

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

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

Project Background

The Biostatistics and Health Data Science Group, is a multi-disciplinary academic research and teaching 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 provides these data for tracking as well as measure to track them, contributing significantly towards achieving the SDG 3 and 5 (The DHS Program, 2025).

However, the DHS-Program has been suspended and currently undergoing review for further funding. During the period of this review, 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 focused on collecting, organizing, merging and analyzing, datasets from DHS-program relevant to our global health projects. While this mitigates the recent suspension of the DHS-program for students and researchers within the team working on global health projects, it also gave me an opportunity to familiarize with global health data and perform exploratory data analysis on 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).

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 carried out in three phases and documentation was ensured for transparency and reproducibility of the workflow and analysis results.

The data was 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 ensure 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 identifiers 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 done for key variables using SPSS, SPSS output tables were cleaned excel with steps documented (see [Appendix 2](#)) and results imported into R to create visual interactive plots.

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

Read data

This plot presents the distribution of women's reported experiences with intimate partner violence (IPV) across countries. The response categories include: "Not ever slapped," "Often

during last 12 months,” “Sometimes during last 12 months,” “Not at all in last 12 months,” and “Yes, timing and frequency unknown.” Most countries show that a significant proportion of women have never been slapped by an intimate partner, but in many settings, notable percentages report being slapped at least sometimes or often within the past year. Variability across countries is visible, with some (e.g., Sao Tome and Principe, Zimbabwe) having higher frequencies of violence, and others (e.g., India, Senegal) showing larger shares of respondents reporting no experience of IPV.

II. FGM: percentage of ever circumcised women within Country, variable code= (FCCIRC)

This plot 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%), 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 significant Country-to-Country differences reflects 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

The chart shows women’s reported autonomy and roles in health care decision-making by Country and response categories response_ahcdm. In many countries, the largest 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.

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 possible future program suspension.
- Continuity of trainings, 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

* 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, KEIR8CFL.SAV and TZIR82FL.SAV were conducted **in** the same year and survey **phase** (Ist Survey conducted **in** DHS Phase 8, **in** 2022).

*KEIR8CFL.SAV however is a continuous DHS Dataset. Create a copy of original dataset as these 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. Save 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 V005.
```

```
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 V005.
```

*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 merged cases from the other dataset (DataSet4 _TZIR82FL.SAV).

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

Appendix 2

* Encoding: UTF-8

*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= (DVPSLAPFQ)

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 dataset.

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 excel

1. remove:

First 3 row headings

2. Name col1: Country

3. Populate Country column

4. Filter and remove:

- All row/col Totals

- cols:

Not in Universe col

Missing

```
-rows: in count/% col
      Blank
      All rows except % within Country (for fgm and ahcdm)
5.Number format is Percentage
```

Appendix 3

Abbreviations - DHS - IPUMS - USAID - LMICs - All IPUMS Variable abbreviations available on [IPUMS-DHS](#).