

Hello,

I am engaged in research examining HIV risk behaviors among men/women who have HIV in sub-Saharan African countries. I have downloaded the necessary datasets and merged them according to processes I outlined in an earlier post:

https://userforum.dhsprogram.com/index.php?t=tree&goto=16859&#msg_16859

Now, I am interested in restricting my sample further, based on the number who are reported to be HIV+ per country/year.

I am not sure what is the best approach to further restrict the data, since there is some variation in the published literature using DHS. For example, some use all available data, while for others, it depends on the number of HIV+s in the country.

Here are 4 options:

1. Use only the most recent survey (regardless of the #men and women)
2. Use all surveys for which data are available , regardless of number of men/women: (for example, see Table 1 here: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5439027/>)
3. Use most recent survey, with at least 50 men/women (e.g., see Data Section here: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5638499/>)
4. Use the most recent survey, with at least 50 men/women who are aware of their status (see Data/methods section here: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4344381/>)

I was going to with option 2 (use all available data), but given the variation in stigma and other questions across surveys, I was thinking of using only recent surveys (option 1).

However, I am attracted to options 3 and 4 too. The only challenge is that I am not sure what the justification is for using 50 individuals.

I would appreciate some clarification on the pros and cons of each approach, and whether there is any established benchmarks to guide selection of sample sizes.

Thanks _ Yawo

The problem with #3 and 4 is that I find no justification for selecting 50 men/women.
