Project 2 Codebook

Yingxi Kong

`Table 1` presents an overall summary statistics of patients' baseline characteristics by their behavioral and pharmacological treatment assignment. Since our study is a 2 × 2, factorial, randomized, placebo-controlled trial, patients are randomly assigned to either behavioral activation for smoking cessation group (BASC) or standard behavioral treatment group (ST) and either varenicline or placebo blister pack. Patients can be categorized into four treatment arm groups: BASC + placebo, BASC + varenicline, ST + placebo, and ST + varenicline. Seeing from `Table 1`, the two placebo groups both have 68 obervations while the two varenicline groups both have 83 observations.

Most variables are evenly distributed across the four treatment arms, which reflects successful randomization in this factorial trial. However, a few key factors, such as socioeconomic indicators (income and education) and specific mental health variables (MDD status, DSM-5 diagnoses), exhibit slight variations that may influence outcomes. Notably, treatment arms with varenicline show higher abstinence rates than placebo groups, suggesting the potential efficacy of this pharmacotherapy in combination with behavioral interventions. While many baseline characteristics are evenly distributed across groups, some may still function as moderators, potentially interacting with treatment assignment to affect abstinence success.

Table 1: Participant Characteristics by Treatment Arm

Characteristic	**Behavioral and Pharmacological Treatment Assignment**				
	BASC + placebo, N = 68	**BASC + varenicline, N = 83**	**ST + placebo, N = 68**	**ST + varenicline, N = 81**	**Overall, N = 300**
Smoking abstinence	4 (5.9%)	26 (31%)	8 (12%)	26 (32%)	64 (21%)
Age	51 (14)	50 (13)	50 (11)	49 (13)	50 (13)
Sex					
Male	30 (44%)	39 (47%)	29 (43%)	37 (46%)	135 (45%)
Female	38 (56%)	44 (53%)	39 (57%)	44 (54%)	165 (55%)
Income					
Less than $20,000	25 (37%)	30 (37%)	26 (38%)	29 (36%)	110 (37%)
$20,000-35,000	16 (24%)	17 (21%)	14 (21%)	21 (26%)	68 (23%)
$35,001-50,000	8 (12%)	13 (16%)	14 (21%)	11 (14%)	46 (15%)
$50,001-75,000	12 (18%)	12 (15%)	8 (12%)	6 (7.5%)	38 (13%)
More than $75,000	6 (9.0%)	10 (12%)	6 (8.8%)	13 (16%)	35 (12%)
Missing	1	1	0	1	3
Education					
Grade School	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)	1 (0.3%)
Some high school	3 (4.4%)	7 (8.4%)	2 (2.9%)	4 (4.9%)	16 (5.3%)
High school graduate or GED	23 (34%)	15 (18%)	11 (16%)	27 (33%)	76 (25%)
Some college/technical school	22 (32%)	32 (39%)	38 (56%)	24 (30%)	116 (39%)
College graduate	19 (28%)	29 (35%)	17 (25%)	26 (32%)	91 (30%)
FTCD score	5 (2)	5 (2)	5 (2)	5 (2)	5 (2)
Missing	0	0	1	0	1
Smoking within 5 mins of waking up	32 (47%)	33 (40%)	35 (51%)	38 (47%)	138 (46%)
BDI score	19 (12)	18 (11)	18 (11)	20 (12)	19 (11)
Cigarettes smoked per day	16 (9)	16 (9)	15 (7)	14 (7)	15 (8)
Cigarette reward value	7 (4)	7 (4)	7 (4)	7 (3)	7 (4)
Missing	1	3	8	6	18
Pleasurable events (substitute reinforcers)	23 (20)	23 (19)	21 (20)	23 (19)	23 (20)

Table 1: Participant Characteristics by Treatment Arm *(continued)*

Characteristic	**Behavioral and Pharmacological Treatment Assignment**				
	BASC + placebo, N = 68	**BASC + varenicline, N = 83**	**ST + placebo, N = 68**	**ST + varenicline, N = 81**	**Overall, N = 300**
Pleasurable events (complementary reinforcers)	28 (22)	22 (17)	27 (20)	25 (19)	25 (19)
Anhedonia	2 (3)	2 (3)	3 (3)	2 (3)	2 (3)
Missing	2	0	1	0	3
Other lifetime DSM-5 diagnosis	35 (51%)	30 (36%)	28 (41%)	40 (49%)	133 (44%)
Taking antidepressant	28 (41%)	24 (29%)	15 (22%)	15 (19%)	82 (27%)
Current vs. past MDD					
Past MDD	36 (53%)	43 (52%)	37 (54%)	37 (46%)	153 (51%)
Current MDD	32 (47%)	40 (48%)	31 (46%)	44 (54%)	147 (49%)
Nicotine metabolism ratio	0.34 (0.18)	0.38 (0.25)	0.37 (0.27)	0.36 (0.21)	0.36 (0.23)
Missing	7	3	2	9	21
Exclusive mentholated cigarette user	40 (59%)	48 (59%)	43 (64%)	47 (58%)	178 (60%)
Missing	0	1	1	0	2
Readiness to quit smoking	7 (1)	7 (1)	7 (1)	7 (1)	7 (1)
Missing	4	5	4	4	17
Race					
Black	37 (54%)	37 (45%)	40 (59%)	43 (53%)	157 (52%)
Hispanic	4 (5.9%)	3 (3.6%)	4 (5.9%)	5 (6.2%)	16 (5.3%)
Non-Hispanic White	24 (35%)	34 (41%)	22 (32%)	25 (31%)	105 (35%)
Other	3 (4.4%)	9 (11%)	2 (2.9%)	8 (9.9%)	22 (7.3%)

[1] Mean (SD) for continuous; n (%) for categorical

To further investigate the variation among groups, we generate a series of stacked bar plots to illustrate the proportions of several key categorical variables by abstinence status and treatment arms, along with distribution plots for selected continuous variables. We can examine potential differences in these baseline characteristics across treatment and abstinence outcomes, understanding how these factors would influence or interact with our treatment effectiveness.

Figure 1: Income Levels by Abstinence Status and Treatment Group

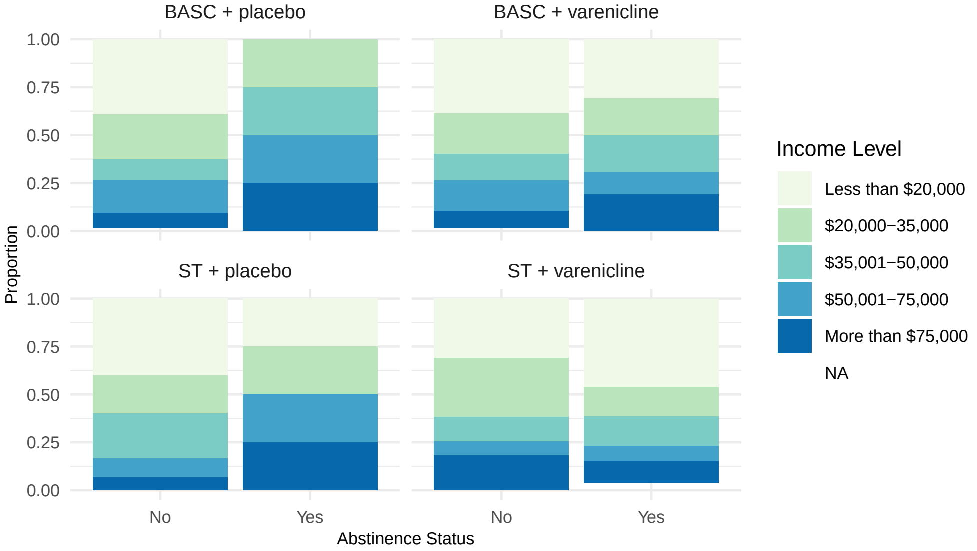


Figure 2: Education Levels by Abstinence Status and Treatment Group

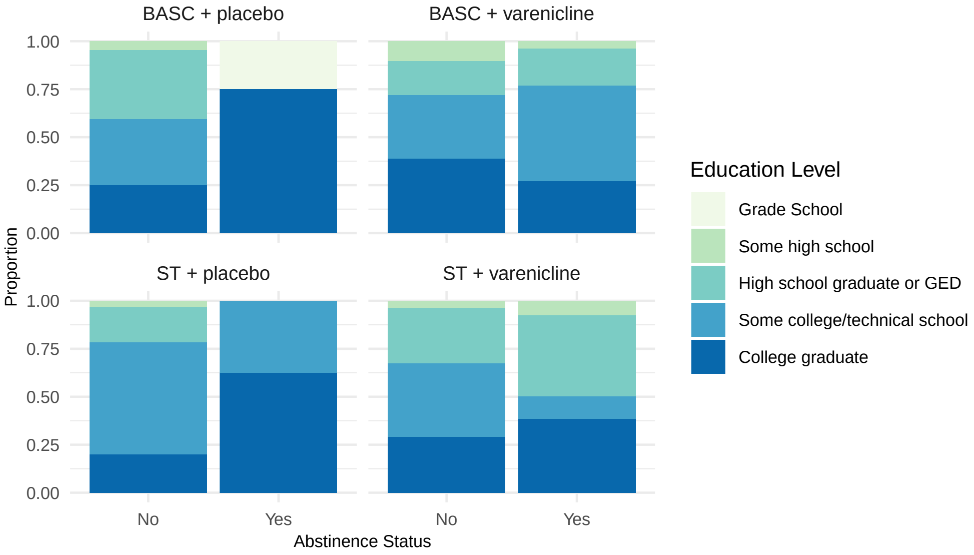
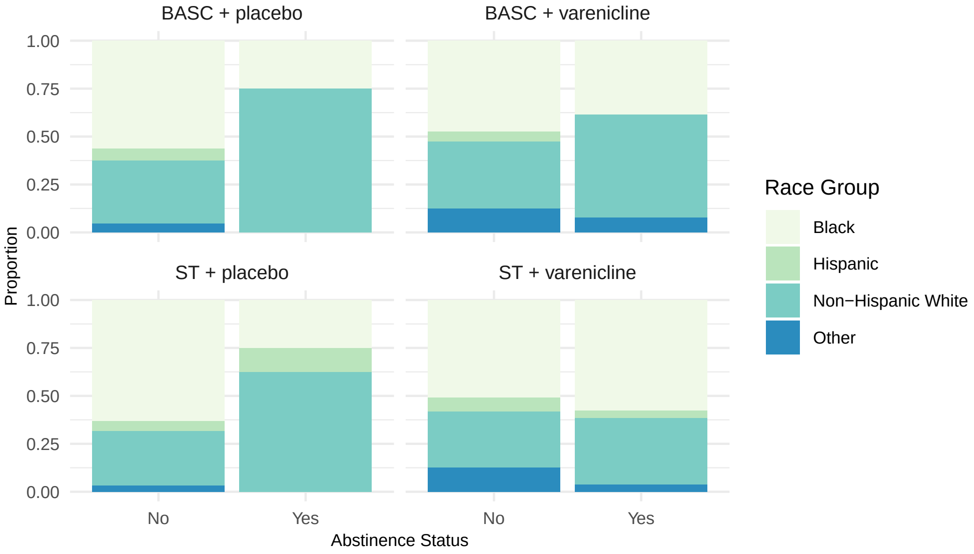
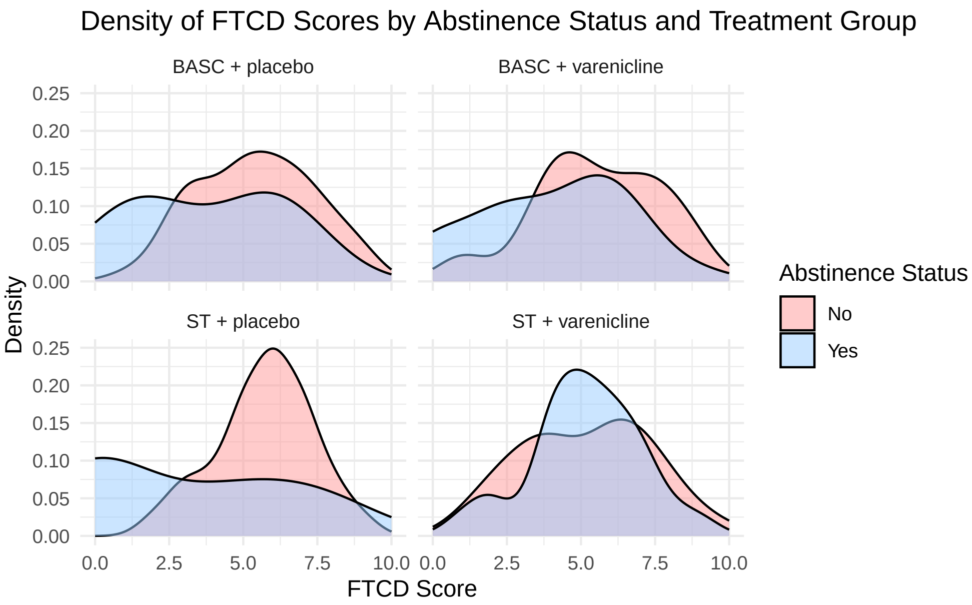
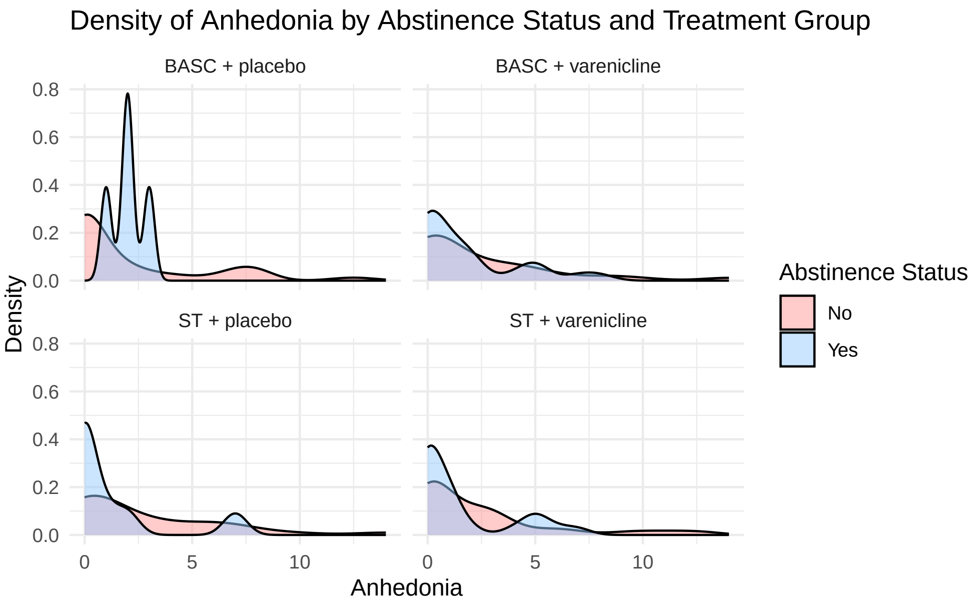
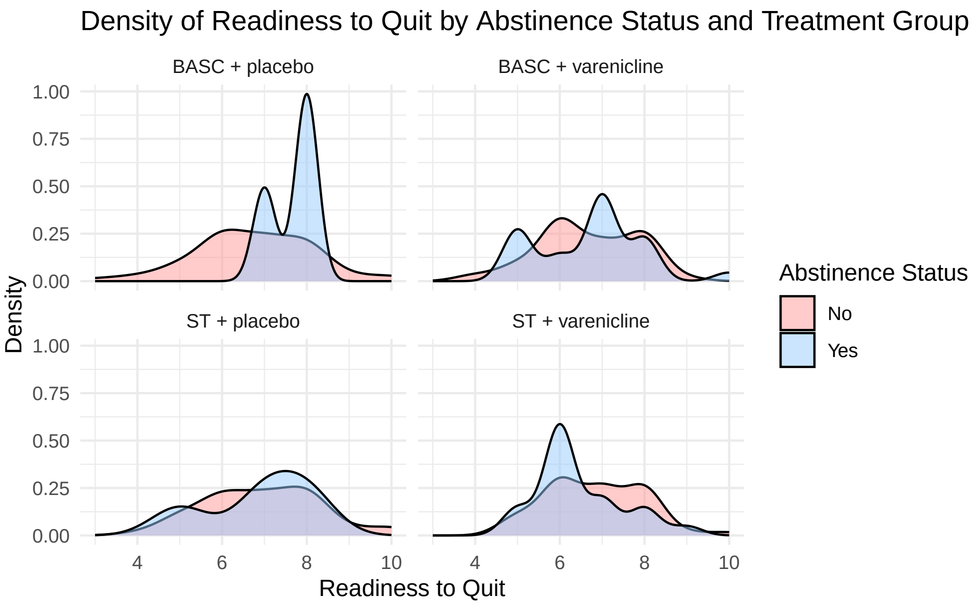


Figure 2: Race by Abstinence Status and Treatment Group




Density of FTCD Scores by Abstinence Status and Treatment Group
BASC + placebo
BASC + varenicline
ST + placebo
ST + varenicline
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.0
2.5
5.0
7.5
10.0
FTCD Score
Abstinence Status
No
Yes


Density of Anhedonia by Abstinence Status and Treatment Group
BASC + placebo
BASC + varenicline
ST + placebo
ST + varenicline
Density
0.0
0.2
0.4
0.6
0.8
0
5
10
Anhedonia
Abstinence Status
No
Yes


Density of Readiness to Quit by Abstinence Status and Treatment Group

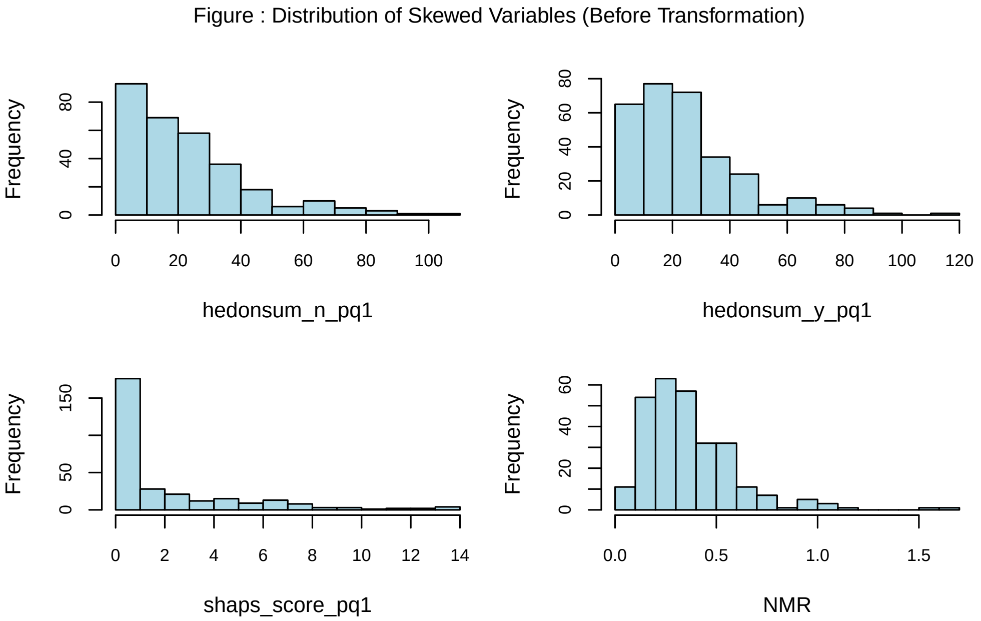


Figure : Distribution of Skewed Variables (Before Transformation)

```
## Area under the curve: 0.8204
```

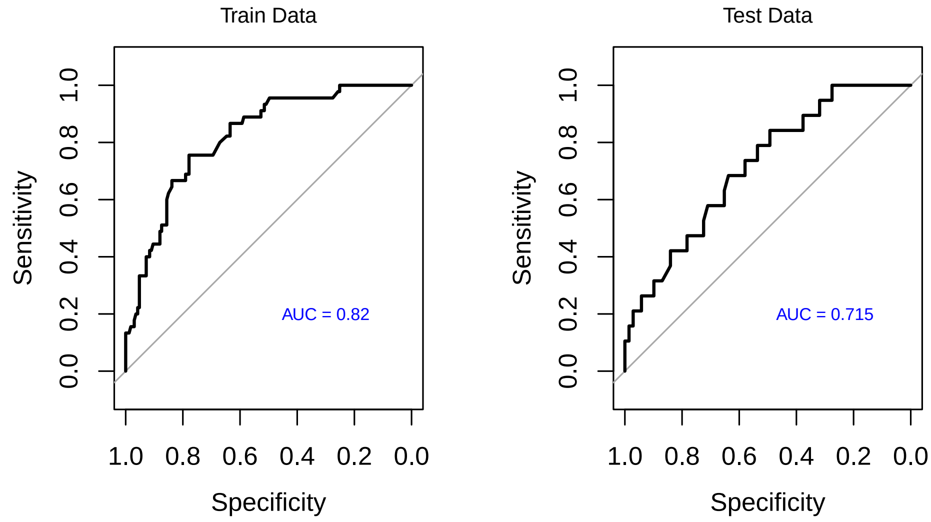
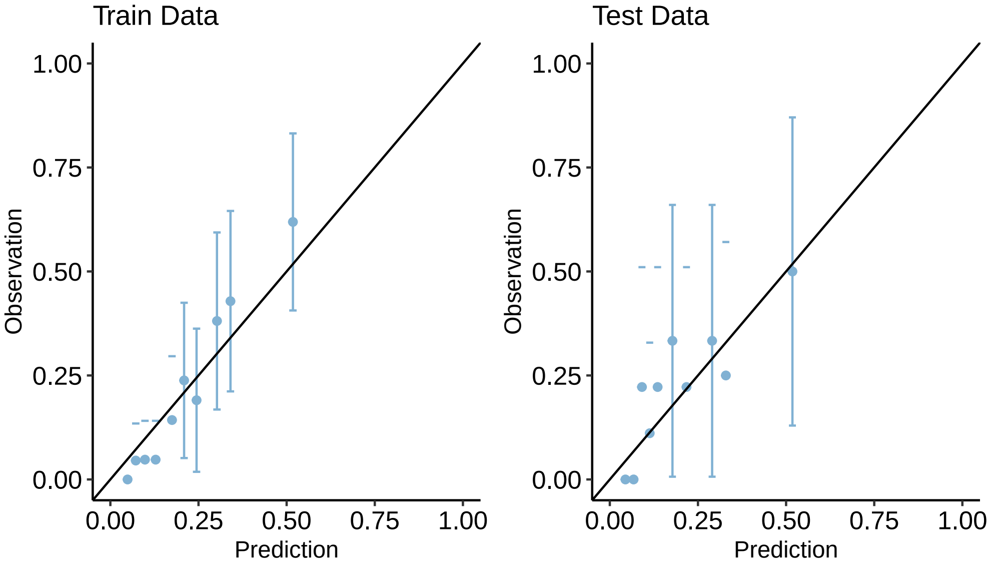


Figure 4: Calibration Plot Comparison



Appendix

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

# load necessary packages
library(tidyverse)
library(mice)
library(gt)
library(gtsummary)
library(kableExtra)
library(RColorBrewer)
library(scico)
library(caret)
library(glmnet)
library(pROC)
library(predtools)
library(gridExtra)
library(ggpubr)
# set working directory
setwd("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Data/")

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

data[, c("abst", "Var", "BA", "sex_ps", "NHW",
         "Black", "Hisp", "inc", "edu", "ftcd.5.mins",
         "otherdiag", "antidepmed", "mde_curr",
         "Only.Menthol")] <- lapply(data[, c("abst", "Var", "BA", "sex_ps", "NHW",
                                             "Black", "Hisp", "inc", "edu",
                                             "ftcd.5.mins", "otherdiag", "antidepmed",
                                             "mde_curr", "Only.Menthol")], as.factor)
# Recode factor levels in the dataset
averaged_data_factor <- data %>%
  mutate(abst = fct_recode(as.factor(abst), "Yes" = "1", "No" = "0"),
         inc = fct_recode(as.factor(inc),
                          "Less than $20,000" = "1",
                          "$20,000-35,000" = "2",
                          "$35,001-50,000" = "3",
                          "$50,001-75,000" = "4",
                          "More than $75,000" = "5"),
         sex_ps = fct_recode(as.factor(sex_ps), "Male" = "1", "Female" = "2"),
         edu = fct_recode(as.factor(edu),
                          "Grade School" = "1",
                          "Some high school" = "2",
                          "High school graduate or GED" = "3",
                          "Some college/technical school" = "4",
                          "College graduate" = "5"),
         ftcd.5.mins = fct_recode(as.factor(ftcd.5.mins), "Yes" = "1", "No" = "0"),
         otherdiag = fct_recode(as.factor(otherdiag), "Yes" = "1", "No" = "0"),
         antidepmed = fct_recode(as.factor(antidepmed), "Yes" = "1", "No" = "0"),
         mde_curr = fct_recode(as.factor(mde_curr), "Current MDD" = "1", "Past MDD" = "0"),
         Only.Menthol = fct_recode(as.factor(Only.Menthol), "Yes" = "1", "No" = "0"),
         race = case_when(Black == 1 ~ "Black",
                          Hisp == 1 ~ "Hispanic",
                          NHW == 1 ~ "Non-Hispanic White",
                          TRUE ~ "Other"),
         trt = case_when(Var == 1 & BA == 1 ~ "BASC + varenicline",
                         Var == 0 & BA == 1 ~ "BASC + placebo",
                         Var == 1 & BA == 0 ~ "ST + varenicline",
                         Var == 0 & BA == 0 ~ "ST + placebo",
                         TRUE ~ NA_character_))

averaged_data_factor <- averaged_data_factor %>%
  mutate(inc = fct_relevel(inc, "Less than $20,000", "$20,000-35,000",
                           "$35,001-50,000", "$50,001-75,000", "More than $75,000"),
         edu = fct_relevel(edu, "Grade School", "Some high school", "High school graduate or GED",
                           "Some college/technical school", "College graduate"))

# Now create the summary table
summary_table <- averaged_data_factor %>%
  select(-c("id", "Var", "BA", "Black", "Hisp", "NHW")) %>%
  tbl_summary(by = trt, label = list(abst ~ "Smoking abstinence",
                                     race ~ "Race",
                                     age_ps ~ "Age",
                                     sex_ps ~ "Sex",
                                     inc ~ "Income",
                                     edu ~ "Education",
                                     ftcd_score ~ "FTCD score",
                                     ftcd.5.mins ~ "Smoking within 5 mins of waking up",
                                     bdi_score_w00 ~ "BDI score",
                                     cpd_ps ~ "Cigarettes smoked per day",
                                     crv_total_pq1 ~ "Cigarette reward value",
                                     hedonsum_n_pq1 ~ "Pleasurable events (substitute reinforcers)",
                                     hedonsum_y_pq1 ~ "Pleasurable events (complementary reinforcers)",
                                     shaps_score_pq1 ~ "Anhedonia",
                                     otherdiag ~ "Other lifetime DSM-5 diagnosis",
                                     antidepmed ~ "Taking antidepressant",
                                     mde_curr ~ "Current vs. past MDD",
                                     NMR ~ "Nicotine metabolism ratio",
                                     Only.Menthol ~ "Exclusive mentholated cigarette user",
                                     readiness ~ "Readiness to quit smoking"),
              type = list(readiness ~ "continuous"),
              statistic = all_continuous() ~ "{mean} ({sd})",
              missing = "ifany",
              missing_text = "Missing") %>%
  add_overall(last = TRUE) %>%
  modify_spanning_header(update = all_stat_cols() ~ "**Behavioral and Pharmacological Treatment Assignme
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categorical") %>%
  bold_labels()

summary_table %>%
  as_kable_extra(booktabs = TRUE, caption = "Participant Characteristics by Treatment Arm",
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 7,
                            latex_options = c("repeat_header", "HOLD_position", "scale_down"))%>%
  column_spec(1, width = "3.5cm") %>%
```

```r
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  column_spec(6, width = "2cm") %>%
  row_spec(0, bold = TRUE, font_size = 7)
ggplot(averaged_data_factor, aes(x = abst, fill = inc)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(title = "Figure 1: Income Levels by Abstinence Status and Treatment Group",
       x = "Abstinence Status",
       y = "Proportion",
       fill = "Income Level") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
  theme(axis.title = element_text(size = 8),
        title = element_text(size = 10),
        axis.text = element_text(size = 8),
        legend.text = element_text(size = 8))

ggplot(averaged_data_factor, aes(x = abst, fill = edu)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(title = "Figure 2: Education Levels by Abstinence Status and Treatment Group",
       x = "Abstinence Status",
       y = "Proportion",
       fill = "Education Level") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
  theme(axis.title = element_text(size = 8),
        title = element_text(size = 10),
        axis.text = element_text(size = 8),
        legend.text = element_text(size = 8))

ggplot(averaged_data_factor, aes(x = abst, fill = race)) +
  geom_bar(position = "fill") +
  facet_wrap(~ trt) +
  labs(title = "Figure 2: Race by Abstinence Status and Treatment Group",
       x = "Abstinence Status",
       y = "Proportion",
       fill = "Race Group") +
  theme_minimal() +
  scale_fill_brewer(palette = "GnBu") +
  theme(axis.title = element_text(size = 8),
        title = element_text(size = 10),
        axis.text = element_text(size = 8),
        legend.text = element_text(size = 8))
ggplot(averaged_data_factor, aes(x = ftcd_score, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(title = "Density of FTCD Scores by Abstinence Status and Treatment Group",
       x = "FTCD Score",
       y = "Density",
       fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF"))

ggplot(averaged_data_factor, aes(x = shaps_score_pq1, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(title = "Density of Anhedonia by Abstinence Status and Treatment Group",
       x = "Anhedonia",
       y = "Density",
       fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF"))

ggplot(averaged_data_factor, aes(x = readiness, fill = abst)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~ trt) +
  labs(title = "Density of Readiness to Quit by Abstinence Status and Treatment Group",
       x = "Readiness to Quit",
       y = "Density",
       fill = "Abstinence Status") +
  theme_minimal() +
  scale_fill_manual(values = c("No" = "#FF9999", "Yes" = "#99CCFF"))
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1), oma = c(0, 0, 4, 0))
hist(averaged_data_factor$hedonsum_n_pq1, main = "", xlab = "hedonsum_n_pq1",
     col = "lightblue", breaks = 15, cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
hist(averaged_data_factor$hedonsum_y_pq1, main = "", xlab = "hedonsum_y_pq1",
     col = "lightblue", breaks = 15, cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
hist(averaged_data_factor$shaps_score_pq1, main = "", xlab = "shaps_score_pq1",
     col = "lightblue", breaks = 15, cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
hist(averaged_data_factor$NMR, main = "", xlab = "NMR",
     col = "lightblue", breaks = 15, cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
mtext("Figure : Distribution of Skewed Variables (Before Transformation)", outer = TRUE, cex = 0.8, font


# par(mfrow = c(2, 2), mar = c(4, 4, 2, 1), oma = c(0, 0, 4, 0))
# hist(averaged_data_factor$hedonsum_n_pq1, main = "", xlab = "hedonsum_n_pq1",
#      col = "lightblue", breaks = 15, cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
# hist(averaged_data_factor$hedonsum_y_pq1, main = "", xlab = "hedonsum_y_pq1",
#      col = "lightblue", breaks = 15, cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
# hist(averaged_data_factor$shaps_score_pq1, main = "", xlab = "shaps_score_pq1",
#      col = "lightblue", breaks = 15, cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
# hist(averaged_data_factor$NMR, main = "", xlab = "NMR",
#      col = "lightblue", breaks = 15, cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
# mtext("Figure : Distribution of Skewed Variables (After Transformation)", outer = TRUE, cex = 0.8, fo
# multiple imputation with m = 5
imputed_data <- mice(data, m = 5, method = 'pmm', maxit = 50, seed = 2550, printFlag = FALSE)

# extract the five imputed datasets
completed_datasets <- list()
for (i in 1:5) {
  completed_datasets[[i]] <- complete(imputed_data, i)
}
```

```r
# calculate average/mode of each missing variable
averaged_data <- completed_datasets[[1]]

for (var in names(averaged_data)) {
  if (any(is.na(data[[var]]))) {
    if (is.numeric(averaged_data[[var]])) {
      averaged_data[[var]] <- rowMeans(sapply(completed_datasets, function(x) x[[var]]))
    } else {
      averaged_data[[var]] <- apply(sapply(completed_datasets, function(x) x[[var]]), 1, function(vals)
        vals <- as.factor(vals)
        unique_vals <- unique(vals)
        unique_vals[which.max(tabulate(match(vals, unique_vals)))]
      })
    }
  }
}
new_data <- averaged_data %>%
  mutate(race = as.factor(case_when(Black == 1 ~ "Black",
                         Hisp == 1 ~ "Hispanic",
                         NHW == 1 ~ "Non-Hispanic White",
                         TRUE ~ "Other")),
         trt = as.factor(case_when(Var == 1 & BA == 1 ~ "BASC + varenicline",
                         Var == 0 & BA == 1 ~ "BASC + placebo",
                         Var == 1 & BA == 0 ~ "ST + varenicline",
                         Var == 0 & BA == 0 ~ "ST + placebo",
                         TRUE ~ NA_character_)),
         inc = fct_recode(as.factor(inc),
                          "Less than $20,000" = "1",
                          "$20,000-35,000" = "2",
                          "$35,001-50,000" = "3",
                          "$50,001-75,000" = "4",
                          "More than $75,000" = "5"),
         edu = fct_recode(as.factor(edu),
                          "Grade School" = "1",
                          "Some high school" = "2",
                          "High school graduate or GED" = "3",
                          "Some college/technical school" = "4",
                          "College graduate" = "5"))

new_data <- new_data %>%
  mutate(inc = fct_relevel(inc, "Less than $20,000", "$20,000-35,000",
                           "$35,001-50,000", "$50,001-75,000", "More than $75,000"),
         edu = fct_relevel(edu, "Grade School", "Some high school", "High school graduate or GED",
                           "Some college/technical school", "College graduate"))

# new_data$hedonsum_n_pq1 <- log(new_data$hedonsum_n_pq1)
# new_data$hedonsum_y_pq1 <- log(new_data$hedonsum_y_pq1)
# new_data$shaps_score_pq1 <- log(new_data$shaps_score_pq1)
# new_data$NMR <- log(new_data$NMR)
set.seed(2550)
train_index <- createDataPartition(new_data$trt, p = 0.7, list = FALSE)
train_data <- new_data[train_index, ]
test_data <- new_data[-train_index, ]
```

```
X <- model.matrix(abst ~ trt + age_ps + sex_ps + inc + edu + ftcd_score + ftcd.5.mins +
                  bdi_score_w00 + cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_y_pq1 +
                  shaps_score_pq1 + otherdiag + antidepmed + mde_curr + NMR + Only.Menthol +
                  readiness + race, data = train_data)[, -1]
y <- train_data$abst
cv_model <- cv.glmnet(X, y, family = "binomial", alpha = 1, nfolds = 10, nlambda = 100)
best_lambda <- cv_model$lambda.min
lasso_model <- glmnet(X, y, family = "binomial", alpha = 1, lambda = best_lambda)

X_interaction <- model.matrix(abst ~ trt * (ftcd_score + NMR + race + mde_curr),
                              data = train_data)[, -1]
y_interaction <- train_data$abst
cv_model_interaction <- cv.glmnet(X_interaction, y_interaction, family = "binomial",
                                  alpha = 1, nfolds = 10, nlambda = 100)
best_lambda_interaction <- cv_model_interaction$lambda.min
lasso_model_interaction <- glmnet(X_interaction, y_interaction, family = "binomial",
                                  alpha = 1, lambda = best_lambda_interaction)

prediction_train <- predict(lasso_model_interaction, X_interaction, type = "response")
roc_train <- roc(y_interaction, prediction_train)
auc_train <- auc(roc_train)
print(auc_train)
X_test_interaction <- model.matrix(abst ~ trt * (ftcd_score + NMR + race + mde_curr),
                                   data = test_data)[, -1]
y_test_interaction <- test_data$abst
prediction_test <- predict(lasso_model_interaction, X_test_interaction, type = "response")

roc_test <- roc(y_test_interaction, prediction_test)
auc_test <- auc(roc_test)
par(mfrow= c(1,2), oma = c(0, 0, 2, 0))
plot(roc_train, main = "Train Data", font.main = 1, cex.main = 0.8)
text(0.3, 0.2, paste("AUC =", round(auc_train, 3)), col = "blue", cex = 0.7)

plot(roc_test, main = "Test Data", font.main = 1, cex.main = 0.8)
text(0.3, 0.2, paste("AUC =", round(auc_test, 3)), col = "blue", cex = 0.7)
train_data$pred <- prediction_train
test_data$pred <- prediction_test
train_data <- train_data %>%
  mutate(abst_num = as.numeric(as.character(abst)))
test_data <- test_data %>%
  mutate(abst_num = as.numeric(as.character(abst)))
cal_plot_train <- calibration_plot(data = train_data, obs = "abst_num", pred = "pred", title = "Train Da
cal_plot_test <- calibration_plot(data = test_data, obs = "abst_num", pred = "pred", title = "Test Data

grid.arrange(cal_plot_train$calibration_plot,
             cal_plot_test$calibration_plot, ncol = 2,
             top = text_grob("Figure 4: Calibration Plot Comparison"))
```