# Baseline Characteristics as Potential Moderators and Predictors of Smoking Cessation in Adults with Major Depressive Disorder

PHP2550 Project 2: A regression analysis

Yanwei (Iris) Tong

2024-10-10

**Abstract**

**Purpose:** Building on a previous randomized, placebo-controlled study exploring factors influencing smoking cessation among adults with major depressive disorder (MDD), this project reexamined data from the same trial to accomplish the following two main objectives: 1) to identify baseline variables that may moderate the impact of behavioral treatments on end-of-treatment (EOT) smoking abstinence, and 2) to evaluate baseline variables as predictors of abstinence outcomes, accounting for the effects of both behavioral treatments and pharmacotherapy.

**Methods**: We applied logistic regression to model the binary smoking abstinence outcome and implemented three selection methods to identify moderating and predictive variables: bidirectional stepwise selection using AIC, elastic net regression with cross-validation, and best subset selection with Mallow's Cp criterion. Interaction terms were carefully incorporated, especially for factors hypothesized to moderate behavioral activation effects on abstinence. *Objective 1* included a comprehensive range of baseline variables and interaction terms with BA, while *Objective 2* focused solely on baseline predictors without interactions to assess their independent predictive effects. Calibration and discrimination metrics were used to evaluate model performance.

**Results and conclusion**: Our analysis identified several key baseline characteristics as potential moderators and predictors for smoking cessation among individuals with MDD. Interactions between BA and FTCD score, Nicotine Metabolism Ratio (NMR), readiness to quit, MDD status, and anhedonia suggested nuanced influences on cessation, potentially acting as moderators of treatment effects. For predictor effects, variables such as education level, race, FTCD score, and NMR had certain impacts on odds of abstinence. In terms of model performance, Elastic Net achieved the best calibration metrics, with the lowest Brier score and calibration error, whereas bidirectional stepwise and best subset exhibited slightly higher AUC values, reflecting marginally better discrimination.

# INTRODUCTION

This regression analysis project, in collaboration with Dr. George Papandonatos from Brown's Department of Biostatistics, seeked to explore factors influencing smoking cessation among adults with major depressive disorder (MDD). Individuals with MDD often demonstrate stronger nicotine dependence and experience more challenging withdrawal symptoms than those without MDD. While varenicline is a proven aid for smoking cessation, addressing psychological factors associated with MDD-related smoking behaviors might also improve quit rates in this population.

Dr. Papandonatos' previous randomized, placebo-controlled study (Hitsman et al. (2023)), which included 300 adult smokers with either current or past MDD, employed a 2x2 factorial design and compared behavioral activation for smoking cessation (BASC) against standard treatment (ST) and varenicline versus placebo. The multi-center study found no significant differences in abstinence outcomes between BASC and ST, regardless of varenicline use. However, varenicline significantly outperformed placebo at the 27-week follow-up, achieving a cessation of 16.2% compared to 7.5% for the placebo group.

Table 1: Participant characteristics by overall sample and treatment arm

| **Characteristic** | **Overall**<br>N = 300 | **BASC+placebo**<br>N = 68 | **BASC+varenicline**<br>N = 83 | **ST+placebo**<br>N = 68 | **ST+varenicline**<br>N = 81 | **p-value** |
|---|---|---|---|---|---|---|
| Demographics | | | | | | |
| Age (years) | 50.0 (12.6) | 50.7 (13.5) | 50.3 (13.2) | 50.3 (10.8) | 48.7 (12.7) | 0.7 |
| Sex (female) | 165 (55%) | 38 (56%) | 44 (53%) | 39 (57%) | 44 (54%) | >0.9 |
| Race | | | | | | |
| Non-Hispanic white | 105 (35%) | 24 (35%) | 34 (41%) | 22 (32%) | 25 (31%) | |
| Black | 157 (52%) | 37 (54%) | 37 (45%) | 40 (59%) | 43 (53%) | |
| Hispanic | 16 (5.3%) | 4 (5.9%) | 3 (3.6%) | 4 (5.9%) | 5 (6.2%) | |
| Other | 22 (7.3%) | 3 (4.4%) | 9 (11%) | 2 (2.9%) | 8 (9.9%) | |
| Income | | | | | | 0.8 |
| Less than $20,000 | 110 (37%) | 25 (37%) | 30 (37%) | 26 (38%) | 29 (36%) | |
| $20,000–35,000 | 68 (23%) | 16 (24%) | 17 (21%) | 14 (21%) | 21 (26%) | |
| $35,001–50,000 | 46 (15%) | 8 (12%) | 13 (16%) | 14 (21%) | 11 (14%) | |
| $50,001–75,000 | 38 (13%) | 12 (18%) | 12 (15%) | 8 (12%) | 6 (7.5%) | |
| More than $75,000 | 35 (12%) | 6 (9.0%) | 10 (12%) | 6 (8.8%) | 13 (16%) | |
| Education | | | | | | |
| Grade school | 1 (0.3%) | 1 (1.5%) | 0 (0%) | 0 (0%) | 0 (0%) | |
| Some high school | 16 (5.3%) | 3 (4.4%) | 7 (8.4%) | 2 (2.9%) | 4 (4.9%) | |
| High school graduate or GED | 76 (25%) | 23 (34%) | 15 (18%) | 11 (16%) | 27 (33%) | |
| Some college/technical school | 116 (39%) | 22 (32%) | 32 (39%) | 38 (56%) | 24 (30%) | |
| College graduate | 91 (30%) | 19 (28%) | 29 (35%) | 17 (25%) | 26 (32%) | |
| **Smoking** | | | | | | |
| Cigarettes per day | 15.1 (7.9) | 15.6 (9.1) | 15.5 (8.5) | 15.0 (7.2) | 14.4 (6.6) | >0.9 |
| FTCD score | 5.2 (2.1) | 5.3 (2.0) | 5.1 (2.3) | 5.4 (2.1) | 5.2 (2.1) | 0.7 |
| Smoking with 5 mins of waking up (Yes) | 138 (46%) | 32 (47%) | 33 (40%) | 35 (51%) | 38 (47%) | 0.5 |
| BDI score | 18.7 (11.5) | 19.0 (12.3) | 18.0 (10.6) | 18.5 (10.8) | 19.5 (12.2) | >0.9 |
| Cigarette reward value | 7.2 (3.7) | 7.4 (3.8) | 7.2 (3.9) | 7.0 (3.7) | 7.1 (3.5) | >0.9 |
| Pleasurable Events Scale (substitute reinforcers) | 22.6 (19.6) | 23.2 (20.3) | 22.9 (19.0) | 20.8 (20.1) | 23.4 (19.5) | 0.6 |
| Pleasurable Events Scale (complementary reinforcers) | 25.4 (19.4) | 27.7 (21.5) | 22.4 (17.0) | 27.4 (19.9) | 25.0 (19.4) | 0.3 |
| Nicotine Metabolism Ratio | 0.4 (0.2) | 0.3 (0.2) | 0.4 (0.2) | 0.4 (0.3) | 0.4 (0.2) | >0.9 |
| Exclusive mentholated cigarette user (Yes) | 178 (60%) | 40 (59%) | 48 (59%) | 43 (64%) | 47 (58%) | 0.9 |
| Readiness to quit smoking | 6.8 (1.2) | 6.8 (1.4) | 6.7 (1.2) | 7.0 (1.3) | 6.7 (1.1) | 0.6 |
| **Psychiatric** | | | | | | |
| Anhedonia | 2.2 (3.2) | 2.2 (3.2) | 2.3 (3.1) | 2.5 (3.4) | 2.1 (3.0) | 0.8 |
| Other lifetime DSM-5 diagnosis (Yes) | 133 (44%) | 35 (51%) | 30 (36%) | 28 (41%) | 40 (49%) | 0.2 |
| Antidepressant medication (Yes) | 82 (27%) | 28 (41%) | 24 (29%) | 15 (22%) | 15 (19%) | 0.013 |
| Current (and past) MDD vs past MDD only (Yes) | 147 (49%) | 32 (47%) | 40 (48%) | 31 (46%) | 44 (54%) | 0.7 |

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

[2] Kruskal-Wallis rank sum test; Pearson's Chi-squared test

Building on the original study, this project reexamined data from the same trial to accomplish the following two main objectives: 1) to identify baseline variables that may moderate the impact of behavioral treatments on end-of-treatment (EOT) smoking abstinence, and 2) to evaluate baseline variables as predictors of smaking cessation, accounting for the effects of both behavioral treatments and pharmacotherapy.

# DATA

## Data Overview

The study population of the RCT consists of 300 adult smokers with a history of current or past MDD. 1) Sociodemographic, 2) smoking-related, and 3) psychiatric characteristics of the participants were collected at baseline and displayed in **Table 1** above. Demographics are age, sex, race, income, education; the smoking behaviors/measurements include key statistics like Fagerstrom Test for Cigarette Dependence (FTCD) score, Nicotine Metabolism Ratio (NMR), and indicator for exclusive Mentholated cigarette user; and psychiatric disgnotic and treatment history . The participants were randomized into four treatment arms: BASC with placebo (`BASC+placebo`), BASC with varenicline (`BASC+Varenicline`), ST with placebo (`ST+placebo`), and ST with varenicline (`ST+Varenicline`).

Overall, the randomization process appears successful, as key variables such as demographic characteristics, smoking intensity (cigarettes per day), and psychiatric measures like the DSM-5 diagnosis and anhedonia scores showed similar distributions across the four treatment arms with $p$-value much greater than 5%, suggesting that participant characteristics were well-balanced across treatment arms, as expected in an RCT. This balance across groups reinforces the original study's internal validity, as any differences in outcomes can be more confidently attributed to the interventions rather than baseline differences in participant characteristics.

Regarding the variable education, as shown in **Table 1**, two of the lowest education levels— grade school and some high school— had very few participant counts (1 and 16, respectively). To ensure adequate sample size and meaningful comparisons across education categories, we merged the first three levels (grade school, some high school, and high school graduate or GED) into a single category, considered as "High School and Below." This aggregation would improve the interpretability of the data by creating a more substantial subgroup and reducing variability, allowing for more reliable statistical analyses.

To explore the potential interaction effects between variables in the dataset, race versus exclusive menthol cigarette use would be a good example. As shown in the contingency table (**Table 2**), the distribution across racial groups is not balanced, as indicated by the significant Chi-square test statistic, suggesting that certain racial groups (Black in this case) might have a stronger inclination towards exclusive menthol use. This imbalance in distribution underscores the importance of considering racial demographics in our analysis, as they may interact with other smoking-related behaviors or biological factors.

Table 2: Contingency Table of Race vs. Only Menthol Use with Chi-Square Test Result

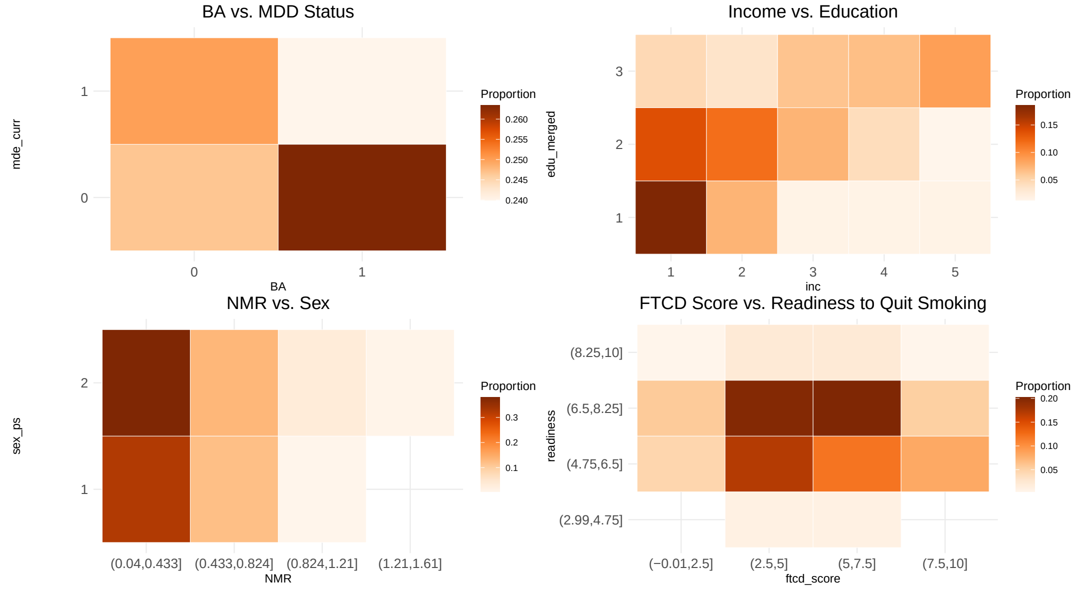
| | Non-Menthol-Only | Menthol-Only |
|---|---:|---:|
| Non-Hispanic white | 72 | 32 |
| Black | 27 | 129 |
| Hispanic | 12 | 4 |
| Other | 9 | 13 |

*Note:*

Chi-Square Statistic 78.49 p-value 0.0000

Similarly, the exploratory interaction heatmaps showcase four selected examples of interactions that would be explored further in our analysis. These examples illustrate potential interactions that may make biological or statistical sense in the context of smoking cessation. For instance, the interaction between behavioral activation and MDD diagnosis could be central to this study's focus on using MDD-targeted treatments to aid smoking cessation. The interaction between income and education could reflect socioeconomic influences on smoking behaviors, while cigarette dependence (measured by FTCD score) and readiness to quit smoking might reveal motivational factors in cessation attempts. Additionally, nicotine metabolism ratio (NMR) versus sex could highlight a possible biological interaction that may affect the body's response to nicotine between genders. The color gradients in **Figure 1** indicating proportion appear to confirm that there might be potential synergies between these variables. This visual evidence supported the inclusion of these interactions for further analysis, as they may capture meaningful relationships that influence smoking cessation. In the methods section, we would discuss a more comprehensive rationale for including a certain set of interaction terms, categorizing some as moderators specifically for BA effects and others as covariates, to capture these multidimensional influences on the smoking abstinence outcome.

Figure 1: Exploratory Interaction Heatmaps–– Moderator and Covariate Examples



## Data Missingness and Imputation

Table 3: Summary of Missing Values

| Variable | Number | Percent |
|---|---|---|
| Participants with any missingness | 59 | 19.67% |
| Nicotine Metabolism Ratio | 21 | 7.00% |
| Cigarette reward value at baseline | 18 | 6.00% |
| Baseline readiness to quit smoking | 17 | 5.67% |
| Income | 3 | 1.00% |
| Anhedonia | 3 | 1.00% |
| Exclusive Mentholated Cigarette User | 2 | 0.67% |
| FTCD score at baseline | 1 | 0.33% |

The overall missingness in the dataset is approximately 20%, which exceeds the commonly accepted 5% threshold for data completeness. This level of missing data could compromise the validity of the analysis if left unaddressed, as it may lead to biased results or loss of valuable information. To preserve the dataset's integrity and maintain statistical power, we opted to impute the missing values with the Multiple Imputation by Chained Equations (MICE) method. It allows for flexible handling of various data types and patterns of missingness, thereby maximizing the use of available information and improving the robustness of the analyses. Specifically, we set the number of imputations to $m = 5$, used predictive mean matching, and iterated the imputation process for a maximum of 50 iterations to ensure reliable imputations.

# METHODS

## Logistic Regression

Given that our smoking cessation outcome variable `abst` is dichotomous, logistic regression is an appropriate modeling choice for this project. Logistic regression models the log-odds of the outcome as a linear combination of predictor variables, as expressed by the equation

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_{ij} x_i x_j + \beta_{ik} x_i x_k + \ldots$$

where the $x_i x_j$ terms denote potential interaction terms (explained in the next subsection).

This transformation ensures that predicted probabilities remain within the range of 0 to 1. To facilitate interpretation, we exponentiate the coefficients ($exp(\beta)$) to express them as odds ratios, which indicate the multiplicative change in the odds of the outcome for each one-unit increase in a predictor, holding other variables constant. When testing the selected model on the test dataset, the `predict()` function in R with `type = "response"` outputs probabilities directly, using the equation

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_{ij} x_i x_j + \beta_{ik} x_i x_k + \ldots)}}$$

This probability-based output provides an intuitive interpretation of the model's predictions in terms of the likelihood of the event occurring given the predictors.

## Variable Inclusion Criteria for Full Models

For *Objective 1*, psychotherapy (`BA`) is the primary predictor of interest, while pharmacotherapy (`Var`) is treated as a control variable. Thus, we ensured to include both of the treatments in all models as consistent factors. The major goal of *Objective 1* is to explore how various baseline characteristics might moderate the effect of behavioral treatment on smoking cessation. To achieve this, we included a range of interaction terms between `BA` and key sociodemographic, smoking-related, and psychiatric variables as potential moderators. These interaction terms allowed us to examine how different participant characteristics may influence the effectiveness of behavoriol activation in promoting abstinence. However, we excluded certain terms, such as interactions between `BA` and variables like indicator of smoking with 5 mins of waking up (`ftcd.5mins`) or cigarettes per day (`cpd_ps`), because these measures are components of the broader FTCD score (`ftcd_score`) and thus would provide redundant information. By prioritizing unique, non-overlapping terms, we aimed to create a comprehensive yet parsimonious model.

In addition to `BA` interaction terms as potential moderators, we included other interaction terms involving `Var` and various baseline characteristics as covariates. We believed these covariate interaction terms shall account for known associations that could impact treatment outcomes. For instance, as previously discussed, there was a recognized relationship between race and menthol-only cigarette use, which could affect smoking cessation success. Similarly, the synergy between factors like baseline readiness to quit smoking (`readiness`), Nicotine Metabolism Ratio (`NMR`), and MDD history (`mde_curr`) could have critical impacts on smoking behavior and mental health status and have plausible biological or statistical interactions with other sociodemographic and smoking-related variables. These interactions reflect meaningful covariate effects that contribute to a more nuanced understanding of how different factors influence treatment outcomes, enabling a more robust analysis of predictors and moderators in the context of smoking cessation for adults with MDD.

For *Objective 2*, the focus shifts to using baseline variables as predictors of smoking cessation outcomes, rather than as covariates or moderators. In this context, `BA` and `Var` are included as control variables across all models to account for the effects of behavioral and pharmacotherapy interventions. However, our goal here is to examine the predictive power of baseline characteristics independently, so we did not include any interaction terms to simplify the model by isolating the main effects of each baseline variable.

## Variable Selection Methods

The entire dataset (N=300) was split into a training set (70%) and a test set (30%) to facilitate model evaluation and to ensure that the model's performance metrics would generalize beyond the training data. This approach allowed us to assess the predictive accuracy and robustness of the selected models on an independent dataset, which is critical for reducing overfitting and improving the reliability of our findings.

### 1) Bidirectional Stepwise Selection

Bidirectional stepwise selection is a model selection method that combines both forward selection and backward elimination to identify the most significant predictors for inclusion in the regression model. By using both directions, this approach benefits from the advantages of both forward and backward selection methods, allowing for the addition of meaningful variables and the removal of redundant ones iteratively. This ensures that the final model includes those predictors that provide substantial contributions to explaining the variability in the outcome. In our analysis, we used the Akaike Information Criterion (AIC) as the selection criterion as AIC balances model fit and complexity by penalizing the number of parameters and so helping to prevent overfitting.

### 2) Elastic Net Regression

Elastic Net regression is a regularization and variable selection technique that incorporates both Lasso (L1 penalty) and Ridge (L2 penalty) regression methods. It addresses some limitations of Lasso, particularly in situations where predictors are highly correlated. By combining L1 and L2 penalties, Elastic Net encourages a grouping effect where correlated variables tend to be selected or excluded together, leading to more stable and reliable models.

In our modeling, Elastic Net was utilized to take advantage of both variable shrinkage and selection properties of Lasso and the grouping effect of Ridge regression. This method helps handle multi-collinearity among predictors while performing variable selection. We used cross-validation to determine the optimal value of the regularization parameter lambda, specifically selecting the lambda that minimizes the cross-validation error ($\lambda_{min}$). This approach ensures that the model achieves the best predictive performance on unseen data by effectively balancing bias and variance.

### 3) Best Subset Selection

Best subset selection is a comprehensive model selection method that involves evaluating all possible combinations of predictor variables to identify the model that best fits the data according to a chosen criterion. For *Objective 1*, due to the complexity of the full model— which included a high number of potential interaction covariate terms— we employed the sequential replacement method, which is a more computationally efficient alternative to exhaustive search for complex models by iteratively replacing variables to find a subset that provides a good balance between model fit and complexity. For *Objective 2*, the model was less complex as it included only baseline predictors without interaction terms. This allowed us to use an exhaustive search strategy, evaluating all possible subsets of the predictors to find the optimal model.

We selected models based on Mallow's $Cp$ criterion. Mallow's $Cp$ provides a measure of the model's predictive error relative to the number of predictors used, helping to select models that are both accurate and parsimonious. By minimizing Mallow's $Cp$, we ensured that the selected model offers the best trade-off between goodness of fit and model simplicity.

## Performance Metrics (Calibration and Discrimination)

To evaluate model performances, we used a combination of calibration and discrimination measures. Calibration plots with error bars and LOESS smoothing allowed us to visually assess the agreement between predicted probabilities and observed outcomes, indicating how well-calibrated the model is across different probability levels. The addition of error bars provides insights into the variability of predictions, while LOESS smoothing offers a flexible fit to better capture trends in calibration. The ROC curve evaluates the model's discrimination ability, reflecting its capability to distinguish between positive and negative outcomes. Furthermore, quantitative metrics such as Brier score and calibration error were included in tables (**Table 5** and **Table 7**) to provide objective assessments of model accuracy and calibration. Brier score combines both calibration and sharpness of probability estimates, while calibration error specifically measures the deviation between predicted and observed probabilities, providing complementary insights into model performance beyond visual assessments.

## RESULTS

### *Objective 1*: Baseline Characteristics as Potential Moderators

In **Table 4**, we observed the coefficients and corresponding odds ratios selected from three model selection methods, with particular focus on behavioral activation and its interaction terms. `BA` was consistently included across all methods, as was pharmacotherapy (`Var`), given their forced inclusion as primary predictor and control. Among the interaction terms with `BA` (`BA1` = BASC) several variables demonstrated potential moderator effects that might influence BA's effectiveness in smoking cessation among individuals with MDD, although there is noticeable variation in terms selected by each method.

Specifically, BASC's interaction with FTCD score with odds ratio (OR) < 1 may suggest that nicotine dependence level could negatively affect how responsive individuals are to BASC. Similarly, current vs. past MDD status (`mde_curr`) appeared as a logical moderator, as BA was designed to alleviate depressive symptoms, which could make it more effective for individuals experiencing active symptoms. The interaction with readiness to quit smoking (`readiness`) with OR slightly higher than 1 in best subset selection may suggest that motivational factors might influence how beneficial BASC is; those who are more prepared to quit may respond differently to BASC. Anhedonia as a potential moderator also aligns well, given that BASC aims to counteract reduced pleasure in activities, a common symptom in depression. Finally, race (Hispanic) as an interaction term may reflect unique socio-cultural factors that could impact the treatment response to BASC. These interaction terms, chosen by at least one method, underscored relevant moderators that may affect psychotherapy's success in achieving smoking cessation.

Noticeably, the inclusion of Anhedonia (`shaps_score_pq1`) and Race Black (`raceBlack`), as non-interaction variables, across three methods may suggest these two factors could play a meaningful role in the overall relationship between BA and smoking cessation. For example, individuals with higher anhedonia levels may require more intensive or targeted interventions to achieve abstinence, highlighting the importance of tailoring BA to address depressive symptoms. Similarly, the repeated selection of race Black as a covariate could imply that racial background could affect treatment outcomes, potentially due to sociocultural or economic factors.

For the majority of other interaction terms, selection varied by method, with little overlap between the chosen variables, reflecting different approaches and model assumptions inherent in bidirectional stepwise, Elastic Net, and best subset selection. This variation potentially emphasizes the differences in variable selection procedure and criteria across the three methods.

Table 4: Summary of Coefficients and Odds Ratios for Potential Moderator Effects across Model Selection Methods

| | Stepwise | | Elastic Net | | Best Subset | |
|---|---|---|---|---|---|---|
| Variable | Coef | OR | Coef | OR | Coef | OR |
| BA1 | 1.6774 | 5.3515 | 0.0273 | 1.0277 | 0.6426 | 1.9013 |
| BA1:ftcd_score | -0.2895 | 0.7486 | | | -0.2466 | 0.7815 |
| BA1:mde_curr1 | | | | | -0.3818 | 0.6826 |
| BA1:raceHispanic | | | -0.2574 | 0.7731 | -15.5255 | |
| BA1:readiness | | | | | 0.0906 | 1.0948 |
| BA1:shaps_score_pq1 | | | | | 0.1701 | 1.1854 |
| NMR | 1.1552 | 3.1747 | 0.3499 | 1.4189 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| NMR:readiness | | | | | 0.1372 | 1.147 |
| NMR:sex_ps2 | | | 0.2227 | 1.2495 | | |
| Var1 | 2.4176 | 11.2191 | 1.725 | 5.6127 | 1.4661 | 4.3322 |
| edu_merged2 | | | | | -0.7765 | 0.46 |
| edu_merged2:Only.Menthol1 | | | -0.6259 | 0.5348 | | |
| edu_merged3:Only.Menthol1 | | | | | 0.4137 | 1.5124 |
| edu_merged3:inc5 | | | 0.3974 | 1.488 | 0.4697 | 1.5996 |
| ftcd.5.mins1 | 1.2624 | 3.5338 | | | | |
| ftcd_score | -0.3949 | 0.6737 | -0.0318 | 0.9687 | | |
| ftcd_score:Var1 | | | | | 0.0765 | 1.0795 |
| ftcd_score:raceBlack | 0.3402 | 1.4053 | | | | |
| ftcd_score:raceHispanic | -15.9128 | 0 | | | -0.1529 | 0.8582 |
| ftcd_score:raceOther | 0.3184 | 1.375 | | | | |
| ftcd_score:readiness | | | -0.0094 | 0.9906 | | |
| ftcd_score:sex_ps2 | | | | | -0.0204 | 0.9799 |
| inc3:Only.Menthol1 | | | | | 0.1433 | 1.1541 |
| mde_curr1 | | | -0.0419 | 0.959 | | |
| mde_curr1:readiness | | | | | -0.0145 | 0.9856 |
| raceBlack | -3.1798 | 0.0416 | -0.112 | 0.894 | -0.9112 | 0.402 |
| raceHispanic | 63.071 | * | | | | |
| raceOther | -3.9604 | 0.0191 | -0.1574 | 0.8544 | | |
| raceOther:Var1 | | | -0.1866 | 0.8298 | | |
| shaps_score_pq1 | -0.2145 | 0.807 | -0.0446 | 0.9564 | -0.266 | 0.7665 |

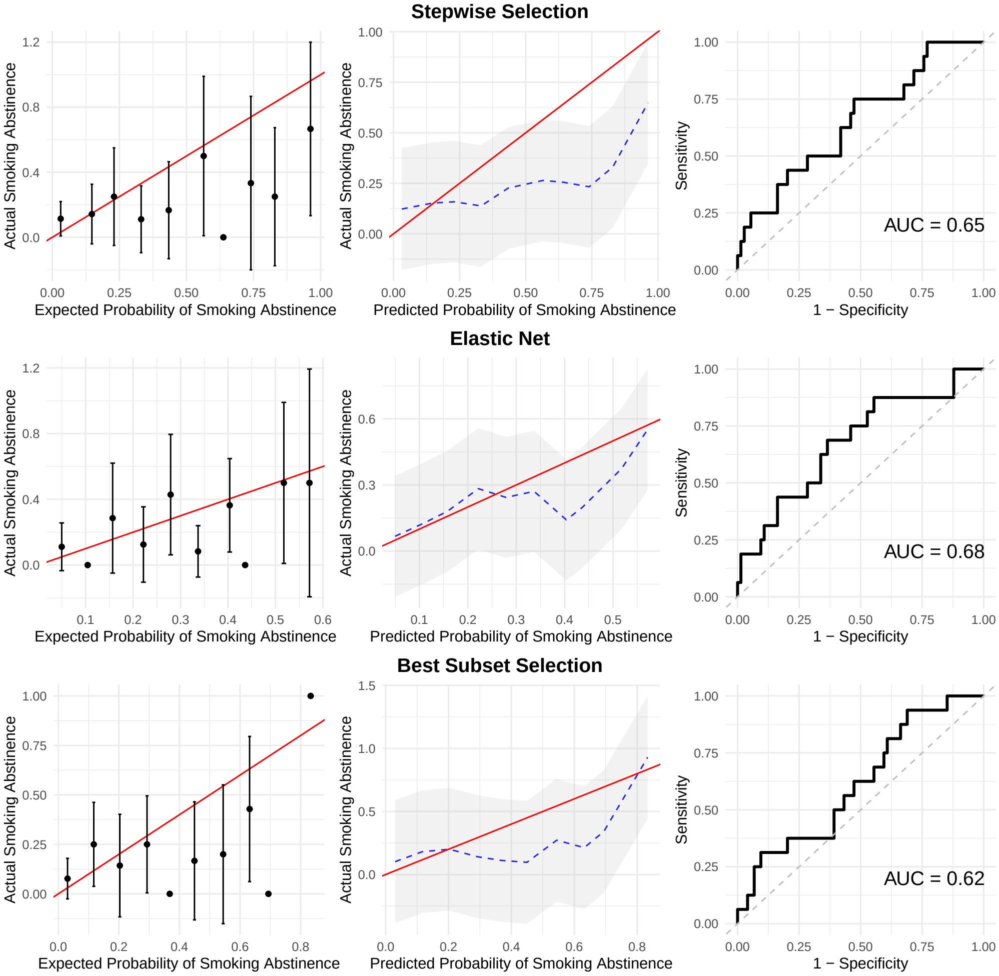
The performance metrics presented in **Table 5** demonstrate the relative strengths and weaknesses of each model selection method for capturing moderator effects on smoking abstinence. Elastic Net exhibited the lowest Brier score (0.1422) and calibration error (0.0677), indicating superior model calibration and a reduced prediction error compared to stepwise and best subset selection. Moreover, Elastic Net achieved a balanced specificity (0.6351) and sensitivity (0.6875), suggesting a more robust discrimination capability across different abstinence probabilities. In contrast, the bidirectional stepwise model demonstrated a higher sensitivity (0.7500) but at the cost of reduced specificity (0.5270), implying a tendency toward over-predicting abstinence outcomes. The best subset model achieved the highest sensitivity (0.9375) but with markedly low specificity (0.3108), indicating that it may overfit to predict abstinence with lower precision across abstinence probabilities. The differences in calibration and discrimination metrics highlight Elastic Net's ability to provide a well-calibrated model with balanced specificity and sensitivity, which may make it a preferable choice for models that aim to generalize well to unseen data.

Table 5: Calibration and Discrimination Metrics for Moderator Effects Modeling

| | Stepwise | Elastic Net | Best Subset |
|---|---|---|---|
| Brier score | 0.1831 | 0.1422 | 0.1696 |
| Calibration error | 0.1678 | 0.0677 | 0.1723 |
| AUC | 0.6503 | 0.6765 | 0.6174 |
| Threshold | 0.1576 | 0.2662 | 0.0737 |
| Specificity | 0.5270 | 0.6351 | 0.3108 |
| Sensitivity | 0.7500 | 0.6875 | 0.9375 |

The calibration and ROC curves in **Figure 2** further illustrate these performance characteristics. The calibration plots for Elastic Net closely align with the ideal diagonal line, suggesting that its predicted abstinence probabilities align well with observed outcomes across different probability levels. Additionally, the LOESS curve for Elastic Net generally tracks the ideal line with only minor deviations, reinforcing its effective calibration. In comparison, bidirectional stepwise and best subset exhibit some misalignment in their calibration plots, with deviations that suggest over- or under-confidence in specific probability ranges. The ROC curves also support Elastic Net's performance, as it demonstrates a moderate AUC of 0.6765, slightly surpassing stepwise and best subset, which achieved AUCs of 0.6503 and 0.6174, respectively. Elastic Net's ROC curve reflects a stronger balance between true positive and false positive rates, contributing to its more consistent predictive performance in modeling moderator effects.

Figure 2: Calibration Plots with Error Bars and LOESS and ROC Curves (Moderator Effects)



Stepwise Selection

Elastic Net

Best Subset Selection

## *Objective 2*: Baseline Characteristics as Potential Predictors

In examining the coefficients and odds ratios for baseline characteristics as predictors, the Nicotine Metabolism Ratio (`NMR`) and FTCD score (`ftcd_score`) stood out as significant predictors across variable selection methods. `NMR` shows relatively high OR $> 1$ in all models indicating its potential strong positive association with smoking cessation outcomes. On the opposite, `ftcd_score` consistently shows OR $< 1$ in all three methods, which may be explained by the rationale that the higher the baseline cigarette dependence, the lower the probability to quit smoking successfully at the end of the trial.

Other demographic variables, such as race, showed generally low OR $<< 1$ (i.e., `raceOther` and `raceBlack`), suggesting possible barriers to cessation in certain groups. Additionally, the education indicators, `edu_merged2` and `edu_merged3`, present odds ratios that differ in directionality, with `edu_merged2` consistently showing OR $< 1$, while `edu_merged3` displays OR $> 1$ across methods. This may suggest that different levels of educational attainment may have opposing effect directionalities on the success likelihood of smoking cessation, potentially reflecting the influence of socioeconomic factors associated with varying educational backgrounds. All these findings highlight the complex interplay between socioeconomic, biological, and demographic factors in influencing smoking cessation success.

Noticeably, pharmacotherapy, as one of the controls, consistently demonstrated a stronger influence on abstinence outcomes than behaviorial treatment, which aligns with previous findings in the original study. `Var` exhibits significantly higher odds ratios across models, particularly in the stepwise and best subset methods, underscoring its greater predictive effect on smoking cessation. In contrast, `BA` shows odds ratios close to 1, suggesting its modest (to negative) impact on the outcome.

Table 6: Summary of Coefficients and Odds Ratios for Potential Predictor Effects across Model Selection Methods

| | Stepwise | | Elastic Net | | Best Subset | |
|---|---|---|---|---|---|---|
| Variable | Coef | OR | Coef | OR | Coef | OR |
| BA1 | 0.1703 | 1.1857 | 0.0346 | 1.0352 | 0.1927 | 1.2125 |
| Var1 | 2.1357 | 8.4632 | 1.778 | 5.9183 | 2.1203 | 8.3333 |
| NMR | 1.4329 | 4.1908 | 0.6433 | 1.9027 | 1.5005 | 4.4841 |
| bdi_score_w00 | | | -0.0047 | 0.9953 | | |
| edu_merged2 | -0.7979 | 0.4503 | -0.3298 | 0.7191 | -0.8995 | 0.4068 |
| edu_merged3 | 0.2209 | 1.2472 | 0.1415 | 1.152 | | |
| ftcd.5.mins1 | 1.1928 | 3.2965 | | | 1.1421 | 3.1334 |
| ftcd_score | -0.3487 | 0.7056 | -0.1006 | 0.9043 | -0.339 | 0.7125 |
| inc5 | | | 0.0886 | 1.0926 | | |
| mde_curr1 | | | -0.0476 | 0.9535 | | |
| raceBlack | -1.2861 | 0.2763 | -0.3877 | 0.6786 | -1.1882 | 0.3048 |
| raceHispanic | -1.0741 | 0.3416 | | | | |
| raceOther | -2.253 | 0.1051 | -0.6816 | 0.5058 | -2.0962 | 0.1229 |
| shaps_score_pq1 | -0.218 | 0.8041 | -0.0671 | 0.9351 | -0.2058 | 0.814 |

In terms of model performance, shown in **Table 7**, Elastic Net still exhibited the best calibration metrics overall, with the lowest Brier score (0.1518) and calibration error (0.1118), indicating it achieves the most accurate probability estimates compared to the other two selection methods. However, both stepwise and best subset methods demonstrated slightly better AUC values (0.6630 and 0.6605, respectively), suggesting they may provide slightly stronger discrimination between abstinence outcomes.

Interestingly, the highest sensitivity was identical across all three methods at 0.8750, in contrast to the findings in *Objective 1*, where sensitivity varied more between methods. This convergence in sensitivity may suggest that when the model complexity is reduced, meaning focusing on baseline predictors without interaction terms, the sensitivity of each method aligns more closely, resulting in more consistent classification performance across methods. Overall, these results suggest that Elastic Net offers the most reliable calibration, while bidirectional stepwise and best subset maintain slightly better discrimination, balancing predictive needs depending on whether calibration or discrimination is prioritized.

Table 7: Calibration and Discrimination Metrics for Predictor Effect Modeling

| | Stepwise | Elastic Net | Best Subset |
|---|---|---|---|
| Brier score | 0.1798 | 0.1518 | 0.1813 |
| Calibration error | 0.1642 | 0.1118 | 0.1743 |
| AUC | 0.6630 | 0.6486 | 0.6605 |
| Threshold | 0.1840 | 0.1481 | 0.1600 |
| Specificity | 0.5676 | 0.4595 | 0.5405 |
| Sensitivity | 0.8750 | 0.8750 | 0.8750 |

# CONCLUSION

This study reexamined factors influencing smoking cessation among adults with MDD, focusing on identifying potential moderators of behavioral activation treatment effects on smoking abstinence and evaluating baseline predictors of cessation outcomes. For both analysis objectives, we utilized stepwise selection, Elastic Net, and best subset methods to select potentially meaningful factors. Key variables like the FTCD score and readiness to quit emerged as plausible moderators, suggesting that the effectiveness of BASC may depend on individual differences in nicotine dependence and motivation to quit. Despite method-specific variations, these results underscore the complexity of identifying consistent moderators, though with a small sample size and numerous interaction terms.

In the predictor-focused analysis, several baseline characteristics stood out for their predictive power in smoking cessation outcomes, including nicotine metabolism ratio (NMR), FTCD score, education, and race. Notably, NMR consistently showed a strong association with cessation, potentially indicating a biological influence on treatment success. Additionally, educational levels presented opposing odds ratios, with some levels being positively associated with cessation while others were negatively associated, possibly reflecting varying socio-economic backgrounds. This predictor analysis highlights the nuanced influences of socioeconomic, biological, and demographic factors on smoking cessation among individuals with MDD.

Regarding prediction performance of the three model selection methods employed, elastic net demonstrated the best calibration performance across both objectives, achieving the lowest Brier score and calibration error, while other methods provided slightly better discrimination as evidenced by higher AUCs. These findings suggest that elastic net would be a robust choice when calibration is prioritized, but stepwise and best subset approaches may offer value in contexts where discrimination is more critical.

Overall, the results of our project provided some insights that could guide future study design when developing smoking cessation interventions for adults with MDD. By highlighting specific baseline moderators and predictors of the cessation outcome, this work may support a more targeted approach to intervention planning in considering baseline smoking behaviors and socio-demographics, potentially improving treatment efficacy for this population.

# LIMITATIONS

A primary limitation of this analysis is the small sample size, with only 300 observations split between training and test sets. This limited data size, combined with the inclusion of multiple interaction terms, exacerbates the “small n, big p” problem, where the number of predictors may outstrip the available data points, potentially leading to overfitting and instability in model estimates. Given the small dataset, further research with larger samples and significance testing of individual predictors would enhance the reliability and generalizability of these conclusions.

Additionally, the predictor selection process was based solely on criteria like AIC for birectional stepwise selection, cross-validation for Elastic Net, or Mallow’s $Cp$ for best subset, focusing on optimizing overall model performance metrics, such as calibration and discrimination, rather than testing the individual significance of each variable. As a result, no formal $p$-values or statistical tests were assigned to assess the significance of individual variables. This approach limits our ability to make definitive claims about the significance of individual variables as verified moderators or predictors and instead relies on their contribution to the model’s overall predictive performance.

## Consent, Data, and Code Availability

Primary data were provided by Dr. George Papandonatos from the Department of Biostatistics at Brown University. The original data cannot be shared directly for privacy. Replication scripts are available at https://github.com/YanweiTong-Iris/PHP2550-Fall24/tree/main/Project2.

# Reference

Hitsman, B., Papandonatos, G. D., Gollan, J. K., Huffman, M. D., Niaura, R., Mohr, D. C., Veluz-Wilkins, A. K., Lubitz, S. F., Hole, A., Leone, F. T., and others (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*, Wiley Online Library, 118, 1710–1725.

# Code Appendix

```
# to prevent scientific notation
options(scipen=999)

# Set up knit environment
knitr::opts_chunk$set(echo = FALSE,
                      message = FALSE,
                      warning = FALSE,
                      error = FALSE)

# Load necessary packages
library(tidyverse)
library(kableExtra)
library(knitr)
library(ggplot2)
library(gridExtra)
library(grid)
library(naniar)
library(gtsummary)
library(gt)
library(patchwork)
library(knitcitations)
library(mice)
library(glmnet)
library(pROC)
library(MASS)
library(leaps)
library(RColorBrewer)
library(cowplot)
# Define data path and import data
data_path = "/Users/yanweitong/Documents/PHP2550-Data/Project2"
data = read.csv(paste0(data_path, "/project2.csv"))

# Data preprocessing
data = data  %>%
  # create race variable
  mutate(race = factor(case_when(
    NHW == 1 ~ "Non-Hispanic white",
    Black == 1 ~ "Black",
    Hisp == 1 ~ "Hispanic",
    TRUE ~ "Other"  # Handle cases where none of the above conditions are met
  ), levels = c("Non-Hispanic white", "Black", "Hispanic", "Other"))) %>%
  # create treatment categories
  mutate(treatment_cat = factor(case_when(BA == 1 & Var == 0 ~ "BASC+placebo",
                                   BA == 0 & Var == 0 ~ "ST+placebo",
                                   BA == 1 & Var == 1 ~ "BASC+varenicline",
                                   BA == 0 & Var == 1 ~ "ST+varenicline"))) %>%
  # factorize categorical/ordinal variables
  mutate(
    abst = factor(abst),
    Var = factor(Var),
    BA = factor(BA),
    sex_ps = factor(sex_ps),
    NHW = factor(NHW),
    Black = factor(Black),

    ftcd.5.mins = factor(ftcd.5.mins),
    otherdiag = factor(otherdiag),
    antidepmed = factor(antidepmed),
    mde_curr = factor(mde_curr),
    Only.Menthol = factor(Only.Menthol),
    edu = factor(edu, levels = c(1, 2, 3, 4, 5)),
    inc = factor(inc, levels = c(1, 2, 3, 4, 5))
  ) %>%
  # make integers numeric
  mutate(across(
    .cols = where(is.integer) & !all_of("id"),
    .fns = as.numeric
  ))
# for sub-tab purpose
table1_data = data %>%
  mutate(
    Demographics = NA,
    Smoking = NA,
    Psychiatric = NA
  ) %>%
  mutate(edu = factor(edu, levels = c(1, 2, 3, 4, 5),
                 labels = c("Grade school",
                            "Some high school",
                            "High school graduate or GED",
                            "Some college/technical school",
                            "College graduate")),
    inc = factor(inc, levels = c(1, 2, 3, 4, 5),
                 labels = c("Less than $20,000",
                            "$20,000-35,000",
                            "$35,001-50,000",
                            "$50,001-75,000",
                            "More than $75,000")))

table1_data %>%
  dplyr::select(
    treatment_cat,
    Demographics,
    age_ps,
    sex_ps,
    race,
    inc,
    edu,
    Smoking,
    cpd_ps,
    ftcd_score,
    ftcd.5.mins,
    bdi_score_w00,
    crv_total_pq1,
    hedonsum_n_pq1,
    hedonsum_y_pq1,
    NMR,
    Only.Menthol,
    readiness,
    Psychiatric,
    shaps_score_pq1,
    otherdiag,
    antidepmed,
    mde_curr
  ) %>%
  tbl_summary(
    statistic = list(all_continuous() ~ c("{mean} ({sd})"),
                     all_categorical() ~ "{n} ({p}%)"),
    by = treatment_cat,
    digits = all_continuous() ~ 1,
    missing = "no",
    type = list(
      age_ps ~ "continuous",
      sex_ps ~ "dichotomous",
      race ~ "categorical",
      inc ~ "categorical",
      edu ~ "categorical",
      cpd_ps ~ "continuous",
      ftcd_score ~ "continuous",
      ftcd.5.mins ~ "dichotomous",
      bdi_score_w00 ~ "continuous",
      crv_total_pq1 ~ "continuous",
      hedonsum_n_pq1 ~ "continuous",
      hedonsum_y_pq1 ~ "continuous",
      NMR ~ "continuous",
      Only.Menthol ~ "dichotomous",
      readiness ~ "continuous",
      shaps_score_pq1 ~ "continuous",
      otherdiag ~ "dichotomous",
      antidepmed ~ "dichotomous",
      mde_curr ~ "dichotomous"
    ),
    label = list(
      age_ps = "Age (years)",
      sex_ps = "Sex (female)",
      race = "Race",
      inc = "Income",
      edu = "Education",
      cpd_ps = "Cigarettes per day",
      ftcd_score = "FTCD score",
      ftcd.5.mins = "Smoking with 5 mins of waking up (Yes)",
      bdi_score_w00 = "BDI score",
      crv_total_pq1 = "Cigarette reward value",
      hedonsum_n_pq1 = "Pleasurable Events Scale (substitute reinforcers)",
      hedonsum_y_pq1 = "Pleasurable Events Scale (complementary reinforcers)",
      Only.Menthol = "Exclusive mentholated cigarette user (Yes)",
      readiness = "Readiness to quit smoking",
      NMR = "Nicotine Metabolism Ratio",
      shaps_score_pq1 = "Anhedonia",
      otherdiag = "Other lifetime DSM-5 diagnosis (Yes)",
      antidepmed = "Antidepressant medication (Yes)",
      mde_curr = "Current (and past) MDD vs past MDD only (Yes)"
    ),
    value = list(
      sex_ps ~ "2",
      Only.Menthol ~ "1",
      otherdiag ~ "1",
      antidepmed ~ "1",
      mde_curr ~ "1",
      ftcd.5.mins ~ "1"
    )
  ) %>%
  add_overall() %>%
  add_p() %>%
  modify_header(label ~ "**Characteristic**") %>%
  modify_caption(caption = "Participant characteristics by overall sample and treatment arm") %>%
  # for sub-tab purpose
  modify_table_body(
    ~ .x %>%
      mutate(across(everything(), ~ ifelse(. == "0 (NA%)", "", .)))
  ) %>%
  modify_table_styling(
    rows = label %in% c("Sociodemographics", "Smoking", "Psychiatric"),
    columns = label,
    text_format = "bold"
  )   %>%
  as_kable_extra(
    booktabs = TRUE,
    longtable = TRUE,
    linesep = "",
    format = "latex"
  ) %>%
  kableExtra::kable_styling(
    position = "center",
    latex_options = c("striped", "repeat_header", "hold_position", "scale_down"),
    stripe_color = "gray!15",
    font_size = 8
  )
# Create a new variable that contains the 3 first levels of edu
data <- data %>%
  mutate(edu_merged = factor(case_when(
    edu %in% c("1", "2", "3") ~ "1",
    edu == "4" ~ "2",
    edu == "5" ~ "3"
  )))

# Create and display the contingency table between race and menthol useage
table_race_menthol <- table(data$race, data$Only.Menthol)

# Perform the chi-square test
chi_square_test <- chisq.test(table_race_menthol)
```

```r
chi_square_text <- paste0(
  sprintf("Chi-Square Statistic   %.2f", chi_square_test$statistic),
  sprintf("  p-value   %.4f", chi_square_test$p.value)
)

kable(table_race_menthol,
      caption = "Contingency Table of Race vs. Only Menthol Use with Chi-Square Test Result",
      col.names = c("Non-Menthol-Only", "Menthol-Only"),
      row.names = TRUE) %>%
   footnote(chi_square_text,
            footnote_as_chunk = FALSE) %>%
  kable_styling(full_width = F, position = "center", font_size = 9)

# Define function to bin continuous variables and create proportion heatmaps
create_heatmap <- function(data, var1, var2, title, bin_var1 = FALSE,
                           bin_var2 = FALSE) {
  # Remove NA values for the specified columns
  data <- data %>% drop_na(all_of(c(var1, var2)))

  # Optionally bin continuous variables
  if (bin_var1) data <- data %>%
      mutate(!!sym(var1) := cut(!!sym(var1), breaks = 4))
  if (bin_var2) data <- data %>%
      mutate(!!sym(var2) := cut(!!sym(var2), breaks = 4))

  # Calculate proportions
  prop_data <- data %>%
    group_by(!!sym(var1), !!sym(var2)) %>%
    summarise(count = n(), .groups = 'drop') %>%
    mutate(prop = count / sum(count))

  ggplot(prop_data, aes_string(x = var1, y = var2)) +
    geom_tile(aes(fill = prop), color = "white") +
    scale_fill_gradientn(colors = brewer.pal(9, "Oranges"), name = "Proportion") +
    labs(title = title, x = var1, y = var2) +
    theme_minimal(base_size = 8) +  # Set smaller base font size
    theme(
      plot.title = element_text(size = 12, hjust = 0.5),
      axis.text.x = element_text(size = 9),
      axis.text.y = element_text(size = 9),
      axis.title.y = element_text(vjust = 0.5),
      legend.key.size = unit(0.45, "cm"),
      plot.margin = margin(1, 1, 1, 1)  # Add space around the plot
    )
}

# Create the plots with adjusted text and layout
p1 <- create_heatmap(data, "BA", "mde_curr", "BA vs. MDD Status")
p2 <- create_heatmap(data, "inc", "edu_merged", "Income vs. Education")
p3 <- create_heatmap(data, "NMR", "sex_ps", "NMR vs. Sex", bin_var1 = TRUE)
p4 <- create_heatmap(data, "ftcd_score", "readiness",
                     "FTCD Score vs. Readiness to Quit Smoking",
                     bin_var1 = TRUE, bin_var2 = TRUE)
```

```
# Arrange the plots in a grid with a title using cowplot
final_plot <- plot_grid(p1, p2, p3, p4, ncol = 2, align = "hv",
                        rel_widths = c(1, 1), rel_heights = c(1, 1))

# Add a title to the entire grid
title <- ggdraw() +
  draw_label("Figure 1: Exploratory Interaction Heatmaps-- Moderator and Covariate Examples",
              size = 16)

# Combine title and grid plot
plot_grid(title, final_plot, ncol = 1, rel_heights = c(0.1, 1))
# Missingness table
# Calculate missing values for each variable
missing_summary <- data %>%
  summarise(across(everything(), ~ sum(is.na(.)), .names = "{col}")) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Number") %>%
  mutate(Percent = (Number / nrow(data)) * 100) %>%
  filter(Number > 0)  # Exclude variables with 0 missingness

# Define a named vector with old and new names for variables
variable_names <- c(
  "inc" = "Income",
  "ftcd_score" = "FTCD score at baseline",
  "crv_total_pq1" = "Cigarette reward value at baseline",
  "shaps_score_pq1" = "Anhedonia",
  "NMR" = "Nicotine Metabolism Ratio",
  "Only.Menthol" = "Exclusive Mentholated Cigarette User",
  "readiness" = "Baseline readiness to quit smoking"
)

# Rename variables in the summary table
missing_summary <- missing_summary %>%
  mutate(Variable = recode(Variable, !!!variable_names))


# Calculate total missing values and total missing percentage
total_rows_with_missing <- sum(rowSums(is.na(data)) > 0)
total_rows_with_missing_pct <- (total_rows_with_missing / nrow(data)) * 100


missing_summary <- missing_summary %>%
  arrange(desc(Percent)) %>%
  mutate(Percent = sprintf("%.2f%%", Percent))

# Combine the total missingness row with the summary table
total_missing_row <- tibble(
  Variable = "Participants with any missingness",
  Number = total_rows_with_missing,
  Percent = sprintf("%.2f%%", total_rows_with_missing_pct)
)

missing_summary <- bind_rows(total_missing_row, missing_summary)
```

```r
# Display the final table
missing_summary %>%
  kable(
    col.names = c("Variable", "Number", "Percent"),
    caption = "Summary of Missing Values"
  ) %>%
  kable_styling(full_width = F, position = "center", font_size = 9)
# Perform MICE imputation
data_mice <- mice(data, m = 5, method = "pmm",
                  maxit = 50, seed = 2024, printFlag = FALSE)


# Complete the data by extracting one of the imputed datasets
data_imp <- complete(data_mice, action = 1)
# and calculate mean for numeric and mode for categorical/binary columns
# data_long <- complete(data_mice, action = "long")

# Define a mode function for categorical/binary variables
# mode_func <- function(x) {
#   ux <- unique(x)
#   ux[which.max(tabulate(match(x, ux)))]
# }

# data_imp = data_long %>%
#   dplyr::select(-.id, -.imp) %>%
#   group_by(id) %>%
#   summarize(across(where(is.numeric), mean, na.rm = TRUE),
#             across(where(~ is.factor(.) || is.character(.)), mode_func)) %>%
#   ungroup() %>%
#   mutate(across(where(is.numeric) &
#                   !matches("NMR"), # NMR is not integer
#                 round))
# Logistic regression with stepwise selection
# Define the outcome and variables in the model
outcome <- data_imp$abst
variable_names <- c("Var", "BA", "age_ps", "sex_ps", "inc", "edu_merged", "race",
                     "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")
variables <- data_imp[, variable_names]
# for Lasso (to break down factors with >2 levels)
variables_dummy <- model.matrix(~ 0 + ., data = variables)
# remove the extra reference group
variables_dummy <- variables_dummy[, -which(colnames(variables_dummy) =="Var0")]

# Split into train and test
set.seed(1)
train_index <- sample(1:nrow(data_imp), 0.7 * nrow(data_imp))
train_data <- data_imp[train_index,]
test_data <- data_imp[-train_index,]
train_outcome <- outcome[train_index]
test_outcome <- outcome[-train_index]
```

```r
# for best subset
train_variables <- variables[train_index, ]
test_variables <- variables[-train_index, ]


# Main effects for `Var` and `BA` (to be considered/controlled in all models)
main_effects <- paste(variable_names, collapse = " + ")

# Define the interaction terms we want to consider
BA_interaction_terms <-
  paste("BA:Var", #interaction between treatments
        #interaction with demographics
        "BA:age_ps + BA:sex_ps + BA:inc+ BA:edu_merged + BA:race",
        # with smoking
        "BA:ftcd_score + BA:Only.Menthol + BA:NMR+ BA:readiness",
        # with MDD
        "BA:mde_curr + BA:bdi_score_w00 + BA:shaps_score_pq1 + BA:antidepmed",
        "BA:otherdiag + BA:shaps_score_pq1",
        sep = " + ")

other_interaction_terms <- paste(
  # pharmacotherapy with demographics
  "Var:age_ps + Var:sex_ps + Var:race + Var:ftcd_score + Var:cpd_ps",
  # between demographics
  "inc:edu_merged",
  # Menthol exclusive with demographics
  "sex_ps:Only.Menthol + race:Only.Menthol + inc:Only.Menthol + edu_merged:Only.Menthol",
  # NMR with demographics, cigarette dependence, and readiness to quit
  "sex_ps:NMR + age_ps:NMR + cpd_ps:NMR + NMR:readiness + ftcd_score:NMR",
  # readiness to quit with cigarette dependence and MDD
  "ftcd_score:readiness + Only.Menthol:readiness + mde_curr:readiness ",
  # cigarette dependence with demographics
  "sex_ps:ftcd_score + race:ftcd_score + age_ps:ftcd_score",
  sep = " + "
)

# Full formula for main effects and interactions
full_formula <- as.formula(paste("abst ~", main_effects, "+", BA_interaction_terms,
                                 "+", other_interaction_terms))


# Define scope with `Var` and `BA` as forced terms in the main model
scope_list <- list(
  lower = as.formula("abst ~ Var + BA"),  # Minimal model with controlled terms
  upper = full_formula                    # Full model with all main and interaction terms
)

# Fit logistic regression model with stepwise selection
set.seed(2024)
stepwise_model <- step(
  glm(formula = abst ~ Var + BA, data = train_data, family = "binomial"),
  scope = scope_list,
  direction = "both",
  trace = 0
)
stepwise_coefs = coef(stepwise_model)
# for L2 + L1
train_variables_dummy <- variables_dummy[train_index, ]
test_variables_dummy <- variables_dummy[-train_index, ]

# Enforce Var and BA as 0 penalty
# Elastic Net
# ^2 generates all pairwise interactions
train_variables_dummy_df <- as.data.frame(train_variables_dummy)
train_variables_dummy_full_interactions <- model.matrix(~ .^2,
                                                         data = train_variables_dummy_df)

test_variables_dummy_df <- as.data.frame(test_variables_dummy)
test_variables_dummy_full_interactions <- model.matrix(~ .^2,
                                                         data = test_variables_dummy_df)

# To identify the potential interaction terms for moderator effects
train_variables_dummy_include_names <- c(
  "Var1", "BA1", "age_ps", "sex_ps2", "inc2", "inc3",
  "inc4", "inc5", "edu_merged2", "edu_merged3",
  "raceBlack", "raceHispanic", "raceOther",
  "ftcd_score", "ftcd.5.mins1", "bdi_score_w00", "cpd_ps",
  "crv_total_pq1", "hedonsum_n_pq1", "hedonsum_y_pq1",
  "shaps_score_pq1", "otherdiag1", "antidepmed1",
  "mde_curr1", "NMR", "Only.Menthol1",
  "readiness", "Var1:BA1",
  #Behavioral treatment (MAIN)
  "BA1:mde_curr1",
  "BA1:age_ps", "BA1:sex_ps2",
  "BA1:raceBlack", "BA1:raceHispanic",
  "BA1:raceOther", "BA1:ftcd_score",
  "BA1:shaps_score_pq1","BA1:bdi_score_w00",
  "BA1:otherdiag1", "BA1:antidepmed1",
  "BA1:mde_curr1","BA1:NMR",
  "BA1:Only.Menthol1", "BA1:readiness",
  # Pharmacotherapy
  "Var1:mde_curr1",
  "Var1:age_ps", "Var1:sex_ps2",
  "Var1:raceBlack", "Var1:raceHispanic",
  "Var1:raceOther", "Var1:ftcd_score",
  # Income*Edu
  "inc2:edu_merged2", "inc2:edu_merged3",
  "inc3:edu_merged2", "inc3:edu_merged3",
  "inc4:edu_merged2", "inc4:edu_merged3",
  "inc5:edu_merged2", "inc5:edu_merged3",
  # Readiness to quit
  "Only.Menthol1:readiness",
  "mde_curr1:readiness", "ftcd_score:readiness",
  # FTCD Score
  "sex_ps2:ftcd_score", "raceBlack:ftcd_score",
  "raceHispanic:ftcd_score", "raceOther:ftcd_score",
  "age_ps:ftcd_score",
  # Menthol exclusive
  "sex_ps2:Only.Menthol1", "raceBlack:Only.Menthol1",
  "raceHispanic:Only.Menthol1", "raceOther:Only.Menthol1",
  "inc2:Only.Menthol1", "inc3:Only.Menthol1",
  "inc4:Only.Menthol1", "inc5:Only.Menthol1",
  "edu_merged2:Only.Menthol1", "edu_merged3:Only.Menthol1",
  # NMR
  "sex_ps2:NMR", "age_ps:NMR", "cpd_ps:NMR",
  "NMR:readiness", "ftcd_score:NMR"
)

train_variables_dummy_include =
  train_variables_dummy_full_interactions[,train_variables_dummy_include_names]
test_variables_dummy_include =
  test_variables_dummy_full_interactions[,train_variables_dummy_include_names]

# Set penalty factors to enforce keeping Var and BA
# Initialize penalty factors to 1 for all variables
penalty_factors <- rep(1, ncol(train_variables_dummy_include))

# Identify columns corresponding exactly to "Var1" and "BA1" (not their interactions)
var1_col <- grep("^Var1$", colnames(train_variables_dummy_include))
ba1_col <- grep("^BA1$", colnames(train_variables_dummy_include))

penalty_factors[c(var1_col, ba1_col)] <- 0
names(penalty_factors) <- colnames(train_variables_dummy_include)
# Fit Elastic Net
set.seed(2024)
enet_model <- cv.glmnet(as.matrix(train_variables_dummy_include), train_outcome,
                        penalty.factor = penalty_factors,
                        alpha = 0.5, family = "binomial")

# Extract coefficients at the optimal lambda (best_lambda)
best_lambda_enet <- enet_model$lambda.min
#remove intercept
optimal_coefs_enet <- as.numeric(coef(enet_model, s = best_lambda_enet)[-1])
coef_names_enet <- rownames(coef(enet_model, s = best_lambda_enet))[-1]

result_table_enet <- data.frame(
  variable = coef_names_enet,
  Coefficient = optimal_coefs_enet
) %>%
  filter(Coefficient != 0)
set.seed(2024)
regsubsets_model <-
    suppressWarnings(regsubsets(
      y = train_outcome,
      x = train_variables_dummy_include,
      nbest = 1, # 1 best model for each number of predictors
      nvmax = 20,
      force.in = c("Var1", "BA1"),
      force.out = NULL,
      really.big = TRUE,
      method = "seqrep",
      warn.dep = FALSE
    ))

reg_summary = summary(regsubsets_model)
cp_min = which.min(reg_summary$cp)
best_subset_coefs = coef(regsubsets_model, cp_min)
best_subset_names = names(coef(regsubsets_model, cp_min))

# Extract variable names from best subset selection and
# fit a glm model with the specified variables
selected_vars <- names(best_subset_coefs)[-1]  # Exclude intercept
formula <- as.formula(paste("abst ~", paste(selected_vars, collapse = " + ")))

model <- glm(formula,
             data = as.data.frame(cbind(train_variables_dummy_include,
                       abst = as.numeric(train_outcome) -1)),
             family = binomial)
best_subset_coefs = coef(model)
# Summary table of coef
large_threshold <- 100

# Calculate OR for each coefficient for each method and rename columns correctly
stepwise_df <- data.frame(
  variable = names(stepwise_coefs),
  `Stepwise Coef` = as.numeric(stepwise_coefs),
  `Stepwise OR` = exp(as.numeric(stepwise_coefs))
)

enet_df <- result_table_enet %>%
  rename(`Elastic Net Coef` = Coefficient) %>%
  mutate(`Elastic Net OR` = exp(`Elastic Net Coef`))

best_subset_df <- data.frame(
  variable = names(best_subset_coefs),
  `Best Subset Coef` = as.numeric(best_subset_coefs),
  `Best Subset OR` = exp(as.numeric(best_subset_coefs))
)

# Merge all data frames based on variable names
combined_df <- full_join(stepwise_df, enet_df, by = "variable") %>%
  full_join(best_subset_df, by = "variable") %>%
  mutate(across(where(is.numeric), ~ round(.x, 4))) %>%
  mutate(across(where(is.numeric), ~ ifelse(as.numeric(.) > large_threshold, "*", .)))

# Remove the intercept row
combined_df <- combined_df[combined_df$variable != "(Intercept)", ]

# Standardize interaction terms by sorting them alphabetically
combined_df <- combined_df %>%
  mutate(
    variable = sapply(variable, function(x) {
      terms <- unlist(strsplit(x, ":"))
      if (length(terms) > 1) {
        paste(sort(terms), collapse = ":")
      } else {
        x
      }
    })
  )

# Combine rows with the same standardized interaction term names
combined_df <- combined_df %>%
  group_by(variable) %>%
  summarize(across(everything(), ~ ifelse(is.numeric(.), sum(as.numeric(.), na.rm = TRUE), .))) %>%
  ungroup()

# Replace zeroes and any remaining NAs with an empty space for readability
combined_df <- combined_df %>%
  mutate(across(where(is.numeric), ~ ifelse(. == 0, "", .))) %>%
  replace(is.na(.), " ")

# Display the final combined table with grouped headers
combined_df %>%
  kable(row.names = F,
        col.names = c("Variable", "Coef", "OR", "Coef", "OR", "Coef", "OR"),
        caption = "Summary of Coefficients and Odds Ratios for Potential Moderator Effects across Model
  add_header_above(c(" " = 1, "Stepwise" = 2, "Elastic Net" = 2, "Best Subset" = 2)) %>%
  kable_styling(full_width = F, position = "center", font_size = 9)

predicted_prob_stepwise <- as.numeric(predict(stepwise_model,
                                newdata = test_data,
                                type = "response"))

# Plot ROC and calculate AUC
roc_stepwise <- roc(test_outcome, predicted_prob_stepwise)

roc_data <- data.frame(
  Specificity = rev(roc_stepwise$specificities),
  Sensitivity = rev(roc_stepwise$sensitivities)
)

# Calculate the AUC
auc_value <- auc(roc_stepwise)

# Plot ROC curve with ggplot2
ROC_stepwise = ggplot(roc_data, aes(x = 1-Specificity, y = Sensitivity)) +
  geom_line(color = "black", size = 1) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey") +
  annotate("text", x = 0.8, y = 0.2,
           label = paste("AUC =", round(auc_value, 2)), size = 5, color = "Black") +
  labs(
    x = "1 - Specificity",
    y = "Sensitivity"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
num_cuts <- 10  # Number of bins for calibration

calib_data <- data.frame(
  prob = predicted_prob_stepwise,  # predicted probabilities
  # binning into `num_cuts` groups
  bin = cut(predicted_prob_stepwise, breaks = num_cuts),
  # observed values (abst outcome in test data)
  class = as.numeric(test_outcome)-1
)

calib_data <- calib_data %>%
  group_by(bin) %>%
  summarise(
    observed = mean(class),
    predicted = mean(prob),
    se = sqrt(observed * (1 - observed) / n())  # Standard error
  )

# Add Loess Fit for Flexible Calibration Line
loess_fit <- loess(observed ~ predicted, data = calib_data, span = 0.75)
calib_data$loess_pred <- predict(loess_fit, calib_data$predicted)

# Plot Calibration Curve with Error Bars
calib_error_bar_stepwise = ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  geom_errorbar(aes(x = predicted, ymin = observed - 1.96 * se,
                    ymax = observed + 1.96 * se),
                colour="black", width=.01)+
  geom_point(aes(x = predicted, y = observed)) +
  labs(x = "Expected Probability of Smoking Abstinence",
       y = "Actual Smoking Abstinence") +
  theme_minimal()


# Plot Calibration Curve with Loess
calib_data <- calib_data %>%
  mutate(loess_ci_lower = loess_pred - 1.96 * sd(loess_pred),
         loess_ci_upper = loess_pred + 1.96 * sd(loess_pred))

calib_loess_stepwise = ggplot(calib_data, aes(x = predicted, y = observed)) +
  # Flexible calibration (Loess)
  geom_line(aes(y = loess_pred), color = "blue", linetype = "dashed") +
  geom_ribbon(aes(ymin = loess_ci_lower, ymax = loess_ci_upper), alpha = 0.2, fill = "grey") +
  geom_abline(intercept = 0, slope = 1, color = "red") +  # Perfect calibration line
  labs(x = "Predicted Probability of Smoking Abstinence",
       y = "Actual Smoking Abstinence") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

predicted_prob_enet <- as.numeric(predict(enet_model,
                                newx = as.matrix(test_variables_dummy_include),
                                s = "lambda.min", type = "response"))

# Plot ROC curve and calculate AUC
roc_enet <- roc(test_outcome, predicted_prob_enet)

roc_data <- data.frame(
  Specificity = rev(roc_enet$specificities),
  Sensitivity = rev(roc_enet$sensitivities)
)

auc_value <- auc(roc_enet)

# Plot ROC curve with ggplot2
ROC_enet = ggplot(roc_data, aes(x = 1-Specificity, y = Sensitivity)) +
  geom_line(color = "black", size = 1) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey") +
  annotate("text", x = 0.8, y = 0.2, label = paste("AUC =", round(auc_value, 2)), size = 5, color = "Bl
  labs(
    x = "1 - Specificity",
    y = "Sensitivity"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
num_cuts <- 10  # Number of bins for calibration

calib_data <- data.frame(
  prob = predicted_prob_enet,  # predicted probabilities
  # binning into `num_cuts` groups
  bin = cut(predicted_prob_enet, breaks = num_cuts),
  # observed values (abst outcome in test data)
  class = as.numeric(test_outcome)-1
)

calib_data <- calib_data %>%
  group_by(bin) %>%
  summarise(
    observed = mean(class),
    predicted = mean(prob),
    se = sqrt(observed * (1 - observed) / n())  # Standard error
  )

# Add Loess Fit for Flexible Calibration Line
loess_fit <- loess(observed ~ predicted, data = calib_data, span = 0.75)
calib_data$loess_pred <- predict(loess_fit, calib_data$predicted)

# Plot Calibration Curve with Error Bars
calib_error_bar_enet = ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  geom_errorbar(aes(x = predicted, ymin = observed - 1.96 * se,
                    ymax = observed + 1.96 * se),
                colour="black", width=.01)+
  geom_point(aes(x = predicted, y = observed)) +
  labs(x = "Expected Probability of Smoking Abstinence",
       y = "Actual Smoking Abstinence") +
  #       title = "Calibration Plot for Elastic Net Model with Error Bars"
  theme_minimal()


# Plot Calibration Curve with Loess
calib_data <- calib_data %>%
  mutate(loess_ci_lower = loess_pred - 1.96 * sd(loess_pred),
         loess_ci_upper = loess_pred + 1.96 * sd(loess_pred))

calib_loess_enet = ggplot(calib_data, aes(x = predicted, y = observed)) +
  # Flexible calibration (Loess)
  geom_line(aes(y = loess_pred), color = "blue", linetype = "dashed") +
  geom_ribbon(aes(ymin = loess_ci_lower, ymax = loess_ci_upper), alpha = 0.2, fill = "grey") +
  geom_abline(intercept = 0, slope = 1, color = "red") +  # Perfect calibration line
  scale_color_manual(values = c("Ideal" = "red",
                                "Flexible calibration" = "blue")) +
  scale_linetype_manual(values = c("Ideal" = "solid",
                                   "Flexible calibration" = "dashed")) +
  labs(x = "Predicted Probability of Smoking Abstinence",
       y = "Actual Smoking Abstinence",
       color = "Legend", linetype = "Legend") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

predict_best_subset <- function(train_data, test_data, best_subset_coefs) {
  train_data <- as.data.frame(train_data)
  test_data <- as.data.frame(test_data)

  # Extract variable names from best subset selection
  selected_vars <- names(best_subset_coefs)[-1]  # Exclude intercept
  formula <- as.formula(paste("abst ~", paste(selected_vars, collapse = " + ")))

  # Fit a glm model with the specified variables
  model <- glm(formula, data = train_data, family = binomial)

  # Predict on test data using type = "response" to get probabilities
  predicted_probabilities <- predict(model, newdata = test_data, type = "response")

  return(predicted_probabilities)
}

predicted_prob_best_subset <-  as.numeric(
  predict_best_subset(
    train_data = cbind(train_variables_dummy_include,
                       abst = as.numeric(train_outcome) -1),
    test_data = cbind(test_variables_dummy_include,
                      abst = as.numeric(test_outcome) -1),
    best_subset_coefs
  )
)
```

```r
# Plot ROC and calculate AUC
roc_best_subset <- roc(test_outcome, predicted_prob_best_subset)

roc_data <- data.frame(
  Specificity = rev(roc_best_subset$specificities),
  Sensitivity = rev(roc_best_subset$sensitivities)
)

auc_value <- auc(roc_best_subset)

# Plot ROC curve with ggplot2
ROC_best_subset = ggplot(roc_data, aes(x = 1-Specificity, y = Sensitivity)) +
  geom_line(color = "black", size = 1) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey") +
  annotate("text", x = 0.8, y = 0.2,
           label = paste("AUC =", round(auc_value, 2)), size = 5, color = "Black") +
  labs(
    x = "1 - Specificity",
    y = "Sensitivity"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
num_cuts <- 10  # Number of bins for calibration

calib_data <- data.frame(
  prob = predicted_prob_best_subset,  # predicted probabilities
  # binning into `num_cuts` groups
  bin = cut(predicted_prob_best_subset, breaks = num_cuts),
  # observed values (abst outcome in test data)
  class = as.numeric(test_outcome)-1
)

calib_data <- calib_data %>%
  group_by(bin) %>%
  summarise(
    observed = mean(class),
    predicted = mean(prob),
    se = sqrt(observed * (1 - observed) / n())  # Standard error
  )

# Add Loess Fit for Flexible Calibration Line
loess_fit <- loess(observed ~ predicted, data = calib_data, span = 0.75)
calib_data$loess_pred <- predict(loess_fit, calib_data$predicted)

# Plot Calibration Curve with Error Bars
calib_error_bar_best_subset = ggplot(calib_data) +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  geom_errorbar(aes(x = predicted, ymin = observed - 1.96 * se,
                    ymax = observed + 1.96 * se),
                colour="black", width=.01)+
  geom_point(aes(x = predicted, y = observed)) +
  labs(x = "Expected Probability of Smoking Abstinence",
       y = "Actual Smoking Abstinence") +
  theme_minimal()


# Plot Calibration Curve with Loess
calib_data <- calib_data %>%
  mutate(loess_ci_lower = loess_pred - 1.96 * sd(loess_pred),
         loess_ci_upper = loess_pred + 1.96 * sd(loess_pred))

calib_loess_best_subset = ggplot(calib_data, aes(x = predicted, y = observed)) +
  # Flexible calibration (Loess)
  geom_line(aes(y = loess_pred), color = "blue", linetype = "dashed") +
  geom_ribbon(aes(ymin = loess_ci_lower, ymax = loess_ci_upper),
              alpha = 0.2, fill = "grey") +
  geom_abline(intercept = 0, slope = 1, color = "red") +  # Perfect calibration line
  labs(x = "Predicted Probability of Smoking Abstinence",
       y = "Actual Smoking Abstinence") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

test_outcome_numeric = as.numeric(test_outcome) - 1

# Calculate Brier Score for each model
brier_stepwise <- mean((predicted_prob_stepwise - test_outcome_numeric)^2)
brier_enet <- mean((predicted_prob_enet - test_outcome_numeric)^2)
brier_best_subset <- mean((predicted_prob_best_subset - test_outcome_numeric)^2)


# calibration error
ce_calculate <- function(predictions, actuals, n_bins = 10) {
  bins <- cut(predictions, breaks = seq(0, 1, length.out = n_bins + 1),
              include.lowest = TRUE)
  bin_means <- tapply(predictions, bins, mean)
  bin_actuals <- tapply(actuals, bins, mean)
  bin_weights <- table(bins) / length(predictions)
  ce <- sum(bin_weights * abs(bin_means - bin_actuals))

  # Remove NA values from bin_means and bin_actuals
  valid_bins <- !is.na(bin_means) & !is.na(bin_actuals)
  bin_means <- bin_means[valid_bins]
  bin_actuals <- bin_actuals[valid_bins]
  bin_weights <- bin_weights[valid_bins]

  # Calculate ECE with valid bins only
  CE <- sum(bin_weights * abs(bin_means - bin_actuals))

  return(CE)
}

CE_stepwise <- ce_calculate(predicted_prob_stepwise, test_outcome_numeric)
CE_enet <- ce_calculate(predicted_prob_enet, test_outcome_numeric)
CE_best_subset <- ce_calculate(predicted_prob_best_subset, test_outcome_numeric)
```

```r
# Calculate AUC for each model
auc_stepwise <- roc_stepwise$auc
auc_enet <- roc_enet$auc
auc_best_subset <- roc_best_subset$auc

# Get optimal threshold, specificity, and sensitivity for each model
threshold_stepwise <- as.numeric(coords(roc_stepwise, "best",
                                        ret = "threshold"))
specificity_stepwise <- as.numeric(coords(roc_stepwise, "best",
                                          ret = "specificity"))
sensitivity_stepwise <- as.numeric(coords(roc_stepwise, "best",
                                          ret = "sensitivity"))

threshold_enet <- as.numeric(coords(roc_enet, "best",
                                    ret = "threshold"))
specificity_enet <- as.numeric(coords(roc_enet, "best",
                                      ret = "specificity"))
sensitivity_enet <- as.numeric(coords(roc_enet, "best",
                                      ret = "sensitivity"))

threshold_best_subset <- as.numeric(coords(roc_best_subset, "best",
                                           ret = "threshold"))
specificity_best_subset <- as.numeric(coords(roc_best_subset, "best",
                                             ret = "specificity"))
sensitivity_best_subset <- as.numeric(coords(roc_best_subset, "best",
                                             ret = "sensitivity"))

# Combine metrics into a table
df_performance <- rbind(
  `Brier score` = round(c(brier_stepwise, brier_enet, brier_best_subset), 4),
  `Calibration error` = round(c(CE_stepwise, CE_enet, CE_best_subset), 4),
  AUC = round(c(auc_stepwise, auc_enet, auc_best_subset), 4),
  Threshold = round(c(threshold_stepwise, threshold_enet, threshold_best_subset), 4),
  Specificity = round(c(specificity_stepwise, specificity_enet, specificity_best_subset), 4),
  Sensitivity = round(c(sensitivity_stepwise, sensitivity_enet, sensitivity_best_subset), 4)
  #`Adjusted R^2` = round(c(adj_r2_stepwise, adj_r2_enet, adj_r2_best_subset), 4)
)

# rename columns
colnames(df_performance) <- c("Stepwise", "Elastic Net", "Best Subset")

# Display the final table
kable(
  df_performance,
  caption = "Calibration and Discrimination Metrics for Moderator Effects Modeling"
)
plots_stepwise = arrangeGrob(
  calib_error_bar_stepwise, calib_loess_stepwise, ROC_stepwise,
  ncol = 3,
  top = textGrob("Stepwise Selection",
                 gp = gpar(fontface = "bold", fontsize = 14)
))
```

```r
plots_enet = arrangeGrob(
  calib_error_bar_enet, calib_loess_enet, ROC_enet,
  ncol = 3,
  top = textGrob("Elastic Net",
                gp = gpar(fontface = "bold", fontsize = 14)
))

plots_best_subset = arrangeGrob(
  calib_error_bar_best_subset, calib_loess_best_subset, ROC_best_subset,
  ncol = 3,
  top = textGrob("Best Subset Selection",
                gp = gpar(fontface = "bold", fontsize = 14)
))

# Bold the main title
main_title <- textGrob(
  "Figure 2: Calibration Plots with Error Bars and LOESS and ROC Curves (Moderator Effects)",
  gp = gpar(fontsize = 16)
)

# Arrange everything with the bold title
grid.arrange(
  plots_stepwise,
  plots_enet,
  plots_best_subset,
  nrow = 3,
  top = main_title
)
predictor_names <- c("Var", "BA", "age_ps", "sex_ps", "inc", "edu_merged", "race",
                     "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")
predictors <- data_imp[, predictor_names]
# for Lasso (to break down factors with >2 levels)
predictors_dummy <- model.matrix(~ 0 + ., data = predictors)
# remove the extra reference group
predictors_dummy <- predictors_dummy[, -which(colnames(predictors_dummy) =="Var0")]


# Full formula for main effects and interactions
full_formula <- as.formula(paste("abst ~", main_effects))

# Define scope with `Var` and `BA` as forced terms in the main model
scope_list <- list(
  lower = as.formula("abst ~ Var + BA"),  # Minimal model with controlled terms
  upper = full_formula                    # Full model with all main terms
)

# Fit logistic regression model with stepwise selection
set.seed(2024)
predictor_stepwise_model <- step(
  glm(formula = abst ~ Var + BA, data = train_data, family = "binomial"),
  scope = scope_list,
  direction = "both",
  trace = 0
)
# Fit Elastic Net
# To identify the potential interaction terms for moderator effects
train_predictors_dummy_include_names <- c(
  "Var1", "BA1", "age_ps", "sex_ps2", "inc2", "inc3",
  "inc4", "inc5", "edu_merged2", "edu_merged3",
  "raceBlack", "raceHispanic", "raceOther",
  "ftcd_score", "ftcd.5.mins1", "bdi_score_w00", "cpd_ps",
  "crv_total_pq1", "hedonsum_n_pq1", "hedonsum_y_pq1",
  "shaps_score_pq1", "otherdiag1", "antidepmed1",
  "mde_curr1", "NMR", "Only.Menthol1",
  "readiness"
)

train_predictors_dummy <- predictors_dummy[train_index, ]
test_predictors_dummy <- predictors_dummy[-train_index, ]

train_predictors_dummy_include =
  train_predictors_dummy[,train_predictors_dummy_include_names]
test_predictors_dummy_include =
  test_predictors_dummy[,train_predictors_dummy_include_names]

# Set penalty factors to enforce keeping Var and BA
# Initialize penalty factors to 1 for all variables
penalty_factors <- rep(1, ncol(train_predictors_dummy_include))

# Identify columns corresponding exactly to "Var1" and "BA1" (not their interactions)
var1_col <- grep("^Var1$", colnames(train_predictors_dummy_include))
ba1_col <- grep("^BA1$", colnames(train_predictors_dummy_include))

penalty_factors[c(var1_col, ba1_col)] <- 0
names(penalty_factors) <- colnames(train_predictors_dummy_include)


set.seed(2024)
predictor_enet_model <- cv.glmnet(as.matrix(train_predictors_dummy_include),
                                  train_outcome,
                         penalty.factor = penalty_factors,
                         alpha = 0.5, family = "binomial")

# Extract coefficients at the optimal lambda (best_lambda)
best_lambda_enet <- predictor_enet_model$lambda.min
#remove intercept
optimal_coefs_enet <- as.numeric(coef(predictor_enet_model, s = best_lambda_enet)[-1])
coef_names_enet <- rownames(coef(predictor_enet_model, s = best_lambda_enet))[-1]

predictor_result_table_enet <- data.frame(
  variable = coef_names_enet,
  Coefficient = optimal_coefs_enet
) %>%
```

```r
  filter(Coefficient != 0)
set.seed(2024)
predictor_regsubsets_model <-
    regsubsets(y = train_outcome,
               x = train_predictors_dummy_include,
               nbest = 1,       # 1 best model for each number of predictors
               nvmax = 20,
               force.in = c("Var1", "BA1"),
               force.out = NULL,
               really.big = T,
               method = "exhaustive")

predictor_reg_summary = summary(predictor_regsubsets_model)
cp_min = which.min(predictor_reg_summary$cp)
predictor_best_subset_coefs = coef(predictor_regsubsets_model, cp_min)
predictor_best_subset_names = names(coef(predictor_regsubsets_model, cp_min))

# Extract variable names from best subset selection and
# fit a glm model with the specified variables
selected_vars <- names(predictor_best_subset_coefs)[-1]  # Exclude intercept
formula <- as.formula(paste("abst ~", paste(selected_vars, collapse = " + ")))

model <- glm(formula,
             data = as.data.frame(cbind(train_predictors_dummy,
                       abst = as.numeric(train_outcome) -1)),
             family = binomial)
predictor_best_subset_coefs = coef(model)
# Summary table of coef
# Stepwise coefficients and OR
stepwise_coefs <- coef(predictor_stepwise_model)
stepwise_df <- data.frame(
  variable = names(stepwise_coefs),
  `Stepwise Coef` = as.numeric(stepwise_coefs),
  `Stepwise OR` = exp(as.numeric(stepwise_coefs))
)

# Elastic Net coefficients and OR
enet_df <- predictor_result_table_enet %>%
  rename(`Elastic Net Coef` = Coefficient) %>%
  mutate(`Elastic Net OR` = exp(`Elastic Net Coef`))

# Best Subset coefficients and OR
best_subset_df <- data.frame(
  variable = names(predictor_best_subset_coefs),
  `Best Subset Coef` = as.numeric(predictor_best_subset_coefs),
  `Best Subset OR` = exp(as.numeric(predictor_best_subset_coefs))
)

# Merge all data frames based on variable names
combined_df <- full_join(stepwise_df, enet_df, by = "variable") %>%
  full_join(best_subset_df, by = "variable") %>%
  mutate(across(where(is.numeric), ~ round(.x, 4)))
```

```r
# Remove the intercept row
combined_df <- combined_df[combined_df$variable != "(Intercept)", ]

# Replace NA values with an empty space
combined_df[is.na(combined_df)] <- " "

# Enforce predictor print-out order, keeping BA1 and Var1 at the top
combined_df <- combined_df %>%
  arrange(
    factor(variable, levels = c("BA1", "Var1")),  # Keep BA1 and Var1 at the top
    variable                                      # Sort the rest alphabetically
  )

# Create the table with grouped headers
combined_df %>%
  kable(
    row.names = F,
    col.names = c("Variable", "Coef", "OR", "Coef", "OR", "Coef", "OR"),
    caption = "Summary of Coefficients and Odds Ratios for Potential Predictor Effects across Model Sele
  ) %>%
  add_header_above(c(
    " " = 1,
    "Stepwise" = 2,
    "Elastic Net" = 2,
    "Best Subset" = 2
  )) %>%
  kable_styling(full_width = F, position = "center", font_size = 9)
predictor_prob_stepwise <- as.numeric(predict(predictor_stepwise_model,
                                newdata = test_data,
                                type = "response"))

# Plot ROC and calculate AUC
predictor_roc_stepwise <- roc(test_outcome, predictor_prob_stepwise)

roc_data <- data.frame(
  Specificity = rev(predictor_roc_stepwise$specificities),
  Sensitivity = rev(predictor_roc_stepwise$sensitivities)
)

auc_value <- auc(predictor_roc_stepwise)

# Plot ROC curve with ggplot2
predictor_ROC_stepwise = ggplot(roc_data, aes(x = 1-Specificity, y = Sensitivity)) +
  geom_line(color = "black", size = 1) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey") +
  annotate("text", x = 0.8, y = 0.2,
           label = paste("AUC =", round(auc_value, 2)), size = 5, color = "Black") +
  labs(
    x = "1 - Specificity",
    y = "Sensitivity"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
predictor_prob_enet <- as.numeric(predict(predictor_enet_model,
                                newx = as.matrix(test_predictors_dummy),
                                s = "lambda.min", type = "response"))

# Plot ROC curve and calculate AUC
predictor_roc_enet <- roc(test_outcome, predictor_prob_enet)

# Convert the ROC object to a data frame for ggplot2
roc_data <- data.frame(
  Specificity = rev(predictor_roc_enet$specificities),
  Sensitivity = rev(predictor_roc_enet$sensitivities)
)

auc_value <- auc(predictor_roc_enet)

# Plot ROC curve with ggplot2
predictor_ROC_enet = ggplot(roc_data, aes(x = 1-Specificity, y = Sensitivity)) +
  geom_line(color = "black", size = 1) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey") +
  annotate("text", x = 0.8, y = 0.2, label = paste("AUC =", round(auc_value, 2)), size = 5, color = "Bl
  labs(
    x = "1 - Specificity",
    y = "Sensitivity"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
predictor_prob_best_subset <-  as.numeric(
  predict_best_subset(
    train_data = cbind(train_predictors_dummy,
                       abst = as.numeric(train_outcome) -1),
    test_data = cbind(test_predictors_dummy,
                      abst = as.numeric(test_outcome) -1),
    predictor_best_subset_coefs
  )
)

# Plot ROC and calculate AUC
predictor_roc_best_subset <- roc(test_outcome, predictor_prob_best_subset)

roc_data <- data.frame(
  Specificity = rev(predictor_roc_best_subset$specificities),
  Sensitivity = rev(predictor_roc_best_subset$sensitivities)
)

auc_value <- auc(predictor_roc_best_subset)

# Plot ROC curve with ggplot2
predictor_ROC_best_subset = ggplot(roc_data, aes(x = 1-Specificity, y = Sensitivity)) +
  geom_line(color = "black", size = 1) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey") +
  annotate("text", x = 0.8, y = 0.2,
           label = paste("AUC =", round(auc_value, 2)), size = 5, color = "Black") +
  labs(
    x = "1 - Specificity",
    y = "Sensitivity"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
# Calculate Brier Score for each model
brier_stepwise <- mean((predictor_prob_stepwise - test_outcome_numeric)^2)
brier_enet <- mean((predictor_prob_enet - test_outcome_numeric)^2)
brier_best_subset <- mean((predictor_prob_best_subset - test_outcome_numeric)^2)


# calibration error
ce_calculate <- function(predictions, actuals, n_bins = 10) {
  bins <- cut(predictions, breaks = seq(0, 1, length.out = n_bins + 1),
              include.lowest = TRUE)
  bin_means <- tapply(predictions, bins, mean)
  bin_actuals <- tapply(actuals, bins, mean)
  bin_weights <- table(bins) / length(predictions)
  ce <- sum(bin_weights * abs(bin_means - bin_actuals))

  # Remove NA values from bin_means and bin_actuals
  valid_bins <- !is.na(bin_means) & !is.na(bin_actuals)
  bin_means <- bin_means[valid_bins]
  bin_actuals <- bin_actuals[valid_bins]
  bin_weights <- bin_weights[valid_bins]

  # Calculate ECE with valid bins only
  CE <- sum(bin_weights * abs(bin_means - bin_actuals))

  return(CE)
}

CE_stepwise <- ce_calculate(predictor_prob_stepwise,
                                      test_outcome_numeric)
CE_enet <- ce_calculate(predictor_prob_enet,
                                  test_outcome_numeric)
CE_best_subset <- ce_calculate(predictor_prob_best_subset,
                                         test_outcome_numeric)
# Calculate AUC for each model
auc_stepwise <- predictor_roc_stepwise$auc
auc_enet <- predictor_roc_enet$auc
auc_best_subset <- predictor_roc_best_subset$auc

# Get optimal threshold, specificity, and sensitivity for each model
threshold_stepwise <- as.numeric(coords(predictor_roc_stepwise, "best",
                                        ret = "threshold"))
specificity_stepwise <- as.numeric(coords(predictor_roc_stepwise, "best",
                                          ret = "specificity"))
sensitivity_stepwise <- as.numeric(coords(predictor_roc_stepwise, "best",
                                          ret = "sensitivity"))
```

```r
threshold_enet <- as.numeric(coords(predictor_roc_enet, "best",
                                    ret = "threshold"))
specificity_enet <- as.numeric(coords(predictor_roc_enet, "best",
                                      ret = "specificity"))
sensitivity_enet <- as.numeric(coords(predictor_roc_enet, "best",
                                      ret = "sensitivity"))

threshold_best_subset <- as.numeric(coords(predictor_roc_best_subset, "best",
                                           ret = "threshold"))
specificity_best_subset <- as.numeric(coords(predictor_roc_best_subset, "best",
                                             ret = "specificity"))
sensitivity_best_subset <- as.numeric(coords(predictor_roc_best_subset, "best",
                                             ret = "sensitivity"))

# Combine metrics into a table
predictor_df_performance <- rbind(
  `Brier score` = round(c(brier_stepwise, brier_enet, brier_best_subset), 4),
  `Calibration error` = round(c(CE_stepwise, CE_enet, CE_best_subset), 4),
  AUC = round(c(auc_stepwise, auc_enet, auc_best_subset), 4),
  Threshold = round(c(threshold_stepwise, threshold_enet, threshold_best_subset), 4),
  Specificity = round(c(specificity_stepwise, specificity_enet, specificity_best_subset), 4),
  Sensitivity = round(c(sensitivity_stepwise, sensitivity_enet, sensitivity_best_subset), 4)
  #`Adjusted R^2` = round(c(adj_r2_stepwise, adj_r2_enet, adj_r2_best_subset), 4)
)

# rename columns
colnames(predictor_df_performance) <- c("Stepwise", "Elastic Net", "Best Subset")

# Display the final table
kable(
  predictor_df_performance,
  caption = "Calibration and Discrimination Metrics for Predictor Effect Modeling"
)
```