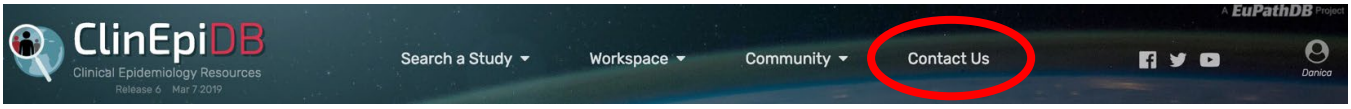
ClinEpiDB Introduction and Variable Exploration

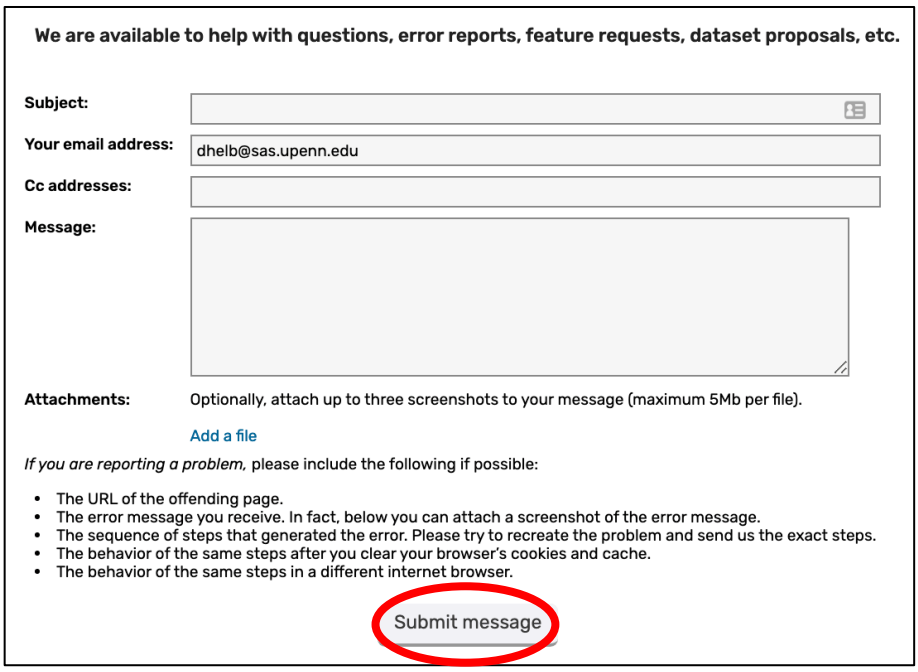
Exercise 1: Getting help or submitting a comment

In this exercise you will learn how to submit comments or ask for help from the ClinEpiDB team.

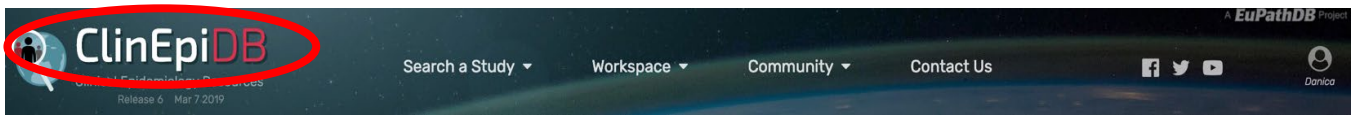
1. To submit a comment or ask a question of the ClinEpiDB team, click on the 'Contact Us' link on the top of the homepage. When you click on 'Contact Us' a form will appear.



2. Fill out the form that appears and click "Submit message". Your message will be sent directly to the ClinEpiDB outreach team, who will be able to answer your question or address your concern.

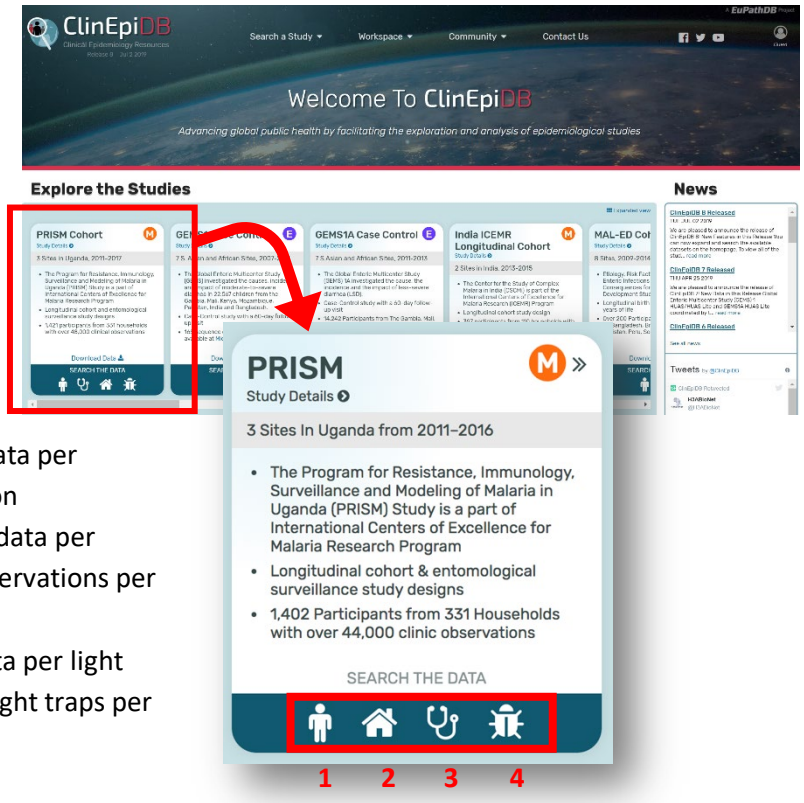


3. When you are finished, return to the front home page again by clicking on the "ClinEpiDB" logo.



Exercise 2: Accessing ClinEpiDB data & search types

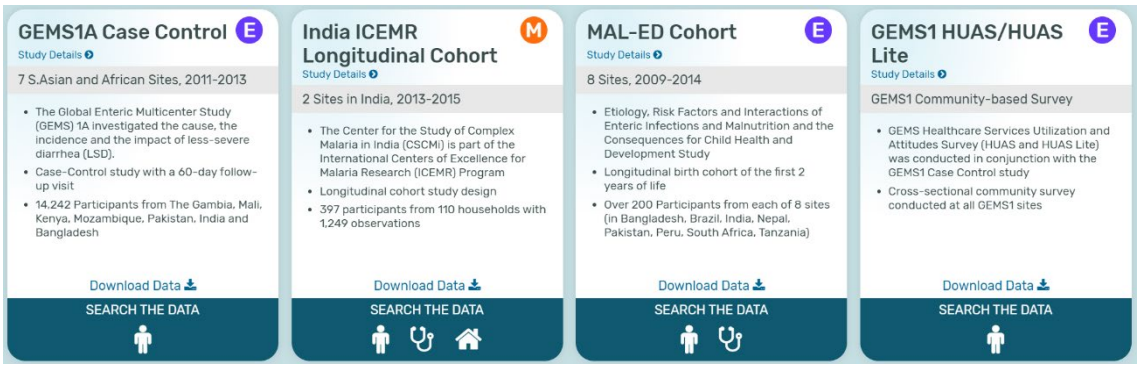
This exercise will introduce you to the ClinEpiDB homepage and how to execute different types of searches on study data.



From the ClinEpiDB homepage, look at the small icons that are located along the bottom of the PRISM study card. The PRISM study has four types of searches you can execute on the data:

1) Participants – returns one row of data per participant
2) Households – returns one row of data per household or household observation
3) Observations – returns one row of data per observation (can have multiple observations per participant)
4) Light traps – returns one row of data per light trap collection (can have multiple light traps per household)

Note that PRISM is a longitudinal cohort study design. If you look at the other study cards; you'll notice that not all search types are available. 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.



In a longitudinal cohort study each participant will have repeated observations and we are generally interested in the status of different measured exposures and outcomes over time. 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.

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? (Circle the search type)

1. How many participants in the study stopped exclusive breastfeeding by 9 months of age?

2. What is the relationship between dwelling type and household wealth index?

3. How many monthly stool samples tested positive for Cryptosporidium?

4. How many moderate-to-severe diarrheal case participants had controls with a positive Shigella microbiology test?

Exercise 3: Explore Participant Search Filters

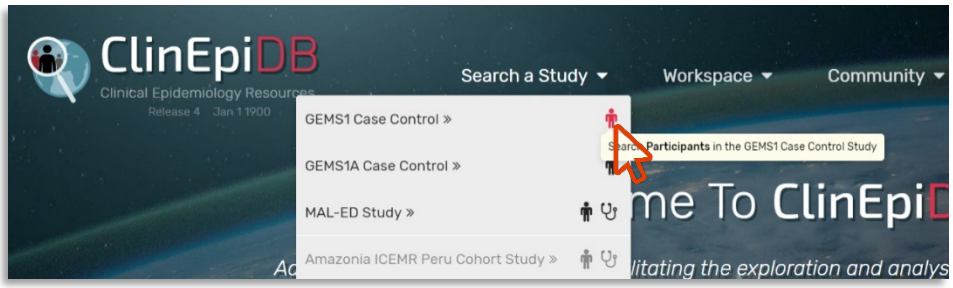
What is the prevalence of wasting and stunting at enrollment in the GEMS Study?

To explore how anthropometric data is distributed in the GEMS population, this exercise will take you extensively through the search wizard. You will discover how you can access data variables and their corresponding histograms and bar graphs.

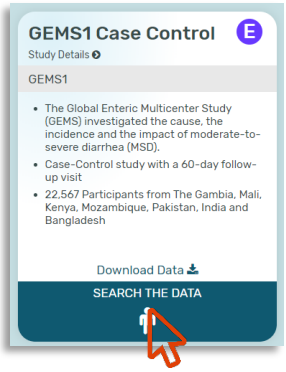
1. Navigate to http://clinepidb.org in your browser.

For GEMS, "Participants" refers to any children who were enrolled in the study. You can access the participant search wizard in a few ways:

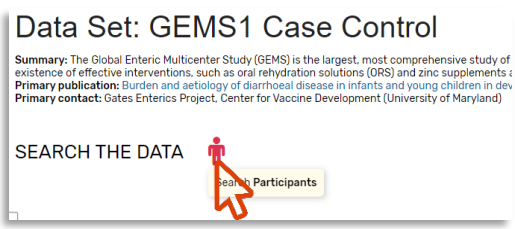
a. Through the navigation bar at the top of the page under Search a Study>GEMS1 Case Control>Search Participants



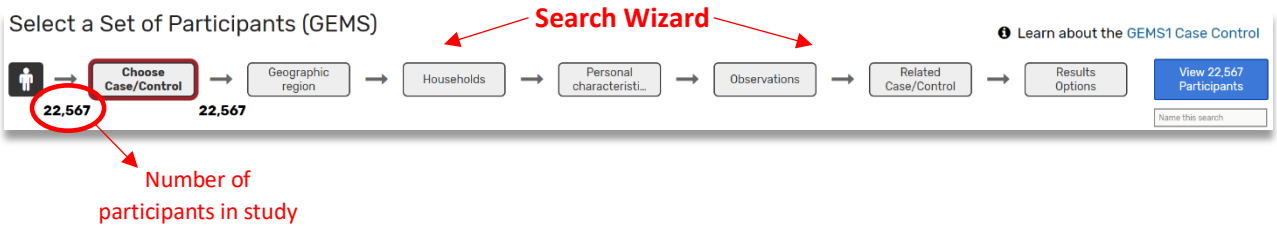
b. Click on the 'Participant' icon on the GEMS1 card:



c. Or click on the main text of the GEMS1 Case Control Card and a new page with information about GEMS1 will load. On this page, find 'SEARCH THE DATA', click on the small human icon next to this.



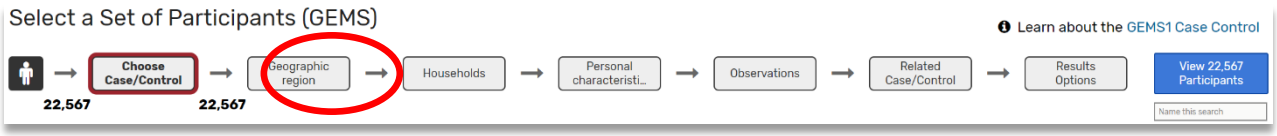
When you open the Participant Search page across the top below the header will be the “Search Wizard”. The purpose of the search wizard is two-fold. First, it creates a simple way to categorize components of the data allowing for a step-wise approach to building searches. Second, it allows you to explore the data to see what the raw number and distribution of characteristics are in both the full dataset or filtered data. Spend some time reading and mouse over the different search wizard boxes.



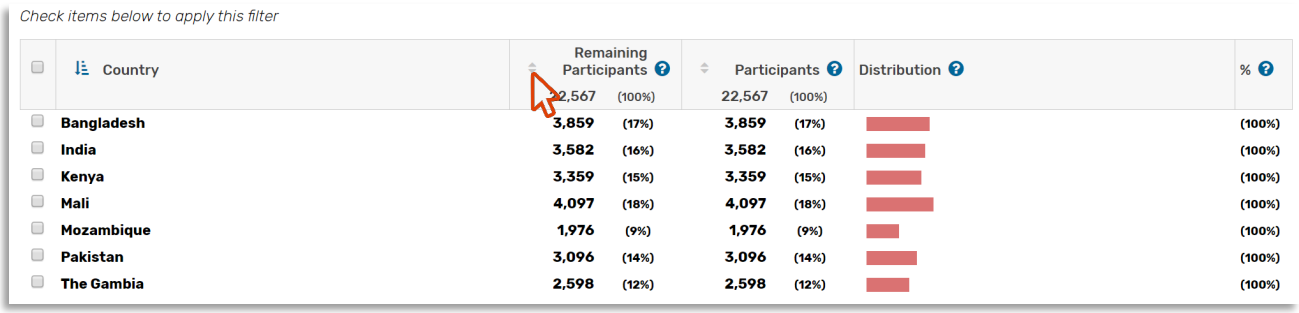
The number below the black square Participant icon represents the total number of Participants that are included in this dataset.

Finding Variables and Examining Data

1. The first open variable listed is ‘Choose Cases/Control’ as a specific selection at the top where you can restrict your search to either cases, controls or the entire dataset. By default, we will search the entire dataset. Leave it this way for now. Next, click on the geographic region search box.



2. Without applying a geographic region selection, can you tell which region had the greatest number of participants?[1] Notice that you can sort the columns of data by clicking the small up and down arrows in the heading of the columns.



3. Next, click on the Personal characteristics search box. This will reveal a slightly more complex table that includes variables grouped into categories in blue text on the left and aggregate data about the selected variable on the right. Variables included in “Personal

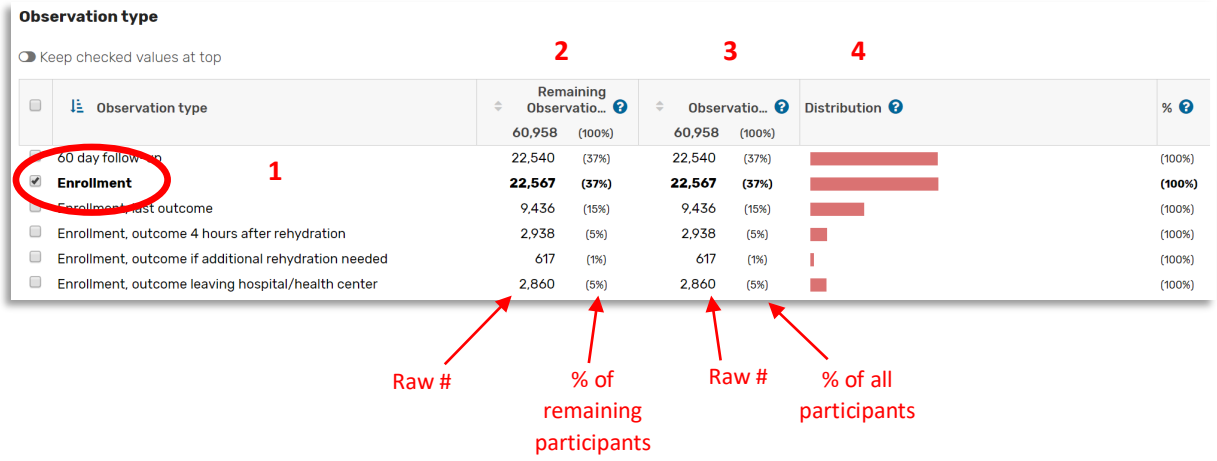
[1] Mali

characteristics" generally are only reported once per participant and can also relate to information about how the study was conducted (administrative information for example).

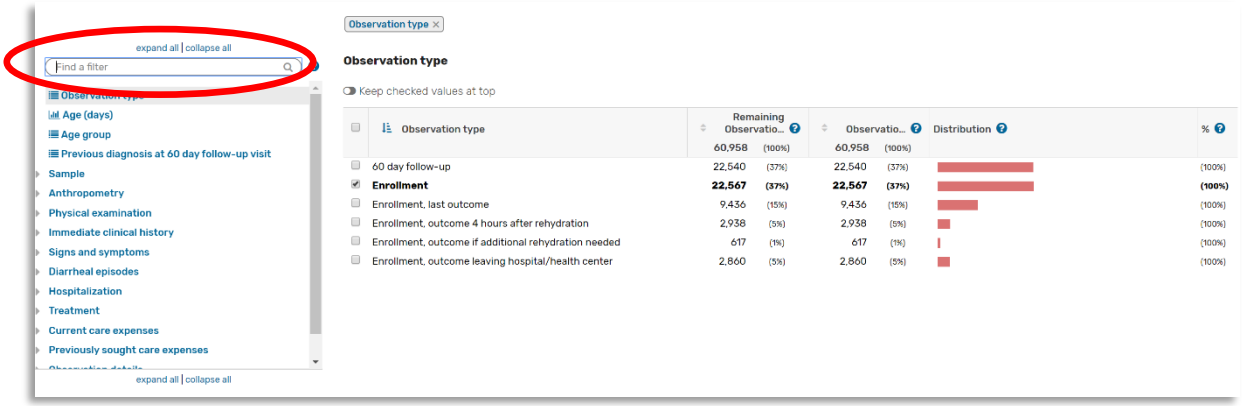
4. Click on the 'Observations' category to find information on variables that were measured during the study. By default, the 'Observation Type' variable is open, select 'Enrollment' from the values.

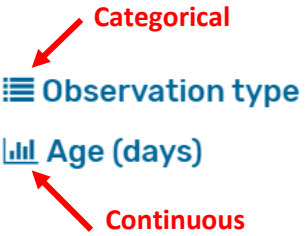
   The Observation type variable is categorical, therefore the data contained in this variable is displayed in a table that includes:

   1. Possible values for the variable
   2. Remaining observations – number of observations with that value after filters have been applied
   3. All observations – the total number of observations with that value
   4. Distribution – length of red bar indicates the number of observations with that value

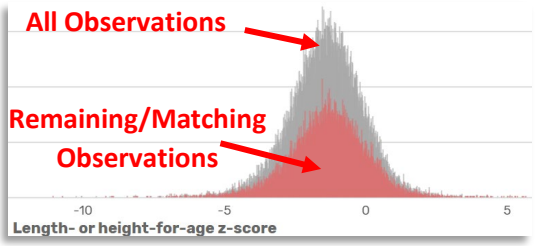


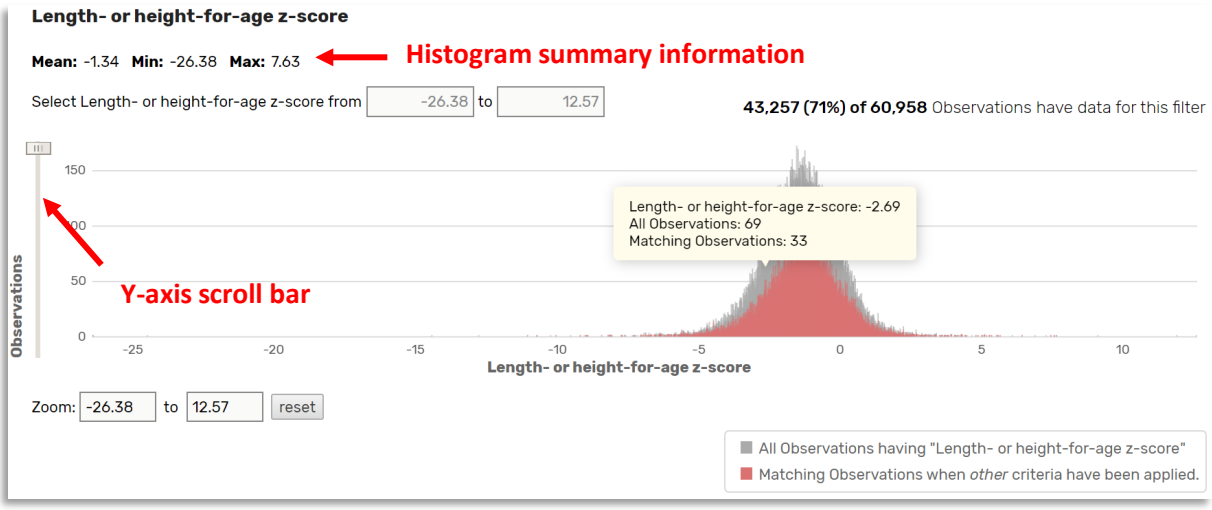
5. To estimate the prevalence of stunting in both cases and controls we first have to find the height-for-age z-score variable. We can do this by typing in the 'Find a filter' box on the left'. If you start typing it will narrow your selection.





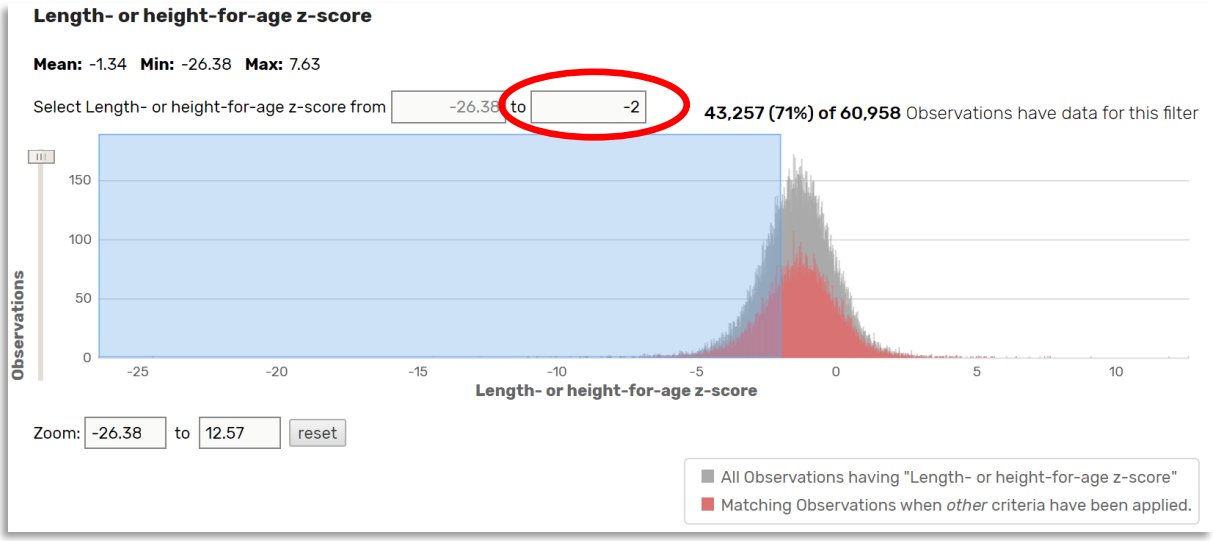
6. In ClinEpiDB we indicate categorical versus continuous data using the icons to the left of the variable. For continuous data in the database that has greater than 10 values, the data is displayed as a histogram rather that in table format. This allows you to see the distribution of values of your selected participants. The x-axis displays the value of the variable that you have selected. In this case, the Length- or height-for-age z-score ranges from -26.38 to 12.57 (Note that some of these z-scores may be improbable and may want to be explored more later as potential outliers or erroneous data points). The y-axis displays the count of the number of observations that match that value. When you hover over the data points, a pop-up will show you the value and the number of observations in the overall sample size (All Observations-grey bars) and in those that match any filter variables you have applied previously (Matching Observations-red bars). The point we have highlighted above indicates that there are 69 Length- or height-for-age z-score measurements in the total number of observations that have the value of -2.69 and 33 of these measurements were done at enrollment. Basic statistics about the selected data is displayed immediately above in the histogram summary information. This is calculated based on any filters you've applied previously.





How many participants had an HAZ less than or equal to -2.00? [Hint: You can select the HAZ range using by highlighting a selected area or by entering it in the boxes above.] You can easily convert this to a proportion by dividing the number of participants by the overall sample size (22,567). This indicates the prevalence of stunting overall in all cases and controls. To see how this differs in Cases versus Controls….you can navigate back to the 'Choose Case/Control' search wizard step and see how your selections update the participant counts.

a. HAZ <=-2 Overall = _____[2]
b. HAZ <=-2 Cases = _____[3]
c. HAZ <-2 Controls = _____[4]



7. Next, determine the prevalence of wasting (weight-for-height) overall at enrollment and in cases and controls. Did wasting or stunting differ more by case control status?

a. WHZ <=-2 Overall = _____[5]
b. WHZ <=-2 Cases = _____[6]
c. WHZ <-2 Controls = _____[7]

8. Examine all available filter categories in blue text on the left-hand side of the page in the Observations section of the search wizard. Click on the 'expand all' link at the top of the list, above the "Find a filter" search box. This will reveal all the types of data in each of the subcategories. Scroll down and read through these variables. Spend some time clicking through and examining the distribution of participant data in each variable.

[2] 28.5%
[3] 28.2%
[4] 28.7%
[5] 15.6%
[6] 23.3%
[7] 10.1%