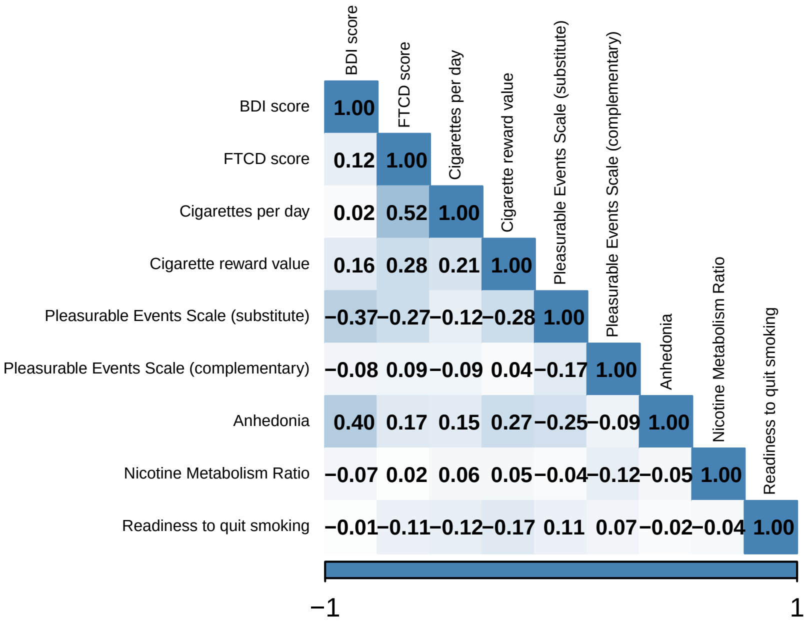
Table 3: Summary of Continuous Variables by Group and Smoking Abstinence

	ST + placebo		BASC + placebo		ST + varenicline		BASC + varenicline	
	No	Yes	No	Yes	No	Yes	No	Yes
Anhedonia	2.68	1.25	2.16	2 (0.41)	2.38	1.54	2.46	1.81
	(0.45)	(0.86)	(0.42)		(0.44)	(0.44)	(0.45)	(0.46)
Cigarette reward value	7.04	6.62	7.44	7.5	6.89	7.55	7.43	6.75
	(0.47)	(1.57)	(0.49)	(1.19)	(0.48)	(0.65)	(0.51)	(0.8)
Cigarettes per day	15.28	13.12	16.06	8.75	14.53	14.23	16.33	13.81
	(0.78)	(4.85)	(1.14)	(3.77)	(0.9)	(1.3)	(1.18)	(1.47)
FTCD score	5.71	3 (1.21)	5.41	3.75	5.18	5.15	5.54	4.04
	(0.21)		(0.25)	(1.31)	(0.3)	(0.36)	(0.28)	(0.49)
Nicotine Metabolism Ratio	0.37	0.36	0.34	0.35	0.33	0.41	0.34	0.46
	(0.04)	(0.04)	(0.02)	(0.06)	(0.03)	(0.04)	(0.02)	(0.07)
Pleasurable Events Scale at	27.52	26.5	27.39	32.5	25.6	23.73	23.09	20.88
baseline–complementary reinforcers	(2.45)	(9.7)	(2.71)	(10.51)	(2.63)	(3.84)	(2.18)	(3.6)
Pleasurable Events Scale at	18.47	37.88	23.06	25.5	25.62	18.81	20.63	27.92
baseline–substitute reinforcers	(2.22)	(11.15)	(2.58)	(8.15)	(2.72)	(3.43)	(2.25)	(4.37)
Readiness to quit smoking	6.96	6.88	6.75	7.67	6.83	6.43	6.66	6.72
	(0.18)	(0.44)	(0.17)	(0.29)	(0.15)	(0.2)	(0.15)	(0.25)

`Table 3` summarizes the mean and standard error (SE) of continuous variables related to smoking behavior, psychological factors (e.g., Anhedonia), and readiness to quit smoking. The table is stratified by treatment group (ST + placebo, BASC + placebo, ST + varenicline, BASC + varenicline) and smoking abstinence status ("Yes" or "No") The comparison between participants who successfully quit smoking and those who did not reveals a consistent pattern across treatment groups: successful smoking cessation is associated with lower nicotine dependence (as measured by FTCD scores), fewer cigarettes smoked per day, and lower anhedonia scores. This trend holds true regardless of the type of treatment group, suggesting that these factors may be the key predictors of smoking cessation success across both pharmacological and behavioral interventions. The Nicotine Metabolism Ratio (NMR) is a measure of how quickly an individual metabolizes nicotine. Faster nicotine metabolizers tend to experience shorter-lasting effects of nicotine, which can lead to smoking more frequently to maintain nicotine levels. NMR is often considered when assessing an individual's likelihood to quit smoking, as those with a higher metabolism rate might find it more challenging to quit due to the need for more frequent dosing. In this table, we observe that the NMR is generally higher in the smoking abstinence ("Yes") group compared to the non-abstinent ("No") group across most treatment conditions, with the exception of the ST + placebo group. This pattern suggests that, for most treatments, individuals with a higher nicotine metabolism rate had better success in achieving abstinence.

`Figure 3` shows the pairwise correlations among continuous clinical variables, revealing a moderate positive correlation between "FTCD score" and "Cigarettes per day" (r = 0.52), indicating some shared information between these variables. Most other correlations, such as those involving "Nicotine Metabolism Ratio" and "Readiness to quit smoking," are weak, suggesting minimal risk of multicollinearity. While the moderate correlation warrants attention to ensure it does not unduly influence the model, the weak correlations imply that the variables can be included with little concern for redundancy.

Figure 3: Correlation Plot among Continuous Clinical Variables



Regression Analysis

Methods

Lasso (Least Absolute Shrinkage and Selection Operator) is well-suited for achieving the goals of this project, as it enables the identification of baseline variables that act as significant predictors or moderators of end-of-treatment (EOT) abstinence while effectively managing the complexity of the data. In this context, the trial data includes a large number of baseline variables, some of which may be highly correlated or only marginally relevant to abstinence outcomes. Lasso helps in this setting by applying a regularization penalty that shrinks less relevant coefficients to zero, effectively selecting only the most predictive and informative variables. This approach addresses the issue of overfitting and improves the model's interpretability by reducing the number of predictors to those most strongly associated with abstinence.

The regression analysis followed a structured sequence of steps. First, missing data was handled using multiple imputations, resulting in five complete datasets. Each imputed dataset was subjected to 10 bootstrapping iterations, where the dataset was resampled with replacement to generate diverse subsets for model evaluation. This approach introduced an additional layer of variability, enhancing the stability of model estimates. Then, each bootstrapped dataset was split into training (70%) and testing (30%) subsets, ensuring that the proportions of behavioral and pharmacotherapy treatment combinations in the training and testing sets were consistent with those in the bootstrapped dataset. LASSO regression with 10-fold cross-validation was applied to the training data to identify the optimal regularization parameter (lambda), which minimized the cross-validation error and selected important features by shrinking irrelevant coefficients to zero. The best-fit LASSO model was then applied to the test data, with model performance assessed through AUC (Area Under the Curve) scores and ROC (Receiver Operating Characteristic) curves for both training and testing datasets. Finally, coefficients from each imputed dataset were averaged to summarize feature importance across imputations, providing a robust model evaluation that accounts for missing data, regularization, and performance validation.

Table 4: Summary of Non-Zero Coefficient Estimates in Exponential Scale

Covariate	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5	**Average**	Proportion Non-Zero
(Intercept)	1.1403	0.4307	0.6256	1.6677	1.5428	1.0996	1.00
BA1	-	-	-	-	-	1.0000	0.00
inc$More than $50,000	-	1.3232	-	2.1038	1.7521	1.3552	0.70
ftcd_score	0.5725	0.9068	0.8269	0.5636	0.7639	0.7194	1.00
NMR_log	1.4696	-	1.2482	1.0719	1.4421	1.3442	0.84
Var1	-	-	-	-	-	1.0000	0.00
BA1:inc$35,001–50,000	3.9902	-	1.0462	4.4054	-	1.9909	0.68
BA1:Only.MentholYes	0.6748	-	-	-	0.7023	0.8071	0.68
age_ps:Var1	1.0253	1.0192	1.0116	1.0016	1.0129	1.0145	0.84
sex_psYes:Var1	1.9163	-	1.0311	2.1114	1.1297	1.4188	0.68
eduHigh school graduate or GED:Var1	1.7813	1.1982	1.8351	2.1017	1.6956	1.6590	0.86

Results

`Table 4` presents the exponentiated non-zero coefficient estimates derived from a Lasso logistic regression model predicting the likelihood of smoking abstinence, incorporating various covariates and interaction terms between treatment groups and baseline characteristics. The coefficient estimates are expressed as odds ratios, where values greater than 1 indicate a positive association with smoking abstinence, while values less than 1 represent a negative association. The columns under each imputation show the average coefficient values across bootstrap samples, while the "Average" column aggregates the averages across all bootstrap samples in the five imputed datasets. The "Proportion Non-Zero" column captures the consistency and importance of each variable, reflecting the proportion of times a variable retained a non-zero coefficient across all imputations and bootstraps. Higher proportions indicate greater stability and significance of a variable, whereas lower proportions suggest a weaker and less consistent association with smoking abstinence.

The variables `FTCD score`, `NMR_log`, and `eduHigh school graduate or GED:Var1` were significant across all timw, with FTCD score and NMR_log showing non-zero coefficients in 100% and 84% of the time, respectively, underscoring their strong and consistent associations with smoking abstinence. Conversely, variables such as `BA1:inc$35,001-50,000` and `BA1:Only.MentholYes` were less stable, showing non-zero coefficients in 68% of the time, suggesting a more variable association with the outcome. Treatment vriables `Var1` and `BA1` consistently showed no meaningful association with smoking abstinence, with a "Proportion Non-Zero" of 0.

For the primary effects, income level (inc More than 50,000) showed an increase in the likelihood of smoking abstinence, with an average odds ratio of 1.3552. This indicates that individuals earning more than 50,000 annually are approximately 35% more likely to achieve smoking abstinence compared to those earning less than 20,000. The FTCD score, representing nicotine dependence, exhibited a negative association with abstinence, with an average odds ratio of 0.7194, meaning that for every unit increase in FTCD score, the odds of smoking abstinence decrease by approximately 28%. Finally, NMR_log, which represents the log-transformed nicotine metabolism ratio (NMR), had an average odds ratio of 1.3442, suggesting that higher values of the log-transformed NMR are associated with a 34% increase in the likelihood of achieving smoking abstinence. This highlights the potential impact of faster nicotine metabolism, captured on a logarithmic scale, on the increased odds of smoking cessation.

Turning to the interactions involving the Behavioral Activation (BA) treatment, the data reveal nuanced patterns. The interaction between BA1 and income level 35,001–50,000 yielded an average odds ratio of 1.99 and was non-zero 68% of the time. This suggests a relatively moderate level of stability and significance, indicating that participants in this income bracket receiving the BA intervention had approximately twice the odds of achieving smoking abstinence compared to those earning less than 20,000. Conversely, the interaction between BA1 and menthol cigarette use (BA1:Only.MentholYes) had an average odds ratio of 0.81. This indicates that menthol cigarette users receiving the BA intervention had about 19% lower odds of achieving abstinence compared to non-menthol users. These findings highlight both the potential benefits of BA for specific demographic subgroups, such as individuals with an income of $35,001–50,000, and the challenges it may pose for others, particularly menthol cigarette users. This suggests that income level and menthol cigarette use are important factors that moderate the effectiveness of BA in achieving smoking abstinence. Consistent with Figure 2, menthol cigarette users in the BA group achieved lower abstinence rates compared to those in the standard treatment (ST) group.

The interaction term between Varenicline and education level (eduHigh school graduate or GED:Var1) demonstrated a strong positive relationship, with an average odds ratio of 1.66, indicating that individuals with a high school diploma or GED had approximately 66% higher odds of achieving smoking abstinence compared to those with only grade school education. This result, supported by a high proportion of non-zero values (86% of the time), underscores the stable and meaningful role of educational attainment in enhancing the effectiveness of pharmacotherapy. This finding aligns with the data in Table 2, where smokers with a high school diploma or GED showed higher cessation rates under the Varenicline treatment arms, particularly in the ST + Varenicline group, reinforcing the importance of education level as a moderator of pharmacotherapy outcomes. Similarly, the interaction between Varenicline and sex (sex_psYes:Var1) showed an average odds ratio of 1.42, suggesting that women may derive additional benefits from Varenicline treatment, with moderately consistent results appearing 68% of the time. This finding aligns with prior evidence suggesting gender-based differences in response to pharmacotherapy, potentially due to physiological or behavioral factors influencing smoking cessation outcomes. The interaction between Varenicline and age (age_ps:Var1) revealed an average odds ratio of 1.01, signifying a slight but positive association, where each additional year of age marginally increases the likelihood of smoking abstinence. This effect, present in 84% of imputations, suggests that older individuals may experience slightly better outcomes with Varenicline, potentially reflecting greater life stage-related motivation or enhanced adherence to cessation programs. These findings collectively highlight the nuanced impact of demographic factors on the success of pharmacotherapy interventions.

Model Performance

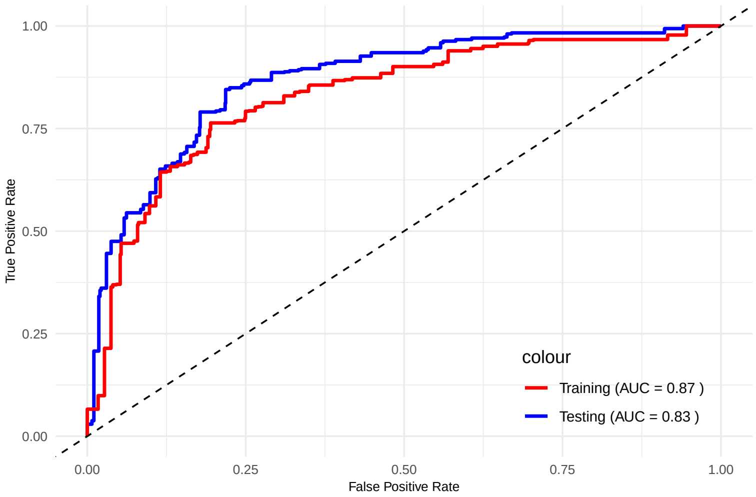
To evaluate the model's performance, ROC and calibration plots were generated, and the area under the curve (AUC), accuracy, sensitivity, and specificity were calculated. The ROC curve and AUC assess the model's classification performance, while the calibration plot evaluates how well the predicted probabilities align with the observed outcome probabilities. Predictions were made using the average coefficient estimates and combined train and test sets from five imputed datasets. As illustrated in `Figure 3`, the training sets achieved an AUC of 0.87, while the testing sets achieved an AUC of 0.83, both indicating that, on average, the model has around an 80% likelihood of correctly distinguishing between positive and negative cases. This AUC reflects a robust and reliable ability to predict smoking abstinence. The consistency of the AUC between training and testing sets suggests that the model has good generalizability, with no substantial drop in classification performance when applied to unseen data.

The performance metrics of the final model, as shown in `Table 5`, reveal strong predictive capability with some trade-offs. The accuracy of 0.81 across both training and testing datasets indicates that the model correctly classifies approximately 81% of cases, reflecting consistent performance across datasets. Additionally, the high sensitivity values of 0.98 and 0.96 for the training and testing sets, respectively, suggest that the model is highly effective at identifying individuals who achieve smoking abstinence (true positives). However, the low specificity values of 0.22 and 0.24 indicate that the model has difficulty correctly identifying individuals who do not achieve abstinence (true negatives), leading to a higher rate of false positives. While the model's high sensitivity may be advantageous for certain applications, the low specificity highlights a limitation in accurately predicting non-abstinent individuals, which could impact its utility depending on the context.

Table 5: Performance of the FInal Model

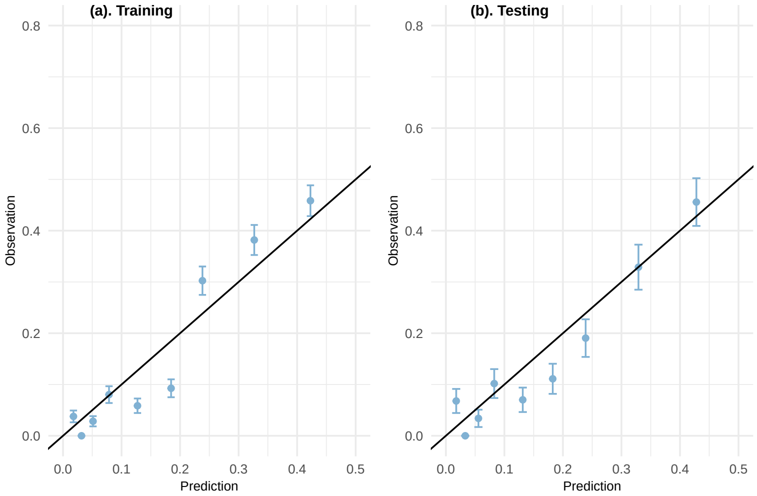
Metric	**Train**	**Test**
AUC	0.87	0.83
Accuracy	0.81	0.81
Sensitivity	0.98	0.96
Specificity	0.22	0.24

Figure 4: ROC Curve Using Average Coefficients



The calibration plot further examines the alignment between predicted and observed probabilities (`Figure 4`). The 45-degree line in the plot represents perfect calibration, where predicted probabilities match the observed outcomes. In the training set, the points are generally close to the calibration line, indicating that the model is well-calibrated. However, there is a slight deviation below the line at higher predicted probability levels, suggesting a mild tendency to over predict the likelihood of smoking abstinence in these cases. In contrast, the calibration plot for the testing set shows greater deviation from the diagonal line, indicating that the model is less well-calibrated on the test data than on the training data. Specifically, the model underpredicts probabilities at lower levels (with points falling below the line) and overpredicts at higher levels (with points above the line). These discrepancies suggest that while the model is reasonably well-calibrated on the training data, it exhibits calibration issues when applied to new, unseen data.

Figure 5: Calibration Plots



Discussion

The EDA revealed that demographic and behavioral factors such as age, education, and baseline nicotine dependence influence cessation outcomes. For instance, higher education levels and moderate income level were associated with greater smoking abstinence rates, suggesting socioeconomic factors may play a role in treatment efficacy. Additionally, individuals with lower nicotine dependence scores and fewer current depressive symptoms exhibited better outcomes, highlighting the importance of addressing dependence levels and mental health symptoms in smoking cessation strategies.

The regression analysis identified several key predictors of smoking abstinence, including the log-scale nicotine metabolism ratio (NMR), FTCD score, interactions of BASC with income ($35,001–$50,000) and only menthol use, and interactions of varenicline with age, sex, and education level (high school graduate or higher). A higher NMR was associated with a 34% increase in the odds of abstinence, while each unit increase in nicotine dependence (FTCD score) reduced the odds of abstinence by 28%. The effectiveness of BASC was slightly diminished for individuals who exclusively smoked menthol cigarettes, while varenicline's effectiveness improved with increasing age, female sex, and higher education levels. The model demonstrated strong classification performance, achieving an AUC of 0.87 in training and 0.83 in testing. However, calibration issues observed in the test data indicate the need for further refinement to improve predictive accuracy and reliability.

However, several limitations should be considered. First, the sample size, while adequate for initial findings, may limit the generalizability of results. Specific subgroups, such as those with varying income levels or education backgrounds, may respond differently to BASC and varenicline, but the sample size in each subgroup was relatively small, potentially affecting the robustness of subgroup analyses. In addition, the need to split the data into training and testing sets, followed by further division of the training set for cross-validation to tune model parameters, reduces the effective data in each subset. The limited data per subset may decrease model stability and generalizability, suggesting that results should be interpreted with caution. A larger sample size in future studies would improve the robustness and predictive reliability of the analysis. Also, splitting the test and train datasets after imputation introduces a data leakage problem, as information from the imputation process, which is influenced by the entire dataset, could inadvertently inform the training or testing stages. Another limitation, rooted in the constraints of the clinical trial, was low treatment adherence, particularly in the BASC-alone group. This low adherence affects the analysis by potentially underestimating the effectiveness of BASC, as participants may not have received sufficient intervention exposure to achieve meaningful outcomes. As a result, the study's findings may not fully reflect the treatment's potential impact, limiting the generalizability and strength of the conclusions. Additionally, the use of variable transformations introduces a limitation in terms of interpretability. Four variables were either square root or log-transformed, which, while effective in improving model performance by reducing skewness and mitigating the influence of outliers, results in a loss of direct interpretability of the coefficients. Further studies are needed to quantify the extent of improvement achieved due to these transformations and determine whether such transformations are necessary for predictive performance. If the improvement in predictive accuracy and model stability outweighs the loss in interpretability, transformations may be deemed a worthwhile step. However, if the benefits are marginal, it may be preferable to maintain raw variable forms to preserve the clarity and interpretability of the model. This evaluation would be essential to balance the trade-offs between model performance and practical usability.

References

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

Cook, J. W., Spring, B., McChargue, D., & Doran, N. (2010). Effects of anhedonia on days to relapse among smokers with a history of depression: A brief report. Nicotine & Tobacco Research, 12(9), 978–982. https://doi.org/10.1093/ntr/ntq118

Evins, A. E., Benowitz, N. L., West, R., Russ, C., McRae, T., Lawrence, D., Krishen, A., St Aubin, L., Maravic, M. C., & Anthenelli, R. M. (2019). Neuropsychiatric safety and efficacy of varenicline, bupropion, and nicotine patch in smokers with psychotic, anxiety, and mood disorders in the EAGLES trial. Journal of Clinical Psychopharmacology, 39(2), 108–116. https://doi.org/10.1097/jcp.0000000000001015

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
# Data manipulation
library(dplyr)
library(tidyr)
library(mice)
library(caret)

# Tables
#library(kableExtra)
library(gtsummary)
library(gt)

# Plots
library(ggplot2)
library(ggpubr)
library(corrplot)
library(moments)
library(predtools)
library(pROC)
library(gridExtra)
library(ggExtra)


# Models
library(glmnet)
library(boot)
library(ISLR)
library(ROCR)
library(pROC)
library(caret)
project2 <- read.csv("~/Desktop/PHP 2550/Data/project2.csv")

# Modify variables
## Binary variables
project2$sex_ps <- ifelse(project2$sex_ps == 1, 0, 1)
binary_vars <- sapply(project2, function(x) {
  unique_vals <- unique(x[!is.na(x)])
  (all(unique_vals %in% c(0, 1)) | all(unique_vals %in% c(1, 2))) & is.numeric(x)
})
project2[binary_vars] <- lapply(project2[binary_vars], function(x) as.factor(x))

## Ordinal variabl
project2 <- project2 %>%
  mutate(edu = factor(edu, levels = 1:5,
                 labels = c("Grade school", "Some high school", "High school graduate or GED",
                            "Some college/technical school", "College graduate"), ordered = TRUE),
         inc = factor(inc, levels = 1:5,
                 labels = c("Less than $20,000", "$20,000-35,000", "$35,001-50,000",
                            "$50,001-75,000", "More than $75,000"), ordered = TRUE))

# Categorical variables
```

```
project2 <-  project2 %>%
  mutate(Group = case_when(Var == 0 & BA == 0 ~ "ST + placebo", Var == 0 & BA == 1 ~ "BASC + placebo",
                           Var == 1 & BA == 0 ~ "ST + varenicline", Var == 1 & BA == 1 ~ "BASC + vareni
         Race = case_when(Black == 1 & Hisp == 0 & NHW == 0 ~ "Balck",
                          Hisp == 1 & Black == 0 & NHW == 0 ~ "Hispanic",
                          NHW == 1 & Black == 0 & Hisp == 0 ~ "White",
                          Black == 1 & Hisp == 1 & NHW == 0 ~ "Balck and Hispanic")) %>%
  mutate(Group = factor(Group, levels = c("ST + placebo", "BASC + placebo", "ST + varenicline", "BASC +
         Race = factor(Race, levels = c("Balck", "Hispanic", "White", "Balck and Hispanic"))) %>%
  mutate(
    abst = factor(abst, levels = c(0, 1), labels = c("No", "Yes")),
    ftcd.5.mins = factor(ftcd.5.mins, levels = c(0, 1), labels = c("No", "Yes")),
    otherdiag = factor(otherdiag, levels = c(0, 1), labels = c("No", "Yes")),
    antidepmed = factor(antidepmed, levels = c(0, 1), labels = c("No", "Yes")),
    mde_curr = factor(mde_curr, levels = c(0, 1), labels = c("No", "Yes")),
    Only.Menthol = factor(Only.Menthol, levels = c(0, 1), labels = c("No", "Yes")),
    sex_ps = factor(sex_ps, levels = c(0, 1), labels = c("No", "Yes"))
  )
project2 %>%
  dplyr::select(-c(id, Var, BA, Black, Hisp, NHW)) %>%
  tbl_summary(
    by = Group,                            # Group by the treatment groups
    label = list(abst ~ "Smoking Abstinence",
                 age_ps ~ "Age",
                 sex_ps ~ "Sex (% female)",
                 inc ~ "Income /yr",
                 edu ~ "Education",
                 ftcd_score ~ "FTCD score",
                 ftcd.5.mins ~ "Smoking with 5 mins of waking up",
                 cpd_ps ~ "Cigarettes /day",
                 crv_total_pq1 ~ "Cigarette reward value",
                 hedonsum_n_pq1 ~ "Substitute reinforcers",
                 hedonsum_y_pq1 ~ "Complementary reinforcers",
                 shaps_score_pq1 ~ "Anhedonia",
                 otherdiag ~ "Lifetime DSM-5 diagnosis",
                 antidepmed ~ "Taking antidepressant medication",
                 mde_curr ~ "Current vs past MDD",
                 NMR ~ "Nicotine Metabolism Ratio",
                 Only.Menthol ~ "Exclusive Mentholated Cigarette User",
                 readiness ~"Readiness to quit smoking",
                 bdi_score_w00 ~ "BDI score"),
    type = list(readiness ~ "continuous"),
    statistic = list(all_continuous() ~ "{mean} ({sd})",
                     all_categorical() ~ "{n} ({p}%)"),
    missing = "ifany",
              missing_text = "Missing",
    digits = all_continuous() ~ 1
  ) %>%
  add_overall(last=TRUE) %>%
  modify_header(label = "**Characteristic**") %>%
  bold_labels() %>%
  as_gt() %>%
  tab_header(
```

```r
    title = md("Table 1: Baseline Characteristics by Treatment Group")) %>%
  tab_options(
    table.font.size = px(8),
    heading.title.font.size = px(8)
  ) %>%
  cols_width(
        vars(label) ~ px(170),
    everything() ~ px(100),
    starts_with("stat_") ~ px(120)
  ) %>%
    tab_style(
      style = cell_text(weight = "bold", align = "center"),
      locations = cells_column_labels(everything())
    )
# Apply log/sqrt transformations to the specified variables
project2_trans <- project2 %>%
  mutate(shaps_score_pq1_log = log(shaps_score_pq1+1),
    hedonsum_n_pq1_sqrt = sqrt(hedonsum_n_pq1),
         hedonsum_y_pq1_sqrt = sqrt(hedonsum_y_pq1),
         NMR_log = log(NMR))

vars_trans <- list("shaps_score_pq1", "hedonsum_n_pq1", "hedonsum_y_pq1", "NMR",
                   "shaps_score_pq1_log", "hedonsum_n_pq1_sqrt", "hedonsum_y_pq1_sqrt", "NMR_log")

vars_name_trans <- list("shaps_score_pq1" = "Anhedonia",
                        "hedonsum_n_pq1" = "Substitute Reinforcers",
                        "hedonsum_y_pq1" = "Complementary Reinforcers",
                        "NMR" = "Nicotine Metabolism Ratio",
                        "shaps_score_pq1_log" = "log (Anhedonia)",
                        "hedonsum_n_pq1_sqrt" = "sqrt (Substitute Reinforcers)",
                        "hedonsum_y_pq1_sqrt" = "sqrt (Complementary Reinforcers)",
                        "NMR_log" = "log (Nicotine Metabolism Ratio)")

plot_list <- list()
for (var in vars_trans) {
  # Set bin specifications
  skew_score <- round(skewness(project2_trans[[var]], na.rm = TRUE), 2)
  if (var == "shaps_score_pq1_log") {
    plot <- ggplot(project2_trans, aes_string(x = var)) +
      geom_histogram(color = "black", fill = "lightblue", breaks = seq(0, max(3.5, na.rm = TRUE), by = 0
      labs(x = vars_name_trans[[var]], y = "Frequency")
  } else {
    plot <- ggplot(project2_trans, aes_string(x = var)) +
      geom_histogram(color = "black", fill = "lightblue", bins = 15) +
      labs(x = vars_name_trans[[var]], y = "Frequency")
  }

  plot <- plot +
    annotate("text", x = Inf, y = Inf, label = paste("Skewness:", skew_score),
             hjust = 1.1, vjust = 1.5, size = 3, color = "blue") +
    theme_minimal() +
    theme(
      plot.title = element_text(size = 8),
      axis.title = element_text(size = 8),
      axis.text = element_text(size = 8))
  plot_list[[var]] <- plot
}

plot_trans <- ggarrange(plot_list[["shaps_score_pq1"]], plot_list[["shaps_score_pq1_log"]],
                        plot_list[["hedonsum_n_pq1"]],
                        plot_list[["hedonsum_n_pq1_sqrt"]],
                        plot_list[["hedonsum_y_pq1"]],
                        plot_list[["hedonsum_y_pq1_sqrt"]],
                        plot_list[["NMR"]], plot_list[["NMR_log"]],
                        ncol=2, nrow=4)
annotate_figure(plot_trans,
                top = text_grob("Figure 1: Variable Distribution Before and After Transformations", fac
# Missing Data Imputation
project2_s <- project2 %>%
  dplyr::select(-c(id, Group, Race))
project2_imp <- mice(project2_s, m = 5, method = 'pmm', seed = 2550, printFlag = FALSE)
project2_imp_trans <- list()

for (i in 1:5) {
  # Extract each imputed dataset
  imputed_data <- complete(project2_imp, action = i)

  # Apply transformations
  imputed_data <- imputed_data %>%
    mutate(
      shaps_score_pq1_log = log(shaps_score_pq1+1),
      hedonsum_n_pq1_sqrt = sqrt(hedonsum_n_pq1),
      hedonsum_y_pq1_sqrt = sqrt(hedonsum_y_pq1),
      NMR_log = log(NMR)) %>%
    mutate(Group = case_when(Var == 0 & BA == 0 ~ "ST + placebo", Var == 0 & BA == 1 ~ "BASC + placebo"
                           Var == 1 & BA == 0 ~ "ST + varenicline", Var == 1 & BA == 1 ~ "BASC + varenic
         Race = case_when(Black == 1 & Hisp == 0 & NHW == 0 ~ "Balck",
                          Hisp == 1 & Black == 0 & NHW == 0 ~ "Hispanic",
                          NHW == 1 & Black == 0 & Hisp == 0 ~ "Non-hispanic White",
                          Black == 1 & Hisp == 1 & NHW == 0 ~ "Other",
                          Black == 0 & Hisp == 0 & NHW == 0 ~ "Other")) %>%
    mutate(Group = factor(Group, levels = c("ST + placebo", "BASC + placebo", "ST + varenicline", "BASC
           edu = factor(case_when(edu=="Grade school" | edu=="Some high school"~ "High school and below
                    edu=="High school graduate or GED"~ "High school graduate or GED",
                    edu=="Some college/technical school"~ "Some college/technical school",
                    edu=="College graduate"~ "College graduate")),
    inc = factor(case_when(inc=="Less than $20,000" ~ "Less than $20,000",
                           inc=="$20,000-35,000" ~ "$20,000-35,000",
                           inc=="$35,001-50,000" ~ "$35,001-50,000",
                           inc=="$50,001-75,000" | inc=="More than $75,000" ~ "$More than $50,000"))
    ) %>%
  mutate(edu = factor(edu, levels = c("High school and below", "High school graduate or GED", "Some coll
                      Race = factor(Race, levels = c("Balck", "Hispanic", "Non-hispanic White", "Other"
                      inc = factor(inc, levels = c("Less than $20,000", "$20,000-35,000", "$35,001-50,0
    dplyr::select(-c(shaps_score_pq1, hedonsum_n_pq1, hedonsum_y_pq1, NMR, Black, Hisp, NHW))
  # Store the transformed dataset
  project2_imp_trans[[i]] <- imputed_data
}
# Prepare data by selecting binary variables and reshaping to long format
project2_bi_summary <- na.omit(project2) %>%
  dplyr::select(abst, ftcd.5.mins, otherdiag, antidepmed, mde_curr, Only.Menthol, Group) %>%
  rename(`Other lifetime DSM-5 diagnosis` = otherdiag,
         `Smoking with 5 mins of waking up` = ftcd.5.mins,
         `Antidepressant medication` = antidepmed,
         `Current vs past MDD` = mde_curr,
         `Mentholated Cigarette User` = Only.Menthol) %>%
  pivot_longer(
    cols = -c(abst, Group),
    names_to = "Variable",
    values_to = "Value"
  ) %>%
  group_by(Group, Variable, Value) %>%
  summarize(sum_abst = sum(abst=="Yes"), .groups = "drop") %>%
  arrange(Group, Variable, Value) %>%
  mutate(Variable = ifelse(Variable == "Other lifetime DSM-5 diagnosis", "Other lifetime DSM-5", Variabl
         Group = ifelse(Group %in% c("BASC + placebo", "BASC + varenicline"), "BASC", "ST")) %>%
  group_by(Group, Variable, Value) %>%  # Group by Group, Variable, and Value
  summarize(sum_abst = sum(sum_abst), .groups = "drop") %>%  # Summarize to remove duplicates
  group_by(Group, Variable) %>%  # Regroup by Group and Variable
  mutate(Proportion = sum_abst / sum(sum_abst))

# Create the plot
ggplot(project2_bi_summary, aes(x = Group, y = sum_abst, fill = Group, alpha = Value)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ Variable, strip.position = "bottom") +  # Separate plots for each variable with facet la
  theme_minimal() +
  labs(
    x = "Group",
    y = "Count",  # Update y-axis label to reflect counts
    title = "Figure 2: Number of Smoking Abstinence by Group",
    fill = "Behavioral activation",
    alpha = "Abstinence"
  ) +
  geom_text(
    aes(
      label = scales::percent(Proportion, accuracy = 1)  # Use pre-computed Proportion
    ),
    position = position_stack(vjust = 0.5),  # Position text in the middle of the bar
    size = 4,
    color = "black",
    fontface = "bold"
  ) +
  scale_alpha_manual(
    values = c("No" = 0.4, "Yes" = 1),
    guide = guide_legend(override.aes = list(alpha = c(0.4, 1)))
  ) +
  theme(
    title = element_text(size = 8, face = "bold"),
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),  # Remove x-axis text
    axis.ticks.x = element_blank(),  # Remove x-axis ticks
    axis.title.y = element_text(size = 8, face = "bold"),
    axis.text.y = element_text(size = 8),
    strip.text = element_text(size = 6, face = "bold"),
    legend.position = "bottom",
    legend.title = element_text(size = 8, face = "bold"),
    legend.text = element_text(size = 6)
  )
# Create summary table for count and percentage of smoking abstinence by Education Level and Income Lev

abstinence_by_edu_group <- project2 %>%
  group_by(edu, Group) %>%
  # Calculate count of abstinent individuals and percentage within each group
  summarize(
    Count_Abstinent = sum(abst == "Yes", na.rm = TRUE),
    Proportion_Abstinent = round(sum(abst == "Yes", na.rm = TRUE)/sum(abst == "Yes" | abst == "No", na.
    .groups = 'drop'
  ) %>%
  mutate(Proportion_Abstinent = paste0(Count_Abstinent, " (", Proportion_Abstinent, "%)")) %>%
  dplyr::select(-Count_Abstinent) %>%
  pivot_wider(names_from = Group, values_from = Proportion_Abstinent, values_fill = "0 (0%)") %>%
  rename(Variables = edu) %>%
  dplyr::select(Variables, `ST + placebo`, `BASC + placebo`, `ST + varenicline`, `BASC + varenicline`)
  mutate(across(c(`ST + placebo`, `BASC + placebo`,`ST + varenicline`, `BASC + varenicline`),
                ~ ifelse(. == "0 (0%)", "-", .))) %>%
  as.data.frame()

header_row <- as.data.frame(matrix(" ", nrow = 1, ncol = ncol(abstinence_by_edu_group)))
colnames(header_row) <- colnames(abstinence_by_edu_group)
header_row$Variables <- "Education Level"
abstinence_by_edu_group <- rbind(header_row, abstinence_by_edu_group)

abstinence_by_inc_group <- na.omit(project2) %>%
  mutate(inc = factor(case_when(inc=="Less than $20,000" ~ "Less than $20,000",
                           inc=="$20,000-35,000" ~ "$20,000-35,000",
                           inc=="$35,001-50,000" ~ "$35,001-50,000",
                           inc=="$50,001-75,000" | inc=="More than $75,000" ~ "$More than $50,000")))
  mutate(inc = factor(inc, levels = c("Less than $20,000", "$20,000-35,000", "$35,001-50,000", "$More t
  group_by(inc, Group) %>%
  # Calculate count of abstinent individuals and percentage within each group
  summarize(
    Count_Abstinent = sum(abst == "Yes", na.rm = TRUE),
    Proportion_Abstinent = round(mean(abst == "Yes", na.rm = TRUE) * 100, 0),
    .groups = 'drop'
  ) %>%
  mutate(Proportion_Abstinent = paste0(Count_Abstinent, " (", Proportion_Abstinent, "%)")) %>%
  dplyr::select(-Count_Abstinent) %>%
  pivot_wider(names_from = Group, values_from = Proportion_Abstinent, values_fill = "0 (0%)") %>%
  rename(Variables = inc) %>%
  dplyr::select(Variables, `ST + placebo`, `BASC + placebo`, `ST + varenicline`, `BASC + varenicline`)
  mutate(across(c( `ST + placebo`, `BASC + placebo`,`ST + varenicline`, `BASC + varenicline`),
                ~ ifelse(. == "0 (0%)", "-", .))) %>%
  as.data.frame()
header_row <- as.data.frame(matrix(" ", nrow = 1, ncol = ncol(abstinence_by_inc_group)))
colnames(header_row) <- colnames(abstinence_by_inc_group)
header_row$Variables <- "Income Level"
abstinence_by_inc_group <- rbind(header_row, abstinence_by_inc_group)

# Combine tables
combined_table <- rbind(abstinence_by_edu_group, abstinence_by_inc_group)

# Format combined table
combined_table %>%
  gt() %>%
  tab_header(title = md("Table 2: Number and proportion of smokers who have achieved cessation")) %>%
  tab_options(
    table.font.size = px(8),
    heading.title.font.size = px(8)
  ) %>%
  cols_width(
    Variables ~ px(120),
    everything() ~ px(60)

  ) %>%
tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_body(
      columns = vars(Variables),
      rows = Variables %in% c("Education Level", "Income Level")
    )
  )
# Create summary table for continuous variables stratified by those who achieved smoking abstinence and
project2_con_summary <- project2 %>%
  dplyr::select(abst, Group, ftcd_score, cpd_ps, crv_total_pq1,
                hedonsum_n_pq1, hedonsum_y_pq1, shaps_score_pq1,
                NMR, readiness) %>%
  rename(`FTCD score` = ftcd_score,
         `Cigarettes per day` = cpd_ps,
         `Cigarette reward value` = crv_total_pq1,
         `Pleasurable Events Scale at baseline-substitute reinforcers` = hedonsum_n_pq1,
         `Pleasurable Events Scale at baseline-complementary reinforcers` = hedonsum_y_pq1,
         `Anhedonia` = shaps_score_pq1,
         `Nicotine Metabolism Ratio` = NMR,
         `Readiness to quit smoking` = readiness) %>%
  pivot_longer(
    cols = -c(abst, Group),
    names_to = "Variable",
    values_to = "Value"
  ) %>%
  group_by(Group, abst, Variable) %>%
  summarize(
    mean_value = mean(Value, na.rm = TRUE),
    se_value = sd(Value, na.rm = TRUE) / sqrt(n()),
    .groups = "drop"
  ) %>%
  mutate(mean_se = paste0(round(mean_value, 2), " (", round(se_value, 2), ")")) %>%
  select(Group, abst, Variable, mean_se) %>%
  pivot_wider(
    names_from = c(Group, abst),
    values_from = mean_se,
    names_glue = "{Group}_{abst}"
  )

# Format the summary table
project2_con_summary %>%
  gt(rowname_col = "Variable") %>%
  tab_header(title = "Table 3: Summary of Continuous Variables by Group and Smoking Abstinence") %>%
  cols_label(
    `ST + placebo_No` = "No",
    `ST + placebo_Yes` = "Yes",
    `BASC + placebo_No` = "No",
    `BASC + placebo_Yes` = "Yes",
    `ST + varenicline_No` = "No",
    `ST + varenicline_Yes` = "Yes",
    `BASC + varenicline_No` = "No",
    `BASC + varenicline_Yes` = "Yes"
  ) %>%
  tab_spanner(
    label = "ST + placebo",
    columns = c(`ST + placebo_No`, `ST + placebo_Yes`)
  ) %>%
  tab_spanner(
    label = "BASC + placebo",
    columns = c(`BASC + placebo_No`, `BASC + placebo_Yes`)
  ) %>%
  tab_spanner(
    label = "ST + varenicline",
    columns = c(`ST + varenicline_No`, `ST + varenicline_Yes`)
  ) %>%
  tab_spanner(
    label = "BASC + varenicline",
    columns = c(`BASC + varenicline_No`, `BASC + varenicline_Yes`)
  ) %>%
  tab_options(
    table.font.size = px(8),
    heading.title.font.size = px(8)
  ) %>%
  cols_width(
    Variable ~ px(200),
    everything() ~ px(50)
  ) %>%
  tab_style(
    style = cell_text(weight = "bold", align = "center"),
    locations = cells_column_labels(everything())
  )
project2_cor <- project2 %>%
  dplyr::select(bdi_score_w00, ftcd_score, cpd_ps, crv_total_pq1,
                hedonsum_n_pq1, hedonsum_y_pq1, shaps_score_pq1,
                NMR, readiness) %>%
  rename(`BDI score` = `bdi_score_w00`,
         `FTCD score` = ftcd_score,
         `Cigarettes per day` = cpd_ps,
         `Cigarette reward value` = crv_total_pq1,
         `Pleasurable Events Scale (substitute)` = hedonsum_n_pq1,
         `Pleasurable Events Scale (complementary)` = hedonsum_y_pq1,
         `Anhedonia` = shaps_score_pq1,
         `Nicotine Metabolism Ratio` = NMR,
         `Readiness to quit smoking` = readiness)

cor_matrix <- cor(project2_cor, use = "complete.obs")

corrplot(cor_matrix, method = "color", type = "lower",
         tl.col = "black", tl.cex = 0.5, addCoef.col = "black",
         number.cex = 0.7, col = colorRampPalette(c("steelblue", "white", "steelblue"))(200))
title("Figure 3: Correlation Plot among Continuous Clinical Variables",
      cex.main = 0.9, line = 0)
perform_cv_lasso_mod <- function(data, seed = 2550, bootstrap_iterations = 10) {
  # Initialize lists to store results
  best_lambdas <- list()
  coef_list <- list()
  auc_list <- list()
  roc_plots <- list()
  bootstrap_results <- list()
  train_data_full <- NULL
  test_data_full <- NULL

  # Loop through each imputed dataset
  for (i in 1:5) {
    # Complete the imputed dataset
    project2_imputed_s <- data[[i]]

    # Perform bootstrap iterations
    for (b in 1:bootstrap_iterations) {
      # Generate bootstrap sample
      set.seed(seed + b)  # Ensure reproducibility for each bootstrap iteration
      bootstrap_sample <- project2_imputed_s[sample(1:nrow(project2_imputed_s), replace = TRUE), ]

      # Split bootstrapped dataset into training and testing sets
      set.seed(seed + b)  # Reuse the same seed for consistency
      train_index <- createDataPartition(bootstrap_sample$Group, p = 0.7, list = FALSE)
      train_data <- bootstrap_sample[train_index, ]
      test_data <- bootstrap_sample[-train_index, ]

      # Combine train and test datasets
      train_data_full <- bind_rows(train_data_full, train_data)
      test_data_full <- bind_rows(test_data_full, test_data)

      # Assign folds for cross-validation in training data
      train_data$foldid <- NA
      for (group in unique(train_data$Group)) {
        group_data <- train_data[train_data$Group == group, ]
        fold_idex <- sample(rep(1:10, length.out = nrow(group_data)))
        train_data$foldid[train_data$Group == group] <- fold_idex
      }

      # Create model matrices for training data
      x_mat <- model.matrix(
        abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins + bdi_score_w00 +
                       cpd_ps + crv_total_pq1 + hedonsum_n_pq1_sqrt + hedonsum_y_pq1_sqrt +
                       shaps_score_pq1_log + otherdiag + antidepmed + mde_curr + NMR_log +
                       Only.Menthol + readiness + Race) +
          Var * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins + bdi_score_w00 +
                   cpd_ps + crv_total_pq1 + hedonsum_n_pq1_sqrt + hedonsum_y_pq1_sqrt +
                   shaps_score_pq1_log + otherdiag + antidepmed + mde_curr + NMR_log +
                   Only.Menthol + readiness + Race),
        data = train_data
      )[, -1]
      y <- train_data$abst

      # Perform LASSO cross-validation to find the best lambda
      lasso_model_cv <- cv.glmnet(
        x_mat, y, alpha = 1, nfolds = 10, foldid = train_data$foldid,
        family = "binomial"
      )

      # Fit LASSO model with the optimal lambda
      best_lambda <- lasso_model_cv$lambda.min
      best_lambdas[[paste(i, b)]] <- best_lambda  # Store the best lambda for this iteration
      lasso_model <- glmnet(
        x_mat, y, alpha = 1, lambda = best_lambda, family = "binomial"
      )

      # Store the coefficients with names
      coef_list[[paste(i, b)]] <- as.matrix(coef(lasso_model))[, , drop = FALSE]

      # Evaluate the model on the test set
      x_mat_test <- model.matrix(
        abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins + bdi_score_w00 +
                       cpd_ps + crv_total_pq1 + hedonsum_n_pq1_sqrt + hedonsum_y_pq1_sqrt +
                       shaps_score_pq1_log + otherdiag + antidepmed + mde_curr + NMR_log +
                       Only.Menthol + readiness + Race) +
          Var * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins + bdi_score_w00 +
                   cpd_ps + crv_total_pq1 + hedonsum_n_pq1_sqrt + hedonsum_y_pq1_sqrt +
                   shaps_score_pq1_log + otherdiag + antidepmed + mde_curr + NMR_log +
                   Only.Menthol + readiness + Race),
        data = test_data
      )[, -1]
      y_test <- test_data$abst

      # Predict probabilities on the test data
      test_predictions <- predict(lasso_model, newx = x_mat_test, type = "response")

      # Calculate AUC
      test_roc <- roc(y_test, as.vector(test_predictions))
```

```
      test_auc <- auc(test_roc)
      auc_list[[paste(i, b)]] <- test_auc

      # Store bootstrap results
      bootstrap_results[[paste(i, b)]] <- list(
        best_lambda = best_lambda,
        coefficients = coef_list[[paste(i, b)]],
        auc = test_auc
      )
    }
  }

  # Combine the coefficient lists into a matrix and calculate the average coefficients
  coef_matrix <- do.call(cbind, coef_list)  # Combine list of named vectors into a matrix
  avg_coef <- rowMeans(coef_matrix, na.rm = TRUE)

  # Return results as a list
  list(
    best_lambdas = best_lambdas,
    coef_list = coef_list,
    auc_list = auc_list,
    bootstrap_results = bootstrap_results,
    avg_coef = avg_coef,
    train_data_full = train_data_full,  # Return the combined dataset
    test_data_full = test_data_full
  )
}

# Run the function with bootstrap
results_mod <- perform_cv_lasso_mod(data = project2_imp_trans, seed = 1234, bootstrap_iterations = 10)
# Create a summary table for coefficients across imputed datasets
create_summary_table <- function(coef_list) {
  # Combine coefficients across imputations
  coef_matrix <- do.call(cbind, coef_list)

  # Calculate the average coefficient for each covariate
  avg_coeff <- rowMeans(coef_matrix, na.rm = TRUE)

  # Calculate the proportion of non-zero coefficients across imputations
  non_zero_counts <- rowSums(coef_matrix != 0, na.rm = TRUE)
  proportion_non_zero <- non_zero_counts / ncol(coef_matrix)

  coef_matrix_exp <- exp(coef_matrix)
  avg_coeff_exp <- exp(avg_coeff)

  # Create a summary dataframe
  summary_table <- data.frame(
    Covariate = rownames(coef_matrix),
    `Imputation 1` = coef_matrix_exp[, 1],
    `Imputation 2` = coef_matrix_exp[, 2],
    `Imputation 3` = coef_matrix_exp[, 3],
    `Imputation 4` = coef_matrix_exp[, 4],
    `Imputation 5` = coef_matrix_exp[, 5],
    Average = avg_coeff_exp,
    `Proportion Non-Zero` = proportion_non_zero
  )

  # Round the values for better readability
  summary_table <- summary_table %>%
    mutate(across(-Covariate, ~ round(., 4)))

  return(summary_table)
}

# Assuming `coef_list` contains the list of coefficient matrices for all imputations
summary_table <- create_summary_table(results_mod$coef_list)
summary_table %>%
  filter(Covariate == "BA1" | Covariate == "Var1" | Proportion.Non.Zero > 0.6) %>%
  mutate(across(starts_with("Imputation"),
                ~ ifelse(. == 1, "-", sprintf("%.4f", .)))) %>%
  #mutate(Proportion_Non_Zero = sprintf("%.2f", Proportion_Non_Zero)) %>%
  gt() %>%
  cols_label(
    `Imputation.1` = "Imputation 1",
    `Imputation.2` = "Imputation 2",
    `Imputation.3` = "Imputation 3",
    `Imputation.4` = "Imputation 4",
    `Imputation.5` = "Imputation 5",
    Proportion.Non.Zero = "Proportion Non-Zero"
  ) %>%
  tab_header(
    title = "Table 4: Summary of Non-Zero Coefficient Estimates in Exponential Scale"
  ) %>%
  tab_style(
    style = list(
      cell_text(weight = "bold")
    ),
    locations = cells_column_labels()
  ) %>%
  tab_options(
    table.font.size = px(10),
    heading.title.font.size = px(10)
  ) %>%
  cols_width(
    Covariate ~ px(150),
    everything() ~ px(60)
  ) %>%
    tab_style(
      style = cell_text(weight = "bold", align = "center"),
      locations = cells_column_labels(everything())
    )
# Plot ROC and AUC
# 1. Prepare model matrices for train and test data
x_mat_train_full <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score
                                                 + ftcd.5.mins + bdi_score_w00 + cpd_ps +
                                                   crv_total_pq1 + hedonsum_n_pq1_sqrt +
                                                   hedonsum_y_pq1_sqrt +
                                                   shaps_score_pq1_log + otherdiag +
                                                   antidepmed + mde_curr + NMR_log +
                                                   Only.Menthol + readiness + Race) + Var * (age_ps + s
                                 data = results_mod$train_data_full)

x_mat_test_full <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
                                           bdi_score_w00 + cpd_ps + crv_total_pq1 +
                                           hedonsum_n_pq1_sqrt + hedonsum_y_pq1_sqrt +
                                           shaps_score_pq1_log + otherdiag + antidepmed +
                                           mde_curr + NMR_log + Only.Menthol + readiness + Race) + Var
                           data = results_mod$test_data_full)

# 2. Calculate predictions for training and testing data using avg_coef
train_linear_predictor <- x_mat_train_full %*% log(summary_table$Average)
test_linear_predictor <- x_mat_test_full %*% log(summary_table$Average)

train_predicted_prob <- 1 / (1 + exp(-train_linear_predictor))
test_predicted_prob <- 1 / (1 + exp(-test_linear_predictor))

# 3: Calculate AUC and ROC
train_roc <- roc(results_mod$train_data_full$abst, as.vector(train_predicted_prob))
test_roc <- roc(results_mod$test_data_full$abst, as.vector(test_predicted_prob))
train_auc <- auc(train_roc)
test_auc <- auc(test_roc)


# Plot ROC for Train and Test Data
roc_plots <- ggplot() +
  geom_line(data = data.frame(specificity = rev(train_roc$specificities),
                              sensitivity = rev(train_roc$sensitivities)),
            aes(x = 1 - specificity, y = sensitivity, color = "Training"), size = 1) +
  geom_line(data = data.frame(specificity = rev(test_roc$specificities),
                              sensitivity = rev(test_roc$sensitivities)),
            aes(x = 1 - specificity, y = sensitivity, color = "Testing"), size = 1) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "black") +
  scale_color_manual(values = c("Training" = "blue", "Testing" = "red"),
                     labels = c(paste("Training (AUC =", round(train_auc, 2), ")"),
                                paste("Testing (AUC =", round(test_auc, 2), ")"))) +
  labs(
       x = "False Positive Rate",
       y = "True Positive Rate") +
  theme_minimal() +
    theme(
      axis.title = element_text(size = 8),
      axis.text = element_text(size = 8)) +
  theme(legend.position = c(0.8, 0.15))

annotate_figure(roc_plots,
                top = text_grob("Figure 4: ROC Curve Using Average Coefficients", face = "bold", size =
# ACC
train_predictions_binary <- ifelse(train_predicted_prob >= 0.5, "Yes", "No")
train_acc <- mean(train_predictions_binary == results_mod$train_data_full$abst)
```

```r
test_predictions_binary <- ifelse(test_predicted_prob >= 0.5, "Yes", "No")
test_acc <- mean(test_predictions_binary == results_mod$test_data_full$abst)

# Sensitivity and Specificity
train_conf_matrix <- table(train_predictions_binary, results_mod$train_data_full$abst)
train_sens <- sensitivity(train_conf_matrix)
train_spec <- specificity(train_conf_matrix)

test_conf_matrix <- table(test_predictions_binary, results_mod$test_data_full$abst)
test_sens <- sensitivity(test_conf_matrix)
test_spec <- specificity(test_conf_matrix)

# 5. Create a Summary Table
summary_table <- data.frame(
  Metric = c("AUC", "Accuracy", "Sensitivity", "Specificity"),
  Train = c(round(train_auc, 2), round(train_acc, 2), round(train_sens, 2), round(train_spec, 2)),
  Test = c(round(test_auc, 2), round(test_acc, 2), round(test_sens, 2), round(test_spec, 2))
)

summary_table %>%
  gt() %>%
  tab_header(
    title = "Table 5: Performance of the FInal Model"
  ) %>%
  tab_style(
    style = list(
      cell_text(weight = "bold")
    ),
    locations = cells_column_labels()
  ) %>%
  tab_options(
    table.font.size = px(10),
    heading.title.font.size = px(10)
  ) %>%
  cols_width(
    everything() ~ px(80)
  ) %>%
    tab_style(
      style = cell_text(weight = "bold", align = "center"),
      locations = cells_column_labels(everything())
    )
# Create a data frame for calibration plot
calibration_data_train <- results_mod$train_data_full %>%
  mutate(abst = ifelse(abst=="Yes", 1, 0),
    pred = train_predicted_prob)
calibration_data_test <- results_mod$test_data_full %>%
  mutate(abst = ifelse(abst=="Yes", 1, 0),
         pred = test_predicted_prob)

calibration_plot_train <- calibration_plot(data = calibration_data_train, obs = "abst", pred = "pred")$
  annotate("text", x = -Inf, y = Inf, label = "(a). Training", hjust = -0.5, vjust = 1, size = 3, fontfa
  xlim(0, 0.5) +
  ylim(0, 0.8) +
  theme_minimal() +
    theme(
      axis.title = element_text(size = 8),
      axis.text = element_text(size = 8))

calibration_plot_test <- calibration_plot(data = calibration_data_test, obs = "abst", pred = "pred")$cal
  annotate("text", x = -Inf, y = Inf, label = "(b). Testing", hjust = -0.5, vjust = 1, size = 3, fontfac
  xlim(0, 0.5) +
  ylim(0, 0.8) +
  theme_minimal() +
    theme(
      axis.title = element_text(size = 8),
      axis.text = element_text(size = 8))

# Arrange plots
grid.arrange(calibration_plot_train, calibration_plot_test, ncol = 2) %>%
  annotate_figure(top = text_grob("Figure 5: Calibration Plots", face = "bold", size = 10))
```