

Evaluating the Impact of Behavioral Activation on Smoking Cessation in Major Depressive Disorder

Zhaoxiang Ding

December 13, 2024

Abstract

Behavioral Activation (BA) has been proposed as a treatment to support smoking cessation in individuals with Major Depressive Disorder (MDD), though evidence regarding its effectiveness remains limited. This study re-evaluates the impact of BA on smoking cessation, utilizing data from a prior 2x2 factorial randomized controlled trial. Our primary objectives are to investigate potential moderators of BA's effect on end-of-treatment (EOT) smoking abstinence and to identify baseline predictors of abstinence while accounting for pharmacotherapy. We first use penalized regression models to select variables and interaction terms, then fit a logistic regression model to estimate the causal effect of BA on smoking cessation. Our results confirm that BA does not significantly influence smoking cessation, consistent with prior findings. However, several baseline predictors, including Non-Hispanic White and FTCD score at baseline, demonstrate strong associations with abstinence. Additionally, Age moderates the effect of pharmacotherapy on cessation. While this model provides valuable insights, limitations such as sample size constraints and potential model dependency suggest a need for further research.

Introduction

Behavioral Activation (BA) is regarded as a promising intervention for aiding smoking cessation in individuals diagnosed with Major Depressive Disorder (MDD), who are known to be more likely to smoke heavily, exhibit greater nicotine dependence, and experience more severe withdrawal symptoms than those without MDD (Hitsman et al. 2023). However, there has been limited research examining the specific impact of BA on smoking cessation outcomes. In one study, Hitsman et al. (2023) employed a 2x2 randomized factorial design (BA versus standard behavioral treatment (ST) and varenicline versus placebo, which is effective drug to help

smoking cessation) to investigate this effect, concluding that BA did not significantly improve smoking cessation outcomes. This project aims to revisit the influence of BA on smoking cessation, using the same dataset from Hitsman et al. (2023) but with an alternative analytical approach.

Our objective is to explore potential moderators influencing the effectiveness of behavioral treatment on end-of-treatment (EOT) abstinence rates. Additionally, we seek to identify baseline characteristics that may predict abstinence, while accounting for the effects of behavioral treatment and pharmacotherapy (varenicline).

Data collection

The randomized, placebo-controlled trial recruited 300 adult daily smokers with current or past major depressive disorder (MDD) across research clinics at Northwestern University and the University of Pennsylvania. Initial eligibility screening was conducted via telephone, followed by final eligibility screening, informed consent, randomization, and baseline assessment at an intake session. Randomization, stratified by clinical site, sex, and depression severity, ensured balanced treatment arms. Participants were allocated to one of four groups using a computer-based system with unequal block sizes to maximize assignment to varenicline arms. Behavioral treatment sessions were standardized to eight 45-minute sessions over 12 weeks, with medication (varenicline or placebo) administered according to FDA-approved guidelines. Adherence and outcomes were monitored through multiple in-person and remote assessments over 27 weeks, with bio-verified smoking abstinence as a primary measure. Participants received compensation for participation and travel to enhance study retention and compliance.

A total of 25 variables were collected in the study, including demographic information, smoking history, and psychological assessments. Variable details are provided in Table 1. The primary outcome of interest was end-of-treatment (EOT) abstinence, defined as self-reported 7-day point prevalence abstinence confirmed by expired carbon monoxide (CO) levels of ≤ 10 ppm. Secondary outcomes included continuous abstinence, time to first lapse, and time to first relapse.

Table 1: An overview of participant characteristics

Variable	Description
abst	Smoking Abstinence
Var (Varenicline)	Pharmacotherapy
BA (Behavioral Activation)	Psychotherapy
age_ps	Age at phone interview
sex_ps	Sex at phone interview
NHW	Non-Hispanic White indicator
Black	Black indicator

Table 1: An overview of participant characteristics

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

Identifying the causal effect

In Hitsman et al. (2023)’s work, the causal effect was estimated by comparing the abstinence rates between the treatment and control groups, following an intent-to-treat (ITT) approach. Rather than solely examining abstinence rates, this project evaluates the odds of abstinence and estimates the causal effect using the odds ratio of abstinence between the treatment and control groups, while adhering to ITT principles. The causal effect, denoted as $\hat{\tau}$, can be formulated as follows:

$$\hat{\tau} = \frac{odds(E[Y^1])}{odds(E[Y^0])}$$

Where Y^1 and Y^0 represent the potential outcomes under treatment and control, respectively.

The dataset originates from a 2x2 randomized factorial design, theoretically balancing all covariates across treatment and control groups. As examined in Table 2, baseline covariate distributions confirm that randomization was successful, as covariates appear balanced between groups. The table also reveals some missing data. While the data may not be missing at

random, the low proportion of missing records suggests a minimal impact on the results, allowing us to treat the missingness as random. Therefore, the assumptions for identifying causal effects are met, and we express the causal effect as:

$$\begin{aligned}\hat{\tau} &= \frac{odds(E[Y^1])}{odds(E[Y^0])} \\ &= \frac{odds(E[Y|A = 1])}{odds(E[Y|A = 0])}\end{aligned}$$

Where A is the whether having behavioral treatment or not. A naive approach to estimate the causal effect is to fit a logistic regression model with the treatment group as the only covariate. The causal effect can be estimated by the coefficient of the treatment group. However, this approach does not consider the potential interaction between the treatment group and other covariates, nor the moderation of other variables. In this project, we fit a logistic regression model with the treatment group and the interaction terms between the treatment group and other covariates.

Except for identifying the causal effect, the logistic regression model can also be used to examine the potential moderators of the effect of behavioral treatment on end-of-treatment (EOT) abstinence and evaluate baseline variables as predictors of abstinence, controlling for behavioral treatment and pharmacotherapy.

Exploratory Data Analysis

Table 2 summarized the distribution of variables among different groups. Except showing a successful randomization as discussed in the previous section, the table also shows that the majority of subjects did not quit smoking by the end of study.

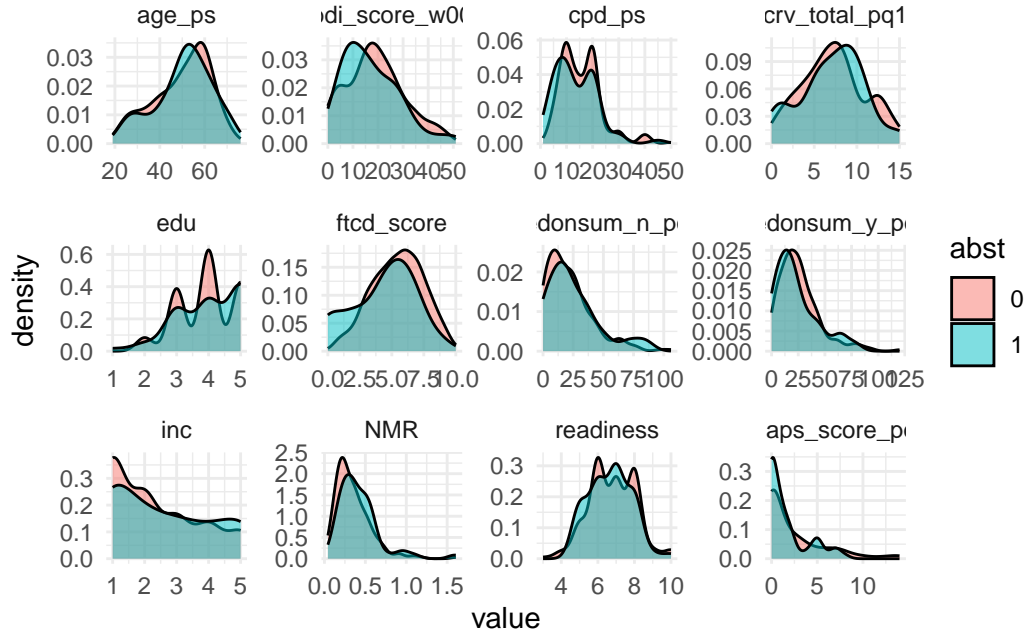
In order to examine the distribution of each variable and their relations with the outcome variable, we first plot the distribution of continuous variables and categorical variables. As shown in Figure 1, most continuous variables' distributions have little change between different outcomes, except that the peak of **age** and **bdi_score** is lower among subjects who quit smoking (**abst** = 1). The distribution of binary variables also shows little difference between different outcomes, except that among those who take Varenicline, more people quit smoking. It is worth to notice that the distribution of **BA** among different outcomes is basically the same, bringing the question of whether **BA** has an effect on smoking cessation.

The correlation among variables are also examined, shown in Table 3. The table shows no variables' VIF bigger than 5, indicating that there is no multicollinearity issue among variables.

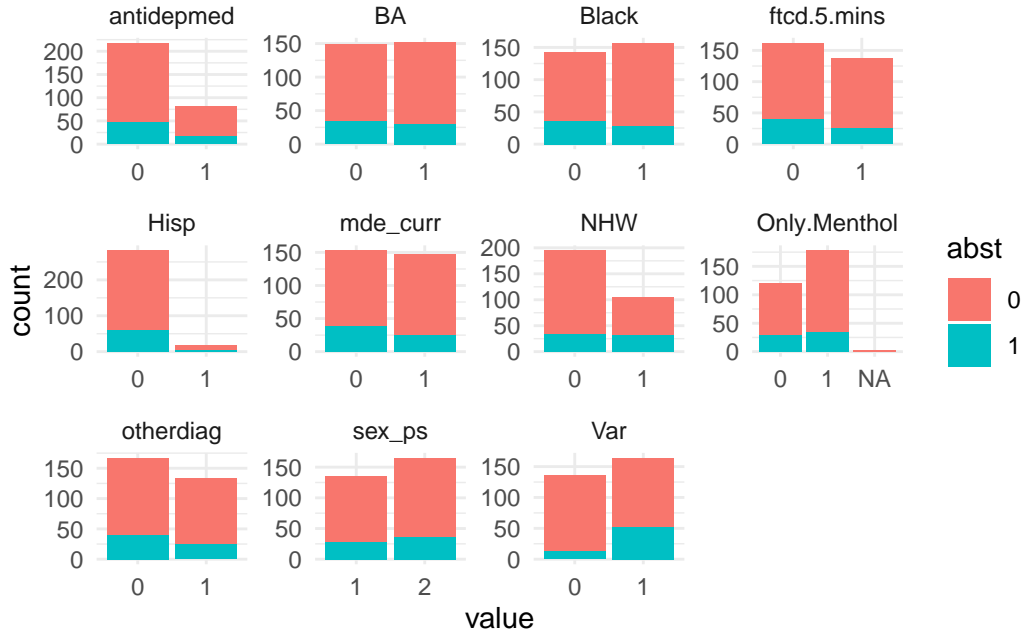
Table 2: Summary of the data

Characteristic	BA + Placebo N = 81	BA + Varenicline N = 83	Control N = 68	ST + Varenicline N = 68
abst				
0	55 (68%)	57 (69%)	60 (88%)	64 (94%)
1	26 (32%)	26 (31%)	8 (12%)	4 (5.9%)
age_ps	52 (41, 59)	53 (40, 60)	51 (45, 58)	54 (42, 61)
sex_ps				
1	37 (46%)	39 (47%)	29 (43%)	30 (44%)
2	44 (54%)	44 (53%)	39 (57%)	38 (56%)
NHW				
0	56 (69%)	49 (59%)	46 (68%)	44 (65%)
1	25 (31%)	34 (41%)	22 (32%)	24 (35%)
Black				
0	38 (47%)	46 (55%)	28 (41%)	31 (46%)
1	43 (53%)	37 (45%)	40 (59%)	37 (54%)
Hisp				
0	76 (94%)	79 (95%)	64 (94%)	63 (93%)
1	5 (6.2%)	4 (4.8%)	4 (5.9%)	5 (7.4%)
inc	2.00 (1.00, 3.00)	2.00 (1.00, 4.00)	2.00 (1.00, 3.00)	2.00 (1.00, 4.00)
Unknown	1	1	0	1
edu	4.00 (3.00, 5.00)	4.00 (3.00, 5.00)	4.00 (4.00, 4.50)	4.00 (3.00, 5.00)
ftcd_score	5.00 (4.00, 7.00)	5.00 (4.00, 7.00)	6.00 (4.00, 7.00)	5.00 (4.00, 7.00)
Unknown	0	0	1	0
ftcd.5.mins				
0	43 (53%)	50 (60%)	33 (49%)	36 (53%)
1	38 (47%)	33 (40%)	35 (51%)	32 (47%)
bdi_score_w00	18 (11, 27)	18 (10, 25)	18 (12, 25)	18 (9, 27)
cpd_ps	15 (10, 20)	15 (10, 20)	13 (10, 20)	15 (10, 20)
crv_total_pq1	7.0 (5.0, 9.0)	8.0 (4.5, 10.0)	7.0 (4.5, 9.0)	7.0 (5.0, 10.0)
Unknown	6	3	8	1
hedonsum_n_pq1	20 (9, 35)	20 (9, 32)	14 (9, 27)	21 (10, 31)
hedonsum_y_pq1	21 (13, 34)	17 (11, 31)	25 (12, 38)	23 (14, 34)
shaps_score_pq1	1.00 (0.00, 3.00)	1.00 (0.00, 4.00)	1.00 (0.00, 5.00)	0.00 (0.00, 3.00)
Unknown	0	0	1	2
otherdiag				
0	41 (51%)	53 (64%)	40 (59%)	33 (49%)
1	40 (49%)	30 (36%)	28 (41%)	35 (51%)
antidepmed				
0	66 (81%)	59 (71%)	53 (78%)	40 (59%)
1	15 (19%)	24 (29%)	15 (22%)	28 (41%)
mde_curr				
0	37 (46%)	43 (52%)	37 (54%)	36 (53%)
1	44 (54%)	40 (48%)	31 (46%)	32 (47%)
NMR	0.29 (0.20, 0.51)	0.33 (0.22, 0.50)	0.32 (0.20, 0.43)	0.32 (0.23, 0.46)
Unknown	9	3	2	7
Only.Menthhol				
0	34 (42%)	34 (41%)	24 (36%)	28 (41%)
1	47 (58%)	48 (59%)	43 (64%)	40 (59%)
Unknown	0	1	1	0
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)
Unknown	4	5	4	4

¹ n (%); Median (Q1, Q3)



(a) Distribution of continuous variables, stratified by outcome (abst)



(b) Distribution of binary variables, stratified by outcome (abst)

Figure 1: Relationship between continuous variables and the outcome variable

Table 3: Generalized variance inflation factor of the data

	VIF
age_ps	1.422728
sex_ps	1.145365
NHW	4.462106
Black	4.736099
Hisp	1.694604
inc	1.468204
edu	1.469299
ftcd_score	2.763104
ftcd.5.mins	1.953346
bdi_score_w00	2.003905
cpd_ps	1.660048
crv_total_pq1	1.569717
hedonsum_n_pq1	1.760080
hedonsum_y_pq1	1.232178
shaps_score_pq1	1.428846
otherdiag	1.282832
antidepmed	1.124861
mde_curr	1.696843
NMR	1.212831
Only.Menthol	1.551892
readiness	1.113905

There are missingness in the dataset, showed in Table 4. 7 out of 25 variables have missing values, with the highest missing percentage is 7%. Indicated that the missingness is not severe, and can be treated as random missingness.

Table 4: Missingness of the data

	Missing Percentage
inc	1.0000
ftcd_score	0.3333
crv_total_pq1	6.0000
shaps_score_pq1	1.0000
NMR	7.0000
Only.Menthol	0.6667
readiness	5.6667

Identifying the interactions

Given that many variables in the dataset are categorical, including all possible interaction terms would produce more terms than the sample size allows, leading to model overfitting. To avoid this, we manually select covariates for inclusion based on their VIF. VIF is a measure of how much the variance of a estimated regression coefficient is due to correlations among predictors. The higher the value, the less benefit of including the variable into interaction terms as the additional information brought by the interaction terms may have been already captured by the main effect of other predictors. We only select variables with VIF less than 1.3 to include in the interaction terms. The threshold is set by balancing the trade-off between including more variables and avoiding overfitting, and the gap of VIF values. The selected variables are `readiness`, `antidepmed`, `sex_ps`, `NMR`, `hedonsum_y_pq1` and `otherdiag`, with corresponding VIF values are 1.11 1.12 1.15 1.21 1.23 1.28. Only up to 2 way interaction terms are included in the model, as the number of interaction terms increase exponentially with the number of variables, and the sample size is limited. We also consider the interaction terms between the treatment group and other covariates, as the effect of the treatment group may be moderated by other variables.

Therefore, the model can be written as:

$$\text{logit}(E[Y_i]) = \beta_0 + \beta_1 A_i + \beta_2 Z_i + \beta_3 X_i^T + \beta_4 X_i^T A_i + \beta_5 X_i^T Z_i + \beta_6 Z_i A_i + \sum_{k=1}^K \beta_7 L_k L_{-k}^T$$

Where L_k is the k th variable with $\text{VIF} < 1.3$ showed in the previous paragraph, and L_{-k} is the rest of the variable with $\text{VIF} < 1.3$

Model selection

The primary goal of this project is to identify potential moderators of behavioral treatment effects on end-of-treatment (EOT) abstinence and to assess baseline predictors of abstinence while controlling for behavioral treatment and pharmacotherapy. We employed various variable selection techniques, including Lasso regression and subset selection with L0, L0L1, and L0L2 penalties, all implemented using 10-fold cross-validation.

To address missing values in the dataset, we applied multiple imputation before splitting the data into training and testing sets. This approach was selected due to the limited sample size, as performing imputation after data splitting could result in unreliable outcomes. The multiple imputation process was conducted using the `mice` package in R, generating a total of five imputed datasets. We randomly assigned 80% of the records to the training set and reserved the remaining 20% for testing. Models were fitted to the training set for each imputed dataset, and the coefficients were pooled to obtain the final results. Variables included in the final model were those retained in at least three out of the five imputed datasets.

For subset selection models using L0, L0L1, and L0L2 penalties, the λ and γ values were chosen to minimize the mean cross-validated error. The coefficients from each model were averaged across imputed datasets to obtain the final results. Using these pooled results, the model was then applied to the test dataset to evaluate its performance.

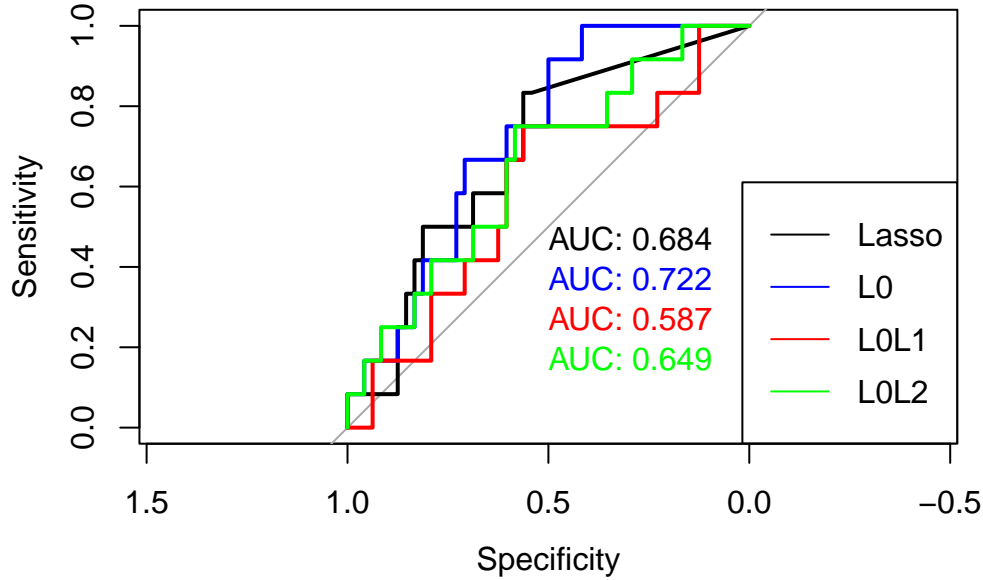


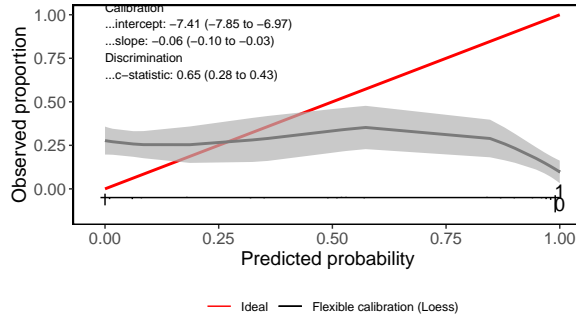
Figure 2: ROC curves of the models in test data

Figure 2 shows the all four models' performance in test data. Subset selection model with L0 penalty have the highest AUC among all models, indicating the model is robust.

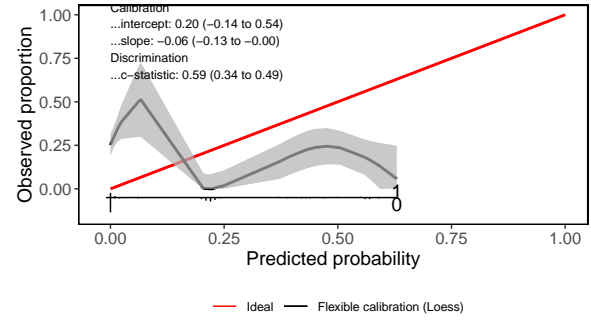
We also examine the calibration curve of the model in the test data, showed in Figure 3. The calibration curve of the L0 model is the closest to the 45 degree line, indicating the model is well calibrated. While the other model's calibration curve is largely deviated from the 45 degree line, indicating the model may not be well calibrated.

Based on the performance of the model in the test data, we use the model with L0 penalty to answer our research question. The model exclude all but 3 terms: `NHW1`, `ftcd_score` and `Var1:age_ps`. All terms with behaviour treatment are excluded from the model, indicating that the behaviour treatment has no effect on addressing smoking cessation among MDD patients.

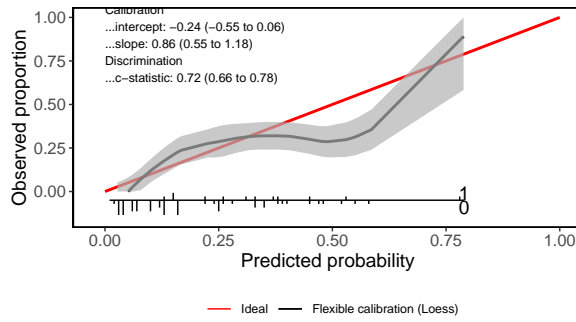
We fit a logistic regression model with the selected terms to examine the effect of the predictors on smoking cessation. As shown in Table 5, the model shows being a non hispanic white (`NHW = 1`) will increase the odds of quitting smoking by 161%, and one unit increase in FTCD score at baseline (`ftcd_score`) will decrease the odds of quitting smoking by 25%. The mean effect



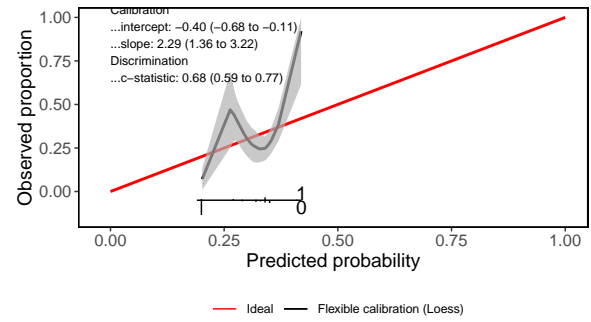
(a) Calibration curve of the L0L2 model



(b) Calibration curve of the L0L1 model



(c) Calibration curve of the L0 model



(d) Calibration curve of the Lasso model

Figure 3: Calibration curves of the models in test data

of **Var** on smoking cessation is negative, but considering the interaction terms, the effect of **Var** on smoking cessation positive when the subject is over 15 years old. Considering that the minimum age in the dataset is 19, exploring the effect of **Var** on smoking cessation for individuals under 19 would be an extrapolation of the model and is not meaningful. Taking varenicline(**Var** = 1) for a subject with 20 years of old will increase the odds of quitting smoking by 24%. When the subject is 40 years of old, the increase of odds is 177%. Indicating the effect is more significant among older people. However, as the p value of both the main effects and the interaction terms are not significant, the effect of **Var** on smoking cessation is still questionable.

Table 5

Characteristic	$\log(\text{OR})^1$	95% CI ¹	p-value
NHW			
0	—	—	
1	0.96	0.33, 1.6	0.003
ftcd_score	-0.28	-0.43, -0.14	<0.001
Var			
0	—	—	
1	-0.59	-3.2, 2.2	0.7
age_ps	-0.01	-0.06, 0.04	0.6
Var * age_ps			
1 * age_ps	0.04	-0.01, 0.10	0.10

¹OR = Odds Ratio, CI = Confidence Interval

Coefficients of the best model

There for, we can conclude that both non hispanic white (NHW) and FTCD score at baseline (**ftcd_score**) are predictors for smoking cessation, with being a non hispanic white will increase the odds of quitting smoking, and higher FTCD score at baseline will decrease the odds of quitting smoking. The effect of **Var** on smoking cessation is postive and moderated by **age**, with the older the subject is, the more significant the effect is. However, the effect of **Var** on smoking cessation is still questionable as the p value indicates the effect may not be significant.

Discussion

The result of this project shows that behaviour treatment has no effect on smoking cessation among MDD patients. This finding corresponds with the result of Hitsman et al. (2023), which also found that BA does not have a significant effect on smoking cessation. Possible explanations for this result has already been discussed in Hitsman et al. (2023).

Two important predictors are identified in this project. `ftcd_score` and `NHW` are predictors for smoking cessation. `Age` is identified as the moderator of the effect of `Var` on smoking cessation.

This project has several limitations. First of all, the model selected to use will have a large impact on the result. The subset selection model with L0 penalty is selected as the best model in this project, but other models may have different results. Second, the data set only contains limited sample size, making examining the effect of interaction terms difficult. As shown in the previous section, only limited interaction terms are included in the model while those interaction terms excluded from the model may also have significant effect.

In conclusion, this project re-evaluates the impact of BA on smoking cessation in MDD patients. The results confirm that BA does not significantly influence smoking cessation, consistent with prior findings. Two baseline predictors, including `ftcd_score` and `NHW`, demonstrate strong associations with abstinence, while `Age` moderates the effect of pharmacotherapy on cessation. Despite limitations, this model offers valuable insights into the predictors of smoking cessation in MDD patients, highlighting the need for further research to confirm these findings.

References

Hitsman, Brian, George D. Papandonatos, Jacqueline K. Gollan, Mark D. Huffman, Raymond Niaura, David C. Mohr, Anna K. Veluz-Wilkins, et al. 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* 118 (9): 1710–25. <https://doi.org/10.1111/add.16209>.

Code Appendix

```
# Example data frame representing the participant characteristics table
participant_characteristics <- data.frame(
  Variable = c("abst", "Var (Varenicline)", "BA (Behavioral Activation)", "age_ps", "sex_ps",
    "inc", "edu", "ftcd_score", "ftcd.5.mins", "bdi_score_pq1", "cpd_ps", "crv_tot",
    "hedonsum_y_pq1", "shaps_score_pq1", "otherdiag", "antidepmed", "mde_curr", "lived",
    "readiness"),
  Description = c("Smoking Abstinence", "Pharmacotherapy", "Psychotherapy", "Age at phone interview",
    "Non-Hispanic White indicator", "Black indicator", "Hispanic indicator", "Education (ordinal categorical, low to high)", "FTCD score at baseline", "BDI score at baseline", "Cigarettes per day at baseline phone survey", "Cigarettes per day at baseline - substitute reinforcers",
    "Pleasurable Events Scale at baseline - substitute reinforcers",
    "Pleasurable Events Scale at baseline - complementary reinforcers", "Anhedonia")
)
```

```

        "Taking antidepressant medication at baseline", "Current vs past MDD", "Ni
        "Exclusive Mentholated Cigarette User", "Baseline readiness to quit smoking
    )

knitr::kable(participant_characteristics)

library(gtsummary)
library(mice)
library(glmnet)
library(pROC)
library(kableExtra)
library(dplyr)
library(LOLearn)
library(ggpubr)
library(tidyr)
library(ggcorrplot)
library(parallel)
library(rms)
library(CalibrationCurves)
# Read data
data <- read.csv("../Data/project2.csv")
num_col <- c(5,12,14:19,23,25)
ordinal_col <- c(10,11)
data[,num_col] <- lapply(data[,num_col], as.numeric)
data[,ordinal_col] <- lapply(data[,ordinal_col], factor, order = T)

data[,-c(num_col, ordinal_col)] <- lapply(data[,-c(num_col, ordinal_col)], factor)
# summary table
data_tbl <- data[, -1]
data_tbl$group <- paste(data_tbl$Var, data_tbl$BA, sep = "_")
data_tbl$group <- case_when(data_tbl$group == "0_0" ~ "Control",
                           data_tbl$group == "1_0" ~ "BA + Placebo",
                           data_tbl$group == "0_1" ~ "ST + Varenicline",
                           data_tbl$group == "1_1" ~ "BA + Varenicline")
data_tbl <- data_tbl[, -c(2,3)]
data_tbl$inc <- as.numeric(data_tbl$inc)
data_tbl$edu <- as.numeric(data_tbl$edu)
tbl_summary(data_tbl, by = group, type = list(readiness ~ 'continuous',
                                              inc ~ 'continuous',
                                              edu ~ 'continuous')) %>%

  as_kable_extra(booktabs = TRUE) %>%
  kableExtra::kable_styling(latex_options = "scale_down")

```

```

# continuous variables
par(mfrow=c(2,5))
data_con <- data[,c(num_col)]
data_con$abst <- data$abst
data_ord <- data[,ordinal_col]
data_ord <- apply(data_ord, 2, as.numeric)
data_con <- cbind(data_con, data_ord)
data_cat <- data[, -c(num_col, ordinal_col)]
data_cat <- data_cat[, -1]

pivot_longer(data_con, cols = -abst) %>%
  ggplot(aes(x = value, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~name, scales = "free") +
  theme_minimal()

pivot_longer(data_cat, cols = -abst) %>%
  ggplot(aes(x = value, fill = abst)) +
  geom_bar() +
  facet_wrap(~name, scales = "free") +
  theme_minimal()

car::vif(glm(abst ~ . -group, data = data_tbl, family = "binomial")) %>%
  as.data.frame() %>%
  kable(col.names = 'VIF')

# missingness
missing_df <- apply(data, 2, function(x) sum(is.na(x)))
missing_df <- missing_df[missing_df != 0]
missing_df <- round(missing_df/nrow(data) * 100, 4)
kable(missing_df, col.names = "Missing Percentage")
vif_res <- car::vif(glm(abst ~ . -group, data = data_tbl, family = "binomial"))
vif_res <- as.data.frame(vif_res)
vif_res$var <- rownames(vif_res)
order_vif <- (vif_res[vif_res < 1.3,])
rownames(order_vif) <- NULL
order_vif <- order_vif[order(order_vif$vif_res),]
#multiple imputation
data_imp <- mice(data[, -1], m = 5, seed = 2550, printFlag = F)
data_imp <- complete(data_imp, action = 'all')
train_index <- sample(1:nrow(data), 0.8*nrow(data))

```

```

model_matrix_fun <- function(data, train_id){
  #' @param data: the data set
  #' @param train_id: the index of the training set
  #' @return: a list of model matrix and response variable
  X <- model.matrix(abst ~ . + (readiness + antidepmed + NMR + sex_ps +
                        hedonsum_y_pq1 + otherdiag)^2
                    + Var*(.) + BA*(.),
                    data = data[train_id,])

  X <- X[,-1]
  Y <- factor(data[train_id,]$abst)

  return(list(X,Y))
}

fit_regression <- function(data, train_id){
  #' @param data: the data set
  #' @param train_id: the index of the training set
  #' @return: a list of the coefficients of the model

  model_data <- model_matrix_fun(data, train_id)

  #model fitting
  lasso_fit <- cv.glmnet(model_data[[1]], model_data[[2]], family = "binomial",
                        alpha = 1, type.measure = "auc",
                        nfolds = 10)

  subset_fit <- L0Learn.cvfit(model_data[[1]], model_data[[2]],
                              nFolds=10, penalty="L0", loss = 'Logistic')
  subset01_fit <- L0Learn.cvfit(model_data[[1]], model_data[[2]],
                                nFolds=10, penalty="L0L1", loss = 'Logistic')
  subset02_fit <- L0Learn.cvfit(model_data[[1]], model_data[[2]],
                                nFolds=10, penalty="L0L2", loss = 'Logistic')

  var_names <- rownames(coef(lasso_fit))
  lasso_coef_id <- which(coef(lasso_fit) != 0)
  lasso_res <- data.frame(var_names[lasso_coef_id], coef(lasso_fit)[lasso_coef_id])

  # choose gamma and lambda so that cvmeans are minimized
  min_l_subset <- which.min(subset_fit$cvMeans[[1]])
  subset_id <- coef(subset_fit, subset_fit$fit$lambda[[1]][min_l_subset], 0)
  id <- which(subset_id != 0)
  subset_res <- data.frame(var_names[id], subset_id[id])

```

```

min_l_subset01 <- sapply(1:10, function (x) which.min(subset01_fit$cvMeans[[x]]))
min_g_subset01 <- sapply(1:10, function (x) subset01_fit$cvMeans[[x]][[min_l_subset01[x]]])
gamma_id <- which.min(min_g_subset01)
min_lambda <- min(min_g_subset01)

subset01_coef <- coef(subset01_fit,
                      min_lambda,
                      subset01_fit$fit$gamma[gamma_id])
id <- which(subset01_coef != 0)

subset01_res <- data.frame(var_names[id],
                          subset01_coef[id])

min_l_subset02 <- sapply(1:10, function (x) which.min(subset02_fit$cvMeans[[x]]))
min_g_subset02 <- sapply(1:10, function (x) subset02_fit$cvMeans[[x]][[min_l_subset02[x]]])
gamma_id <- which.min(min_g_subset02)
min_lambda <- min(min_g_subset02)
subset02_coef <- coef(subset02_fit,
                      min_lambda,
                      subset02_fit$fit$gamma[gamma_id])
id <- which(subset02_coef != 0)
subset02_res <- data.frame(var_names[id],
                          subset02_coef[id])

return(list(lasso_res, subset_res, subset01_res, subset02_res))
}
set.seed(2550)

fit_res <- mclapply(1:5, function(x) fit_regression(data_imp[[x]], train_index),
                  mc.cores = 5)

#choose variable and pooled the result
lasso_var <- sapply(1:5, function(x) fit_res[[x]][[1]][,1])
lasso_var <- as.factor(unlist(lasso_var))
lasso_res <- lapply(1:5, function(x) fit_res[[x]][[1]])
lasso_res <- do.call(rbind, lasso_res)
#summary(lasso_var)

lasso_coef <- data.frame(var = c('Intercept', 'Var1:age_ps', 'Var1:NMR'),
                        value = c(mean(lasso_res[lasso_res$var_names.lasso_coef_id. == '(Intercept)',

```



```

mean(lasso_res[lasso_res$var_names.lasso_coef_id. == 'Var1:age_ps'],
mean(lasso_res[lasso_res$var_names.lasso_coef_id. == 'Var1:age_ps'],

subset_var <- sapply(1:5, function(x) fit_res[[x]][[2]][,1])
subset_var <- as.factor(unlist(subset_var))
subset_res <- lapply(1:5, function(x) fit_res[[x]][[2]])
subset_res <- do.call(rbind, subset_res)
#summary(subset_var)
subset_coef <- data.frame(var = c('Intercept', 'ftcd_score', 'NHW1', 'Var1:age_ps'),
value = c(mean(subset_res[subset_res$var_names.id. == '(Intercept)',
mean(subset_res[subset_res$var_names.id. == 'ftcd_score',
mean(subset_res[subset_res$var_names.id. == 'NHW1',2]),
mean(subset_res[subset_res$var_names.id. == 'Var1:age_ps',

subset01_var <- sapply(1:5, function(x) fit_res[[x]][[3]][,1])
subset01_var <- as.factor(unlist(subset01_var))
subset01_res <- lapply(1:5, function(x) fit_res[[x]][[3]])
subset01_res <- do.call(rbind, subset01_res)
#summary(subset01_var)
subset01_coef <- data.frame(var = c('Intercept', 'BA1:bdi_score_w00', 'ftcd_score', 'Var1:age_ps'),
value = c(mean(subset01_res[subset01_res$var_names.id. == '(Intercept)',
mean(subset01_res[subset01_res$var_names.id. == 'BA1:bdi_score_w00',
mean(subset01_res[subset01_res$var_names.id. == 'ftcd_score',
mean(subset01_res[subset01_res$var_names.id. == 'Var1:age_ps',

subset02_var <- sapply(1:5, function(x) fit_res[[x]][[4]][,1])
subset02_var <- as.factor(unlist(subset02_var))
subset02_res <- lapply(1:5, function(x) fit_res[[x]][[4]])
subset02_res <- do.call(rbind, subset02_res)
#summary(subset02_var)
subset02_coef <- data.frame(var = c('Intercept', 'antidepressant1:NMR', 'BA1:bdi_score_w00', 'BA1:inc.L',
'ftcd_score', 'mde_curr1', 'NHW1', 'Var1:age_ps', 'Var1:Black'),
value = c(mean(subset02_res[subset02_res$var_names.id. == '(Intercept)',
mean(subset02_res[subset02_res$var_names.id. == 'antidepressant1:NMR',
mean(subset02_res[subset02_res$var_names.id. == 'BA1:bdi_score_w00',
mean(subset02_res[subset02_res$var_names.id. == 'BA1:inc.L',
mean(subset02_res[subset02_res$var_names.id. == 'ftcd_score',
mean(subset02_res[subset02_res$var_names.id. == 'mde_curr1',
mean(subset02_res[subset02_res$var_names.id. == 'NHW1',2],
mean(subset02_res[subset02_res$var_names.id. == 'Var1:age_ps',
mean(subset02_res[subset02_res$var_names.id. == 'Var1:Black',

```

```

# test data set
test_data <- lapply(1:5, function(x) data_imp[[x]][-train_index,])
test_data <- do.call(rbind, test_data)

# fit model on test dataset
lasso_x <- model.matrix(abst ~ Var:age_ps + Var:NMR, data = test_data)
lasso_x <- lasso_x[,c(1,3,5)]
t <- lasso_coef$value %*% t(lasso_x)
lasso_p <- exp(t)/(1+exp(t))
auc_test_l <- roc(test_data$abst, lasso_p)

subset_x <- model.matrix(abst ~ ftcd_score + NHW + Var:age_ps, data = test_data)
subset_x <- subset_x[, -4]
t <- subset_coef$value %*% t(subset_x)
subset_p <- exp(t)/(1+exp(t))
auc_test_sub <- roc(test_data$abst, subset_p)

subset01_x <- model.matrix(abst ~ BA:bdi_score_w00 + ftcd_score + Var:age_ps, data = test_data)
subset01_x <- subset01_x[,c(1,2,4,6)]
t <- subset01_coef$value %*% t(subset01_x)
subset01_p <- exp(t)/(1+exp(t))
auc_test_sub01 <- roc(test_data$abst, subset01_p)

subset02_x <- model.matrix(abst ~ antidepmed:NMR + BA:bdi_score_w00 + BA:inc + ftcd_score + Var:age_ps, data = test_data)
subset02_x <- subset02_x[, -c(5,7,9,11,13,15,17,19)]
subset02_x <- subset02_x[, -c(8:10)]
t <- subset02_coef$value %*% t(subset02_x)
subset02_p <- exp(t)/(1+exp(t))
auc_test_sub02 <- roc(test_data$abst, subset02_p)

saveRDS(list(lasso_coef, subset_coef, subset01_coef, subset02_coef), "coef.rds")

plot.roc(auc_test_l, print.auc=TRUE, col = 'black')
plot.roc(auc_test_sub, print.auc=TRUE, col = 'blue', add = TRUE, print.auc.x = 0.5, print.auc.y = 0.5)
plot.roc(auc_test_sub01, print.auc=TRUE, col = 'red', add = TRUE, print.auc.x = 0.5, print.auc.y = 0.5)
plot.roc(auc_test_sub02, print.auc=TRUE, col = 'green', add = TRUE, print.auc.x = 0.5, print.auc.y = 0.5)
legend("bottomright", legend = c("Lasso", "L0", "L0L1", "L0L2"), col = c("black", "blue", "red", "green"))

p1 <- valProbggplot(as.numeric(subset02_p), as.numeric(test_data$abst)-1)
p1$ggPlot
p2 <- valProbggplot(as.numeric(subset01_p), as.numeric(test_data$abst)-1)
p2$ggPlot

```

```
p3 <- valProbggplot(as.numeric(subset_p), as.numeric(test_data$abst)-1)
p4 <- valProbggplot(as.numeric(lasso_p), as.numeric(test_data$abst)-1)
p3$ggPlot
p4$ggPlot
mod <- glm(abst ~ NHW + ftcd_score + Var*age_ps, data = data, family = "binomial")
tbl_regression(mod)
```