Influence of Baseline Characteristics on Smoking Cessation in MDD: A Study of Behavioral and Pharmacological Treatment Effects

Yingxi Kong

Abstract

Background: People with Major Depressive Disorder (MDD) are more likely to engage in tobacco use. However, most smoking cessation clinical trials had excluded this group from the enrollment, limiting available information to support smoking cessation for this group. This study, collaborating with Dr. George Papandonatos, aims to evaluates behavioral and pharmacological treatment for smoking cessation in adults with MDD to identify baseline characteristics that may moderate the behavioral treatment or predict smoking cessation outcome.

Methods: In a 2 × 2 factorial, randomized, placebo-controlled trial, 300 adult smokers with current or past MDD were randomly assigned to either Behavioral Activation for Smoking Cessation (BASC) or standard treatment (ST), combined with either varenicline or placebo. Participants' demographic characteristics, smoking behavior, mental health measurement, and the abstinence outcome at the week 27 follow-up were collected. Lasso regression was employed to identify significant baseline characteristics and potential interaction effects with treatment, controlling for both behavioral and pharmacotherapy assignments. In addition, bootstrapping is applied to provide a more robust and stable analysis.

Results: Findings suggest that as predictors, controlling for treatment groups and other factors, having higher nicotine metabolism ratio and identifying as Non-Hispanic White were associated with higher likelihood of abstinence, while higher nicotine dependence score and current MDD experience were associated with lower odds of abstinence. Moreover, income, menthol cigarette users indicators, and nicotine dependece score serve as moderators of the behavioral treatment effect on the end-of-treatment (EOT) abstinence.

Conclusion: This study provides insights into how smoking cessation treatment outcomes among adults with MDD interact with baseline characteristics, identifying demographic factors, nicotine dependence, and MDD status as key predictors and moderators. However, variation in treatment adherence and several underrepresented groups could have influenced the estimate of treatment effect. Further research should explore strategies to improve engagement and broaden data collection to achieve better representation of these groups, enhancing the accuracy and generalizability of the findings. Further researches focusing on relaxed lasso or L0 + L2 penalty with more bootstrap iterations are also recommended.

Introduction

Major Depressive Disorder (MDD) has been one of the most prevalent mental health disorders in the world, with rates that have continued to rise, particularly during the COVID-19 pandemic[1]. Individuals with MDD are not only at risk for a range of adverse health outcomes but are also more likely to engage in harmful health behaviors, including tobacco use. The rate of smoking among individuals with MDD is 2–3 times higher than in the general population[2]. However, most smoking cessation clinical trials have excluded this important group from the trial enrollment[3], thereby limiting available information and recommendations to support smoking cessation for this group.

A recent randomized, placebo-controlled trial led by Dr. George Papandonatos evaluated the efficacy and safety of combining behavioral and pharmacological treatment for smoking cessation among individuals with current or past MDD. The study involved 300 participants and employed a 2 × 2 factorial design to compare Behavioral Activation for Smoking Cessation (BASC) with standard treatment (ST), and varenicline with placebo. BASC, a behavioral intervention designed to enhance engagement in rewarding activities and reduce avoidance behavior, was paired with varenicline, a pharmacotherapy shown to reduce cravings and mitigate nicotine's rewarding effects. The results indicated that while varenicline significantly improved abstinence rates compared to placebo, BASC did not outperform ST, suggesting that while pharmacotherapy may provide substantial benefits for smokers with MDD, the behavioral component of cessation treatment may require further refinement.

Collaborating with Dr. George Papandonatos, this study aims to investigate the role of baseline characteristics as potential moderators of the effectiveness of behavioral treatment on end-of-treatment (EOT) abstinence outcomes. Furthermore, we aim to assess these baseline characteristics as predictors of abstinence, while controlling for both behavioral treatment and pharmacotherapy. By identifying factors that may influence the efficacy of cessation interventions, this analysis would contribute to inform targeted treatment strategies to enhance smoking cessation outcomes among individuals with MDD.

Methods

Our sample population consists of 300 adult smokers with or previously with MDD, without psychotic feaures, and are interested in quitting smoking. Initial eligibility screening was completed by telephone. Final eligibility screening, informed consent, treatment randomization and the baseline assessment was completed at the intake session (week 0). Patients were randomly assigned to either behavioral activation for smoking cessation (BASC) or standard behavioral treatment (ST) and either varenicline or placebo groups. That is, participants were assigned to four distinct intervention groups, including `ST + placebo`, `ST + varenicline`, `BASC + placebo`, and `BASC + varenicline.` Randomization was stratified by clinical site, sex, and level of depressive symptoms to ensure balanced representation across these factors.

Follow-up data was collected at week 27 to assess smoking cessation outcomes, along with relevant baseline characteristics. Key variables include smoking abstinence status, demographic characteristics (sex, age, income, and education), smoking behaviors (number of cigarettes per day, time to first cigarette after getting up, and nicotine dependence score), and psychiatric measures (MDD status, anhedonia score, other diagnoses, and antidepressant usage).

Data Preprocessing

To prepare dataset for analysis, we firstly converted all categorical variables into factors. For socioeconomic factors, income and education, we recoded levels to improve readability and interpretability. In addition, we combined race and ethnicity indicators into a single race variable with categories including Black, Hispanic, Non-Hispanic White, Mixed Race, and Unknown.

Table 1: Summary of Missing Data Patterns Across Variables

Variable	Missing Count	Missing Percentage
NMR	21	7 %
crv_total_pq1	18	6 %
readiness	17	5.67 %
inc	3	1 %
shaps_score_pq1	3	1 %
Only.Menthol	2	0.67 %
ftcd_score	1	0.33 %

The dataset contains varying levels of missingness across several variables as presented in `Table 1`. Nicotine Metabolism Ratio (`NMR`) has the highest missing rate, with 7% of observations missing. The FTCD score at baseline (`ftcd_score`) has the lowest missing rate, 0.33%, with only one patient missing this information. Given our limited sample size, we prefer to retain as many observations as possible for our analysis. Therefore, to address the missingness, we applied a multiple imputation approach using the `mice()` function from the `mice` package in R, generating five imputed datasets to provide plausible values for all missing entries before proceeding to the primary analysis.

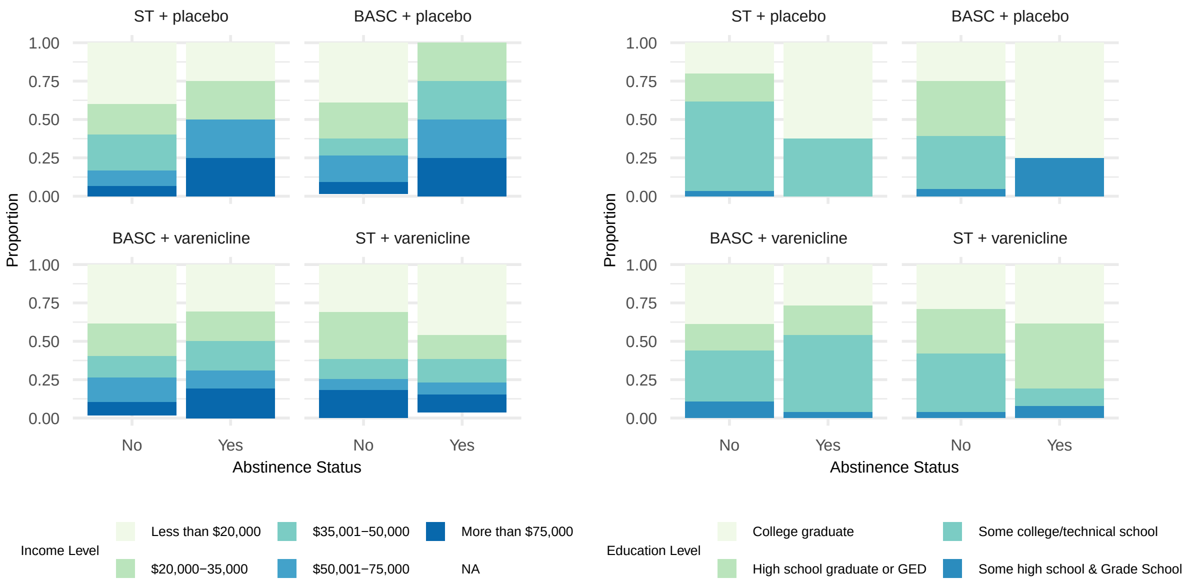
Data Exploration and Transformation

To explore potential interactions between baseline characteristics and treatment assignment on EOT abstinence, we examined the distribution of each baseline variable across treatment groups and abstinence outcomes.

For categorical variables, bar charts show patterns across treatment groups and abstinence groups in `Figure 1` and `Figure 2.` Income in `Figure 1` exhibits different distributions among groups and abstinence outcomes. Within each treatment group, different income levels show various proportions of abstinence, suggesting that income level may be a potential predictor of treatment effect on abstinence.

In the `BASC + placebo` group, most participants with incomes lower than $20,000 did not achieve abstinence at the week 27 follow-up, while nearly half of those in the `ST + placebo` group with similar income level achieved abstinence. In addition, people with income ranging from $35,001 to $50,000 are more likely to quit smoking at the week 27 follow-up in the `BASC + placebo` group while most participants with this income level assigned to the `ST + placebo` group did not achieve abstinence. Similar variations in abstinence outcomes within income levels were observed when comparing the two behavioral treatments combined with varenicline, indicating that income level might be a potential moderator of the behavioral treatment effectiveness on the EOT abstinence among people with MDD.

Figure 1: Baseline Characteristics by Abstinence Status and Treatment Group (Categorical 1)



Similarly, as shown in `Figure 1`, education level appears to impact abstinence rates across treatment groups. Within each treatment group, participants with different education levels demonstrate varying abstinence rates at follow-up. For example, observing the `BASC + placebo` group, college graduated participants are more likely to achieve abstinence at the follow up compared to those with lower education levels, suggesting that income level may be a predictor of treatment effect on abstinence.

Additionally, comparing the two behavioral treatments with varenicline, college graduates in the `ST + varenicline` group show higher abstinence rates at week 27 than those in the `BASC + varenicline` group.

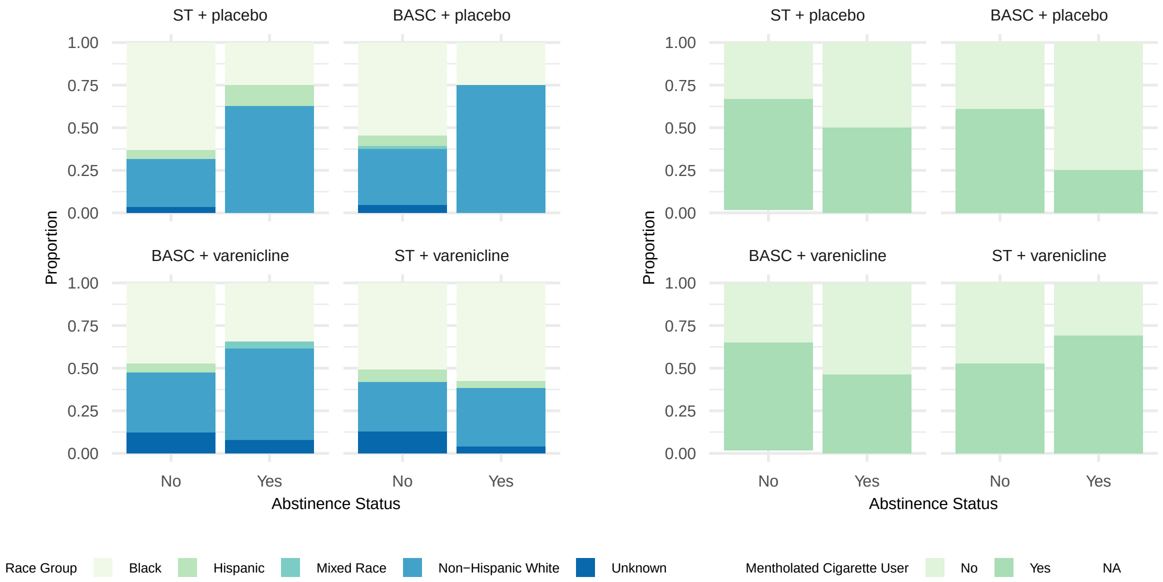
Also, patients with a high school diploma or GED exhibit a higher likelihood of smoking cessation with BASC while their abstinence rate becomes much lower with ST. These findings suggest that education levels moderate the effects of behavioral treatment on EOT abstinence among individuals with MDD.

Race and the indicator of exclusive mentholated cigarette users (`Only.Menthol`) also show differences in distribution across treatment groups and outcome values, as shown in `Figure 2`. For example, in the `ST + placebo`, `BASC + placebo`, and `BASC + varenicline` groups, non-Hispanic White participants are more likely to stop smoking compared to other racial groups. This indicates that race might be a predictor of the treatment effect on EOT abstinence for people with MDD.

Additionally, comparing the two behavioral treatments with placebo, Hispanic participants with ST are more likely to achieve smoking cessation while this pattern reverses when they were assigned to BASC. Moreover, comparing the two varenicline groups, black people with BASC show less likelihood of stopping smoking while they show a larger abstinence rate in the ST group.

Similar pattern observed for the indicator of exclusive mentholated cigarette users (`Only.Menthol`). In the two varenicline groups, mentholated cigarette users (`Only.Menthol` $= 1$) with ST are more likely to achieve abstinence while these users with BASC exhibit much lower abstinence rate at week 27 follow-up. These findings suggest that race and the indicator of exclusive mentholated cigarette users could be potential predictors or moderators of the treatment effects on the EOT abstinence for people with MDD.

Figure 2: Baseline Characteristics by Abstinence Status and Treatment Group (Categorical 2)



We also examine the distribution of continuous variables by treatment groups and outcome status, as shown in `Figure 3` and `Figure 4`. Among continuous variables, age (`age_ps`), FTCD score, NMR, and BDI score (`bdi_score_W00`) exhibited differences in distribution across treatment groups and abstinence status at the week 27 follow-up.

Seeing `Figure 3`, the abstinence rate varies with age within the same treatment group, suggesting age as a predictor of treatment effect on EOT abstinence. Moreover, in the two placebo groups, younger participants in the `BASC + placebo` group show higher abstinence rate while this group of individuals exhibits lower abstinence rate in the `ST + placebo` group. Within the varenicline groups, middle-aged participants in ST + varenicline demonstrate significantly higher abstinence rates than middle-aged participants in BASC + varenicline.

Figure 3: Baseline Characteristics by Abstinence Status and Treatment Group (Continuous 1)

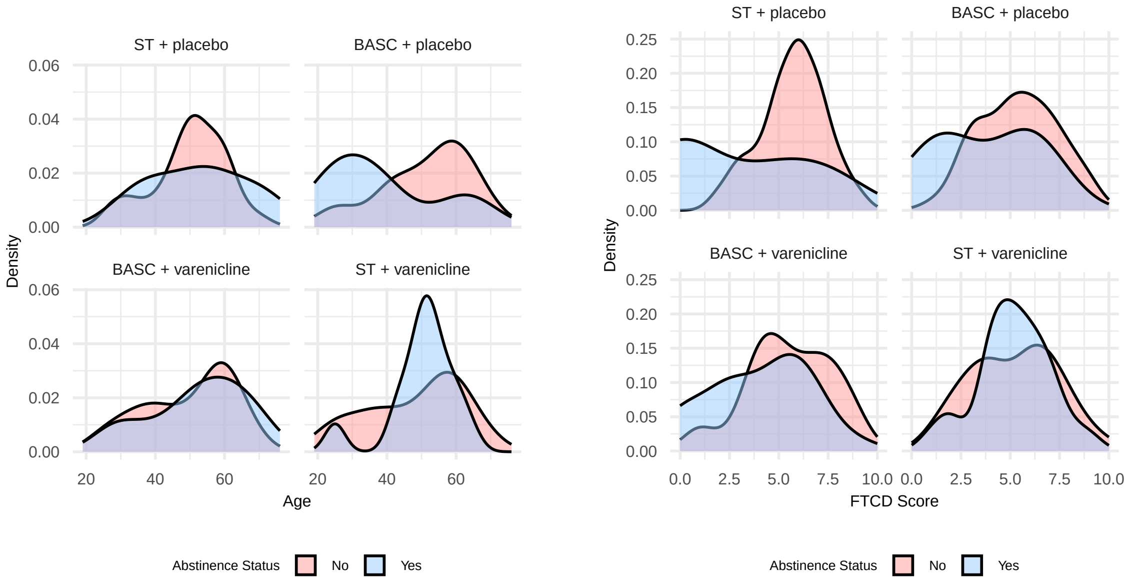
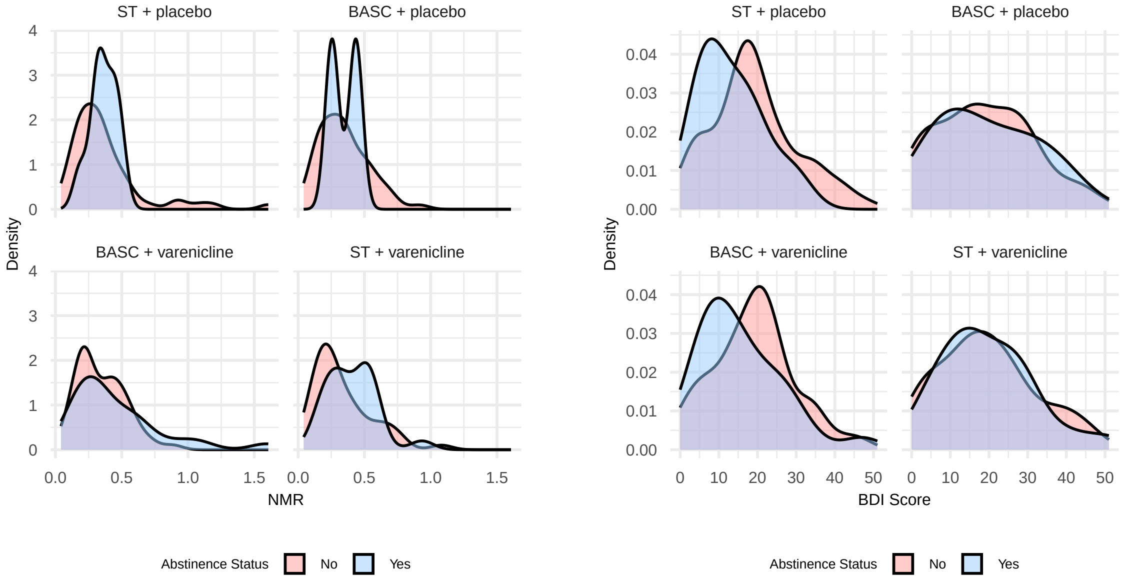


Figure 4: Baseline Characteristics by Abstinence Status and Treatment Group (Continuous 2)



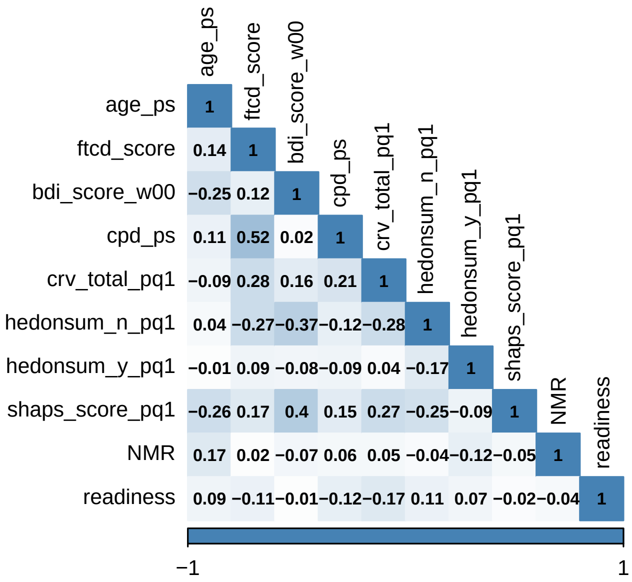
The FTCD score also appears to influence abstinence outcomes, with abstinence rates changing as FTCD scores vary. Moreover, participants with higher FTCD score in the `ST + placebo` group show significantly higher likelihood to continue smoking compared to those in the `BASC + placebo` group. Moreover, participants with FTCD score around 5 in the `ST + varenicline` group show higher abstinence rate compared to those in the `BASC + varenicline` group. These findings suggest potential interactions between age and behavioral treatment, as well as FTCD score-treatment and treatment. Age and FTCD score might serve as predictors and moderators of the treatment effect on the EOT abstinence.

Seeing `Figure 4`, the distribution of NMR is right skewed towards participants with higher values across all treatment groups. In the placebo groups, participants with lower NMR are more likely to quit smoking at the week 27 follow-up while this gap is less pronounced in the varenicline groups. Within the `BASC + placebo` group, there is a huge gap in abstinence rate between participants with NMRs ranging from 0.3 to 0.5. Moreover, comparing the varenicline groups, participants with NMRs between 0.3 and 0.7 show higher abstinence rate in the `ST + varenicline` group but lower abstinence rate in the `BASC + varenicline` group.

For the BDI score (measure of depression), which reflects depressive symptom severity, distinct patterns also emerge. In the placebo groups, participants in the ST group with lower BDI scores (indicating less severe depressive symptoms) are more likely to stop smoking, while those with moderate BDI scores show lower abstinence rates. However, in the BASC group, participants show a more uniform distribution regardless of their depressive severity. The same pattern was observed when we compared the two varenicline groups, suggesting a complex interacting relationship between BDI score and the combination of behavioral and pharmacological treatments. These findings suggest that both NMR and BDI scores may interact with behavioral treatment types, influencing smoking cessation outcomes and potentially serving as moderators of treatment effects.

Additionally, examining the correlation among continuous variables shown in `Figure 5`, we observed that most variables show low to moderate correlations with each other, with both positive and negative relationships present.

Figure 5: Correlation Plot among Environmental Condition Characteristics



To analyze the impact of behavioral treatment on end-of-treatment abstinence and examine the moderating role of baseline characteristics, we selected Lasso regression as our primary model. Since our data has limited observations compared to number of covariates, Lasso was chosen for its ability to perform both variable selection and regularization, making it particularly suited for our study, which involves numerous baseline predictors and interaction terms. By applying an L1 penalty, Lasso shrinks less relevant coefficients to zero, effectively selecting a subset of the most influential predictors and interactions.

Results

Before conducting the primary analysis, we performed exploratory data analysis (EDA) to examine baseline characteristics, assess data distributions, and identify potential relationships within the dataset.

`Table 2` presents an overall summary of statistics of patients' baseline characteristics by their behavioral and pharmacological treatment assignment. Since our study is a 2 × 2, factorial, randomized, placebo-controlled trial, patients are randomly assigned to either behavioral activation for smoking cessation group (BASC) or standard behavioral treatment group (ST) and either varenicline or placebo blister pack. Patients can be categorized into four treatment arm groups: BASC + placebo, BASC + varenicline, ST + placebo, and ST + varenicline. From `Table 2`, the two placebo groups both have 68 observations while the two varenicline groups have 83 and 81 observations, respectively.

Most variables are evenly distributed across the four treatment arms, which reflects successful randomization in this factorial trial. However, a few key factors, such as socioeconomic indicators (income and education) and specific mental health variables (MDD status, DSM-5 diagnoses), exhibit slight variations that may influence outcomes. Notably, treatment arms with varenicline show higher abstinence rates than placebo groups, suggesting the potential efficacy of this pharmacotherapy in combination with behavioral interventions. While many baseline characteristics are evenly distributed across groups, some may still function as moderators, potentially interacting with treatment assignment to affect abstinence success. In addition, only one observation falls into the Grade School level in the education variable. To ensure the appropriate representation of categories, we combined the grade school level with the next level, some high school, during the regression analysis. This adjustment ensures sufficient sample sizes across categories when we split the data.

Table 2: Participant Characteristics by Treatment Arm

	Behavioral and Pharmacological Treatment Assignment				
Characteristic	**ST + placebo, N = 68**	**BASC + placebo, N = 68**	**BASC + varenicline, N = 83**	**ST + varenicline, N = 81**	**Overall, N = 300**
Smoking abstinence	8 (12%)	4 (5.9%)	26 (31%)	26 (32%)	64 (21%)
Age	50 (11)	51 (14)	50 (13)	49 (13)	50 (13)
Sex					
Male	29 (43%)	30 (44%)	39 (47%)	37 (46%)	135 (45%)
Female	39 (57%)	38 (56%)	44 (53%)	44 (54%)	165 (55%)
Income					
Less than $20,000	26 (38%)	25 (37%)	30 (37%)	29 (36%)	110 (37%)
$20,000-35,000	14 (21%)	16 (24%)	17 (21%)	21 (26%)	68 (23%)
$35,001-50,000	14 (21%)	8 (12%)	13 (16%)	11 (14%)	46 (15%)
$50,001-75,000	8 (12%)	12 (18%)	12 (15%)	6 (7.5%)	38 (13%)
More than $75,000	6 (8.8%)	6 (9.0%)	10 (12%)	13 (16%)	35 (12%)
Missing	0	1	1	1	3
Education					
College graduate	17 (25%)	19 (28%)	29 (35%)	26 (32%)	91 (30%)
High school graduate or GED	11 (16%)	23 (34%)	15 (18%)	27 (33%)	76 (25%)
Some college/technical school	38 (56%)	22 (32%)	32 (39%)	24 (30%)	116 (39%)
Some high school & Grade School	2 (2.9%)	4 (5.9%)	7 (8.4%)	4 (4.9%)	17 (5.7%)
FTCD score	5 (2)	5 (2)	5 (2)	5 (2)	5 (2)
Missing	1	0	0	0	1
Smoking within 5 mins of waking up	35 (51%)	32 (47%)	33 (40%)	38 (47%)	138 (46%)
BDI score	18 (11)	19 (12)	18 (11)	20 (12)	19 (11)
Cigarettes smoked per day	15 (7)	16 (9)	16 (9)	14 (7)	15 (8)
Cigarette reward value	7 (4)	7 (4)	7 (4)	7 (3)	7 (4)
Missing	8	1	3	6	18
Pleasurable events (substitute reinforcers)	21 (20)	23 (20)	23 (19)	23 (19)	23 (20)
Pleasurable events (complementary reinforcers)	27 (20)	28 (22)	22 (17)	25 (19)	25 (19)

Table 2: Participant Characteristics by Treatment Arm *(continued)*

Characteristic	**Behavioral and Pharmacological Treatment Assignment**				
	ST + placebo, N = 68	**BASC + placebo, N = 68**	**BASC + varenicline, N = 83**	**ST + varenicline, N = 81**	**Overall, N = 300**
Anhedonia	3 (3)	2 (3)	2 (3)	2 (3)	2 (3)
Missing	1	2	0	0	3
Other lifetime DSM-5 diagnosis	28 (41%)	35 (51%)	30 (36%)	40 (49%)	133 (44%)
Taking antidepressant	15 (22%)	28 (41%)	24 (29%)	15 (19%)	82 (27%)
Current vs. past MDD					
Past MDD	37 (54%)	36 (53%)	43 (52%)	37 (46%)	153 (51%)
Current MDD	31 (46%)	32 (47%)	40 (48%)	44 (54%)	147 (49%)
Nicotine metabolism ratio	0.37 (0.27)	0.34 (0.18)	0.38 (0.25)	0.36 (0.21)	0.36 (0.23)
Missing	2	7	3	9	21
Exclusive mentholated cigarette user	43 (64%)	40 (59%)	48 (59%)	47 (58%)	178 (60%)
Missing	1	0	1	0	2
Readiness to quit smoking	7 (1)	7 (1)	7 (1)	7 (1)	7 (1)
Missing	4	4	5	4	17
Race					
Black	40 (59%)	36 (53%)	36 (43%)	43 (53%)	155 (52%)
Hispanic	4 (5.9%)	4 (5.9%)	3 (3.6%)	5 (6.2%)	16 (5.3%)
Mixed Race	0 (0%)	1 (1.5%)	1 (1.2%)	0 (0%)	2 (0.7%)
Non-Hispanic White	22 (32%)	24 (35%)	34 (41%)	25 (31%)	105 (35%)
Unknown	2 (2.9%)	3 (4.4%)	9 (11%)	8 (9.9%)	22 (7.3%)

[1] Mean (SD) for continuous; n (%) for categorical

During the data exploration, we observed that some continuous variables were highly skewed. Since applying transformations can reduce interpretability, we decided to perform Lasso regression on both the non-transformed and transformed datasets to evaluate whether transformation is necessary in this context.

For transformed data, we applied specific transformations based on the distributional characteristics of each variable, with transformations performed after the imputation step. Among the continuous variables, pleasurable events scale of substitute reinforcers (`hedonsum_n_pq1`), pleasurable events scale of complementary reinforcers (`hedonsum_y_pq1`), measure of anhedonia (`shaps_score_pq1`), and NMR exhibit right-skewed distributions. `Table 3` summarizes the skewness value of these variables before and after transformation.

For the two pleasurable event scale variables, `hedonsum_n_pq1` and `hedonsum_y_pq1`, we applied a square root transformation to reduce high positive skewness values (1.34 and 1.39, respectively). This transformation brought their skewness close to zero (-0.06 and 0.06, respectively), resulting in more symmetric distributions. Additionally, given that `shaps_score_pq1` had nearly 50% zero entries and a high positive skewness (1.71), we explored several transformations, including log, square root, and inverse hyperbolic sine (`asinh()`). The inverse hyperbolic sine transformation produced the lowest skewness (0.52), making it the most suitable choice for this variable. Finally, we applied a log transformation on NMR which presents the highest skewness value before transformation (1.92). The log transformation successfully reduced the skewness to a nearly symmetric value.

Table 3: Variable Transformation on Skewness

Variable	Transformation	Skewness before Transformation	Skewness after Transformation
hedonsum_n_pq1	Square Root Transformation	1.338843	-0.0591728
hedonsum_y_pq1	Square Root Transformation	1.391398	0.0620129
shaps_score_pq1	Inverse Hyperbolic Sine Transformation	1.707230	0.5217093
NMR	Log Transformation	1.915358	-0.2241582

As mentioned earlier, we performed multiple imputation using the `mice()` function to generate five different imputed data to address missingness. We then applied transformations listed in `Table 2` to the corresponding skewed variables in each imputed dataset.

With transformed and non-transformed data list, each imputed dataset was then split into a 70% training set and a 30% test set, stratified by treatment group using the `createDataPartition()` function in the `caret` package. Lasso regression was conducted on each training set using `cv.glmnet()` with a design matrix that included all baseline characteristics and their interactions with the behavioral and pharmacological treatment, respectively. To ensure consistent treatment group distribution across each cross-validation fold in lasso regression, we created custom fold assignments by treatment level and specified these assignments through the `foldid` argument in `cv.glmnet()`.

During cross-validation, we identified the optimal regularization parameter, `lambda.min`, which minimized the cross-validated error, and extracted the coefficient estimates for each lasso model at this optimal lambda value. Finally, we averaged the coefficient estimates from all five Lasso models to obtain the final pooled estimates, presented in `Table 4` and `Table 5`.

Table 4: Lasso Model Coefficient Estimate (With Transformation)

Variable	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5	Pooled Estimate
(Intercept)	-0.9284816	-0.8559755	-1.0279108	-0.9061058	-1.0172281	-0.9471404
ftcd_score	-0.1291703	-0.1160073	-0.1303573	-0.1194138	-0.1323784	-0.1254654
hedonsum_n_pq1	0.0000000	0.0000000	0.0000000	0.0000000	-0.0093333	-0.0018667
mde_curr1	-0.2278997	-0.1922978	-0.2446255	-0.2033876	-0.2283386	-0.2193099
NMR	0.2767557	0.3616764	0.1371997	0.3070181	0.1543496	0.2473999
raceNon-Hispanic White	0.3132327	0.2796795	0.3692644	0.2622257	0.4057589	0.3260323
readiness	0.0000000	0.0000000	-0.0105551	0.0000000	0.0000000	-0.0021110
BA1:inc$35,001-50,000	0.0000000	0.0000000	0.0039396	0.0000000	0.0621926	0.0132264
BA1:Only.Menthol1	-0.4293390	-0.4048011	-0.3786728	-0.3862525	-0.4467786	-0.4091688
BA1:raceNon-Hispanic White	0.0238877	0.0023216	0.0278458	0.0101141	0.0176939	0.0163726
age_ps:Var1	0.0167030	0.0168513	0.0169254	0.0156714	0.0201370	0.0172576
crv_total_pq1:Var1	0.0329422	0.0254871	0.0280000	0.0346775	0.0025389	0.0247291
eduHigh school graduate or GED:Var1	0.0710722	0.0259790	0.0897544	0.0211769	0.0997030	0.0615371
raceHispanic:Var1	0.0000000	0.0000000	-0.0053140	0.0000000	0.0000000	-0.0010628
raceMixed Race:Var1	1.0083121	0.7832045	1.0211446	0.8111792	1.0186898	0.9285060
sex_ps2:Var1	0.0910193	0.0558703	0.1151808	0.0702265	0.1108091	0.0886212

Table 5: Lasso Model Coefficient Estimate (No Transformation)

Variable	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5	Pooled Estimate
(Intercept)	-1.3610413	-1.4173454	-1.3699316	-1.4846438	-1.3237450	-1.3913414
ftcd_score	-0.0900852	-0.0709199	-0.0917031	-0.0487933	-0.0994217	-0.0801846
mde_curr1	-0.2262826	-0.2125496	-0.2235050	-0.1049150	-0.2424306	-0.2019366
raceNon-Hispanic White	0.1536878	0.0894440	0.2222718	0.0000000	0.2626255	0.1456058
BA1:Only.Menthol1	-0.3135925	-0.2884218	-0.2640911	-0.1173353	-0.3622487	-0.2691379
age_ps:Var1	0.0068256	0.0036695	0.0096917	0.0025406	0.0117753	0.0069005
crv_total_pq1:Var1	0.0249885	0.0120130	0.0214356	0.0181417	0.0000000	0.0153157
eduHigh school graduate or GED:Var1	0.0000000	0.0000000	0.0069655	0.0000000	0.0521435	0.0118218
NMR:Var1	1.4601455	1.9477758	1.0639002	1.5778488	1.2010092	1.4501359
raceMixed Race:Var1	0.5982123	0.3790549	0.5574714	0.0000000	0.7024706	0.4474419

Examining the results of Lasso regression on both the transformed and non-transformed datasets, we decided to retain the transformation. Comparing `Table 4` and `Table 5`, we found key skewed variables such as `NMR` and `hedonsum_n_pq1` became significant with transformation while are dropped with non-transformed data, indicating that the transformation better captured the relationships between these variables and the outcome. In addition, we generate the ROC curve for training and test set with both datasets shown in `Figure 6`. The

AUC exhibits a slight increase of 0.02 for the training set after applying transformation and it increases by 0.01 for the test set with transformation. The calibration plots shown in `Figure 7` do not provide clear or conclusive insights into whether transformations significantly improve the model. While transformations can reduce interpretability, the improved model performance and increased significance of skewed variables justify their use in this context. By applying transformations, we aim to achieve a more robust and accurate model without compromising key insights.

Figure 6: ROC Curves Comparison

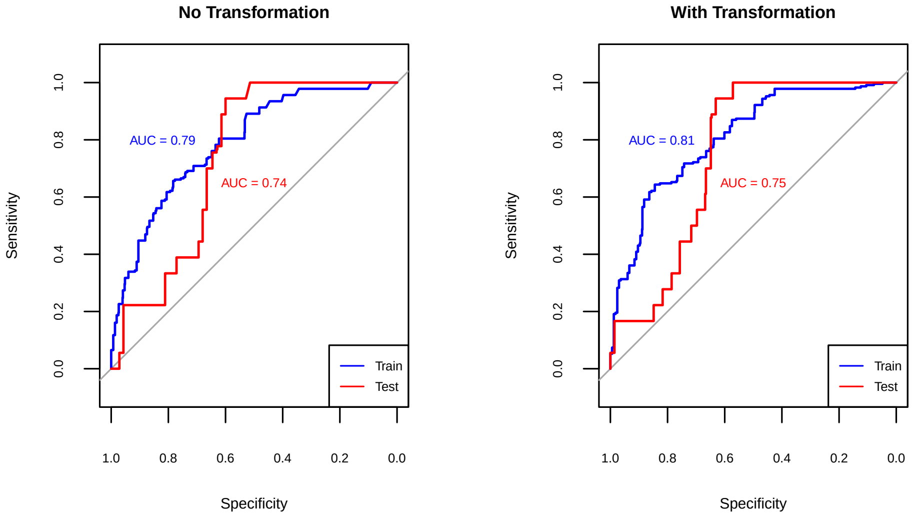
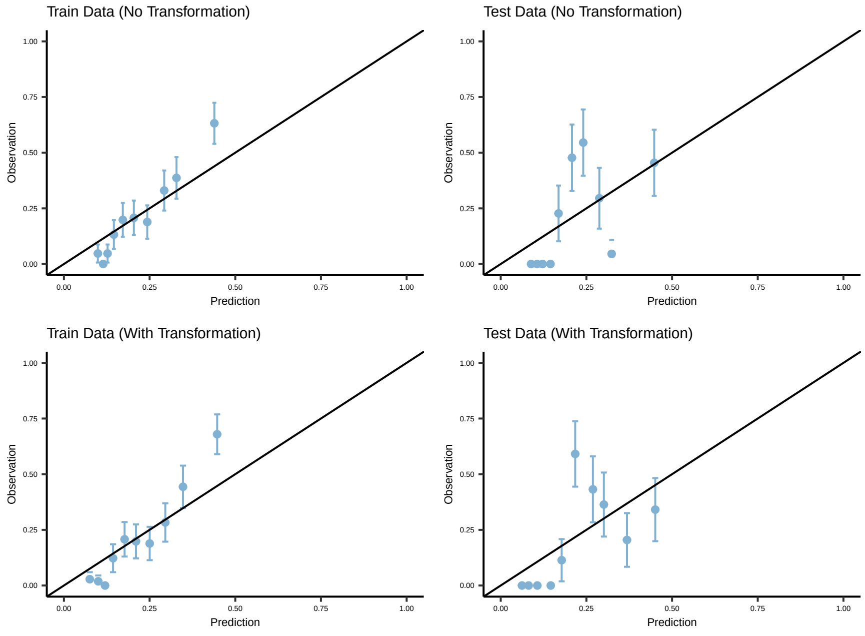


Figure 7: Calibration Plot Comparison



In addition to the primary analysis, we employed bootstrap resampling to evaluate the stability and reliability of the model coefficients and predictions (code can be found in `Bootstrap.Rmd`). We performed a bootstrap procedure with 200 iterations on the imputed datasets. In each iteration, we generated a bootstrap sample from the dataset and split it into training and testing subsets. The Lasso regression model was fitted to the training data using 5-fold cross-validation to identify the optimal regularization parameter (λ), ensuring robust variable selection. Coefficients from the final Lasso model were extracted for each bootstrap iteration and stored for further analysis. This method allowed us to quantify the variability of coefficient estimates and identify consistently important predictors across the bootstrap samples. To summarize the results, we calculated the mean coefficient estimates, 95% confidence intervals, and the proportion of times each variable was selected as significant across iterations with a csv file saved within the `Boostrap_Results` folder named as `summary_results.csv`. Variables selected more than one-third of the time were identified as significant. These significant variables, along with their pooled estimates, confidence intervals, and selection proportions, are presented in `Table 6`.

Observing the model coefficient without bootstrap shown in `Table 4`, we find that among the main effect variables, `ftcd_score`, `hedonsum_n_pq1`, and `readiness` present significant negative coefficient estimates as a predictor of smoking abstinence. This indicates that, holding other factors, higher nicotine dependence levels, higher values of the square root-transformed pleasurable events scale of substitute reinforcers, and greater baseline readiness to quit smoking are associated with reduced odds of smoking cessation at week 27. However, with `hedonsum_n_pq1` and `readiness` only exhibit significant in one imputed dataset, the effect of these two factors on abstinence might be relatively minor. Similarly, the variable `mde_curr`, indicating current Major Depressive Disorder (MDD), also has a significant negative coefficient estimate, suggesting that individuals with MDD are less likely to achieve abstinence, all else being equal.

In contrast, again as predictors, `NMR` and the race variable (Non-Hispanic White) have significant positive coefficient estimates. These findings suggest that higher log-transformed NMR values and identifying as Non-Hispanic White are positively associated with higher odds of smoking cessation, holding other factors constant.

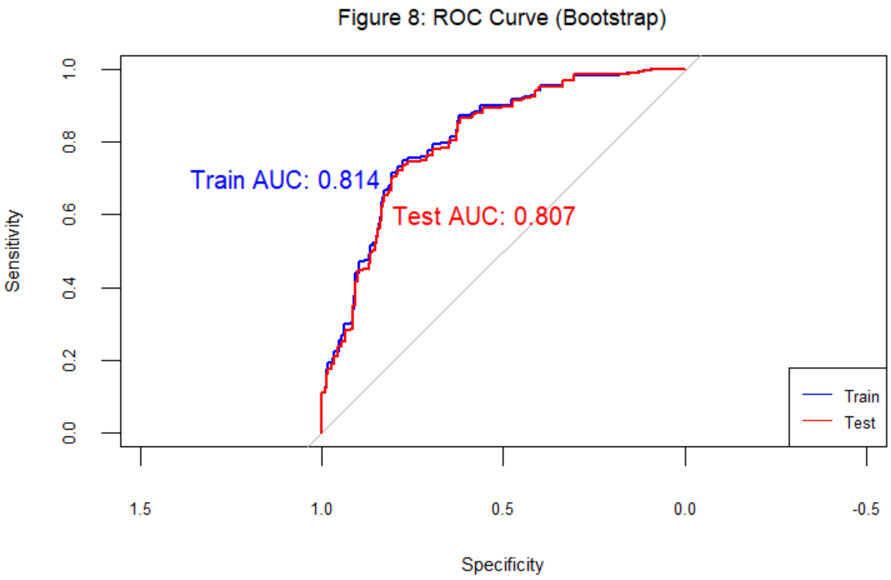
The interaction terms between behavioral treatment and the income variable ($35,001–50,000), the binary indicator of exclusive mentholated cigarette use, and the race variable (Non-Hispanic White) are significant in our results, identifying as the potential moderators of the effect of behavioral treatment. This suggests that participants receiving the BASC treatment with an income level of $35,001–50,000 or identifying as Non-Hispanic White may experience an additional increase in their odds of smoking cessation. Conversely, exclusive mentholated cigarette users undergoing the BASC treatment might experience an additional decrease in their odds of abstinence.

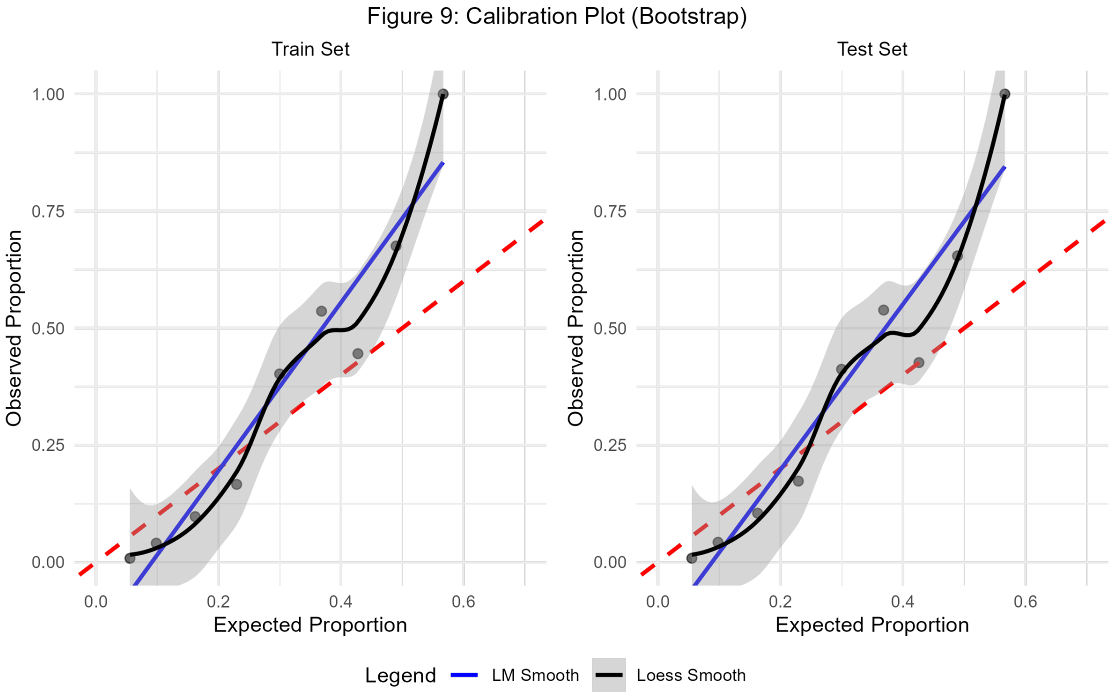
The model also has several interaction terms with the indicator of varenicline treatment which identify potential moderators for the pharmacological treatment on the EOT abstinence for people with MDD. Specifically, age, cigarette reward value (`crv_total_pq1`), education level (high school graduate or GED), race (Hispanic & Mixed Race), and sex serve as potential moderators. These interactions suggest that the effectiveness of varenicline may vary depending on these baseline characteristics, highlighting the nuanced ipact of pharmacological treatment across different demographic and behavioral profiles.

Table 6: Bootstrap Lasso Model Coefficient Estimate

Variable	Pooled.Estimate	Lower.CI	Upper.CI	Siginificant.Proportion
(Intercept)	-0.4622382	-2.3170828	2.3347997	1.000
ftcd_score	-0.2238994	-0.5670841	0.0000000	0.938
mde_curr1	-0.1396745	-0.7928705	0.0000000	0.418
NMR	0.2286429	0.0000000	0.8491399	0.647
raceNon-Hispanic White	0.3644262	0.0000000	1.4701028	0.648
BA1:ftcd_score	-0.0356967	-0.2615361	0.0000000	0.362
BA1:incMore than $75,000	0.2751041	0.0000000	1.7869354	0.380
BA1:Only.Menthol1	-0.2309910	-1.1583771	0.0000000	0.445
age_ps:Var1	0.0093747	0.0000000	0.0308070	0.757
antidepmed1:Var1	0.1497947	0.0000000	0.9060176	0.382
crv_total_pq1:Var1	0.0284297	0.0000000	0.1480517	0.471
eduHigh school graduate or GED:Var1	0.3050588	0.0000000	1.3186101	0.561
ftcd.5.mins1:Var1	0.4474196	0.0000000	1.7131313	0.598
raceHispanic:Var1	-0.2201501	-1.8094060	0.2680946	0.351
raceMixed Race:Var1	0.4918142	0.0000000	2.9191000	0.520

In comparison, the bootstrapped coefficient results in `Table 6` provide additional insights into predictor stability. Variables like hedonsum_n_pq1 and readiness, which exhibited limited significance in the original analysis, are dropped from the bootstrapped results, reflecting their minimal contribution. Conversely, new significant interactions emerge, such as between behavioral treatment and Anhedonia(`shaps_score_pq1`) and the indicator of whether taking antidepressant medication (`antidepmed`), both showing positive coefficient estimates. The previously observed significant interaction between behavioral treatment and sex is no longer present. Additionally, most variables exhibit confidence intervals that do not contain zero, suggesting they consistently demonstrate an effect across bootstrap iterations. However, the interaction term between varenicline and the race variable has a confidence interval that includes zero, indicating that its effect may not be as robust or consistent as the others.





We performed model evaluation using the ROC and calibration plots across the bootstrapped training and testing sets, as presented in `Figure 8` and `Figure 9`. The train AUC is 0.814, and the test AUC is 0.807, indicating strong predictive performance of the model. The close proximity of these AUC values between the train and test sets demonstrates that the model generalizes well, as it exhibits similar performance on both datasets. Moreover, the calibration curves in both sets generally align with the ideal line, with some deviations at higher predicted probabilities. There is no significant differences in calibration are observed between the training and testing sets, further supporting the model's robustness and reliability.

Discussion

Collaborating with the study of smoking cessation for people with MDD led by Dr. George Papandonatos, this study aims to further investigate potential moderators of behavioral treatment (BASC vs. ST) on the EOT abstinence for this group of people and dedicates to identify baseline characteristics as predictors of abstinence, controlling for behavioral and pharmacotherapy treatment. Lasso regression is chosen to perform variable selection that helps to focus on the most influential baseline characteristics and their interaction terms, balancing predictivity with interpretability. In addition, we perform a bootstrap approach aiming to provide a more robust and reliable analysis.

Control for treatment and other factors, as predictors, higher nicotine dependence (higher FTCD score), higher and current MDD status both associate with lower likelihood of abstinence. Conversely, having faster nicotine metabolism (higher NMR in log scale) and identifying as Non-Hispanic White were associated with higher odds of abstinence, adjusting for treatment and other factors.

Additionally, FTCD score emerged as both a predictor and moderator, with participants with higher nicotine dependence score showing lower odds of abstinence and experiencing an additional reduce from BASC. Menthol cigarette use and income level also moderated the effects of BASC, with menthol users experiencing lower abstinence odds and individuals with incomes more than $70,000 benefiting more from BASC. Furthermore, significant interaction terms with varenicline suggest that the efficacy of pharmacotherapy varies by factors like cigarette reward value, education, race, and age. The model evaluation using ROC and calibration plots reveals our model's strong discriminative power and exhibits well-calibrated results.

As noted in the original paper, a limitation of this study is the low adherence to both behavioral treatment and pharmacotherapy, particular in the BASC-only group. The use of an ITT approach under this scenario might lead to an underestimation of treatment effects due to low adherence. An additional limitation of our analysis is that, although our sample provides a diverse phase of patients with MDD, certain levels within some categorical variables have limited sample sizes. For example, as shown in `Table 3`, only one observation falls within the "Grade school" category for education. This limited representation may reduce the statistical power and limit the generalizability of our study.

Further research could explore strategies to enhance engagement and adherence and expand data collection to achieve more balanced representation of those underrepresented groups to improve the accuracy and generalizability of our model. Additionally, a future direction would be to apply a multilevel modeling approach which allows for the analysis of time-varying and group-level predictors, such as provider expertise or site-specific support. This method would allow us to assess how these factors interact with individual characteristics to impact abstinence outcome. Moreover, due to computational limitation and convergence issue, our analysis only focused on Lasso regression and we performed the bootstrap approach with only 200 iterations. Further researches using relaxed lasso or L0 + L2 penalty with more bootstrap iterations are recommended.

```r
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

# load necessary packages
library(tidyverse)
library(mice)
library(gt)
library(gtsummary)
library(kableExtra)
library(RColorBrewer)
library(scico)
library(caret)
library(glmnet)
library(pROC)
library(predtools)
library(gridExtra)
library(ggpubr)
library(patchwork)
library(e1071)
library(corrplot)
library(L0Learn)
library(MASS)
library(magick)
library(grid)
library(gridExtra)
# set working directory
# Windows
setwd("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Data/")

# Mac
# setwd("~/Desktop/Fall 2024/PHP 2550/Data/")

# read in data
data <- read.csv("project2.csv")
# factor categorical variables
data[, c("abst", "Var", "BA", "sex_ps", "NHW",
         "Black", "Hisp", "inc", "edu", "ftcd.5.mins",
         "otherdiag", "antidepmed", "mde_curr",
         "Only.Menthol")] <- lapply(data[, c("abst", "Var", "BA", "sex_ps", "NHW",
                                             "Black", "Hisp", "inc", "edu",
                                             "ftcd.5.mins", "otherdiag", "antidepmed",
                                             "mde_curr", "Only.Menthol")], as.factor)

# Recode factor levels in the dataset
averaged_data_factor <- data %>%
  mutate(abst = fct_recode(as.factor(abst), "Yes" = "1", "No" = "0"),
         inc = fct_recode(as.factor(inc),
                          "Less than $20,000" = "1",
                          "$20,000-35,000" = "2",
                          "$35,001-50,000" = "3",
                          "$50,001-75,000" = "4",
                          "More than $75,000" = "5"),
         sex_ps = fct_recode(as.factor(sex_ps), "Male" = "1", "Female" = "2"),
         edu = fct_recode(as.factor(edu),
                          "Grade School" = "1",
                          "Some high school" = "2",
                          "High school graduate or GED" = "3",
                          "Some college/technical school" = "4",
                          "College graduate" = "5"),
         ftcd.5.mins = fct_recode(as.factor(ftcd.5.mins), "Yes" = "1", "No" = "0"),
         otherdiag = fct_recode(as.factor(otherdiag), "Yes" = "1", "No" = "0"),
         antidepmed = fct_recode(as.factor(antidepmed), "Yes" = "1", "No" = "0"),
         mde_curr = fct_recode(as.factor(mde_curr), "Current MDD" = "1", "Past MDD" = "0"),
         Only.Menthol = fct_recode(as.factor(Only.Menthol), "Yes" = "1", "No" = "0"),
         race = as.factor(case_when(Black == 0 & Hisp == 0 & NHW == 0 ~ "Unknown",
                                    Black == 1 & Hisp == 1 & NHW == 1 ~ "Mixed Race",
                                    Black == 1 & Hisp == 1 ~ "Mixed Race",
                                    Black == 1 & NHW == 1 ~ "Mixed Race",
                                    NHW == 1 & Hisp == 1 ~ "Mixed Race",
                                    Black == 1 ~ "Black",
                                    Hisp == 1 ~ "Hispanic",
                                    NHW == 1 ~ "Non-Hispanic White",
                                    TRUE ~ "Other")),
         trt = as.factor(case_when(Var == 1 & BA == 1 ~ "BASC + varenicline",
                         Var == 0 & BA == 1 ~ "BASC + placebo",
                         Var == 1 & BA == 0 ~ "ST + varenicline",
                         Var == 0 & BA == 0 ~ "ST + placebo",
                         TRUE ~ NA_character_)))

averaged_data_factor$trt <- relevel(factor(averaged_data_factor$trt), ref = "ST + placebo")

averaged_data_factor <- averaged_data_factor %>%
  mutate(inc = fct_relevel(inc, "Less than $20,000", "$20,000-35,000",
                           "$35,001-50,000", "$50,001-75,000", "More than $75,000"),
         edu = fct_relevel(edu, "Grade School", "Some high school", "High school graduate or GED",
                           "Some college/technical school", "College graduate"))

averaged_data_factor$edu <- case_when(averaged_data_factor$edu == "Grade School" ~ "Some high school &
                                      averaged_data_factor$edu == "Some high school" ~ "Some high schoo
                                      TRUE ~ averaged_data_factor$edu)
missingness_df <- averaged_data_factor %>%
    summarise(across(everything(), ~ sum(is.na(.)))) %>%
    pivot_longer(cols = everything(), names_to = "Variable", values_to = "Missing_Count") %>%
    mutate(Total_Count = nrow(averaged_data_factor),
           Missing_Percentage = paste(round((Missing_Count / Total_Count) * 100, 2), "%")) %>%
    arrange(desc(Missing_Percentage)) %>%
  filter(Missing_Count != 0) %>%
  dplyr::select(-Total_Count)

colnames(missingness_df) <- c("Variable", "Missing Count", "Missing Percentage")
missingness_df %>%
  kable(booktabs = TRUE, caption = "Summary of Missing Data Patterns Across Variables ") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down"))
income_stackplot <- ggplot(averaged_data_factor, aes(x = abst, fill = inc)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(x = "Abstinence Status",
```

```r
       y = "Proportion",
       fill = "Income Level") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6)) +
  guides(fill = guide_legend(nrow = 2))

edu_stackplot <- ggplot(averaged_data_factor, aes(x = abst, fill = edu)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(x = "Abstinence Status",
       y = "Proportion",
       fill = "Education Level") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6)) +
  guides(fill = guide_legend(nrow = 2))

combined_plot_eduinc <- (wrap_elements(panel = income_stackplot + theme(legend.position = "bottom")) /
                           wrap_elements(panel = edu_stackplot + theme(legend.position = "bottom"))) +
  plot_layout(ncol = 2, guides = 'collect') +
  plot_annotation(title = "Figure 1: Baseline Characteristics by Abstinence Status and Treatment Group
                  theme = theme(plot.title = element_text(size = 8, hjust = 0.5)))

combined_plot_eduinc <- combined_plot_eduinc & theme(plot.margin = margin(10, 10, 10, 10),
                                                     legend.position = c(0.5, 0.1))

combined_plot_eduinc
race_stackplot <- ggplot(averaged_data_factor, aes(x = abst, fill = race)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(x = "Abstinence Status",
       y = "Proportion",
       fill = "Race Group") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
   theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))


only.menthol_stackplot <- ggplot(averaged_data_factor, aes(x = abst, fill = Only.Menthol)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(x = "Abstinence Status",
       y = "Proportion",
       fill = "Mentholated Cigarette User") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
   theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

combined_plot_racementhol <- (wrap_elements(panel = race_stackplot + theme(legend.position = "bottom"))
                           wrap_elements(panel = only.menthol_stackplot + theme(legend.position = "botto
  plot_layout(ncol = 2, guides = 'collect') +
  plot_annotation(title = "Figure 2: Baseline Characteristics by Abstinence Status and Treatment Group
                  theme = theme(plot.title = element_text(size = 8, hjust = 0.5)))

combined_plot_racementhol <- combined_plot_racementhol & theme(plot.margin = margin(10, 10, 10, 10),
                                                               legend.position = c(0.5, 0.1))

combined_plot_racementhol
ftcd_score_stackplot <- ggplot(averaged_data_factor, aes(x = ftcd_score, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(x = "FTCD Score",
       y = "Density",
       fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF")) +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

age_stackplot <- ggplot(averaged_data_factor, aes(x = age_ps, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(title = "",
       x = "Age",
       y = "Density",
       fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF")) +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

combined_plot_ftcdage <- (wrap_elements(panel = age_stackplot + theme(legend.position = "bottom")) /
                            wrap_elements(panel = ftcd_score_stackplot + theme(legend.position = "botto
  plot_layout(ncol = 2, guides = 'collect') +
  plot_annotation(title = "Figure 3: Baseline Characteristics by Abstinence Status and Treatment Group
                  theme = theme(plot.title = element_text(size = 8, hjust = 0.5)))

combined_plot_ftcdage <- combined_plot_ftcdage & theme(plot.margin = margin(10, 10, 10, 10),
                                                       legend.position = c(0.5, 0.1))

combined_plot_ftcdage
NMR_stackplot <- ggplot(averaged_data_factor, aes(x = NMR, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(x = "NMR",
       y = "Density",
       fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF")) +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

bdi_stackplot <- ggplot(averaged_data_factor, aes(x = bdi_score_w00, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(x = "BDI Score",
       y = "Density",
       fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF")) +
  theme(axis.title = element_text(size = 6),
        title = element_text(size = 6),
        axis.text = element_text(size = 6),
        legend.title = element_text(size = 5),
        legend.text = element_text(size = 5),
        legend.key.size = unit(0.3, "cm"),
        legend.position = "bottom",
        strip.text = element_text(size = 6))

combined_plot_NMRbdi <- (wrap_elements(panel = NMR_stackplot + theme(legend.position = "bottom")) /
                           wrap_elements(panel = bdi_stackplot + theme(legend.position = "bottom"))) +
  plot_layout(ncol = 2, guides = 'collect') +
  plot_annotation(title = "Figure 4: Baseline Characteristics by Abstinence Status and Treatment Group
                  theme = theme(plot.title = element_text(size = 8, hjust = 0.5)))

combined_plot_NMRbdi <- combined_plot_NMRbdi & theme(plot.margin = margin(10, 10, 10, 10),
                                                     legend.position = c(0.5, 0.1))

combined_plot_NMRbdi
# create a correlation matrix among environmental condition factors
cor_matrix <- cor(averaged_data_factor[, c(5, 12, 14, 15, 16, 17, 18, 19, 23, 25)], use = "complete.obs

# correlation plot of environmental condition factors
corrplot(cor_matrix, method = "color", type = "lower",
         tl.col = "black", tl.cex = 0.8, addCoef.col = "black",
         number.cex = 0.7, col = colorRampPalette(c("steelblue", "white", "steelblue"))(200))
title("Figure 5: Correlation Plot among Environmental Condition Characteristics",
      cex.main = 0.9, line = 3)
# create the summary table
summary_table <- averaged_data_factor %>%
  dplyr::select(-c("id", "Var", "BA", "Black", "Hisp", "NHW")) %>%
  tbl_summary(by = trt, label = list(abst ~ "Smoking abstinence",
                                     race ~ "Race",
                                     age_ps ~ "Age",
                                     sex_ps ~ "Sex",
                                     inc ~ "Income",
                                     edu ~ "Education",
                                     ftcd_score ~ "FTCD score",
                                     ftcd.5.mins ~ "Smoking within 5 mins of waking up",
                                     bdi_score_w00 ~ "BDI score",
                                     cpd_ps ~ "Cigarettes smoked per day",
                                     crv_total_pq1 ~ "Cigarette reward value",
                                     hedonsum_n_pq1 ~ "Pleasurable events (substitute reinforcers)",
                                     hedonsum_y_pq1 ~ "Pleasurable events (complementary reinforcers)",
                                     shaps_score_pq1 ~ "Anhedonia",
                                     otherdiag ~ "Other lifetime DSM-5 diagnosis",
                                     antidepmed ~ "Taking antidepressant",
                                     mde_curr ~ "Current vs. past MDD",
                                     NMR ~ "Nicotine metabolism ratio",
                                     Only.Menthol ~ "Exclusive mentholated cigarette user",
                                     readiness ~ "Readiness to quit smoking"),
              type = list(readiness ~ "continuous"),
              statistic = all_continuous() ~ "{mean} ({sd})",
              missing = "ifany",
              missing_text = "Missing") %>%
  add_overall(last = TRUE) %>%
  modify_spanning_header(update = all_stat_cols() ~ "**Behavioral and Pharmacological Treatment Assignme
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categorical") %>%
  bold_labels()

summary_table %>%
  as_kable_extra(booktabs = TRUE, caption = "Participant Characteristics by Treatment Arm",
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 7,
                            latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3.5cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  column_spec(6, width = "2cm") %>%
  row_spec(0, bold = TRUE, font_size = 7)
# Take transformation
averaged_data_factor_transformed <- averaged_data_factor
averaged_data_factor_transformed$shaps_score_pq1 <- asinh(averaged_data_factor$shaps_score_pq1)
averaged_data_factor_transformed$hedonsum_n_pq1 <- sqrt(averaged_data_factor$hedonsum_n_pq1)
averaged_data_factor_transformed$hedonsum_y_pq1 <- sqrt(averaged_data_factor$hedonsum_y_pq1)
averaged_data_factor_transformed$NMR <- log(averaged_data_factor$NMR)

skewness_df <- data.frame(Variable = c("hedonsum_n_pq1", "hedonsum_y_pq1", "shaps_score_pq1", "NMR"),
                          transformation = c("Square Root Transformation",
                                             "Square Root Transformation",
                                             "Inverse Hyperbolic Sine Transformation",
                                             "Log Transformation"),
                          skewness_before = c(skewness(averaged_data_factor$hedonsum_n_pq1),
                                              skewness(averaged_data_factor$hedonsum_y_pq1),
                                              skewness(averaged_data_factor$shaps_score_pq1, na.rm = TRU
                                              skewness(averaged_data_factor$NMR, na.rm = TRUE)),
                          skewness_after = c(skewness(averaged_data_factor_transformed$hedonsum_n_pq1),
                                             skewness(averaged_data_factor_transformed$hedonsum_y_pq1),
                                             skewness(averaged_data_factor_transformed$shaps_score_pq1,
                                             skewness(averaged_data_factor_transformed$NMR, na.rm = TRU

colnames(skewness_df) <- c("Variable", "Transformation",
                           "Skewness before Transformation", "Skewness after Transformation" )

skewness_df %>%
  kable(booktabs = TRUE, caption = "Variable Transformation on Skewness") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "2cm") %>%
  column_spec(2, width = "4cm") %>%
  column_spec(3, width = "3.5cm") %>%
  column_spec(4, width = "3.5cm")
# Set working directory
# Windows
setwd("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Data/")

# Mac
```

```r
# setwd("~/Desktop/Fall 2024/PHP 2550/Data/")

# Read in data
data <- read.csv("project2.csv")

# Factor categorical variables
data[, c("abst", "Var", "BA", "sex_ps", "NHW",
         "Black", "Hisp", "inc", "edu", "ftcd.5.mins",
         "otherdiag", "antidepmed", "mde_curr",
         "Only.Menthol")] <- lapply(data[, c("abst", "Var", "BA", "sex_ps", "NHW",
                                             "Black", "Hisp", "inc", "edu",
                                             "ftcd.5.mins", "otherdiag", "antidepmed",
                                             "mde_curr", "Only.Menthol")], as.factor)

# Multiple imputation with m = 5
new_data <- data[, -1]
imputed_data <- mice(new_data, m = 5, method = 'pmm', seed = 2550, printFlag = FALSE)

# Extract the five imputed datasets into a list
completed_datasets <- list()
for (i in 1:5) {
  completed_datasets[[i]] <- complete(imputed_data, i)
}

for (i in 1:length(completed_datasets)) {
  completed_datasets[[i]] <- completed_datasets[[i]] %>%
    mutate(trt = as.factor(case_when(Var == 1 & BA == 1 ~ "BASC + varenicline",
                                     Var == 0 & BA == 1 ~ "BASC + placebo",
                                     Var == 1 & BA == 0 ~ "ST + varenicline",
                                     Var == 0 & BA == 0 ~ "ST + placebo",
                                     TRUE ~ NA_character_)),
           race = as.factor(case_when(Black == 0 & Hisp == 0 & NHW == 0 ~ "Unknown",
                                      Black == 1 & Hisp == 1 & NHW == 1 ~ "Mixed Race",
                                      Black == 1 & Hisp == 1 ~ "Mixed Race",
                                      Black == 1 & NHW == 1 ~ "Mixed Race",
                                      NHW == 1 & Hisp == 1 ~ "Mixed Race",
                                      Black == 1 ~ "Black",
                                      Hisp == 1 ~ "Hispanic",
                                      NHW == 1 ~ "Non-Hispanic White",
                                      TRUE ~ "Other")),
           inc = fct_recode(as.factor(inc),
                            "Less than $20,000" = "1",
                            "$20,000-35,000" = "2",
                            "$35,001-50,000" = "3",
                            "$50,001-75,000" = "4",
                            "More than $75,000" = "5"),
           edu = fct_collapse(as.factor(edu),
                             "Some high school & Grade School" = c("1", "2"),
                             "High school graduate or GED" = "3",
                             "Some college/technical school" = "4",
                             "College graduate" = "5")) %>%
    mutate(inc = fct_relevel(inc, "Less than $20,000", "$20,000-35,000",
                             "$35,001-50,000", "$50,001-75,000", "More than $75,000"),
           edu = fct_relevel(edu, "Some high school & Grade School",
                             "High school graduate or GED",
                             "Some college/technical school", "College graduate")) %>%
    mutate(trt = relevel(factor(trt), ref = "ST + placebo"))

  # Apply transformations
  completed_datasets[[i]]$shaps_score_pq1 <- asinh(completed_datasets[[i]]$shaps_score_pq1)
  completed_datasets[[i]]$hedonsum_n_pq1 <- sqrt(completed_datasets[[i]]$hedonsum_n_pq1)
  completed_datasets[[i]]$hedonsum_y_pq1 <- sqrt(completed_datasets[[i]]$hedonsum_y_pq1)
  completed_datasets[[i]]$NMR <- log(completed_datasets[[i]]$NMR)
}
# Extract the five imputed datasets into a list
completed_datasets_nontrans <- list()
for (i in 1:5) {
  completed_datasets_nontrans[[i]] <- complete(imputed_data, i)
}

for (i in 1:length(completed_datasets_nontrans)) {
  completed_datasets_nontrans[[i]] <- completed_datasets_nontrans[[i]] %>%
    mutate(trt = as.factor(case_when(Var == 1 & BA == 1 ~ "BASC + varenicline",
                                     Var == 0 & BA == 1 ~ "BASC + placebo",
                                     Var == 1 & BA == 0 ~ "ST + varenicline",
                                     Var == 0 & BA == 0 ~ "ST + placebo",
                                     TRUE ~ NA_character_)),
           race = as.factor(case_when(Black == 0 & Hisp == 0 & NHW == 0 ~ "Unknown",
                                      Black == 1 & Hisp == 1 & NHW == 1 ~ "Mixed Race",
                                      Black == 1 & Hisp == 1 ~ "Mixed Race",
                                      Black == 1 & NHW == 1 ~ "Mixed Race",
                                      NHW == 1 & Hisp == 1 ~ "Mixed Race",
                                      Black == 1 ~ "Black",
                                      Hisp == 1 ~ "Hispanic",
                                      NHW == 1 ~ "Non-Hispanic White",
                                      TRUE ~ "Other")),
           inc = fct_recode(as.factor(inc),
                            "Less than $20,000" = "1",
                            "$20,000-35,000" = "2",
                            "$35,001-50,000" = "3",
                            "$50,001-75,000" = "4",
                            "More than $75,000" = "5"),
           edu = fct_collapse(as.factor(edu),
                             "Some high school & Grade School" = c("1", "2"),
                             "High school graduate or GED" = "3",
                             "Some college/technical school" = "4",
                             "College graduate" = "5")) %>%
    mutate(inc = fct_relevel(inc, "Less than $20,000", "$20,000-35,000",
                             "$35,001-50,000", "$50,001-75,000", "More than $75,000"),
           edu = fct_relevel(edu, "Some high school & Grade School",
                             "High school graduate or GED",
                             "Some college/technical school", "College graduate")) %>%
    mutate(trt = relevel(factor(trt), ref = "ST + placebo"))
}
lasso_model_function_moderator <- function(data_list) {
  lasso_coef <- list()
```

```r
  for (index in seq_along(data_list)) {
    data <- data_list[[index]]

    # Split train and test sets
    set.seed(2550)
    train_index <- createDataPartition(data$trt, p = 0.7, list = FALSE)
    train_data <- data[train_index, ]
    test_data <- data[-train_index, ]

    # Create fold IDs for cross-validation
    train_data$foldid <- NA
    for (trt_level in unique(train_data$trt)) {
      treatment_data <- train_data[train_data$trt == trt_level, ]
      fold_ids <- sample(rep(1:5, length.out = nrow(treatment_data)))
      train_data$foldid[train_data$trt == trt_level] <- fold_ids
    }

    # Define model matrix
    X <- 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 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    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 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    readiness + race), data = train_data)[, -1]
    y <- train_data$abst

    # Fit lasso with cross-validation using custom foldid
    cv_model <- cv.glmnet(X, y, family = "binomial", alpha = 1, nfolds = 5,
                          foldid = train_data$foldid, nlambda = 100)
    best_lambda <- cv_model$lambda.min

    # Fit final lasso model using best lambda
    lasso_model <- glmnet(X, y, family = "binomial", alpha = 1, lambda = best_lambda)

    # Extract coefficients and store in a data frame
    coefficients <- as.data.frame(as.matrix(coef(lasso_model)))
    coefficients$Variable <- rownames(coefficients)
    rownames(coefficients) <- NULL
    colnames(coefficients)[1] <- "Estimates"

    # Store coefficients in list
    lasso_coef[[index]] <- coefficients
  }

  # Return list of coefficients for all imputed datasets
  return(lasso_coef)
}
# Run the lasso model function with transformed data
lasso_coef_results_moderator <- lasso_model_function_moderator(completed_datasets)

coef_imputation_df <- Reduce(function(x, y) merge(x, y, by = "Variable"), lasso_coef_results_moderator)
```

```r
names(coef_imputation_df) <- c("Variable", "Imputation1", "Imputation2", "Imputation3", "Imputation4",
coef_imputation_df <- coef_imputation_df %>%
  filter(Imputation1 != 0 | Imputation2 != 0 | Imputation3 != 0 | Imputation4 != 0 | Imputation5 != 0)

coef_imputation_df$SortCategory <- c(1, 4, 3, 3, 3, 4, 4, 2, 2, 2, 2, 4, 4, 2, 2, 4)

coef_imputation_df <- coef_imputation_df[order(coef_imputation_df$SortCategory, coef_imputation_df$Vari
  dplyr::select(-c("SortCategory"))

names(coef_imputation_df) <- c("Variable", "Imputation 1", "Imputation 2", "Imputation 3", "Imputation 

# coef_imputation_df %>%
#   kable(booktabs = TRUE, caption = "Lasso Model Coefficient Estimate (With Transformation)") %>%
#   kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
#   column_spec(1, width = "0.5cm") %>%
#   column_spec(2, width = "2.5cm") %>%
#   column_spec(3, width = "1.5cm") %>%
#   column_spec(4, width = "1.5cm") %>%
#   column_spec(5, width = "1.5cm") %>%
#   column_spec(6, width = "1.5cm") %>%
#   column_spec(7, width = "1.5cm")
# Generate combined coefficient data frame
imputed_coefs_list_moderator <- list()
for (i in seq_along(lasso_coef_results_moderator)) {
  coefs <- lasso_coef_results_moderator[[i]]
  colnames(coefs)[colnames(coefs) == "Estimates"] <- paste0("Estimates_", i)
  imputed_coefs_list_moderator[[i]] <- coefs[, c("Variable", paste0("Estimates_", i))]
}

# Combine and pool estimates
wide_format_coefficients_moderator <- Reduce(function(x, y) merge(x, y, by = "Variable", all = TRUE), in
wide_format_coefficients_moderator$Pooled_Estimate <- rowMeans(wide_format_coefficients_moderator[ , gr

coef_table_moderator <- wide_format_coefficients_moderator %>%
  filter(Pooled_Estimate != 0) %>%
  dplyr::select(c("Variable", "Pooled_Estimate"))

colnames(coef_table_moderator)[2] <- "Estimate"
coef_table_moderator$exp_estimate <- exp(coef_table_moderator$Estimate)
# Order main effects and interactions
main_effects <- coef_table_moderator[!grepl(":", coef_table_moderator$Variable), ]
interaction_terms <- coef_table_moderator[grepl(":", coef_table_moderator$Variable), ]
age_interaction <- coef_table_moderator[coef_table_moderator$Variable == "age_ps:Var1", ]

# Combine for final ordered result
ordered_model_results <- rbind(main_effects, interaction_terms)[-8,]
ordered_model_results <- rbind(ordered_model_results, age_interaction)
rownames(ordered_model_results) <- NULL
colnames(ordered_model_results)[3] <- "Exponential Estimate"

trans_coef <- left_join(coef_imputation_df, ordered_model_results, by = "Variable") %>%
  dplyr::select(-8)
```

```r
names(trans_coef)[7] <- "Pooled Estimate"

# Display results
trans_coef %>%
  kable(booktabs = TRUE, caption = "Lasso Model Coefficient Estimate (With Transformation)") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "4cm") %>%
  column_spec(2, width = "1.5cm") %>%
  column_spec(3, width = "1.5cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1.5cm") %>%
  column_spec(6, width = "1.5cm") %>%
  column_spec(7, width = "2cm")
# Run the lasso model function with data none transformation
lasso_coef_results_moderator_nontrans <- lasso_model_function_moderator(completed_datasets_nontrans)

coef_imputation_df_nontrans <- Reduce(function(x, y) merge(x, y, by = "Variable"), lasso_coef_results_mo
names(coef_imputation_df_nontrans) <- c("Variable", "Imputation1", "Imputation2", "Imputation3", "Imputa
coef_imputation_df_nontrans <- coef_imputation_df_nontrans %>%
  filter(Imputation1 != 0 | Imputation2 != 0 | Imputation3 != 0 | Imputation4 != 0 | Imputation5 != 0)

# coef_imputation_df_nontrans$SortCategory <- c(1, 4, 3, 4, 4, 2, 2, 4, 4, 4, 2, 4)
coef_imputation_df_nontrans$SortCategory <- c(1, 4, 3, 4, 4, 2, 2, 4, 4, 2)

coef_imputation_df_nontrans <- coef_imputation_df_nontrans[order(coef_imputation_df_nontrans$SortCatego
  dplyr::select(-c("SortCategory"))

names(coef_imputation_df_nontrans) <- c("Variable", "Imputation 1", "Imputation 2", "Imputation 3", "Imp

# coef_imputation_df_nontrans %>%
#   kable(booktabs = TRUE, caption = "Lasso Model Coefficient Estimate (No Transformation)") %>%
#   kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
#   column_spec(1, width = "0.5cm") %>%
#   column_spec(2, width = "2.5cm") %>%
#   column_spec(3, width = "1.5cm") %>%
#   column_spec(4, width = "1.5cm") %>%
#   column_spec(5, width = "1.5cm") %>%
#   column_spec(6, width = "1.5cm") %>%
#   column_spec(7, width = "1.5cm")
# Generate combined coefficient data frame
imputed_coefs_list_moderator_nontrans <- list()
for (i in seq_along(lasso_coef_results_moderator_nontrans)) {
  coefs <- lasso_coef_results_moderator_nontrans[[i]]
  colnames(coefs)[colnames(coefs) == "Estimates"] <- paste0("Estimates_", i)
  imputed_coefs_list_moderator_nontrans[[i]] <- coefs[, c("Variable", paste0("Estimates_", i))]
}

# Combine and pool estimates
wide_format_coefficients_moderator_nontrans <- Reduce(function(x, y) merge(x, y, by = "Variable", all =
wide_format_coefficients_moderator_nontrans$Pooled_Estimate <- rowMeans(wide_format_coefficients_modera

coef_table_moderator_nontrans <- wide_format_coefficients_moderator_nontrans %>%
  filter(Pooled_Estimate != 0) %>%
  dplyr::select(c("Variable", "Pooled_Estimate"))

colnames(coef_table_moderator_nontrans)[2] <- "Estimate"
coef_table_moderator_nontrans$exp_estimate <- exp(coef_table_moderator_nontrans$Estimate)
# Order main effects and interactions
main_effects_nontrans <- coef_table_moderator_nontrans[!grepl(":", coef_table_moderator_nontrans$Variabl
interaction_terms_nontrans <- coef_table_moderator_nontrans[grepl(":", coef_table_moderator_nontrans$Va
age_interaction_nontrans <- coef_table_moderator_nontrans[coef_table_moderator_nontrans$Variable == "ag

# Combine for final ordered result
ordered_model_results_nontrans <- rbind(main_effects_nontrans, interaction_terms_nontrans)[-5,]
ordered_model_results_nontrans <- rbind(ordered_model_results_nontrans, age_interaction_nontrans)
rownames(ordered_model_results_nontrans) <- NULL
colnames(ordered_model_results_nontrans)[3] <- "Exponential Estimate"

nontrans_coef <- left_join(coef_imputation_df_nontrans, ordered_model_results_nontrans, by = "Variable"
  dplyr::select(-8)

names(nontrans_coef)[7] <- "Pooled Estimate"

# Display results
nontrans_coef %>%
  kable(booktabs = TRUE, caption = "Lasso Model Coefficient Estimate (No Transformation)") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "4cm") %>%
  column_spec(2, width = "1.5cm") %>%
  column_spec(3, width = "1.5cm") %>%
  column_spec(4, width = "1.5cm") %>%
  column_spec(5, width = "1.5cm") %>%
  column_spec(6, width = "1.5cm") %>%
  column_spec(7, width = "2cm")
long_data_train <- data.frame()
long_data_test <- data.frame()

# get stratified training index based on treatment group
set.seed(2550)
train_index <- createDataPartition(completed_datasets[[1]]$trt, p = 0.7, list = FALSE)

# generate long format of train and test dataframe from the five imputed datasets
for (i in 1:length(completed_datasets)) {
  imputed_dataset <- completed_datasets[[i]]
  train_set <- imputed_dataset[train_index, ]
  test_set <- imputed_dataset[-train_index, ]

  long_data_train <- rbind(long_data_train, train_set)
  long_data_test <- rbind(long_data_test, test_set)
}
# create the design matrix with interaction terms
long_data_matrix_train <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mi
                                    bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    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 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    readiness + race),
                                       data = long_data_train)

# convert the design matrix to a data frame
long_data_trainset <- as.data.frame(long_data_matrix_train)

# extract the intercept from pooled coefficients
pooled_intercept <- wide_format_coefficients_moderator %>%
  filter(Variable == "(Intercept)") %>%
  pull(Pooled_Estimate)

# extract only non-intercept pooled coefficients
pooled_coefs <- wide_format_coefficients_moderator %>%
  filter(Variable != "(Intercept)")

# ensure the predictor variables in the data match those in pooled coefficients
predictor_vars <- pooled_coefs$Variable
long_data_trainset <- long_data_trainset[, predictor_vars, drop = FALSE]

# calculate log-odds using matrix multiplication with pooled coefficients
long_data_trainset$log_odds <- pooled_intercept + as.matrix(long_data_trainset) %*% pooled_coefs$Pooled_

# convert log-odds to probabilities
long_data_trainset$predicted_prob <- 1 / (1 + exp(-long_data_trainset$log_odds))
# create the design matrix with interaction terms
long_data_matrix_test <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mi
                                    bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    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 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    readiness + race),
                                 data = long_data_test)

# convert the design matrix to a data frame
long_data_testset <- as.data.frame(long_data_matrix_test)

# ensure the predictor variables in the data match those in pooled coefficients
long_data_testset <- long_data_testset[, predictor_vars, drop = FALSE]

# calculate log-odds using matrix multiplication with pooled coefficients
long_data_testset$log_odds <- pooled_intercept + as.matrix(long_data_testset) %*% pooled_coefs$Pooled_E

# convert log-odds to probabilities
long_data_testset$predicted_prob <- 1 / (1 + exp(-long_data_testset$log_odds))
long_data_train_nontrans <- data.frame()
long_data_test_nontrans <- data.frame()

# get stratified training index based on treatment group
set.seed(2550)
```

```
train_index <- createDataPartition(completed_datasets[[1]]$trt, p = 0.7, list = FALSE)

# generate long format of train and test dataframe from the five imputed datasets
for (i in 1:length(completed_datasets_nontrans)) {
  imputed_dataset <- completed_datasets_nontrans[[i]]
  train_set <- imputed_dataset[train_index, ]
  test_set <- imputed_dataset[-train_index, ]

  long_data_train_nontrans <- rbind(long_data_train_nontrans, train_set)
  long_data_test_nontrans <- rbind(long_data_test_nontrans, test_set)
}
# create the design matrix with interaction terms
long_data_matrix_train_nontrans <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score +
                                    bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    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 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    readiness + race),
                                       data = long_data_train_nontrans)

# convert the design matrix to a data frame
long_data_trainset_nontrans <- as.data.frame(long_data_matrix_train_nontrans)

# extract the intercept from pooled coefficients
pooled_intercept_nontrans <- wide_format_coefficients_moderator_nontrans %>%
  filter(Variable == "(Intercept)") %>%
  pull(Pooled_Estimate)

# extract only non-intercept pooled coefficients
pooled_coefs_nontrans <- wide_format_coefficients_moderator_nontrans %>%
  filter(Variable != "(Intercept)")

# ensure the predictor variables in the data match those in pooled coefficients
predictor_vars_nontrans <- pooled_coefs_nontrans$Variable
long_data_trainset_nontrans <- long_data_trainset_nontrans[, predictor_vars_nontrans, drop = FALSE]

# calculate log-odds using matrix multiplication with pooled coefficients
long_data_trainset_nontrans$log_odds <- pooled_intercept_nontrans + as.matrix(long_data_trainset_nontra

# convert log-odds to probabilities
long_data_trainset_nontrans$predicted_prob <- 1 / (1 + exp(-long_data_trainset_nontrans$log_odds))
# create the design matrix with interaction terms
long_data_matrix_test_nontrans <- model.matrix(abst ~ BA * (age_ps + sex_ps + inc + edu + ftcd_score +
                                    bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    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 + hedonsum_
                                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Me
                                    readiness + race),
                                 data = long_data_test_nontrans)
```

```
# convert the design matrix to a data frame
long_data_testset_nontrans <- as.data.frame(long_data_matrix_test_nontrans)

# ensure the predictor variables in the data match those in pooled coefficients
long_data_testset_nontrans <- long_data_testset_nontrans[, predictor_vars_nontrans, drop = FALSE]

# calculate log-odds using matrix multiplication with pooled coefficients
long_data_testset_nontrans$log_odds <- pooled_intercept_nontrans + as.matrix(long_data_testset_nontrans

# convert log-odds to probabilities
long_data_testset_nontrans$predicted_prob <- 1 / (1 + exp(-long_data_testset_nontrans$log_odds))
# do roc on train and test sets
auc_result_nontrans <- roc(long_data_train_nontrans$abst, long_data_trainset_nontrans$predicted_prob)
auc_result_test_nontrans <- roc(long_data_test_nontrans$abst, long_data_testset_nontrans$predicted_prob

auc_result <- roc(long_data_train$abst, long_data_trainset$predicted_prob)
auc_result_test <- roc(long_data_test$abst, long_data_testset$predicted_prob)

# plot roc for both sets
par(mfrow= c(1,2), oma = c(0, 0, 2, 0))
plot(auc_result_nontrans, main = "No Transformation", col = "blue", lwd = 1.5,
     cex.main = 0.7, cex.lab = 0.6, cex.axis = 0.5)
plot(auc_result_test_nontrans, add = TRUE, col = "red", lwd = 1.5)
text(0.82, 0.8, paste("AUC =", round(auc(auc_result_nontrans), 2)), col = "blue", cex = 0.5)
text(0.5, 0.65, paste("AUC =", round(auc(auc_result_test_nontrans), 2)), col = "red", cex = 0.5)
legend("bottomright", legend = c("Train", "Test"), col = c("blue", "red"), lwd = 1, cex = 0.5)

plot(auc_result, main = "With Transformation", col = "blue", lwd = 1.5,
     cex.main = 0.7, cex.lab = 0.6, cex.axis = 0.5)
plot(auc_result_test, add = TRUE, col = "red", lwd = 1.5)
text(0.82, 0.8, paste("AUC =", round(auc(auc_result), 2)), col = "blue", cex = 0.5)
text(0.5, 0.65, paste("AUC =", round(auc(auc_result_test), 2)), col = "red", cex = 0.5)
legend("bottomright", legend = c("Train", "Test"), col = c("blue", "red"), lwd = 1, cex = 0.5)
mtext("Figure 6: ROC Curves Comparison", outer = TRUE, cex = 0.8)
long_data_trainset <- long_data_trainset %>%
  mutate(abst_num = as.numeric(as.character(long_data_train$abst)))
long_data_testset <- long_data_testset %>%
  mutate(abst_num = as.numeric(as.character(long_data_test$abst)))

cal_plot_train <- calibration_plot(data = long_data_trainset, obs = "abst_num",
                                   pred = "predicted_prob", title = "Train Data (With Transformation)",
                                   y_lim = c(0, 1), x_lim=c(0, 1))
cal_plot_train <- cal_plot_train$calibration_plot
cal_plot_train <- cal_plot_train +
  theme(plot.title = element_text(size = 8),
        axis.title.x = element_text(size = 6), axis.title.y = element_text(size = 6),
        axis.text.x = element_text(size = 4), axis.text.y = element_text(size = 4))

cal_plot_test <- calibration_plot(data = long_data_testset, obs = "abst_num",
                                  pred = "predicted_prob", title = "Test Data (With Transformation)",
                                  y_lim = c(0, 1), x_lim=c(0, 1))
cal_plot_test <- cal_plot_test$calibration_plot
cal_plot_test <- cal_plot_test +
  theme(plot.title = element_text(size = 8),
        axis.title.x = element_text(size = 6), axis.title.y = element_text(size = 6),
        axis.text.x = element_text(size = 4), axis.text.y = element_text(size = 4))


long_data_trainset_nontrans <- long_data_trainset_nontrans %>%
  mutate(abst_num = as.numeric(as.character(long_data_train_nontrans$abst)))
long_data_testset_nontrans <- long_data_testset_nontrans %>%
  mutate(abst_num = as.numeric(as.character(long_data_test_nontrans$abst)))

cal_plot_train_nontrans <- calibration_plot(data = long_data_trainset_nontrans, obs = "abst_num",
                                            pred = "predicted_prob", title = "Train Data (No Transforma
                                            y_lim = c(0, 1), x_lim=c(0, 1))
cal_plot_train_nontrans <- cal_plot_train_nontrans$calibration_plot
cal_plot_train_nontrans <- cal_plot_train_nontrans +
  theme(plot.title = element_text(size = 8),
        axis.title.x = element_text(size = 6), axis.title.y = element_text(size = 6),
        axis.text.x = element_text(size = 4), axis.text.y = element_text(size = 4))

cal_plot_test_nontrans <- calibration_plot(data = long_data_testset_nontrans, obs = "abst_num",
                                           pred = "predicted_prob", title = "Test Data (No Transformati
                                           y_lim = c(0, 1), x_lim=c(0, 1))

cal_plot_test_nontrans <- cal_plot_test_nontrans$calibration_plot
cal_plot_test_nontrans <- cal_plot_test_nontrans +
  theme(plot.title = element_text(size = 8),
        axis.title.x = element_text(size = 6), axis.title.y = element_text(size = 6),
        axis.text.x = element_text(size = 4), axis.text.y = element_text(size = 4))

grid.arrange(cal_plot_train_nontrans, cal_plot_test_nontrans,
             cal_plot_train, cal_plot_test, ncol = 2,
             top = text_grob("Figure 7: Calibration Plot Comparison"))
bootstrap_summary <- read.csv("Bootstrap_Results/summary_results.csv")
bootstrap_summary %>%
  kable(booktabs = TRUE, caption = "Bootstrap Lasso Model Coefficient Estimate ") %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down"))
# Read in the bootstrap ROC
bootstrap_roc_image <- image_read("BootStrap_Results/roc_curve.png")
bootstrap_roc_image
# Read in the bootstrap calibration
bootstrap_calibration_image <- image_read("BootStrap_Results/calibration_plot.png")
bootstrap_calibration_image
```

Reference

1. Santomauro, D. F., Herrera, A. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Vos, T., Whiteford, H., Ferrari, A. J., Charlson, F. J., et al. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, *398*(10312), 1700–1712.
2. Weinberger, A. H., Chaiton, M. O., Zhu, J., Wall, M. M., Hasin, D. S., & Goodwin, R. D. (2020). Trends in the prevalence of current, daily, and nondaily cigarette smoking and quit ratios by depression status in the u.s.: 2005–2017. *American Journal of Preventive Medicine*, *58*(5), 691–698.
3. Hitsman, B., Papandonatos, G. D., McChargue, D. E., & al., et. (2013). Past major depression and

smoking cessation outcome: A systematic review and meta-analysis update. *Addiction*, *108*(2), 294–306.