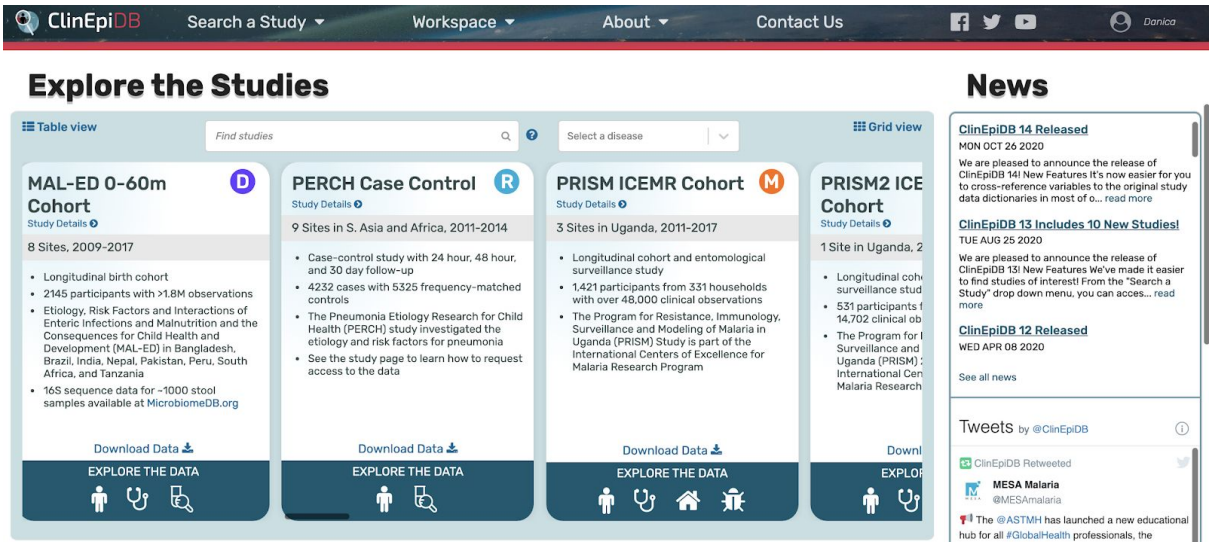
Exercise 2: Review of search types, units of analysis, study designs

This exercise will introduce you to how to execute different types of searches on study data on ClinEpiDB. We will also review units of analysis and study designs.

On the ClinEpiDB homepage, notice the small icons that are located along the bottom of each of the study cards. These icons represent the 5 types of searches you can execute on the data:



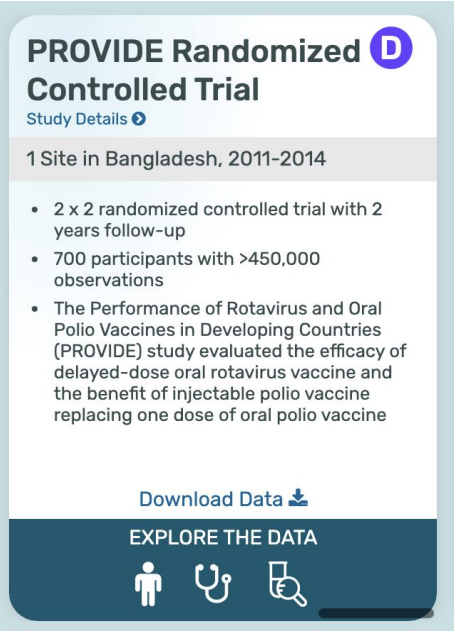
When data are collected in an epidemiologic study, the unit of analysis is an important consideration. It helps to define what is being studied and how we can interrogate the data. Each of the different searches in ClinEpiDB represent a different unit of analysis.

	Initiates a participant-level search	Returns one row of data per participant.
	Initiates an observation-level search	Returns one row of data per observation. There can be multiple observations (days of data collection) per participant. If each participant was observed no more than 1x, all data can be found using a participant-level search and a separate observation-level search will not be possible.
	Initiates a household-level search	Returns one row of data per household or household observation. If only one participant per household was enrolled, all data can be found using a participant-level search and a separate household-level search will not be possible.

	Initiates a sample-level search	Returns one row of data per sample. There can be multiple samples collected at each participant's observation. If no more than 1 sample was collected from each participant, all data can be found using a participant-level search and a separate sample-level search will not be possible.
	Initiates a light trap-level search	Returns one row of data per light trap collection. There can be multiple light traps per household.

If you look at the various study cards, you will notice that not all of these search types are available for each study.

Take a look at the PROVIDE Randomized Controlled Trial study card, and notice that you can choose to initiate either a participant-level search, an observation-level search, or a sample-level search. Even though data about the household was collected in the PROVIDE study (for example, the drinking water source for the household), you are not able to initiate a household-level search. This is because each household in the study enrolled exactly one participant, and so a separate household-level search is not needed to return data with granularity at the household-level.



Why might an observation-level search be necessary for a longitudinal cohort and not for a case control study? What about for a cross-sectional survey?

In a longitudinal cohort study each participant will have repeated observations of the same variable and we are generally interested in the status of different measured exposures and outcomes over time. The study may have some information that is only collected once and applies generally to the participant (sex, date of birth, month of weaning initiation) or information that is collected repeatedly over time (illness status, height, weight, breastfeeding status). We can ask questions about the participants in the study or the repeated observations of those participants or a combination of both. For a cohort study you might ask, how many *participants* had at least one positive test for streptococcus? In this case, you would want to conduct a participant level search. If you were asking, how many positive tests for streptococcus occurred in the entire cohort, you would want to conduct an *observation* level search. Depending on your approach, you may most frequently use the participant level search however as you develop skills building complex queries you may discover the additional advantages of this flexibility.

For a case-control study without any follow-up, individuals are enrolled based on their outcome and their exposure status may be retrospectively assessed but resulting in only one measurement of each variable of interest. Conversely, we may want to ask questions about aggregate groups of participants within households. In this case, we can use the household level search.

Note that by selecting an observation level search this doesn't mean that you will be restricted to exploring variables that were only collected repeatedly; you will still have access to all variables in the study but the primary unit of analysis will be on the observation level.

In practice, study designs may not always be as straightforward. A case-control study may conduct follow-up, for example, or a survey may have some repeated measurements for some individuals. In these scenarios it's best to initiate the search you are most comfortable with or available to you and explore the data and examine the number of participants compared to the number of observations and see how they differ. This will help you to understand the structure of the data.

For the following questions, which search type might you want to initiate?

1. How many monthly MAL-ED stool samples tested positive for Cryptosporidium?[1]

2. How many participants in the PROVIDE study had at least 10 total diarrheal episodes?[2]

3. What is the relationship between dwelling type and household wealth index in the PRISM study?[3]

4. In the GEMS study, how many moderate-to-severe diarrheal cases had controls with a positive Shigella microbiology test?[4]

[1] Sample-level search
[2] Participant-level search
[3] Household-level search
[4] Participant-level search