Investigating Moderators of Success of Behavioral Activation for Smoking Cessation

Rachel Yost

2024-11-11

Abstract

Individuals with Major Depressive Disorder (MDD) often struggle with smoking, facing more uncomfortable withdrawal symptoms and a higher probability of relapse that those without MDD. However, they are often are excluded from trials examining possible treatments for smokers. Between 2015 and 2020, researchers at Feinberg School of Medicine conducted a randomized, 2x2 factorial design trial comparing Behavioral Activation Therapy to a standard treatment, and Varenicline to a placebo for smoking cessation in participants with lifetime MDD.

They found that Behavioral Activation Therapy (BA) did not outperform the standard treatment with or without the inclusion of Varenicline. We used the data collected during this study to investigate possible moderators of BA on smoking cessation, as well as the main effects of other covariates. To determine which covariates and interactions were important, we used Lasso for variable selection. We split the data into a training and test set (70/30), and selected lambda using 5-fold cross validation. The full model included the main effects of BA, Varenicline, and all additional covariates, interactions of all variables with BA, and interactions of Varenicline with Black and with an indicator for if the participant is on antidepressants. We evaluated AUC-ROC and model calibration on our test set to check the fit of our model. The model produced an AUC of 0.760 on our training set and 0.715 on our test set. The variables selected by Lasso, other than the treatments-which we chose not to penalize, were Non-Hispanic_White, FTCD score, current MDD, and log(Nicotine Metabolite Ratio). No interactions were selected. The standard treatment slightly outperformed BA, Varenicline outperformed the placebo, non-Hispanic White participants had higher odds of smoking cessation than those who were not non-hispanic White, higher FTCD scores and currently having MDD decreased odds of smoking cessation, and higher log(Nicotine Metabolite Ratio) increased odds of smoking cessation. The results of this analysis suggest that having MDD, a higher dependence on nicotine, being non-Hispanic White, and Nicotine Metabolite Ratio can be predictors of success at quitting smoking, reaffirms that Varenicline is effective, and suggests that BA may be less effective than the standard treatment for smoking cessation.

Introduction

Although smoking rates have recently declined among individuals with depression, those with depression are still more likely to smoke than individuals without it (Han et al., 2022). Additionally, people with mood disorders experience more withdrawal symptoms, with higher probability of withdrawal-related discomfort and relapse (Weinberger et al. 2010). Despite the heightened challenges with smoking cessation experienced by individuals with Major Depressive Disorder (MDD), they have historically been excluded from studies investigating treatment for smokers (Talukder et al. 2021). Varenicline, a medication used to treat nicotine dependence (Burke and Ebbert, 2016), and Behavioral Activation (BA) for smoking cessation are two treatments used in a randomized, placebo-controlled trial conducted by Feinberg School of Medicine. In this study, Behavioral Activation (BA) therapy is compared to standard behavioral treatment (ST), and Varenicline is compared to a placebo in a 2x2 factorial design trial on patients with a lifetime diagnosis of MDD.

The results from this trial showed that BA did not outperform the standard treatment, with or without the inclusion of Varenicline (Hitsman et al. 2023). Using the data collected from this trial, we wanted to investigate how other covariates might predict smoking cessation, and if they moderate the effects of BA on smoking cessation.

A previous study found that participants who had a higher level of anhedonia and received BA were abstinent for significantly more days than those who received the standard treatment. They also found that BA was more effective at improving abstinence rates when individuals had less depressive symptoms. (Martínez-Vispo, 2020). Due to these previous findings, we hypothesized that BA would have differential effects on smoking cessation based on the participants level of anhedonia and whether or not the individual currently has MDD or had it in the past. To determine which interactions and covariates were relevant predictors of smoking cessation, we used Lasso for variable selection.

Methods

Data Description

The data was provided by the Feinberg School of Medicine at Northwestern University, and was collected between 2015 and 2020. Participants smoked at least one cigarette per day, were diagnosed with MDD in their lifetime, and had interest in quitting smoking. Participants were randomly assigned to receive BA or standard treatment, and Varenicline or placebo. The dataset contained 300 observations and included information on treatment assignment and 21 other covariates describing the participants age, sex, race, and information about their health and smoking habits. The full list of variables can be seen in Table 1.

Preprocessing and Missing Data

We split the data into a training and test set (70/30), randomly sampling within each of the four treatment options. This resulted in 211 training observations and 89 test observations. Data was missing for the variables for income, FTCD score, cigarette reward value at baseline, anhedonia, nicotine metabolism ratio, an indicator for smoking only menthol cigarettes, and baseline readiness to quit smoking. We performed multiple imputation on each training/test set separately using the mice package (van Buuren and Groothuis-Oudshoorn, 2011), with 5 imputed datasets for each test/train set. Number of missing values per variable can be seen in Table 1.

Binary variables were transformed from numeric to factors. A description of the variables stratified by treatment group is shown in Table 1. Based on Pearson chi-squared test, Kruskal-Wallis rank sum test, and Fisher's exact test, there were no significant differences in covariates between treatment groups, except for whether or not the participants were taking antidepressant medication. For those in the Varenicline only group, 19% were on antidepressant medication, compared to 41% of participants in the Behavioral Activation group.

Table 1: Data Characteristics

Characteristic	**Placebo & ST** N = 68	**Placebo & BA** N = 68	**Varenicline & ST** N = 81	**Varenicline & BA** N = 83	**p-value**
Age	51 (45, 58)	54 (42, 61)	52 (41, 59)	53 (40, 60)	0.7
Sex					>0.9
1	29 (43%)	30 (44%)	37 (46%)	39 (47%)	
2	39 (57%)	38 (56%)	44 (54%)	44 (53%)	
Non-Hispanic White					0.5
0	46 (68%)	44 (65%)	56 (69%)	49 (59%)	
1	22 (32%)	24 (35%)	25 (31%)	34 (41%)	
Black					0.3
0	28 (41%)	31 (46%)	38 (47%)	46 (55%)	
1	40 (59%)	37 (54%)	43 (53%)	37 (45%)	
Hispanic					>0.9

Table 1: Data Characteristics *(continued)*

Characteristic	**Placebo & ST** N = 68	**Placebo & BA** N = 68	**Varenicline & ST** N = 81	**Varenicline & BA** N = 83	**p-value**
0	64 (94%)	63 (93%)	76 (94%)	79 (95%)	
1	4 (5.9%)	5 (7.4%)	5 (6.2%)	4 (4.8%)	
Income	2.00 (1.00, 3.00)	2.00 (1.00, 4.00)	2.00 (1.00, 3.00)	2.00 (1.00, 4.00)	>0.9
Unknown	0	1	1	1	
Education	4.00 (4.00, 4.50)	4.00 (3.00, 5.00)	4.00 (3.00, 5.00)	4.00 (3.00, 5.00)	0.4
FTCD score at baseline	6.00 (4.00, 7.00)	5.00 (4.00, 7.00)	5.00 (4.00, 7.00)	5.00 (4.00, 7.00)	0.7
Unknown	1	0	0	0	
Smoking with 5 mins of waking up					0.5
0	33 (49%)	36 (53%)	43 (53%)	50 (60%)	
1	35 (51%)	32 (47%)	38 (47%)	33 (40%)	
BDI score at baseline	18 (12, 25)	18 (9, 27)	18 (11, 27)	18 (10, 25)	>0.9
Cigarettes per day at baseline	13 (10, 20)	15 (10, 20)	15 (10, 20)	15 (10, 20)	>0.9
Cigarette reward value at baseline	7.0 (4.5, 9.0)	7.0 (5.0, 10.0)	7.0 (5.0, 9.0)	8.0 (4.5, 10.0)	>0.9
Unknown	8	1	6	3	
PES – Substitute reinforcers	14 (9, 27)	21 (10, 31)	20 (9, 35)	20 (9, 32)	0.6
PES – Complementary reinforcers	25 (12, 38)	23 (14, 34)	21 (13, 34)	17 (11, 31)	0.3
Anhedonia	1.00 (0.00, 5.00)	0.00 (0.00, 3.00)	1.00 (0.00, 3.00)	1.00 (0.00, 4.00)	0.8
Unknown	1	2	0	0	
Other lifetime DSM-5 diagnosis					0.2
0	40 (59%)	33 (49%)	41 (51%)	53 (64%)	
1	28 (41%)	35 (51%)	40 (49%)	30 (36%)	
Taking antidepressants					0.013
0	53 (78%)	40 (59%)	66 (81%)	59 (71%)	
1	15 (22%)	28 (41%)	15 (19%)	24 (29%)	
Current vs past MDD					0.7
0	37 (54%)	36 (53%)	37 (46%)	43 (52%)	
1	31 (46%)	32 (47%)	44 (54%)	40 (48%)	
Nicotine Metabolism Ratio	0.32 (0.20, 0.43)	0.32 (0.23, 0.46)	0.29 (0.20, 0.51)	0.33 (0.22, 0.50)	>0.9
Unknown	2	7	9	3	
Exclusive Mentholated Cigarette User					0.9
0	24 (36%)	28 (41%)	34 (42%)	34 (41%)	
1	43 (64%)	40 (59%)	47 (58%)	48 (59%)	
Unknown	1	0	0	1	
Readiness	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)	7.00 (6.00, 8.00)	0.6
Unknown	4	4	4	5	

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

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

We also examined variable distributions, and checked to see if any variables were skewed. We found that age, income, education, BDI score at baseline, cigarettes per day, cigarette reward value, both Pleasant Events Schedule variables, anhedonia, and nicotine metabolism ratio had skewed distributions. After visually evaluating square root, squared, and log transformations, we decided to include age squared (age^2), the log transformed nicotine metabolism ratio, and square root transformations of BDI score, cigarettes per day, and the Pleasant Events Schedule-complimentary reinforcers in one of our possible models. We performed our analyses on the model with transformations and the model without and compared AUC and model calibration.

To inform how we select the variables to include in our full model, and to check for collinearity among the variables, we examined the spearman correlations using the data without imputation (Figure 1). Notable correlations include BDI score with whether or not the participant currently has MDD (r = .60), the indicator variable "Black" with smoking only menthol cigarettes (r=.46), BDI score with anhedonia (r = .45), income with education (r = .43), FTCD score at baseline with smoking within 5 minutes of waking up (r = .64), and cigarettes per day with FTCD score (r = .55). Varenicline had a correlation of .25 with abstinence, and behavioral activation was not correlated with abstinence (r = -.02).

We also checked interactions of other covariates with BA, to gain an understanding of what we might expect

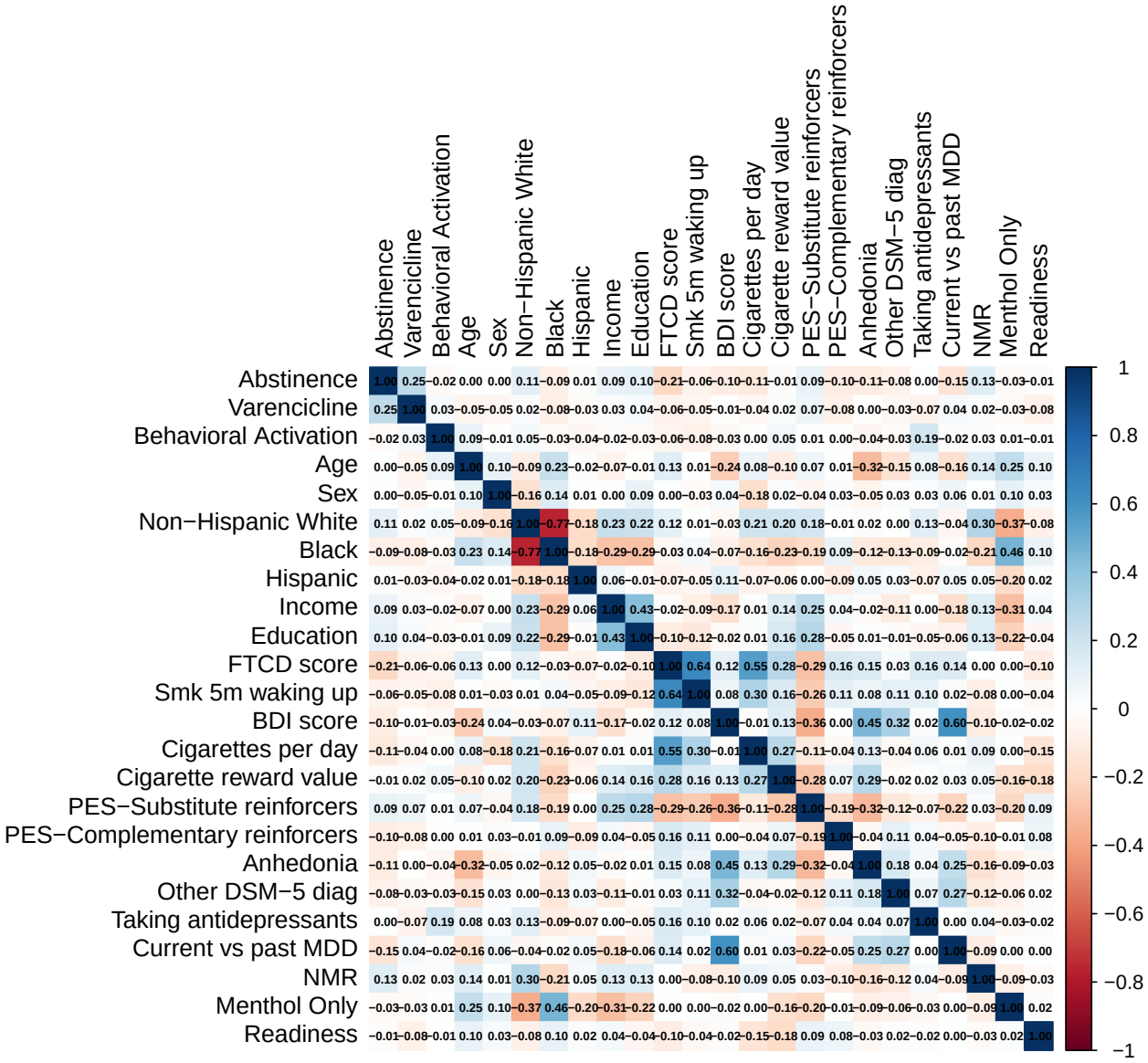

Figure 1: Correlation Plot For All Variables

to see from our model. Additionally, we checked interactions with Varenicline, to see if any other interactions should be included, aside from just the ones with BA. Most interactions that we investigated did not appear important (Figure 2).

We decided to include the interaction of Varenicline with Black in our full model, since the proportion abstinent from smoking was significantly higher in those that received Varenicline among the black population, but there was not a significant increase in abstinence rate for non black participants (Figure 2a). We also chose to include the interaction of antidepressants with Varenicline (Figure 2h), in case there were any interactions between the medications.

Variable Selection via Lasso

We chose to use Lasso since we have many possible variables under consideration, and we would like to select only some of the variables to include in our model, so that our result is interpretable and not overfitted. Lasso works by minimizing the sum of the squared residuals plus a penalty term: $\lambda \sum_{j=1}^{p} |\beta_j|$. This ℓ_1 penalty allows the coefficients to be shrunk to exactly zero, in contrast to the penalty used by ridge regression, $\lambda \sum_{j=1}^{p} \beta_j^2$, which shrinks the coefficients towards zero, but they do not reach zero. Best subset selection is another method to consider for variable selection. Best subset selection works by testing each possible combination of p predictors and identifying the best subset of variables based on certain criteria such as cross-validated predicted error, AIC, BIC, or adjusted R^2. However, since the number of possible models increases rapidly as the number of possible variables increases, it can become computationally difficult with large p (Gareth et al., 2013). Since we have many variables and interactions to test, we chose to use Lasso instead of best subset selection.

Within each of the 5 imputed data sets we fit the full model including BA, Varenicline and the other covariates as main effects, and interactions with all covariates and Varencline with BA. Additionally, we included interactions between Varenicline and Black and Varenicline and the indicator for if the participant is taking antidepressants. We chose not to penalize BA and Varenicline, to ensure they would be kept in the model. We then used the glmnet package (Friedman et al., 2010), to find the best value of lambda via 5-fold cross validation on the training data, based on the minimum mean cross validated error. We then fit the Lasso model on the training set using the selected lambda and obtained parameter estimates for the returned variables. We then obtained predictions for abstinence on the test set and AUC estimates. After repeating these steps for the 5 imputed data sets, we averaged AUC estimates and model parameter estimates. Using the pooled parameter estimates based on Rubin's rules, we predicted abstinence on the stacked imputed data and plotted ROC curves and calibration. This process was repeated again, instead using the variable transformations described previously.

Results

The model without transformations selected the following variables (other than Varenicline and BA, since they were not penalized in the model) : non-Hispanic White, FTCD score, current MDD, and log(Nicotine Metabolite Ratio) (NMR). The model with transformations selected non-Hispanic White, FTCD score, and current MDD, and no interactions.

The AUC obtained by the model with transformations was 0.760 on the training set, and 0.715 on the test set. The ROC curves shown in figure 3 indicate that the model performed better on the training data than the test data, which is to be expected. The calibration plots for the model fit with the transformed data show that the model is calibrated fairly well, since the 95% confidence interval for the loess fit overlaps with the ideal fit for the test data. The model performs similarly well for the training data. (Figure 4).

On the data without transformations, the AUC was 0.760 on the training set and 0.714 on the test set (Figure 5). The calibration plots fit on the data without transformations also show good calibration for the test data and training data (Figure 6).

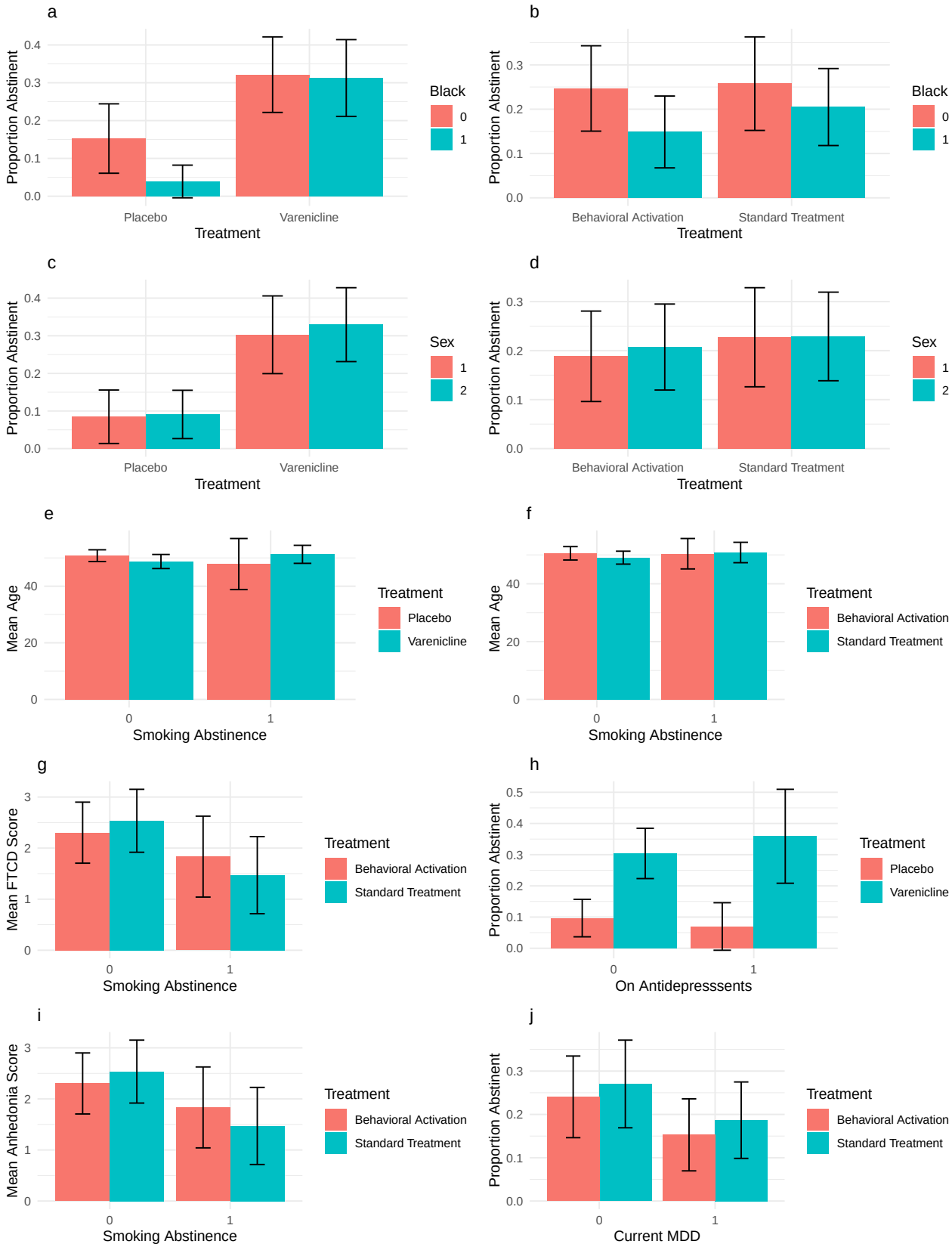

Figure 2: Interactions with Covariates and Treatments. Error bars represent 95% confidence intervals

Since the model with transformations performed better based on AUC, we chose the model with transformations as our best model. An AUC of 0.715 indicates that the model had moderate discrimination ability.

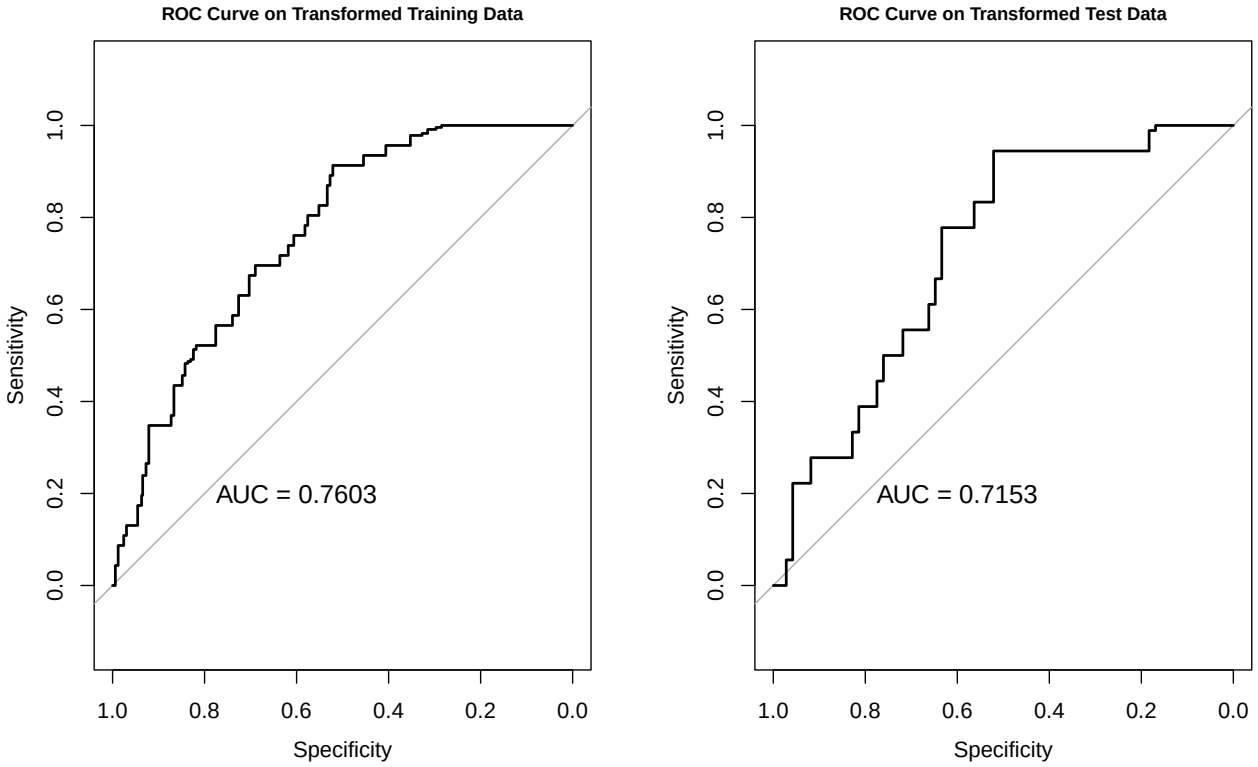

Figure 3: ROC Curves for Transformed Data Model

The variables selected by Lasso using the model fit with transformations are shown in Table 2.

Table 2: Variables Selected By Lasso

	OR
(Intercept)	1.23955
Behavioral Activation	0.93985
Varenicline	1.25664
Non-Hispanic White	1.04621
FTCD Score	0.98418
Current MDD	0.95263
log(NMR)	1.00269

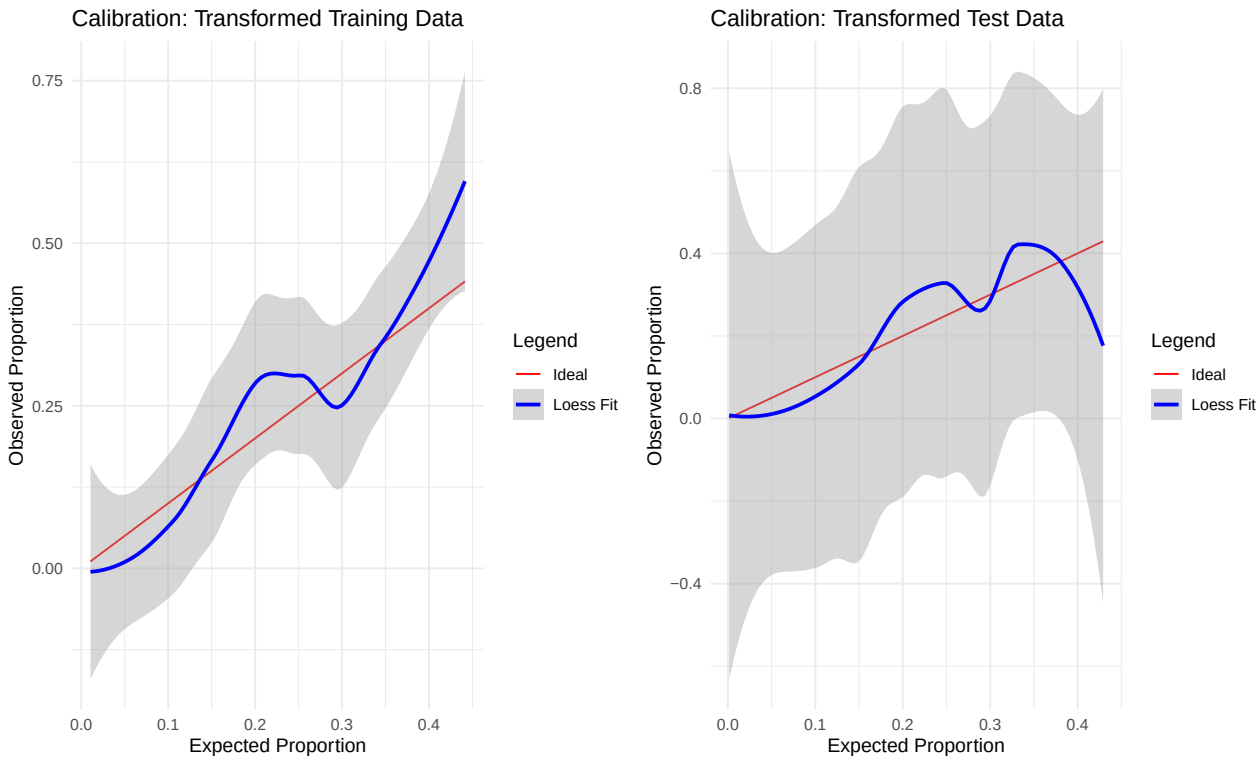

Figure 4: Calibration Plots for Transformed Data Model

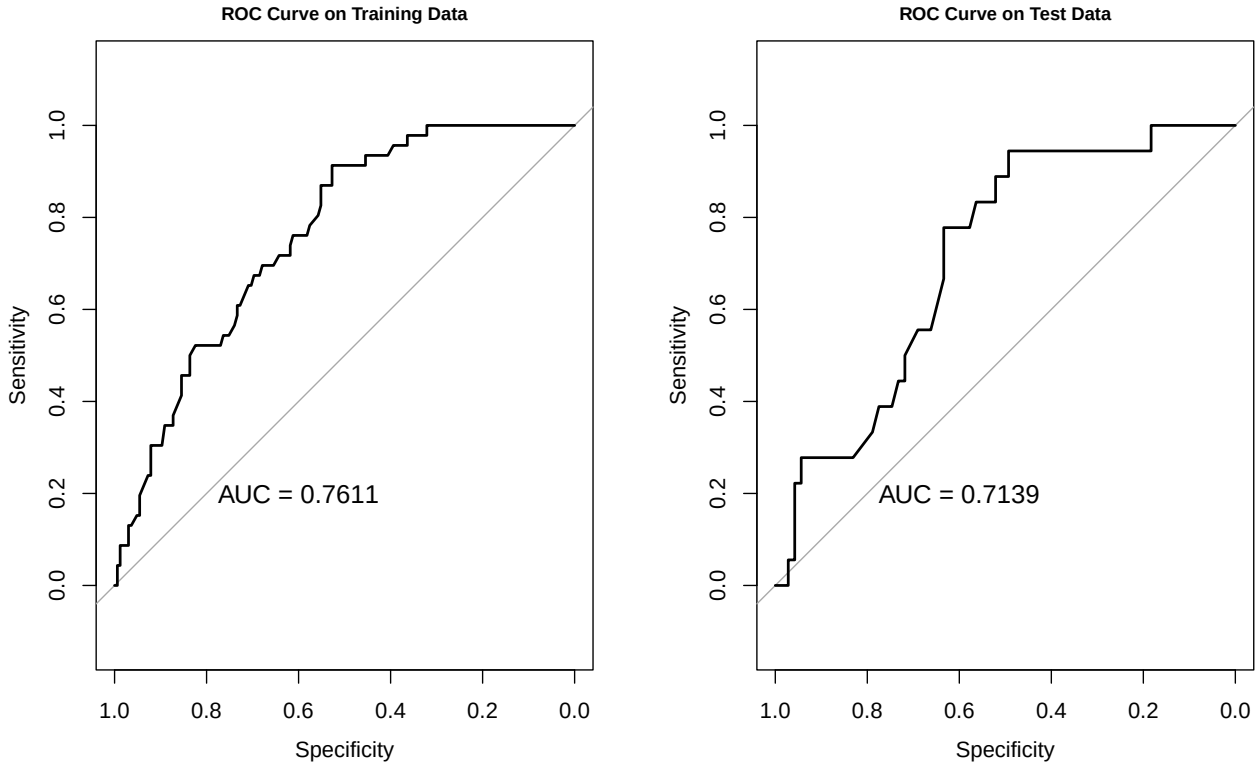

Figure 5: ROC Curves for Data Modeled Without Transformations

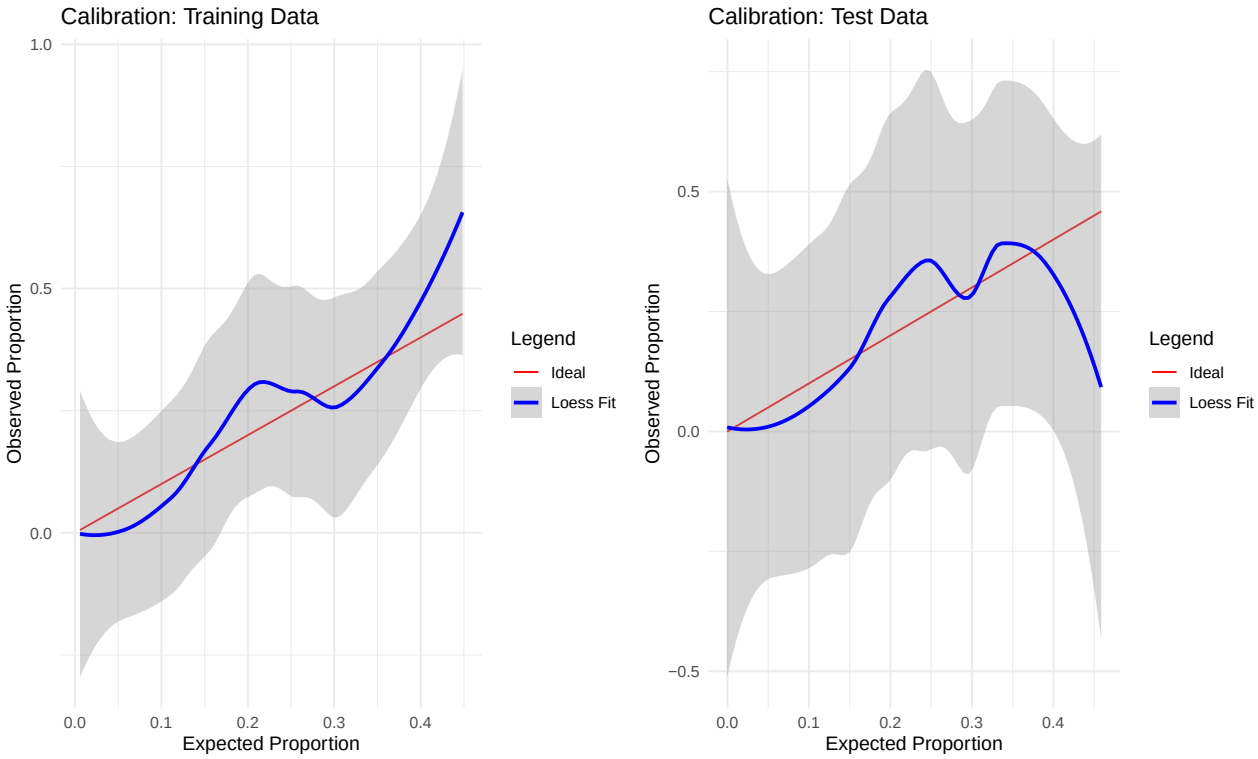

Figure 6: ROC Curves for Data Modeled Without Transformations

Discussion and Conclusion

Based on these results, we reaffirm that Varenicline is effective at improving the odds of smoking cessation (OR = 1.256; Table 1.). We also see that those who had BA treatment had lower odds of smoking cessation than those who received the standard treatment. This matches the results of the study our data is sourced from, which found that BA did not outperform the standard treatment (Hitsman et al. 2023). Our hypothesis that current MDD would decrease odds of smoking cessation was correct (OR = .953), although we did not see an interaction with BA. We also found that being non-Hispanic white was associated with higher odds of smoking cessation (OR = 1.05), which matches with the results of the correlation plot in Figure 1, where non-Hispanic white had a correlation of r = 0.11 with abstinence. Odds also increased when log(NMR) increases (OR = 1.003), matching the correlation of NMR with abstinence shown in Figure 1, r = 0.13. This seems surprising, since previous studies have shown that lower NMR is associated with more likely to remain abstinent (Lerman et al. 2006, Kaufmann et al. 2015). However, in our model, the OR is very near 1, suggesting that the effect may not be very strong. Additionally, since Lasso does not produce standard errors, we cannot calculate p-values to determine if any of our effects are significant or not. A possible solution to this could have been to fit the model on bootstrapped samples, so that we could obtain standard errors for the output coefficients.

Some other limitations of our study are that the sample size was somewhat small (n=300). Additionally, we did not have an external validation set to use, so our measures of discrimination and calibration may be overfit to our dataset. Our data split further reduced the amount of information used to build our model.

Overall, we found that being Non-Hispanic White, FTCD score, having current MDD, and log(NMR) are possible predictors for smoking abstinence when considering the effects of Behavioral Activation and Varenicline. We confirm that Varenicline is an effective treatment and that Behavioral Activation may not be very effective. Future research should be performed to confirm that the other predictors we found are significant, since we were unable to obtain p-values. Additionally, the effect of NMR on smoking cessation should be further investigated, since our results do not match those of previous studies.

References

Burke, M. V., Hays, J. T., & Ebbert, J. O. (2016). Varenicline for smoking cessation: a narrative review of efficacy, adverse effects, use in at-risk populations, and adherence. *Patient preference and adherence, 10*, 435–441. https://doi.org/10.2147/PPA.S83469

Friedman J, Tibshirani R, Hastie T (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software, 33(1), 1-22. doi:10.18637/jss.v033.i01 https://doi.org/10.18637/jss.v033.i01.

Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). "An Introduction to Statistical Learn- ing with Application in R."

Han, B., Volkow, N. D., Blanco, C., Tipperman, D., Einstein, E. B., & Compton, W. M. (2022). Trends in Prevalence of Cigarette Smoking Among US Adults With Major Depression or Substance Use Disorders, 2006-2019. *JAMA*, *327*(16), 1566–1576. https://doi.org/10.1001/jama.2022.4790

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., Khan, S. S., Fox, E. N., Bauer, A. M., Wileyto, E. P., Bastian, J., & Schnoll, R. A. (2023). Efficacy and safety of combination behavioral activation for smoking cessation and varenicline for treating tobacco dependence among individuals with current or past major depressive disorder: A 2×2 factorial, randomized, placebo-controlled trial. *Addiction (Abingdon, England)*, *118*(9), 1710–1725. https://doi.org/10.1111/add.16209

Kaufmann A, Hitsman B, Goelz P, et al. Rate of nicotine metabolism and smoking cessation outcomes in a community-based sample of treatment-seeking smokers. Addictive Behaviors. 2015;51:93–9.

Lerman C, Tyndale R, Patterson F, et al. Nicotine metabolite ratio predicts efficacy of transdermal nicotine for smoking cessation. Clin Pharmacol Ther. 2006;79(6):600–608.

Martínez-Vispo C, López-Durán A, Senra C, Rodríguez-Cano R, Fernández Del Río E, Becoña E. Behavioral activation and smoking cessation outcomes: The role of depressive symptoms. Addict Behav. 2020 Mar;102:106183. doi: 10.1016/j.addbeh.2019.106183. Epub 2019 Oct 19. PMID: 31809878.

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. DOI 10.18637/jss.v045.i03.

Talukder, S. R., Lappin, J. M., Boland, V., McRobbie, H., & Courtney, R. J. (2021). Inequity in smoking cessation clinical trials testing pharmacotherapies: Exclusion of smokers with mental health disorders. *Tobacco Control*, *32*(4), 489–496. https://doi.org/10.1136/tobaccocontrol-2021-056843

Weinberger, A. H., Desai, R. A., & McKee, S. A. (2010). Nicotine withdrawal in U.S. smokers with current mood, anxiety, alcohol use, and substance use disorders. *Drug and alcohol dependence*, *108*(1-2), 7–12. https://doi.org/10.1016/j.drugalcdep.2009.11.004

Code Appendix

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

library(tidyverse)
library(gtsummary) #summary table
library(kableExtra)
library(corrplot)
library(gridExtra)
library(glmnet)
library(mice)
library(glmnet)
```

```r
library(pROC)
data <- read_csv("/Users/rachelyost/Downloads/project2.csv")

#save the data while everything is numeric for use later on
data_numeric <- read_csv("/Users/rachelyost/Downloads/project2.csv")
#summarize missing data

#examine missingness in the main data
sum(is.na(data))

#see which columns are missing data and how much
colSums(is.na(data))[colSums(is.na(data)) > 0]

#see if any people have missing values for multiple columns
rowSums(is.na(data))

#missing some income data, ftcd score, crv_total_pq1, anhedonia, NMR,
#only menthol, and readiness
#need to make a summary table for each group
data <- data %>%
  mutate(group = case_when(Var == 0 & BA == 0 ~ "Placebo and Standard Treatment",
                           Var== 0 & BA == 1 ~ "Placebo and BA",
                           Var== 1 & BA == 0 ~ "Varenicline and Standard Treatment",
                           Var== 1 & BA == 1 ~ "Varenicline and BA"))

#make sure everything is the correct data type
numeric_vars <- c("abst","age_ps", "inc", "edu", "ftcd_score", "bdi_score_w00",
                  "cpd_ps", "crv_total_pq1", "hedonsum_n_pq1", "hedonsum_y_pq1",
                  "shaps_score_pq1", "NMR", "readiness")

factor_vars <- colnames(data[,(!colnames(data) %in% numeric_vars)])

#if variable is in factor_vars, transform to a factor
for (vars in factor_vars){
  data[[vars]] <- factor(data[[vars]])
}

#set seed for replication
set.seed(100)

#sample for the training and test sets within each treatment group
placebo_ids <- data[data$group == "Placebo and Standard Treatment",]$id
training_ids <- sample(placebo_ids, round(length(placebo_ids)*.70))

var_ids <- data[data$group == "Varenicline and Standard Treatment",]$id
training_ids <- c(training_ids, sample(var_ids, round(length(var_ids)*.70)))

BA_ids <- data[data$group == "Placebo and BA",]$id
training_ids <- c(training_ids, sample(BA_ids, round(length(BA_ids)*.70)))

both_ids <- data[data$group == "Varenicline and BA",]$id

training_ids <- c(training_ids, sample(both_ids, round(length(both_ids)*.70)))
```

```r
#get the rows that have ids in the training set
training_rows <- which(data$id %in% training_ids)

data <- data %>% select(!group)
#don't use ID column for prediction
predmat <- make.predictorMatrix(data)
predmat[,"id"] <- 0

training <- data[training_rows,]

test <- data[-training_rows,]

#impute the missing data 5 times
miced_data_train <- mice(training, m=5, predictorMatrix = predmat)
miced_data_test <- mice(test, m=5, predictorMatrix = predmat)
#remove variables that we don't want in the table
data_for_tbl <- data %>% mutate(group = case_when(Var == 0 & BA == 0 ~ "Placebo & ST",
                           Var== 0 & BA == 1 ~ "Placebo & BA",
                           Var== 1 & BA == 0 ~ "Varenicline & ST",
                           Var== 1 & BA == 1 ~ "Varenicline & BA")) %>% select(-id, -abst, -Var, -BA)

#change levels for "group"
data_for_tbl$group <- factor(data_for_tbl$group,
                             levels = c("Placebo & ST",
                                        "Placebo & BA",
                                        "Varenicline & ST",
                                        "Varenicline & BA"))


colnames(data_for_tbl) = c("Age", "Sex", "Non-Hispanic White",
                           "Black", "Hispanic", "Income", "Education",
                           "FTCD score at baseline", "Smoking with 5 mins of waking up",
                           "BDI score at baseline", "Cigarettes per day at baseline",
                           "Cigarette reward value at baseline",
                           "PES - Substitute reinforcers",
                           "PES - Complementary reinforcers",
                           "Anhedonia", "Other lifetime DSM-5 diagnosis",
                           "Taking antidepressants",
                           "Current vs past MDD", "Nicotine Metabolism Ratio",
                           "Exclusive Mentholated Cigarette User",
                           "Readiness", "group")

#create table 1 stratified by group, make sure eduction and readiness are
#treated as continuous
tbl_summary(data_for_tbl,
            by = group,
            type = list(Income ~ "continuous",
            Education ~ "continuous",
            Readiness ~ "continuous"
            )) %>%
  add_p() %>%
   as_kable_extra(booktabs = TRUE,
```

```
                 caption = "Data Characteristics",
                longtable = TRUE, linesep = "") %>%
   kable_styling(font_size = 8,
                latex_options = c("repeat_header", "HOLD_position"))
#for each column in the data, if the column is numeric, create a histogram
for (col in colnames(data)){
  hist(as.numeric(data[[col]]), main = col)
}

#based on the histograms:
#consider transformations of age_ps, inc, edu, bdi_score_w00, cpd_ps, cvr_total_pq1,
#hedonsum_n_pq1, hedonsum_y_pq1, shaps_score_pq1, NMR

#function to plot each of the transformations next to each other
plots <- function(var){
  p1 <- ggplot(data) + geom_histogram(aes(x= data[[var]]), bins=30) +
  labs(main = var)
  p2 <- ggplot(data) + geom_histogram(aes(x= log(data[[var]] + 1)), bins=30) +
  labs(main = var)
  p3 <- ggplot(data) + geom_histogram(aes(x= sqrt(data[[var]])), bins=30) +
  labs(main = var)
  p4 <- ggplot(data) + geom_histogram(aes(x= (data[[var]])^2), bins=30) +
  labs(main = var)

  grid.arrange(p1, p2, p3, p4)
}

plots("age_ps") #^2
plots("inc")
#plots("edu")
plots("bdi_score_w00") #sqrt
plots("cpd_ps") #sqrt
#plots("cvr_total_pq1")
#plots("hedonsum_n_pq1")
plots("hedonsum_y_pq1") #sqrt
plots("shaps_score_pq1")
plots("NMR") #log

#remove id variable
data_for_mat <- data_numeric %>% select(-id)

#rename columns for the plot
colnames(data_for_mat) <-  c("Abstinence", "Varencicline", "Behavioral Activation", "Age", "Sex", "No
                           "Black", "Hispanic", "Income", "Education",
                           "FTCD score", "Smk 5m waking up",
                           "BDI score", "Cigarettes per day",
                           "Cigarette reward value",
                           "PES-Substitute reinforcers",
                           "PES-Complementary reinforcers",
                           "Anhedonia", "Other DSM-5 diag",
                           "Taking antidepressants",
                           "Current vs past MDD", "NMR",
                           "Menthol Only",
```

```
                          "Readiness")

#correlation matrix
corrplot(cor(na.omit(data_for_mat), method = "spearman"), method = 'color', addCoef.col = 'black', tl
         number.cex = .5, tl.cex = 1)
#what interactions might make sense?

data_numeric %>% group_by(sex_ps, Var) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) #probably not

data_numeric %>% group_by(sex_ps, BA) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) #probably not

data_numeric %>% group_by(Black, Var) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) #yes
data_numeric %>% group_by(Black, BA) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) #yes

data_numeric %>% group_by(Only.Menthol, Var) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) #yes

data_numeric %>% group_by(Hisp, Var) %>%  #not really enough hispanic to say
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n)
data_numeric %>% group_by(Hisp, BA) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n)

data_numeric %>% group_by(abst, Var) %>%
  summarize(mean_age = mean(age_ps)) #yes?
data_numeric %>% group_by(abst, BA) %>%
  summarize(mean_age = mean(age_ps)) #nah

data_numeric %>% group_by(abst, BA) %>%
  summarize(mean_score = mean(shaps_score_pq1, na.rm=TRUE)) #not sure

data_numeric %>% group_by(antidepmed, Var) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) #yes?

data_numeric %>% group_by(abst, BA) %>%
  summarize(mean_readiness = mean(readiness, na.rm=TRUE)) #can't tell, probably not
#create plots for all interactions
black_var_plot_df <- data_numeric %>% group_by(Black, Var) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n)  %>%
  mutate(Var = ifelse(Var == 1, "Varenicline", "Placebo"),
         SE = sqrt((prop * (1-prop)/n)))

black_BA_plot_df <- data_numeric %>% group_by(Black, BA) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n)%>%
  mutate(BA = ifelse(BA == 1, "Behavioral Activation", "Standard Treatment"),
         SE = sqrt((prop * (1-prop)/n)))

sex_var_plot_df <- data_numeric %>% group_by(sex_ps, Var) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) %>%
  mutate(Var = ifelse(Var == 1, "Varenicline", "Placebo"),
```

```
         SE = sqrt((prop * (1-prop)/n)))

sex_BA_plot_df <- data_numeric %>% group_by(sex_ps, BA) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) %>%
  mutate(BA = ifelse(BA == 1, "Behavioral Activation", "Standard Treatment"),
         SE = sqrt((prop * (1-prop)/n)))

age_var_plot_df <- data_numeric %>% group_by(abst, Var) %>%
  summarize(mean_age = mean(age_ps), sd_age = sd(age_ps), n = n()) %>%
  mutate(SE = sd_age/sqrt(n), Var = ifelse(Var == 1, "Varenicline", "Placebo"))

age_BA_plot_df <- data_numeric %>% group_by(abst, BA) %>%
  summarize(mean_age = mean(age_ps), sd_age = sd(age_ps), n = n()) %>% mutate(SE = sd_age/sqrt(n), BA

shaps_BA_plot_df <- data_numeric %>% group_by(abst, BA) %>%
  summarize(mean_score = mean(shaps_score_pq1, na.rm=TRUE),
            sd_score = sd(shaps_score_pq1, na.rm=TRUE), n = n()) %>%
  mutate(SE = sd_score/sqrt(n), BA = ifelse(BA == 1, "Behavioral Activation", "Standard Treatment"))

antidepmed_Var_plot_df <- data_numeric %>% group_by(antidepmed, Var) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n) %>%
  mutate(Var = ifelse(Var == 1, "Varenicline", "Placebo"),
         SE = sqrt((prop * (1-prop)/n)))

anhendonia_BA_plot_df <- data_numeric %>% group_by(abst, BA) %>%
  summarize(mean_score = mean(shaps_score_pq1, na.rm=TRUE), sd_age = sd(shaps_score_pq1, na.rm=TRUE),

currentMDD_BA_plot_df <- data_numeric %>% group_by(mde_curr, BA) %>%
  summarize(n = n(), n_abst = sum(abst), prop= n_abst/n)%>%
  mutate(BA = ifelse(BA == 1, "Behavioral Activation", "Standard Treatment"),
         SE = sqrt((prop * (1-prop)/n)))

p1 <- ggplot(black_var_plot_df, aes(x = Var, y = prop, fill = as.factor(Black))) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = prop - 1.96 * SE, ymax = prop + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Proportion Abstinent", fill = "Black", x = "Treatment", title = "a") +
  theme_minimal()


p2 <- ggplot(black_BA_plot_df, aes(x = BA, y = prop, fill = as.factor(Black))) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = prop - 1.96 * SE, ymax = prop + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Proportion Abstinent", fill = "Black", x = "Treatment", title = "b") +
  theme_minimal()

p3 <- ggplot(sex_var_plot_df, aes(x = Var, y = prop, fill = as.factor(sex_ps))) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = prop - 1.96 * SE, ymax = prop + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Proportion Abstinent", fill = "Sex", x = "Treatment", title = "c") +
  theme_minimal()
```

```r
p4 <- ggplot(sex_BA_plot_df, aes(x = BA, y = prop, fill = as.factor(sex_ps))) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = prop - 1.96 * SE, ymax = prop + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Proportion Abstinent", fill = "Sex", x = "Treatment", title = "d") +
  theme_minimal()

p5 <- ggplot(age_var_plot_df, aes(x=as.factor(abst), y = mean_age, fill= Var)) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = mean_age - 1.96 * SE, ymax = mean_age + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Mean Age", fill = "Treatment", x = "Smoking Abstinence", title = "e") +
  theme_minimal()

p6 <- ggplot(age_BA_plot_df, aes(x=as.factor(abst), y = mean_age, fill= BA)) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = mean_age - 1.96 * SE, ymax = mean_age + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Mean Age", fill = "Treatment", x = "Smoking Abstinence", title = "f") +
  theme_minimal()


p7 <- ggplot(shaps_BA_plot_df, aes(x=as.factor(abst), y = mean_score, fill= BA)) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = mean_score - 1.96 * SE, ymax = mean_score + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Mean FTCD Score", fill = "Treatment", x = "Smoking Abstinence", title = "g") +
  theme_minimal()

p8 <- ggplot(antidepmed_Var_plot_df,
             aes(x = as.factor(antidepmed), y = prop, fill = Var)) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = prop - 1.96 * SE, ymax = prop + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Proportion Abstinent", fill = "Treatment",
       x = "On Antidepresssents", title = "h") +
  theme_minimal()

p9 <- ggplot(anhendonia_BA_plot_df, aes(x=as.factor(abst), y = mean_score, fill= BA)) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = mean_score - 1.96 * SE, ymax = mean_score + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Mean Anhedonia Score", fill = "Treatment", x = "Smoking Abstinence", title = "i") +
  theme_minimal()

p10 <- ggplot(currentMDD_BA_plot_df,
             aes(x = as.factor(mde_curr), y = prop, fill = BA)) +
  geom_col(position = "dodge") +
  geom_errorbar(aes(ymin = prop - 1.96 * SE, ymax = prop + 1.96 * SE),
                position = position_dodge(width = 0.9), width = 0.25) +
  labs(y = "Proportion Abstinent", fill = "Treatment",
```

```
       x = "Current MDD", title = "j") +
  theme_minimal()
grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10, ncol=2)
###use the multiple imputation data sets

#matrix for storing results
res <- matrix(nrow=48, ncol=1)
colnames(res) <- "s1"

#vector for storing auc
auc_result <- c()
auc_result_train <- c()

#list to hold rocs
rocs <- list()
rocs_train <- list()

#set seed for reproducibility
set.seed(100)

#for each of the 5 data sets
for (i in 1:5){
  #get the imputed training data set
  training_data <- complete(miced_data_train, action = i)

  #get the imputed test data set
  test_data <- complete(miced_data_test, action = i)

  #remove id column
  data_for_X <- training_data %>% dplyr:: select(-id)

  #get matrix to use for glmnet
  X <- model.matrix(abst ~ BA*Var + BA*age_ps + BA*sex_ps + BA*NHW +
                      BA*Black + BA*Hisp + BA*inc + BA*edu + BA*ftcd_score +
                      BA*ftcd.5.mins + BA*bdi_score_w00 + BA*cpd_ps +
                      BA*crv_total_pq1 + BA*hedonsum_n_pq1 + BA*hedonsum_y_pq1 +
                      BA*shaps_score_pq1 + BA*otherdiag +
                      BA*antidepmed + BA*mde_curr + BA*NMR +
                      BA*Only.Menthol + BA*readiness + Var*Black + Var*antidepmed, data=data_for_X)[,

  #get outcome
  y <- as.numeric(training_data$abst)


  #prevent var and BA from being penalized
  pen.f <- rep(1, ncol(X))
  pen.f[c(1,2)] <- 0

  #fit Lasso model on the training set
  fit_cv <- cv.glmnet(X,y, alpha=1, nfolds = 5, penalty.factor=pen.f)

  #get best value of lambda based on cv
  s <- fit_cv$lambda.min
```

```
  #store result
  res <- cbind(res,coef(fit_cv, s=s))

  data_for_X_test <- test_data %>% dplyr:: select(-id)

  X_test <- model.matrix(abst ~ BA*Var + BA*age_ps + BA*sex_ps + BA*NHW +
                       BA*Black + BA*Hisp + BA*inc + BA*edu + BA*ftcd_score +
                       BA*ftcd.5.mins + BA*bdi_score_w00 + BA*cpd_ps +
                       BA*crv_total_pq1 + BA*hedonsum_n_pq1 + BA*hedonsum_y_pq1 +
                       BA*shaps_score_pq1 + BA*otherdiag +
                       BA*antidepmed + BA*mde_curr + BA*NMR +
                       BA*Only.Menthol + BA*readiness + Var*Black + Var*antidepmed, data=data_for_X_te

  #get test outcome
  y_test <- data_for_X_test$abst

  #get predicted values from the test set
  preds <- predict(fit_cv, X_test , s = s)

  #get auc
  roc <- roc(y_test, preds)
  auc_result <- c(auc_result, roc$auc)

  rocs[[i]] <- roc


  #get predicted values from the training set
  preds_train <- predict(fit_cv, X , s = s)

  #get auc training
  roc_train <- roc(y, preds_train)
  auc_result_train <- c(auc_result_train, roc_train$auc)

  rocs_train[[i]] <- roc_train
}

#take average of coefficients and aucs
results <- apply(res,1,mean, na.rm=TRUE)
auc <- mean(auc_result)
auc_train <- mean(auc_result_train)
#repeat the process from above but with the transformed data

#matrix to hold coefficients
res <- matrix(nrow=48, ncol=1)
colnames(res) <- "s1"

#vector to hold auc
auc_result <- c()
auc_result_train_t <- c()

#list to hold roc
rocs_transformed <- list()
rocs_train_t <- list()
```

```r
#set seed for reproducibility
set.seed(100)

#for each imputed data set
for (i in 1:5){

  #get the imputed training data set
  training_data <- complete(miced_data_train, action = i)

  #get the imputed test data set
  test_data <- complete(miced_data_test, action = i)

  #perform variable transformations
  data_transformed <- training_data %>% mutate(age_ps_squared = age_ps^2,
                            sqrt.bdi_score_w00 = sqrt(bdi_score_w00),
                            sqrt.cpd_ps = sqrt(cpd_ps),
                            sqrt.hedonsum_y_pq1 = sqrt(hedonsum_y_pq1 + 1),
                            log.NMR = log(NMR))

  #remove id column
  data_for_X <- data_transformed %>% dplyr:: select(-id)

  #get matrix to use for glmnet
  X <- model.matrix(abst ~ BA*Var + BA*age_ps_squared + BA*sex_ps + BA*NHW +
                      BA*Black + BA*Hisp + BA*inc + BA*edu + BA*ftcd_score +
                      BA*ftcd.5.mins + BA*sqrt.bdi_score_w00 + BA*sqrt.cpd_ps +
                      BA*crv_total_pq1 + BA*hedonsum_n_pq1 + BA*sqrt.hedonsum_y_pq1 +
                      BA*shaps_score_pq1 + BA*otherdiag +
                      BA*antidepmed + BA*mde_curr + BA*log.NMR +
                      BA*Only.Menthol + BA*readiness + Var*Black + Var*antidepmed, data=data_for_X)[,


  #get outcome
  y <- as.numeric(training_data$abst)

  #prevent var and BA from being penalized
  pen.f <- rep(1, ncol(X))
  pen.f[c(1,2)] <- 0

  #fit Lasso model on the data set
  fit_cv <- cv.glmnet(X,y, alpha=1, nfolds = 5, penalty.factor=pen.f)
  s <- fit_cv$lambda.min
  res <- cbind(res,coef(fit_cv, s=s))

  #get transformed test data
  data_transformed_test <- test_data %>% mutate(age_ps_squared = age_ps^2,
                            sqrt.bdi_score_w00 = sqrt(bdi_score_w00),
                            sqrt.cpd_ps = sqrt(cpd_ps),
                            sqrt.hedonsum_y_pq1 = sqrt(hedonsum_y_pq1 + 1),
                            log.NMR = log(NMR))

  data_for_X_test <- data_transformed_test %>% dplyr:: select(-id)
```

```
  #create matrix for glm net for test data
 X_test <- model.matrix(abst ~ BA*Var + BA*age_ps_squared + BA*sex_ps + BA*NHW +
                        BA*Black + BA*Hisp + BA*inc + BA*edu + BA*ftcd_score +
                        BA*ftcd.5.mins + BA*sqrt.bdi_score_w00 + BA*sqrt.cpd_ps +
                        BA*crv_total_pq1 + BA*hedonsum_n_pq1 + BA*sqrt.hedonsum_y_pq1 +
                        BA*shaps_score_pq1 + BA*otherdiag +
                        BA*antidepmed + BA*mde_curr + BA*log.NMR +
                        BA*Only.Menthol + BA*readiness + Var*Black + Var*antidepmed, data=data_for_X_te

 #get outcomes for test data
  y_test <- as.numeric(data_transformed_test$abst)

  #get predictions on test set
  preds <- predict(fit_cv, X_test , s = s)
  roc <- roc(y_test, preds)

  auc_result <- c(auc_result, roc$auc)

  rocs_transformed[[i]] <- roc

  #get predicted values from the training set
  preds_train_t <- predict(fit_cv, X , s = s)

  #get auc training
  roc_train_t <- roc(y, preds_train_t)
  auc_result_train_t <- c(auc_result_train_t, roc_train_t$auc)

  rocs_train_t[[i]] <- roc_train_t

}

#take average of coefficients and aucs
results_transformed <- apply(res,1,mean, na.rm=TRUE)
auc_transformed <- mean(auc_result)
auc_train_t <- mean(auc_result_train_t)

#get long format data
train <- complete(miced_data_train, action = "long")
test <- complete(miced_data_test, action = "long")


#get transformed test data
data_transformed_test <- test %>% mutate(age_ps_squared = age_ps^2,
                            sqrt.bdi_score_w00 = sqrt(bdi_score_w00),
                            sqrt.cpd_ps = sqrt(cpd_ps),
                            sqrt.hedonsum_y_pq1 = sqrt(hedonsum_y_pq1 + 1),
                            log.NMR = log(NMR))

#get transformed training data
data_transformed_train <- train %>% mutate(age_ps_squared = age_ps^2,
                            sqrt.bdi_score_w00 = sqrt(bdi_score_w00),
                            sqrt.cpd_ps = sqrt(cpd_ps),
```

```
                          sqrt.hedonsum_y_pq1 = sqrt(hedonsum_y_pq1 + 1),
                          log.NMR = log(NMR))

#remove id variables
data_for_X_test <- data_transformed_test %>% dplyr:: select(-id)
data_for_X_train <- data_transformed_train %>% dplyr:: select(-id)

#get test data matrix
X_test <- model.matrix(abst ~ BA*Var + BA*age_ps_squared + BA*sex_ps + BA*NHW +
                       BA*Black + BA*Hisp + BA*inc + BA*edu + BA*ftcd_score +
                       BA*ftcd.5.mins + BA*sqrt.bdi_score_w00 + BA*sqrt.cpd_ps +
                       BA*crv_total_pq1 + BA*hedonsum_n_pq1 + BA*sqrt.hedonsum_y_pq1 +
                       BA*shaps_score_pq1 + BA*otherdiag +
                       BA*antidepmed + BA*mde_curr + BA*log.NMR +
                       BA*Only.Menthol + BA*readiness + Var*Black + Var*antidepmed, data= data_for_X_t

#get train data matrix
X_train <- model.matrix(abst ~ BA*Var + BA*age_ps_squared + BA*sex_ps + BA*NHW +
                       BA*Black + BA*Hisp + BA*inc + BA*edu + BA*ftcd_score +
                       BA*ftcd.5.mins + BA*sqrt.bdi_score_w00 + BA*sqrt.cpd_ps +
                       BA*crv_total_pq1 + BA*hedonsum_n_pq1 + BA*sqrt.hedonsum_y_pq1 +
                       BA*shaps_score_pq1 + BA*otherdiag +
                       BA*antidepmed + BA*mde_curr + BA*log.NMR +
                       BA*Only.Menthol + BA*readiness + Var*Black + Var*antidepmed, data= data_for_X_t

#predict on test and train
preds_test <- X_test %*% as.data.frame(results_transformed)[,1]
preds_train <- X_train %*% as.data.frame(results_transformed)[,1]

#get test and train outcomes
y_test <- as.numeric(data_for_X_test$abst)
y_train <- as.numeric(data_for_X_train$abst)

#get roc on test and training data
roc_test <- roc(y_test, preds_test)
roc_train <- roc(y_train, preds_train)

#start plotting arrangement
par(mfrow = c(1, 2))

plot(roc_train, main = "ROC Curve on Transformed Training Data", cex.main =.8)
# Add AUC value as text on the plot
text(x = 0.6, y = 0.2, labels = paste("AUC =", round(auc_train_t, 4)), cex = 1.2)

plot(roc_test, main = "ROC Curve on Transformed Test Data", cex.main = .8)
# Add AUC value as text on the plot
text(x = 0.6, y = 0.2, labels = paste("AUC =", round(auc_transformed, 4)), cex = 1.2)

#go back to normal plotting arrangement
par(mfrow = c(1, 1))

#set number of cuts for calibration plot
num_cuts <- 10
```

```r
#create data frame with probabilities, bins, and class
calib_test <-  data.frame("prob" = preds_test,
                          bin = cut(preds_test, breaks = num_cuts),
                          class = y_test)

calib_test <- calib_test %>%
              group_by(bin) %>%
              summarise(observed = sum(class)/n(),
                        expected = sum(prob)/n(),
                        se = sqrt(observed * (1-observed) / n()))

#get calibration plot for transformed data
calib_test_plot <- ggplot(calib_test) +
  geom_line(aes(x = expected, y = expected, color = "Ideal")) +
  geom_smooth(aes(x = expected, y = observed, color = "Loess Fit"),
              method = "loess") +
  labs(x = "Expected Proportion", y = "Observed Proportion",
       title = "Calibration: Transformed Test Data", color = "Legend") +
  theme(plot.title = element_text(size=10)) +
  theme_minimal() +
  scale_color_manual(values = c("Ideal" = "red", "Loess Fit" = "blue"))

#get calibration for training data
calib_train <-  data.frame("prob" = preds_train,
                           bin = cut(preds_train, breaks = num_cuts),
                           class = y_train)

calib_train <- calib_train %>%
              group_by(bin) %>%
              summarise(observed = sum(class)/n(),
                        expected = sum(prob)/n(),
                        se = sqrt(observed * (1-observed) / n()))

#get calibration plot for training data
calib_train_plot <- ggplot(calib_train) +
  geom_line(aes(x = expected, y = expected, color = "Ideal")) +
  geom_smooth(aes(x = expected, y = observed, color = "Loess Fit"),
              method = "loess") +
  labs(x = "Expected Proportion", y = "Observed Proportion",
       title = "Calibration: Transformed Training Data", color = "Legend") +
  theme(plot.title = element_text(size=10)) +
  theme_minimal() +
  scale_color_manual(values = c("Ideal" = "red", "Loess Fit" = "blue"))

grid.arrange(calib_train_plot, calib_test_plot, ncol=2)


#get long format data
train <- complete(miced_data_train, action = "long")
test <- complete(miced_data_test, action = "long")

#remove id column
```

```r
data_for_X_test <- test %>% dplyr:: select(-id)
data_for_X_train <- train %>% dplyr:: select(-id)

#get X test matrix
X_test <- model.matrix(abst ~ BA*Var + BA*age_ps + BA*sex_ps + BA*NHW +
                       BA*Black + BA*Hisp + BA*inc + BA*edu + BA*ftcd_score +
                       BA*ftcd.5.mins + BA*bdi_score_w00 + BA*cpd_ps +
                       BA*crv_total_pq1 + BA*hedonsum_n_pq1 + BA*hedonsum_y_pq1 +
                       BA*shaps_score_pq1 + BA*otherdiag +
                       BA*antidepmed + BA*mde_curr + BA*NMR +
                       BA*Only.Menthol + BA*readiness + Var*Black + Var*antidepmed, data=data_for_X_te

#get X train matrix
X_train <- model.matrix(abst ~ BA*Var + BA*age_ps + BA*sex_ps + BA*NHW +
                       BA*Black + BA*Hisp + BA*inc + BA*edu + BA*ftcd_score +
                       BA*ftcd.5.mins + BA*bdi_score_w00 + BA*cpd_ps +
                       BA*crv_total_pq1 + BA*hedonsum_n_pq1 + BA*hedonsum_y_pq1 +
                       BA*shaps_score_pq1 + BA*otherdiag +
                       BA*antidepmed + BA*mde_curr + BA*NMR +
                       BA*Only.Menthol + BA*readiness + Var*Black + Var*antidepmed, data=data_for_X_tr

#get predictions
preds_test <- X_test %*% as.data.frame(results)[,1]
preds_train <- X_train %*% as.data.frame(results)[,1]

#get outcomes
y_test <- as.numeric(data_for_X_test$abst)
y_train <- as.numeric(data_for_X_train$abst)

#get rocs for test and training sets
roc_test <- roc(y_test, preds_test)
roc_train <- roc(y_train, preds_train)

par(mfrow = c(1, 2))

plot(roc_train, main = "ROC Curve on Training Data", cex.main = .8)
text(x = 0.6, y = 0.2, labels = paste("AUC =", round(auc_train, 4)), cex = 1.2)

plot(roc_test, main = "ROC Curve on Test Data", cex.main = .8)
text(x = 0.6, y = 0.2, labels = paste("AUC =", round(auc, 4)), cex = 1.2)

par(mfrow = c(1, 1))
num_cuts <- 10

#get calibration results
calib_test <-  data.frame("prob" = preds_test,
                          bin = cut(preds_test, breaks = num_cuts),
                          class = y_test)
calib_test <- calib_test %>%
             group_by(bin) %>%
             summarise(observed = sum(class)/n(),
                       expected = sum(prob)/n(),
                       se = sqrt(observed * (1-observed) / n()))
```

```r
calib_test_plot <- ggplot(calib_test) +
  geom_line(aes(x = expected, y = expected, color = "Ideal")) +
  geom_smooth(aes(x = expected, y = observed, color = "Loess Fit"),
              method = "loess") +
  labs(x = "Expected Proportion", y = "Observed Proportion",
       title = "Calibration: Test Data", color = "Legend") +
  theme(plot.title = element_text(size=10)) +
  theme_minimal() +
  scale_color_manual(values = c("Ideal" = "red", "Loess Fit" = "blue"))

#get training calibration results
calib_train <-  data.frame("prob" = preds_train,
                           bin = cut(preds_train, breaks = num_cuts),
                           class = y_train)
calib_train <- calib_train %>%
              group_by(bin) %>%
              summarise(observed = sum(class)/n(),
                        expected = sum(prob)/n(),
                        se = sqrt(observed * (1-observed) / n()))

#get training calibration plot
calib_train_plot <- ggplot(calib_train) +
  geom_line(aes(x = expected, y = expected, color = "Ideal")) +
  geom_smooth(aes(x = expected, y = observed, color = "Loess Fit"),
              method = "loess") +
  labs(x = "Expected Proportion", y = "Observed Proportion",
       title = "Calibration: Training Data", color = "Legend") +
  theme(plot.title = element_text(size=10)) +
  theme_minimal() +
  scale_color_manual(values = c("Ideal" = "red", "Loess Fit" = "blue"))

grid.arrange(calib_train_plot, calib_test_plot, ncol = 2)

results <- as.data.frame(results_transformed)
results <- results %>% filter(abs(results_transformed) > 0)
colnames(results) <- "OR"

results$OR <- round(exp(results$OR), digits = 5)

rownames(results) <- c("(Intercept)", "Behavioral Activation", "Varenicline", "Non-Hispanic White", '

#results <- readRDS(file = "results.RData")

results %>%
    kable(., caption = "Variables Selected By Lasso")
```