NHANES 2021-2023: BMI and Blood Pressure Analysis

2025-10-17

Contents


```
# Introduction & Setup
# Purpose: Observe association between BMI and mean SBP among adults >=20 in NHANES 2021-2023 and wheth
# Also explore BMI distribution by education & race, and BP trial variability.

# Reproducibility notes: edit `data_dir` to point to local folder with NHANES XPT files
data_dir <- "C:/Users/Edison/Downloads/      "

pkgs <- c("tidyverse","haven","janitor","stringr","scales","skimr","naniar","broom","ggpubr","knitr")
new_pkgs <- setdiff(pkgs, installed.packages()[,"Package"])
if(length(new_pkgs)) install.packages(new_pkgs, repos = "https://cloud.r-project.org")

lapply(pkgs, library, character.only = TRUE)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'janitor'
##
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
##
##
##
## Attaching package: 'scales'
##
##
## The following object is masked from 'package:purrr':
##
##     discard
##
##
## The following object is masked from 'package:readr':
##
##     col_factor
##
##
##
## Attaching package: 'naniar'
##
##
## The following object is masked from 'package:skimr':
##
##     n_complete

## [[1]]
##  [1] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"
##  [7] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics"
## [13] "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[2]]
##  [1] "haven"     "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"
##  [7] "readr"     "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"
## [13] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[3]]
##  [1] "janitor"   "haven"     "lubridate" "forcats"   "stringr"   "dplyr"
##  [7] "purrr"     "readr"     "tidyr"     "tibble"    "ggplot2"   "tidyverse"
## [13] "stats"     "graphics"  "grDevices" "utils"     "datasets"  "methods"
## [19] "base"
##
## [[4]]
##  [1] "janitor"   "haven"     "lubridate" "forcats"   "stringr"   "dplyr"
##  [7] "purrr"     "readr"     "tidyr"     "tibble"    "ggplot2"   "tidyverse"
## [13] "stats"     "graphics"  "grDevices" "utils"     "datasets"  "methods"
## [19] "base"
```

```
##
## [[5]]
##  [1] "scales"    "janitor"   "haven"     "lubridate" "forcats"   "stringr"
##  [7] "dplyr"     "purrr"     "readr"     "tidyr"     "tibble"    "ggplot2"
## [13] "tidyverse" "stats"     "graphics"  "grDevices" "utils"     "datasets"
## [19] "methods"   "base"
##
## [[6]]
##  [1] "skimr"     "scales"    "janitor"   "haven"     "lubridate" "forcats"
##  [7] "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"     "tibble"
## [13] "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices" "utils"
## [19] "datasets"  "methods"   "base"
##
## [[7]]
##  [1] "naniar"    "skimr"     "scales"    "janitor"   "haven"     "lubridate"
##  [7] "forcats"   "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"
## [13] "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices"
## [19] "utils"     "datasets"  "methods"   "base"
##
## [[8]]
##  [1] "broom"     "naniar"    "skimr"     "scales"    "janitor"   "haven"
##  [7] "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"     "readr"
## [13] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics"
## [19] "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[9]]
##  [1] "ggpubr"    "broom"     "naniar"    "skimr"     "scales"    "janitor"
##  [7] "haven"     "lubridate" "forcats"   "stringr"   "dplyr"     "purrr"
## [13] "readr"     "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"
## [19] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[10]]
##  [1] "knitr"     "ggpubr"    "broom"     "naniar"    "skimr"     "scales"
##  [7] "janitor"   "haven"     "lubridate" "forcats"   "stringr"   "dplyr"
## [13] "purrr"     "readr"     "tidyr"     "tibble"    "ggplot2"   "tidyverse"
## [19] "stats"     "graphics"  "grDevices" "utils"     "datasets"  "methods"
## [25] "base"
```

```r
options(dplyr.summarise.inform = FALSE)

# Create outputs folder
if(!dir.exists("outputs")) dir.create("outputs")

# Helper: safe read XPT
safe_read_xpt <- function(path){
  if(file.exists(path)) read_xpt(path) %>% clean_names() else stop(paste0("File not found: ", path))
}

# Check for LaTeX engine (TinyTeX) for PDF rendering
if(!requireNamespace("tinytex", quietly = TRUE)){
  message("tinytex not installed. Rendering to PDF requires a LaTeX distribution.\nInstalling tinytex pa
  install.packages("tinytex")
}
if(!tinytex::is_tinytex()){
```

```
  message("tinytex distribution not installed. You can install it via tinytex::install_tinytex() if you
}
## Global knitr chunk options to improve PDF layout
knitr::opts_chunk$set(
  echo = TRUE,
  message = FALSE,
  warning = FALSE,
  fig.width = 7,
  fig.height = 4,
  out.width = "\\linewidth",
  dpi = 150
)

options(width = 80)

# Set a slightly smaller base font for ggplot to fit PDF
theme_set(ggplot2::theme_minimal(base_size = 10))
```

Data loading

```
# Read NHANES files (edit filenames if necessary)
demo <- safe_read_xpt(file.path(data_dir, "DEMO_L.XPT"))
bmx  <- safe_read_xpt(file.path(data_dir, "BMX_L.XPT"))
bpx  <- safe_read_xpt(file.path(data_dir, "BPXO_L.XPT"))

# Quick glimpse
skimr::skim(demo)
```

Table 1: Data summary

Name	demo
Number of rows	11933
Number of columns	27
Column type frequency:	
numeric	27
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
seqn	0	1.00	136344.00	3444.90	130378.00	133361.00	136344.00	139327.00	142310.0	
sddsrvyr	0	1.00	12.00	0.00	12.00	12.00	12.00	12.00	12.0	
ridstatr	0	1.00	1.74	0.44	1.00	1.00	2.00	2.00	2.0	
riagendr	0	1.00	1.53	0.50	1.00	1.00	2.00	2.00	2.0	
ridageyr	0	1.00	38.32	25.60	0.00	13.00	37.00	62.00	80.0	
ridagemn	11556	0.03	11.63	6.81	0.00	6.00	11.00	17.00	24.0	
ridreth1	0	1.00	3.10	1.08	1.00	3.00	3.00	4.00	5.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ridreth3	0	1.00	3.32	1.52	1.00	3.00	3.00	4.00	7.0	
ridexmon	3073	0.74	1.52	0.50	1.00	1.00	2.00	2.00	2.0	
ridexagm	9146	0.23	121.91	67.16	0.00	66.00	122.00	179.50	239.0	
dmqmiliz	3632	0.70	1.92	0.28	1.00	2.00	2.00	2.00	7.0	
dmdborn4	19	1.00	1.16	0.36	1.00	1.00	1.00	1.00	2.0	
dmdyrusr	10058	0.16	7.33	15.83	1.00	3.00	6.00	6.00	99.0	
dmdeduc2	4139	0.65	3.80	1.15	1.00	3.00	4.00	5.00	9.0	
dmdmartz	4141	0.65	1.78	3.10	1.00	1.00	1.00	2.00	99.0	
ridexprg	10430	0.13	2.24	0.49	1.00	2.00	2.00	3.00	3.0	
dmdhhsiz	0	1.00	3.24	1.70	1.00	2.00	3.00	4.00	7.0	
dmdhrgnd	7818	0.34	1.56	0.50	1.00	1.00	2.00	2.00	2.0	
dmdhragz	7809	0.35	2.54	0.64	1.00	2.00	2.00	3.00	4.0	
dmdhredz	8187	0.31	2.17	0.66	1.00	2.00	2.00	3.00	3.0	
dmdhrmaz	7913	0.34	1.38	0.68	1.00	1.00	1.00	2.00	3.0	
dmdhsedz	9806	0.18	2.28	0.69	1.00	2.00	2.00	3.00	3.0	
wtint2yr	0	1.00	27404.14	19449.16	4584.46	14331.75	21670.19	33831.33	170968.3	
wtmec2yr	0	1.00	27404.14	27962.96	0.00	0.00	21717.85	38341.15	227108.3	
sdmvstra	0	1.00	179.92	4.31	173.00	176.00	180.00	184.00	187.0	
sdmvpsu	0	1.00	1.49	0.50	1.00	1.00	1.00	2.00	2.0	
indfmpir	2041	0.83	2.71	1.67	0.00	1.18	2.50	4.50	5.0	

```
skimr::skim(bmx)
```

Table 3: Data summary

Name	bmx
Number of rows	8860
Number of columns	22
Column type frequency:	
numeric	22
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
seqn	0	1.00	136345.83	3453.78	130378.0	133319.75	136377.5	139336.2	142310.0	
bmdstats	0	1.00	1.13	0.50	1.0	1.00	1.0	1.0	4.0	
bmxwt	106	0.99	70.55	30.39	2.7	54.20	71.7	89.1	248.2	
bmiwt	8515	0.04	2.88	0.62	1.0	3.00	3.0	3.0	4.0	
bmxrecum	8406	0.05	84.33	14.06	48.5	73.48	84.7	96.1	118.8	
bmirecum	8842	0.00	1.00	0.00	1.0	1.00	1.0	1.0	1.0	
bmxhead	8790	0.01	41.93	2.80	34.4	40.20	42.4	44.0	46.5	
bmihead	8860	0.00	NaN	NA	NA	NA	NA	NA	NA	
bmxht	361	0.96	159.66	19.86	79.1	154.40	163.6	172.1	200.7	
bmiht	8726	0.02	2.31	0.95	1.0	1.00	3.0	3.0	3.0	
bmxbmi	389	0.96	27.25	8.14	11.1	21.60	26.4	31.7	74.8	
bmdbmic	6368	0.28	2.56	0.88	1.0	2.00	2.0	3.0	4.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
bmxleg	1525	0.83	38.13	3.86	24.9	35.50	38.1	40.8	51.6	
bmileg	8464	0.04	1.00	0.00	1.0	1.00	1.0	1.0	1.0	
bmxarml	292	0.97	35.11	6.18	10.0	33.60	36.5	39.0	49.2	
bmiarml	8660	0.02	1.00	0.00	1.0	1.00	1.0	1.0	1.0	
bmxarmc	298	0.97	30.56	7.37	12.0	26.40	31.2	35.4	63.3	
bmiarmc	8655	0.02	1.00	0.00	1.0	1.00	1.0	1.0	1.0	
bmxwaist	670	0.92	92.12	22.05	39.8	77.50	92.7	107.0	187.0	
bmiwaist	8513	0.04	1.00	0.00	1.0	1.00	1.0	1.0	1.0	
bmxhip	2084	0.76	106.26	14.66	69.9	96.40	103.7	113.5	187.1	
bmihip	8499	0.04	1.00	0.00	1.0	1.00	1.0	1.0	1.0	

```
skimr::skim(bpx)
```

Table 5: Data summary

Name	bpx
Number of rows	7801
Number of columns	12
Column type frequency:	
character	1
numeric	11
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
bpaoarm	0	1	0	1	147	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
seqn	0	1.00	136349.49	3449.49	130378	133335	136382	139325	142310	
bpaocsz	190	0.98	3.52	0.67	2	3	4	4	5	
bpxosy1	284	0.96	119.29	18.56	61	106	117	130	232	
bpxodi1	284	0.96	72.75	11.90	33	64	72	80	142	
bpxosy2	296	0.96	119.08	18.57	59	106	116	129	233	
bpxodi2	296	0.96	72.09	11.85	32	64	71	79	139	
bpxosy3	321	0.96	118.92	18.50	50	106	116	129	232	
bpxodi3	321	0.96	71.81	11.77	24	64	71	79	136	
bpxopls1	284	0.96	72.34	12.72	35	63	71	80	158	
bpxopls2	296	0.96	73.09	12.78	32	64	72	81	141	
bpxopls3	321	0.96	73.69	12.89	31	65	73	82	154	

Week 5: BMI & SBP cleaning

Build raw variables and compute mean BP

```
# Detect BP columns
sbp_cols <- names(bpx)[stringr::str_detect(names(bpx), "^bpxo?sy[1-3]$")]
dbp_cols <- names(bpx)[stringr::str_detect(names(bpx), "^bpxo?di[1-3]$")]

bpx_summary <- bpx %>%
  transmute(seqn,
            mean_sbp = rowMeans(select(., any_of(sbp_cols)), na.rm = TRUE),
            mean_dbp = rowMeans(select(., any_of(dbp_cols)), na.rm = TRUE)) %>%
  mutate(mean_sbp = ifelse(is.nan(mean_sbp), NA_real_, mean_sbp),
         mean_dbp = ifelse(is.nan(mean_dbp), NA_real_, mean_dbp))

# Prepare demographics and BMI
demo <- demo %>% mutate(riagendr = as.numeric(riagendr))
demo_sex <- demo %>% transmute(seqn, age = ridageyr, sex = factor(riagendr, levels=c(1,2), labels=c("Ma

bmi_raw <- bmx %>% transmute(seqn, bmi_raw = bmxbmi)

dat_raw <- demo_sex %>% left_join(bmi_raw, by = "seqn") %>% left_join(bpx_summary, by = "seqn") %>% fil

# Save a small head to outputs for inspecting
write.csv(head(dat_raw, 50), file = "outputs/dat_raw_head.csv", row.names = FALSE)
```

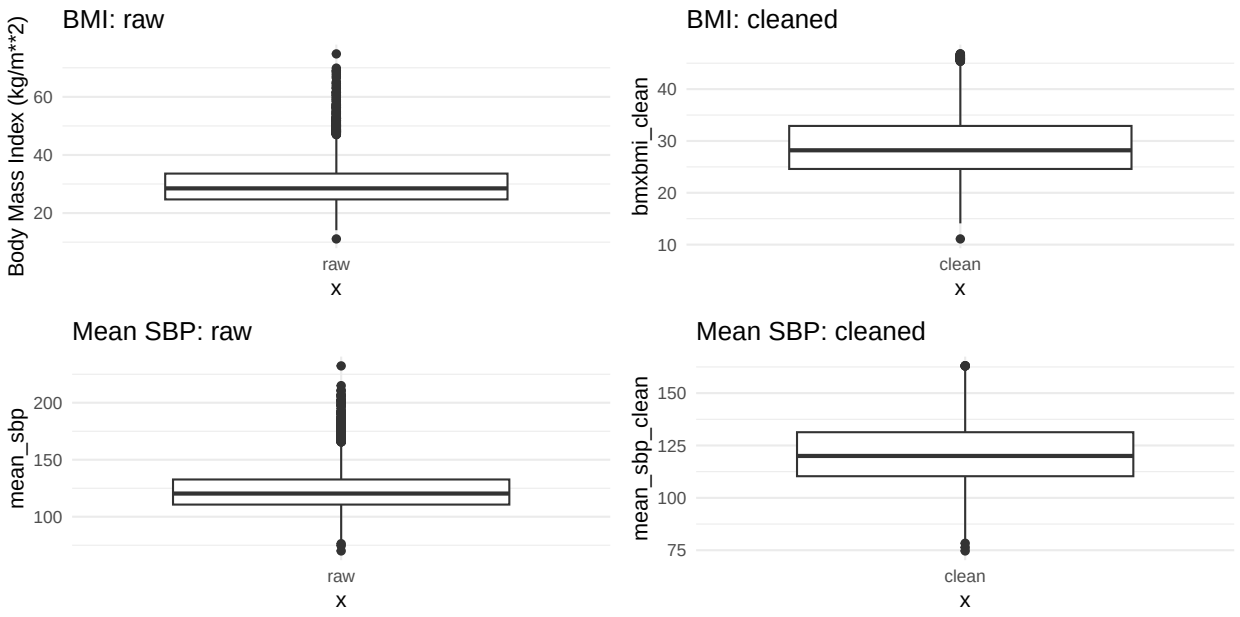
Outlier cleaning (physiologic + IQR + MAD)

```
# BMI cleaning
BMI_LO <- 10; BMI_HI <- 80
bmi_clean <- bmx %>% transmute(seqn, bmxbmi) %>% mutate(
  q1 = quantile(bmxbmi, 0.25, na.rm = TRUE),
  q3 = quantile(bmxbmi, 0.75, na.rm = TRUE),
  iqr = q3 - q1,
  lo_iqr = q1 - 1.5*iqr,
  hi_iqr = q3 + 1.5*iqr,
  med = median(bmxbmi, na.rm = TRUE),
  madv = mad(bmxbmi, na.rm = TRUE),
  z = ifelse(madv > 0, (bmxbmi - med)/madv, 0),
  flag = (bmxbmi < BMI_LO | bmxbmi > BMI_HI) | (bmxbmi < lo_iqr | bmxbmi > hi_iqr) | (abs(z) > 3.5),
  bmxbmi_clean = ifelse(flag, NA_real_, bmxbmi)
) %>% select(seqn, bmxbmi_clean)

# SBP cleaning
SBP_LO <- 70; SBP_HI <- 260
sbp_clean <- bpx_summary %>% transmute(seqn, mean_sbp) %>% mutate(
  q1 = quantile(mean_sbp, 0.25, na.rm = TRUE),
  q3 = quantile(mean_sbp, 0.75, na.rm = TRUE),
  iqr = q3 - q1,
  lo_iqr = q1 - 1.5*iqr,
  hi_iqr = q3 + 1.5*iqr,
```

```
  med = median(mean_sbp, na.rm = TRUE),
  madv = mad(mean_sbp, na.rm = TRUE),
  z = ifelse(madv > 0, (mean_sbp - med)/madv, 0),
  flag = (mean_sbp < SBP_LO | mean_sbp > SBP_HI) | (mean_sbp < lo_iqr | mean_sbp > hi_iqr) | (abs(z) > 3
  mean_sbp_clean = ifelse(flag, NA_real_, mean_sbp)
) %>% select(seqn, mean_sbp_clean)

# Build cleaned analytic dataset
anal <- demo_sex %>% left_join(bmi_clean, by = "seqn") %>% left_join(sbp_clean, by = "seqn") %>% filter
```

Plots: Boxplots before vs after

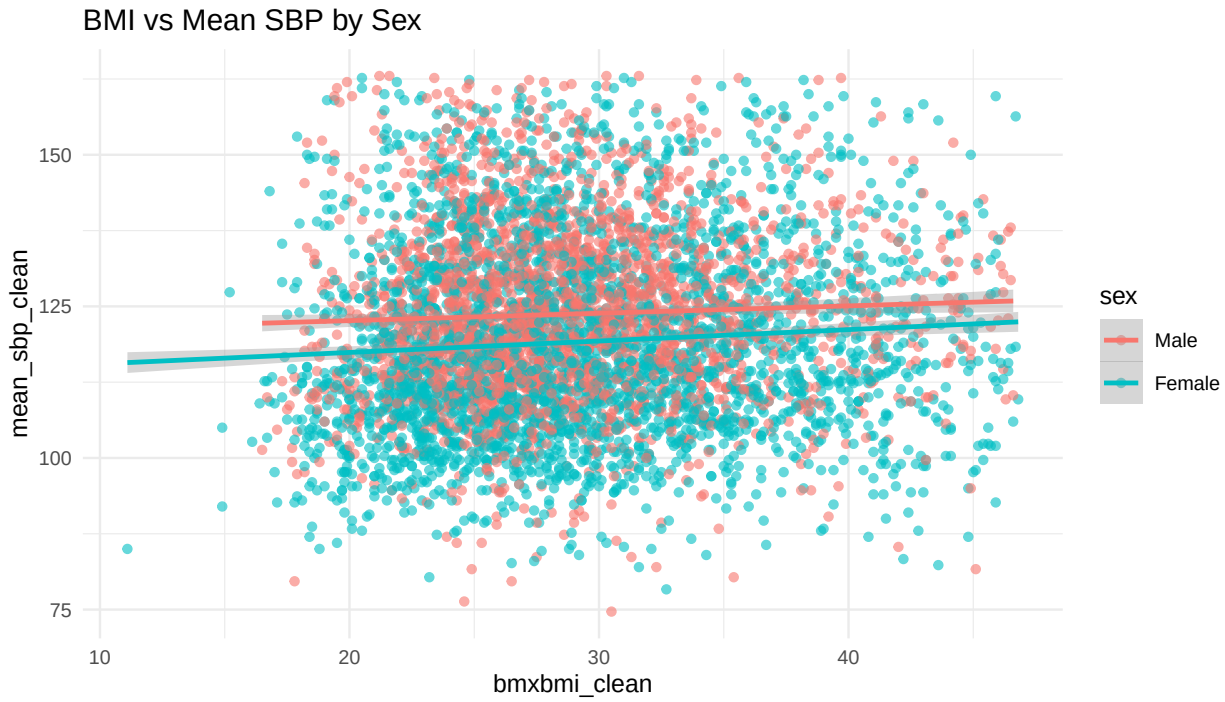
```
# BMI before vs after
p_bmi_before <- ggplot(dat_raw, aes(x = "raw", y = bmi_raw)) + geom_boxplot() + labs(title = "BMI: raw"
p_bmi_after  <- ggplot(anal, aes(x = "clean", y = bmxbmi_clean)) + geom_boxplot() + labs(title = "BMI: c

# SBP before vs after
p_sbp_before <- ggplot(dat_raw, aes(x = "raw", y = mean_sbp)) + geom_boxplot() + labs(title = "Mean SBP
p_sbp_after  <- ggplot(anal, aes(x = "clean", y = mean_sbp_clean)) + geom_boxplot() + labs(title = "Mean

# Arrange
ggpubr::ggarrange(p_bmi_before, p_bmi_after, p_sbp_before, p_sbp_after, ncol = 2, nrow = 2)
```



Scatter: BMI vs SBP by sex and regression

```
anal2 <- anal %>% mutate(sex = factor(sex)) %>% filter(!is.na(bmxbmi_clean) & !is.na(mean_sbp_clean))

p_scatter <- ggplot(anal2, aes(x = bmxbmi_clean, y = mean_sbp_clean, color = sex)) +
```

```
  geom_point(alpha = 0.6) + geom_smooth(method = "lm") + labs(title = "BMI vs Mean SBP by Sex")
print(p_scatter)
```




```
# Stratified models
models <- anal2 %>% group_by(sex) %>% nest() %>% mutate(model = map(data, ~ lm(mean_sbp_clean ~ bmxbmi_
models %>% unnest(tidy)
```

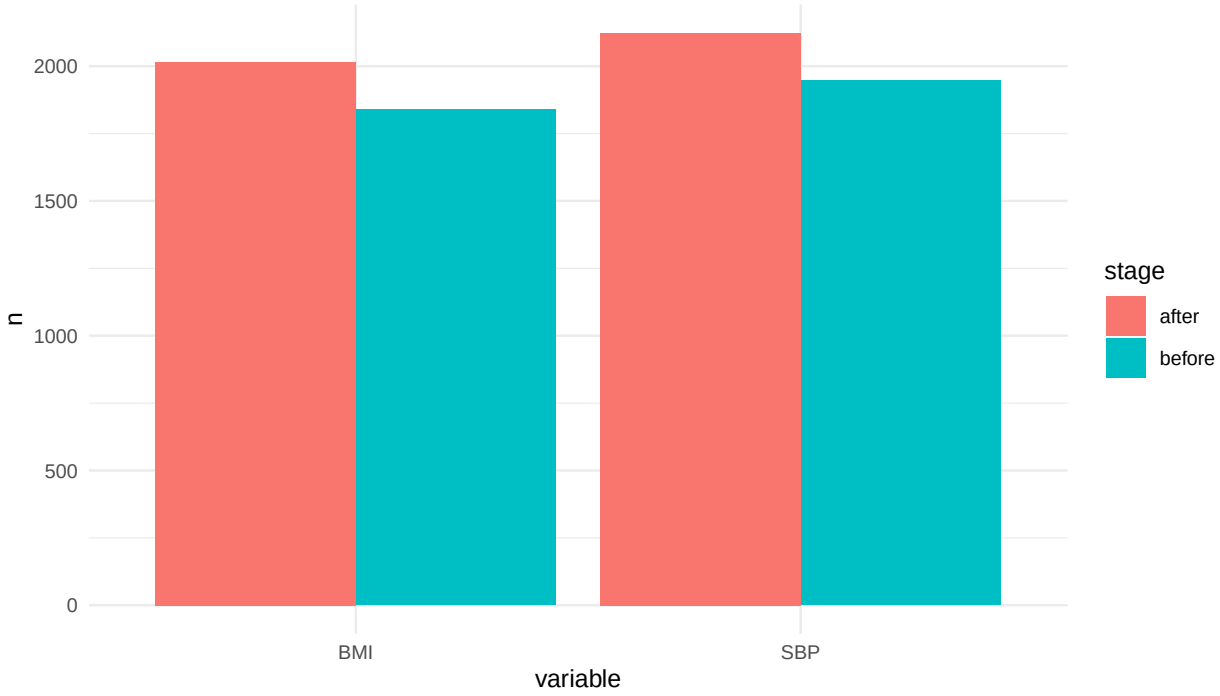
```
## # A tibble: 6 x 8
## # Groups:   sex [2]
##   sex    data                model term  estimate std.error statistic   p.value
##   <fct>  <list>              <lis> <chr>    <dbl>     <dbl>     <dbl>     <dbl>
## 1 Male   <tibble [2,492 x 4]> <lm>  (Int~ 111.        1.66       67.0  0
## 2 Male   <tibble [2,492 x 4]> <lm>  bmxb~   0.0963    0.0495      1.95 5.18e-  2
## 3 Male   <tibble [2,492 x 4]> <lm>  age     0.187     0.0159     11.8  4.55e- 31
## 4 Female <tibble [2,986 x 4]> <lm>  (Int~  90.5       1.40       64.7  0
## 5 Female <tibble [2,986 x 4]> <lm>  bmxb~   0.131     0.0392      3.34 8.42e-  4
## 6 Female <tibble [2,986 x 4]> <lm>  age     0.465     0.0152     30.6  1.59e-179
```

Missingness before/after

```
miss_before <- tibble(variable = c("BMI","SBP"), before = c(sum(is.na(dat_raw$bmi_raw)), sum(is.na(dat_
miss_after  <- tibble(variable = c("BMI","SBP"), after  = c(sum(is.na(anal$bmxbmi_clean)), sum(is.na(ana
miss_tab <- left_join(miss_before, miss_after, by = "variable") %>% mutate(total = nrow(anal), before_p
knitr::kable(miss_tab)
```

variable	before	after	total	before_pct	after_pct
BMI	1839	2016	7809	0.2354975	0.2581637
SBP	1946	2123	7809	0.2491996	0.2718658

```
p_miss <- miss_tab %>% pivot_longer(cols = c(before, after), names_to = "stage", values_to = "n") %>% gg
print(p_miss)
```



Week 6: EDU, Race, and BP trials

Recode EDU and Race; distribution tables

```
# EDU (dmdeduc2)
dat_demo_edu <- demo %>% transmute(seqn, age = ridageyr, dmdeduc2)
edu_tab <- dat_demo_edu %>% mutate(edu = case_when(
  dmdeduc2 == 1 ~ "<9th",
  dmdeduc2 == 2 ~ "9-11th",
  dmdeduc2 == 3 ~ "HS/GED",
  dmdeduc2 == 4 ~ "Some college",
  dmdeduc2 == 5 ~ "College+",
  TRUE ~ NA_character_
)) %>% count(edu) %>% mutate(prop = n/sum(n))
knitr::kable(edu_tab)
```

edu	n	prop
9-11th	666	0.0558116

edu	n	prop
<9th	373	0.0312579
College+	2625	0.2199782
HS/GED	1749	0.1465683
Some college	2370	0.1986089
NA	4150	0.3477751

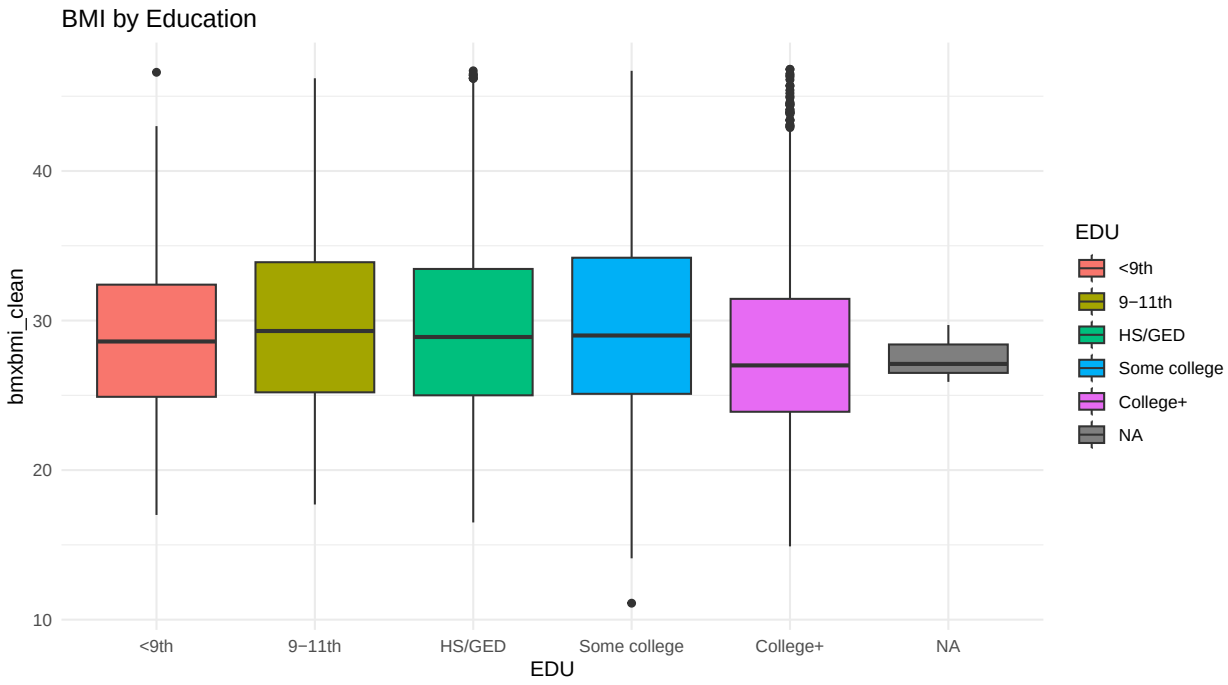
```
write.csv(edu_tab, file = "outputs/EDU_distribution.csv", row.names = FALSE)

# Race (ridreth3)
race_tab <- demo %>% transmute(seqn, ridreth3) %>% mutate(ridreth3 = as.integer(ridreth3)) %>% mutate(ra
knitr::kable(race_tab)
```

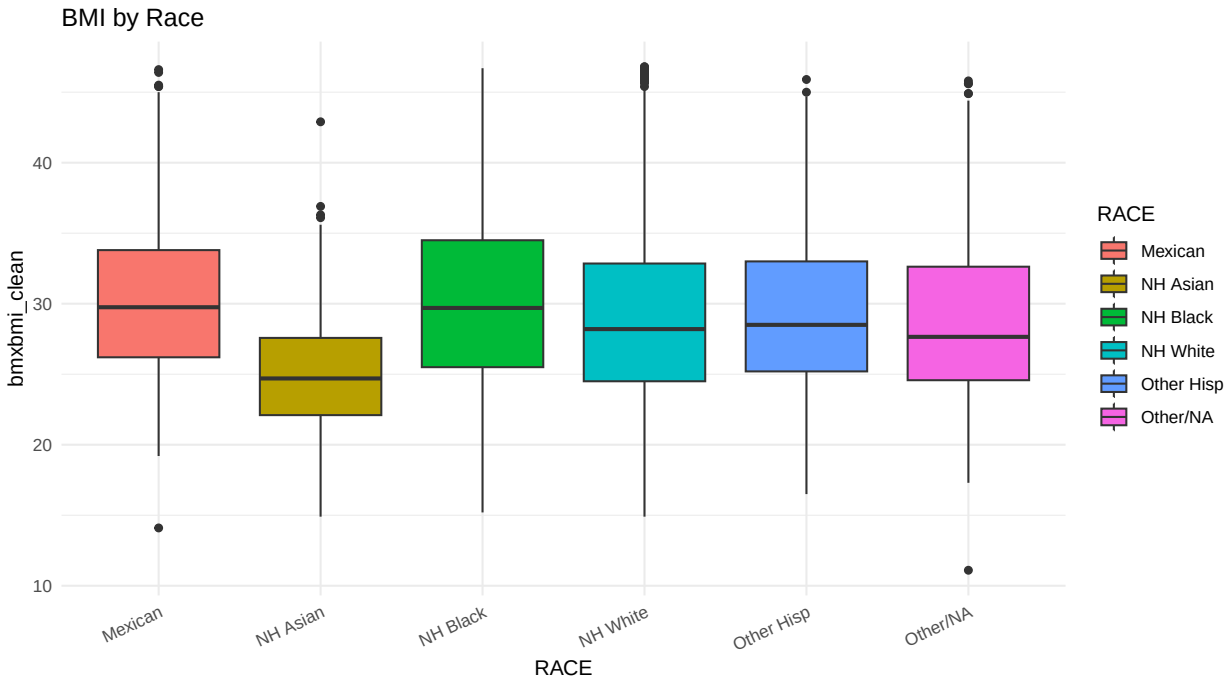
race	n	prop
Mexican	1117	0.0936060
NH Asian	681	0.0570686
NH Black	1597	0.1338306
NH White	6217	0.5209922
Other Hisp	1373	0.1150591
Other/NA	948	0.0794436

```
write.csv(race_tab, file = "outputs/RACE_distribution.csv", row.names = FALSE)
```

BMI distribution by EDU and Race (boxplots)

```
# attach cleaned BMI
bmi_for_plots <- anal %>% select(seqn, bmxbmi_clean) %>% left_join(demo %>% select(seqn, dmdeduc2, ridre
  EDU = case_when(
    dmdeduc2 == 1 ~ "<9th",
    dmdeduc2 == 2 ~ "9-11th",
    dmdeduc2 == 3 ~ "HS/GED",
    dmdeduc2 == 4 ~ "Some college",
    dmdeduc2 == 5 ~ "College+",
    TRUE ~ NA_character_
  ),
  RACE = case_when(ridreth3==1 ~ "Mexican", ridreth3==2 ~ "Other Hisp", ridreth3==3 ~ "NH White", ridre
) %>%
  mutate(EDU = factor(EDU, levels = c("<9th","9-11th","HS/GED","Some college","College+"))) %>%
  drop_na(bmxbmi_clean)

p_bmi_edu  <- ggplot(bmi_for_plots, aes(x = EDU, y = bmxbmi_clean, fill = EDU)) + geom_boxplot() + labs
p_bmi_race <- ggplot(bmi_for_plots, aes(x = RACE, y = bmxbmi_clean, fill = RACE)) + geom_boxplot() + la

p_bmi_edu; p_bmi_race
```

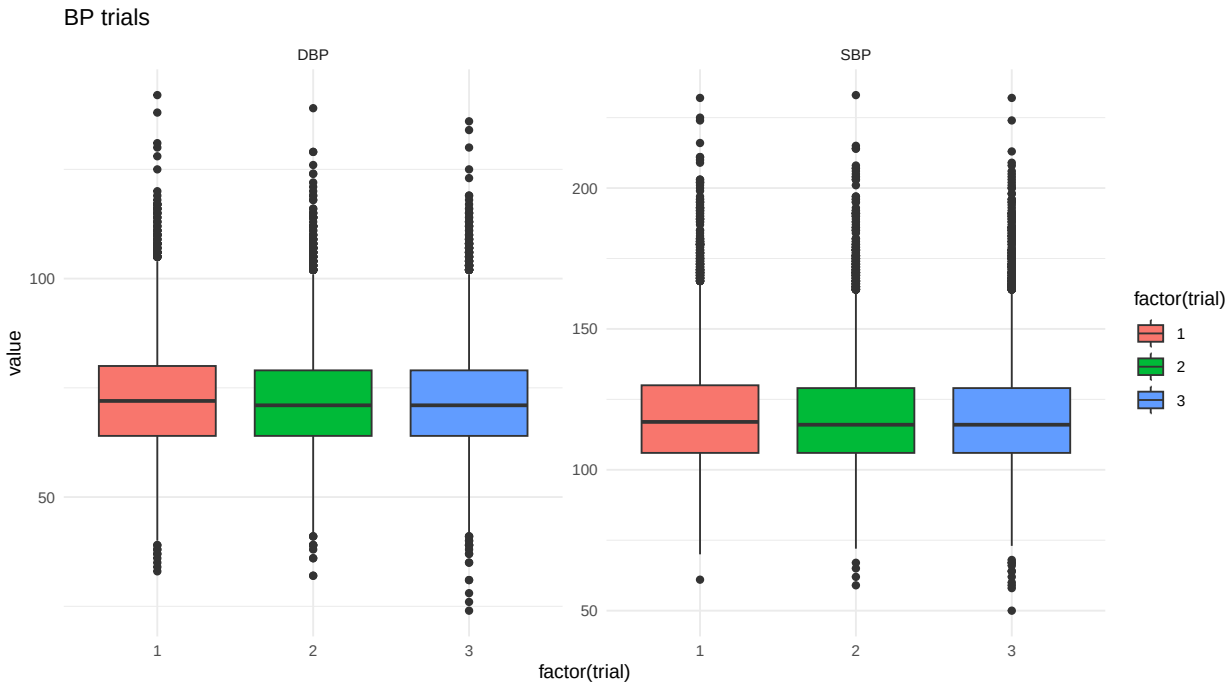




Reshape BP trials (wide → long) and plots

```
sbp_cols <- names(bpx)[stringr::str_detect(names(bpx), "^bpxo?sy[1-3]$")]
dbp_cols <- names(bpx)[stringr::str_detect(names(bpx), "^bpxo?di[1-3]$")]

bpx_long <- bpx %>% select(seqn, any_of(c(sbp_cols, dbp_cols))) %>% pivot_longer(cols = -seqn, names_to
```

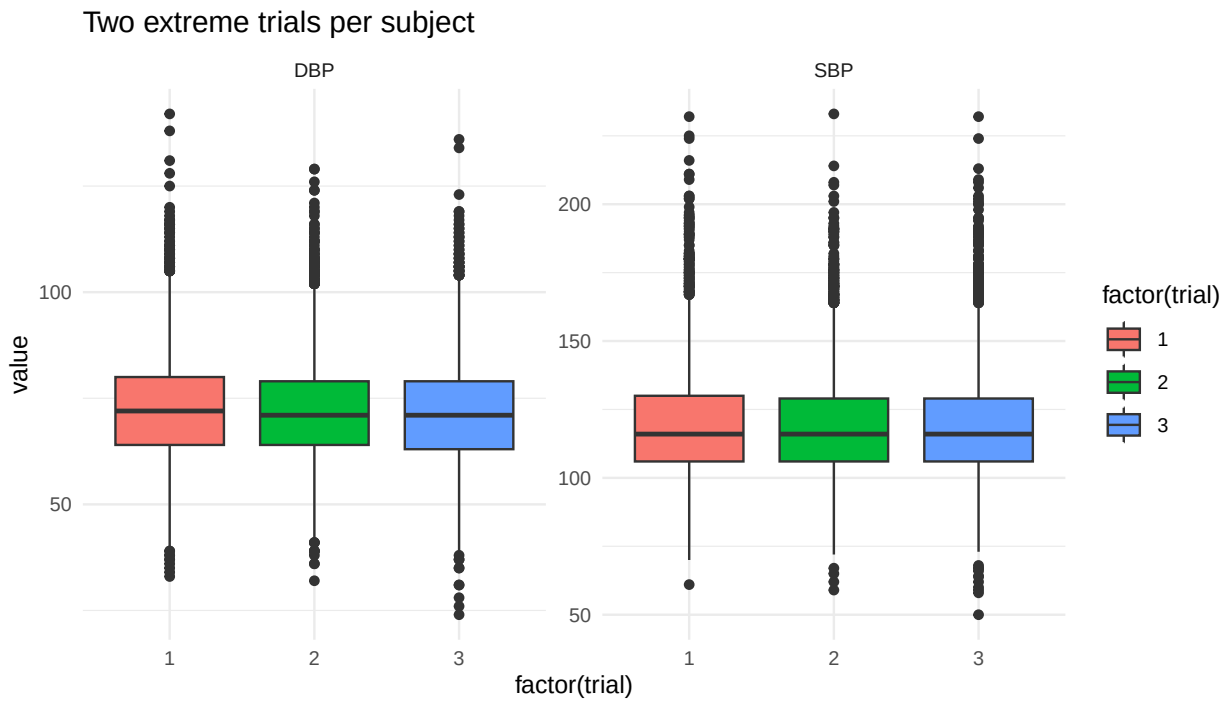
```
p_bp_trials <- ggplot(bpx_long, aes(x = factor(trial), y = value, fill = factor(trial))) + geom_boxplot
print(p_bp_trials)
```




```
# Save
ggsave("outputs/BP_trials_boxplot_rmd.png", p_bp_trials, width = 9, height = 5)
```

Homework extension: select two trials with largest within-subject difference

```
bpx_two <- bpx_long %>% group_by(seqn, measure) %>% filter(n()>=2) %>% mutate(vmin = min(value, na.rm =

p_bp_two <- ggplot(bpx_two, aes(x = factor(trial), y = value, fill = factor(trial))) + geom_boxplot() +
print(p_bp_two)
```




```
ggsave("outputs/BP_two_extreme_rmd.png", p_bp_two, width = 9, height = 5)

# Summary stat: mean absolute within-subject difference per measure
within_diff <- bpx_long %>% group_by(seqn, measure) %>% filter(n()>=2) %>% summarise(ma_diff = max(valu
knitr::kable(within_diff)
```

measure	mean_ma_diff	median_ma_diff	n
DBP	4.999334	4	7506
SBP	7.175460	6	7506

Conclusion

```
## Summary:

## - BMI cleaning removed physiologic and statistical outliers; SBP cleaned similarly.

## - Scatter and regression suggest (describe findings after running script with real data).

## - BMI distributions vary by education and race (see plots).

## - BP trial variability (mean within-subject differences) summarized above; inspect plots to assess t

## Reproducible workflow: this Rmd includes all code to reproduce analyses; change `data_dir` and knit
```