Exploring Predictors and Moderators of Smoking Abstinence in Behavioral and Pharmacotherapy Treatments for Smokers with Major Depressive Disorder

Jizhou Tian

2024-11

Introduction

Smoking cessation remains a formidable challenge, particularly for individuals with major depressive disorder (MDD), who are more likely to smoke heavily, display stronger nicotine dependence, and experience severe withdrawal symptoms. These difficulties are compounded by depression-related psychological factors, which can amplify the rewards of nicotine and reduce motivation to abstinence. To address these issues, a randomized, placebo-controlled, 2x2 factorial study examined the effects of Behavioral Activation for Smoking Cessation (BASC) versus standard treatment (ST), along with adjunctive varenicline versus placebo, among adult smokers with current or past MDD. The study aimed to assess whether behavioral and pharmacological interventions could enhance smoking cessation rates in this high-risk population.

The findings from this study indicated that while varenicline—a widely used pharmacotherapy for smoking cessation—was effective in improving abstinence rates relative to placebo, BASC alone did not outperform standard behavioral treatment, regardless of whether varenicline was used. These results suggest that while pharmacotherapy can aid smoking cessation in individuals with MDD, researchers found no evidence that BASC was more effective than ST in increasing cessation.

This project aims to build upon these findings by examining baseline variables as potential moderators of the effects of behavioral treatment on end-of-treatment (EOT) abstinence, and evaluating baseline variables as predictors of abstinence while controlling for behavioral treatment and pharmacotherapy. By identifying these moderators, the project seeks to enhance our understanding of how behavioral activation might promote smoking cessation, especially for adults with MDD, and provide insights into tailoring treatment strategies to improve success rates among this vulnerable population.

Methods

Study setting and population

The study was a randomized, placebo-controlled, 2x2 factorial trial conducted at research clinics in Northwestern University and the University of Pennsylvania. It aimed to assess the efficacy and safety of BASC combined with varenicline for adults with current or past MDD. Participants included **300** adult smokers who had smoked daily (>=1 cigarette per day) and had a lifetime diagnosis of MDD without psychotic features. Eligible individuals expressed an interest in quitting smoking and underwent both initial and final eligibility screening, including informed consent, randomization, and a baseline assessment at intake (week 0). Randomization assigned participants to one of four arms: BASC with varenicline, BASC with placebo, ST with varenicline, and ST with placebo.

A total of 25 variables, including a column for participant ID, were collected in this study. Smoking abstinence is our primary outcome of interest. Varenicline and Behavioral Activation (BA) are the pharmacotherapy

and psychotherapy treatments, respectively. Demographic variables include age, sex, race, ethnicity (i.e., whether participants are non-Hispanic white, Black, or Hispanic), income, and education.

Smoking-related baseline variables include the FTCD score (which measures the level of nicotine dependence), whether participants smoke within 5 minutes of waking up, cigarettes per day, cigarette reward value, scores on the pleasurable events scale for substitute and complementary reinforcers, nicotine metabolism ratio (NMR), exclusive menthol cigarette use, and baseline readiness to quit smoking.

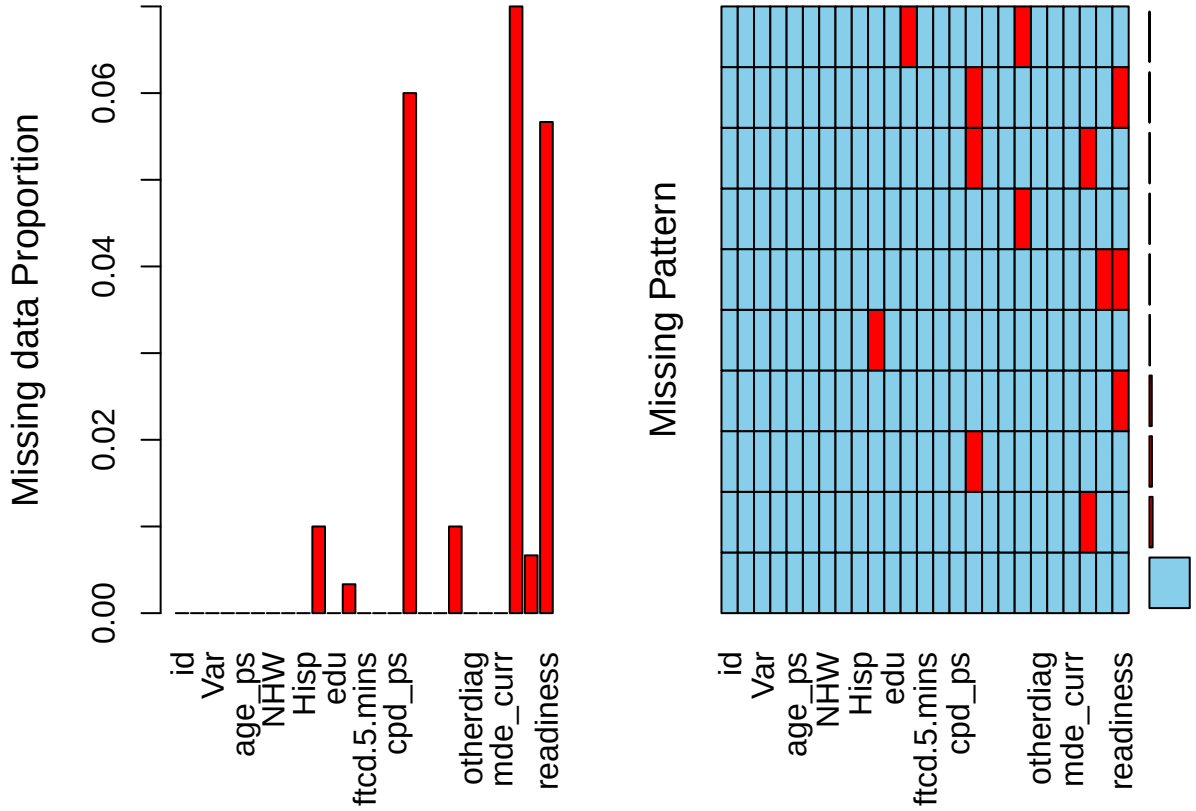
Mental health-related baseline variables include the BDI score (assessing the severity of depressive symptoms), SHAPS score (measuring anhedonia), whether participants have any other lifetime DSM-5 diagnosis (i.e., other mental health disorders), whether they are taking antidepressant medication, and whether they currently have major depressive disorder (MDD).

An overview of participant characteristics

Variable	Description
abst	Smoking Abstinence
Var (Varenicline)	Pharmacotherapy
BA (Behavioral Activation)	Psychotherapy
age_ps	Age at phone interview
sex_ps	Sex at phone interview
NHW	Non-Hispanic White indicator
Black	Black indicator
Hisp	Hispanic indicator
inc	Income (ordinal categorical, low to high)
edu	Education (ordinal categorical, low to high)
ftcd_score	FTCD score at baseline
ftcd.5.mins	Smoking with 5 mins of waking up
bdi_score_pq1	BDI score at baseline
cpd_ps	Cigarettes per day at baseline phone survey
crv_total_pq1	Cigarette reward value at baseline
hedonsum_n_pq1	Pleasurable Events Scale at baseline – substitute reinforcers
hedonsum_y_pq1	Pleasurable Events Scale at baseline – complementary reinforcers
shaps_score_pq1	Anhedonia
otherdiag	Other lifetime DSM-5 diagnosis
antidepmed	Taking antidepressant medication at baseline
mde_curr	Current vs past MDD
NMR	Nicotine Metabolism Ratio
Only.Menthol	Exclusive Mentholated Cigarette User
readiness	Baseline readiness to quit smoking

Data processing

We began by assessing the extent of missing data. Nicotine Metabolism Ratio had the highest proportion of missing data, with 21 cases missing (7%). This was followed by Cigarette Reward Value at baseline, with 18 missing cases (6%), and Baseline Readiness to Quit Smoking, with 17 missing cases (5.7%). For all other variables with missing data, the number of missing cases was no more than 3. No specific missing data pattern was identified. We removed participants with missing data and proceeded with complete cases only for further analysis, resulting in a final sample size of 241.



Since this study is a 2x2 factorial trial, we then examined the distribution of participant characteristics across the different intervention groups. As shown in the table, aside from education level and baseline antidepressant medication use, there were no significant differences in the distribution of other variables among the groups."

Table 2: Participant Characteristics Summary Table

Variables	**placebo + BASC** N = 53	**placebo + ST** N = 54	**varenicline + BASC** N = 71	**varenicline + ST** N = 63	**p-value**
age_ps	54 (43, 60)	51 (46, 58)	53 (40, 60)	51 (38, 58)	0.4
sex_ps					0.9
1	22 (42%)	23 (43%)	34 (48%)	29 (46%)	
2	31 (58%)	31 (57%)	37 (52%)	34 (54%)	
NHW	18 (34%)	19 (35%)	29 (41%)	20 (32%)	0.7
Black	30 (57%)	30 (56%)	32 (45%)	32 (51%)	0.6
Hisp	3 (5.7%)	4 (7.4%)	3 (4.2%)	4 (6.3%)	>0.9
inc					0.4
1	20 (38%)	18 (33%)	25 (35%)	22 (35%)	
2	13 (25%)	12 (22%)	15 (21%)	16 (25%)	
3	5 (9.4%)	13 (24%)	12 (17%)	8 (13%)	
4	11 (21%)	5 (9.3%)	10 (14%)	5 (7.9%)	
5	4 (7.5%)	6 (11%)	9 (13%)	12 (19%)	
edu					0.037
1	1 (1.9%)	0 (0%)	0 (0%)	0 (0%)	
2	2 (3.8%)	1 (1.9%)	6 (8.5%)	3 (4.8%)	
3	20 (38%)	9 (17%)	13 (18%)	21 (33%)	

Variables	placebo + BASC N = 53	placebo + ST N = 54	varenicline + BASC N = 71	varenicline + ST N = 63	p-value
4	16 (30%)	30 (56%)	27 (38%)	18 (29%)	
5	14 (26%)	14 (26%)	25 (35%)	21 (33%)	
ftcd_score	5 (4, 7)	6 (4, 7)	5 (4, 7)	5 (4, 7)	0.5
ftcd.5.mins	24 (45%)	27 (50%)	27 (38%)	30 (48%)	0.5
bdi_score_w00	18 (8, 27)	18 (9, 24)	17 (9, 23)	18 (12, 27)	>0.9
cpd_ps	15 (10, 20)	14 (10, 20)	15 (10, 20)	15 (10, 20)	>0.9
crv_total_pq1	7 (5, 10)	7 (4, 9)	8 (5, 10)	7 (5, 9)	0.9
hedonsum_n	20 (8, 30)	14 (9, 27)	19 (9, 33)	20 (9, 36)	0.7
hedonsum_y	24 (15, 36)	24 (10, 36)	18 (12, 31)	19 (14, 34)	0.6
shaps_score	0 (0, 3)	1 (0, 5)	1 (0, 4)	1 (0, 3)	0.7
otherdiag	26 (49%)	23 (43%)	27 (38%)	30 (48%)	0.6
antidepmed	23 (43%)	10 (19%)	21 (30%)	12 (19%)	0.010
mde_curr	26 (49%)	22 (41%)	31 (44%)	34 (54%)	0.5
NMR	0.32 (0.23, 0.49)	0.32 (0.20, 0.42)	0.31 (0.21, 0.48)	0.28 (0.20, 0.51)	>0.9
Only.Menthol	33 (62%)	33 (61%)	42 (59%)	37 (59%)	>0.9
readiness					0.7
3	1 (1.9%)	0 (0%)	0 (0%)	0 (0%)	
4	1 (1.9%)	1 (1.9%)	2 (2.8%)	0 (0%)	
5	4 (7.5%)	9 (17%)	9 (13%)	6 (9.5%)	
6	16 (30%)	9 (17%)	21 (30%)	25 (40%)	
7	14 (26%)	15 (28%)	18 (25%)	17 (27%)	
8	14 (26%)	16 (30%)	19 (27%)	12 (19%)	
9	2 (3.8%)	1 (1.9%)	1 (1.4%)	2 (3.2%)	
10	1 (1.9%)	3 (5.6%)	1 (1.4%)	1 (1.6%)	

We also checked the correlation between continuous variables in our data. For variables whose measurement scale for data is ordinal, we treated them as continuous because all of them have at least five categories. In the correlation plot, we found that FTCD score at baseline and Cigarettes per day at baseline phone survey have strong positive correlation (0.52), so is BDI score at baseline and SHAPS score (Anhedonia) (0.41). A relative strong negative correlation is found between BDI score at baseline and Pleasurable Events Scale at baseline – substitute reinforcers (-0.39).

Model derivation

The goal of the study is to examine baseline variables as potential moderators of the effects of behavioral treatment on abstinence, and to evaluate baseline variables as predictors of abstinence while controlling for behavioral treatment and pharmacotherapy. Thus, variable selection techniques and the addition of interaction terms between baseline variables and behavioral treatment need to be considered in model building.

Since our primary outcome of interest is binary, our model is structured based on logistic regression. We first conducted Logistic LASSO regression with L1 regularization using the **glmnet** package in R (Friedman, Hastie, and Tibshirani, 2010). **Model 1** was created by applying LASSO to all variables without adding interaction terms. This model aims to roughly identify which variables are important and retained in the model. **Model 2** was then built using LASSO, incorporating all variables along with the interactions between BASC and all smoking-related and mental health-related variables, as well as the interactions between Varenicline and all smoking-related and mental health-related variables. This model is intended to examine baseline variables as potential moderators. By comparing the two models, we can assess the impact of adding interaction terms on the model results.

To account for nonlinearity and potential interactions, we also applied random forest using the **randomForest** package in R (Liaw and Wiener, 2002). **Model 3** was created by applying random forest to all variables

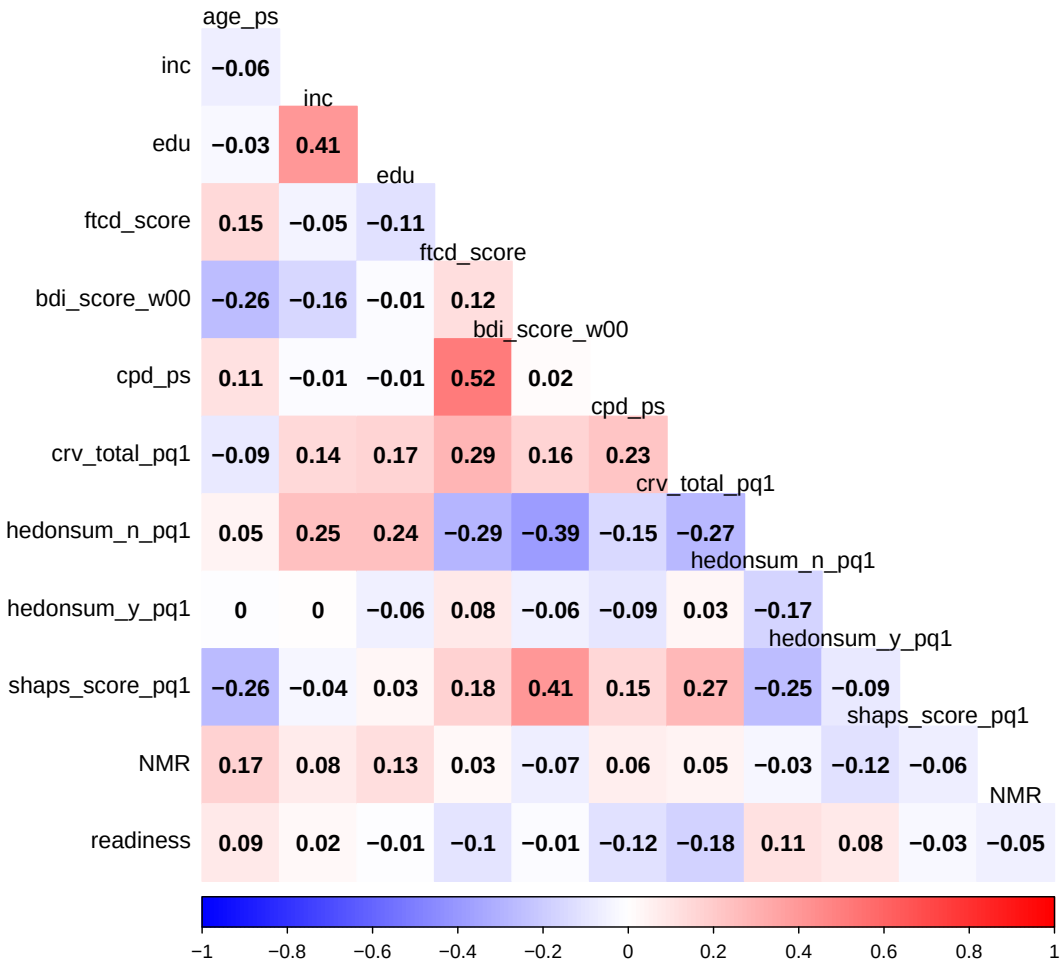


Figure 1: Correlation Plot

without adding interaction terms. Model 4 is similar to Model 2, including the same interaction terms, but employs random forest instead of Logistic LASSO to address the classification problem.

Model performance

We fitted Models 1 through 4 using the entire data set. For Models 1 and 2, cross-validation was used to determine the optimal λ for each logistic LASSO regression. For Models 3 and 4, we used the default settings for the random forest parameters: the number of trees grown was set to 500, the minimum size of terminal nodes was 1 for the classification problem, and the number of variables randomly sampled as candidates at each split was $\sqrt{p}$, where p is the number of variables.

To evaluate model performance, we conducted 5-fold cross-validation for Model 1 and Model 2, and using out-of-bag prediction for Model 3 and Model 4. We recorded the AUC value for each model. We also obtained importance plots separately from the two random forest models.

Results

Logistic LASSO regression

As shown in Table 3, which presents the coefficient tables for Models 1 and 2, we observed that in Model 1, the coefficients for Varenicline (0.976), Non-Hispanic White indicator (0.281), FTCD score at baseline (-0.189), Anhedonia (-0.02), Current MDD (-0.28), and Nicotine Metabolism Ratio (0.369) were not eliminated by LASSO. Model 2 has similar significant variables, with the addition of the income variable (0.017) and the exclusion of Nicotine Metabolism Ratio. Notably, the coefficient for Varenicline decreased substantially from 0.976 in Model 1 to 0.111 in Model 2, while changes in the coefficients for other variables were minimal. Most interaction terms were eliminated in Model 2, with only three remaining: BASC with FTCD score at baseline (-0.014), Varenicline with Smoking within 5 minutes of waking up (0.707), and Varenicline with Nicotine Metabolism Ratio (1.734).

Based on these results, we can conclude that Varenicline is an important predictor (present in both models), whereas BASC appears to be less relevant (with a coefficient of 0 in both models). This is consistent with the conclusions in the article. The effect of Varenicline on abstinence may be moderated by Smoking within 5 minutes of waking up and Nicotine Metabolism Ratio, both of which are smoking-related variables. Non-Hispanic White status, FTCD score at baseline, and Current MDD status may be important predictors for abstinence; specifically, being Non-Hispanic White, having a low FTCD score at baseline, or not currently having MDD may contribute positively to abstinence. Anhedonia might also play a role in abstinence, though its effect is minor (coefficient around -0.02 in both models). Additionally, baseline FTCD score may act as a moderator for BASC.

The original article stated that 'BA was predicted to increase abstinence by addressing anhedonia, especially in individuals with current MDD.' However, we observed no significant patterns related to it in Model 1 or Model 2. To further investigate, we conducted an exploratory analysis. As shown in Figure 2, one finding is that the group without BASC but who achieved abstinence had very low SHAPS scores. This may suggest that a low baseline SHAPS score (indicating lower levels of anhedonia) could contribute to abstinence without the need for BASC, which aligns with the conclusion in the original article. However, to explore the relationship among BASC, anhedonia, and abstinence more thoroughly, we may need longitudinal data on anhedonia after treatment, as this might better reveal the impact of BASC.

Random Forest

Figure 3 shows the importance plot for random forest without adding interaction terms. Foucusing on the mean decrease accuracy measurement, FTCD score at baseline is the most important variable, then follows Varenicline, Cigarettes per day at baseline. Current MDD status is in the fourth important place, but it doesn't have significant difference in importance than others.

Table 3: Coefficients for Lasso Model 1 and Model 2

Variable	Model 1	Model 2
(Intercept)	-1.182	-0.755
Var	0.976	0.111
BA	0	0
age_ps	0	0
sex_ps	0	0
NHW	0.281	0.241
Black	0	0
Hisp	0	0
inc	0	0.017
edu	0	0
ftcd_score	-0.189	-0.256
ftcd.5.mins	0	0
bdi_score_w00	0	0
cpd_ps	0	0
crv_total_pq1	0	0
hedonsum_n_pq1	0	0
hedonsum_y_pq1	0	0
shaps_score_pq1	-0.02	-0.021
otherdiag	0	0
antidepmed	0	0
mde_curr	-0.28	-0.33
NMR	0.369	0
Only.Menthol	0	0
readiness	0	0
BA:ftcd_score	NA	-0.014
BA:ftcd.5.mins	NA	0
BA:cpd_ps	NA	0
BA:crv_total_pq1	NA	0
BA:hedonsum_n_pq1	NA	0
BA:hedonsum_y_pq1	NA	0
BA:NMR	NA	0
BA:Only.Menthol	NA	0
BA:readiness	NA	0
BA:bdi_score_w00	NA	0
BA:shaps_score_pq1	NA	0
BA:otherdiag	NA	0
BA:antidepmed	NA	0
BA:mde_curr	NA	0
Var:ftcd_score	NA	0
Var:ftcd.5.mins	NA	0.707
Var:cpd_ps	NA	0
Var:crv_total_pq1	NA	0
Var:hedonsum_n_pq1	NA	0
Var:hedonsum_y_pq1	NA	0
Var:NMR	NA	1.734
Var:Only.Menthol	NA	0
Var:readiness	NA	0
Var:bdi_score_w00	NA	0
Var:shaps_score_pq1	NA	0
Var:otherdiag	NA	0
Var:antidepmed	NA	0
Var:mde_curr	NA	0

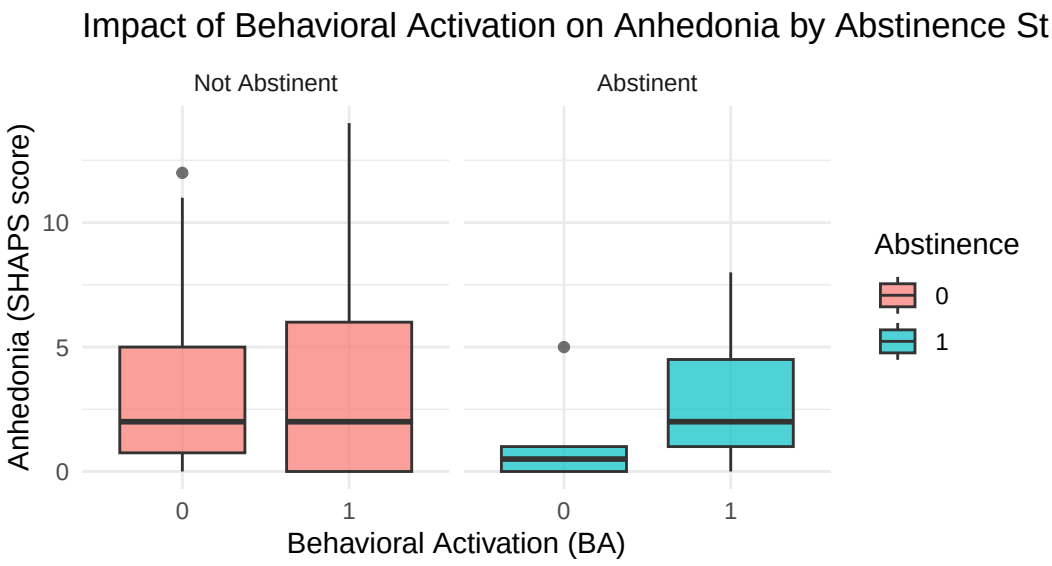


Figure 2: Impact of Behavioral Activation on Anhedonia by Abstinence Status

Figure 4 demonstrates the importance plot after adding interaction terms into the model. FTCD score at baseline is still the most important variable, following by 7 interaction terms of Varenicline with baseline variables, then comes to the interaction term of BASC with FTCD score, and then Varenicline.

The two importance plots reveals the importance of baseline FTCD score, also prove that Varenicline is important but its impact on abstinence is moderated by other baseline variables.

Figure 3 shows the importance plot for the random forest model without interaction terms. Focusing on the mean decrease in accuracy measurement, FTCD score at baseline is the most important variable, followed by Varenicline and cigarettes per day at baseline. Current MDD status ranks fourth in importance, though its increase in importance compared to the rest of variables is not significant.

Figure 4 displays the importance plot after adding interaction terms to the model. FTCD score at baseline remains the most important variable, followed by seven interaction terms involving Varenicline and baseline variables, then the interaction term of BASC with FTCD score, and finally Varenicline.

These two importance plots highlight the significance of baseline FTCD score and indicate that Varenicline is an important predictor, with its impact on abstinence moderated by other baseline variables.

ROC Plots for Model 1 to Model 4

The four ROC plots are shown together in Figure 5. Model 4 achieves the highest AUC value of 0.719 (0.639, 0.799). Models 2 and 1 have similar AUC values, at 0.699 (0.619, 0.780) and 0.691 (0.614, 0.767), respectively. Model 3 has the lowest AUC, at 0.666 (0.577, 0.754).

Discussion

This study aimed to identify baseline variables as potential moderators of behavioral treatment effects on smoking abstinence and to evaluate these variables as predictors of abstinence while controlling for both behavioral treatment and pharmacotherapy. We constructed two logistic LASSO models and two random forest models: one with interaction terms and one without, for each method.

Our results showed that certain baseline factors, such as FTCD score, Non-Hispanic White ethnicity, and Current MDD status, may play significant roles in predicting abstinence, and that Varenicline's effectiveness

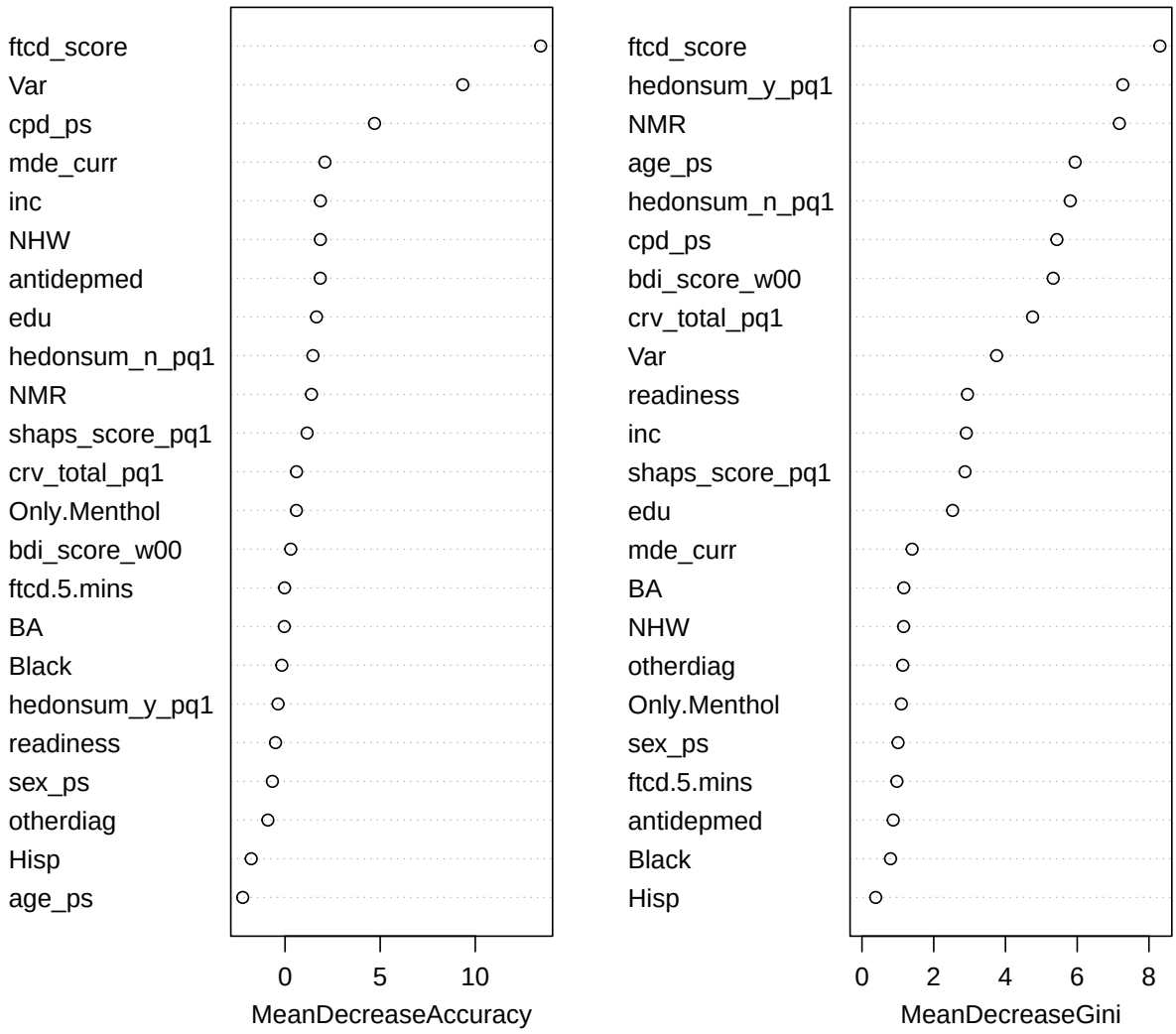


Figure 3: Variable Importance Plot without Interactions

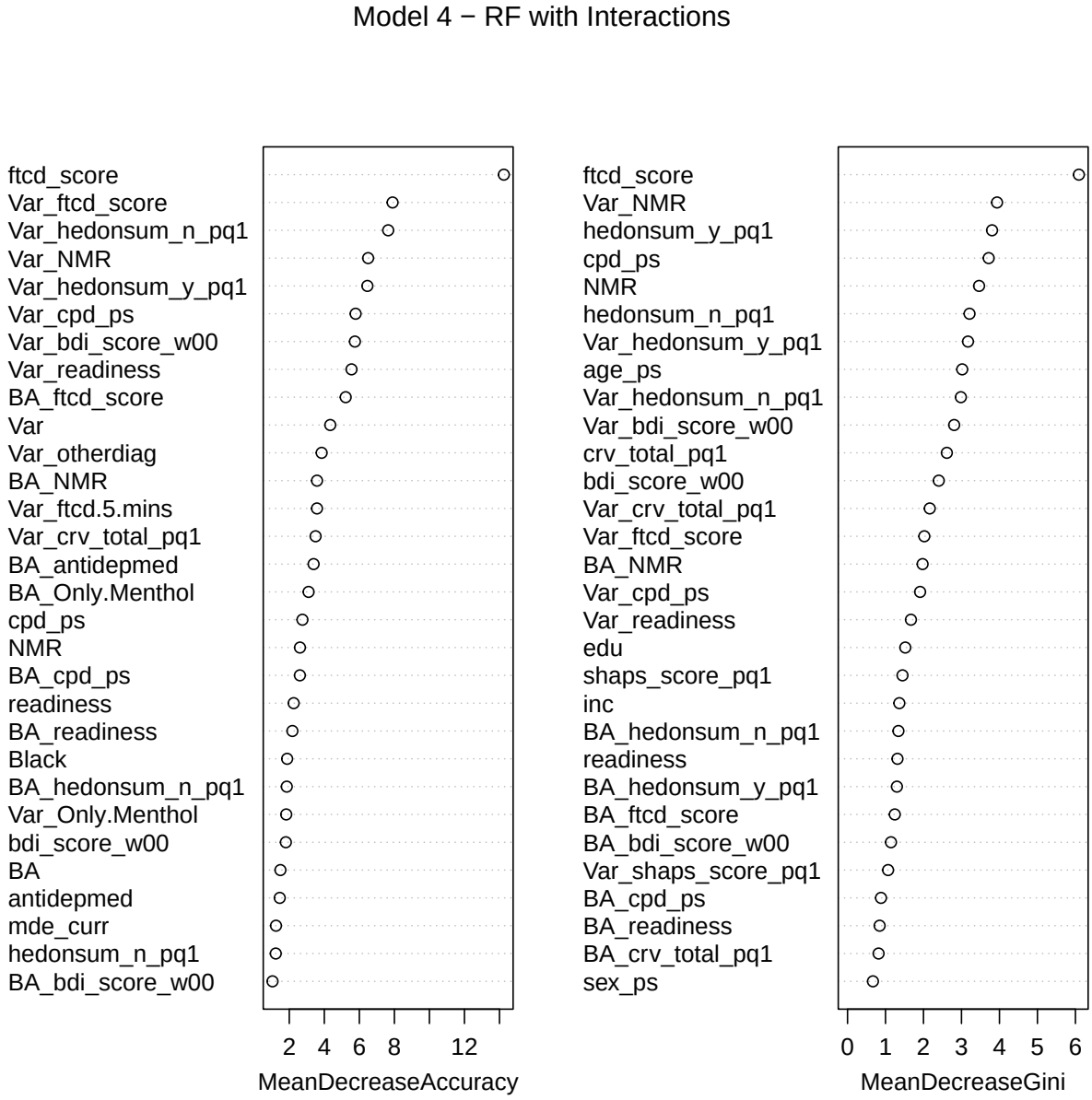


Figure 4: Variable Importance Plot with Interactions

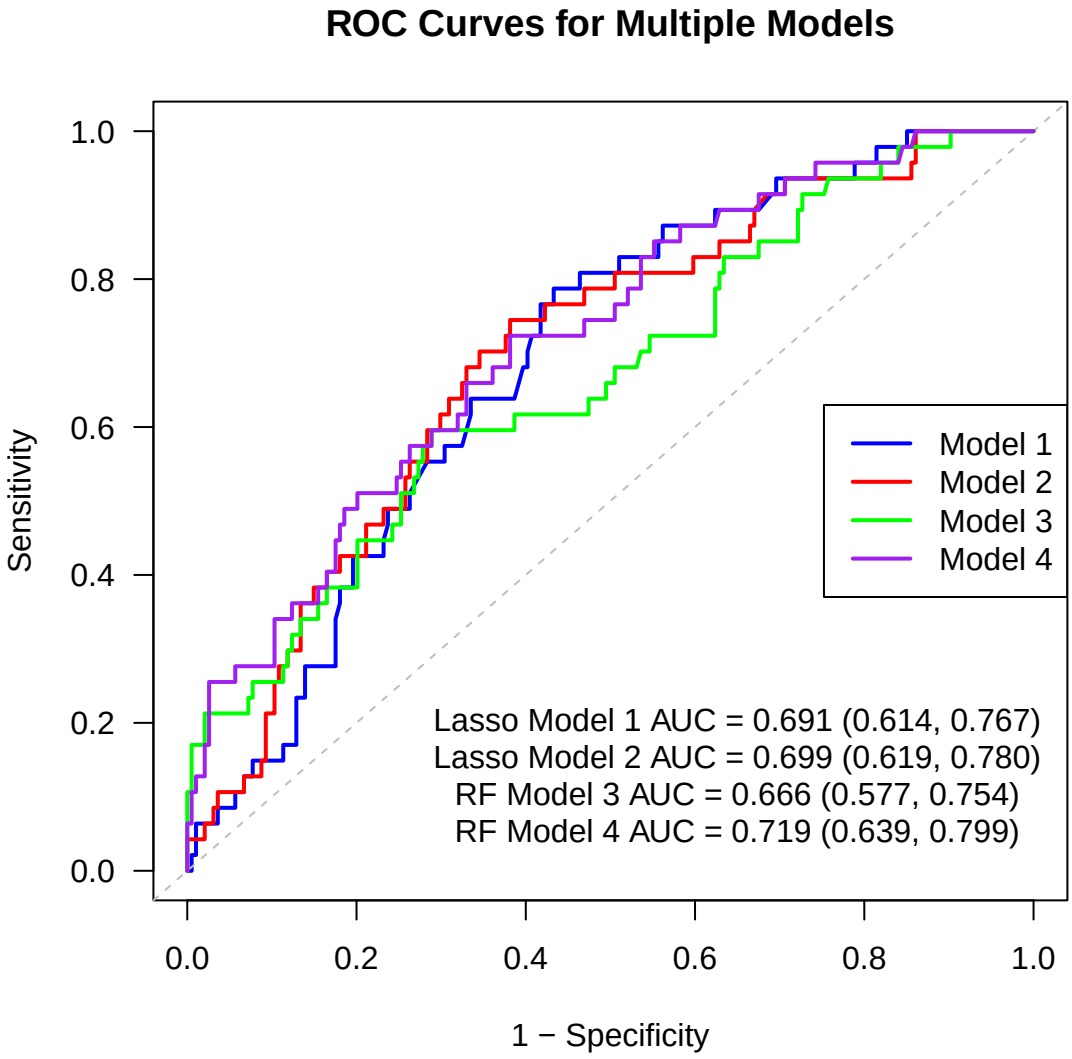


Figure 5: ROC Plot

might be moderated by some smoking-related factors, like smoking within 5 minutes of waking up and nicotine metabolism ratio. These findings align with the original article's conclusions, particularly regarding Varenicline's role as a key predictor of abstinence.

We found that FTCD score may be a moderator of behavioral treatment effects on smoking abstinence. However, the moderating effects of other mental health-related variables were not significant. Additionally, we found that BASC did not contribute as much as expected to abstinence, which aligns with the conclusions in the article.

The article states that BASC might improve abstinence by addressing anhedonia, especially in individuals with current MDD, our models did not reveal significant interactions between BA and anhedonia. This may suggest that the relationship among BA, anhedonia, and abstinence is complex and may require additional data for deeper exploration. Specifically, longitudinal measurements of anhedonia after treatment may be needed to fully understand the role of BASC.

The logistic LASSO regression model may be limited by its linear structure, even with the inclusion of interaction terms. By incorporating interaction terms, the random forest model offers a more flexible structure. Its higher AUC value through out-of-bag prediction demonstrates the improved performance. However, a drawback of using random forest is the difficulty in obtaining precise quantitative estimates of predictor effects.

This project has several limitations. First, the missing data rate is not low enough, considering the sample size of only 300. Multiple imputation could be considered as an alternative approach to address this issue. Second, we treated ordinal variables as continuous variables since each has at least five categories; however, this assumption may not always be reasonable. Additionally, we only considered interactions between BASC or Varenicline and other smoking-related or mental health-related variables. There may be important interactions between other variables that were not included. Our random forest model partially addresses this limitation by capturing complex interactions automatically.

Conclusion

This analysis highlights the importance of FTCD score as a baseline predictor of abstinence. Non-Hispanic white ethnicity and current MDD status may also contribute to predicting abstinence, while controlling for behavioral treatment and pharmacotherapy. The analysis also confirms Varenicline's significance in promoting abstinence among smokers with MDD. However, the anticipated effect of BASC on abstinence, potentially through anhedonia, was not evident, indicating a need for further research. We found that FTCD score may moderate the effects of behavioral treatment on smoking abstinence, while several baseline smoking-related variables may moderate the effects of pharmacotherapy on abstinence.

Appendix

```
library(tidyverse)
library(VIM)
library(gtsummary)
library(kableExtra)
library(corrplot)
library(glmnet)
library(pROC)
library(randomForest)
library(knitr)
# Loading data
dat = read.csv("project2.csv")
dim(dat)
length(unique(dat$id))
colnames(dat)
# Investigate missing data
apply(dat,2,function(x) sum(is.na(x)))
sum(complete.cases(dat))
# Overview of missingness
aggr(dat,ylab = c("Missing data Proportion", "Missing Pattern"))
# Data Cleaning --- Drop all the missingness
dat = dat %>%
  filter(rowSums(is.na(.)) == 0) %>%
  select(-id)
# Summary table for participants in different intervention groups
dat %>%
  mutate(
    group = case_when(
      Var == 0 & BA == 0 ~ "placebo + ST",
      Var == 0 & BA == 1 ~ "placebo + BASC",
      Var == 1 & BA == 0 ~ "varenicline + ST",
      Var == 1 & BA == 1 ~ "varenicline + BASC")) %>%
  select(-c(abst,Var,BA)) %>%
  rename(hedonsum_n = hedonsum_n_pq1,
         hedonsum_y = hedonsum_y_pq1,
         shaps_score = shaps_score_pq1) %>%
  tbl_summary(
    by = group
  ) %>%
  add_p(
    test = list(
      edu ~ "chisq.test",
      readiness ~ "chisq.test"
    )) %>%
  modify_header(label = "**Variables**") %>%
  modify_caption("Participant Characteristics Summary Table") %>%
  as_kable()
# Correlation Plot
# Continuous variables here also include ordinal variables
continuous_vars <- dat[, c("age_ps", "inc", "edu", "ftcd_score", "bdi_score_w00", "cpd_ps", "crv_total_
cor_matrix <- cor(continuous_vars)
corrplot(
```

```
  cor_matrix,
  method = "color",
  col = colorRampPalette(c("blue", "white", "red"))(200),
  type = "lower",
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 0,
  diag = FALSE
)
# Data preparation for lasso
X = model.matrix(abst ~ ., data = dat)[, -1]
Y = factor(dat$abst)
# Lasso --- Model 1 --- No interactions
set.seed(1)
cv.lasso.m1 = cv.glmnet(X, Y, alpha = 1, family="binomial")
best.lambda.m1 = cv.lasso.m1$lambda.min
lasso.m1.coef = coef(cv.lasso.m1, s = best.lambda.m1)
 # Output data frame for lasso coefficient
 lasso.m1.coef.df = data.frame(
   Variable = rownames(lasso.m1.coef),
   Coefficient = as.vector(lasso.m1.coef)
 ) %>%
 mutate(Coefficient = ifelse(Coefficient==0, "0", round(Coefficient, 3)))
# 5-fold-cross-validation
k = 5
set.seed(123)
folds.m1 = sample(1:k, nrow(dat), replace = TRUE)
#auc.5cv.m1 = numeric(k)
predictions.5cv.m1 = numeric(nrow(dat))


for (i in 1:k) {
  train_index = which(folds.m1 != i)
  test_index = which(folds.m1 == i)

  #We have define X and Y above
  #X = model.matrix(abst ~ ., data = dat)[, -1]
  #Y = factor(dat$abst)
  x_train = X[train_index, ]
  y_train = Y[train_index]
  x_test = X[test_index, ]
  y_test = Y[test_index]

  cv_lasso_5cv = cv.glmnet(x_train, y_train, family = "binomial", alpha = 1)
  pred_prob_5cv = predict(cv_lasso_5cv, newx = x_test,
                          s = "lambda.min", type = "response")

  predictions.5cv.m1[test_index] = pred_prob_5cv
}

roc.5cv.m1 = roc(Y, predictions.5cv.m1)
auc.5cv.m1 = auc(roc.5cv.m1)
# LASSO regression with interactions --- data preparation
```

```r
X.m2 = model.matrix(
  abst ~
   # Intervention
   BA + Var +
   # Demographic
   age_ps + sex_ps + NHW + Black + Hisp + inc + edu +
   # Smoking-related variables
   ftcd_score + ftcd.5.mins + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 +
   hedonsum_y_pq1 + NMR + Only.Menthol + readiness +
   # Mental Health related variables
   bdi_score_w00 + shaps_score_pq1 + otherdiag + antidepmed + mde_curr +
   # Interaction terms with BA
   BA:ftcd_score + BA:ftcd.5.mins + BA:cpd_ps + BA:crv_total_pq1 +
   BA:hedonsum_n_pq1 + BA:hedonsum_y_pq1 + BA:NMR + BA:Only.Menthol +
   BA:readiness +
   BA:bdi_score_w00 + BA:shaps_score_pq1 + BA:otherdiag +
   BA:antidepmed + BA:mde_curr +
   # Interaction terms with Var
   Var:ftcd_score + Var:ftcd.5.mins + Var:cpd_ps + Var:crv_total_pq1 +
   Var:hedonsum_n_pq1 + Var:hedonsum_y_pq1 + Var:NMR + Var:Only.Menthol +
   Var:readiness +
   Var:bdi_score_w00 + Var:shaps_score_pq1 + Var:otherdiag +
   Var:antidepmed + Var:mde_curr,
 data = dat)[, -1]
# Lasso --- Model 2 --- all interactions
set.seed(1)
cv.lasso.m2 = cv.glmnet(X.m2, Y, alpha = 1, family="binomial")
best.lambda.m2 = cv.lasso.m2$lambda.min
lasso.m2.coef = coef(cv.lasso.m2, s = best.lambda.m2)

#predictions = predict(cv_lasso, newx = X, s = best_lambda, type="response")
#roc(Y, predict(cv_lasso, newx = X, s = best_lambda, type="response"))

 # Output data frame for lasso coefficient
 lasso.m2.coef.df = data.frame(
   Variable = rownames(lasso.m2.coef),
   Coefficient = as.vector(lasso.m2.coef)
 ) %>%
 mutate(Coefficient = ifelse(Coefficient==0, "0", round(Coefficient, 3)))
# Merge lasso coefficient data frame
merged.lasso.coef.df =
  right_join(lasso.m1.coef.df, lasso.m2.coef.df, by = "Variable") %>%
  rename(
    `Model 1` = Coefficient.x,
    `Model 2` = Coefficient.y
  )

merged.lasso.coef.df %>%
  kbl(caption = "Coefficients for Lasso Model 1 and Model 2") %>%
  kable_styling(
    bootstrap_options = c("striped", "hover", "condensed"),
    full_width = F,
    position = "center"
```

```r
  )
# 5-fold-cross-validation
k = 5
set.seed(123)
folds.m2 = sample(1:k, nrow(dat), replace = TRUE)
#auc.5cv.m1 = numeric(k)
predictions.5cv.m2 = numeric(nrow(dat))


for (i in 1:k) {
  train_index = which(folds.m2 != i)
  test_index = which(folds.m2 == i)

  x_train = X.m2[train_index, ]
  y_train = Y[train_index]
  x_test = X.m2[test_index, ]
  y_test = Y[test_index]

  cv_lasso_5cv = cv.glmnet(x_train, y_train, family = "binomial", alpha = 1)
  pred_prob_5cv = predict(cv_lasso_5cv, newx = x_test,
                          s = "lambda.min", type = "response")

  predictions.5cv.m2[test_index] = pred_prob_5cv
}

roc.5cv.m2 = roc(Y, predictions.5cv.m2)
auc.5cv.m2 = auc(roc.5cv.m2)
# Plot Anhedonia vs. BA by Abstinence Status
dat %>%
  filter(mde_curr == 1) %>%
  ggplot(aes(x = factor(BA), y = shaps_score_pq1, fill = factor(abst))) +
  geom_boxplot(alpha = 0.7, position = position_dodge(0.8)) +
  labs(
    title = "Impact of Behavioral Activation on Anhedonia by Abstinence Status",
    x = "Behavioral Activation (BA)",
    y = "Anhedonia (SHAPS score)",
    fill = "Abstinence"
  ) +
  facet_wrap(~ abst, labeller = labeller(abst = c("0" = "Not Abstinent",
                                                  "1" = "Abstinent"))) +
  theme_minimal()
# Sample size in the above plot
dat %>%
  filter(mde_curr == 1, abst == 0) %>%
  select(BA) %>%
  table()

dat %>%
  filter(mde_curr == 1, abst == 1) %>%
  select(BA) %>%
  table()
# random forest model
set.seed(1)
```

```r
rf.m1 = randomForest(
  x = dat[,
   # Intervention
   c("BA","Var",
   # Demographic
   "age_ps", "sex_ps", "NHW", "Black", "Hisp", "inc", "edu",
   # Smoking-related variables
   "ftcd_score", "ftcd.5.mins", "cpd_ps", "crv_total_pq1", "hedonsum_n_pq1",
   "hedonsum_y_pq1", "NMR", "Only.Menthol", "readiness",
   # Mental Health related variables
   "bdi_score_w00", "shaps_score_pq1", "otherdiag", "antidepmed", "mde_curr")],
  y = Y,
  keep.forest = TRUE,
  importance = TRUE
)

prob.pred.rf.m1 = predict(rf.m1, type="prob")[,2]
roc.rf.m1 = roc(Y, prob.pred.rf.m1)
round(roc.rf.m1$auc,3)
ci.auc(roc.rf.m1)
rf.m1.imp = varImpPlot(rf.m1, main="Model 3 - RF without Interactions")
# random forest model
dat.rf.m2 = dat %>%
  mutate(abst = as.factor(abst),
         BA_ftcd_score = BA*ftcd_score,
         BA_ftcd.5.mins = BA*ftcd.5.mins,
         BA_cpd_ps = BA*cpd_ps,
         BA_crv_total_pq1 = BA*crv_total_pq1,
         BA_hedonsum_n_pq1 = BA*hedonsum_n_pq1,
         BA_hedonsum_y_pq1 = BA*hedonsum_y_pq1,
         BA_NMR = BA*NMR,
         BA_Only.Menthol = BA*Only.Menthol,
         BA_readiness = BA*readiness,
         BA_bdi_score_w00 = BA*bdi_score_w00,
         BA_shaps_score_pq1 = BA*shaps_score_pq1,
         BA_otherdiag = BA*otherdiag,
         BA_antidepmed = BA*antidepmed,
         BA_mde_curr = BA*mde_curr,
         Var_ftcd_score = Var*ftcd_score,
         Var_ftcd.5.mins = Var*ftcd.5.mins,
         Var_cpd_ps = Var*cpd_ps,
         Var_crv_total_pq1 = Var*crv_total_pq1,
         Var_hedonsum_n_pq1 = Var*hedonsum_n_pq1,
         Var_hedonsum_y_pq1 = Var*hedonsum_y_pq1,
         Var_NMR = Var*NMR,
         Var_Only.Menthol = Var*Only.Menthol,
         Var_readiness = Var*readiness,
         Var_bdi_score_w00 = Var*bdi_score_w00,
         Var_shaps_score_pq1 = Var*shaps_score_pq1,
         Var_otherdiag = Var*otherdiag,
         Var_antidepmed = Var*antidepmed,
         Var_mde_curr = Var*mde_curr,)
```

```r
set.seed(1)
rf.m2 = randomForest(
  abst ~
   # Intervention
   BA + Var +
   # Demographic
   age_ps + sex_ps + NHW + Black + Hisp + inc + edu +
   # Smoking-related variables
   ftcd_score + ftcd.5.mins + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 +
   hedonsum_y_pq1 + NMR + Only.Menthol + readiness +
   # Mental Health related variables
   bdi_score_w00 + shaps_score_pq1 + otherdiag + antidepmed + mde_curr +
   # Interaction terms with BA
   BA_ftcd_score + BA_ftcd.5.mins + BA_cpd_ps + BA_crv_total_pq1 +
   BA_hedonsum_n_pq1 + BA_hedonsum_y_pq1 + BA_NMR + BA_Only.Menthol +
   BA_readiness +
   BA_bdi_score_w00 + BA_shaps_score_pq1 + BA_otherdiag +
   BA_antidepmed + BA_mde_curr +
   # Interaction terms with Var
   Var_ftcd_score + Var_ftcd.5.mins + Var_cpd_ps + Var_crv_total_pq1 +
   Var_hedonsum_n_pq1 + Var_hedonsum_y_pq1 + Var_NMR + Var_Only.Menthol +
   Var_readiness +
   Var_bdi_score_w00 + Var_shaps_score_pq1 + Var_otherdiag +
   Var_antidepmed + Var_mde_curr,
  data = dat.rf.m2,
  keep.forest = TRUE,
  importance = TRUE
)

prob.pred.rf.m2 = predict(rf.m2, type="prob")[,2]
roc.rf.m2 = roc(Y, prob.pred.rf.m2)
round(roc.rf.m2$auc,3)
ci.auc(roc.rf.m2)
rf.m2.imp = varImpPlot(rf.m2, main="Model 4 - RF with Interactions")
# ROC Plot
plot(1 - roc.5cv.m1$specificities, roc.5cv.m1$sensitivities,
     xlim = c(0, 1), ylim = c(0, 1),
     las = 1, type = "l", lwd = 2, col = "blue",
     main = "ROC Curves for Multiple Models",
     xlab = "1 - Specificity", ylab = "Sensitivity")

abline(0, 1, lty = 2, col = "gray")
lines(1 - roc.5cv.m2$specificities, roc.5cv.m2$sensitivities,
      col = "red", lwd = 2)
lines(1 - roc.rf.m1$specificities, roc.rf.m1$sensitivities,
      col = "green", lwd = 2)
lines(1 - roc.rf.m2$specificities, roc.rf.m2$sensitivities,
      col = "purple", lwd = 2)


text(0.65,0.2, paste0("Lasso Model 1 AUC = ",
                      formatC(roc.5cv.m1$auc, format="f", digits=3),
                      " (", formatC(ci(roc.5cv.m1)[1], format="f", digits=3),
```

```
                      ", ", formatC(ci(roc.5cv.m1)[3], format="f", digits=3),
                      ")"))
text(0.65,0.15, paste0("Lasso Model 2 AUC = ",
                       formatC(roc.5cv.m2$auc, format="f", digits=3),
                       " (", formatC(ci(roc.5cv.m2)[1], format="f", digits=3),
                      ", ", formatC(ci(roc.5cv.m2)[3], format="f", digits=3),
                      ")"))
text(0.65,0.10, paste0("RF Model 3 AUC = ",
                       formatC(roc.rf.m1$auc, format="f", digits=3),
                       " (", formatC(ci(roc.rf.m1)[1], format="f", digits=3),
                      ", ", formatC(ci(roc.rf.m1)[3], format="f", digits=3),
                      ")"))
text(0.65,0.05, paste0("RF Model 4 AUC = ",
                       formatC(roc.rf.m2$auc, format="f", digits=3),
                       " (", formatC(ci(roc.rf.m2)[1], format="f", digits=3),
                      ", ", formatC(ci(roc.rf.m2)[3], format="f", digits=3),
                      ")"))

legend("right", legend = c("Model 1", "Model 2", "Model 3", "Model 4"),
       col = c("blue", "red", "green", "purple"), lwd = 2)
```