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 age at phone interview and sex at phone interview 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 and interactions. AUC and Briei were used to evaluate models.

Results: The results indicate that Non-Hispanic White status, lower FTCD scores (indicating lower nicotine dependence), and varenicline use were significant predictors of smoking abstinence. Varenicline demonstrated a strong positive effect on cessation rates ($\beta = 1.20$, $p < 0.001$), while behavioral therapy alone did not show significant subgroup variation. Nicotine Metabolism Ratio (NMR) exhibited a positive trend ($\beta = 0.54$, $p = 0.075$) but did not reach statistical significance. Overall, the models showed good discrimination (pooled AUC = 0.76) and calibration (pooled Brier score = 0.14), indicating reliable predictive performance across imputations.

Conclusion: This study highlights the critical role of nicotine dependence, demographic factors, and pharmacotherapy in smoking cessation among individuals with MDD. Varenicline consistently improved cessation outcomes, regardless of baseline characteristics, underscoring its importance in smoking cessation strategies. While behavioral activation alone did not uniformly enhance cessation rates, the observed trends in associations with baseline characteristics, such as NMR, warrant further investigation. These findings emphasize the need for personalized smoking cessation strategies that incorporate individual demographic and smoking-related baseline characteristics to optimize outcomes.

Introduction

Smoking is a leading cause of preventable death, associated with severe health outcomes such as cardiovascular disease and cancer [10]. 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

Table 1: Summary of Missing Data by Variable

Variable	Missing	Total	MissingPercent
Nicotine Metabolism Ratio	21	300	7.00
Cigarette reward value at baseline	18	300	6.00
Baseline readiness to quit smoking	17	300	5.67
Income	3	300	1.00
Anhedonia	3	300	1.00
Exclusive Mentholated Cigarette User	2	300	0.67
FTCD score at baseline	1	300	0.33

Using these 300 observations, the number of missing values for all 24 variables (except id) is plotted. Observing Table 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(n = 59).

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.

All our subsequent analyses used the imputed data.

Exploratory data analysis

Raw data

Table 2: Population Characteristic by Therapy Combination

Characteristic	**BASC + Placebo** N = 68	**BASC + Varenicline** N = 83	**ST + Placebo** N = 68	**ST + Varenicline** 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)

Table 2: Population Characteristic by Therapy Combination *(continued)*

Characteristic	**BASC + Placebo**	**BASC + Varenicline**	**ST + Placebo**	**ST + Varenicline**
	N = 68	N = 83	N = 68	N = 81
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

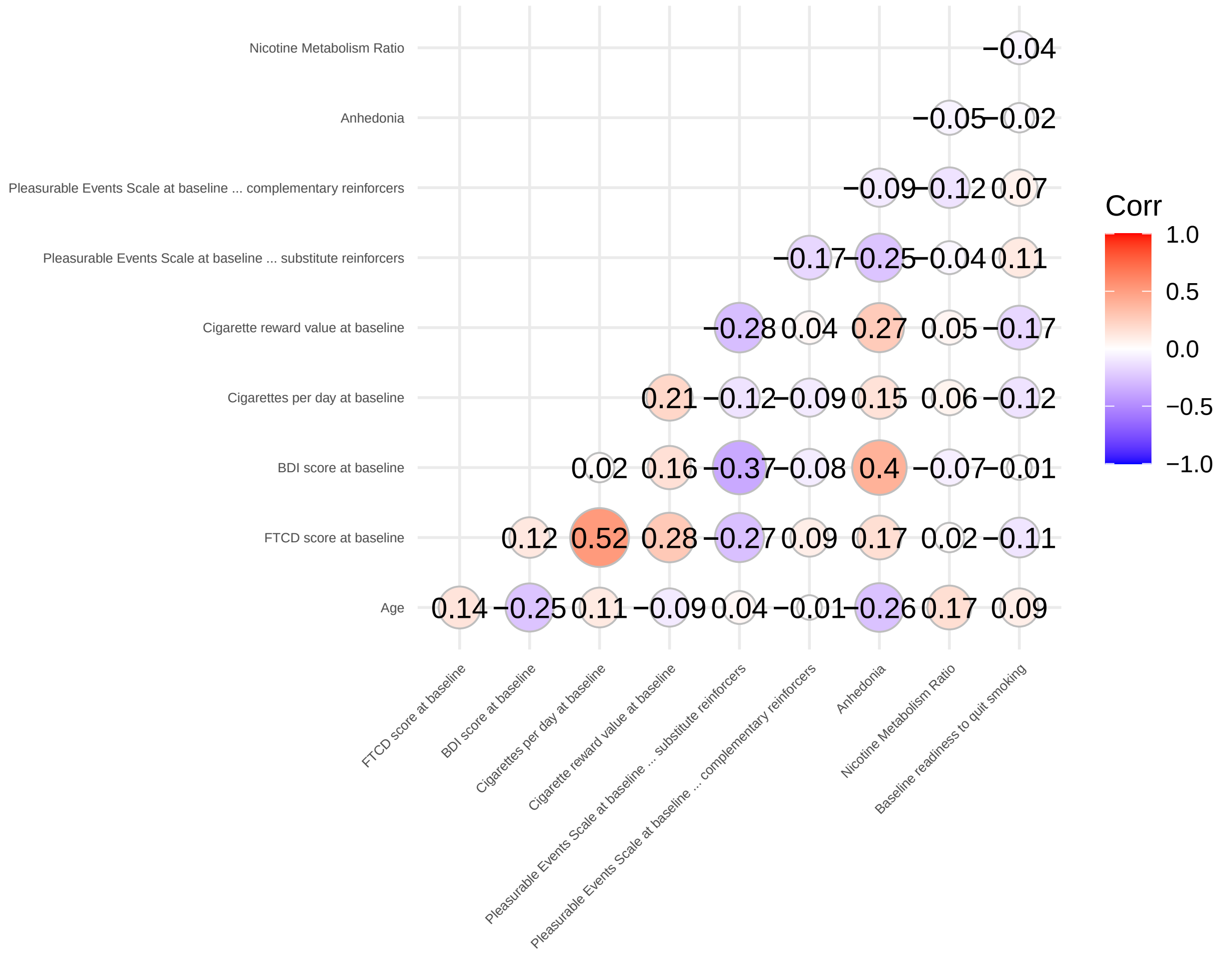
[1] n (%); Median (Q1, Q3)

Table 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

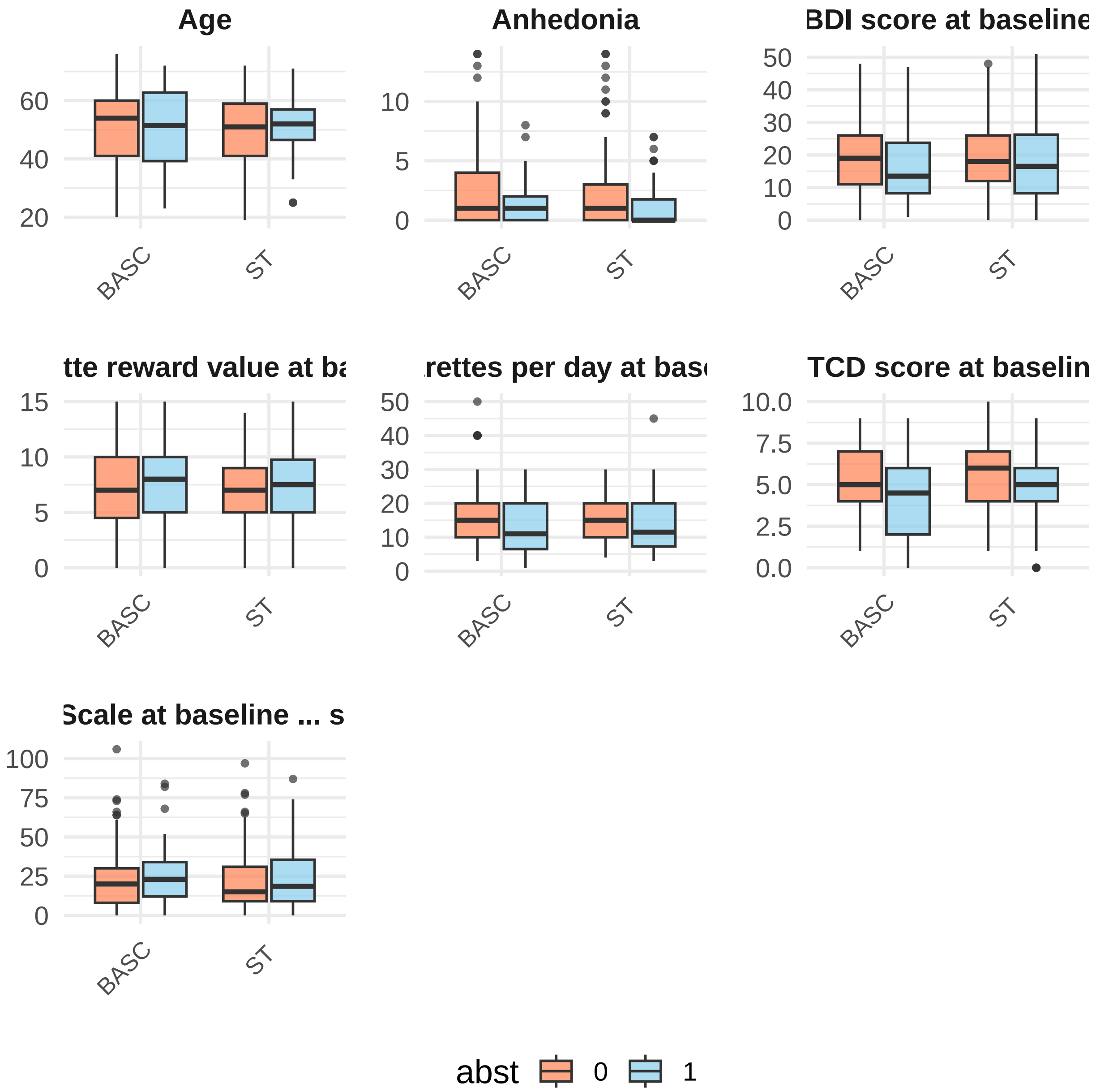
Correlation

Figure1. Correlation Heatmap of Continuous Variables



Figures 1 illustrate the correlation patterns of the raw data. 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.

Figure 2. Distribution of Baseline Variables by Behavioral T



No strong correlation was found between the continuous variables in the baseline features. Figure 2 shows the box plots grouped by behavioral activation and abstinence status to provide insights into potential moderators. Each box plot compares the smoking quitting status of the treatment groups (ST and BA). There is no significant difference in these characteristics between the treatment group and the non abstinence group, as the distribution overlaps to a large extent. Therefore, these may not be moderators that activate behavior.

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 supports the use of Lasso, as Ridge regression is more appropriate for data with high multicollinearity.

Furthermore, we considered the potential impact of high variance settings, where predictors may exhibit wide variability or where the relationship between predictors and the outcome is highly variable across the dataset. In such settings, Lasso is advantageous because it applies strong regularization to stabilize coefficient estimates, especially for less informative predictors. This helps prevent overfitting and enhances model interpretability by excluding variables with highly fluctuating or noisy contributions. In contrast, Ridge regression may retain all predictors, including those with high variance, leading to less interpretable results. Therefore, Lasso is particularly suited for identifying key moderators in this analysis, even in the presence of potentially high variability among baseline variables.

In this analysis, pharmacotherapy (Varenicline or Placebo) is included as a covariate to control for its potential effects, but no interactions between pharmacotherapy and Psychotherapy are considered. The primary focus remains on the effect of Psychotherapy and its potential moderators on smoking cessation outcomes.

The analysis will proceed as follows: (1) Relaxed Lasso regression is applied across all imputed datasets to select important baseline variables and their interactions with Psychotherapy. Variables selected in at least one imputed dataset are retained for further analysis. (2) Modified Poisson regression models are fitted to each imputed dataset using the selected variables as predictors to estimate the effect of Psychotherapy and its moderators on the risk of smoking cessation. Robust standard errors are computed for each model using the sandwich estimator (vcovHC with type HC0). (3) Results from all imputed datasets are pooled to calculate the mean estimates, standard errors, z-values, p-values, and 95% confidence intervals.

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) Relaxed Lasso regression is applied across all imputed datasets to select important baseline variables and their interactions with Psychotherapy. Variables selected in at least one imputed dataset are retained for further analysis. (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. Robust standard errors are computed for each model using the sandwich estimator (vcovHC with type HC0). (3) Results from all imputed datasets are pooled to calculate the mean estimates, standard errors, z-values, p-values, and 95% confidence intervals.

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 combination of key performance metrics to assess both discrimination and calibration. The evaluation process consisted of the following components:

1. Train-Test Split Validation: Each imputed dataset was split into 70% training data and 30% testing data. Models were trained on the training data and evaluated on the testing data to ensure robustness and avoid overfitting. For each imputation, the Area Under the Receiver Operating Characteristic Curve (AUC) and Brier Score were computed.
2. Discrimination (AUC): Discrimination was assessed using the AUC, which measures the model's ability to distinguish between positive (smoking cessation) and negative (continued smoking) cases. Higher AUC values indicate better discrimination. The AUC values across the five imputations were plotted to visualize their consistency. A pooled AUC was calculated across all imputations by combining the predictions, with an average AUC of 0.76, indicating good discrimination performance.
3. Calibration (Brier Score): Calibration was evaluated using the Brier Score, which measures the mean squared difference between predicted probabilities and observed outcomes. Lower Brier Scores indicate better calibration, reflecting accurate probability predictions. The pooled Brier Score across imputations was 0.14, suggesting good alignment between predicted and observed probabilities.

Predictions from all imputations were pooled to calculate overall AUC and Brier Score. This provided a summary metric to assess model performance across all imputations. The pooled AUC and Brier Score reflect the overall discrimination and calibration performance of the model.

5. AUC Plot: The AUC values across imputations were plotted to visualize model consistency. The plot revealed stable AUC values around 0.76, highlighting consistent discriminatory power across imputations.
6. Calibration Plot: The calibration plot visualized the alignment between predicted probabilities and observed outcomes. Points closely aligned with the diagonal indicate good calibration, confirming that the model's predicted probabilities closely match the observed probabilities.

Results

Research Question 1: Moderators of Psychotherapy

We fit a relaxed lasso regression using all 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 4.

Table 3: Selected Variables for Each Imputed Dataset

Variable	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5
(Intercept)	-1.393	-1.134	-1.697	-1.395	-1.398
Var1	1.385	1.567	1.370	1.386	1.372
NHW1	0.655	NA	0.526	0.656	0.641
ftcd_score	-0.216	-0.244	-0.219	-0.216	-0.213
NMR	NA	NA	0.980	NA	NA

Table 4: Modified Poisson Results with Robust Standard Errors

Term	Estimate	Std.Error	z.value	p.value	ci_low	ci_high
(Intercept)	-1.972	0.396	-4.986	0.000	-2.748	-1.197
NHW1	0.466	0.222	2.104	0.036	0.032	0.901
NMR	0.539	0.291	1.886	0.075	-0.032	1.110
Var1	1.198	0.291	4.119	0.000	0.628	1.768
ftcd_score	-0.167	0.049	-3.392	0.001	-0.264	-0.071

The analysis aimed to investigate the effect of baseline covariates on smoking abstinence and whether these effects differ based on behavioral therapy (BASC vs. ST). The results indicate that several baseline variables significantly influence smoking cessation outcomes.

Specifically, Non-Hispanic White individuals (NHW1: β = 0.47, p = 0.036) had higher rates of smoking abstinence compared to other racial groups, suggesting that race may play a role in cessation success. Additionally, higher Nicotine Metabolism Ratio (NMR: β = 0.54, p = 0.075) showed a positive association with abstinence, though it was marginally significant, indicating that individuals with faster nicotine metabolism may have an increased likelihood of quitting.

Lower FTCD scores, which measure nicotine dependence, were significantly associated with better cessation outcomes (FTCD: β = -0.17, p = 0.001), underscoring the importance of lower nicotine dependence for smoking cessation success. Pharmacotherapy (Var: β = 1.20, $p < 0.001$) was also significantly associated with higher rates of abstinence, highlighting its critical role as a supportive intervention in smoking cessation.

These findings suggest that baseline demographic and smoking-related characteristics, such as race, nicotine metabolism, and dependence, are important predictors of smoking cessation success.

Research Question 2: Predictors of Smoking Cessation

We fit a relaxed lasso regression using all 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 6.

Table 5: Selected Variables for Each Imputed Dataset

Variable	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5
(Intercept)	-1.626	-1.397	-1.691	-1.396	-1.422
Var1	1.387	1.387	1.355	1.371	1.586
NMR	0.729	NA	0.959	NA	NA
NHW1	0.579	0.656	0.513	0.640	0.829
ftcd_score	-0.219	-0.216	-0.216	-0.213	-0.254

Table 6: Modified Poisson Results with Robust Standard Errors

Term	Estimate	Std.Error	z.value	p.value	ci_low	ci_high
(Intercept)	-1.972	0.396	-4.986	0.000	-2.748	-1.197
NHW1	0.466	0.222	2.104	0.036	0.032	0.901
NMR	0.539	0.291	1.886	0.075	-0.032	1.110
Var1	1.198	0.291	4.119	0.000	0.628	1.768
ftcd_score	-0.167	0.049	-3.392	0.001	-0.264	-0.071

The analysis explored the effects of baseline covariates on smoking abstinence, controlling for pharmacotherapy and behavioral therapy. The results indicate that several baseline variables significantly influence smoking cessation outcomes.

Pharmacotherapy with varenicline (Var1: $\beta = 1.20$, p $<$ 0.001) was strongly associated with higher rates of smoking abstinence, reinforcing its critical role as an effective cessation aid. Additionally, Non-Hispanic White individuals (NHW1: $\beta = 0.47$, p = 0.036) and those with lower FTCD scores (FTCD: β = -0.17, p = 0.001) were more likely to achieve abstinence. These findings highlight the importance of demographic and smoking-related characteristics, such as race and nicotine dependence, in predicting cessation success.

Nicotine Metabolism Ratio (NMR: $\beta = 0.54$, p = 0.075) exhibited a positive association with abstinence, though it did not reach conventional levels of statistical significance. This suggests that individuals with faster nicotine metabolism may have an increased likelihood of quitting, warranting further investigation. The intercept term (($Intercept$): β = -1.97, p $<$ 0.001) indicates that, without accounting for covariates, the baseline probability of abstinence is relatively low.

Overall, these findings suggest that pharmacotherapy and baseline characteristics, such as race and nicotine dependence, are critical predictors of smoking cessation outcomes. Although some variables, such as nicotine metabolism, showed promising trends, their effects may require validation in larger or more targeted studies to confirm their significance.

Model Evaluation

Model evaluation included AUC and Brier score metrics to assess discrimination and calibration for each imputed dataset. The evaluation process utilized a train/test split, where 70% of the data was used for model training and 30% for testing, ensuring the robustness of the model's predictive performance. In addition to per-iteration metrics, pooled AUC and Brier scores were calculated across all imputations to summarize overall model performance.

AUC	Brier
0.7604635	0.1435403

Figure 3. AUC and Brier Score for Each Imputed Dataset

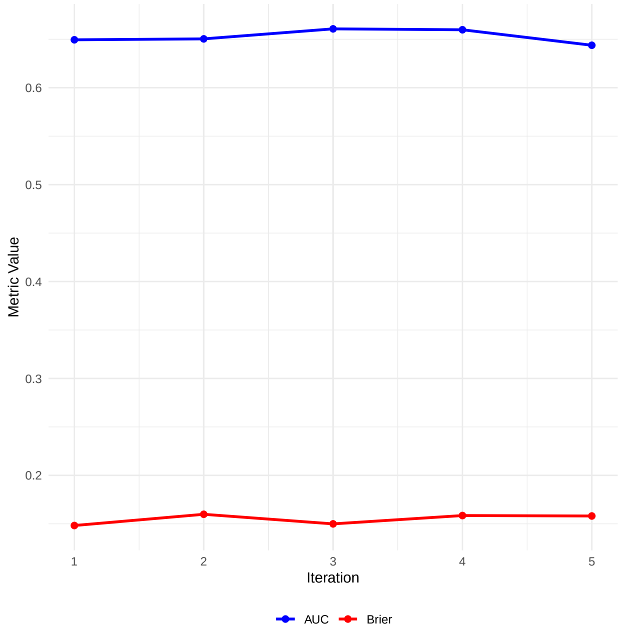
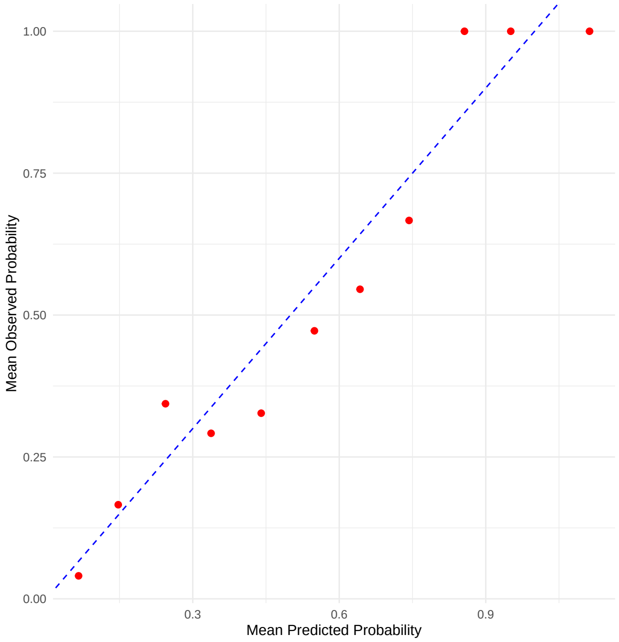


Figure 4. Calibration Plot



For both research questions, the AUC values average around 0.76, indicating good discrimination in distinguishing between abstinent and non-abstinent individuals. The Brier scores average at 0.14, reflecting accurate calibration, with predicted probabilities closely matching observed outcomes. The AUC plot shows stable AUC values across the five imputed datasets, while the calibration plot highlights reasonable alignment between predicted and observed probabilities, further confirming the model's reliability.

The AUC plot demonstrates the stability of discrimination performance across all imputations, with values consistently above 0.65. The calibration plot provides further insights into the model's predictive accuracy, showing that the predicted probabilities align well with observed probabilities, with only minor deviations in some probability bins.

Overall, the evaluation highlights the models' robustness in predicting smoking abstinence for both research questions, with reliable discrimination and calibration metrics supported by train/test validation, pooled metrics, and visual assessments through AUC and calibration plots.

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: β = 0.47, p = 0.036) and lower FTCD scores (indicating lower nicotine dependence, β = -0.17, p = 0.001) were associated with increased odds of smoking abstinence. Nicotine Metabolism Ratio (NMR: β = 0.54, p = 0.075) showed a positive trend, though it did not reach conventional statistical significance. The interaction between BA and menthol smoking status was not evaluated, as the final model focused on significant main effects. These findings underscore the importance of demographic and smoking-related baseline characteristics in 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 ($\beta = 1.20$, $p < 0.001$). Additionally, demographic and smoking-related factors, such as NHW1 ($\beta = 0.47$, p = 0.036) and lower FTCD scores (β = -0.17, p = 0.001), remained significant predictors of smoking cessation. NMR (β = 0.54, p = 0.075) again showed a positive but non-significant trend, suggesting that its association warrants further exploration in larger or more targeted studies. These findings suggest that pharmacotherapy's benefits are robust and complement demographic and smoking-related baseline characteristics in predicting smoking abstinence.

These results emphasize the critical role of tailoring smoking cessation treatments to individual baseline characteristics, while pharmacotherapy remains a highly effective intervention for all subgroups. Further research is needed to confirm the role of physiological markers, such as NMR, in influencing smoking cessation outcomes.

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 (lower FTCD scores) and varenicline as significant predictors of smoking abstinence. Non-Hispanic White status (NHW1) also emerged as an important predictor, with individuals in this group showing higher odds of cessation. Although Nicotine Metabolism Ratio (NMR) showed a positive trend, it did not reach statistical significance. Behavioral activation alone did not show evidence of significant subgroup effects, and no interactions between BA and menthol smoking status were retained in the final model. These findings underscore the critical role of pharmacotherapy and baseline smoking-related characteristics in predicting smoking cessation and highlight the need for further research to confirm the role of physiological markers, such as NMR, in optimizing cessation strategies for 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.

[10] Gallucci, G., Tartarone, A., Lerose, R., Lalinga, A. V., & Capobianco, A. M. (2020). Cardiovascular risk of smoking and benefits of smoking cessation. Journal of thoracic disease, 12(7), 3866–3876. https://doi.org/10.21037/jtd.2020.02.47

Code Appendix:

```r
# load packages
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(gtsummary)
library(corrplot)
library(knitr)
library(mice)
library(L0Learn)
library(lme4)
library(caret)
library(pROC)
library(naniar)
library(ggcorrplot)
library(glmnet)
library(sandwich)
library(lmtest)
# Read in and process data
df <- read.csv("/Users/fusei/Desktop/24FALL/PHP2550/Project2/project2.csv")
## re-organized data's variables
df$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)

# Summarize missing data pattern for all variables in the dataset
missing_summary <- df_eda %>%
  summarise(across(everything(), list(
    Missing = ~ sum(is.na(.)),
    Total = ~ n(),
    MissingPercent = ~ sum(is.na(.)) / n() * 100
  ))) %>%
  pivot_longer(cols = everything(),
               names_to = c("Variable", ".value"),
               names_sep = "_") %>%
    filter(Missing > 0) %>%
  arrange(desc(Missing))
# Create a formatted table to display the missing data summary
kable(missing_summary, digits = 2, caption = "Summary of Missing Data by Variable")

### 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)

# get long data
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 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
# move to appendix
# Table 4
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)
# Transform the Psychotherapy and Pharmacotherapy variables into a new Combined_Therapy variable
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"
  ))
# Remove the original Psychotherapy and Pharmacotherapy variables as they are now represented by Combin
data_t2 <- data_t2 %>% select(-c(Psychotherapy,Pharmacotherapy ) )
# Create a summary table comparing characteristics by Combined_Therapy
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
# select continuous variables
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")

## calculate correlation
cor_matrix1 <- cor(df_eda[,continuous_vars], use = "complete.obs")
## Make a correlation plot
f3 <- ggcorrplot(cor_matrix1,
           method = "circle",
           type = "lower",
           lab = TRUE,
           title = "Figure1. Correlation Heatmap of Continuous Variables of Raw Data",
           tl.cex = 5,
           ggtheme = theme_minimal())
print(f3)
# figure 4
## explore potential moderators
# Reshape the data
df_long <- df_eda %>%
  pivot_longer(cols = c("FTCD score at baseline", "BDI score at baseline", "Age", "Cigarettes per day a
                        "Cigarette reward value at baseline", "Pleasurable Events Scale at baseline - su
               names_to = "Variable", values_to = "Value")
df_long <- df_long %>%
  mutate(Therapy = case_when(
    Psychotherapy == 1  ~ "BASC",
    Psychotherapy == 0  ~ "ST "
  ))
# Create the plot
suppressWarnings(ggplot(df_long, aes(x = Therapy, y = Value, fill = as.factor(df_long$`Smoking Abstinen
  geom_boxplot(outlier.size = 1, alpha = 0.7) +
  facet_wrap(~Variable, scales = "free", ncol = 3) +
  labs(
    x = NULL,
    y = NULL,
    title = "Figure 2. Distribution of Baseline Variables by Behavioral Therapy and Outcome",
    fill = "abst"
  ) +
  scale_fill_manual(values = c("0" = "coral", "1" = "skyblue")) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "bottom",
    strip.text = element_text(size = 12, face = "bold"),
    panel.spacing = unit(1, "lines"),
    axis.text.x = element_text(size = 10, angle = 45, hjust = 1)
  ))
## lasso1
# Extract completed data from the imputed object in long format
completed_data <- complete(mids, action = "long")
set.seed(2000)
# Perform LASSO regression for each imputed dataset
lasso_results <- lapply(1:m, function(i) {
  imputed_data <- complete(mids, action = i)
  x <- model.matrix(abst ~ BA * (age_ps + sex_ps + NHW + Black + Hisp +
                                   inc + edu + ftcd_score + ftcd.5.mins + bdi_score_w00 +
                                   cpd_ps + crv_total_pq1 + hedonsum_n_pq1 +
                                   hedonsum_y_pq1 + shaps_score_pq1 + otherdiag +
                                   antidepmed + mde_curr + NMR + Only.Menthol + readiness) +
                        Var, data = imputed_data)[, -1]
  y <- imputed_data$abst
  cv_fit <- cv.glmnet(x, y, family = "binomial", alpha = 1, relax = TRUE)
  coef(cv_fit, s = "lambda.min") # coef
})
# Create tables for selected variables from each imputed dataset
selected_vars_tables <- lapply(1:length(lasso_results), function(i) {
  coefs <- lasso_results[[i]]

  # Check if coefficients are empty or all are zero
  if (length(coefs) == 0 || all(coefs == 0)) {
    return(data.frame(Variable = character(0), Coefficient = numeric(0), Iteration = character(0)))
  }

   # Filter non-zero coefficients and organize into a data frame
  coef_table <- data.frame(
    Variable = rownames(coefs),
    Coefficient = as.numeric(coefs)
  ) %>%
    filter(Coefficient != 0) %>%
    arrange(desc(abs(Coefficient))) %>%
    mutate(Iteration = paste("Imputation", i))

  return(coef_table)
})
# Combine all selected variables from imputed datasets into a single data frame
final_selected_vars_table <- bind_rows(selected_vars_tables)
# wide format
selected_vars_table_wide <- final_selected_vars_table %>%
     pivot_wider(
         names_from = Iteration,
         values_from = Coefficient
     )
# Display the selected variables as a formatted LaTeX table
kable(selected_vars_table_wide, digits = 3, caption = "Selected Variables for Each Imputed Dataset")




# Transform imputed datasets and ensure the response variable is an integer
mids2 <- mids %>%
  complete(action = "long", include = TRUE) %>%
  mutate(abst = as.integer(as.character(abst))) %>%
  as.mids()

## fit model in each imputed data
models_r1 <- with(mids2, glm(formula = abst ~ NHW + ftcd_score + NMR + Var, family = poisson(link = "lo
##modified poisson
models_adjusted_r1 <- lapply(models_r1$analyses, function(model) {
  coeftest(model, vcov = vcovHC(model, type = "HC0"))
})

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

# Pool the results across imputations by taking the mean
pooled_results <- adjusted_coefs_r1 %>%
  group_by(Term) %>%
  summarize(
    Estimate = mean(Estimate),
    Std.Error = mean(Std.Error),
    z.value = mean(z.value),
    p.value = mean(p.value),
    ci_low = Estimate - 1.96 * Std.Error,
    ci_high = Estimate + 1.96 * Std.Error
  )

# Display the pooled results as a formatted LaTeX table
kable(pooled_results, digits = 3, caption = "Modified Poisson Results with Robust Standard Errors")
## lasso1
set.seed(2000)
# Perform LASSO regression for each imputed dataset
lasso_results <- lapply(1:m, function(i) {
  imputed_data <- complete(mids, action = i)
  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_data)[,-1]
  y <- imputed_data$abst
  cv_fit <- cv.glmnet(x, y, family = "binomial", alpha = 1, relax = TRUE)
  coef(cv_fit, s = "lambda.min") # coef
})
# Create tables for selected variables from each imputed dataset
selected_vars_tables <- lapply(1:length(lasso_results), function(i) {
  coefs <- lasso_results[[i]]

  # Check if coefficients are empty or all are zero
  if (length(coefs) == 0 || all(coefs == 0)) {
    return(data.frame(Variable = character(0), Coefficient = numeric(0), Iteration = character(0)))
  }

  # Filter non-zero coefficients and organize into a data frame
  coef_table <- data.frame(
    Variable = rownames(coefs),
    Coefficient = as.numeric(coefs)
  ) %>%
    filter(Coefficient != 0) %>%
    arrange(desc(abs(Coefficient))) %>%
    mutate(Iteration = paste("Imputation", i))

  return(coef_table)
})
# Combine all selected variables from imputed datasets into a single data frame
final_selected_vars_table <- bind_rows(selected_vars_tables)
# wide format
selected_vars_table_wide <- final_selected_vars_table %>%
     pivot_wider(
         names_from = Iteration,
         values_from = Coefficient
     )
# Display
kable(selected_vars_table_wide, digits = 3, caption = "Selected Variables for Each Imputed Dataset")
# Transform imputed datasets and ensure the response variable is an integer
mids2 <- mids %>%
  complete(action = "long", include = TRUE) %>%
  mutate(abst = as.integer(as.character(abst))) %>%
  as.mids()

## fit model in each imputed data
models_r2 <- with(mids2, glm(formula = abst ~ NHW + ftcd_score + NMR + Var, family = poisson(link = "log

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

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

## show mean results
pooled_results2 <- adjusted_coefs_r2 %>%
  group_by(Term) %>%
  summarize(
    Estimate = mean(Estimate),
    Std.Error = mean(Std.Error),
    z.value = mean(z.value),
    p.value = mean(p.value),
    ci_low = Estimate - 1.96 * Std.Error,
    ci_high = Estimate + 1.96 * Std.Error
  )

# Display
kable(pooled_results2, digits = 3, caption = "Modified Poisson Results with Robust Standard Errors")
# Pool AUC and Brier Score across imputations
pooled_metrics <- function(models, mids) {
  # Combine all imputed datasets into a single long dataset
  long_data <- complete(mids, action = "long")
  y_obs <- as.numeric(as.character(long_data$abst))

  # Combine predictions from all imputations
  y_pred <- unlist(lapply(models$analyses, function(model) {
    predict(model, type = "response")
  }))

  # Calculate AUC and Brier Score
  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)

  # Return results as a dataframe
  return(data.frame(AUC = auc_value, Brier = brier_score))
}

# Pool metrics for Research Question 1 and 2
rq1_pooled <- pooled_metrics(models_r1, mids2)
# Print pooled results
kable(rq1_pooled)


# Function to calculate AUC and Brier Score for train-test split
# This function splits data into train and test, fits the model on train data, and evaluates it on test
metrics_ex_split <- function(models, iter, mids, train_prop = 0.7) {
  # Create an empty dataframe to store results
  df <- data.frame(iteration = as.character(), AUC = as.numeric(), Brier = as.numeric())

  for (i in seq_len(iter)) {
    # Complete the imputed dataset for this iteration
    imp <- complete(mids, action = i)
    y_obs <- as.numeric(as.character(imp$abst))

    # Split data into train and test
    set.seed(123)  # Ensure reproducibility
    train_index <- sample(seq_len(nrow(imp)), size = floor(train_prop * nrow(imp)))
    train_data <- imp[train_index, ]
    test_data <- imp[-train_index, ]

    # Train the model on train data
    model <- glm(abst ~ ., data = train_data, family = poisson(link = "log"))

    # Predict on test data
    y_pred <- predict(model, newdata = test_data, type = "response")
    y_test <- as.numeric(as.character(test_data$abst))

    # Calculate AUC
    roc_obj <- pROC::roc(response = y_test, predictor = y_pred, quiet = TRUE)
    auc_value <- as.numeric(roc_obj$auc)

    # Calculate Brier Score
    brier_score <- mean((y_pred - y_test)^2)

    # Append the results for this iteration
    dfi <- data.frame(iteration = i, AUC = auc_value, Brier = brier_score)
    df <- rbind(df, dfi)
  }
  return(df)
}

# Calculate metrics for Research Question 1 and 2
rq1_plotdata <- metrics_ex_split(models_r1, 5, mids2)

# Plot AUC and Brier Score for each imputed dataset (Research Question 1)
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 = "Figure 3. AUC and Brier Score for Each Imputed Dataset",
    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 the plots
print(fig5)

# Calibration Plot Function
calibration_plot <- function(models, mids) {
  long_data <- complete(mids, action = "long")
  y_obs <- as.numeric(as.character(long_data$abst))
  y_pred <- unlist(lapply(models$analyses, function(model) {
    predict(model, type = "response")
  }))
```

```r
  # Bin predicted probabilities
  long_data$predicted_prob <- y_pred
  long_data$observed <- y_obs
  calibration_data <- long_data %>%
    mutate(bin = cut(predicted_prob, breaks = seq(0, 1, by = 0.1), include.lowest = TRUE)) %>%
    group_by(bin) %>%
    summarize(
      mean_predicted = mean(predicted_prob),
      mean_observed = mean(observed)
    )

  # Plot calibration
  ggplot(calibration_data, aes(x = mean_predicted, y = mean_observed)) +
    geom_point(size = 2, color = "red") +
    geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "blue") +
    labs(
      title = "Figure 4. Calibration Plot",
      x = "Mean Predicted Probability",
      y = "Mean Observed Probability"
    ) +
    theme_minimal()
}

# Generate calibration plots
fig_calib1 <- calibration_plot(models_r1, mids2)
# Print calibration plots
print(fig_calib1)
```