Milestone 4

2025-03-06

Setup

Install packages and data

Installing important packages for the project and downloading the dataset

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```
library(gt)
library(tinytex)
library(naniar)
library(ggthemes)
dhs <- read_csv("~/Documents/GitHub/bios/453/bios453/IAIR7EFL.csv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 724115 Columns: 423
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (1): caseid
## dbl (277): v001, v002, v003, v004, v005, v012, v015, v020, v021, v022, v023,...
## lgl (145): midx_4, midx_5, midx_6, m3a_4, m3a_5, m3a_6, m2n_2, m2n_3, m2n_4,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1) Description of data management utilized and cleaning

As part of the data management process fo this project, I significantly reduced its size to include only variables that I believe are relevant for my research question. This includes all of the relevant variables, in addition to each confounder listed in my concept map (both primary and secondary). Once I had each of these variables selected, In went through and ensured that they were all formatted correctly for the types of regression I wanted to run during my analysis. I also made sure that certain items were filtered out. For example,

some continuous reportings like birth weight in kg clearly contained severe misinputs. Since I am also only concerned with it as a numerical variable, I removed some of the categorical sections on the upper end like not weighed at birth, dont know, or missing. Additionally, I chose to adhere to UNICEF's validity threshold: 250g <= birthweight <=5,500g for a more realistic analysis. I also chose to keep variables like time pregnant and age numeric because that makes sense given what I am interested in exploring. Binary variables like whether or not someone took antiparasitics) were limited to only response values, as I am not interested in those who couldnt answer the question. This leaves us with a hefty 153,582 observations for analysis.

```
dhs_clean <- dhs %>%
  select(m60_1, v481, m18_1, m19_1, m17_1, v457, m15_1, v228, v208, v190a, v190, v149, v131, v024, v463
  rename(
    paradrug = m60_1,
    insured = v481,
    sizechild = m18_1,
    bwkg = m19_1,
    csect = m17_1,
    anemia = v457,
    delivplace = m15_1,
    termpreg = v228,
    bpast5 = v208,
    wind_urbrur = v190a,
    wind = v190,
    educ = v149,
    ethnic = v131,
    state = v024,
    smokes = v463aa,
    married = v501,
    age = v447a,
  ) %>%
  filter(
    paradrug!=8,
    paradrug!=9,
    bwkg <5501,
    bwkg > 249,
    bwkg != 9996,
    bwkg != 9998,
    bwkg != 9999,
    sizechild!=8,
    insured != 9,
    sizechild != 9,
    csect != 9,
    anemia != 9,
    termpreg != 9,
    educ != 9,
    smokes != 9,
    married != 9,
    age != 99,
  ) %>%
  mutate(
    paradrug = as.factor(paradrug),
    insured = as.factor(insured),
    sizechild = factor(sizechild,
                       levels = c("very large" = "1",
                                  "larger than average" = "2",
```

```
                                    "average" = "3",
                                    "smaller than average" = "4",
                                    "very small" = "5")),
    csect = factor(csect),
    anemia = factor(anemia,
                    levels = c("severe" = "1",
                               "moderate" = "2",
                               "mild" = "3",
                               "not anemic" = "4")),
    delivplace = factor(delivplace),
    termpreg = factor(termpreg),
    bpast5 = factor(bpast5),
    wind = factor(wind,
                  levels = c("poorest" = "1",
                             "poorer" = "2",
                             "middle" = "3",
                             "richer" = "4",
                             "richest" = "5")),
    educ = factor(educ,
                  levels = c("none" = "0",
                             "incomplete primary" = "1",
                             "complete primary" = "2",
                             "incomplete secondary" = "3",
                             "complete secondary" = "4",
                             "higher" = "5")),
    ethnic = factor(ethnic,
                    levels = c("caste" = "991",
                               "tribe" = "992",
                               "no caste/tribe" = "993",
                               "dont know" = "998")),
    smokes = factor(smokes)
  )
```

2) Missing data report

According to a missing data summary, there are no missing observations within the cleaned dataset. After removing "dont know" values from majority of the variables of interest (due to their being irrelevent) we no longer have any data that are not available and a final cleaned dataset of 153582 observations.

```
miss_var_summary(dhs_clean) %>%
  arrange(desc(pct_miss))
```

```
## # A tibble: 17 x 3
##    variable    n_miss pct_miss
##    <chr>        <int>    <num>
##  1 paradrug         0        0
##  2 insured          0        0
##  3 sizechild        0        0
##  4 bwkg             0        0
##  5 csect            0        0
##  6 anemia           0        0
##  7 delivplace       0        0
##  8 termpreg         0        0
##  9 bpast5           0        0
```

```
## 10 wind_urbrur      0        0
## 11 wind             0        0
## 12 educ             0        0
## 13 ethnic           0        0
## 14 state            0        0
## 15 smokes           0        0
## 16 married          0        0
## 17 age              0        0
```

3) Concept map

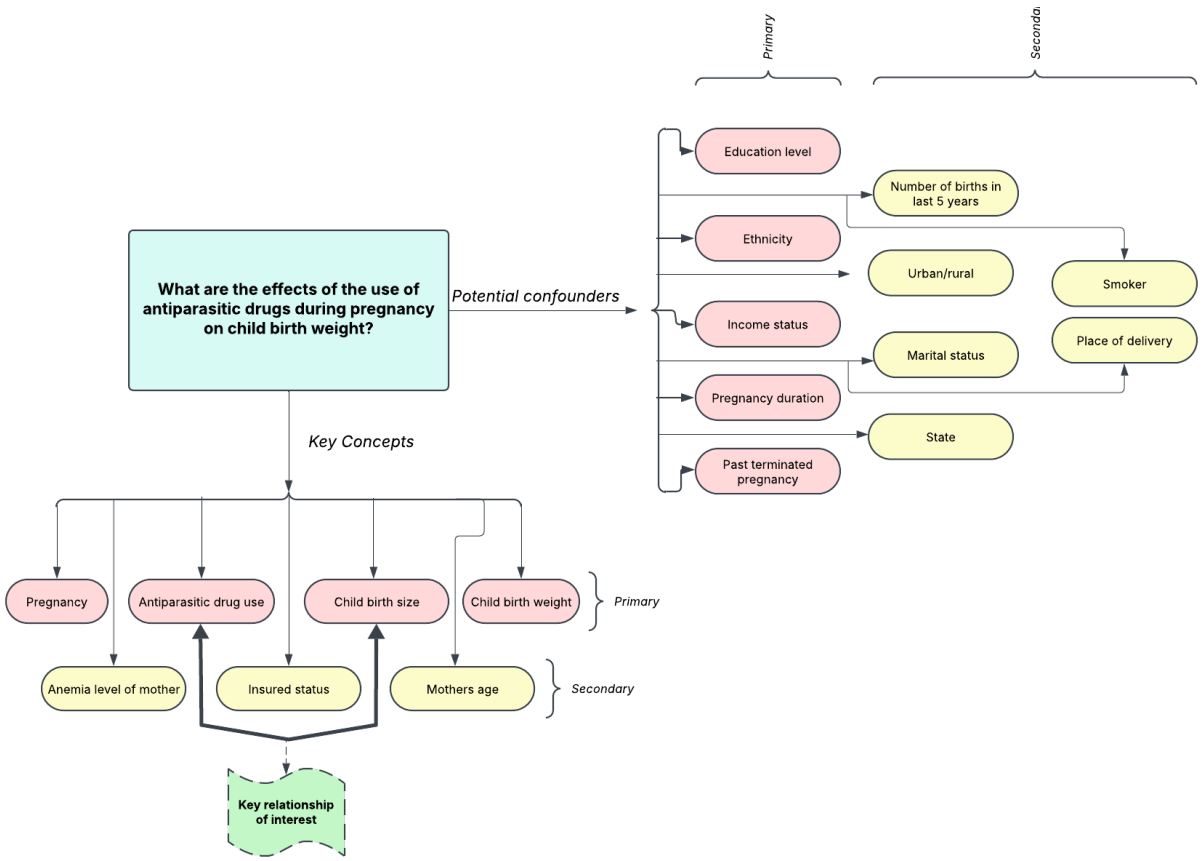


Figure 1: Concept Map

4) Summary table of characteristics

```
library(dplyr)
library(tidyr)
library(gt)

# categorical summary
cat_summary <- dhs_clean %>%
  select(paradrug, insured, educ) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "category") %>%
  group_by(variable) %>%
  mutate(total = n()) %>%
```

```
  group_by(variable, category) %>%
  summarise(
    count = n(),
    percent = paste0(round(100 * count / first(total), 1), "%"),
    .groups = "drop"
  ) %>%
  arrange(variable, category) %>%
  group_by(variable) %>%
  mutate(row_id = row_number()) %>%
  ungroup()

# summary for numerical variables
num_summary <- dhs_clean %>%
  summarise(across(c(age, bwkg),
    list(mean = ~round(mean(., na.rm = TRUE), 2),
         median = ~round(median(., na.rm = TRUE), 2),
         sd = ~round(sd(., na.rm = TRUE), 2)),
    .names = "{.col}_{.fn}")) %>%
  pivot_longer(everything(), names_to = c("variable", "statistic"), names_sep = "_") %>%
  pivot_wider(names_from = statistic, values_from = value) %>%
  mutate(row_id = 1)

# merge and gt table
bind_rows(
  cat_summary,
  num_summary %>% mutate(category = NA, count = NA, percent = NA)
) %>%
  arrange(factor(variable, levels = c("paradrug", "insured", "educ", "age", "bwkg"))) %>%
  gt(groupname_col = "variable") %>%
  tab_header(
    title = "Summary Statistics for DHS Dataset",
  ) %>%
  cols_label(
    category = "Category",
    count = "Count",
    percent = "Percentage",
    mean = "Mean",
    median = "Median",
    sd = "Std Dev"
  ) %>%
  fmt_number(
    columns = c(mean, median, sd),
    decimals = 2
  ) %>%
  cols_align(
    align = "center",
    columns = c(category, count, percent, mean, median, sd)
  ) %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels()
  ) %>%
  tab_style(
```

Summary Statistics for DHS Dataset

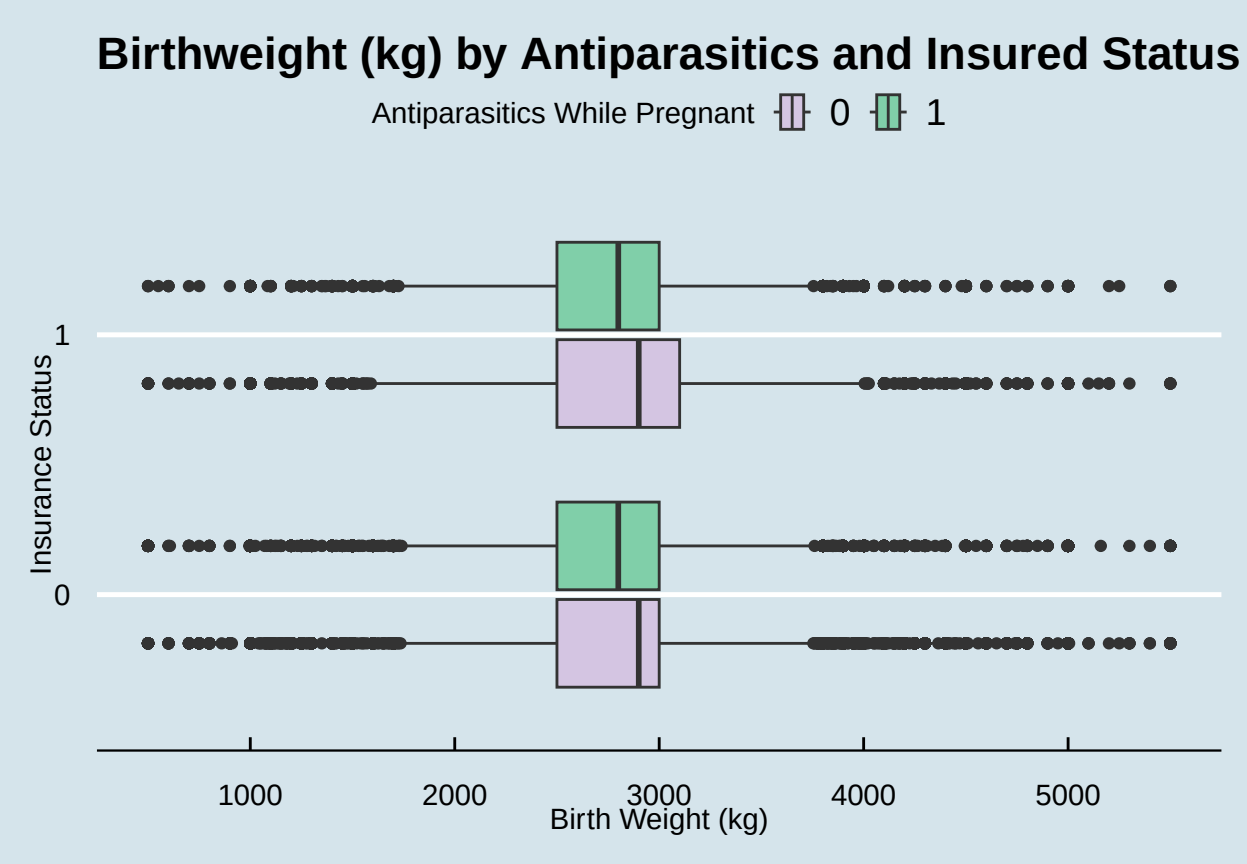
Category	Count	Percentage	Mean	Median	Std Dev
paradrug					
0	104771	68.2%	-	-	-
1	48811	31.8%	-	-	-
insured					
0	109570	71.3%	-	-	-
1	44012	28.7%	-	-	-
educ					
0	28084	18.3%	-	-	-
1	18238	11.9%	-	-	-
3	80116	52.2%	-	-	-
4	3000	2%	-	-	-
5	24144	15.7%	-	-	-
age					
-	-	-	27.38	27.00	5.12
bwkg					
-	-	-	2,816.83	2,900.00	554.07

```
    style = cell_fill(color = "#C5DECD"),
    locations = cells_body(
      rows = variable %in% c("paradrug", "insured", "educ")
    )
  ) %>%
  fmt_missing(columns = everything(), missing_text = "-") %>%
  cols_hide(columns = c(row_id))%>%
  tab_options(table.background.color = "#F1F7ED")
```

```
## Warning: Since gt v0.6.0 `fmt_missing()` is deprecated and will soon be removed.
## i Use `sub_missing()` instead.
## This warning is displayed once every 8 hours.
```

5) Graphical representation of characteristics

```
ggplot(data = dhs_clean, mapping = aes(y = insured, x = bwkg, fill = paradrug)) +
  geom_boxplot() +
  theme_economist() +
  scale_fill_manual(values = c("#D4C5E2", "#80CFA9")) +
  labs(title = "Birthweight (kg) by Antiparasitics and Insured Status",
       y = "Insurance Status",
       x = "Birth Weight (kg)",
       fill = "Antiparasitics While Pregnant")
```



6) Summary and interpretation of characteristics

Based on the characteristics obseved in the table and graphic, a few things stand out. To start, given the fact that the core questions involves antiparasitic drug use, its important to point out that majority of people to not use antiparasitics during pregnancy. Additionally, less than 30% of people have some form of health insurance which may act as a major confounder when trying to predict birth weight. Majority of people surveyed have completed primary school (around 52.2%) but only 2% have some form of higher education. As for age, most of those who participated in the survey were in their late 20s, with a median age of 27 (SD of 5.12). The median birth weight was 2900g, which is less than the international average of roughly 3300g. Based on the graphic, it is evident that on average, children born to mothers who received antiparasitics during pregnancy weighed less at birth than those of mothers who did not. There is also a much larger interquartile range of birthweights for children born to mothers who were insured but did not receive antiparasitics during pregnancy than any other combined category. The Lowest average birth weights were among children born to mothers who were both insured and received antiparasitics during pregnancy. It is possible that due to the large discrepancy between number of insured vs uninsured, this relationship is due to noise, but it is interesting nonetheless. Overall, the spread is relatively consistent across categories.