Evaluating Baseline Variables as Moderators and Predictors for Smoking Cessation

Maia Lindner-Liaw

2024-11-07

Abstract

Background: More than 30% of people with major depressive disorder (MDD) are reported as daily smokers, yet smoking cessation treatments for this group are not well studied. The drug varenicline improved cessation rates in one study, but was less effective for MDD smokers, suggesting additional behavioral treatment is needed to help MDD patients quit. A follow up 2x2 factorial design study of both behavioral and drug treatment found no significant improvement in abstinence rates for the combined treatment. The goal of this study is to expand on the analysis of the 2x2 study to determine potential moderators for the behavioral treatment effect among baseline variables, as well as explore possible predictors of abstinence among the baseline variables.

Methods: Missing values were imputed using multiple imputation, and we fit a logistic regression with a lasso penalty for variable selection on a 75%/25% train-test data split for each imputed data set. 10-fold cross-validation was used to determine the optimal penalty for each lasso model, and the model coefficients were pooled. The model was evaluated using ROC curves, AUC, and a calibration plot.

Results: Current MDD and menthol cigarette use were possible moderators, and race, education, nicotine dependence, cigarette reward value, anhedosis, nicotine metabolism rate, and quit readiness were baseline predictors of abstinence.

Conclusion: Some baseline variables were found to be possible moderators for behavioral therapy treatment and predictors of abstinence, but the small sample size and limitations of the methods may affect these conclusions.

Introduction

Major depressive disorder (MDD) is one of the most common mental health disorders in the world, and more than 30% of people with MDD have been reported as daily smokers. This group often show greater dependence on tobacco products and experience stronger withdrawal symptoms, smoke more heavily, and experience a greater reward from smoking than people without MDD. These smoking trends highlight the importance of effective smoking cessation methods for people with MDD. However, most existing smoking cessation studies have excluded this group, leading to a lack of literature and evidence. There exists one study in with MDD smokers were included, which showed that the drug varenicline improved abstinence, but cessation was higher for smokers without MDD. This suggests that the addition of behavioral treatment may is necessary to obtain comparable cessation results for MDD sufferers. [1]

The study by Hitsman et al investigates the effectiveness of behavioral treatment in addition to varenicline for smoking cessation among patients with MDD. The study used a 2x2 factorial design to compare control, drug only, behavior only, and drug and behavioral treatment groups. Poisson regression was used to estimate abstinance rate ratios across the treatments using intent-to-treat framework. Results showed no significant interaction between the behavioral treatment and the varenicline. [1]

In this report, we will use the same data to conduct further analysis on the effect of the behavioral treatment for smoking cessation. In particular, we are interested in determining if baseline covariates in the data are moderators for the behavioral treatment effects. Additionally, we will be assessing whether these baseline covariates are predictors of abstinance, controlling for the drug and behavioral treatments.

EDA

We will begin by conducting exploratory data analysis to inform data pre-processing and modeling. There are 300 observations in the data set, with 25 variables. The covariates are a mix of binary, continuous, and categorical variables. The `sex_ps` covariates is coded as 1 for male and 2 for female, so we subtract 1 to get a binary indicator for female. All binary and categorical variables are factored to ensure they are not read as continuous variables. The four treatment types used in the original paper were behavioral only, drug only, both drug and behavioral, and no treatment control. To that end, we create a new variable called `trt` that summarizes the treatment group for each observation from the existing indicators for behavioral treatment and drug treatment. Table 1 shows the basic demographic information from the study participants, grouped by this new `trt` variable.

The average age across all groups is about 50 and females make up about 55% of the sample as indicated by the `sex_ps` variable, where 1 indicates female. Black subjects make up roughly 45%-60% of the sample, followed by non-hispanic whites (NHW) at 30%-40%, and finally hispanics with less than 8% in each group. Most subjects fall into the lowest income category earning less than $20,000 a year, and the majority have some college or technical school education. Overall, the baseline covariates are balanced across the treatment groups.

Additional variables available in the data set are shown in Table 2, stratified by abstinence status. These include recorded measurements such as level of nicotine dependence (FTCD score), number of cigarettes per day (`cpd_ps`), depression severity score (`bdi_score_w00`), and antidepressant medication status (`antidepmed`). In this table, we do see some differences in these measures between the abstinent and non-abstinent group. In particular, baseline nicotine dependence as measured by FTCD score is higher in the non-abstinent group with small standard error, which makes sense as higher dependence will make quitting more difficult. The `mde_curr` variable, which is the number of subjects currently vs previously suffering from MDD, is also much higher in the non-abstinent group, at 52% compared to 39%. This agrees with existing findings that MDD makes smoking cessation more challenging. In the baseline readiness to quit, scores from 5 to 8 were the most common in both outcome groups.

We can also look at the correlation between covariates to get an idea of which variables are related to each other. Figure 1 visualizes these correlations. We can see that there are some covariates that are highly correlated. The very strong negative correlation between `Black` and `NHW` is because subjects are either one or the other, and there are not many Hispanic subjects. Education and income have a strong positive correlation, which is expected, as higher education generally leads to higher pay. The indicator for smoking only menthol cigarettes is notably correlated with race, showing Blacks are more likely to smoke only menthols compared to non-hispanic whites. FTCD score, smoking within 5 minutes of waking up, and cigarettes per day all have strong positive correlations. Finally, the indicator for currently having MDD vs previously having it (`mde_curr`) also has a strong positive correlation with depression severity (`bdi_score_w00`). Most of these strong correlations are not unexpected, as the covariates has similarities in what they are measuring.

Table 1: Demographics by Treatment Type

Characteristic	**Behavioral Only** N = 68	**Both** N = 83	**Control** N = 68	**Drug Only** N = 81
age_ps	51 (14)	50 (13)	50 (11)	49 (13)
sex_ps				
0	30 (44%)	39 (47%)	29 (43%)	37 (46%)
1	38 (56%)	44 (53%)	39 (57%)	44 (54%)
NHW				
0	44 (65%)	49 (59%)	46 (68%)	56 (69%)
1	24 (35%)	34 (41%)	22 (32%)	25 (31%)
Black				
0	31 (46%)	46 (55%)	28 (41%)	38 (47%)
1	37 (54%)	37 (45%)	40 (59%)	43 (53%)
Hisp				
0	63 (93%)	79 (95%)	64 (94%)	76 (94%)
1	5 (7.4%)	4 (4.8%)	4 (5.9%)	5 (6.2%)
income				
1	25 (37%)	30 (37%)	26 (38%)	29 (36%)
2	16 (24%)	17 (21%)	14 (21%)	21 (26%)
3	8 (12%)	13 (16%)	14 (21%)	11 (14%)
4	12 (18%)	12 (15%)	8 (12%)	6 (7.5%)
5	6 (9.0%)	10 (12%)	6 (8.8%)	13 (16%)
education				
1	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)
2	3 (4.4%)	7 (8.4%)	2 (2.9%)	4 (4.9%)
3	23 (34%)	15 (18%)	11 (16%)	27 (33%)
4	22 (32%)	32 (39%)	38 (56%)	24 (30%)
5	19 (28%)	29 (35%)	17 (25%)	26 (32%)

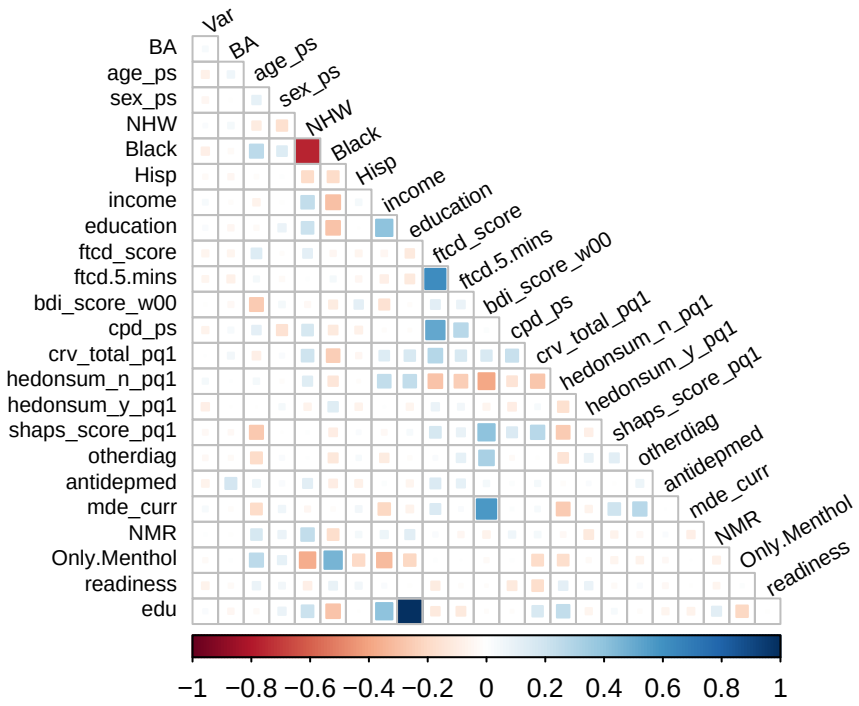
[1] Mean (SD); n (%)

Table 2: Measured Variables by Abstinance

Characteristic	**0** N = 236	**1** N = 64
ftcd_score	5 (2)	4 (2)
ftcd.5.mins		
0	123 (52%)	39 (61%)
1	113 (48%)	25 (39%)
bdi_score_w00	19 (12)	17 (11)
cpd_ps	16 (8)	14 (8)
crv_total_pq1	7 (4)	7 (4)
hedonsum_n_pq1	22 (19)	25 (22)
hedonsum_y_pq1	26 (19)	23 (20)
shaps_score_pq1	2 (3)	2 (2)
otherdiag		
0	127 (54%)	40 (63%)
1	109 (46%)	24 (38%)
antidepmed		
0	171 (72%)	47 (73%)
1	65 (28%)	17 (27%)
mde_curr		
0	114 (48%)	39 (61%)
1	122 (52%)	25 (39%)
NMR	0.35 (0.22)	0.42 (0.26)
Only.Menthol		
0	91 (39%)	29 (45%)
1	143 (61%)	35 (55%)
readiness		
3	1 (0.4%)	0 (0%)
4	5 (2.2%)	0 (0%)
5	24 (11%)	11 (19%)
6	68 (30%)	15 (25%)
7	53 (24%)	18 (31%)
8	61 (27%)	13 (22%)
9	6 (2.7%)	1 (1.7%)
10	6 (2.7%)	1 (1.7%)

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

Figure 1. Correlation Between Covariates



Checking missingness in the data is another important consideration in data pre-processing. In this data set, there is low missingness, with only 7 of the 25 variables containing missing values. For each of these variables, the missingness is less than 6%, as seen in Table 3. In a large data set, this level of missingness would likely be inconsequential, and complete-case analysis would be unlikely to bias the results. However, given the sample size of 300 in this case, it would be advantageous to retain those observations to make use of every observation in the sample. Therefore, we will perform multiple imputation using the `mice` package to create 5 complete data sets for analysis. Multiple imputation has been shown to have lower bias than other imputation methods, which is why we choose it for this analysis. After imputation, we note there are no longer missing values in any of the imputed data sets.

Table 3: Variables with Missingness

Variable	Proportion
income	0.0100
ftcd_score	0.0033
crv_total_pq1	0.0600
shaps_score_pq1	0.0100
NMR	0.0700
Only.Menthol	0.0067
readiness	0.0567

Modeling Methods

Before we begin model fitting, some considerations are testing the model and properly using our imputation. A common way to ensure data for model validation is to split the data set into a training and test set. We will do so using 75% of the data in the training set, and 25% in the test set, randomly sampling to obtain this split. We then fit the model on the training set and validate it on the test set. However, after multiple

imputation, we actually have 5 full data sets to analysis instead of just 1. Therefore, each of the imputed data sets gets split in the same way, so that the training set for each data set contains the same observations. We fit the model on each of these training sets and then pool the coefficient estimates at the end by taking the mean. We then validate the pooled coefficients on the test set.

For our analysis, the outcome of interest is the binary indicator for smoking abstinance `abst`. This means that we need to use a logistic regression model to fit this outcome. More specifically, we are interested in moderators of the effects of the behavioral treatment and baseline predictors of abstinance. This means that we are interested in any significant main effects and interactions between baseline covariates and the behavioral treatment variable. The upper limit of possible variables in the model is then quite large, which makes variable selection important if we are to obtain a parsimonious model. Regularized regression, such as best subset, lasso, and ridge regression are all methods that can assist in variable selection without the need for an additional algorithm, step or otherwise.

Best subset regression uses the L0 penalty, which puts a constraint on the number of non-zero coefficients. Lasso uses the the L1 penalty, which puts a constraint on the sum of the absolute values of the coefficients. Ridge uses the L2 penalty, which puts a constraint on the sum of the squared coefficients. In each of these cases, the penalty term is controlled by a lambda value, which increases the penalty as it gets bigger. As the penalty increases, best subset regression will remove variables from the model, lasso will snap coefficient values to zero, and ridge will force coefficients closer to zero, but they will never reach zero. Because a parsimonious model will be most helpful in assessing significant predictors and moderators, we want to use method that will actually reduce the number of coefficients, which rules out ridge regression. Both best subset and lasso will drop variables, but best subset is a stricter penalty and also more computationally demanding in practice, so lasso regression is a nice middle ground to use. Therefore, we will fit a logistic regression with a lasso penalty to the data.

An important step in fitting a lasso regression is selecting the optimal lambda value to use. To do so, we will use 10-fold cross-validation to determine the lambda value that minimizes the error of the model. We then refit the lasso model to the training set using this optimal lambda. We will include interactions between behavioral treatment and all other covariates, which will allow lasso to force all non-significant interactions to zero. As previously stated, we will perform this procedure on each of the 5 imputed data sets and then average the coefficients from each of the 5 models.

To evaluate the model, we will assess its performance on both the training data and the test data. We will assess and visualize its accuracy and discrimination using AUC and a ROC curve. We will also evaluate the model's calibration by plotting the predicted and observed probabilities for the test data. The calibration plot will also include the ideal line where the observed and predicted values are equal to better compare the performance of the model.

Modeling Results

Table 4 shows the pooled coefficients and corresponding odds ratios from each of the 5 models. Counting the levels from the categorical covariates and their interactions, there were 56 possible coefficients for the model, which the lasso regression shrunk to 11, excluding the intercept term. There are 2 interactions remaining in the model: the interaction between behavioral treatment and currently having MDD, and the interaction between behavioral treatment and consuming only menthol cigarettes. This means that currently having MDD and smoking only menthol cigarettes are potential moderators for the effects of the behavioral treatment. The remaining main effects in the model are predictors of abstinence.

Table 4: Pooled Coefficient Estimates From Lasso

	coef	OR
(Intercept)	-0.9799	0.3753
Var1	0.8200	2.2706
NHW1	0.3035	1.3546
edu4	-0.1673	0.8460

	coef	OR
ftcd_score	-0.1537	0.8575
crv_total_pq1	0.0073	1.0073
shaps_score_pq1	-0.0105	0.9895
mde_curr1	-0.1650	0.8479
NMR	0.0409	1.0417
readiness	-0.0027	0.9973
BA1:mde_curr1	-0.1083	0.8973
BA1:Only.Menthol1	-0.0038	0.9962

Figure 2 shows the ROC curves for the fitted values on both the test and the training data. In both cases, we see that the model performs better than 45 degree representing random prediction. The AUC on the training data is 0.78 and on the test data is 0.76.

Figure 3 shows the calibration of the model on the test data. The red line is the ideal line where the observed and predicted values are the same. The blue line shows calibration with linear smoothing, and the black line shows the calibration with `loess` smoothing, which allows more flexibility. The grey area represents the confidence interval for the `loess` smoothing. Both types of smoothed curves roughly follow the red line, indicating adequate calibration.

Figure 2. Lasso Model ROC

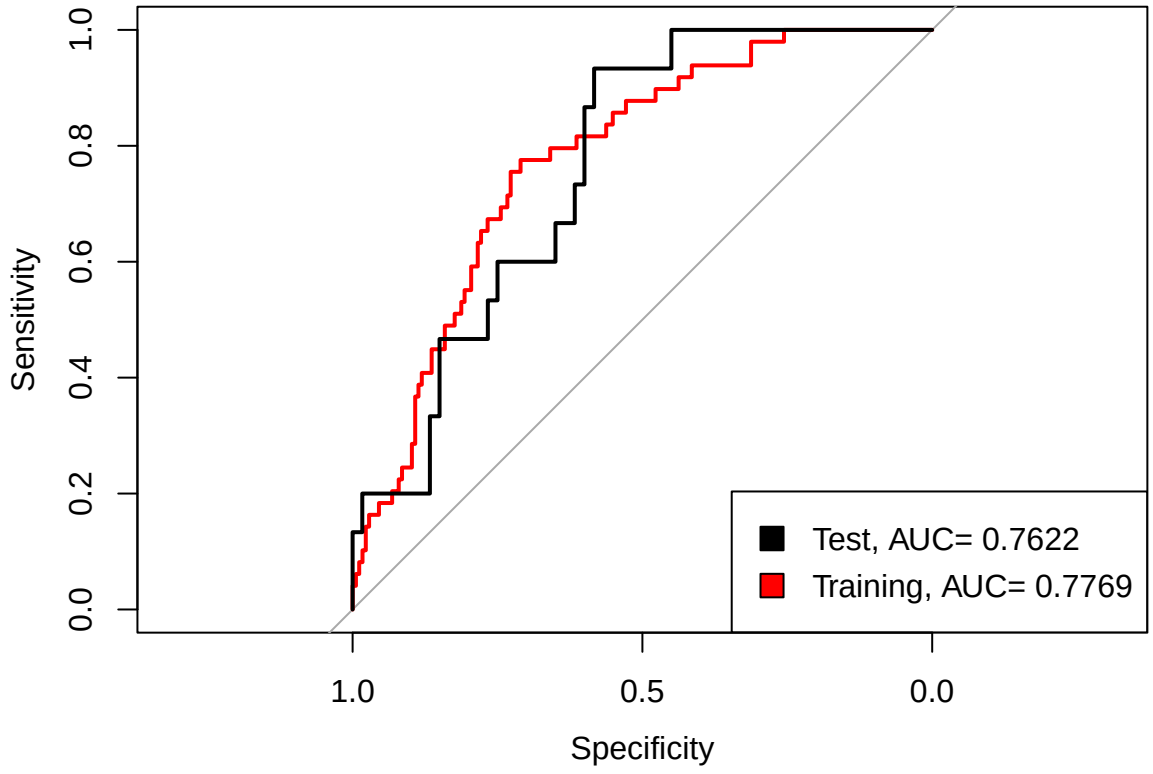
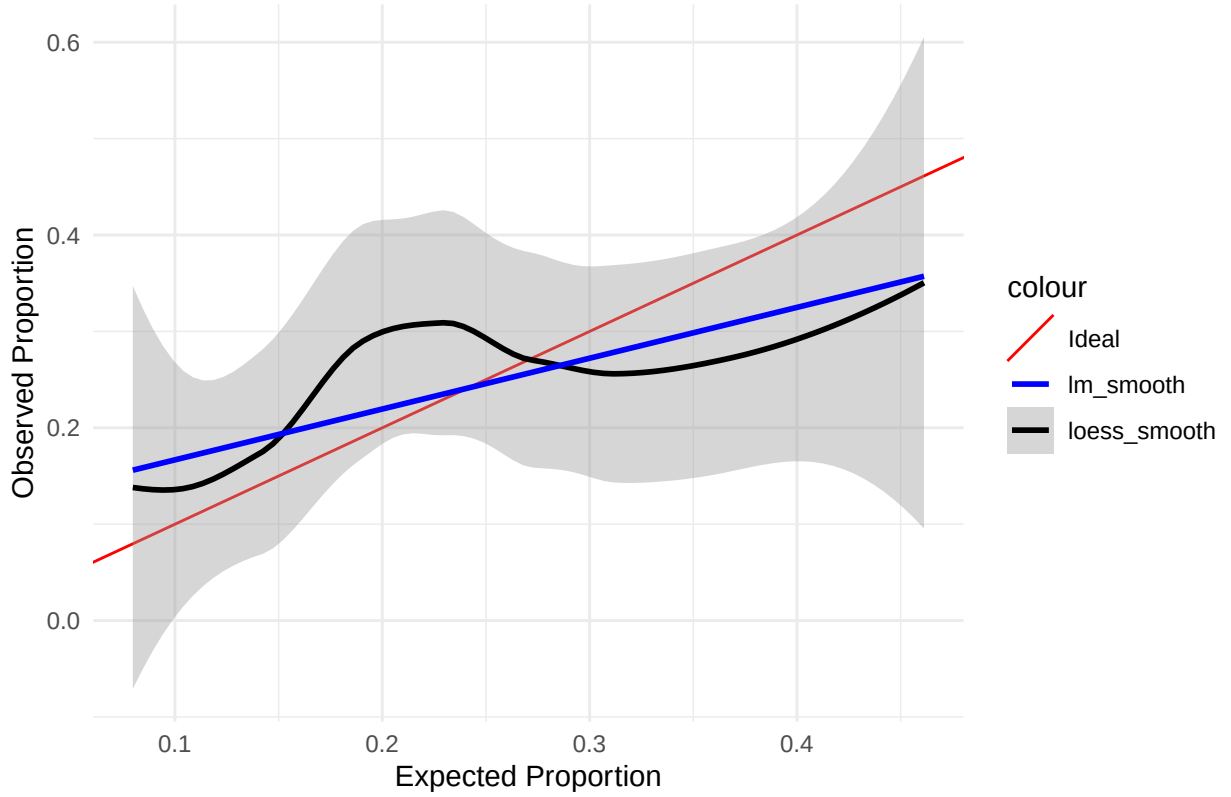


Figure 3. Lasso Calibration Plot on Test Data



Discussion

In our results, we saw that the lasso penalty shrunk the model from a possible 56 terms down to 12 total terms, indicating a successful variable selection procedure for a parsimonious model. We found that currently having MDD and smoking only menthol cigarettes are potential moderators for the behavioral treatment effect. Both of these interaction terms had negative coefficients, indicating that the behavioral treatment was less effective for only menthol smokers and current MDD patients. This aligns with existing evidence that people with depression have a harder time quitting smoking, and that menthol cigarettes are more addictive than normal cigarettes. Although these are moderators according to the model, it is also important to note the magnitude of the coefficients, which are both small. Because the interaction coefficients are so small, although they were significant enough to include in the model, they may not be large enough to be clinically significant in practice. However, it provides a starting point for further investigation into moderators for the treatment effect.

The logistic model uses the logit link function, so we can interpret the other coefficients as odds ratios after exponentiation. The baseline covariates remaining in the model are non-hispanic white, education at level 4, FTCD score, cigarette reward value, anhedosis, current MDD, NMR, and readiness to quite, as well as the varenicline treatment indicator. The odds of abstinance are 2.27 times higher for a person on varenicline compared to a person not on varenicline, controlling for the other variables left in the model. Being non-hispanic white increases the odds of quitting by 1.35 times, controlling for the other variables. Having some college or technical school education decreases the odds of quitting by 0.85 times, controlling for the other variables. For a unit increase in FTCD score, the odds of abstinence are 0.86 times lower, controlling for the other variables. The other coefficients all have similar interpretations. Based on this model, varenicline, being white, higher cigarette reward value, and higher NMR results in improved odds of quitting, whereas some college or technical school education, higher FTCD score, higher anhedosis score, currently having MDD, and higher quit readiness decrease the odds of abstinence. Most of these are not surprising, although we may have expected something different for education. Education at level 4 is the second highest education

level, and it is reasonable to think that more education would make people understand more strongly why they should quit. On the other hand, level 4 indicates incomplete higher education, which means that these people dropped out or failed to graduate. It is not unlikely that MDD is a factor in why a person didn't graduate, and we have seen that MDD decreases the likelihood of quitting smoking. As with the interaction terms, some of the main effect coefficients are very small, and although they were included in the model, they aren't necessarily clinically significant.

Note that the main effect for the behavior treatment was dropped from the model. We typically always want to include main effects in a model when we include their interactions, which is not the case here. Because we don't know the effect of the behavioral treatment outside of the interactions, our conclusions on potential moderators may not be true. Additionally, this means that our coefficient interpretations did not control for the behavior treatment, which could affect their significance as predictors. Therefore, a clear next step would be to refit the model excluding the behavioral treatment term from the penalty, which will always keep it in the model.

In terms of model performance, the ROC curves and AUC values are relatively large, indicating decent discrimination and accuracy on both the training data and the test data. The calibration plot also shows that the smoothed curves roughly follow the ideal line, indicating no glaring lack of model calibration.

Other than the aforementioned exclusion of the behavioral treatment term from the model, another limitation is the sample size of the data. Because we only have 300 total observations, and only 75% are used for model training, the results are likely to be dependent on the seed choice for the random sampling. Not only could this affect the coefficients included in the model and their magnitude, it could also affect the model performance and evaluation measures, for better or worse. To address this issue, future work could include additional cross-validation or bootstrap sampling methods in the model fitting to get a better idea of the model's true performance and generalizability.

The pooling of the coefficients across the 5 imputed data sets may also be a limitation. The models fit to each of the 5 data sets did not always include the same variables, so some zeros were included in the mean calculations. This could be a source of bias, and it is unclear whether this was the best way to handle this. Another possibility would be to have excluded those zero values, which reduces the denominator of the mean by 1, although that may upweight the pooled coefficient too much. This is an area requiring further investigation and literature review to determine what the proper pooling method would be.

Conclusion

Our modeling results found that currently having MDD and smoking only menthol cigarettes were possible moderators for the effect of behavioral treatment for smoking cessation in patients with depression. Additionally, varenicline, race, education, nicotine dependence, anhedosis, NMR, quit readiness, current MDD, and cigarette reward value were predictors of abstinance. However, there are several limitations that could affect these conclusions and require further research.

References

[1] Hitsman B, Papandonatos GD, Gollan JK, Huffman MD, Niaura R, Mohr DC, Veluz-Wilkins AK, Lubitz SF, Hole A, Leone FT, Khan SS, Fox EN, Bauer AM, Wileyto EP, Bastian J, Schnoll RA. 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. 2023 Sep;118(9):1710-1725. doi: 10.1111/add.16209. Epub 2023 May 3. Erratum in: Addiction. 2024 Sep;119(9):1669. doi: 10.1111/add.16609. PMID: 37069490.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE,
                      fig.pos = "H" ,
                      fig.align = 'center')
library(tidyverse)
library(glmnet)
library(gtsummary)
library(gt)
library(knitr)
library(kableExtra)
library(mice)
library(pROC)
library(corrplot)

# read data
dat<-read.csv("project2.csv")

# rename income variable
names(dat)[10]<-"income"
names(dat)[11]<-"education"

#make sex binary
dat$sex_ps<-dat$sex_ps-1

# create treatment status variable
dat<-dat%>%mutate(trt=case_when(Var==1 & BA==0~"Drug Only",
                                Var==0 & BA==1~"Behavioral Only",
                                Var==1 & BA==1~"Both",
                                TRUE~"Control"))

# factor categoricals
dat$income<-factor(dat$income)
dat$edu<-factor(dat$education)
dat$sex_ps<-factor(dat$sex_ps)
dat$abst<-factor(dat$abst)
dat$Var<-factor(dat$Var)
dat$BA<-factor(dat$BA)
dat$NHW<-factor(dat$NHW)
dat$Black<-factor(dat$Black)
dat$Hisp<-factor(dat$Hisp)
dat$ftcd.5.mins<-factor(dat$ftcd.5.mins)
dat$otherdiag<-factor(dat$otherdiag)
dat$antidepmed<-factor(dat$antidepmed)
dat$mde_curr<-factor(dat$mde_curr)
dat$Only.Menthol<-factor(dat$Only.Menthol)


# group columns by demographic and medical details for separate tables
demographic<-names(dat)[5:11]
med<-names(dat)[12:25]
```

```r
# demographic table
t1<-tbl_summary(select(dat, c(all_of(demographic), trt)), by=trt, missing="no",
            statistic = list(all_continuous()~ "{mean} ({sd})",
                             all_categorical()~ "{n} ({p}%)"))%>%
  modify_header(all_stat_cols() ~ "**{level}**  \nN = {n}")%>%
  modify_caption("Demographics by Treatment Type")%>%
  as_kable_extra(booktabs=TRUE)%>%kable_styling(font_size = 12)

t1
# measurements table
t2<-tbl_summary(select(dat, c(all_of(med), abst)), by=abst, missing="no",
            statistic = list(all_continuous()~ "{mean} ({sd})",
                             all_categorical()~ "{n} ({p}%)"))%>%
  modify_caption("Measured Variables by Abstinance")%>%
  as_kable_extra(booktabs=TRUE)%>%kable_styling(font_size = 12)

t2
# Create correlation matrix
cor_mat <- dat%>%
  select(-c(id, abst, trt)) %>% apply(2, as.numeric)%>%
  cor(use = "complete.obs")

# Plot correlation matrix
corrplot(cor_mat, method = "square", type = "lower", diag = FALSE,mar=c(0,0,2,0),
         tl.cex = 0.7, tl.col = "black", tl.srt = 30, title = " Figure 1. Correlation Between Covariates

# deal with missing data
num_missing<-function(data){
  #' Returns the number of missing values in the given data
  #' @param data, data vector or data frame of any type
  #' @return the number of NA values in the data

  return(length(which(is.na(data))))
}

# table of missingness proportion
prop<-apply(dat, 2, num_missing)/nrow(dat)
# non-zero missingness
prop<-prop[prop>0]%>%round(digits=4)
kable(prop, caption = "Variables with Missingness",
      col.names = c("Variable", "Proportion"))

# run and save multiple imputation
#dat_mice<-mice(select(dat, -c(id, trt)), seed=1)
#saveRDS(dat_mice, "PDA2Mice")

# read in imputed data
dat_mice<-readRDS("PDA2Mice.txt")

set.seed(2550)
# split data into train/validate set
train_id<-sample(1:300, 300*.75)
```

```r
# initialize objects to store lambdas and coefficients
lambda<-c()
coefficients<-list()

for (i in 1:5){
  # extract imputed dataset and split data
  dat<-mice::complete(dat_mice,i)
  train<-dat[train_id,]
  test<-dat[-train_id,]

  # create model matrix and outcome
  x <- model.matrix(abst~BA*(Var+age_ps+sex_ps+NHW+Black+Hisp+income+edu+
                               ftcd_score+ ftcd.5.mins+ bdi_score_w00+cpd_ps+
                               crv_total_pq1+hedonsum_n_pq1+shaps_score_pq1+
                               otherdiag+antidepmed+mde_curr+NMR+Only.Menthol+
                               readiness), data=train)[,-1]
  y <- train$abst

  # use cv to find optimal lambda for lasso
  lasso <- cv.glmnet(x, y, alpha=1, family="binomial")
  l<-lasso$lambda.min

  # fit lasso with optimal lambda
  lasso.fit <- glmnet(x, y, alpha=1, lambda=l, family="binomial")

  # extract coefficients and save
  coefs<-coef(lasso.fit)
  lambda<-c(lambda, l)
  coefficients[i]<-coefs

}

# get data frame of coefficients
coefficients<-as.data.frame(as.matrix(cbind(coefficients[[1]], coefficients[[2]], coefficients[[3]], co
# get mean across imputations
pooled_coef<-rowMeans(coefficients)

# extract non-zero coefficients
no0coef<-pooled_coef[pooled_coef!=0]

# return coefs and odds ratios
kable(data.frame(coef=no0coef, OR=exp(no0coef)), caption="Pooled Coefficient Estimates From Lasso", dig
# predict on training data
pred_train<-plogis(as.numeric(x%*%pooled_coef[2:56]+pooled_coef[1]))

# create model matrix for test data
x_test <- model.matrix(abst~BA*(Var+age_ps+sex_ps+NHW+Black+Hisp+income+edu+
                               ftcd_score+ ftcd.5.mins+ bdi_score_w00+cpd_ps+
                               crv_total_pq1+hedonsum_n_pq1+shaps_score_pq1+
                               otherdiag+antidepmed+mde_curr+NMR+Only.Menthol+
                               readiness), data=test)[,-1]
# predict on test data
pred_test<-plogis(as.numeric(x_test%*%pooled_coef[2:56]+pooled_coef[1]))
```

```r
# roc curve for training data
roc_train<-roc(dat$abst[train_id], pred_train)
roc_test<-roc(dat$abst[-train_id], pred_test)

# plot roc curves
plot(roc_train, col="red", main="Figure 2. Lasso Model ROC")
plot(roc_test, add=TRUE)
legend("bottomright",
       legend = c(paste("Test, AUC=", as.character(round(roc_test$auc, 4))),
                  paste("Training, AUC=", as.character(round(roc_train$auc, 4)))),
       fill = c("black", "red"))
### calibration plot
num_cuts <- 50

test_calib<-data.frame(prob = pred_test,
                       bin = cut(pred_test, breaks = num_cuts),
                       # converting to numeric from factor makes it 1,2, so
                       # subtract 1 to get binary
                       class = as.numeric(train$abst)-1)
test_calib <- test_calib %>%
             group_by(bin) %>%
             summarize(observed = sum(class)/n(),
                       expected = sum(prob)/n(),
                       se = sqrt(observed * (1-observed) / n()))

cols <- c("Ideal"="red","loess_smooth"="black","lm_smooth"="blue")
ggplot(test_calib) +
  geom_abline(aes(intercept = 0, slope = 1, color="Ideal")) +
  geom_smooth(aes(x = expected, y = observed, color="loess_smooth"), se=TRUE)+
  geom_smooth(aes(x = expected, y = observed, color="lm_smooth"), se=FALSE, method="lm")+
  scale_color_manual(values=cols)+
  labs(x = "Expected Proportion", y = "Observed Proportion", title="Figure 3. Lasso Calibration Plot on
  theme_minimal()
```