Smoking and Depression – A Closer Look at Causal Mediation Mechanisms

PHP2550 Project 2 | Due: November 10th at 11:59pm

Daniel Posmik (daniel_posmik@brown.edu)

Overview

This project is a collaboration with Dr. George Papandonatos in the Biostatistics Department and looks at smoking cessation in adults with major depressive disorder (MDD). Individuals with MDD are more likely to smoke heavily, exhibit greater nicotine dependence, and experience more severe withdrawal symptoms than those without MDD. While varenicline is an effective drug used to help people stop smoking, treatment of depression-related psychological factors related to smoking behavior may also improve rates of smoking cessation among adults with major depressive disorder (MDD).

A previous study used a randomized, placebo-controlled, 2x2 factorial design comparing behavioral activation for smoking cessation (BASC) versus standard behavioral treatment (ST) and varenicline versus placebo. Participants comprised 300 hundred adult smokers with current or past MDD. The study found that behavioral activation for smoking cessation did not outperform standard behavioral treatment, with or without adjunctive varenicline therapy.

The goal of this project is to use the data from this trial to examine baseline variables as potential moderators of the effects of behavioral treatment on end-of-treatment (EOT) abstinence and evaluate baseline variables as predictors of abstinence, controlling for behavioral treatment and pharmacotherapy. More concretely, we can formulate the following study aims:

- **Aim 1:** Determining controls: Which baseline variables predict smoking cessation?
- **Aim 2:** The Black Race Indicator: How does Black Race introduce collider bias?
- **Aim 3:** Further Mediation mechanisms: Which mediation paths are relevant?

Summary of the Data

Before we begin examining the study aims, let us take a look at the data. Exploratory data analysis showed that missingness is relatively low across the data. The most missing data is in

the Nicotine Metabolism Ratio variable, which has 7% missing data. This variable is followed by cigarette reward value at baseline (6%), baseline readiness to quit smoking (5.7%), income and anhedonia (1%, respectively), and excl. menthol cigarette use and FTCD score (<1%, respectively). Our questions in this project will largely center around possible mediation and moderation effects between the treatment therapies and the outcome, smoking abstinence. A helpful first step is to examine the balance of the data across the treatment arms.

Table 1: Balance Across Treatment Arms

		Treatment Arms (var & ba)			
Description	Variable	0 & 0	0 & 1	1 & 0	1 & 1
Patient ID	id	**146.8**	**151.5**	**157.6**	**158.1**
Smoking Abstinence (Yes)	abst1	**0.1**	**0.0**	**0.3**	**0.3**
Age at phone interview	age_ps	**50.6**	**51.4**	**47.9**	**50.2**
Sex at phone interview	sex_ps	**1.6**	**1.6**	**1.5**	**1.5**
Non-Hispanic White indicator	nhw1	**0.4**	**0.3**	**0.3**	**0.4**
Black indicator	black1	**0.6**	**0.6**	**0.5**	**0.5**
Hispanic indicator	hisp1	**0.1**	**0.1**	**0.1**	**0.0**
Income (Low)	inc.L	**-0.2**	**-0.2**	**-0.2**	**-0.2**
Income (Medium)	inc.Q	**0.0**	**0.1**	**0.1**	**0.1**
Income (High)	inc.C	**0.0**	**-0.1**	**0.1**	**0.0**
Income (Very High)	inc^4	**0.1**	**-0.1**	**0.0**	**0.0**
Education (Low)	edu.L	**0.3**	**0.2**	**0.3**	**0.3**
Education (Medium)	edu.Q	**-0.1**	**-0.1**	**-0.1**	**0.0**
Education (High)	edu.C	**-0.3**	**-0.1**	**0.0**	**-0.1**
Education (Very High)	edu^4	**-0.1**	**0.1**	**0.1**	**0.0**
FTCD score at baseline	ftcd_score	**5.5**	**5.2**	**5.2**	**5.0**
Smoking with 5 mins of waking up	ftcd5mins	**0.5**	**0.5**	**0.5**	**0.4**
BDI score at baseline	bdi_score_w00	**18.5**	**18.5**	**19.1**	**17.5**
Cigarettes per day at baseline	cpd_ps	**15.5**	**16.1**	**14.4**	**15.2**
Cigarette reward value at baseline	crv_total_pq1	**7.0**	**7.5**	**7.0**	**7.4**
Pleasurable Events Scale (subst)	hedonsum_n_pq1	**21.4**	**22.4**	**23.9**	**23.3**
Pleasurable Events Scale (compl)	hedonsum_y_pq1	**26.3**	**28.3**	**24.8**	**23.3**
Anhedonia	shaps_score_pq1	**2.8**	**2.2**	**2.2**	**2.3**
Other lifetime DSM-5 diagnosis	otherdiag	**0.4**	**0.5**	**0.5**	**0.4**
Antidepressant meds at baseline	antidepmed1	**0.2**	**0.4**	**0.2**	**0.3**
Current vs past MDD	mde_curr1	**0.4**	**0.5**	**0.5**	**0.4**
Nicotine Metabolism Ratio	nmr	**0.4**	**0.4**	**0.4**	**0.4**
Excl. Menthol Cigarette User	onlymenthol1	**0.6**	**0.6**	**0.6**	**0.6**

Baseline readiness to quit smoking	readiness	**6.9**	**6.8**	**6.7**	**6.7**

The following rows are interesting candidates for mediation.

- **Age**: There are noticeable differences in age across the treatment arms. Further, it is reasonable to suspect that the age of the patient could influence the causal mechanism.
- **Black race**: Although relatively well-balanced in the data, contextual knowledge about the tobacco industry's history of targeting Black communities makes this variable an interesting candidate for mediation.
- **Antidepression Medication**: Antidepression medication at baseline is an interesting candidate for mediation since there are noticeable differences across the treatment arms. Intuitively, we would expect a mediating relationship to be plausible.
- **Cigarette reward value**: This variable is a measure of the perceived reward of cigarettes. In conjunction with the other reward variables, there are noticeable differences across the treatment arms.

We believe examining these four variables will give us well-rounded insights into the causal mediation mechanisms at play. We could certainly be more exhaustive in our examination, such as taking into account variables with even starker differences across arms (e.g., pleasurable events scale). However, given the scope of this project, we believe our choice of mediation candidates strikes a good balance between information from the data and contextual knowledge.

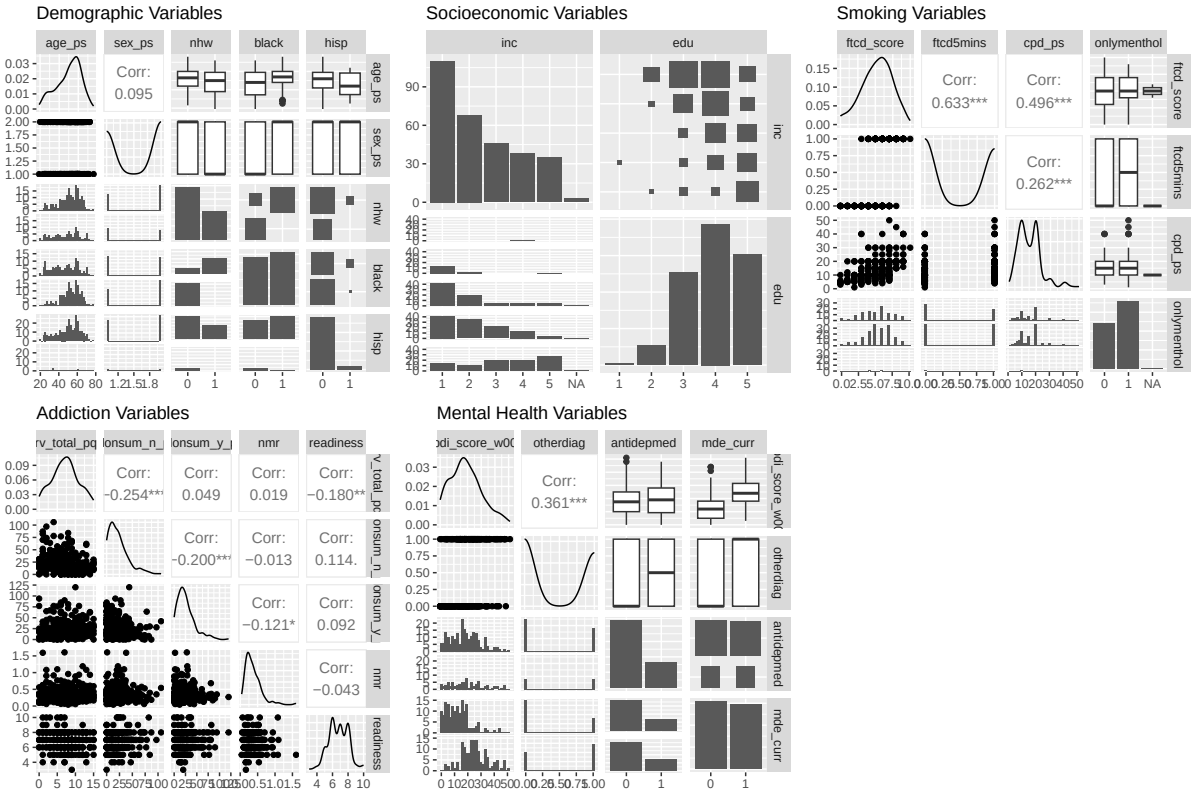
Aim 1: Determining Controls

Our first aim concerns the identification of relevant baseline variables that predict smoking cessation. Note that we are not yet concerned with establishing causality, rather we want to identify a sensible set of baseline adjustment variables that will help us in our causal identification later.

We can begin by designating our baseline variables into categories:

- Demographic information: `age_ps`, `sex_ps`, `nhw`, `black`, `hisp`
- Socioeconomic information: `inc`, `edu`
- Baseline smoking behavior: `ftcd_score`, `ftcd5mins`, `cpd_ps`, `onlymenthol`
- Addiction profile: `crv_total_pq1`, `hedonsum_n_pq1`, `hedonsum_y_pq1`, `nmr`, `readiness`
- Mental health profile: `bdi_scorew00`, `otherdiag`, `antidepmed`, `mde_curr`

Here, we discern between active smoking metrics (e.g., cigarettes per day) with a patient's addiction profile, i.e. a broader look on the patients motivations and exhibited traits as it relates to their cigarette/ nicotine addiction. Now, let us take a look at relationships between the data. We will report separate analyses of correlation by category to avoid clutter and emphasize clarity.



We can see that there are some interesting relationships between the variables. The FTCD score variable is visibly correlated with smoking within 5min of waking up and cigarettes per day. This makes sense since FTCD is a composite score for nicotine dependence that takes these factors into account. Now, regarding the addiction variables, the correlations are relatively low across the board. There is a slight relationship between cigarette reward value at baseline and pleasurable events scale (substitute reinforcers). This makes sense since both are a measure of gratification and reward and we would have even reasonably expected a higher correlation. In terms of mental health variables, we see that there is significant separation between being on antidepression medication and having another depression diagnosis. This phenomenon is even higher when comparing being on antidepression medication and the measure for MDD current vs. past. The BDI score at baseline is slightly related to having another diagnosis, but not necessarily to being on antidepression medicine or the measure for MDD current vs. past.

Initial analysis shows that–including both treatment variables–the FTCD score (Estimate: $-3.090e-01$, p: 0.0012) is the only significant variable in predicting abstinence. The following table shows the results of two logistic regression models, one without and one with a regularizatio penalty, when including FTCD Score and the mediation candidates. We choose not to conduct stepwise subset selection because of three reasons: 1) only one variable is significant in the full model, 2) we have sufficient contextual knowledge thanks to the paper, and

3) stepwise selection runs a high risk of overfitting. We decided against imputation because the number of missing values is very low in this reduced model.

```
[1] "Our error-minimizing value of lambda is 0.0039"
```

Here, the LASSO model serves a largely prescriptive purpose. Rather than being interested in the LASSO model itself, we seek to gather evidence on the stability of the variables chosen from the EDA step. We can see that the logit coefficients are relatively stable across the LASSO model, although the cigarette reward value and age variables were shrunk significantly. On the contrary, if we had observed significant shrinkage across categories, this could have warranted even more reduction in the set of controls. The LASSO regression here was used as a check and I have revised the report to better communicate this. In terms of diagnostics, we can see that our lambda is relatively small, suggesting a smaller regularization effect in the model. The LASSO model imposes fairly significant shrinkage on the coefficients. Now, no variable is significant in the LASSO model. This is not surprising given that the FTCD score was the only significant variable in the full model.

Aim 2: The Black Race Indicator and Collider Bias

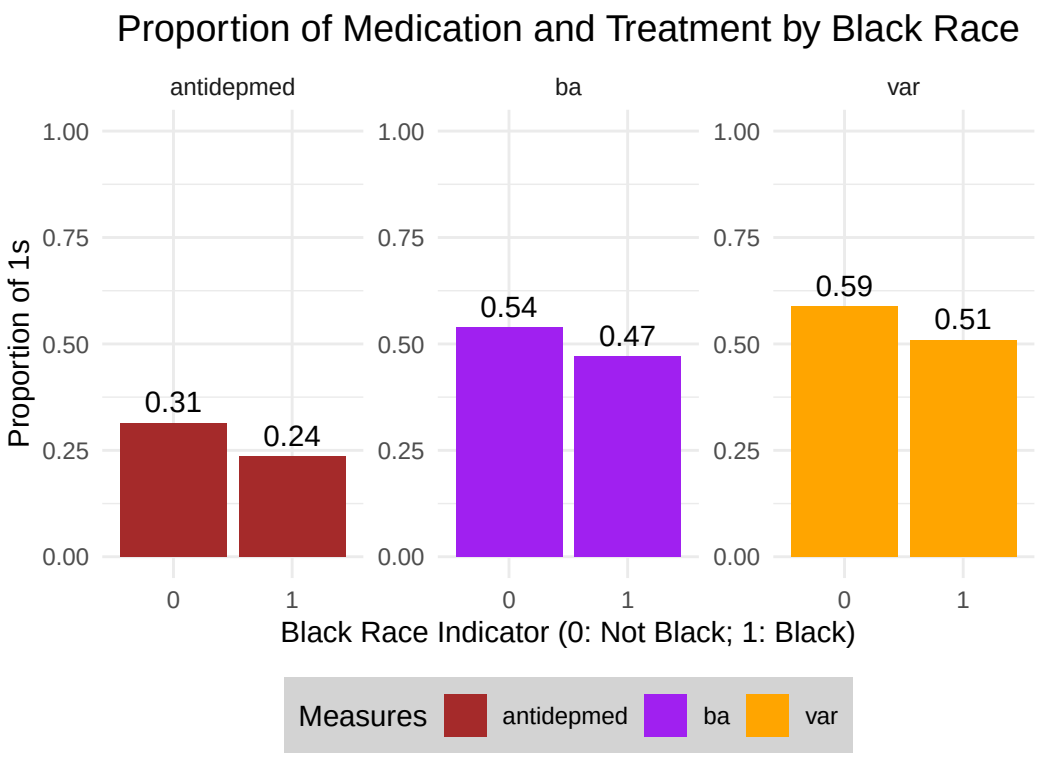
We now examine an interesting hypothesis around the role of race as a mediator in the causal relationship between behavioral treatment and smoking cessation. Now, by the definition of a mediator, race would have an incoming causal path from the treatment and an outgoing path to the outcome. The latter, i.e. race has an effect on smoking cessation, is plausible. However, the former, i.e. the treatment affecting race, is not. That being said, we are not ready to discard the hypothesis that race plays an interesting role in the causal relationship just yet.

While it is impossible that the treatment would affect race, the reverse certainly seems plausible. That means that it is likely that race affects the treatment. Under this hypothesis, race would no longer be classified as a mediator, but as a collider. If we controlled for race, we would introduce causal paths from race to the treatment and the outcome variables, effectively introducing collider bias. More specifically, we believe race being a collider is a reasonable suspicicion since marginalized populations are often more likely to experience MDD (evidence in visualization below as measured by baseline antidepression medication), and thus experience different treatment regimes. In our data, though treatment was randomized, we can see slight evidence of this claim when analyzing the proportion of treatment by the black race indicator.

Table 2: Logistic Regression and LASSO Regularization: Predicting Abstinence by Baseline Variables

	Logit Model	LASSO Model
(Intercept)	−1.829*	1.240
	(0.859)	
var1	1.446***	0.181
	(0.372)	
ba1	−0.361	−0.039
	(0.328)	
ftcd_score	−0.342***	−0.048
	(0.081)	
age_ps	0.026+	0.003
	(0.014)	
black1	−0.671+	−0.085
	(0.360)	
antidepmed1	0.375	0.043
	(0.373)	
crv_total_pq1	0.039	0.004
	(0.048)	
Num.Obs.	281	281
F	4.820	
RMSE	0.37	
nulldev		46.0284697508895
npasses		6

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001



We are now ready to assess how this problem manifests in causal estimation. Assuming that the black race indicator is indeed a collider, introducing it as a control would introduce bias. In the below visualization, the red arrow in the left diagram "collides" with the treatment.

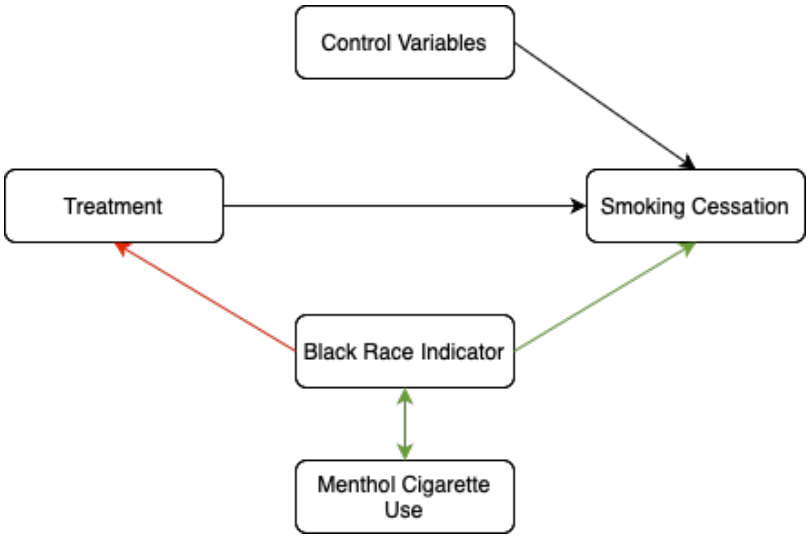


Figure 1: Causal Directed Acyclic Graph (DAG) with the Race Collider

We are interested in resolving this problem with an instrumental variable (IV) approach, using the use of menthol cigarettes to instrument for the Black race indicator. By using menthol cigarettes as an IV, we can eliminate the colliding path from race to treatment. The causal DAG is shown below.

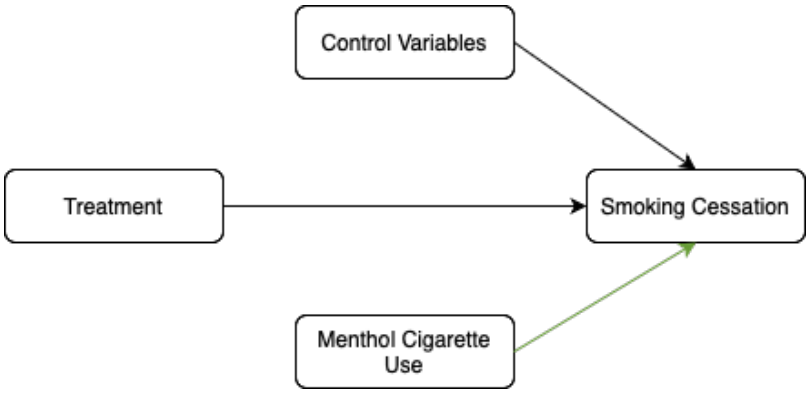


Figure 2: Causal Directed Acyclic Graph (DAG) with the Menthol Cigarettes IV

Importantly, for a successful IV estimation, we have to fulfill two key assumptions:

- **Relevance Assumption**: The IV variable must be highly predictive of the treatment variable.
- **Exclusion Restriction**: The IV variable must not be directly related to the outcome variable.

In the below contingency table, we can see that the black race indicator coincides strongly with the use of menthol cigarettes. This validates our reasoning to use menthol cigarettes as an IV. Now, regarding the exclusion restriction, this is difficult to test. However, we shall assume that the use of menthol cigarettes is not directly related to smoking cessation for the sake of this analysis. In practice, this assumption is often a point of contention and scrutiny.

Table 3: Contingency Table of Black Race Indicator and Menthol Cigarette Use

Black	Only Menthol	Count
0	**0**	93
1	**0**	27
0	**1**	49
1	**1**	129

Using an IV framework, we will proceed with causal effect estimation using Menthol Cigarettes as an instrument for the black race indicator. For a concise overview, we will report the black race interaction results in the next section, Aim 3.

Aim 3: Further Examination of Mediation Mechanisms

We will know turn to the second aim of our study, which is to examine mediation mechanisms. For each hypothesized mediation effect, we will fit a logistic regression model similar to the above, but including the respective interaction terms.

The results are interesting. The interaction term between the IV for black race and not receiving the varicline treatment is highly significant ($p < 0.01$). This suggests that for individuals who use menthol cigarettes, the effect of the varicline treatment on smoking cessation is significantly different. This is a strong indication that there are race-specific moderation mechanisms at play, e.g., the treatment is more effective for non-Black individuals.

In the other interaction groups, we observe interaction effects at a significance level of $p < 0.05$. This suggests that we can be relatively confident that mediation mechanisms are plausible, given our observed data. The interaction term between age and the varicline treatment is significant, suggesting that the effect of the treatment on smoking cessation is moderated by age. For individuals that receive varicline, all else equal, a one year increase in age corresponds to a 0.035 increase in log odds (+ ~3% probability) of smoking cessation.

The interaction term between the cigarette reward value and the treatment is also significant, suggesting that the effect of the treatment on smoking cessation is moderated by the perceived reward of cigarettes. Finally, the interaction term between antidepression medication and the treatment is significant, suggesting that the effect of the treatment on smoking cessation is moderated by the use of antidepression medication.

It is important to note that all significant interaction terms can be found on the varicline treatment side, suggesting that behavioral treatment is not moderated (nor significant) with respect to smoking cessation. We can conclude that the varicline treatment is the main driver of smoking cessation in our data and is moderated strongly by black race and also by age, cigarette reward value, and antidepression medication.

Limitations

This project presented many limitations that we want to acknowledge. Due to the time constraints, there are many sets and subsets of interaction terms that can be explored. With the time and space given, we believe our analysis does provide a good foundation for understanding smoking cessation predictors. However, one major limitation was the low sample size relative to the number of predictors, which necessitated a drastic reduction in variables to avoid overfitting and spurious results. While this reduction was justified qualitatively through the grouping of variables into logical clusters, it remains worrisome in terms of capturing the complexity of smoking cessation behaviors. Additionally, while LASSO regression was used as a diagnostic tool to assess variable importance and potential shrinkage, the limited shrinkage observed does not entirely mitigate concerns about the stability of variable selection given the

Table 4: Summary of Mediation Models with Interaction Terms

	Mediation Models			
	Black Race Interaction	Age Interaction	CRV Interaction	Medication Interaction
(Intercept)	−0.308	−1.077	−1.251	−0.408
	(0.957)	(0.849)	(0.843)	(1.088)
ftcd_score	−0.335***	−0.352***	−0.331***	−0.346***
	(0.089)	(0.089)	(0.087)	(0.089)
age_ps	0.021		0.023	0.023
	(0.016)		(0.015)	(0.016)
antidepmed1	0.432	0.406	0.393	
	(0.418)	(0.414)	(0.411)	
crv_total_pq1	0.032	0.048		0.045
	(0.054)	(0.053)		(0.053)
var0 × onlymenthol0	−1.703*			
	(0.686)			
var1 × onlymenthol0	−0.502			
	(0.585)			
var0 × onlymenthol1	−1.853**			
	(0.594)			
ba1 × onlymenthol0	0.215			
	(0.553)			
ba1 × onlymenthol1	−0.810+			
	(0.481)			
onlymenthol1		−0.218	−0.242	−0.186
		(0.386)	(0.387)	(0.386)
var0 × age_ps		0.004		
		(0.017)		
var1 × age_ps		0.035*		
		(0.016)		
ba1 × age_ps		−0.009		
		(0.007)		
var0 × crv_total_pq1			−0.059	
			(0.075)	
var1 × crv_total_pq1			0.128*	
			(0.060)	
ba1 × crv_total_pq1			−0.052	
			(0.046)	
var0 × antidepmed0				−1.629*
				(0.785)
var1 × antidepmed0				−0.075
				(0.709)
var0 × antidepmed1				−1.623*
				(0.745)
ba1 × antidepmed0				−0.544
				(0.425)
ba1 × antidepmed1				0.080
				(0.722)
Num.Obs.	241	241	241	241
F		4.027	3.571	
RMSE	0.37	0.37	0.37	0.37

+ p \num{< 0.1}, * p \num{< 0.05}, ** p \num{< 0.01}, *** p \num{< 0.001}

dataset's constraints. Finally, the inherent limitations of this dataset, compounded by the challenges of low n and high p, underscore the need for caution when interpreting the results and highlight areas for future research with larger and more representative samples.

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

```
# Set up knit environment
knitr::opts_chunk$set(echo = F)
knitr::opts_chunk$set(error = F)
knitr::opts_chunk$set(warning = F)
knitr::opts_chunk$set(message = F)

# Load necessary packages
library(tidyverse)
library(magrittr)
library(lubridate)
library(GGally)
library(broom)
library(corrplot)
library(kableExtra)
library(grid)
library(gridExtra)
library(glmnet)
library(ggridges)
library(knitr)
library(ggplot2)
library(naniar)
library(gtsummary)
library(cowplot)
library(gt)

# Define folders
base_folder <-
  "/Users/posmikdc/Documents/brown/classes/php2550-pda/"

input_folder <-
  paste0(base_folder, "php2550-projects/project2/data/")

output_folder <-
  paste0(base_folder, "php2550-projects/project2/output/")


# Define function to rename columns
my_rename <- function(df) {
  names(df) <- names(df) %>%
```

```
    tolower() %>%
    gsub(" ", "_", .) %>%
    gsub("[^[:alnum:]_]", "", .)
  return(df)
}

# Load data
p2_dta <- read_csv(paste0(input_folder, "project2.csv")) %>%
  my_rename() %>%
  mutate(
    var = as.factor(var),
    ba = as.factor(ba),
    nhw = as.factor(nhw),
    black = as.factor(black),
    hisp = as.factor(hisp),
    inc = ordered(inc),
    edu = ordered(edu),
    antidepmed = as.factor(antidepmed),
    mde_curr = as.factor(mde_curr),
    onlymenthol = as.factor(onlymenthol),
    abst = as.factor(abst)
)

# Expand factors into binary columns
p2_matrix <- p2_dta %>%
  model.matrix(~ . - 1, data = .) %>%
  as.data.frame() %>%
  select(- abst0)

tbl.bal <- p2_matrix %>%
  group_by(var1, ba1) %>%
  summarise(across(everything(), mean, na.rm = TRUE)) %>%
  filter(
    (var1 == 1 & ba1 == 1) |
    (var1 == 1 & ba1 == 0) |
    (var1 == 0 & ba1 == 1) |
    (var1 == 0 & ba1 == 0)
  ) %>%
  pivot_longer(
    cols = -c(var1, ba1),
    names_to = "Variable",
    values_to = "mean_value"
```

```
  ) %>%
  pivot_wider(
    names_from = c(var1, ba1),
    names_sep = " & ",
    values_from = "mean_value"
  ) %>%
  mutate(across(where(is.numeric), ~ round(.x, 1)))

# Merge table with other descriptors
Description <- c(
  "Patient ID", #1
  "Smoking Abstinence (Yes)",
  "Age at phone interview", #5
  "Sex at phone interview", #6
  "Non-Hispanic White indicator", #7
  "Black indicator", #8
  "Hispanic indicator", #9
  "Income (Low)", #10
  "Income (Medium)", #10
  "Income (High)", #10
  "Income (Very High)", #10
  "Education (Low)", #11
  "Education (Medium)",
  "Education (High)",
  "Education (Very High)",
  "FTCD score at baseline", #12
  "Smoking with 5 mins of waking up", #13
  "BDI score at baseline", #14
  "Cigarettes per day at baseline", #15
  "Cigarette reward value at baseline", #16
  "Pleasurable Events Scale (subst)", #17
  "Pleasurable Events Scale (compl)", #18
  "Anhedonia", #19
  "Other lifetime DSM-5 diagnosis", #20
  "Antidepressant meds at baseline", #21
  "Current vs past MDD", #22
  "Nicotine Metabolism Ratio", #23
  "Excl. Menthol Cigarette User", #24
  "Baseline readiness to quit smoking" #25
  )

tbl.bal <- cbind(Description, tbl.bal)
```

```r
# Print table
knitr::kable(tbl.bal,
  caption = "Balance Across Treatment Arms") %>%
  kable_styling("striped", full_width = F) %>%
  add_header_above(c(" " = 2,
                     "Treatment Arms (var & ba)" = 4)) %>%
  column_spec(3:6, bold = T)

# Define baseline variables
baseline_vars <- c("age_ps",
                   "sex_ps",
                   "nhw",
                   "black",
                   "hisp",
                   "inc",
                   "edu",
                   "ftcd_score",
                   "ftcd5mins",
                   "bdi_score_w00",
                   "cpd_ps",
                   "crv_total_pq1",
                   "hedonsum_n_pq1",
                   "hedonsum_y_pq1",
                   "otherdiag",
                   "antidepmed",
                   "mde_curr",
                   "nmr",
                   "onlymenthol",
                   "readiness")

# Define subgroups
demo_vars <- c("age_ps",
               "sex_ps",
               "nhw",
               "black",
               "hisp")

se_vars <- c("inc",
             "edu")

smoking_vars <- c("ftcd_score",
                  "ftcd5mins",
```

```
                  "cpd_ps",
                  "onlymenthol")

addiction_vars <- c("crv_total_pq1",
                    "hedonsum_n_pq1",
                    "hedonsum_y_pq1",
                    "nmr",
                    "readiness")

mentalhealth_vars <- c("bdi_score_w00",
                       "otherdiag",
                       "antidepmed",
                       "mde_curr")

# Plot correlations
p.demo <-
  GGally::ggpairs(p2_dta %>% select(all_of(demo_vars)),
                  title = "Demographic Variables")

p.se <-
  GGally::ggpairs(p2_dta %>% select(all_of(se_vars)),
                  title = "Socioeconomic Variables")

p.smoking <-
  GGally::ggpairs(p2_dta %>% select(all_of(smoking_vars)),
                  title = "Smoking Variables")

p.addiction <-
  GGally::ggpairs(p2_dta %>% select(all_of(addiction_vars)),
                  title = "Addiction Variables")

p.mentalhealth <-
  GGally::ggpairs(p2_dta %>% select(all_of(mentalhealth_vars)),
                  title = "Mental Health Variables")

# Arrange plots in 1x3 format
plot_grid(
   ggmatrix_gtable(p.demo),
   ggmatrix_gtable(p.se),
   ggmatrix_gtable(p.smoking),
   ggmatrix_gtable(p.addiction),
   ggmatrix_gtable(p.mentalhealth),
```

```
   nrow = 2)

# Create reduced data frame
p2_red <- p2_dta %>%
  select(abst, ftcd_score, age_ps,
         black, antidepmed, crv_total_pq1,
         var, ba) %>%
  drop_na()

# Create formula
formula <-
  as.formula(abst ~ var + ba + ftcd_score +
                    age_ps + black + antidepmed +
                    crv_total_pq1)

# Fit logistic regression model
logit_fit <- glm(formula,
                 data = p2_red,
                 family = binomial)

# Fit LASSO model
n.cv = 10 # Number of cross-validation folds
reg.type <- 1 # Type of regularization

lasso_cv <-
  glmnet::cv.glmnet(x = model.matrix(formula, data = p2_red),
                    y = as.numeric(p2_red$abst),
                    alpha = reg.type,
                    family = "binomial",
                    nfolds = n.cv)

# Find optimal lambda value
lambda.min <- lasso_cv$lambda.min
#plot(lasso_cv)

# Print
print(paste0("Our error-minimizing value of lambda is ",
      round(lambda.min,5)))

# LASSO model
lasso_fit <-
  glmnet::glmnet(x = model.matrix(formula, data = p2_red),
```

```
                  y = as.numeric(p2_red$abst),
                  alpha = reg.type,
                  lambda = lambda.min)

# Print table
modelsummary::modelsummary(list("Logit Model" = logit_fit,
                                "LASSO Model" = lasso_fit),
  style = "stars",
  stars = TRUE,
  title = "Logistic Regression and LASSO Regularization: Predicting Abstinence by Baseline Va
  gof_omit = 'IC|Log|Adj')

# Create a summary data frame
pbar_df <- p2_dta %>%
  filter(!is.na(black)) %>%
  pivot_longer(cols = c(var, ba, antidepmed),
    names_to = "variable", values_to = "value") %>%
  filter(!is.na(value)) %>%
  group_by(black, variable) %>%
  summarize(proportion_1 = mean(value == 1), .groups = "drop")

# Plot
p.bar <- ggplot(pbar_df,
    aes(x = factor(black), y = proportion_1, fill = variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = round(proportion_1, 2)),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 4) +
  facet_wrap(~ variable, scales = "free_y") +
  labs(title = "Proportion of Medication and Treatment by Black Race",
       x = "Black Race Indicator (0: Not Black; 1: Black)",
       y = "Proportion of 1s",
       fill = "Measures") +
  scale_fill_manual(values = c("brown", "purple", "orange")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14),
    legend.position = "bottom",
    legend.box = "horizontal",
    legend.background = element_rect(fill = "lightgray",
                                     color = NA)
  ) +
```

```r
  ylim(0, 1)

# Display the plot
p.bar

ggsave("../fig/bar-plot.png", plot = p.bar, width = 8, height = 6, dpi = 300)

# Create a contingency table
ctgy.tbl <-
  table(p2_dta$black,
        p2_dta$onlymenthol)

# Convert the table to a data frame
ctgy.df <- as.data.frame(ctgy.tbl)

ctgy.df %>%
  kable(col.names = c("Black", "Only Menthol", "Count"),
    caption = "Contingency Table of Black Race Indicator and Menthol Cigarette Use") %>%
  kable_styling("striped", full_width = F) %>%
  column_spec(1, bold = TRUE) %>%
  column_spec(2, bold = TRUE)
# Create interaction models
black.x_fit <-
  glm(abst ~ (var + ba):onlymenthol + ftcd_score +
              age_ps + antidepmed + crv_total_pq1,
                   data = p2_dta %>% drop_na(),
                   family = binomial)

age.x_fit <-
  glm(abst ~ (var + ba):age_ps + ftcd_score +
              onlymenthol + antidepmed + crv_total_pq1,
                   data = p2_dta %>% drop_na(),
                   family = binomial)

crv.x_fit <-
  glm(abst ~ (var + ba):crv_total_pq1 + ftcd_score +
              onlymenthol + antidepmed + age_ps,
                   data = p2_dta %>% drop_na(),
                   family = binomial)

antidepmed.x_fit <-
  glm(abst ~ (var + ba):antidepmed + ftcd_score +
```

```
              onlymenthol + crv_total_pq1 + age_ps,
                   data = p2_dta %>% drop_na(),
                   family = binomial)

# Create the summary table
model.tbl <- modelsummary::modelsummary(
    list("Black Race Interaction" = black.x_fit,
         "Age Interaction" = age.x_fit,
         "CRV Interaction" = crv.x_fit,
         "Medication Interaction" = antidepmed.x_fit),
    style = "stars",
    stars = TRUE,
    title = "Summary of Mediation Models with Interaction Terms",
    gof_omit = 'IC|Log|Adj',
    output = "kableExtra"
)

# Customize table to fit on one page
model.tbl %>%
    kable_styling(font_size = 10,
                  bootstrap_options = "condensed",
                  full_width = TRUE) %>%
    add_header_above(c(" " = 1, "Mediation Models" = 4)) %>%
    column_spec(1, width = "11em") %>%
    row_spec(c(10,22,28,34), hline_after = TRUE) %>%
    scroll_box(width = "100%", height = "400px")
```