Title: Challenges and Solutions to Combining Data from Leading Global Health Surveys: An IPUMS Perspective

Authors: Anna Bolgrien, Mirian King, Devon Kristiansen

Abstract (150 words)
IPUMS Global Health freely provides integration and documentation for three leading global health surveys: Demographic and Health Surveys (DHS), UNICEF Multiple Indicator Cluster Surveys (MICS), and Performance Monitoring for Action (PMA), with nationally representative surveys in over 110 countries (42 in Africa). IPUMS Global Health has increased comparative scholarship by helping researchers analyze multiple samples *within* each survey collection.

MICS, DHS, and PMA surveys often cover the same topics and employ similar questions and sampling. Pooling data *across* these IPUMS collections could extend analyses' geographic and temporal scope, but surveys' differences make such pooling labor-intensive and error-prone.

In this paper, we identify the main barriers to combining data *across* IPUMS Global Health data collections and describe IPUMS' ongoing work to increase the three surveys' interoperability. We illustrate the gains from interoperability by presenting results for three Sustainable Development Goal (SDG) indicators, using IPUMS DHS, MICS, and PMA data from Africa.

Extended Abstract

IPUMS is the leading source of census and survey data integrated across time and place (Kugler & Fitch, 2018). The data and documentation are accessible through interactive websites that allow registered users to create customized data and syntax extracts free of charge, significantly simplifying comparative research. IPUMS Global Health provides integrated data and documentation for three global health surveys: Demographic and Health Surveys (DHS), UNICEF Multiple Indicator Cluster Surveys (MICS), and Performance Monitoring for Action (PMA), encompassing nationally representative surveys from over 110 countries (42 in Africa). The IPUMS versions of the DHS, MICS, and PMA data have increased scholarship on a variety of topics on women and children's health, by helping researchers analyze multiple samples *within* each survey collection.

To further facilitate comparative global health research, IPUMS Global Health is now exploring ways in which researchers can leverage commonalities *across* the three data collections. In this paper, we first outline the scope of possible interoperability among the three IPUMS Global Health data collections. Second, we describe three approaches to achieving cross-survey-type interoperability - via new Global Health variables, new summary documentation, and efforts by individual researchers (with tips on issues to keep in mind). Finally, we present results for three Sustainable Development Goal (SDG) indicators, created using data from IPUMS Global Health surveys from Africa.

Possibilities and challenges

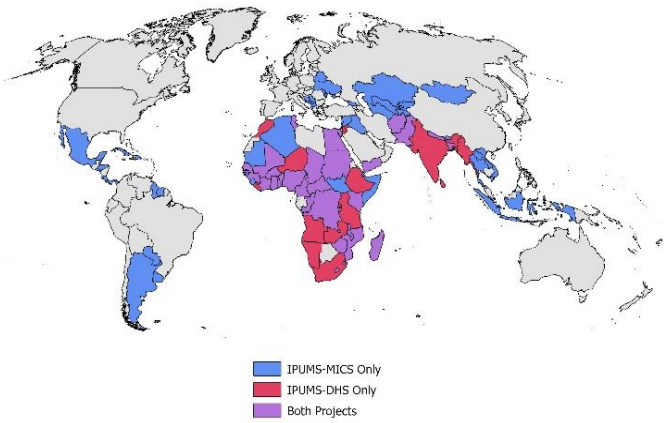
While there are many unique features of each data collection, the demographic microdata collected by DHS and MICS have been used to track progress towards a variety of indicators since the 1980s (for DHS) and early 2000s (for MICS). These surveys typically cover common topics, including household characteristics, marriage and fertility, education, maternal and newborn child health, economic characteristics and work, disability, and migration.

Before IPUMS Global Health, the expansive scope and richness of the DHS and MICS data proved to be a barrier to studying trends across time and countries. For example, changes in question wording or data

coding would require researchers to refer to lengthy documentation for each individual sample. Each IPUMS data project has removed this burden for a given survey type, by using consistent variable names and codes across samples, displaying variable availability, and documenting variable-specific question wording, universes, and comparability issues. Harmonized microdata can be downloaded to create a customized data file including the researchers' chosen countries, sample years, and variables. These innovations save researchers hours of data exploration, language translation, data cleaning and recoding, and file manipulation. At this writing, extensive harmonized health survey data are freely available from IPUMS: for 45 countries, 180 samples, and over 15,000 variables for IPUMS DHS; for 89 countries, 211 samples, and over 1000 variables for IPUMS MICS[1], and for 11 countries, over 200 samples, and over 9000 variables for IPUMS PMA.

Data from IPUMS Global Health are, however, only integrated within a given survey type (DHS, MICS, or PMA). While IPUMS staff sometimes collaborate, each of these harmonized data collections has a separate starting point, staff, source of funding, and partner supplying source data. Moreover, each partner who supplied their original data is committed to retaining the distinct identity of their survey, which might have been compromised by using standard variable names and losing response detail through imposing common codes. Thus, for example, while DHS, MICS, and PMA all collect information from women of childbearing age about their fertility and their knowledge and use of family planning, the names of relevant variables and their coding schemes are unique to each survey data collection. Combining data across survey types could greatly expand the temporal and geographic scope of analyses, as shown in Figure 1, which displays countries included in IPUMS MICS (blue), IPUMS DHS (red) and in both IPUMS DHS and IPUMS MICS (purple)[2].

Figure 1. Geographic scope of IPUMS MICS (blue), IPUMS DHS (red), or both (purple)



Three types of solutions to cross-data project challenges

At IPUMS Global Health, we are first working to increase interoperability by creating some new "Global Health" (_GH) harmonized variables that share names and codes across the three survey types. The original data collections retain all the detail of variable codes, while using composite coding to maximize comparability across samples. The new interoperable Global Health variables follow a different mandate: to **identify major categories** and impose consistent codes across IPUMS DHS, MICS, and PMA. For example, the TOILETTYE variable in IPUMS DHS uses four-digit composite coding to retain all detail across DHS samples, but the Global Health variable on type of household toilet facility would use only 1

[1] Please note that IPUMS MICS does not disseminate data directly and instead facilitates harmonization by providing syntax for registered UNICEF MICS users to create harmonized data.

[2] IPUMS PMA omitted for brevity.

or 2 digits, fitting responses into broad categories of improved versus unimproved sanitation facilities. We expect to release at least 60 consistently named and coded Global Health variables in each IPUMS Global Health dataset by summer 2024.

The second approach to increasing interoperability across data collections is researching and posting detailed documentation, such as User Notes or journal articles, about overarching challenges to comparability and how to address those challenges. For example, for a 2023 ISI conference paper and forthcoming journal article, we found common ground for studying women's experience of intimate partner violence across IPUMS DHS, MICS, and PMA (Kristiansen et al. 2022). We will be looking into comparability issues for child vaccination variables this summer, and we plan to document tricky differences in collecting data on children in DHS versus MICS. For example, DHS collects information on child health issues from biological mothers, while MICS takes reports from guardians who did not birth the child, and the two surveys use a different pre-survey time reference period to identify children of interest.

While shared Global Health (_GH) variables and broad guidance on interoperability challenges and solutions will help the research community, individual researchers must take on much of the work of imposing consistency for their variables of interest (the third approach). We suggest seven topics that researchers should keep in mind when planning and carrying out research across IPUMS Global Health data collections:

1) Consider the availability (countries, years, collection) for samples needed for your research question.
2) Identify and select the appropriate unit of analysis.
3) Consider unique design features of each data collection (e.g., Service Delivery Points and longitudinal data in PMA, reproductive calendars and GPS sample points in DHS).
4) Identify relevant variable names and codes within each data collection.
5) Review online documentation on comparability issues (including question wording).
6) Review universes and modify your dataset to impose consistency within and across data collections.
7) Identify and conduct additional data formatting to facilitate pooling across data collections

Some of these steps are also useful for researchers selecting material for analysis *within* an IPUMS Global Health data collection, since IPUMS documents but does not resolve issues such as different variable universes (who was asked a question) and sampling frames (e.g., all women versus ever-married women only). IPUMS Global Health systematically displays the information on these issues for each data collection to simplify the task of cross-collection harmonization.

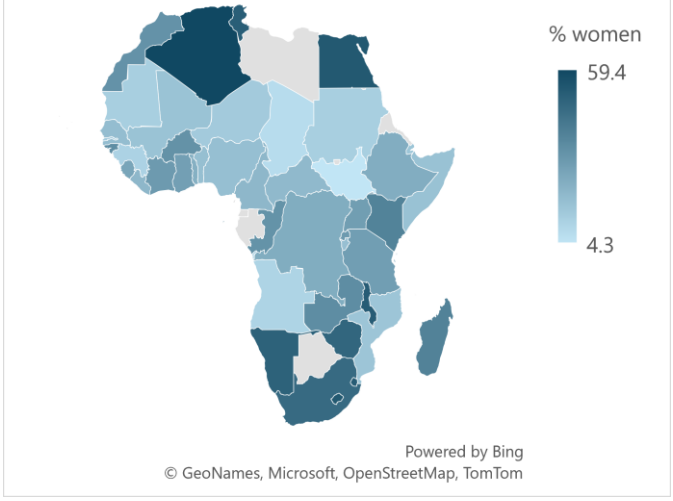
Three examples of interoperability across IPUMS Global Health data collections

The geographic and temporal scope of results obtainable by combining data across survey types is impressive and sometimes worth the extra work of harmonizing across IPUMS Global Health data collections. We support this claim by calculating three SDG indicators: 1) Percent of women who own a mobile phone (in DHS and MICS); 2) Percent of women using family planning at time of survey (in DHS, MICS, and PMA); 3) Percent of households with access to water from an improved source (in DHS, MICS, and PMA). Space limitations preclude our presenting and discussing all three results in this abstract in depth. However, we identify three areas in which interoperability across IPUMS Global Health projects can provide new perspectives in research.

First, comparing data in different projects within a country in similar years can help validate results of data collection efforts. For example, mobile phone ownership in Zimbabwe in 2015 from DHS and in

2019 from MICS was calculated to be 72% and 72.1%, respectively [Figure not shown due to space limitations].

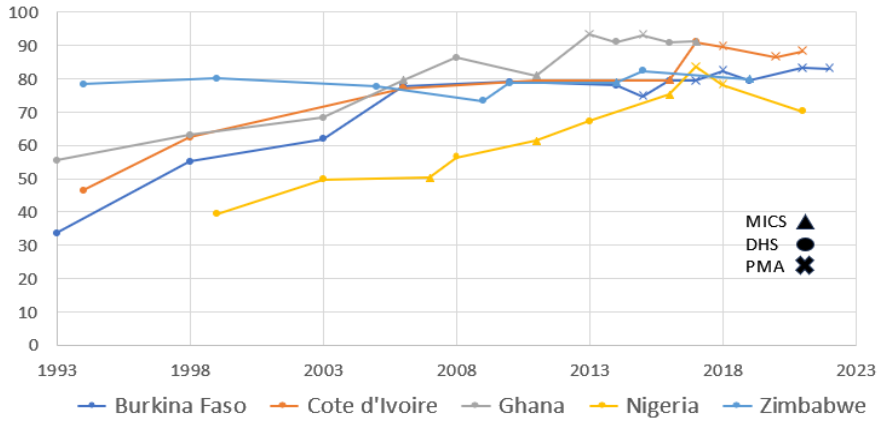
Figure 2: Percent of women using family planning at time of survey (in DHS, MICS, and PMA)



Second, different data collections partner and collect data in different countries. By pooling results across data collections, results can reflect a greater geographic scope. The map in Figure 2 (left) shows the percentage of women who report using a family planning method at the time of the survey. We identified the most recent sample from each country (range between 2003-2022) and this figure represents 19 samples from IPUMS DHS, 23 samples from IPUMS MICS, and 3 samples from IPUMS PMA.

Finally, surveys within countries occur every 5-10 years since the 1980s for DHS and early 2000s for MICS. Starting in 2013, PMA conducts annual data collection in select countries. This means that within any one data collection there may only be a few surveys over time. The usefulness of IPUMS Global Health interoperable variables allows for longer and more comprehensive data series to be constructed. For example, Figure 3 (below) shows the percentage of households with access to improved drinking water for five countries. Each data point is identified as coming from DHS, MICS, or PMA.

Figure 3: Percent of households with access to water from an improved source (in DHS, MICS, and PMA)



While these examples demonstrate the power of interoperable microdata from IPUMS DHS, MICS, and PMA for tracking national-level SDGs, the possibilities we will generate with cross-project interoperability extend beyond indicator calculation and tracking. Global Health interoperable variables will also bolster analyses in global health research such as regression analyses by enabling the study of small subpopulations frequently hidden in aggregate data or small sample sizes. Overall, we anticipate interoperability across the three data collections to generate novel research studies which might otherwise be too complex to undertake.

Citations:

Bolgrien, Anna, Elizabeth Heger Boyle, Matthew Sobek, and Miriam King. IPUMS MICS Data Harmonization Code. Version 1.1 [Stata syntax]. IPUMS: Minneapolis, MN., 2024. https://doi.org/10.18128/D082.V1.1

Boyle, Elizabeth Heger, Miriam King, and Matthew Sobek. IPUMS-Demographic and Health Surveys: Version 9 [dataset]. IPUMS and ICF, 2022. https://doi.org/10.18128/D080.V9

Kristiansen, Devon, Elizabeth Heger Boyle, Kathryn L. Grace, and Matthew Sobek. IPUMS PMA: Version 8.0 [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D081.V8.0

Kristiansen, Devon, Maya Luetke, Matt Gunther, Miriam King, Anna Bolgrien, and Mehr Munir. "A Review of Domestic Violence Data in Free, Harmonized International Survey Data from IPUMS." Paper presented at the International Statistical Institute (ISI) World Statistics Congress, 2023

Kugler, Tracy A. & Fitch Catherine A. Interoperable and accessible census and survey data from IPUMS. Sci. Data 5:180007 doi: 10.1038/sdata.2018.7 (2018).