Aim 1 Task 2

Yiming Miao & Kexin Wang

Load dataset and libraries

```r
library(mice)
library(MatchIt)
library(tidyr)
library(glue)
library(splines)
library(rlang)
library(ggplot2)
library(ggpubr)
set.seed(123)
data <- readRDS("data/ukbiobank.rds")
```

Select participants with 4 records after report of Parkinson's disease

```r
report_year <- as.numeric(format(as.Date(data$date_of_parkinsons_disease_report_f42032_0_0), "%Y"))
demo_feature <- c("eid", "sex_f31_0_0", "townsend_deprivation_index_at_recruitment_f189_0_0",
                  "ethnic_background_f21000_0_0", "age_at_recruitment_f21022_0_0")
no_data <- data[is.na(report_year), demo_feature]
no_data <- no_data[complete.cases(no_data),]
no_data$parkinson <- 0
yes_data <- data[(!is.na(report_year)), ]
year_recruitment <- yes_data$age_at_recruitment_f21022_0_0 + yes_data$year_of_birth_f34_0_0
report_year <- as.numeric(format(as.Date(yes_data$date_of_parkinsons_disease_report_f42032_0_0),
                                 "%Y"))
yes_data <- yes_data[year_recruitment >= report_year, demo_feature]
yes_data <- yes_data[complete.cases(yes_data),]
yes_data$parkinson <- 1
data_full <- rbind(yes_data, no_data)
```

Propensity score matching

```r
m1 <- glm(parkinson ~ sex_f31_0_0 + townsend_deprivation_index_at_recruitment_f189_0_0 +
            ethnic_background_f21000_0_0 + age_at_recruitment_f21022_0_0,
          family = binomial, data = data_full)
data_full$propensity_score <- predict(m1, type = "response")
m2 <- matchit(parkinson ~ propensity_score, data = data_full, method = "nearest")
matched_data <- match.data(m2)
```

Select features for longitudinal analysis

```r
var_table <- read.csv("variable_table.csv")
use_category <- setdiff(levels(as.factor(var_table$category)),
                        c("Actigraphy", "Demographic"))
demo_data <- data[data$eid %in% matched_data$eid, ]
demo_data$parkinson <- ifelse(is.na(demo_data$date_of_parkinsons_disease_report_f42032_0_0),
                           "No Parkinson", "Parkinson")
demo_data <- demo_data[, c("eid", "parkinson", "sex_f31_0_0",
                           "townsend_deprivation_index_at_recruitment_f189_0_0",
                           "ethnic_background_f21000_0_0", "age_at_recruitment_f21022_0_0")]
colnames(demo_data) <- c("eid", "parkinson", "sex", "index", "ethnic", "age")

data1 <- data[data$eid %in% matched_data$eid,]
data1 <- data1[, colnames(data1) %in% var_table$variable[var_table$category %in% use_category]]
data1 <- data1[sapply(data1, is.numeric)]
feature <- gsub("(.*)_[^_]*_[^_]*$", "\\1", colnames(data1))
feature <- names(table(feature))[table(feature) == 4]
data2 <- data1[, grep(paste(c(feature, "^parkinson"), collapse="|"),
                      colnames(data1))]
```

Process some negative values

```r
process_feature <- c(paste0("tea_intake_f1488", "_", 0:3, "_0"),
                     paste0("time_spent_driving_f1090", "_", 0:3, "_0"),
                     paste0("longest_period_of_depression_f4609", "_", 0:3, "_0"),
                     paste0("average_weekly_intake_of_other_alcoholic_drinks_f5364",
                            "_", 0:3, "_0"),
                     paste0("average_weekly_red_wine_intake_f1568", "_", 0:3, "_0"),
                     paste0("sleep_duration_f1160", "_", 0:3, "_0"),
                     paste0("time_spent_driving_f1090", "_", 0:3, "_0"),
                     paste0("time_spent_using_computer_f1080", "_", 0:3, "_0"))
for (col in process_feature) {
  data2[[col]][data2[[col]] == -10] <- 0
  data2[[col]][data2[[col]] == -1] <- NA
  data2[[col]][data2[[col]] == -3] <- NA
}
```

Example

```r
var_select <- feature[27]
form2 <- formula(paste0(var_select,
                        "~bs(event, degree =2)*parkinson + sex + index + ethnic + age"))
form1 <- formula(paste0(var_select,
                        "~bs(event, degree =2):parkinson + sex + index + ethnic + age"))
form0 <- formula(paste0(var_select,
                        "~bs(event, degree =2)+parkinson + sex + index + ethnic + age"))
# fit2 <- nlme::lme(form2, random = ~1|eid, data = df, na.action = na.exclude, method = "ML")
# fit1 <- nlme::lme(form1, random = ~1|eid, data = df, na.action = na.exclude, method = "ML")
# fit0 <- nlme::lme(form0, random = ~1|eid, data = df, na.action = na.exclude, method = "ML")
```

Helpers

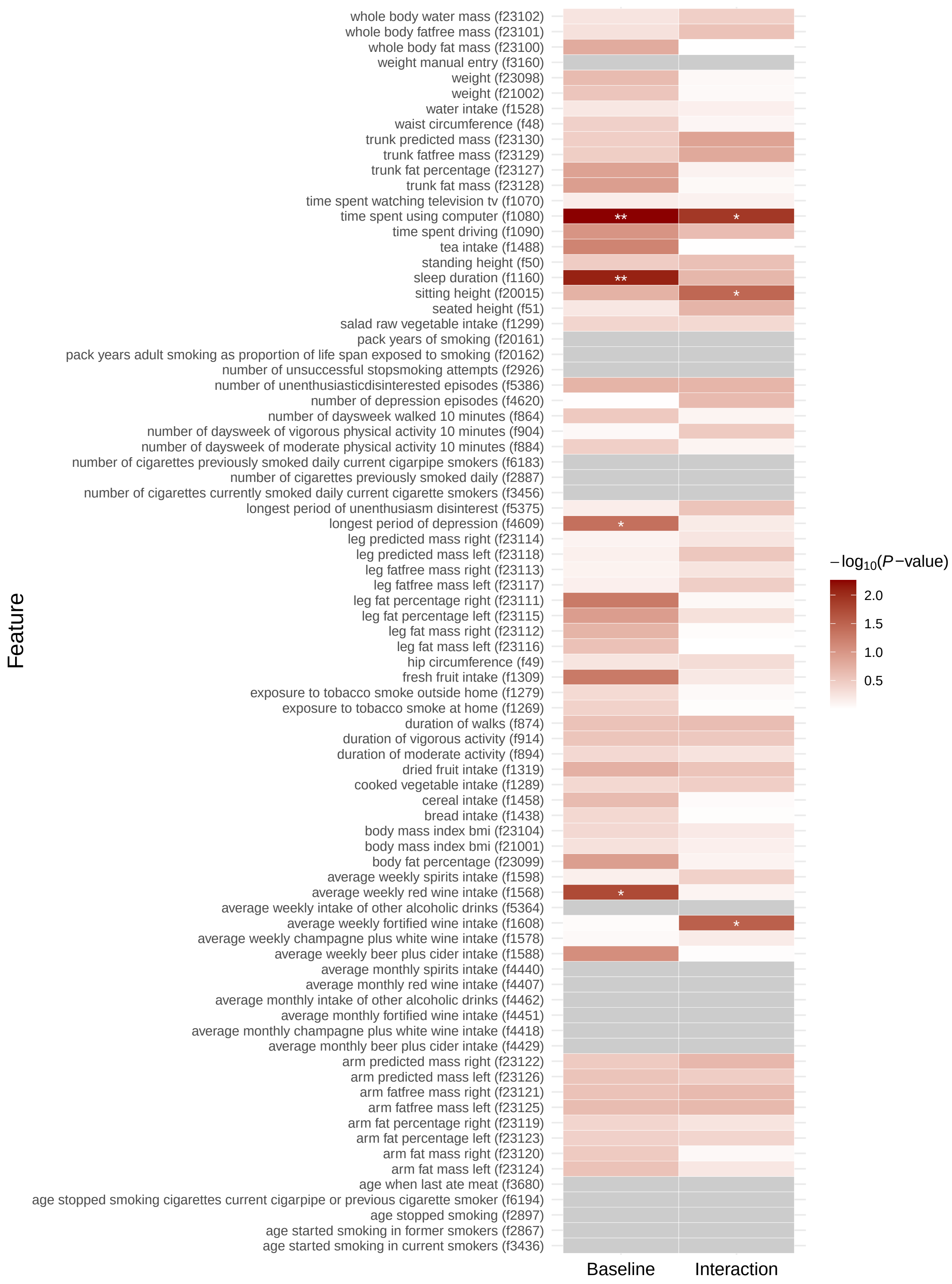
```r
mixed_model <- function(var_select, var_data, demo_data) {
  var_data <- var_data[, grep(paste0("^", var_select), colnames(var_data))]
  df <- cbind(demo_data, var_data)
  df <- pivot_longer(df,
                     cols = starts_with(var_select),
                     names_to = c(".value", "event"),
                     names_pattern = glue::glue("^({var_select})_(.)_.*"))
  df$event_id <- paste0(df$eid, "_", df$event)
  df$eid <- as.factor(df$eid)
  df$parkinson <- as.factor(df$parkinson)
  levels(df$parkinson) = c("Yes", "No")
  df$sex <- as.factor(df$sex)
  df$index <- as.numeric(df$index)
  df$ethnic <- as.factor(df$ethnic)
  df$age <- as.numeric(df$age)
  df$event <- as.integer(df$event)

  df <- df[!is.na(df[[var_select]]) & !is.na(df$parkinson) & !is.na(df$event) &
           !is.na(df$sex) & !is.na(df$index) &
           !is.na(df$ethnic) & !is.na(df$age),]

  if (dim(df)[1] >= 30) {
    form2 <- formula(paste0(var_select,
                            "~bs(event, degree=2)*parkinson + sex + index + ethnic + age"))
    form1 <- formula(paste0(var_select,
                            "~bs(event, degree=2):parkinson + sex + index + ethnic + age"))
    form0 <- formula(paste0(var_select,
                            "~bs(event, degree=2)+parkinson + sex + index + ethnic + age"))

    fit2 <- tryCatch({
      nlme::lme(form2, random = ~1|eid, data = df, na.action = na.exclude, method = "ML")},
      error = function(e) NULL)

    fit1 <- tryCatch({
      nlme::lme(form1, random = ~1|eid, data = df, na.action = na.exclude, method = "ML")},
      error = function(e) NULL)

    fit0 <- tryCatch({
      nlme::lme(form0, random = ~1|eid, data = df, na.action = na.exclude, method = "ML")},
      error = function(e) NULL)

    # Only proceed if all models are successfully fitted
    if (!is.null(fit2) && !is.null(fit1)) {
      pval_base <- anova(fit2, fit1)$`p-value`[2]
    } else {
      pval_base <- NA
    }
    if (!is.null(fit2) && !is.null(fit0)) {
      pval_int <- anova(fit2, fit0)$`p-value`[2]
    } else {
      pval_int <- NA
    }
    if (!is.null(fit2) && !is.null(fit1) && !is.null(fit0)) {
      df_plot <- df
      df_plot$fit <- predict(fit2, level = 0)
      df_plot$sex <- factor(levels(df_plot$sex)[1],
                            levels = levels(df_plot$sex))
      df_plot$index <- mean(df_plot$index)
      df_plot$ethnic <- factor(levels(df_plot$ethnic)[1],
                               levels = levels(df_plot$ethnic))
      df_plot$age <- mean(df_plot$age)
      df_plot$pred_adj <- predict(fit2, level = 0, newdata = df_plot)
      var_name_parts <- strsplit(var_select, "_")[[1]]
      var_name_parts[length(var_name_parts)] <-
        paste0("(", var_name_parts[length(var_name_parts)], ")")
      var_name <- paste(var_name_parts, collapse = " ")

      trend_plot <- ggplot(data = df_plot, aes(x = event, y = !!sym(var_select),
                                               color = parkinson)) +
        geom_jitter(width = 0, height = 0, size = 1.0, alpha = 0.4) +
        geom_line(aes(y = pred_adj), size = 0.9) +
        theme_minimal() +
        theme(legend.position = "top") +
        labs(x = "Event", y = var_name, color = "Parkinson's disease") +
        xlim(c(0, 3))
    } else {
      trend_plot <- NULL
    }
  } else {
    trend_plot <- NA
    pval_base <- NA
    pval_int <- NA
  }
  return(list(pval_base = pval_base, pval_int = pval_int, trend_plot = trend_plot))
}
```

Longitudinal trend analysis

```
feature_name <- sapply(strsplit(feature, "_"), function(name_parts) {
  name_parts[length(name_parts)] <- paste0("(", name_parts[length(name_parts)], ")")
  paste(name_parts, collapse = " ")
})
pval_mat <- matrix(NA, nrow = length(feature), ncol = 2)
rownames(pval_mat) <- feature_name
colnames(pval_mat) <- c("Baseline", "Interaction")
trend_plot_list <- list()
for (i in 1:length(feature)) {
  pval_tmp <- mixed_model(feature[i], data2, demo_data)
  pval_mat[i, 1] <- pval_tmp$pval_base
  pval_mat[i, 2] <- pval_tmp$pval_int
  trend_plot_list[[i]] <- pval_tmp$trend_plot
}
```

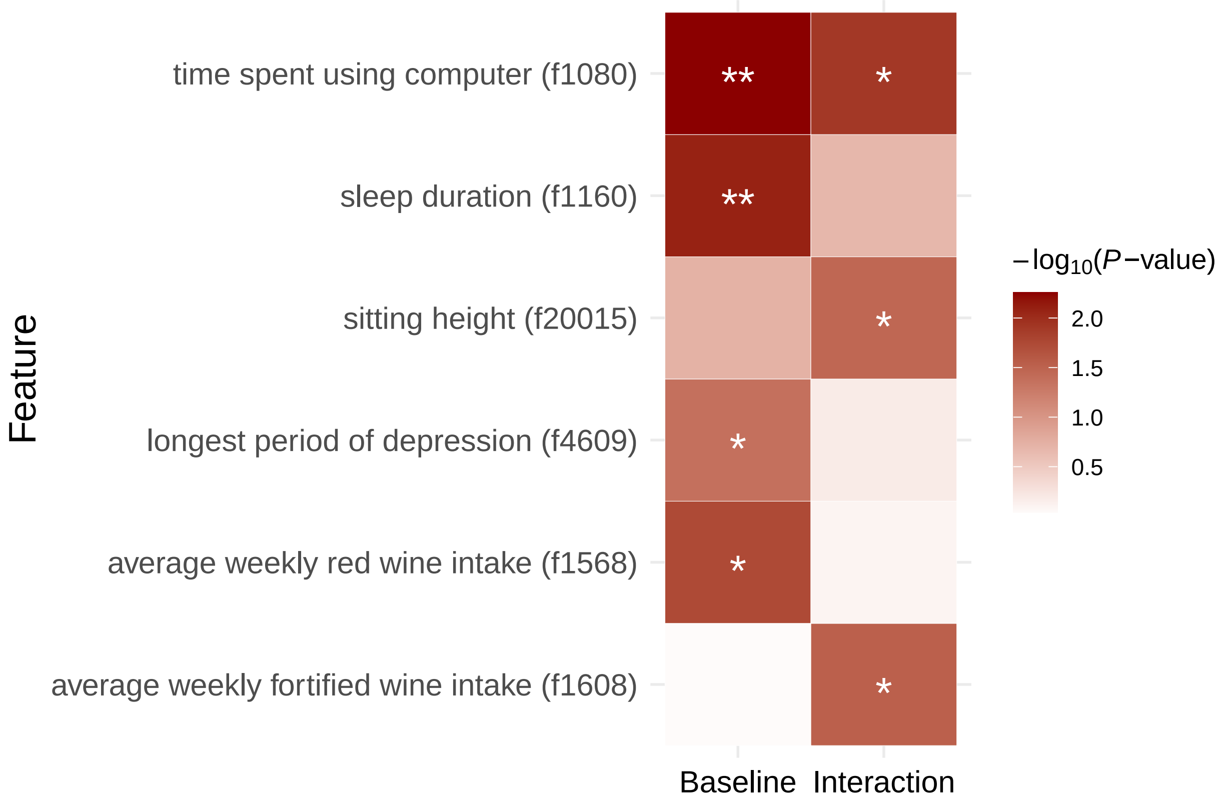
Trend comparison heatmap

```r
log_pval <- -log10(pval_mat)
df_pval <- pivot_longer(as.data.frame(log_pval), cols = everything())
df_pval$feature <- factor(rep(rownames(pval_mat), each = 2))
df_pval$name <- factor(df_pval$name)

ggplot(df_pval, aes(x = name, y = feature, fill = value)) +
  geom_tile(color = "white", size = 0.1) +
  scale_fill_gradient2(low = "white", high = "darkred", na.value = "grey80",
                       name = expression(-log[10]*"("*italic("P")*"-value"*")")) +
  theme_minimal() +
  ylab("Feature") +
  theme(axis.title.y = element_text(size = 15),
        axis.title.x = element_blank(),
        axis.text.x = element_text(size = 12, color = "black")) +
  geom_text(aes(label = ifelse(value >= -log10(0.001), "***",
                               ifelse(value >= -log10(0.01), "**",
                                      ifelse(value >= -log10(0.05), "*", "")))),
              size = 4, vjust = 0.75, color = "white")
```


whole body water mass (f23102)
whole body fatfree mass (f23101)
whole body fat mass (f23100)
weight manual entry (f3160)
weight (f23098)
weight (f21002)
water intake (f1528)
waist circumference (f48)
trunk predicted mass (f23130)
trunk fatfree mass (f23129)
trunk fat percentage (f23127)
trunk fat mass (f23128)
time spent watching television tv (f1070)
time spent using computer (f1080)
time spent driving (f1090)
tea intake (f1488)
standing height (f50)
sleep duration (f1160)
sitting height (f20015)
seated height (f51)
salad raw vegetable intake (f1299)
pack years of smoking (f20161)
pack years adult smoking as proportion of life span exposed to smoking (f20162)
number of unsuccessful stopsmoking attempts (f2926)
number of unenthusiasticdisinterested episodes (f5386)
number of depression episodes (f4620)
number of daysweek walked 10 minutes (f864)
number of daysweek of vigorous physical activity 10 minutes (f904)
number of daysweek of moderate physical activity 10 minutes (f884)
number of cigarettes previously smoked daily current cigarpipe smokers (f6183)
number of cigarettes previously smoked daily (f2887)
number of cigarettes currently smoked daily current cigarette smokers (f3456)
longest period of unenthusiasm disinterest (f5375)
longest period of depression (f4609)
leg predicted mass right (f23114)
leg predicted mass left (f23118)
leg fatfree mass right (f23113)
leg fatfree mass left (f23117)
leg fat percentage right (f23111)
leg fat percentage left (f23115)
leg fat mass right (f23112)
leg fat mass left (f23116)
hip circumference (f49)
fresh fruit intake (f1309)
exposure to tobacco smoke outside home (f1279)
exposure to tobacco smoke at home (f1269)
duration of walks (f874)
duration of vigorous activity (f914)
duration of moderate activity (f894)
dried fruit intake (f1319)
cooked vegetable intake (f1289)
cereal intake (f1458)
bread intake (f1438)
body mass index bmi (f23104)
body mass index bmi (f21001)
body fat percentage (f23099)
average weekly spirits intake (f1598)
average weekly red wine intake (f1568)
average weekly intake of other alcoholic drinks (f5364)
average weekly fortified wine intake (f1608)
average weekly champagne plus white wine intake (f1578)
average weekly beer plus cider intake (f1588)
average monthly spirits intake (f4440)
average monthly red wine intake (f4407)
average monthly intake of other alcoholic drinks (f4462)
average monthly fortified wine intake (f4451)
average monthly champagne plus white wine intake (f4418)
average monthly beer plus cider intake (f4429)
arm predicted mass right (f23122)
arm predicted mass left (f23126)
arm fatfree mass right (f23121)
arm fatfree mass left (f23125)
arm fat percentage right (f23119)
arm fat percentage left (f23123)
arm fat mass right (f23120)
arm fat mass left (f23124)
age when last ate meat (f3680)
age stopped smoking cigarettes current cigarpipe or previous cigarette smoker (f6194)
age stopped smoking (f2897)
age started smoking in former smokers (f2867)
age started smoking in current smokers (f3436)
Feature
Baseline
Interaction
**
*
**
*
*
*
*
$-\log_{10}(P\text{−value})$
2.0
1.5
1.0
0.5

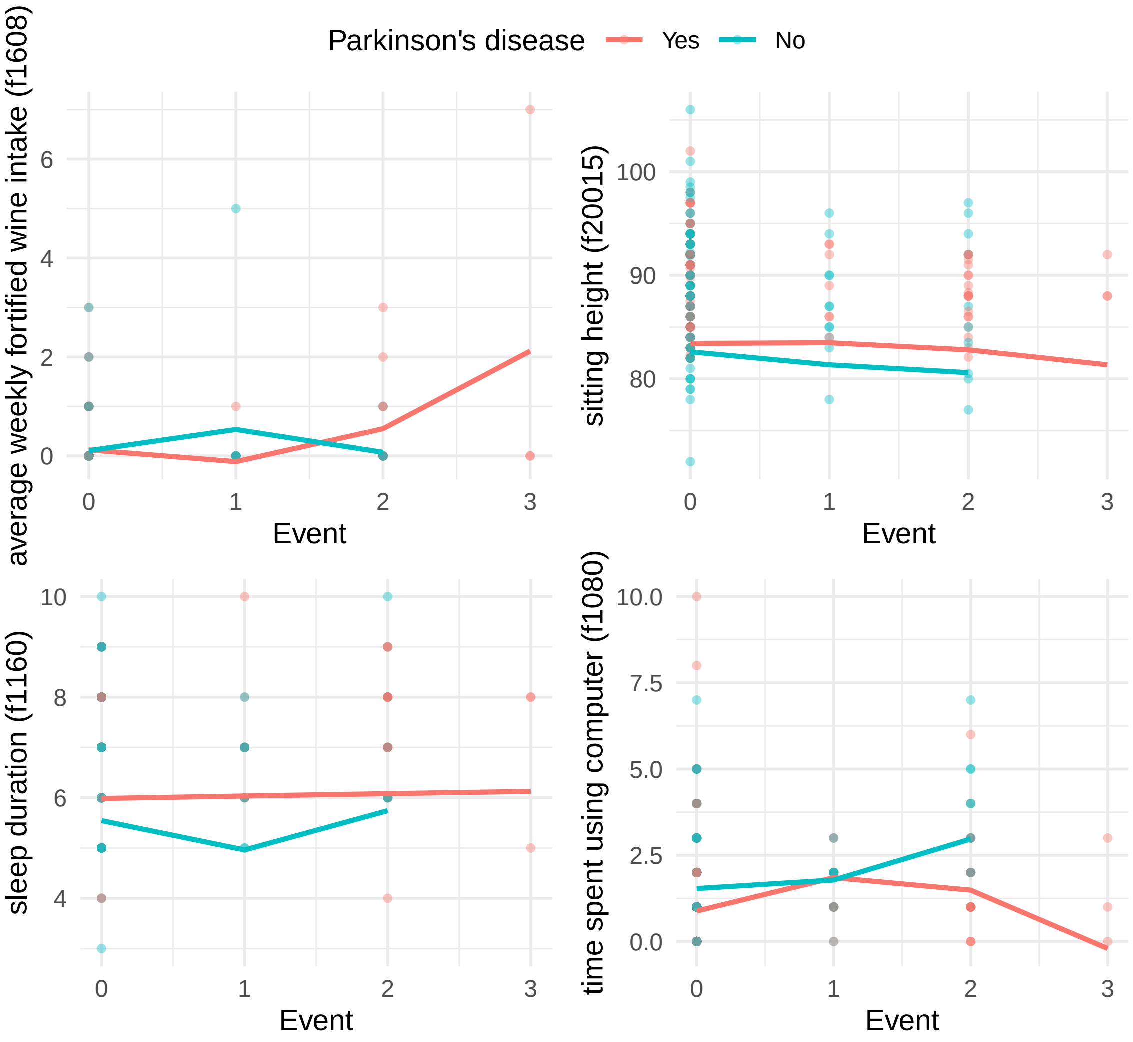
Trend comparison heatmap (significant features)

```r
idx_sig <- which(apply(pval_mat, 1, function(x) any(x <= 0.05, na.rm = TRUE)))
sig_pval_mat <- pval_mat[idx_sig, ]
log_pval <- -log10(sig_pval_mat)
df_pval <- pivot_longer(as.data.frame(log_pval), cols = everything())
df_pval$feature <- factor(rep(rownames(sig_pval_mat), each = 2))
df_pval$name <- factor(df_pval$name)

ggplot(df_pval, aes(x = name, y = feature, fill = value)) +
  geom_tile(color = "white", size = 0.1) +
  scale_fill_gradient2(low = "white", high = "darkred", na.value = "grey80",
                       name = expression(-log[10]*"("*italic("P")*"-value"*")")) +
  theme_minimal() +
  ylab("Feature") +
  theme(axis.title.y = element_text(size = 15),
        axis.title.x = element_blank(),
        axis.text.x = element_text(size = 12, color = "black"),
        axis.text.y = element_text(size = 12)) +
  geom_text(aes(label = ifelse(value >= -log10(0.001), "***",
                               ifelse(value >= -log10(0.01), "**",
                                      ifelse(value >= -log10(0.05), "*", "")))),
            size = 6, vjust = 0.75, color = "white")
```



Longitudinal trend (significant features)

```
idx_plot <- c(22, 63, 64, 68)
ggarrange(plotlist = trend_plot_list[idx_plot],
          nrow = 3, ncol = 2, common.legend = TRUE, legend = "top")
```


Parkinson's disease
Yes
No
average weekly fortified wine intake (f1608)
Event
sitting height (f20015)
Event
sleep duration (f1160)
Event
time spent using computer (f1080)
Event