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 therapy, and Varenicline to a placebo.

They found that Behavioral Activation Therapy (BA) did not outperform the standard therapy 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 10-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 variables selected by LASSO were Varenicline, an indicator for Non-Hispanic White, FTCD score, anhedonia, presence of other mental health diagnoses, an indicator for if the participant is currently a smoker, Nicotine Metabolism Ratio (NMR), smoking only menthol cigarettes, BA with income, BA with NMR, BA with smoking only menthol cigarettes, and Varenicline with an indicator for Black. The results of this analysis suggest that the effectiveness of Behavioral Activation Therapy for smoking cessation may vary based on participant characteristics.

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, and Varenicline is compared to a placebo in a 2x2 factorial design. 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. We hypothesized that age, sex, readiness to quit smoking, and race, would be moderators of BA. 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.

Missing Data and Preprocessing

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. Missing data was imputed with multiple imputation using the mice package (van Buuren and Groothuis-Oudshoorn, 2011), resulting in 5 imputed data sets. 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** N = 68	**Behavioral Activation** N = 68	**Varenicline** N = 81	**Both** 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
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%)	

Table 1: Data Characteristics *(continued)*

Characteristic	**Placebo** N = 68	**Behavioral Activation** N = 68	**Varenicline** N = 81	**Both** N = 83	**p-value**
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 pleasurable events scale variables, anhedonia, and nicotine metabolism ratio were skewed. After 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 Pleasurable Events Scale-complimentary reinforcers, in one of our possible models.

To inform how we select the variables to include in our full model, and to check for collinearity among the variables, we examined the pearson correlations using the data without imputation (Figure 1). Notable correlations include BDI score with whether or not the participant currently has MDD (r = .58), the indicator variable "Black" with smoking only menthol cigarettes (r=.46), BDI score with anhedonia (r = .41), Black with Non-Hispanic White (r = -.76), income with education (r = .40), Black with Non Hispanic White (-.77), FTCD score at baseline with smoking within 5 minutes of waking up (r = .63), and cigarettes per day with FTCD score (r = .52). 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 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 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.

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

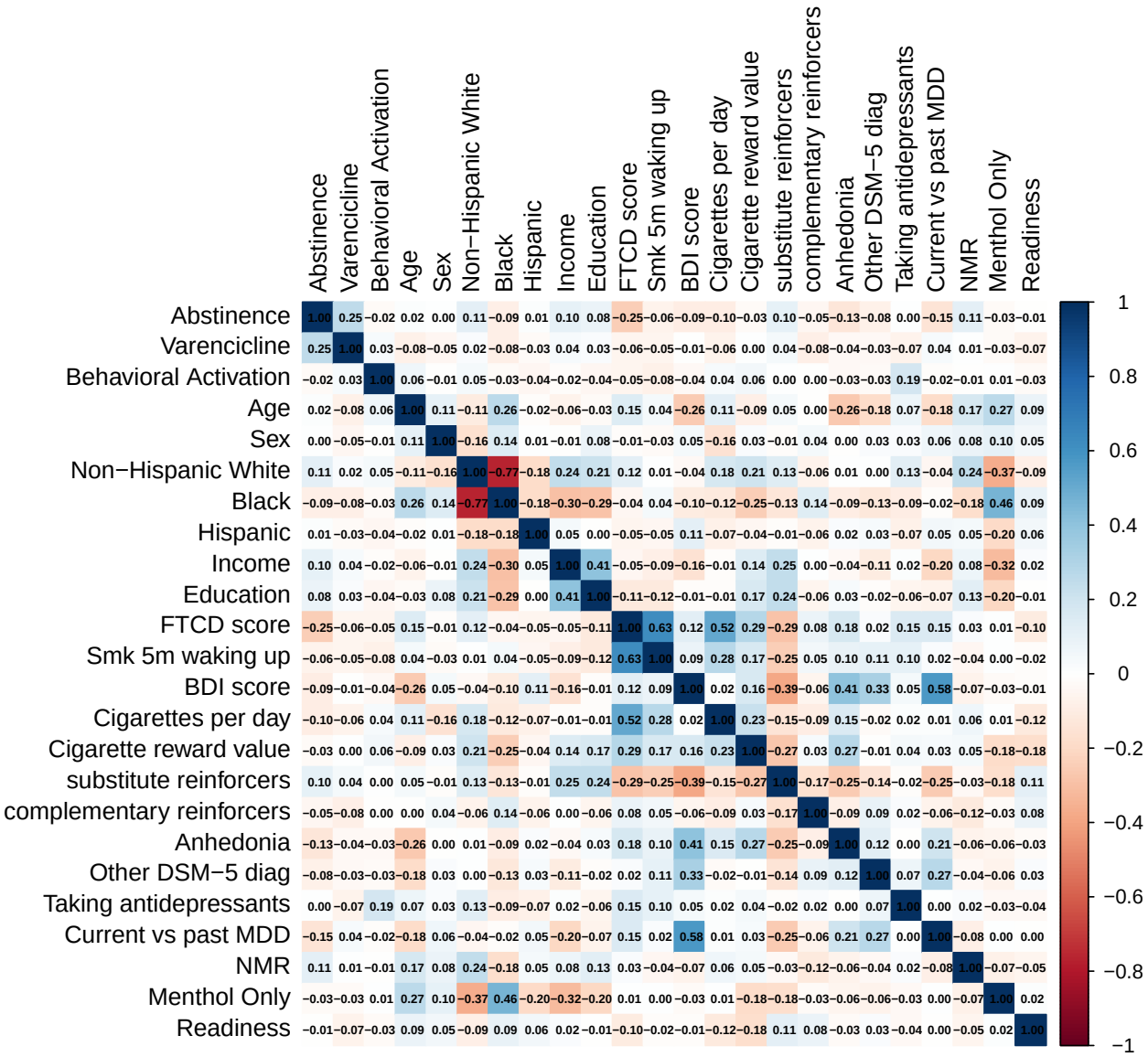

Figure 1: Correlation Plot For All Variables

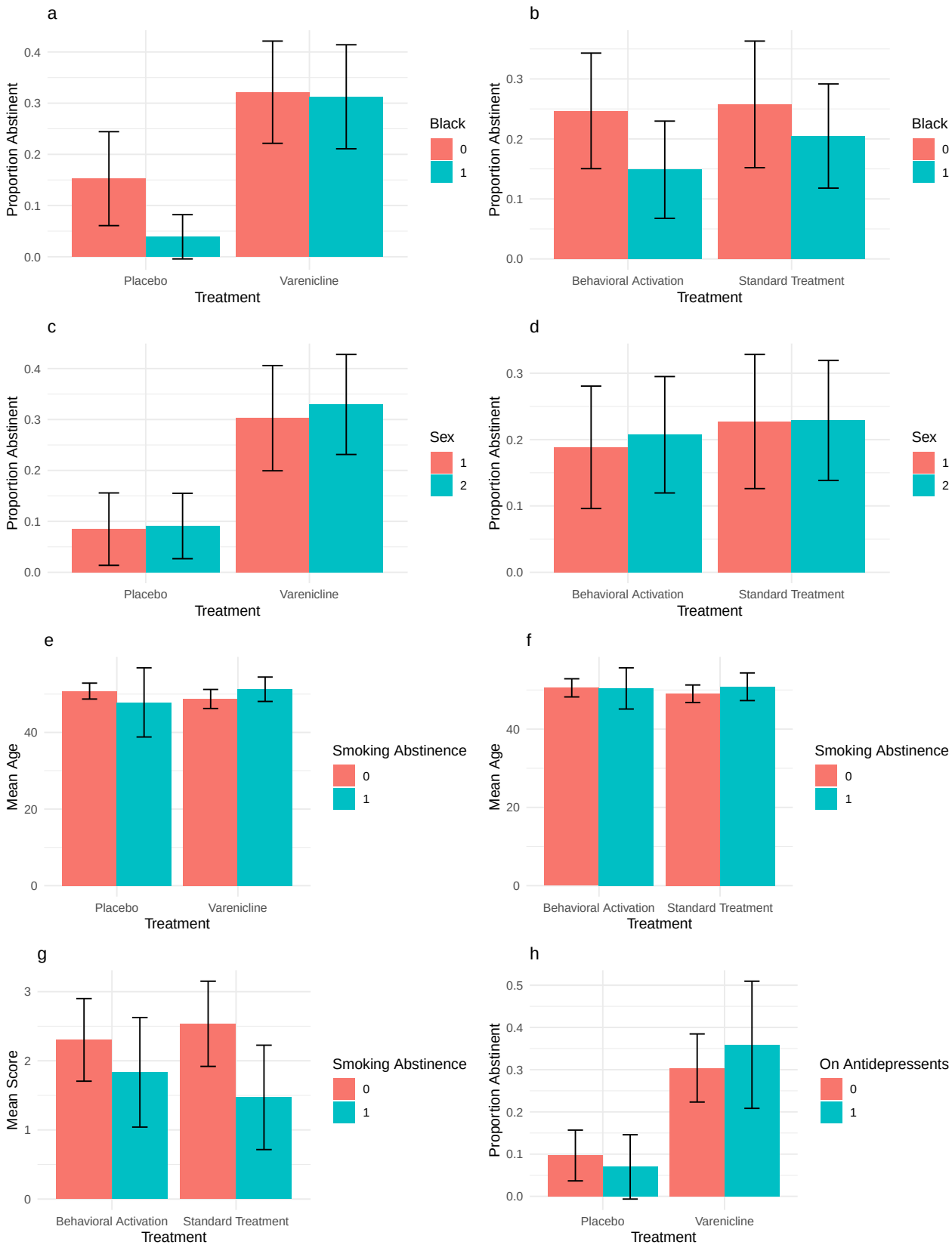

Figure 2: Interactions with Covariates and Treatments. Error bars represent 95% confidence intervals

is taking antidepressants. We then used the glmnet package (Friedman et al., 2010), to find the best value of lambda via 10-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: Var, NHW, FTCD score, Anhedonia, Other diagnoses, current MDE, NMR, Only.Menthol, Readiness, and BA:income, BA:NMR, BA:Only Menthol, and Var:Black.

The model with transformations selected Var, sex, NHW, FTCD score, anhedonia, current MDE, log(NMR), readiness to quit, BA:FTCD score, BA: antidepressants, Var: Black, and Var: antidepressants.

The AUC obtained by the model without transformations was 0.747, and 0.690 with transformations, both on the test set. Interestingly, the ROC curves shown in figure 3 indicate that the model with the transformed data performed better on the test set than the training set. We would expect that the model would be overfit to the training set, thus producing a better ROC curve. The model that did not consider variable transformations performed similarly on the test and training data (Figure 5.)

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, although the proportion of people in the abstinence class appears to be overestimated a bit when the expected proportion is 0.5 (Figure 4).

The calibration plots fit on the data without transformations also show a good fit for the test data, with an overestimate of abstinence in the training set (Figure 5).

Since the model without the transformations performed better based on AUC, we chose the model with no transformations as our best model. An AUC of 0.747 indicates that the model performs better than if the model randomly assigned treatment groups, which would be an AUC of 0.5.

The variables selected by LASSO using the model fit without transformations are shown in Table 2.

Table 2: Variables Selected By Lasso

	Coefficients
(Intercept)	0.24
Var	0.165
NHW	0.0587
ftcd_score	-0.0151
shaps_score_pq1	-0.0038
otherdiag	-0.0544
mde_curr	-0.0284
NMR	0.0149
Only.Menthol	0.0017
readiness	-0.00211
BA:inc	0.000326
BA:NMR	0.0027
BA:Only.Menthol	-0.0639
Var:Black	0.0127

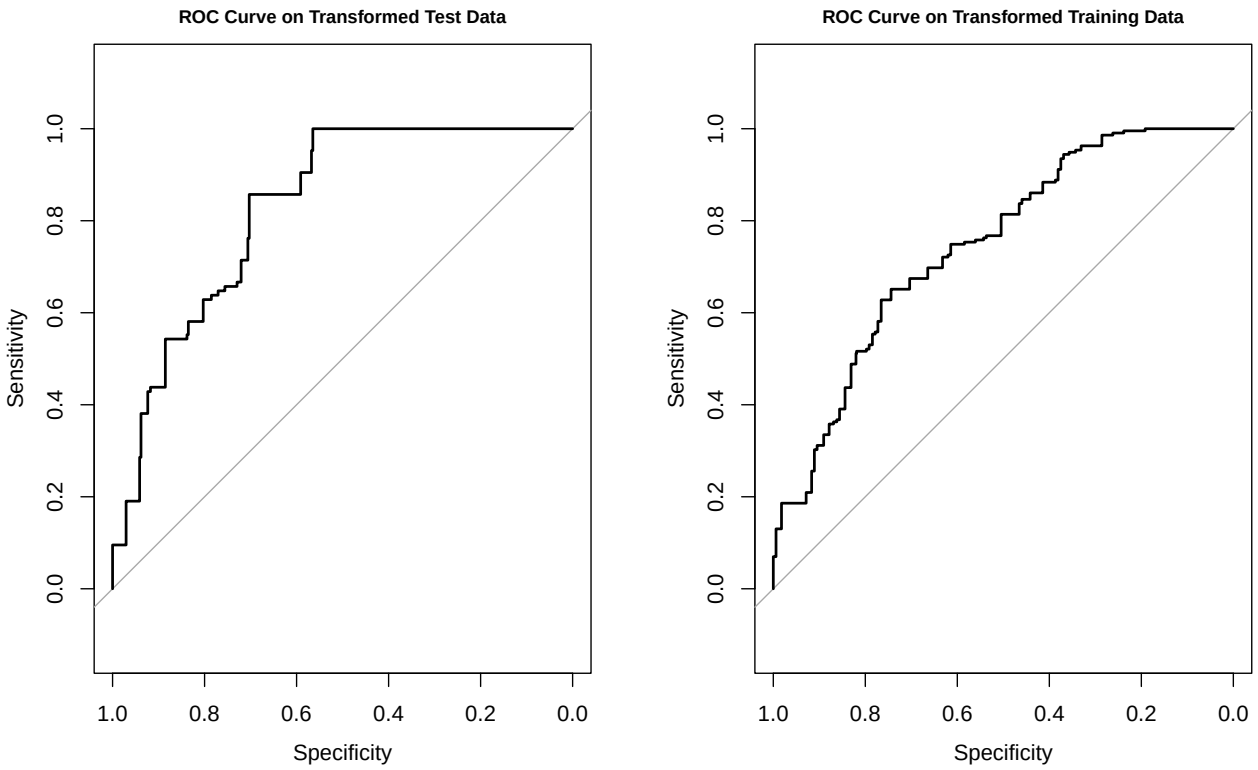

Figure 3: ROC Curves for Transformed Data Model

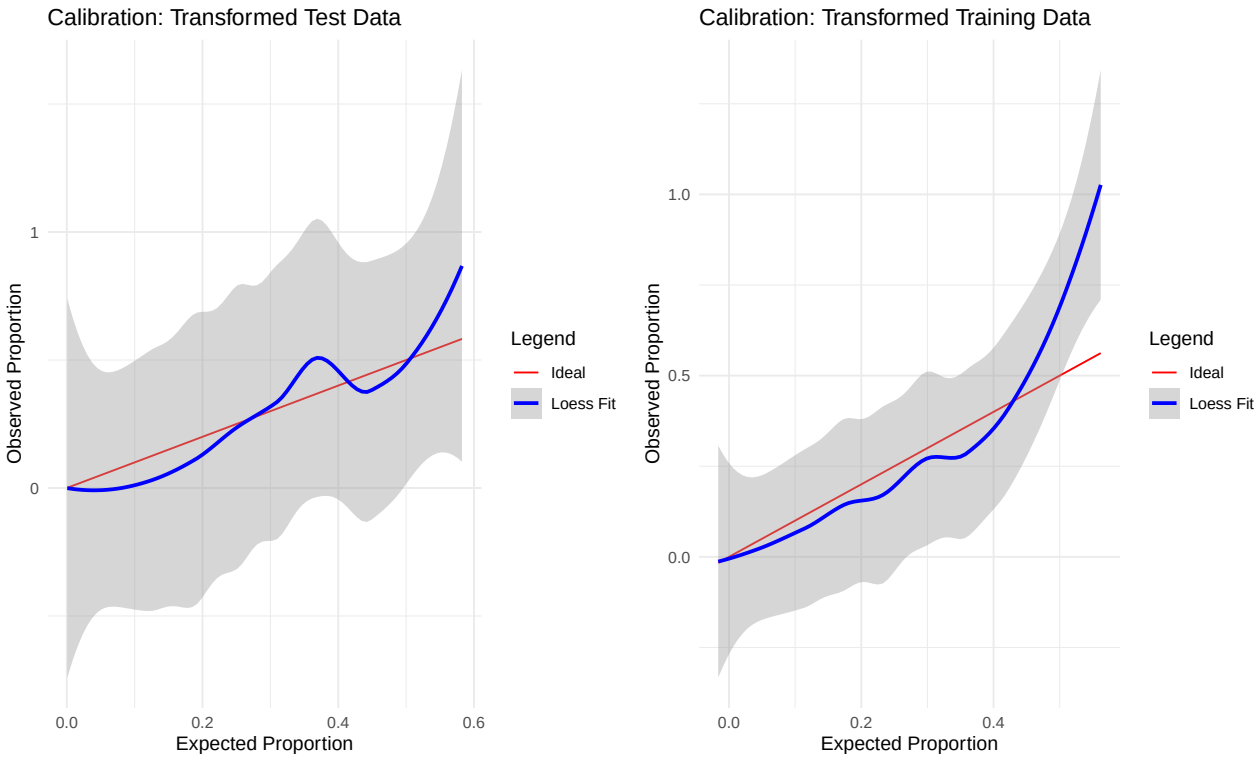

Figure 4: Calibration Plots for Transformed Data Model

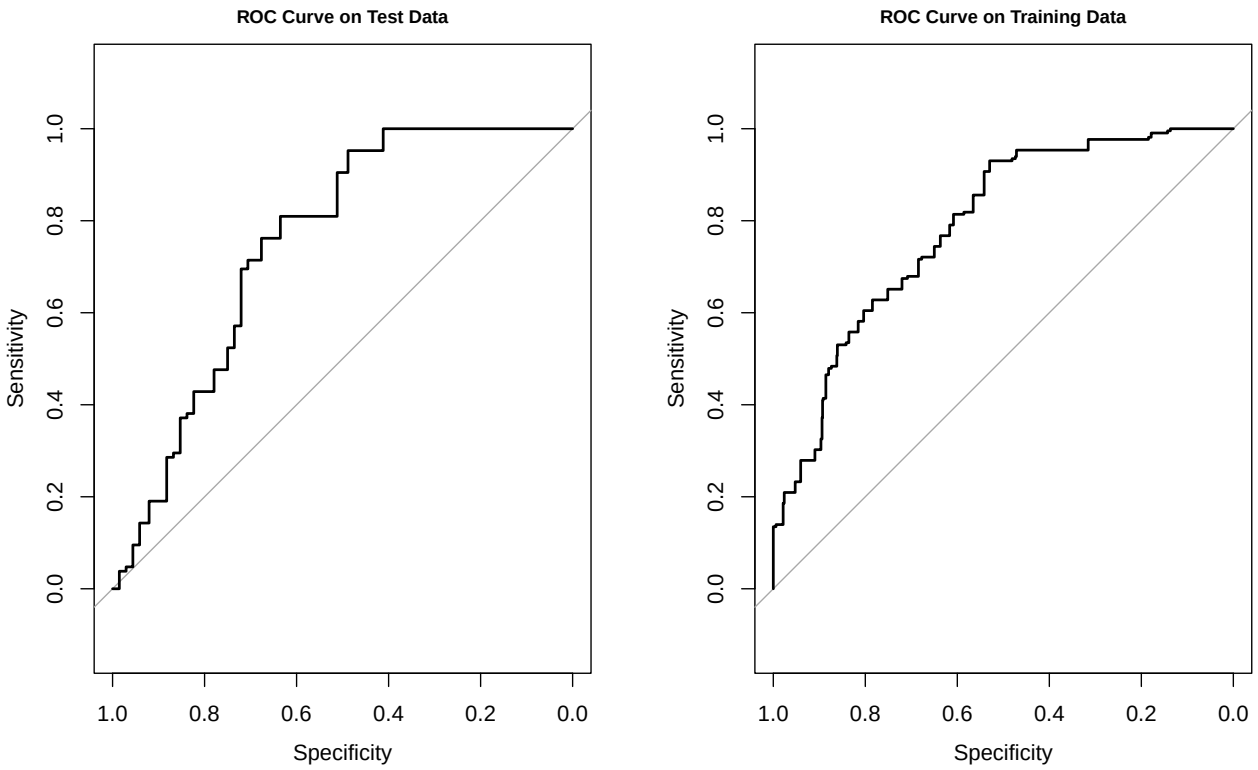

Figure 5: ROC Curves for Data Modeled Without Transformations

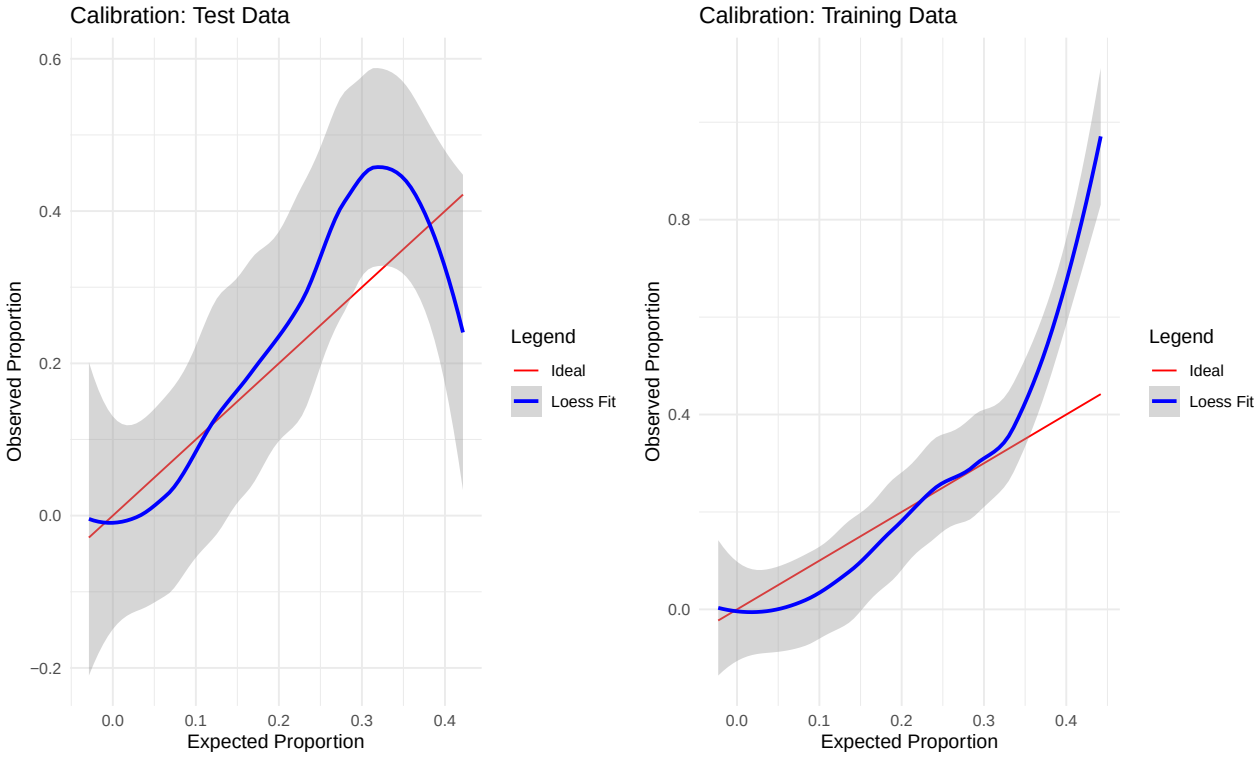

Figure 6: ROC Curves for Data Modeled Without Transformations

Discussion and Conclusion

Based on these results, we reaffirm that Varenicline is effective at increasing smoking cessation. The other variables selected by LASSO did not align with our previous hypothesis that race, sex, age, and readiness would interact with BA. The log odds for abstinence are predicted to increase when a participant is assigned Varenicline. The log odds of abstinence also increase when a participant is non-hispanic white, and also increase with nicotine metabolism ratio, and use of only menthol cigarettes. The effect of BA on abstinence is increased with income and NMR, and the effect of Varenicline is increased when the participant is black.

FTCD score, anhedonia, the presence of other diagnoses, currently having MDD, and readiness to quit smoking reduced the log odds of abstinence. These results make sense since these variables indicate greater struggles with mental health and nicotine addiction, which may make it more difficult to abstain from smoking. These variables were also negatively correlated with abstinence in Figure 1.

The positive effect of NMR and being non-Hispanic white on smoking abstinence agrees with the correlations we saw in Figure 1. However, the use of menthol only cigarettes was not correlated with abstinence in Figure 1, and menthol cigarettes have previously been suggested to be harder to quit (Foulds et al., 2010). The coefficient for smoking only menthol cigarettes is rather small, even though it was selected by LASSO. Future research should further investigate the effect of menthol cigarettes on the effectiveness of BA before any conclusions can be made. Since menthol was correlated with the age, black indicator, and income variables, the coefficient for menthol in the model may have absorbed some of the effects of these variables, since LASSO penalizes the inclusion of many variables.

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. Furthermore, LASSO selected interaction terms but not necessarily the corresponding main effects, which reduces the interpretability of the model, especially since the main-effect of BA was not included.

Overall, we found that the effectiveness of Behavioral Activation therapy may be moderated by participant income, NMR, and smoking only menthol cigarettes. Further research should be conducted to corroborate these findings and investigate how Behavioral Activation may be effective in these populations.

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

Foulds, J., Hooper, M. W., Pletcher, M. J., & Okuyemi, K. S. (2010). Do smokers of menthol cigarettes find it harder to quit smoking?. Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco, 12 Suppl 2(Suppl 2), S102–S109. https://doi.org/10.1093/ntr/ntq166

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.

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

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)
library(pROC)
library(gglasso)
data <- read_csv("Downloads/project2.csv")

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

```r
#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

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

#impute the missing data 5 times
miced_data <- mice(data, m=5)
#need to make a table 1 for each group
data <- data %>%
  mutate(group = case_when(Var == 0 & BA == 0 ~ "Placebo",
                           Var== 0 & BA == 1 ~ "Behavioral Activation",
                           Var== 1 & BA == 0 ~ "Varenicline",
                           Var== 1 & BA == 1 ~ "Both"))

#make sure everything is the correct data type
numeric_vars <- c("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]])
}
#remove varibles that we don't want in the table
data_for_tbl <- data %>% dplyr::select(-id, -Var, -BA, -abst)

#change levels for "group"
data_for_tbl$group <- factor(data_for_tbl$group,
                             levels = c("Placebo",
                                        "Behavioral Activation",
                                        "Varenicline",
                                        "Both" ))

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
```

```r
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",
                           "substitute reinforcers",
                           "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 = "pearson"), 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)))

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()


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=Var, y = mean_age, fill= as.factor(abst))) +
  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 = "Smoking Abstinence", x = "Treatment", title = "e") +
  theme_minimal()

p6 <- ggplot(age_BA_plot_df, aes(x=BA, y = mean_age, fill= as.factor(abst))) +
  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 = "Smoking Abstinence", x = "Treatment", title = "f") +
  theme_minimal()


p7 <- ggplot(shaps_BA_plot_df, aes(x=BA, y = mean_score, fill= as.factor(abst))) +
  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 Score", fill = "Smoking Abstinence", x = "Treatment", title = "g") +
  theme_minimal()

p8 <- ggplot(antidepmed_Var_plot_df,
             aes(x = Var, y = prop, fill = as.factor(antidepmed))) +
  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 = "On Antidepressents",
       x = "Treatment", title = "h") +
  theme_minimal()

grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8, ncol=2)
#set seed for replication
set.seed(190)
```

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

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

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

both_ids <- data[data$group == "Both",]$id
training_ids <- c(training_ids, sample(both_ids, round(length(both_ids)*.70)))

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

###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()

#list to hold rocs
rocs <- list()

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

  #remove test set
  training_data <- complete_data[training_rows,]

  #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 <- training_data$abst

  #set seed for reproducibility
```

```
  set.seed(10)

  #fit lasso model on the training set
  fit_cv <- cv.glmnet(X,y, alpha=1, nfolds = 10)
  #get best value of lambda based on cv
  s <- fit_cv$lambda.min
  #store result
  res <- cbind(res,coef(fit_cv, s=s))

  #create test set
  test_data <- complete_data[-training_rows,]

  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
}

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

#make training set
training_rows <- sample(1:nrow(data), nrow(data)*.65)

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

#vector to hold auc
auc_result <- c()

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

```r
#for each imputed data set
for (i in 1:5){
  #get imputed data set
  complete_data <- complete(miced_data, action = i)

  #get training data set
  training_data <- complete_data[training_rows,]

  #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 <- training_data$abst

  #set seed for reproducibility
  set.seed(10)

  #fit lasso model on the data set
  fit_cv <- cv.glmnet(X,y, alpha=1, nfolds = 10)
  s <- fit_cv$lambda.min
  res <- cbind(res,coef(fit_cv, s=s))

  #get test set
  test_data <- complete_data[-training_rows,]

  #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 <- 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

}

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

#get long format data
longdata <- complete(miced_data, action = "long")

#get test data
test <- longdata[!(longdata$id %in% training_ids),]

#get training data
train <- longdata[longdata$id %in% training_ids,]

#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 <- data_for_X_test$abst
y_train <- data_for_X_train$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_test, main = "ROC Curve on Transformed Test Data", cex.main = .8)
plot(roc_train, main = "ROC Curve on Transformed Training Data", cex.main =.8)

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

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

#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_test_plot, calib_train_plot, ncol=2)

#get long format data
longdata <- complete(miced_data, action = "long")

#test data set
test <- longdata[!(longdata$id %in% training_ids),]

#get training data set
train <- longdata[longdata$id %in% training_ids,]

#remove id column
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 <- data_for_X_test$abst
y_train <- data_for_X_train$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_test, main = "ROC Curve on Test Data", cex.main = .8)

plot(roc_train, main = "ROC Curve on Training Data", cex.main = .8)

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()))

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_test_plot, calib_train_plot, ncol = 2)

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

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

results %>% dplyr::mutate_if(is.numeric, funs(as.character(signif(., 3)))) %>%
    kable(., caption = "Variables Selected By Lasso")
```