Evaluate baseline variables as predictors of abstinence

Jing Fu

November 2024

Abstract

Background: Smoking cessation remains a significant challenge, particularly for individuals with a history of major depressive disorder (MDD), who often experience higher nicotine dependence and more severe withdrawal symptoms. This study investigates the moderating effects of baseline factors, including behavioral activation (BA) and pharmacotherapy (varenicline), on smoking abstinence outcomes in this high-risk population.

Methods: A randomized, placebo-controlled, 2x2 factorial design was used, involving 300 adult smokers with current or past MDD. Modified Poisson regression models were applied to multiply imputed datasets to assess the impact of demographic, psychological, and physiological baseline factors. AUC and Briei were used to evaluate models.

Results: The results indicate that Non-Hispanic White status, lower FTCD scores, and higher Nicotine Metabolism Ratio were significant predictors of smoking abstinence, with varenicline showing a strong positive effect on cessation rates. Behavioral therapy alone did not exhibit significant variation across subgroups, while pharmacotherapy consistently improved outcomes. Overall, the models demonstrated moderate to good discrimination and calibration.

Conclusion: This study highlights the critical role of nicotine dependence and pharmacotherapy in smoking cessation among individuals with MDD. While behavioral activation may not uniformly enhance cessation outcomes, its interaction with smoking behaviors warrants further investigation. These findings underscore the need for personalized smoking cessation strategies tailored to individual baseline characteristics.

Introduction

Smoking is a leading cause of preventable death, associated with severe health outcomes such as cardiovascular disease and cancer. Despite widespread public health initiatives, quitting smoking remains a significant challenge, particularly for individuals with major depressive disorder (MDD). This population is not only more likely to smoke heavily and exhibit stronger nicotine dependence but also experiences more intense withdrawal symptoms and higher relapse rates [1-4]. For these individuals, effective smoking cessation requires addressing both physiological dependence and the psychological barriers imposed by depression.

Pharmacotherapy and behavioral interventions are two primary approaches to smoking cessation. Among pharmacotherapies, varenicline has proven effective by reducing nicotine cravings and withdrawal symptoms [5]. However, pharmacological treatment alone may not sufficiently address the psychological challenges faced by individuals with MDD. To complement this, Behavioral Activation for Smoking Cessation (BASC) integrates traditional smoking cessation techniques with strategies to improve mood and reduce anhedonia, a common symptom in MDD. BASC aims to enhance overall well-being while simultaneously addressing smoking-related behaviors [6-8].

A recent randomized controlled trial explored the efficacy of BASC combined with varenicline compared to standard treatment (ST) and placebo [9]. Surprisingly, the study found no significant difference between BASC and ST in smoking cessation rates, raising questions about potential individual differences in treatment effectiveness. Specifically, baseline characteristics such as depression severity, nicotine dependence, or socioeconomic factors may influence how individuals respond to these interventions. Understanding these factors is critical to tailoring treatments for optimal outcomes.

This project aims to investigate two key questions: (1) whether baseline variables moderate the effects of behavioral treatments (BASC vs. ST) on smoking cessation, and (2) which baseline variables predict smoking abstinence, controlling for pharmacotherapy and behavioral treatment. These questions are crucial for developing personalized smoking cessation strategies, particularly for individuals with MDD, to improve treatment efficacy and long-term outcomes.

Methods

Study population

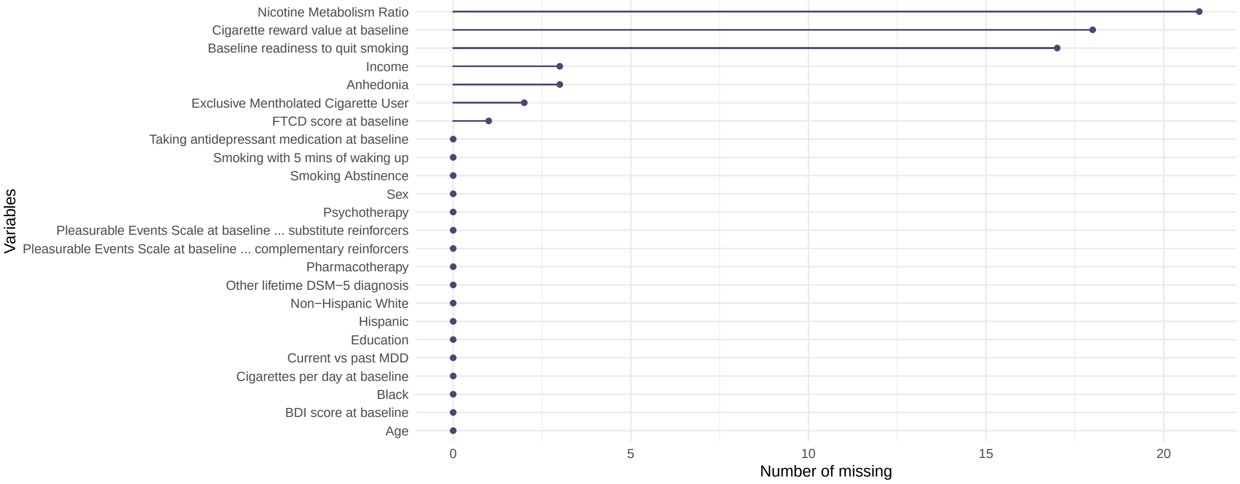
The study population consisted of 300 adults recruited from two research sites: Northwestern University in Chicago, Illinois, and the University of Pennsylvania in Philadelphia, Pennsylvania. Participants were daily smokers (>= 1 cigarette/day) with a lifetime diagnosis of major depressive disorder (MDD) without psychotic features, based on DSM-5 criteria.

Initial eligibility screening was conducted via telephone, followed by a comprehensive baseline assessment at the intake session. Participants were randomized into one of four treatment arms, stratified by site, sex, and depressive symptom severity (minimal/mild versus moderate/severe), as measured by the Beck Depression Inventory-II (BDI-II). This randomization ensured balanced distribution across treatment conditions, which included Behavioral Activation for Smoking Cessation (BASC) or Standard Treatment (ST), combined with either varenicline or placebo.

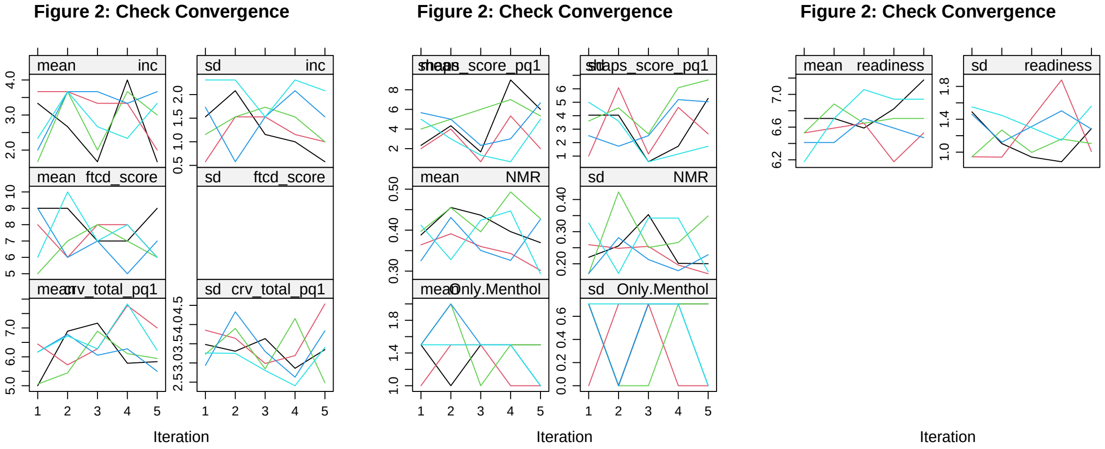
The study captured a wide range of baseline characteristics, including demographic variables (e.g., age, sex, race/ethnicity, income, and education), smoking-related measures (e.g., FTCD scores, cigarettes per day, time to first cigarette), and psychological factors (e.g., BDI-II scores, anhedonia, and readiness to quit).

Missing data

Figure1. Missing Pattern



Using these 300 observations, the number of missing values for all 24 variables (except id) is plotted. Observing Figure 1, we can see that there are 7 variables including Anhedonia, Income, FTCD score at baseline, Exclusive Mentholated Cigarette User, Baseline readiness to quit smoking, Cigarette reward value at baseline and Nicotine Metabolism Ratio that have missing values. In order to retain as much data as possible, multiple imputation was performed using the mice package in R to address the missing data in the study.



The MICE algorithm iteratively imputes missing values for each variable based on observed values of other variables. We generated five imputed datasets. This process allows for robust statistical analysis by reducing bias and improving efficiency. All covariates from the original dataset were included in the imputation model.

For continuous variables, we used predictive mean matching (PMM), which ensures that the imputed values are plausible and fall within the range of observed values. PMM was applied to variables such as FTCD score at baseline, Cigarette reward value, Anhedonia, Baseline readiness to quit smoking and Nicotine Metabolism Ratio. This method is particularly effective in handling skewed or non-normally distributed continuous variables.

Categorical variables were imputed using appropriate logistic regression-based methods. For binary variables like Exclusive Mentholated Cigarette User, logistic regression (logreg) was employed, while multi-category variable Income was imputed using polytomous regression (polyreg). These methods ensure that the imputed values respect the categorical nature of these variables, maintaining their integrity in subsequent analyses.

The multiple imputation procedure assumes that the missing data mechanism follows the missing at random (MAR) assumption. This means the likelihood of missing data is related to other observed variables in the dataset but not to the unobserved values themselves. For instance, missingness in variables like Income or Nicotine Metabolism Ratio appeared to be associated with demographic and smoking behavior variables rather than the underlying value of these variables. This assumption aligns with the MAR framework, ensuring that the imputation process remains valid.

Figure 2 displays the mean and standard deviation of imputed values across five iterations for multiple variables, highlighting the convergence and stability of the multiple imputation process. Each line represents one of the imputed datasets, showing how the means and SDs fluctuate across iterations. For most variables, both mean and SD plots exhibit convergence, indicating that the imputation algorithm has stabilized. However, for FTCD score, no SD plot is generated because this variable has only one missing value. These plots confirm that the multiple imputation procedure has effectively filled in missing values while maintaining reasonable consistency with the observed data.

All our subsequent analyses used the imputed data.

Exploratory data analysis

Raw data

Table 1: Population Characteristic by Behavioral Therapy

Characteristic	**BASC** N = 151	**ST** N = 149
Smoking Abstinence		
0	121 (80%)	115 (77%)
1	30 (20%)	34 (23%)
Pharmacotherapy		
0	68 (45%)	68 (46%)
1	83 (55%)	81 (54%)
Age	53 (41, 60)	52 (43, 58)
Sex		
1	69 (46%)	66 (44%)
2	82 (54%)	83 (56%)
Non-Hispanic White		
0	93 (62%)	102 (68%)
1	58 (38%)	47 (32%)
Black		
0	77 (51%)	66 (44%)
1	74 (49%)	83 (56%)
Hispanic		
0	142 (94%)	140 (94%)
1	9 (6.0%)	9 (6.0%)
Income		
1	55 (37%)	55 (37%)
2	33 (22%)	35 (24%)
3	21 (14%)	25 (17%)
4	24 (16%)	14 (9.5%)
5	16 (11%)	19 (13%)
Missing	2	1
Education		
1	1 (0.7%)	0 (0%)
2	10 (6.6%)	6 (4.0%)
3	38 (25%)	38 (26%)
4	54 (36%)	62 (42%)
5	48 (32%)	43 (29%)
FTCD score at baseline	5.00 (4.00, 7.00)	6.00 (4.00, 7.00)
Missing	0	1
Smoking with 5 mins of waking up		
0	86 (57%)	76 (51%)
1	65 (43%)	73 (49%)
BDI score at baseline	18 (9, 26)	18 (11, 26)
Cigarettes per day at baseline	15 (10, 20)	14 (10, 20)
Cigarette reward value at baseline	7.0 (5.0, 10.0)	7.0 (5.0, 9.0)
Missing	4	14
Pleasurable Events Scale at baseline – substitute reinforcers	20 (9, 31)	16 (9, 32)
Pleasurable Events Scale at baseline – complementary reinforcers	22 (12, 32)	22 (12, 37)
Anhedonia	1.00 (0.00, 3.00)	1.00 (0.00, 3.00)
Missing	2	1
Other lifetime DSM-5 diagnosis		
0	86 (57%)	81 (54%)
1	65 (43%)	68 (46%)
Taking antidepressant medication at baseline		
0	99 (66%)	119 (80%)
1	52 (34%)	30 (20%)
Current vs past MDD		
0	79 (52%)	74 (50%)
1	72 (48%)	75 (50%)
Nicotine Metabolism Ratio	0.32 (0.22, 0.48)	0.32 (0.20, 0.46)
Missing	10	11
Exclusive Mentholated Cigarette User		
0	62 (41%)	58 (39%)
1	88 (59%)	90 (61%)
Missing	1	1

Table 1: Population Characteristic by Behavioral Therapy *(continued)*

Characteristic	**BASC**	**ST**
	N = 151	N = 149
Baseline readiness to quit smoking	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)
Missing	9	8

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

Table 2: Population Characteristic by Therapy Combination

Characteristic	**BASC + Placebo**	**BASC + Varenicline**	**ST + Placebo**	**ST + Varenicline**
	N = 68	N = 83	N = 68	N = 81
Smoking Abstinence				
0	64 (94%)	57 (69%)	60 (88%)	55 (68%)
1	4 (5.9%)	26 (31%)	8 (12%)	26 (32%)
Age	54 (42, 61)	53 (40, 60)	51 (45, 58)	52 (41, 59)
Sex				
1	30 (44%)	39 (47%)	29 (43%)	37 (46%)
2	38 (56%)	44 (53%)	39 (57%)	44 (54%)
Non-Hispanic White				
0	44 (65%)	49 (59%)	46 (68%)	56 (69%)
1	24 (35%)	34 (41%)	22 (32%)	25 (31%)
Black				
0	31 (46%)	46 (55%)	28 (41%)	38 (47%)
1	37 (54%)	37 (45%)	40 (59%)	43 (53%)
Hispanic				
0	63 (93%)	79 (95%)	64 (94%)	76 (94%)
1	5 (7.4%)	4 (4.8%)	4 (5.9%)	5 (6.2%)
Income				
1	25 (37%)	30 (37%)	26 (38%)	29 (36%)
2	16 (24%)	17 (21%)	14 (21%)	21 (26%)
3	8 (12%)	13 (16%)	14 (21%)	11 (14%)
4	12 (18%)	12 (15%)	8 (12%)	6 (7.5%)
5	6 (9.0%)	10 (12%)	6 (8.8%)	13 (16%)
Missing	1	1	0	1
Education				
1	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)
2	3 (4.4%)	7 (8.4%)	2 (2.9%)	4 (4.9%)
3	23 (34%)	15 (18%)	11 (16%)	27 (33%)
4	22 (32%)	32 (39%)	38 (56%)	24 (30%)
5	19 (28%)	29 (35%)	17 (25%)	26 (32%)
FTCD score at baseline	5.00 (4.00, 7.00)	5.00 (4.00, 7.00)	6.00 (4.00, 7.00)	5.00 (4.00, 7.00)
Missing	0	0	1	0
Smoking with 5 mins of waking up				
0	36 (53%)	50 (60%)	33 (49%)	43 (53%)
1	32 (47%)	33 (40%)	35 (51%)	38 (47%)
BDI score at baseline	18 (9, 27)	18 (10, 25)	18 (12, 25)	18 (11, 27)
Cigarettes per day at baseline	15 (10, 20)	15 (10, 20)	13 (10, 20)	15 (10, 20)
Cigarette reward value at baseline	7.0 (5.0, 10.0)	8.0 (4.5, 10.0)	7.0 (4.5, 9.0)	7.0 (5.0, 9.0)
Missing	1	3	8	6
Pleasurable Events Scale at baseline – substitute reinforcers	21 (10, 31)	20 (9, 32)	14 (9, 27)	20 (9, 35)
Pleasurable Events Scale at baseline – complementary reinforcers	23 (14, 34)	17 (11, 31)	25 (12, 38)	21 (13, 34)
Anhedonia	0.00 (0.00, 3.00)	1.00 (0.00, 4.00)	1.00 (0.00, 5.00)	1.00 (0.00, 3.00)
Missing	2	0	1	0
Other lifetime DSM-5 diagnosis				
0	33 (49%)	53 (64%)	40 (59%)	41 (51%)
1	35 (51%)	30 (36%)	28 (41%)	40 (49%)
Taking antidepressant medication at baseline				
0	40 (59%)	59 (71%)	53 (78%)	66 (81%)
1	28 (41%)	24 (29%)	15 (22%)	15 (19%)
Current vs past MDD				
0	36 (53%)	43 (52%)	37 (54%)	37 (46%)
1	32 (47%)	40 (48%)	31 (46%)	44 (54%)
Nicotine Metabolism Ratio	0.32 (0.23, 0.46)	0.33 (0.22, 0.50)	0.32 (0.20, 0.43)	0.29 (0.20, 0.51)
Missing	7	3	2	9
Exclusive Mentholated Cigarette User				
0	28 (41%)	34 (41%)	24 (36%)	34 (42%)
1	40 (59%)	48 (59%)	43 (64%)	47 (58%)
Missing	0	1	1	0
Baseline readiness to quit smoking	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)
Missing	4	5	4	4

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

Table 1 presents the population characteristics stratified by behavioral therapy groups. Among the participants, 151 individuals received Behavioral Activation for Smoking Cessation (BASC), while 149 received Standard Therapy (ST), resulting in nearly equal group sizes. The two groups exhibit similar characteristics across most variables, with the exception of the Cigarette Reward Value at Baseline variable. Notably, 14 participants in the ST group had missing data for this variable, compared to only 4 participants in the BASC group.

Table 2 provides the population characteristics stratified by the combination of behavioral and pharmacological therapies. Of the total participants, 164 received varenicline, while 136 received placebo. Notably, the groups receiving varenicline, regardless of behavioral therapy type, demonstrated the highest smoking abstinence rates at 32%. Across other baseline characteristics, the four groups exhibited comparable distributions.

Imputed data

Table 3: Population Characteristic by Behavioral Therapy

Characteristic	**BASC**	**ST**
	N = 151	N = 149
.imp		
1	151 (100%)	149 (100%)
Smoking Abstinence		
0	121 (80%)	115 (77%)
1	30 (20%)	34 (23%)
Pharmacotherapy		
0	68 (45%)	68 (46%)
1	83 (55%)	81 (54%)
Age	53 (41, 60)	52 (43, 58)
Sex		
1	69 (46%)	66 (44%)
2	82 (54%)	83 (56%)
Non-Hispanic White		
0	93 (62%)	102 (68%)
1	58 (38%)	47 (32%)
Black		
0	77 (51%)	66 (44%)
1	74 (49%)	83 (56%)
Hispanic		
0	142 (94%)	140 (94%)
1	9 (6.0%)	9 (6.0%)
Income		
1	56 (37%)	55 (37%)
2	34 (23%)	36 (24%)
3	21 (14%)	25 (17%)
4	24 (16%)	14 (9.4%)
5	16 (11%)	19 (13%)
Education		
1	1 (0.7%)	0 (0%)
2	10 (6.6%)	6 (4.0%)
3	38 (25%)	38 (26%)
4	54 (36%)	62 (42%)
5	48 (32%)	43 (29%)
FTCD score at baseline	5.00 (4.00, 7.00)	6.00 (4.00, 7.00)
Smoking with 5 mins of waking up		
0	86 (57%)	76 (51%)
1	65 (43%)	73 (49%)
BDI score at baseline	18 (9, 26)	18 (11, 26)
Cigarettes per day at baseline	15 (10, 20)	14 (10, 20)
Cigarette reward value at baseline	7.0 (5.0, 10.0)	7.0 (5.0, 9.0)
Pleasurable Events Scale at baseline – substitute reinforcers	20 (9, 31)	16 (9, 32)
Pleasurable Events Scale at baseline – complementary reinforcers	22 (12, 32)	22 (12, 37)
Anhedonia	1.00 (0.00, 4.00)	1.00 (0.00, 3.00)
Other lifetime DSM-5 diagnosis		
0	86 (57%)	81 (54%)
1	65 (43%)	68 (46%)
Taking antidepressant medication at baseline		
0	99 (66%)	119 (80%)
1	52 (34%)	30 (20%)

Table 3: Population Characteristic by Behavioral Therapy *(continued)*

Characteristic	**BASC** N = 151	**ST** N = 149
Current vs past MDD		
0	79 (52%)	74 (50%)
1	72 (48%)	75 (50%)
Nicotine Metabolism Ratio	0.32 (0.22, 0.49)	0.32 (0.20, 0.47)
Exclusive Mentholated Cigarette User		
0	62 (41%)	59 (40%)
1	89 (59%)	90 (60%)
Baseline readiness to quit smoking	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)

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

Table 4: Population Characteristic by Therapy Combination

Characteristic	**BASC + Placebo** N = 68	**BASC + Varenicline** N = 83	**ST + Placebo** N = 68	**ST + Varenicline** N = 81
.imp				
1	68 (100%)	83 (100%)	68 (100%)	81 (100%)
Smoking Abstinence				
0	64 (94%)	57 (69%)	60 (88%)	55 (68%)
1	4 (5.9%)	26 (31%)	8 (12%)	26 (32%)
Age	54 (42, 61)	53 (40, 60)	51 (45, 58)	52 (41, 59)
Sex				
1	30 (44%)	39 (47%)	29 (43%)	37 (46%)
2	38 (56%)	44 (53%)	39 (57%)	44 (54%)
Non-Hispanic White				
0	44 (65%)	49 (59%)	46 (68%)	56 (69%)
1	24 (35%)	34 (41%)	22 (32%)	25 (31%)
Black				
0	31 (46%)	46 (55%)	28 (41%)	38 (47%)
1	37 (54%)	37 (45%)	40 (59%)	43 (53%)
Hispanic				
0	63 (93%)	79 (95%)	64 (94%)	76 (94%)
1	5 (7.4%)	4 (4.8%)	4 (5.9%)	5 (6.2%)
Income				
1	25 (37%)	31 (37%)	26 (38%)	29 (36%)
2	17 (25%)	17 (20%)	14 (21%)	22 (27%)
3	8 (12%)	13 (16%)	14 (21%)	11 (14%)
4	12 (18%)	12 (14%)	8 (12%)	6 (7.4%)
5	6 (8.8%)	10 (12%)	6 (8.8%)	13 (16%)
Education				
1	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)
2	3 (4.4%)	7 (8.4%)	2 (2.9%)	4 (4.9%)
3	23 (34%)	15 (18%)	11 (16%)	27 (33%)
4	22 (32%)	32 (39%)	38 (56%)	24 (30%)
5	19 (28%)	29 (35%)	17 (25%)	26 (32%)
FTCD score at baseline	5.00 (4.00, 7.00)	5.00 (4.00, 7.00)	6.00 (4.50, 7.00)	5.00 (4.00, 7.00)
Smoking with 5 mins of waking up				
0	36 (53%)	50 (60%)	33 (49%)	43 (53%)
1	32 (47%)	33 (40%)	35 (51%)	38 (47%)
BDI score at baseline	18 (9, 27)	18 (10, 25)	18 (12, 25)	18 (11, 27)
Cigarettes per day at baseline	15 (10, 20)	15 (10, 20)	13 (10, 20)	15 (10, 20)
Cigarette reward value at baseline	7.0 (5.0, 10.0)	8.0 (5.0, 10.0)	7.0 (4.0, 9.0)	7.0 (5.0, 9.0)
Pleasurable Events Scale at baseline – substitute reinforcers	21 (10, 31)	20 (9, 32)	14 (9, 27)	20 (9, 35)
Pleasurable Events Scale at baseline – complementary reinforcers	23 (14, 34)	17 (11, 31)	25 (12, 38)	21 (13, 34)
Anhedonia	0.00 (0.00, 3.00)	1.00 (0.00, 4.00)	1.00 (0.00, 5.00)	1.00 (0.00, 3.00)
Other lifetime DSM-5 diagnosis				
0	33 (49%)	53 (64%)	40 (59%)	41 (51%)
1	35 (51%)	30 (36%)	28 (41%)	40 (49%)
Taking antidepressant medication at baseline				
0	40 (59%)	59 (71%)	53 (78%)	66 (81%)
1	28 (41%)	24 (29%)	15 (22%)	15 (19%)
Current vs past MDD				
0	36 (53%)	43 (52%)	37 (54%)	37 (46%)
1	32 (47%)	40 (48%)	31 (46%)	44 (54%)
Nicotine Metabolism Ratio	0.31 (0.21, 0.45)	0.34 (0.22, 0.51)	0.32 (0.20, 0.44)	0.30 (0.20, 0.51)
Exclusive Mentholated Cigarette User				
0	28 (41%)	34 (41%)	25 (37%)	34 (42%)
1	40 (59%)	49 (59%)	43 (63%)	47 (58%)
Baseline readiness to quit smoking	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)

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

Using the first imputed dataset, we generated updated versions of Table 1 and Table 2, resulting in Table 3 and Table 4, respectively. Upon examining Tables 3 and 4, it is evident that the imputed data maintain a balance across both classification schemes. The characteristics of various variables remain comparable across the different groups, demonstrating consistency and similarity post-imputation.

Correlation

Figure3. Correlation Heatmap of Continuous Variabl

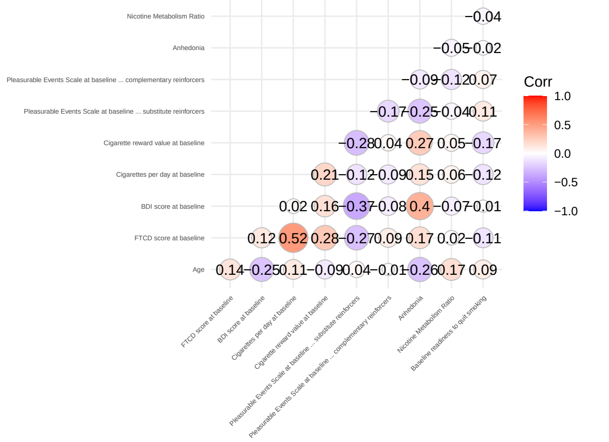
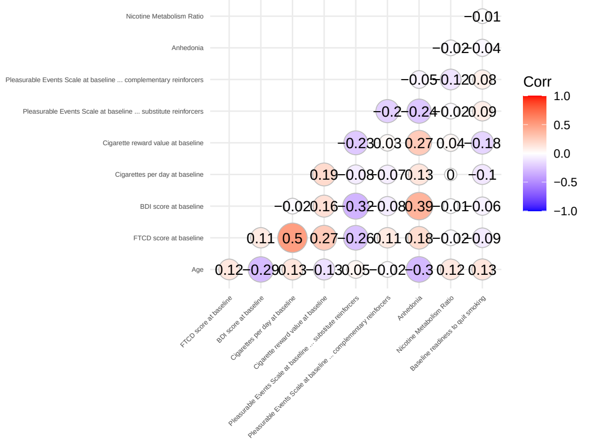


Figure4. Correlation Heatmap of Continuous Variabl



Figures 3 and 4 illustrate the correlation patterns of the raw data and imputed data, respectively, showing a high similarity between the two. Given this consistency, the imputed data will be used for subsequent analyses. As the maximum correlation observed in the figures is approximately 0.5, indicating only moderate collinearity among variables, Lasso regression will be employed for variable selection in the next steps.

Model Development

Research Question 1: Moderators of Psychotherapy: The goal of Research Question 1 is to identify baseline variables that may moderate the effect of behavioral therapy (BASC vs. ST) on smoking cessation. We opted for Lasso regression over Ridge regression due to its ability to perform variable selection by shrinking the coefficients of less important variables to zero through L1 regularization. This approach simplifies the model while retaining key predictors. Additionally, our correlation analysis revealed that the highest correlation between variables is only 0.5, indicating low multicollinearity. This further supports the use of Lasso, as Ridge regression is more appropriate for data with high multicollinearity.

In this analysis, pharmacotherapy (Varenicline or Placebo) is not included as a covariable. The focus is solely on the effect of Psychotherapy and its potential moderators. We assume that pharmacotherapy does not significantly interact with Psychotherapy in influencing smoking cessation outcomes.

The analysis will proceed as follows: (1) Lasso regression will be used on the first imputed data set to select important interactions between baseline variables and Psychotherapy. (2) Modified Poisson regression will be fitted to each imputed data set to estimate the effect of Psychotherapy and its moderators on the risk of smoking cessation, and the result of the first imputed data will be shown.

Research Question 2: Predictors of Smoking Cessation: The goal of Research Question 2 is to identify baseline variables that directly predict smoking cessation, controlling for both behavioral and pharmacological treatments. We again chose Lasso regression for variable selection due to its ability to handle numerous predictors and automatically select the most relevant ones.

In this analysis, we exclude interaction terms between baseline variables and treatment variables (Psychotherapy and pharmacological therapies). The focus here is on the direct predictive ability of baseline variables for smoking cessation, rather than exploring treatment effect moderation. Including interaction terms would add complexity and reduce model interpretability.

The analysis will proceed as follows: (1) Lasso regression will be used on the first imputed data set to select key baseline predictors. (2) Modified Poisson regression will be fitted to each imputed data set to estimate the direct effect of these predictors on smoking cessation, controlling for pharmacotherapy and Psychotherapy, and the result of the first imputed data will be shown.

Rationale for Using Modified Poisson Regression: We chose Modified Poisson regression to estimate risk ratios (RRs) for the binary outcome of smoking cessation. Risk ratios are more interpretable in this context compared to odds ratios from logistic regression. Furthermore, Modified Poisson regression, coupled with robust standard errors, provides reliable estimates even when the response variable is binary, making it well-suited for our data structure and research objectives.

Evaluation Metrics

Models were evaluated using a set of key performance metrics to assess both discrimination and calibration. Model discrimination was measured using the Area Under the Receiver Operating Characteristic Curve (AUC), which captures the model's ability to distinguish between positive (smoking cessation) and negative (continued smoking) cases. A higher AUC value indicates better discrimination performance. Model calibration was assessed using the Brier score, which measures the mean squared difference between predicted probabilities and actual outcomes. Lower Brier scores indicate better calibration, reflecting the model's accuracy in predicting class probabilities.

Results

Research Question 1: Moderators of Psychotherapy

We fit a lasso regression using the first imputed data. In this model we considered interaction terms between all baseline variables and Psychotherapy. A summary of the model coefficients are reported in Table 6.

Table 5: Lasso Selected Variables and Coefficients

	Variable	Coefficient
1	(Intercept)	-0.7675651
5	NHW1	0.2915827
16	ftcd_score	-0.1493408
26	mde_curr1	-0.0094625
27	NMR	0.3183828
55	BA1:Only.Menthol1	-0.0182745

Table 6: Modified Poisson Results from the First Imputation

	Estimate	Std.Error	z.value	p.value
(Intercept)	-1.265	0.386	-3.273	0.001
NHW1	0.528	0.246	2.147	0.032
ftcd_score	-0.179	0.047	-3.798	0.000
NMR	0.763	0.334	2.284	0.022
mde_curr1	-0.294	0.219	-1.348	0.178
Only.Menthol1	0.499	0.310	1.609	0.108
BA1	0.198	0.300	0.660	0.509
Only.Menthol1:BA1	-0.761	0.432	-1.761	0.078

The analysis aimed to investigate the effect of baseline covariates on smoking abstinence and whether these effects differ based on behavioral therapy. The results indicate that several baseline variables significantly influence smoking cessation outcomes. Specifically, the odds of smoking abstinence were significantly higher for Non-Hispanic White individuals (NHW1:$\beta = 0.53$, p<0.05) and for those with lower FTCD scores (FTCD: $\beta = -0.18$, p<0.001), suggesting that these factors are associated with better cessation outcomes.

Moreover, Nicotine Metabolism Ratio (NMR: $\beta = 0.76$, p<0.05) was positively associated with abstinence, indicating that individuals with higher nicotine metabolism rates may have an increased likelihood of quitting. However, other variables such as current major depressive disorder (mde_curr1: $\beta = -0.29$, p=0.18) and the interaction between behavioral therapy and menthol smoking (Only.Menthol1:BA1: $\beta = -0.76$, p=0.08) were not statistically significant, though they showed trends towards influencing abstinence outcomes.

These findings suggest that baseline demographic and smoking-related characteristics, such as nicotine dependence and metabolic rate, are important predictors of smoking cessation success, while the interaction effects between behavioral and pharmacological interventions warrant further investigation.

Research Question 2: Predictors of Smoking Cessation

We fit a lasso regression using the first imputed data. In this model we considered all baseline variables, Psychotherapy and Pharmacotherapy as main effects. A summary of the model coefficients are reported in Table 8.

Table 7: Lasso Selected Variables and Coefficients

	Variable	Coefficient
1	(Intercept)	-1.4040141
3	Var1	1.0835397
6	NHW1	0.3643686
17	ftcd_score	-0.1549119
24	shaps_score_pq1	-0.0049782
27	mde_curr1	-0.1341249
28	NMR	0.3725556

Table 8: Modified Poisson Results from the First Imputation

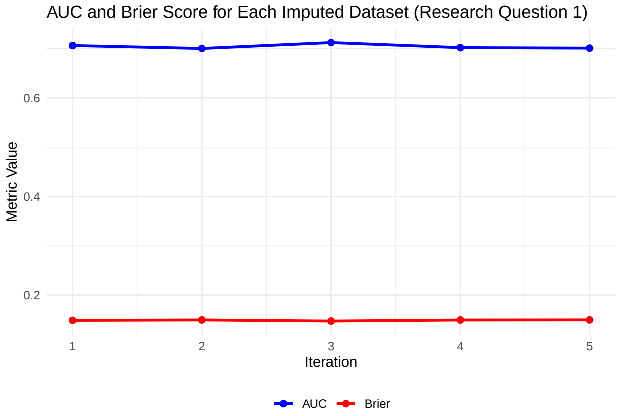
	Estimate	Std.Error	z.value	p.value
(Intercept)	-1.866	0.412	-4.533	0.000
Var1	1.207	0.288	4.184	0.000
NHW1	0.463	0.221	2.100	0.036
ftcd_score	-0.145	0.050	-2.893	0.004
shaps_score_pq1	-0.047	0.040	-1.190	0.234
mde_curr1	-0.297	0.221	-1.345	0.179
NMR	0.557	0.322	1.727	0.084

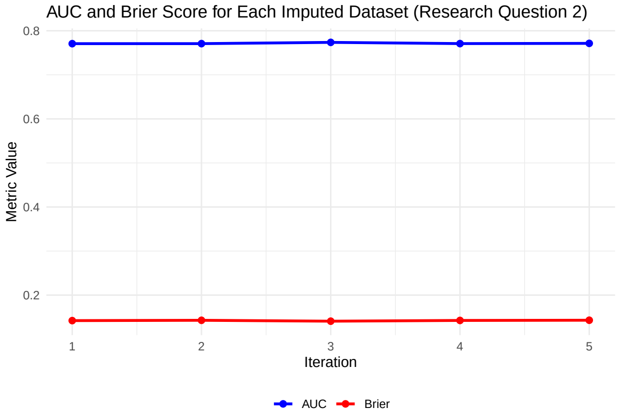
The analysis explored the effects of baseline covariates on smoking abstinence, controlling for pharmacotherapy and behavioral therapy. The results show that several baseline variables significantly influence smoking cessation. Pharmacotherapy with varenicline (Var1: $\beta = 1.21$, p<0.001) was strongly associated with higher odds of smoking abstinence, highlighting its effectiveness in supporting cessation. Additionally, Non-Hispanic White individuals (NHW1: $\beta = 0.46$, p<0.05) and those with lower FTCD scores (FTCD: $\beta = -0.15$, p<0.01) were more likely to achieve abstinence.

However, other variables, including SHAPS scores (anhedonia measure) (shaps_score_pq1: $\beta = -0.05$, p=0.23), current major depressive disorder (mde_curr1:$\beta = -0.3$, p=0.18), and Nicotine Metabolism Ratio (NMR: $\beta = 0.56$, p=0.08), did not reach statistical significance. These findings suggest that while pharmacotherapy and demographic factors play a crucial role in smoking cessation, the impact of certain psychological and physiological characteristics may require further exploration in larger or more targeted studies.

Model Evaluation

Model evaluation included AUC and Brier score metrics to assess discrimination and calibration for each imputed data set.





For both research questions, the models demonstrate reasonable performance based on AUC and Brier scores across the imputed datasets. In Research Question 1, the AUC values range around 0.70, indicating moderate discrimination, with Brier scores around 0.15, suggesting acceptable calibration. In Research Question 2, the AUC values are higher, averaging around 0.77, indicating better discrimination, and the Brier scores around 0.14 reflect slightly improved calibration compared to Research Question 1.

Discussion

This study aimed to identify baseline variables that moderate the effectiveness of behavioral activation (BA) and pharmacotherapy (varenicline) on smoking cessation among individuals with a history of major depressive disorder (MDD). By employing a Modified Poisson regression model on multiply imputed datasets, we assessed the impact of various demographic, psychological, and physiological baseline factors. Our findings contribute to the understanding of how personalized treatment approaches can be optimized to improve smoking cessation outcomes.

For Research Question 1, the analysis revealed that several baseline characteristics significantly predicted smoking cessation among individuals receiving behavioral therapy. Non-Hispanic White status (NHW1), lower FTCD scores (indicating lower nicotine dependence), and higher Nicotine Metabolism Ratio (NMR) were associated with increased odds of smoking abstinence. However, the interaction between BA and menthol smoking status was not statistically significant, suggesting that BA's effect may not vary substantially across these subgroups. These findings highlight the importance of considering baseline characteristics when tailoring behavioral interventions for smoking cessation.

For Research Question 2, pharmacotherapy with varenicline was found to significantly improve smoking abstinence rates, with a strong positive effect. Additionally, demographic and smoking-related factors, such as NHW1 and lower FTCD scores, continued to show significant associations with cessation. Although current major depressive disorder (mde_curr1) and NMR approached significance, their impact was less pronounced, indicating that pharmacotherapy's benefits might be somewhat independent of these baseline characteristics.

Limitations Despite these findings, the study has several limitations. First, the generalizability of the results may be limited by the specific study population, which consisted of smokers with a history of MDD recruited from two sites. This demographic may not fully represent the broader population of smokers, particularly those without a history of MDD or from different socioeconomic backgrounds. Future studies should include more diverse samples to validate the generalizability of these findings.

Second, while we used multiple imputation to address missing data, the assumption of missingness at random (MAR) may not always hold. If missingness is related to unmeasured factors, the imputation process might introduce bias, potentially impacting the robustness of our findings. Sensitivity analyses using different missing data assumptions could help evaluate the extent of this issue.

Third, the study's observational nature and reliance on self-reported smoking abstinence could introduce recall and reporting biases. Self-reported outcomes may be subject to social desirability, leading participants to underreport smoking behavior. Incorporating biochemical verification of abstinence in future research could enhance the accuracy of the outcome assessment.

Conclusions

This project explored the baseline factors influencing the effectiveness of behavioral activation (BA) and varenicline in promoting smoking cessation among individuals with a history of major depressive disorder (MDD). Using Modified Poisson regression and multiple imputation to handle missing data, the analysis identified nicotine dependence and varenicline as significant predictors of smoking abstinence. Although behavioral activation alone did not show a significant effect, its interaction with menthol smoking status suggested potential subgroup variability. These findings underscore the importance of personalized treatment approaches and highlight the need for further research to optimize smoking cessation strategies in this high-risk population.

References

[1] Breslau, N., Kilbey, M. M., & Andreski, P. (1992). Nicotine withdrawal symptoms and psychiatric disorders: findings from an epidemiologic study of young adults. The American journal of psychiatry, 149(4), 464-469.

[2] Spring, B., Pingitore, R., & McChargue, D. E. (2003). Reward value of cigarette smoking for comparably heavy smoking schizophrenic, depressed, and nonpatient smokers. American Journal of Psychiatry, 160(2), 316-322.

[3] Weinberger, A. H., Desai, R. A., & McKee, S. A. (2010). Nicotine withdrawal in US smokers with current mood, anxiety, alcohol use, and substance use disorders. Drug and alcohol dependence, 108(1-2), 7-12.

[4] Lyons, M., Hitsman, B., Xian, H., Panizzon, M. S., Jerskey, B. A., Santangelo, S., . . . & Tsuang, M. T. (2008). A twin study of smoking, nicotine dependence, and major depression in men. Nicotine & Tobacco Research, 10(1), 97-108.

[5] Robert M. Anthenelli, Chad Morris, Tanya S. Ramey, et al. Effects of Varenicline on Smoking Cessation in Adults With Stably Treated Current or Past Major Depression: A Randomized Trial. Ann Intern Med.2013;159:390-400. [Epub 17 September 2013]. doi:10.7326/0003-4819-159-6-201309170-00005

[6] Cuijpers, P., Van Straten, A., & Warmerdam, L. (2007). Behavioral activation treatments of depression: A meta-analysis. Clinical psychology review, 27(3), 318-326.

[7] Dimidjian, S., Barrera Jr, M., Martell, C., Muñoz, R. F., & Lewinsohn, P. M. (2011). The origins and current status of behavioral activation treatments for depression. Annual review of clinical psychology, 7(1), 1-38.

[8] Hopko, D. R., Lejuez, C. W., Ruggiero, K. J., & Eifert, G. H. (2003). Contemporary behavioral activation treatments for depression: Procedures, principles, and progress. Clinical psychology review, 23(5), 699-717.

[9] Hitsman, B., Papandonatos, G. D., Gollan, J. K., Huffman, M. D., Niaura, R., Mohr, D. C., . . . & 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.

Code Appendix:

```r
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(gtsummary)
library(corrplot)
library(knitr)
library(mice)
library(L0Learn)
library(lme4)
library(caret)
library(pROC)
library(naniar)
# Read in and process data
df <- read.csv("/Users/fusei/Desktop/24FALL/PHP2550/Project2/project2.csv")
df$abst <- as.factor(df$abst)
df$Var <- as.factor(df$Var)
df$BA <- as.factor(df$BA)
df$age_ps <- as.numeric(df$age_ps)
df$sex_ps <- as.factor(df$sex_ps)
df$NHW <- as.factor(df$NHW)
df$Black <- as.factor(df$Black)
df$Hisp <- as.factor(df$Hisp)
df$inc <- as.factor(df$inc)
df$edu <- as.factor(df$edu)
df$ftcd_score <- as.numeric(df$ftcd_score)
df$ftcd.5.mins <- as.factor(df$ftcd.5.mins)
df$bdi_score_w00 <- as.numeric(df$bdi_score_w00)
df$cpd_ps <- as.numeric(df$cpd_ps)
df$crv_total_pq1 <- as.numeric(df$crv_total_pq1)
df$hedonsum_n_pq1 <- as.numeric(df$hedonsum_n_pq1)
df$hedonsum_y_pq1 <- as.numeric(df$hedonsum_y_pq1)
df$shaps_score_pq1 <- as.numeric(df$shaps_score_pq1)
df$otherdiag <- as.factor(df$otherdiag)
df$antidepmed <- as.factor(df$antidepmed)
df$mde_curr <- as.factor(df$mde_curr)
df$NMR <- as.numeric(df$NMR)
df$Only.Menthol <- as.factor(df$Only.Menthol)
df$readiness <- as.numeric(df$readiness)
df$id <- NULL
# Define variable names for correlation plot
df_eda <- df %>%
  rename(`Smoking Abstinence` = abst,
         `Pharmacotherapy` = Var,
         `Psychotherapy` = BA,
         `Age` = age_ps,
         `Sex` = sex_ps,
         `Non-Hispanic White` = NHW,
         `Black` = Black,
         `Hispanic` = Hisp,
         `Income` = inc,
         `Education` = edu,
         `FTCD score at baseline` = ftcd_score,
         `Smoking with 5 mins of waking up` = ftcd.5.mins,
         `BDI score at baseline` = bdi_score_w00,
         `Cigarettes per day at baseline` = cpd_ps,
         `Cigarette reward value at baseline` = 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,
         `Other lifetime DSM-5 diagnosis` = otherdiag,
         `Taking antidepressant medication at baseline` = antidepmed,
         `Current vs past MDD` = mde_curr,
         `Nicotine Metabolism Ratio` = NMR,
         `Exclusive Mentholated Cigarette User` = Only.Menthol,
         `Baseline readiness to quit smoking` = readiness)

### Missing Pattern
gg_miss_var(df_eda) + labs(y = "Number of missing", title  = "Figure1. Missing Pattern")


### Multiple Imputation
# Set the number of imputations
m = 5

# Run multiple imputation
mids <- mice(df, m = 5, method = c(
    "",
    "",
    "",
    "",
    "",
    "",
    "",
    "",
    "polyreg", #income
    "",
    "pmm", #"FTCD score at baseline"
    "",
    "",
    "",
    "pmm", #"Cigarette reward value at baseline"
    "",
    "",
    "pmm", #"Anhedonia"
    "",
    "",
    "",
    "pmm", #"Nicotine Metabolism Ratio"
    "logreg", #"Exclusive Mentholated Cigarette User"
    "pmm" #"Baseline readiness to quit smoking"
), seed = 2024, printFlag = FALSE)

completed_data <- complete(mids, action = "long")

## check convergence
```

```
par(mfrow = c(6, 2), mar = c(3, 3, 2, 1))
print(plot(mids, main = "Figure 2: Check Convergence"))
# Table 1
data_t1 <- df_eda %>%
  mutate(Therapy = case_when(
    Psychotherapy == 1  ~ "BASC",
    Psychotherapy == 0  ~ "ST "
  ))

data_t1 <- data_t1 %>% select(-Psychotherapy)
t1 <-
    tbl_summary(data_t1,
            by = Therapy,
            type = list(`Baseline readiness to quit smoking` ~ "continuous"),
            missing_text = "Missing") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Population Characteristic by Behavioral Therapy",
                 longtable = TRUE,
                 linesep = "") %>%
  kableExtra::kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))

t1
# Table 2
# data
data_t2 <- df_eda %>%
  mutate(Combined_Therapy = case_when(
    Psychotherapy == 1 & Pharmacotherapy == 1 ~ "BASC + Varenicline",
    Psychotherapy == 1 & Pharmacotherapy == 0 ~ "BASC + Placebo",
    Psychotherapy == 0 & Pharmacotherapy == 1 ~ "ST + Varenicline",
    Psychotherapy == 0 & Pharmacotherapy == 0 ~ "ST + Placebo"
  ))
data_t2 <- data_t2 %>% select(-c(Psychotherapy,Pharmacotherapy ) )

t2 <-
    tbl_summary(data_t2,
            by = Combined_Therapy,
            type = list(`Baseline readiness to quit smoking` ~ "continuous"),
            missing_text = "Missing") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Population Characteristic by Therapy Combination",
                 longtable = TRUE,
                 linesep = "") %>%
  kableExtra::kable_styling(font_size = 5,
                            latex_options = c("repeat_header", "HOLD_position"))

t2
df_eda_imp1<- completed_data[completed_data$.imp == 1,]
df_eda_imp1$.id <- NULL

# Define variable names for correlation plot
df_eda_imp1 <- df_eda_imp1 %>%
  rename(`Smoking Abstinence` = abst,
         `Pharmacotherapy` = Var,
         `Psychotherapy` = BA,
         `Age` = age_ps,
         `Sex` = sex_ps,
         `Non-Hispanic White` = NHW,
         `Black` = Black,
         `Hispanic` = Hisp,
         `Income` = inc,
         `Education` = edu,
         `FTCD score at baseline` = ftcd_score,
         `Smoking with 5 mins of waking up` = ftcd.5.mins,
         `BDI score at baseline` = bdi_score_w00,
         `Cigarettes per day at baseline` = cpd_ps,
         `Cigarette reward value at baseline` = 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,
         `Other lifetime DSM-5 diagnosis` = otherdiag,
         `Taking antidepressant medication at baseline` = antidepmed,
         `Current vs past MDD` = mde_curr,
         `Nicotine Metabolism Ratio` = NMR,
         `Exclusive Mentholated Cigarette User` = Only.Menthol,
         `Baseline readiness to quit smoking` = readiness)
# Table 3
data_t1 <- df_eda_imp1 %>%
  mutate(Therapy = case_when(
    Psychotherapy == 1  ~ "BASC",
    Psychotherapy == 0  ~ "ST "
  ))

data_t1 <- data_t1 %>% select(-Psychotherapy)

tbl_summary(data_t1,
            by = Therapy,
            type = list(`Baseline readiness to quit smoking` ~ "continuous"),
            missing_text = "Missing") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Population Characteristic by Behavioral Therapy",
                 longtable = TRUE,
                 linesep = "") %>%
  kableExtra::kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))



# Table 4
# data
data_t2 <- df_eda_imp1 %>%
  mutate(Combined_Therapy = case_when(
    Psychotherapy == 1 & Pharmacotherapy == 1 ~ "BASC + Varenicline",
    Psychotherapy == 1 & Pharmacotherapy == 0 ~ "BASC + Placebo",
    Psychotherapy == 0 & Pharmacotherapy == 1 ~ "ST + Varenicline",
    Psychotherapy == 0 & Pharmacotherapy == 0 ~ "ST + Placebo"
```

```r
  ))
data_t2 <- data_t2 %>% select(-c(Psychotherapy,Pharmacotherapy ) )

t4 <-
    tbl_summary(data_t2,
            by = Combined_Therapy,
            type = list(`Baseline readiness to quit smoking` ~ "continuous"),
            missing_text = "Missing") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Population Characteristic by Therapy Combination",
                 longtable = TRUE,
                 linesep = "") %>%
  kableExtra::kable_styling(font_size = 5,
                            latex_options = c("repeat_header", "HOLD_position"))

t4
library(ggcorrplot)
#continuous
continuous_vars <- c("Age",
                     "FTCD score at baseline",
                     "BDI score at baseline",
                     "Cigarettes per day at baseline",
                     "Cigarette reward value at baseline",
                     "Pleasurable Events Scale at baseline - substitute reinforcers",
                     "Pleasurable Events Scale at baseline - complementary reinforcers",
                     "Anhedonia",
                     "Nicotine Metabolism Ratio",
                     "Baseline readiness to quit smoking")

## raw
cor_matrix1 <- cor(df_eda[,continuous_vars], use = "complete.obs")

f3 <- ggcorrplot(cor_matrix1,
           method = "circle",
           type = "lower",
           lab = TRUE,
           title = "Figure3. Correlation Heatmap of Continuous Variables of Raw Data",
           tl.cex = 5,
           ggtheme = theme_minimal())
## imputed
cor_matrix2 <- cor(df_eda_imp1[continuous_vars], use = "complete.obs")
f4 <- ggcorrplot(cor_matrix2,
           method = "circle",
           type = "lower",
           lab = TRUE,
           title = "Figure4. Correlation Heatmap of Continuous Variables of Imputed Data",
           tl.cex = 5,
           ggtheme = theme_minimal())
print(f3)
print(f4)

metrics_ex <- function(models,iter,mids){
    df <- data.frame(iteration = as.character(), AUC = as.numeric(), Brier = as.numeric())
    for(i in seq_len(iter)){
        imp <- complete(mids, action = i)
        y_obs <- as.numeric(as.character(imp$abst))
        y_pred <- predict(models$analyses[[i]], type = "response")
        roc_obj <- pROC::roc(response = y_obs, predictor = y_pred)
        auc_value <- roc_obj$auc
        brier_score <- mean((y_pred - y_obs)^2)
        dfi <- data.frame(iteration = i, AUC = auc_value, Brier = brier_score)
        df <- rbind(df,dfi)
    }
    return(df)
}
## lasso1
library(glmnet)
imputed_data1 <- complete(mids, action = 1)
x <- model.matrix(abst ~  BA * (age_ps + sex_ps + NHW + Black + Hisp +
                  inc + edu + ftcd_score + ftcd.5.mins + bdi_score_w00 +
                  cpd_ps + crv_total_pq1 + hedonsum_n_pq1 +
                  hedonsum_y_pq1 + shaps_score_pq1 + otherdiag +
                  antidepmed + mde_curr + NMR + Only.Menthol + readiness),
                  data = imputed_data1)[,-1]
y <- imputed_data1$abst
set.seed(2000)
lasso_fit1 <- cv.glmnet(x, y, family = "binomial", alpha = 1)

coef_lasso1 <- as.matrix(coef(lasso_fit1, s = "lambda.min"))
coef_table1 <- data.frame(
  Variable = rownames(coef_lasso1),
  Coefficient = as.numeric(coef_lasso1)
)

## select coef != 0
## these variables should be put in modified model
selected_vars1 <- coef_table1[coef_table1$Coefficient != 0, ]
kable(selected_vars1, caption = "Lasso Selected Variables and Coefficients")

library(sandwich)
library(lmtest)

mids2 <- mids %>%
  complete(action = "long", include = TRUE) %>%
  mutate(abst = as.integer(as.character(abst))) %>%
  as.mids()

models_r1 <- with(mids2, glm(formula = abst ~ NHW + ftcd_score + NMR + mde_curr + Only.Menthol + BA + BA

#pooled_results <- pool(models)

##modified poisson
models_adjusted_r1 <- lapply(models_r1$analyses, function(model) {
  coeftest(model, vcov = vcovHC(model, type = "HC0"))
})
```

```r
adjusted_coefs_r1 <- do.call(rbind,lapply(models_adjusted_r1, function(x) {
  data.frame(
    Estimate = x[, "Estimate"],
    Std.Error = x[, "Std. Error"],
    z.value = x[, "z value"],
    p.value = x[, "Pr(>|z|)"]
  )
}))

# pooled_table <- data.frame(
#   Term = pooled_summary$term,
#   Estimate = pooled_summary$estimate,
#   `Std.Error` = pooled_summary$std.error,
#   `z value` = pooled_summary$statistic,
#   `p value` = pooled_summary$p.value
# )


kable(adjusted_coefs_r1[1:8,], digits = 3, caption = "Modified Poisson Results from the First Imputatio
## lasso1
library(glmnet)
x <- model.matrix(abst ~ BA + Var + BA*Var + age_ps + sex_ps + NHW + Black + Hisp +
                  inc + edu + ftcd_score + ftcd.5.mins + bdi_score_w00 +
                  cpd_ps + crv_total_pq1 + hedonsum_n_pq1 +
                  hedonsum_y_pq1 + shaps_score_pq1 + otherdiag +
                  antidepmed + mde_curr + NMR + Only.Menthol +
                  readiness,
                  data = imputed_data1)[,-1]
y <- imputed_data1$abst
lasso_fit <- cv.glmnet(x, y, family = "binomial", alpha = 1)

coef_lasso <- as.matrix(coef(lasso_fit, s = "lambda.min"))
coef_table <- data.frame(
  Variable = rownames(coef_lasso),
  Coefficient = as.numeric(coef_lasso)
)

## select coef != 0
## these variables should be put in modified model
selected_vars <- coef_table[coef_table$Coefficient != 0, ]
kable(selected_vars, caption = "Lasso Selected Variables and Coefficients")


models_r2 <- with(mids2, glm(formula = abst ~ Var + NHW + ftcd_score + shaps_score_pq1 + mde_curr + NMR

models_adjusted_r2 <- lapply(models_r2$analyses, function(model) {
  coeftest(model, vcov = vcovHC(model, type = "HC0"))
})

adjusted_coefs_r2 <- do.call(rbind,lapply(models_adjusted_r2, function(x) {
  data.frame(
    Estimate = x[, "Estimate"],
    Std.Error = x[, "Std. Error"],
    z.value = x[, "z value"],
    p.value = x[, "Pr(>|z|)"]
  )
}))

kable(adjusted_coefs_r2[1:7,], digits = 3, caption = "Modified Poisson Results from the First Imputatio

metrics_ex <- function(models,iter,mids){
    df <- data.frame(iteration = as.character(), AUC = as.numeric(), Brier = as.numeric())
    for(i in seq_len(iter)){
        imp <- complete(mids, action = i)
        y_obs <- as.numeric(as.character(imp$abst))
        y_pred <- predict(models$analyses[[i]], type = "response")
        roc_obj <- pROC::roc(response = y_obs, predictor = y_pred, quiet = TRUE)
        auc_value <- as.numeric(roc_obj$auc)
        brier_score <- mean((y_pred - y_obs)^2)
        dfi <- data.frame(iteration = i, AUC = auc_value, Brier = brier_score)
        df <- rbind(df,dfi)
    }
    return(df)
}

rq1_plotdata <- metrics_ex(models_r1,5,mids2)
rq2_plotdata <- metrics_ex(models_r2,5,mids2)

rq1_long <- rq1_plotdata %>%
  pivot_longer(cols = c(AUC, Brier), names_to = "Metric", values_to = "Value")

fig5 <- ggplot(rq1_long, aes(x = iteration, y = Value, color = Metric)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(
    title = "AUC and Brier Score for Each Imputed Dataset (Research Question 1)",
    x = "Iteration",
    y = "Metric Value"
  ) +
  scale_color_manual(values = c("AUC" = "blue", "Brier" = "red")) +
  theme_minimal() +
  theme(
    legend.title = element_blank(),
    legend.position = "bottom"
  )

rq2_long <- rq2_plotdata %>%
  pivot_longer(cols = c(AUC, Brier), names_to = "Metric", values_to = "Value")

fig6 <- ggplot(rq2_long, aes(x = iteration, y = Value, color = Metric)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(
    title = "AUC and Brier Score for Each Imputed Dataset (Research Question 2)",
    x = "Iteration",
    y = "Metric Value"
  ) +
  scale_color_manual(values = c("AUC" = "blue", "Brier" = "red")) +
  theme_minimal() +
  theme(
    legend.title = element_blank(),
    legend.position = "bottom"
  )
print(fig5)
print(fig6)
```