EXECUTIVE REPORT (PART A): PROFESSIONAL WORK BASED PLACEMENT IN HEALTH DATA SCIENCE

Author: Ijeoma Nwachukwu
Date: 2025-08-26

Project Topic: DHS Data Management and Analysis of Gender Inequality in Reproductive Women across LMICs using IPUMS-DHS Dataset

Project Background

My project focused on collecting, organizing, merging and analyzing datasets from DHS-program relevant to The Biostatistics and Health Data Science (BHDS) group's global health projects, at the University of Aberdeen. BDHS group is a multi-disciplinary academic research and teaching unit under the IAHS characteristic by collaborative research, consultancy and training across clinical, biological and global health domains. In the global health domain where I was assigned to, the data used to conduct the research as well as for training purposes are collected from a number of secure sources, including the The DHS-Program.

The DHS-Program, funded by USAID collects nationally representative global health data, to monitor and evaluate population, health, and nutrition programs, providing data to track approximately 30 SDG indicators. They provide these data for tracking as well as measures to track them, contributing significantly towards achieving the SDG 3 and 5 (The DHS Program, 2025).

However, the DHS-Program has been suspended and is being reviewed for further funding. As a result, new registrations are not being accepted, hence restricting access to datasets commonly used by undergraduate and post graduate students for their theses and training, especially in LMICs, thereby significantly hampering preparations for future national and global health leadership training in addition to other far-reaching effects.

My project aims to address this challenge and also explore aspects of Gender Inequality, including Female Genital Mutilation, Intimate Partner Violence and Autonomy of Health Care Decision Making which are often intertwined and are prevalent issues for women of child bearing age in LMICs(Wessells & Kostelny, 2022).

I would like to thank my placement supervisors, Dr. Aravinda Guntupalli and Dr. Caroline Franco, for their valuable guidance, support and feedback throughout this project.

Project Aim

This project achieved two aims

1. Created a global health data repository of DHS Datasets for 38 years (1984-2022)
2. Pooled Cross Country Exploratory Data Analysis of Gender Inequalities in women of child bearing age.

Methods

The project was conducted in three phases and well documented for transparency and reproducibility of the workflow and analysis results. The data was accessed from DHS-Program website using my supervisor's login. To access datasets, new users must register for an account on the The DHS-Program website.

Phase 1: Auto-download of DHS Datasets

A structured reproducible workflow was scripted using R Markdown which serves as a comprehensive toolkit for accessing, processing, and locally managing DHS downloads, enabling seamless data retrieval for collaborative research in support of global health studies. It ensures secure data access, automates downloads, and systematically unzips, organizes and saves the datasets in hierarchical file structure.`FileName/CountryName/SurveyYear/DataType`. The workflow is specifically for DHS Datasets in SPSS and STATA formats as specified in my project tasks.

Phase 2: DHS IR File Merge (Pilot merge)

A structured, reproducible workflow was developed to merge DHS Individual Recode (IR) datasets for 2 countries (Kenya and Tanzania 2022) using SPSS Syntax. A cross-Country unique identifier `UCASEID` was created by concatenating Country-cluster and case IDs. Subsets containing the `UCASEID` and relevant IPV variables were saved and merged using SPSS commands. This workflow can be adapted for additional countries and survey rounds, and replicated for different variables, provided that the variable names, labels, and meanings are first confirmed to be consistent according to the DHS Recode Manual (The DHS Program, 2025). See syntax of workflow in Appendix 1.

Phase 3: Exploratory Data Analysis and Results using IPUMS-DHS Data

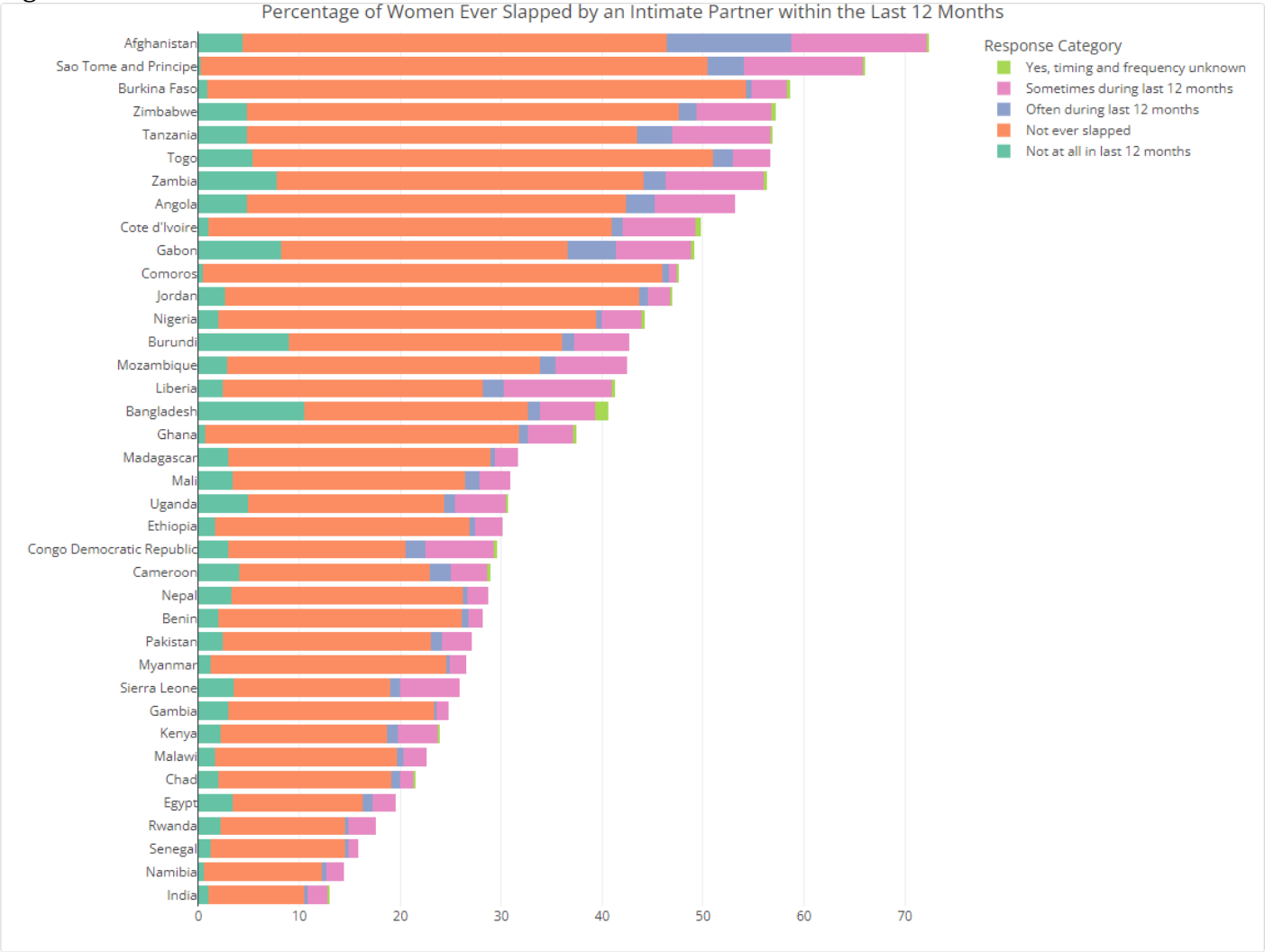
Exploratory data analysis was done using datasets from IPUMS-DHS website which are harmonized DHS survey datasets across countries and over time. Initial Cross-tabulations was conducted for key variables using SPSS (Appendix 2), SPSS output tables were saved as csv files and results imported into R to create visual interactive plots. See codes to reproduce results, plots, R objects and packages in Appendix 5.

The analysis is descriptive and unweighted; they do not account for the size of each country's population or difference over the years. Difference between countries may partly result from sampling variations rather than true prevalence variations. Not all countries have data for all years.

Interpretation is therefore based on descriptive patterns rather than causal claims.

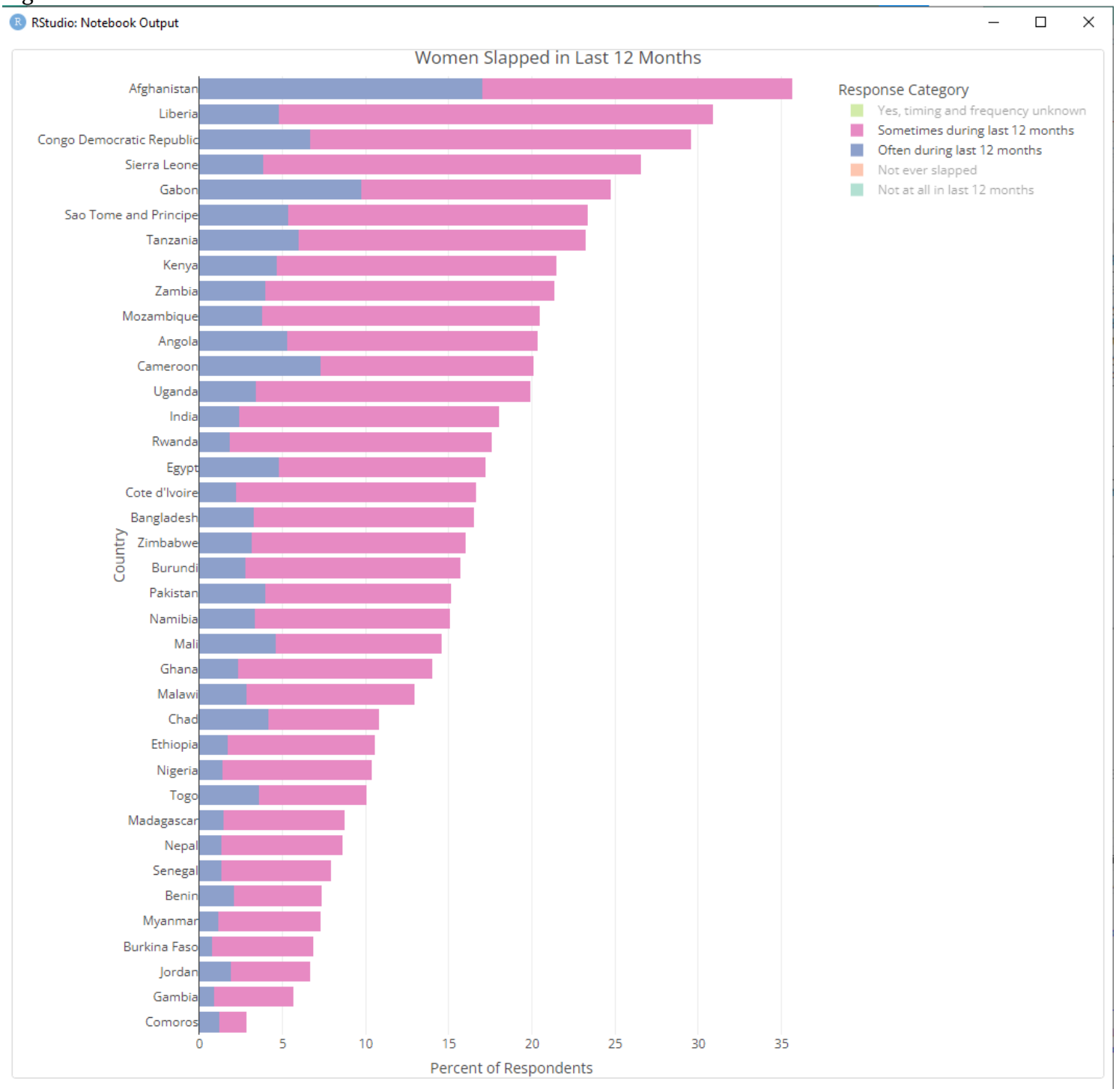
I. IPV: Percentage of Women Slapped in Last 12 month (frequency) by an Intimate Partner, variable code= (DVPSLAPFQ)

Fig.1



This plot presents the distribution of women's reported experiences with intimate partner violence (IPV) across countries. Variability across countries is visible, with some (e.g., Afghanistan, Sao Tome and Principe, Zimbabwe) having higher frequencies of violence, and others (e.g., India, Senegal) showing larger shares of respondents (within country) reporting no experience of IPV.

Fig.2



Notable percentages report being slapped at least sometimes or often within the past year before the survey (over 35%) which reflects a widespread challenge regarding this type of gender-based violence.

II. FGM: Percentage of Ever Circumcised Women within Country, Variable Code= (FCCIRC)

Fig.3

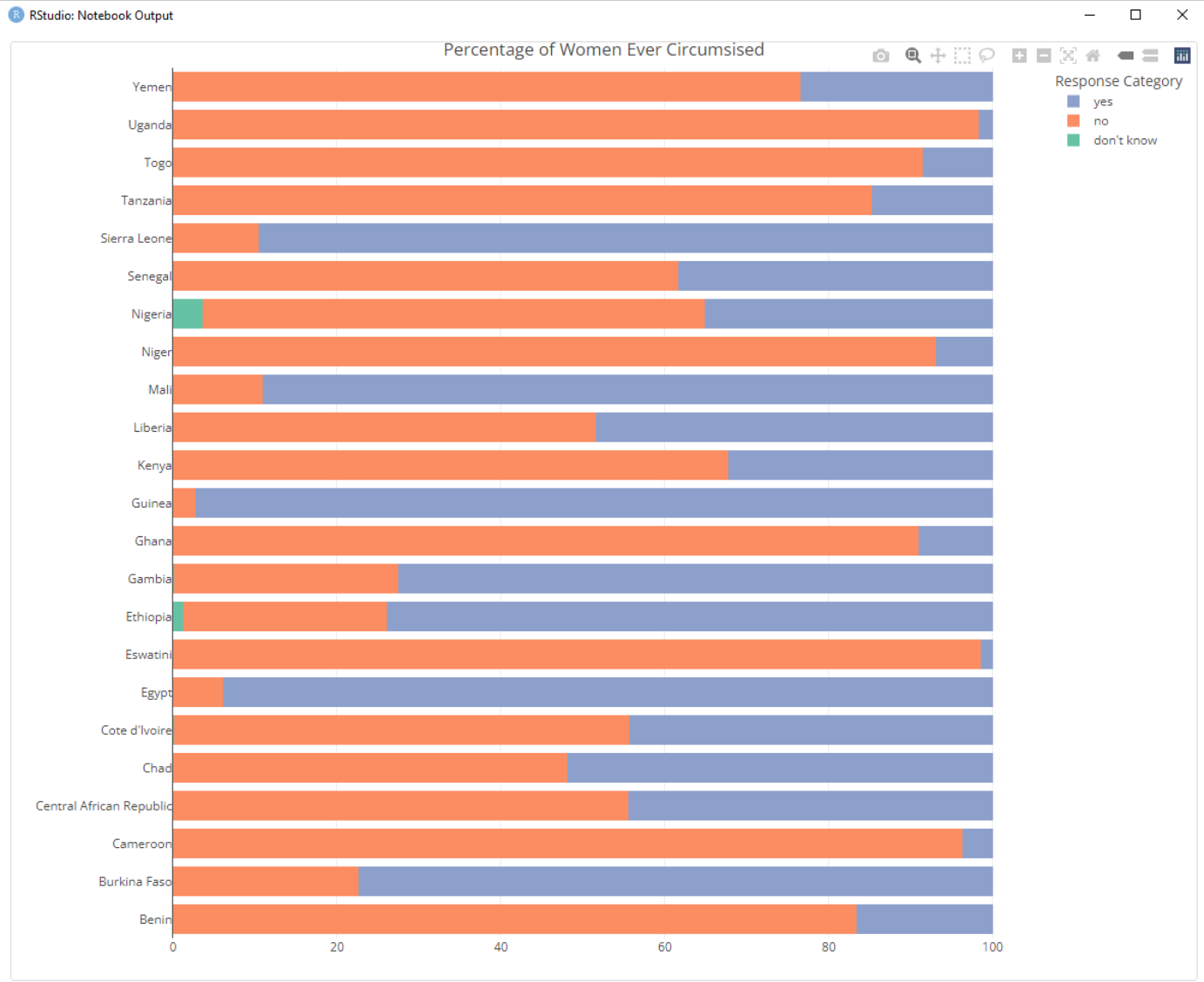
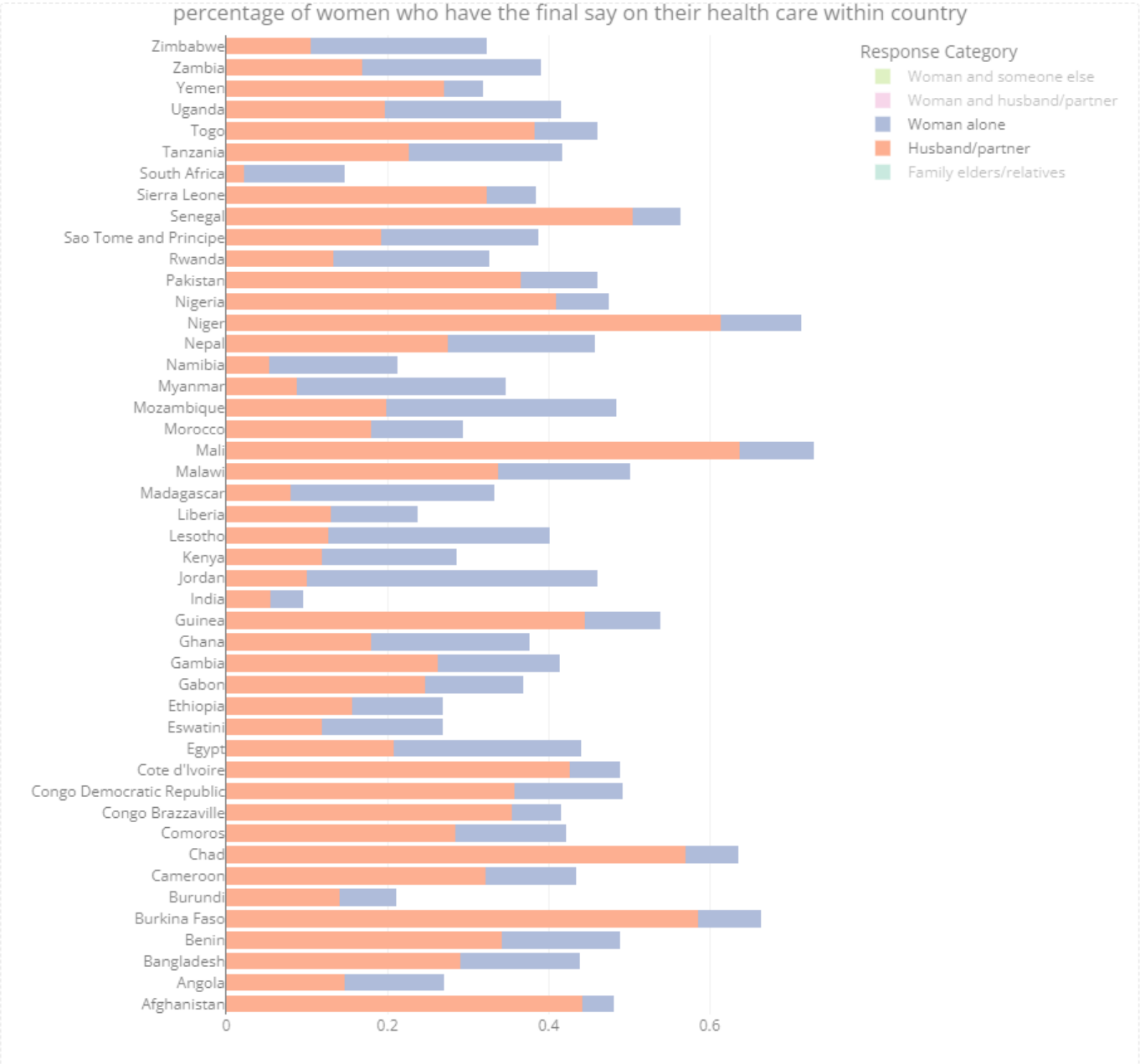


Fig.3 presents the percentage response of women who have experienced female genital mutilation/cutting (FGM/C) within Country, with responses categorized as "yes," "no," and "don't know." There is wide Country variation: nations like Guinea, Sierra Leone, Mali, Gambia, and Egypt show extremely high percentages of women reporting being circumcised (often over 80%) compared to those responding no or don't know, while countries such as Ghana, Cameroon, Tanzania, and others report relatively low response rate. The "don't know" response is almost negligible in most contexts. The Country-to-Country varying responses reflect varying cultural, legal, and historical norms about FGM/C practices.

III. AHCDM: percentage of women who have the final say on their health care within Country, variable code= DECFEMHCARE

Fig.4

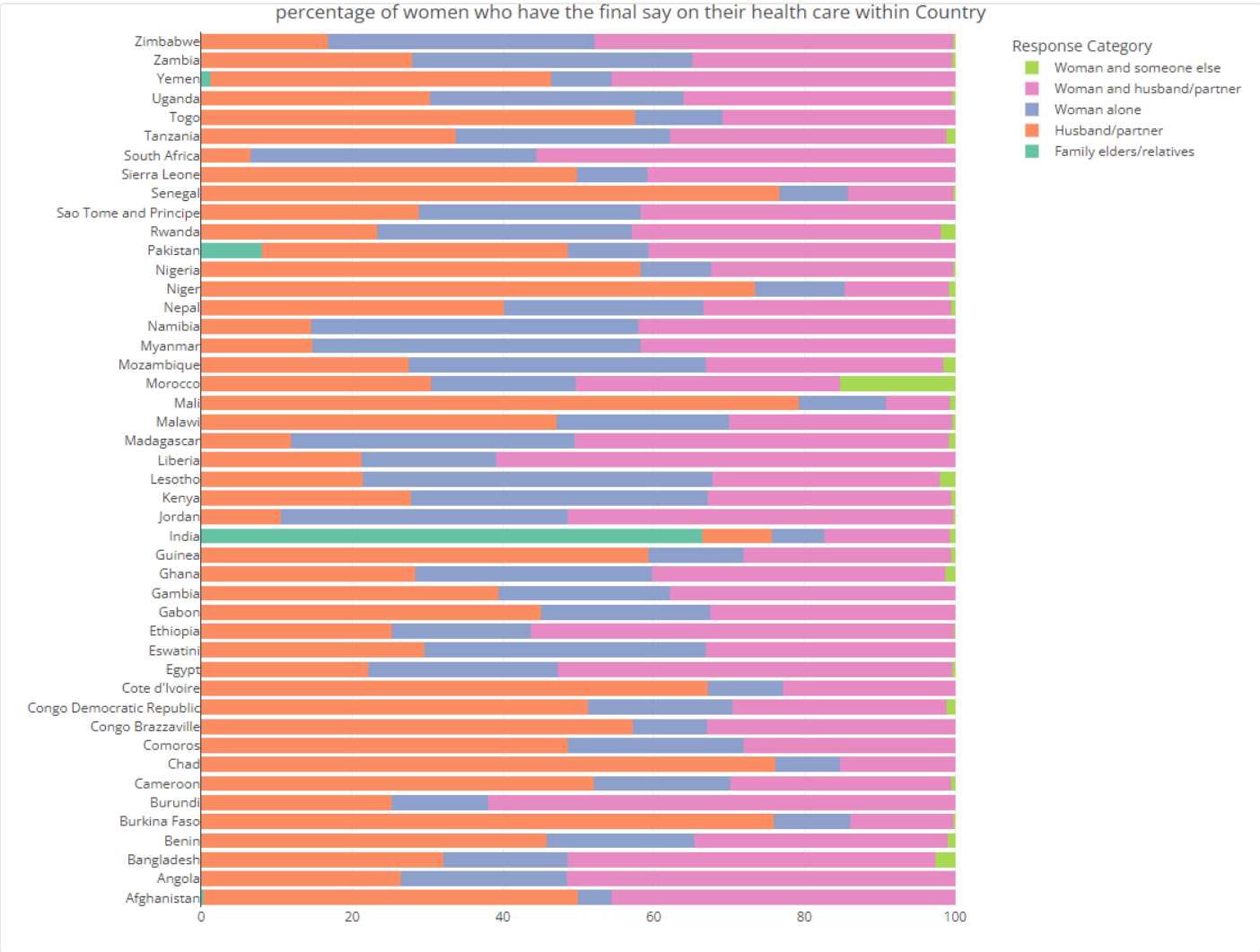


Shows women's reported autonomy and roles in health care decision-making by `Country` and response categories. The plot shows that in most countries, healthcare decisions for women are often made by their "husbands/partners," with relatively fewer women making these decisions entirely on their own.

From Fig.5, the larger proportion of women say decisions are made "with their husband/partner" or by their "husband/partner" alone, reflecting persistent gender norms around health autonomy. However, countries such as Mozambique, Lesotho, and Madagascar display higher shares for "Woman alone," indicating stronger female decision-making autonomy. "Woman and someone else" and "Family

elders/relatives" are minor categories in most contexts, suggesting these are less common arrangements for household health decisions.

Fig.5



In conclusion, the results show a pattern of women facing violence over time from intimate partners, female genital mutilation and often not in control over their own healthcare, with differences depending on the country.

Output

Output files including dataset, analysis and results are saved to One drive folder in the below order

```
DHS-Download Task
├── [DHS_Downloads]
└── [Downloads report, metadata, log]

[Gender Inequalities]
├── [DHS]
│   ├── [dhs-ir-piolt-merge-KE8_TZ8]
│   └── [planning-and-var-map]
└── [IPUMS]
    ├── [ipums-analysis]
    │   ├── [r-project-files-exec-report]
    │   └── [spss-analysis]
    ├── [ipums-data-extracts-comd-files]
    ├── [ipums-ir-dataset]
    └── [ipums-planning-and-var-map]
```

Implications for the Organisation:

- Easy access to DHS Global Datasets.
- Data Preservation pending the resumption of the DHS program and mitigation against probable future program suspension.
- Continuity of training, including future and ongoing training and projects by students and researchers working on global health projects.

References

The DHS Program. (2025).*Sustainable Development Goals.* https://dhsprogram.com/topics/sdgs/index.cfm (Accessed August 28, 2025)

The DHS Program. (2025). *Merging datasets.* https://dhsprogram.com/data/Merging-datasets.cfm (Accessed September 1, 2025)

Wessells, M. G., & Kostelny, K. (2022). The psychosocial impacts of intimate partner violence against women in LMIC contexts: Toward a holistic approach. *International Journal of Environmental Research and Public Health, 19(21), 14488.* https://doi.org/10.3390/ijerph192114488

Appendix 1

SPSS pilot file merge workflow

```
*SPSS
* Encoding: UTF-8.
*SPSS Version 30.0.0.0(172)
*  Encoding: UTF-8.


*Check Recode file to confirm  variable names context match. For this pilot merging, KEIR
8CFL.SAV and TZIR82FL.SAV were conducted in the same year and survey phase (Ist Survey co
nducted in DHS Phase 8, in 2022).
```

```
*KEIR8CFL.SAV however is a continuous DHS Dataset. Create a copy of original dataset as t
hese changes will over-write the original dataset. Unless otherwise specified as in Step
2

*Step1: Create Unique ID using V000 and Case ID variables from both files. to merge from
Dataset 1( KEIR8CFL.SAV )

*Unique ID for Kenya; Dataset 1( KEIR8CFL.SAV ).

DATASET ACTIVATE DataSet1.
STRING  UCASEID (A20).
COMPUTE UCASEID=CONCAT(V000,CASEID).
VARIABLE LABELS  UCASEID 'Unique Case ID'.
EXECUTE.


*Unique ID for Tanzania; Dataset 2( TZIR82FL.SAV ).

DATASET ACTIVATE DataSet2.
STRING  UCASEID (A20).
COMPUTE UCASEID=CONCAT(V000,CASEID).
VARIABLE LABELS  UCASEID 'Unique Case ID'.
EXECUTE.



*Step 2: Select Unique case ID along with IPV variables from both datasets for merging. S
ave them with a different name. Modify file path.

DATASET ACTIVATE DataSet1.
SAVE OUTFILE='C:\Users\Desktop\_KEIR8CFL.SAV'
  /KEEP UCASEID V000 V001 V003 V004 V005 V006 V007 G100 G101 G102 G103 G104 G105 G107 V00
5.

DATASET ACTIVATE DataSet2.
SAVE OUTFILE='C:\Users\Desktop\_TZIR82FL.SAV'
  /KEEP UCASEID V000 V001 V003 V004 V005 V006 V007 G100 G101 G102 G103 G104 G105 G107 V00
5.

*Open _KEIR8CFL.SAV and _TZIR82FL.SAV as Datasets 3 and 4 respectively


*Step 3: Merge all variables.
DATASET ACTIVATE DataSet3.
ADD FILES /FILE=*
  /FILE='DataSet4'.
EXECUTE.

*By default, the active dataset (Dataset3 _KEIR8CFL.SAV) is modified to contain the merge
d cases from the other dataset (Dataset4  _TZIR82FL.SAV).
```

```
SAVE OUTFILE='C:\Users\Desktop\KE8-TZ8-ir-ipv.SAV'
  /COMPRESSED.
```

Appendix 2

```
* Encoding: UTF-8
* SPSS Version 29.0.2.0 (20)
Naming conventions for CROSS TABULATIONS results for further analysis
1. ipv: percentage of women slapped in last 12 month (frequency), variable code= (DVPSLAP
FQ)
2. fgm: percentage of ever circumcised women within Country, variable code= (FCCIRC)
3. ahcdm: percentage of women who have the final say on their health care within Country,
variable code= (FCCIRC)


*Load datset.
 GET
  FILE='C:\Users\Desktop\ipums-ir-dataset.sav'.



DATASET ACTIVATE DataSet1.

CROSSTABS
  /TABLES=Country BY DVPSLAPFQ
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW COLUMN
  /COUNT ROUND CELL.


CROSSTABS
  /TABLES= Country BY FCCIRC
  /FORMAT=AVALUE TABLES
  /CELLS=COLUMN
  /COUNT ROUND CELL.


CROSSTABS
  /TABLES=Country BY DECFEMHCARE
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW COLUMN
  /COUNT ROUND CELL..



*------------------------------------------------------------------.


*For data cleaning in R
 1. Remove first 3 row headings
 2. Rename col1: Country
 3. Filter and remove:
    - All row/col Totals
    - cols:
```

```
            Not in Universe col
            Missing
```

Appendix 3

Abbreviations

- DHS: Demographic Health Surveys

- IPUMS: Integrated Public Use Microdata Series

- USAID: United States Agency for International Development

- LMICs: Low- and Middle-Income Countries

- IPV: Intimate Partner Violence

- FGM/C: Female Genital Mutilation/Cutting

- AHCDM: Autonomy of Healthcare Decision Making

- IAHS: Institute of Applied Health Sciences

All IPUMS-DHS Variable abbreviations available on IPUMS-DHS.

Appendix 4

Data cleaning in R

```r
library(plotly)
library(dplyr)
library(readxl)
library(tidyr)
library(forcats)

# Read data

library(readxl)
ipv_fgm_ahcdm_spss_output <- read_excel("ipv-fgm-ahcdm-spss-output.xlsx",
    sheet = "ipv")

#Clean data
library(readxl)
library(dplyr)

# List of sheet names
sheet_names <- c("ipv", "fgm", "ahcdm")

# Define cleaning function
library(readxl)
library(dplyr)

# List of sheet names
sheet_names <- c("ipv", "fgm", "ahcdm")
```

```r
# Define a cleaning function
clean_spss_output <- function(df) {
  df %>%
    # Remove first 3 rows and last row
    slice(-(1:3), -n()) %>%
    # Remove first column and last column
    select(-1, -ncol(.)) %>%
    # Rename second column to 'Country'
    rename(Country = 1) %>%
    # Convert columns 2 to end to numeric
    mutate(across(2:ncol(.), as.numeric))
}

# Read and clean each sheet
cleaned_data <- lapply(sheet_names, function(sheet) {
  df_raw <- read_excel("ipv-fgm-ahcdm-spss-output.xlsx", sheet = sheet)
  clean_spss_output(df_raw)
})

# Assign cleaned data to named objects for easy access
names(cleaned_data) <- sheet_names
ipv_clean <- cleaned_data$ipv
fgm_clean <- cleaned_data$fgm
ahcdm_clean <- cleaned_data$ahcdm

#Rename categories columns and delete Missing, NIU and Total Values

df_ipv <- ipv_clean %>%
  select(
    1:ncol(.)
  ) %>%
  rename(Country=1,
    `Not ever slapped` = 2,
    `Often during last 12 months` = 3,
    `Sometimes during last 12 months` = 4,
    `Not at all in last 12 months` = 5,
    `Yes, timing and frequency unknown` = 6
  )
# Delete columns 8:9
df_ipv <- df_ipv %>% select(-any_of(c("...8","...9","...10")))



df_fgm <- fgm_clean %>%
  select(
    1:ncol(.)
  ) %>%
  rename(Country=1,
  `no`=2,
  `yes`=3,
  `don't know`=4,
  )
# Delete columns 6:8,10:12
df_fgm <- df_fgm %>% select(-any_of(c("...6","...7","...8")))
```

```r
df_ahcdm <- ahcdm_clean %>%
  select(
    1:ncol(.)
  ) %>%
  rename(Country=1,
`Woman alone`=2,
  `Woman and husband/partner`=3,
  `Woman and someone else`=4,
  `Husband/partner`=5,
  `Family elders/relatives`=8
)

# Delete columns 6:7,9:10
df_ahcdm <- df_ahcdm %>%
  select(-any_of(c("...7","...8","...9","...10","...11")))
```

Appendix 5

IPV percentage by country using R Plotly

```r
library(plotly)
library(forcats)

df_ipv <- df_ipv
response_ipv <- c(
  "Not ever slapped",
  "Often during last 12 months",
  "Sometimes during last 12 months",
  "Not at all in last 12 months",
  "Yes, timing and frequency unknown"
)

# Calculate total (include 'Missing'), then get proportions
df_ipv <- df_ipv %>%
  rowwise() %>%
  mutate(Total = sum(c_across(all_of(response_ipv)))) %>%
  ungroup()

# Calculate percentages
for (col in response_ipv) {
  df_ipv[[paste0(col, " %")]] <- 100 * df_ipv[[col]] / df_ipv$Total
}

# Reshape to long format for plotting
df_ipv_long <- df_ipv %>%
  select(Country, ends_with("%")) %>%
  pivot_longer(
    cols = -Country,
    names_to = "Response",
    values_to = "Percent"
  ) %>%
  mutate(Response = gsub(" %", "", Response))  # Clean up response label
```

```r
#Generate interactive plot using plotly

fig_ipv <- plot_ly(
  df_ipv_long,
  y = ~Country,
  x = ~Percent,
  color = ~Response,
  type = "bar",
  orientation = "h"
) %>%
  layout(
    barmode = "stack",
    title = "Women Slapped in Last 12 Months",
    yaxis = list(title = "Country", categoryorder = "total ascending"),
    xaxis = list(title = "Percent of Respondents"),
    legend = list(title = list(text = "Response Category"))
  )

fig_ipv

#save_plotly_screenshot(fig1_ipv, "fig1_ipv.png")

knitr::include_graphics("fig1_ipv.png")
```

FGM percentage by country using R Plotly

```r
library(plotly)
library(forcats)

df_fgm <- df_fgm

fgm_response_cols <- c(
  "no",
  "yes",
  "don't know"
)

# Calculate total (include 'Missing'), then get proportions
df_fgm <- df_fgm %>%
  rowwise() %>%
  mutate(Total = sum(c_across(all_of(fgm_response_cols)))) %>%
  ungroup()

# Calculate percentages
for (col in fgm_response_cols) {
  df_fgm[[paste0(col, " %")]] <- 100 * df_fgm[[col]] / df_fgm$Total
}


# Reshape to long format for plotting
```

```r
df_fgm_long <- df_fgm %>%
  select(Country, ends_with("%")) %>%
  pivot_longer(
    cols = -Country,
    names_to = "Response",
    values_to = "Percent"
  ) %>%
  mutate(Response = gsub(" %", "", Response))  # Clean up response label

#Generate interactive plot using Plotly

fig_fgm <- plot_ly(
  df_fgm_long,
  y = ~Country,
  x = ~Percent,
  color = ~Response,
  type = "bar",
  orientation = "h"
) %>%
  layout(
    barmode = "stack",
    title = "Percentage of Women Ever Circumsised",
    xaxis = list(title = " "),
    yaxis = list(title = " "),
    legend = list(title = list(text = "Response Category"))
  )


#save_plotly_screenshot(fig_fgm, "fig_fgm.png")

knitr::include_graphics("fig3_fgm.png")
```

AHCDM percentage by country using R Plotly

```r
library(plotly)
library(forcats)

df_ahcdm <- df_ahcdm
ahcdm_response_col <- c(
  "Woman alone",
  "Woman and husband/partner",
  "Woman and someone else",
  "Husband/partner",
  "Family elders/relatives"
)

# Calculate total (include 'Missing'), then get proportions
df_ahcdm <- df_ahcdm %>%
  rowwise() %>%
  mutate(Total = sum(c_across(all_of(ahcdm_response_col)))) %>%
  ungroup()

# Calculate percentages
```

```r
for (col in ahcdm_response_col) {
  df_ahcdm [[paste0(col, " %")]] <- 100 * df_ahcdm[[col]] / df_ahcdm$Total
}



# Reshape to long format for plotting
df_ahcdm_long <- df_ahcdm %>%
  select(Country, ends_with("%")) %>%
  pivot_longer(
    cols = -Country,
    names_to = "Response",
    values_to = "Percent"
  ) %>%
  mutate(Response = gsub(" %", "", Response))  # Clean up response label

#Generate interactive plot using Plotly

fig3_ahcdm <- plot_ly(
  df_ahcdm_long,
  y = ~Country,
  x = ~Percent,
  color = ~Response,
  type = "bar"
) %>%
  layout(
    barmode = "stack",
    title = "percentage of women who have the final say on their health care within Count
ry",
    xaxis = list(title = " "),
    yaxis = list(title = " "),
    legend = list(title = list(text = "Response Category"))
  )

#save_plotly_screenshot(fig_ahcdm, "fig_ahcdm.png")

knitr::include_graphics("fig4_ahcdm.png")
```