**METHODS**

*Data sources*

The data used for this visualization come from 2 household surveys conducted in Malawi in 2015/2016:

* The Population-Based HIV Impact Assessment (PHIA) is described in detail at the following link: <https://phia.icap.columbia.edu/countries/malawi/>
* The Demographic and Health Survey (DHS) is described in detail at the following link: <https://dhsprogram.com/publications/publication-FR319-DHS-Final-Reports.cfm>

Both surveys are publicly available, but access to survey datasets requires user registration. The calculations required for data visualization are performed using the “Household member recode” of the DHS dataset. For the PHIA, these calculations require combining the household, children and adult datasets. An R script detailing these calculations is included in this submission.

*Data analysis*

In both the PHIA and DHS datasets, I first grouped the female *de facto* population into 5-year age groups, and I calculated the proportion of the population in each of these groups. Then, after combining the PHIA and DHS age distribution into a single data frame, I calculated the ratio of these two proportion for each age group. In doing so, I accounted for survey design, using weights provided by each survey program. A ratio below indicates that the proportion of the population in a specific age group is higher in DHS than in PHIA, whereas a ratio above 1 indicates that the proportion of the population in a specific age group is higher in PHIA than in DHS.

I calculated confidence intervals around these ratios using standard formulae for the variance of a ratio of two quantities (e.g., see equation 20 in <https://www.stat.cmu.edu/~hseltman/files/ratio.pdf>). In calculating confidence intervals, I assumed that the PHIA and DHS samples were drawn independently.

*Hypothesis*

I formulated the following hypothesis: if interviewers manipulate reported ages, then computed age ratios should be in favor of DHS (i.e., <1) at ages where individuals are eligible for additional data collection only in PHIA (e.g., 5-14 years, 50-64 years). As a corollary, they should be in favor of PHIA (>1) in other age groups where individuals are eligible for both surveys, or are not eligible for any of the two surveys.

*Data visualization*

I plotted the ratios and their confidence intervals along a vertical axis starting with the youngest age group (0-4 years old) and ending with the 75-79 years old age group. I omitted individuals aged 80+ years old from the visualization because a) there are few individuals at such ages in each dataset, and b) the true duration in years of this open interval is unknown. I filled various areas of the plot according to eligibility for additional data collection in each survey. On the x-axis, I displayed age ratios describing the relative age composition of the two surveys. The x-axis uses a logarithmic scale.

*Limitations*

My analysis has several limitations. First, the visualization documents the results of a quasi-experiment. It assumes that distortions in the relative age composition of household populations observed between the two surveys are primarily attributable to differences in age-related eligibility thresholds. There are other aspects of survey implementation that might have differed between the two surveys. However, it is unlikely that these characteristics were also related to the age of household members. Second, I did not investigate whether all interviewers engage in age manipulation, or whether only some of the interviewers engage in this practice. The extent of age manipulation might also vary across data collection teams, depending on the implementation of verification protocols by study supervisors. Third, I did not investigate age manipulation among the male population, due to more complex sampling designs including different probabilities of selection in various sub-samples of the DHS survey. Finally, the main outcome of this quasi-experiment is an assessment of the relative age distribution of household residents and visitors. I did not have access to a gold standard measurement of age for each reported household member (e.g., a birth certificate).