Frequently Asked Questions (FAQs)

1. How do I work with Demographic and Health Survey (DHS) data in the PPA Wizard?

Phase 2 contains instructions for **obtaining DHS data** in the required .dta format, and describes the **DHS in comparison with other potential care seeking data sources**. The **Care Seeking Data Requirements** section of Phase 3 features a fictional DHS from Country X as the example care seeking data source.

DHS datasets, like many survey datasets, contain numeric codes for variable names and/or values. This means it is essential to consult the survey documentation to determine which variables and which values to select while working in the wizard.

The .dta file for the DHS Individual Recode will be accompanied by a FRW file, along with several other documentation files. The FRW file contains the labels for all variable and value codes, and provides summary tables for each variable, at the national level. For an example, see the fictional **FRW file accompanying the Country X DHS** which provides weighted frequency tables for a few key variables.

Variable codes for the DHS are standardized, so they are the same for all countries. However, not all countries have data on all variables. Some variables that may be helpful for your PPA include:

Sample Weights (to select in Step 2.2 of the wizard)

- v005 – Sample weights
- For accurate sample size reporting enter .0000001 as the weight multiplier

Geography (to select in Step 3.1 of the wizard for subnational PPAs)

- v024 – Region (or highest administrative unit below national level)
- v025 – Type of place of residence (e.g., urban/rural)
- SDIST – District
- SREGION – Administrative level below region
- SREG1, SREG2, etc. – Custom country geographies

Demographics (to subset by and generate separate PPAs for different groups of people)

- v013 – Age in 5-year groups
- v106 – Highest educational level
- v190 – Wealth index

Care Seeking

- h44a – Place of care seeking for child diarrhea
- h46a – Place of care seeking for child fever
- v829 – Place of HIV test (general)
- v842 – Place of antenatal HIV test
- s1113 – Place of care seeking for any illness or treatment in the last 30 days, among adults

You must subset DHS datasets according to your care seeking variable of choice in Step 2.3 of the wizard, excluding NA values. You may optionally subset by an additional variable. In Step 3.1, select the same care seeking variable for Facility Type. *Do not* select a variable for Health Sector.

Many countries have the care seeking variables split out further with a numeric extension, e.g. h46a_1, h46a_2, h46a_3, etc. In this case, select h46a_1 only. This corresponds to the first child for whom a given woman sought care. Since the PPA is concerned with adult care seeking preferences, only one child per woman should be counted. *Because of this, care seeking numbers and proportions at the national level may be slightly different in a PPA than in a DHS summary report or in the FRW file.*

2. What are some of the most common data cleaning challenges relevant to working in the PPA Wizard?

Misspellings, multiple alternate spellings, or inconsistent capitalization

You will notice these as soon as you see your data summarized in the wizard. For example:

- Step 3.1 summarizes the columns you select for Health Sector, Facility Type, and Level of Geographic Aggregation (if applicable). It may not look too bad here (e.g. a few alternate spellings like Centre, Center, Ctr). However, when you get to Step 4.2, you may expect to see a list of 15-20 Health

Sector/Facility Type combinations but instead you see hundreds! All those alternate spellings multiply, costing you time.

- In Step 3.2 you may expect to see values of "yes" and "no" for a column designating Xpert Availability. However, instead you see 0, 1, Yes, yes, YES, yyes, NO, N0, no, n. While it's easy enough to check the values corresponding to "yes" if you're aiming to run a PPA quickly, it may be too messy when presenting to colleagues or collaborators
- In Steps 3.1, 5.1, and 5.2, alternate or misspellings for geographies will appear. If there are a lot of these, it will become frustrating to map all of them in Step 5.2

In all the above situations there is a threshold at which it is well worth it to go back to your original dataset and clean the columns containing your PPA variables—and it may be a lower threshold than you initially think.

Too many categories for Health Sector and Facility Type

Sometimes datasets are "clean" in that they record values in an internally consistent manner. Yet, if there are too many categories (factor levels/unique values) it can make the data source unwieldly to work with in the wizard. This challenge arises most commonly when:

- A data source contains columns for Health Sector and Facility Type *and*
- There are many possible categories for Health Sector, Facility Type, or both

You need at most four categories for Health Sector (Public, Private, Informal Private, and a custom sector). It is fine to have a *few* more categories in your raw data than you plan to use in the wizard. For example, NGO, faith-based, company, and humanitarian are sectors commonly grouped into the private sector in the PPA. However, if your dataset has highly specified values for Health Sector it will have the same multiplicative effect as many alternate spellings, leading to too many Health Sector/Facility Type combinations to map in Step 4.2. If the sheer number of Health Sectors and/or Facility Types step seems it will lead to too much work, or seems like it will result in a health facility mapping scheme that is too complex to interpret at a glance, it may be worthwhile to revisit the raw data and create a new column for Health Sector, based on a higher-level classification.

3. What if I want separate PPAs for different groups of people (split by age, gender, wealth index, etc.)?

In this situation you need to create a separate PPA for each group, for example one for men and one for women. This is a good time to make use of the "duplicate" feature on the Team PPAs page. First, create the PPA for one gender by sub-setting your data to include women only in Step 2.3 and carrying out the rest of the steps as you otherwise would. Then duplicate the women's PPA, change the name to reflect men, and edit Step 2.3 to subset to only men. Make sure to check Step 4.2 to see if any new values (health facility types) come into scope based on the new subset, as they will need to be mapped to PPA Sectors and Levels. No additional changes to the inputs should be required to generate the men's PPA.

Depending on the groups of people a team wishes to compare, multiple care seeking data sources may be required. For example, the DHS Individual Recode would not work to differentiate between men and women since it is a women's survey. Alternatively, the DHS women's survey could be a good choice for a pediatric PPA since it includes care seeking *for children's illness*. If available, a different data source such as a prevalence survey could be used for an adult PPA based on adult care seeking for their own illness. A prevalence survey could be used to create separate men's and women's PPAs, if desired.

4. What if I want to create a PPA for just one subnational area (Region, State, City, etc.)?

In this case, create a "national" PPA. Technically, a national PPA is a PPA aggregated at the highest level to produce a single visual. This can represent any geographic area that the data covers. If necessary, subset the raw data to include *only* the single geography of interest in Step 2.3.

5. What happens if I change my subset for a data source after completing all the steps? Will the wizard retain my inputs for that data source?

Yes! So you will not need to redo any inputs you already provided. Changing the subset means that some values (people or health facilities) in your raw data will come into scope for your analysis, and/or some values will go out of scope. For those previously in scope that remain in scope, you've already done the work—the wizard will retain the selections you've made and the inputs you've provided. If new values come into scope as a result of changing the subset, you will need to provide additional input specific to those values. For example, in Step 3.2 new values for Service Availability may need to be selected (but your previous selections will be retained). In Step 4.2, new Health Sectors and/or Facility Types may need to be mapped (again, your previous mapping will be retained). The same applies to geographies in Steps 5.1 and 5.2.

6. What happens if I edit a data source and then re-upload it into the wizard? Will the wizard retain the information I entered for the previous version of the data source?

No! So be careful. Firstly, you may not have two data sources with the same name stored in the PPA Wizard. If you edit a data source and upload it without renaming it, you will receive an error message. To keep the original name for the data source you must delete the old version in the PPA Wizard and then re-upload the new version. Or, secondly, if you want two different datasets with two different names, upload the newly named version afresh. Either way, you must start over with providing inputs for the new version of the data source if you wish to use it.

If you notice problems with your raw data while you're working in the wizard, the best course of action is to STOP, go back to your raw data, and fix the problems. The sooner you do this, the less work you will need to redo in the wizard after you upload the corrected dataset.

7. Is the PPA Wizard secure?

Yes. The site is secured with https, providing secure transfer of data over the internet. Any user of the PPA Wizard is required to log in with a username and password. To ensure the security of you PPA data, it is recommended that you do not share your password with anyone. Personal identifiable fields should be removed from data sources prior to uploading them into the wizard. While these fields are not used in the PPA, administrators on the team will be able to see the values in these fields if they choose. Read-only team members will only see values for the variables that their team has selected for the PPAs in their Team Space.

8. Is it possible to download the raw data from the PPA Wizard?

No. For data security and privacy reasons, it is not possible to download the raw data from the PPA Wizard.