Exercise 2.1: LLINEUP: perform a simple participant-level search

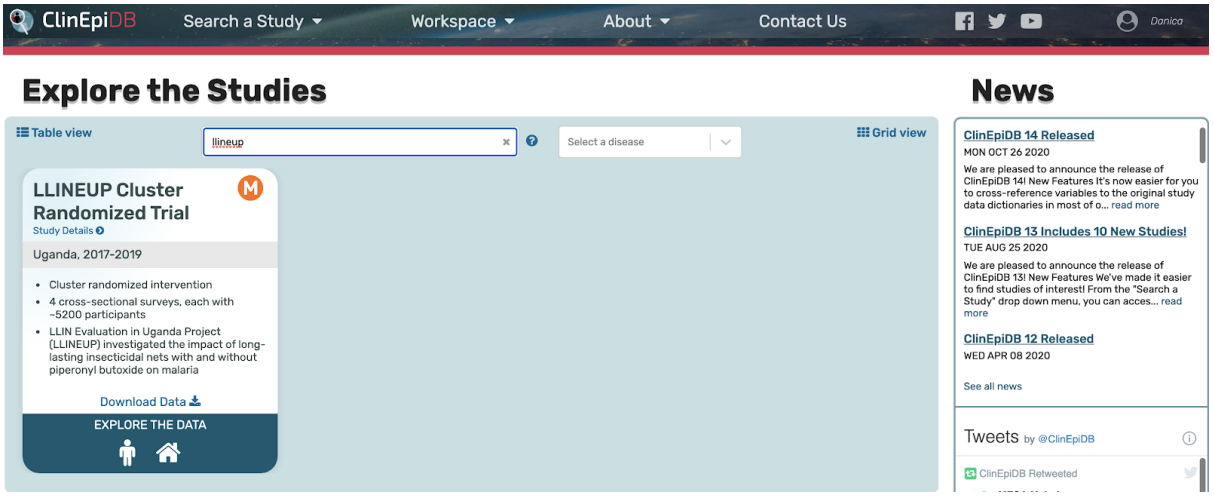
6 months after the long-lasting insecticidal net (LLIN) distribution campaign, how did the parasite prevalence differ amongst children 2-10 years of age in households that received combination LLINS versus those that received conventional LLINS?

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.

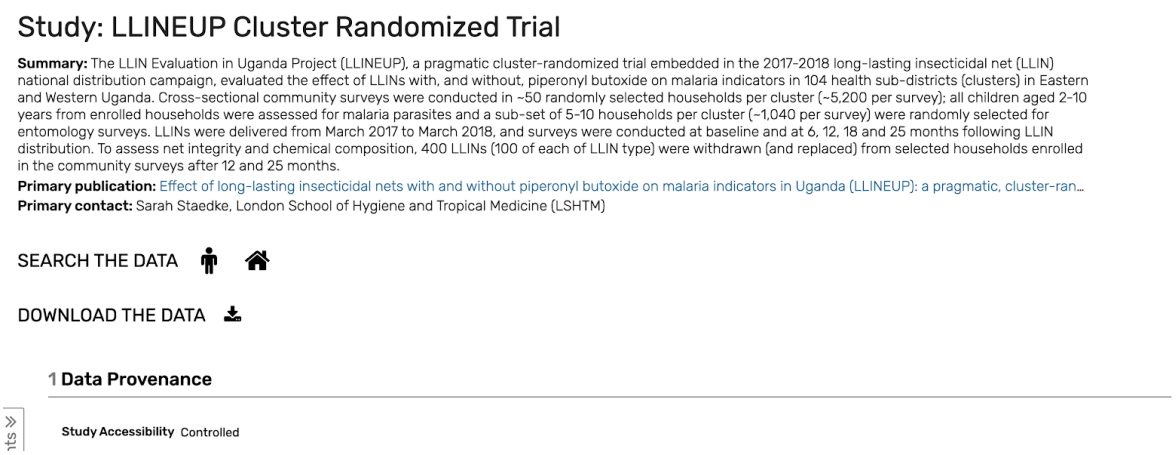
Parasite prevalence describes the fraction of the host population infected with a parasite.

*For this analysis, use "**Plasmodium, by thick smear microscopy**" to determine which participants were infected with Plasmodium parasites. Use "**Cluster as treated, dichotomized**" to identify the set of participants who received either combination or conventional LLINs.*

1. Navigate to http://clinepidb.org in your browser and type "LLINEUP" into the "find studies" search bar to quickly navigate to the LLINEUP Cluster Randomized Trial study card.



2. Before we initiate our search, click on "Study details" on the LLINEUP study card to review the design of this cluster randomized trial.



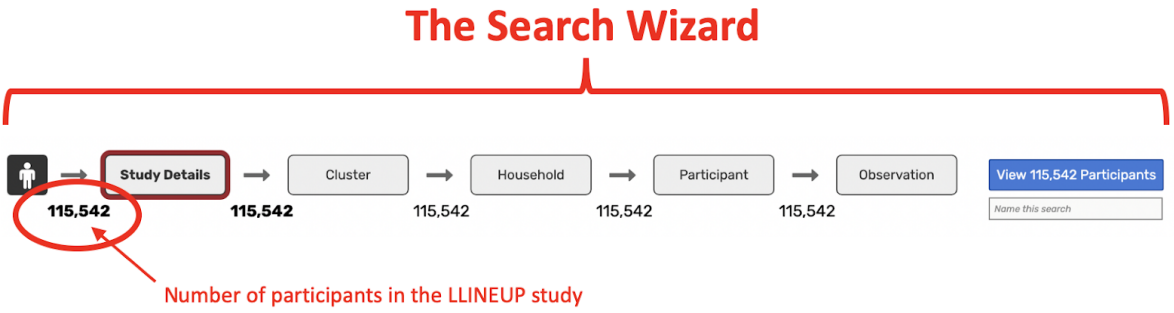
3. Use the information on the LLINEUP study page to answer the following questions and get a better understanding of how this study was conducted:

    - What were the intervention clusters? What were the control clusters?[1]
    - How do conventional LLINS differ from combination LLINS?[2]
    - Were the participants followed longitudinally after receiving the intervention?[3]

4. Since we want to compare the proportion of individuals infected with *Plasmodium* in two populations, the primary unit of analysis will be at the participant-level. Click on the "participant" icon to initiate a participant-level search.

The Search Wizard

5. 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 clicking through the different search wizard boxes.



6. The number below the black square Participant icon represents the total number of Participants that are included in this dataset. If the display on your screen does not show 115,542 participants, make sure that you have clicked on the LLINEUP study and initiated a participant search.

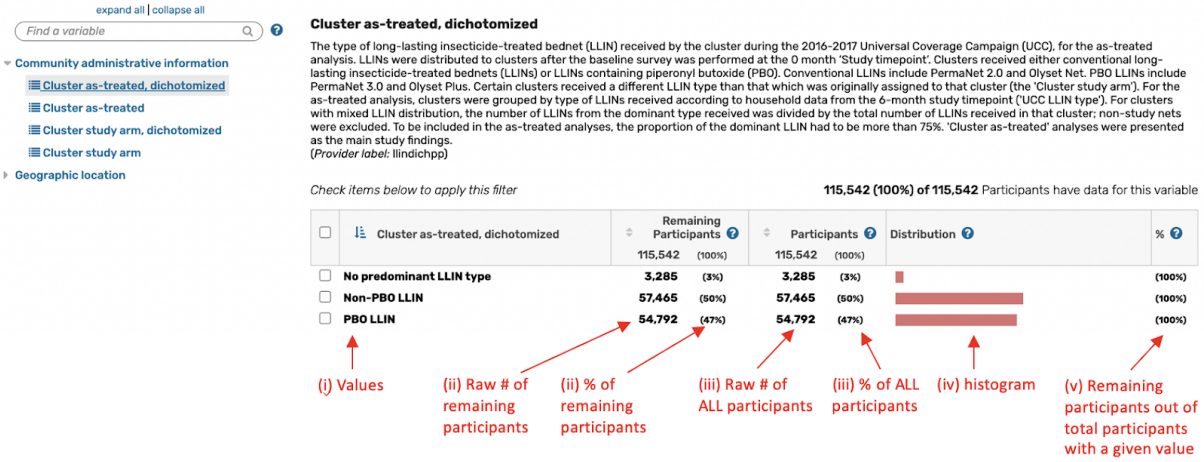
[1] Intervention clusters were health sub-districts randomized to receive combination LLINs. Control clusters were health sub-districts randomized to conventional LLINs.

[2] Combination LLINs combine a pyrethroid insecticide with piperonyl butoxide (PBO), a synergist to potentially overcome pyrethroid resistance in anopheline vectors. Conventional LLINs do not contain PBO.

[3] Even though data was collected at multiple time points following LLIN distribution, participants were NOT followed longitudinally in this study. At each follow-up time point, all eligible children from 50 households per cluster were ***randomly*** selected to take part in a cross-sectional household survey. Because households were enrolled at random for each follow-up cross-sectional survey, each time point surveyed included different participants.

Let's start by determining the parasite prevalence in children 2-10 years who lived in households that received combination LLINs (containing PBO), 6 months after LLINs were distributed. To do this, we will need to select the set of participants who fall within the correct age range and who received PBO LLINs.

7. We will use the search wizard to create the subset of participants that fit our interests. Click on "Study details", the first step in the search wizard. For experimental studies, "Study details" is used to organize key variables such as the study arm. Notice the way the page is set up, with a variable tree table that lists variables grouped into categories in blue text on the left and aggregate data about the selected variable on the right. What categories of variables are included in the "Study details" step for LLINEUP?[4]



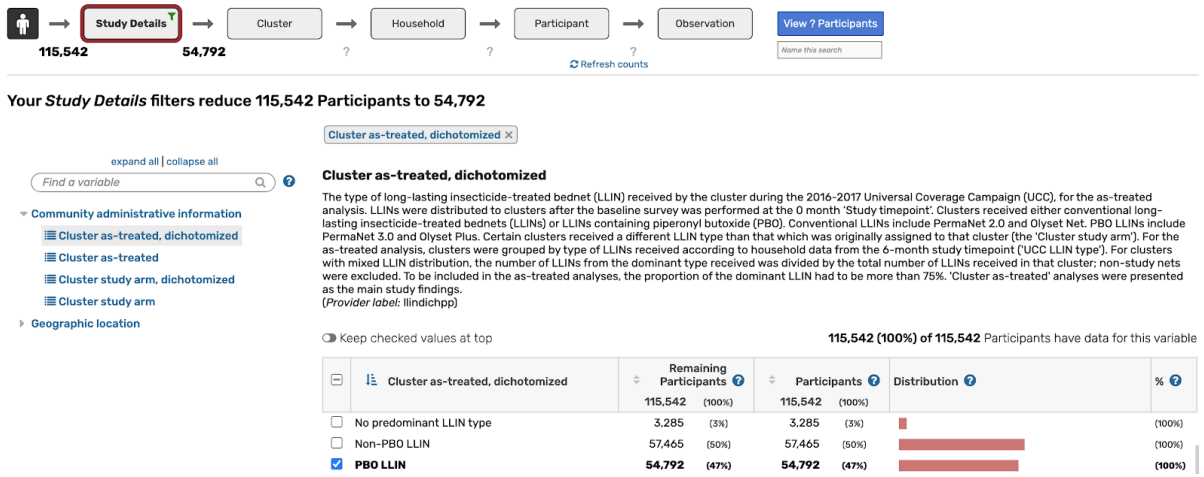
8. Click on the "Cluster as-treated" variable. Notice that a study-specific definition of the variable is provided. You can sort the columns of data by clicking the small up and down arrows on the heading of the columns. Without applying a selection, can you tell which type of LLIN was distributed to the greatest number of participants?[5]

9. We are interested in participants who actually received combination LLINs. Click on "Cluster as-treated, dichotomized" in the variable tree. This will bring up the distribution of participants who belonged to clusters that received either PBO LLINs or non-PBO LLINs, or who belonged to a cluster that did not receive a predominant LLIN type. Select "PBO LLIN" for "Cluster as-treated, dichotomized". What happened to the number of participants indicated in the search wizard? Why did this happen?[6]

[4] The "Study details" step for LLINEUP includes variables organized into two categories: (1) Community administrative information and (2) Geographic location.

[5] PermaNet 2.0 was distributed to the greatest number of participants (n = 43,081).

[6] When you click on the box next to "PBO LLIN", the number of participants that meet your selection criteria has been reduced from 115,542 to 54,792. This is because you have filtered out all of the participants who either belonged to clusters that received non-PBO LLINs or that did not receive a predominant LLIN type.



10. Next, click on the "Cluster" and the "Household" search boxes. Without applying any selections, explore the types of variables that are organized in each of these steps in the search wizard. 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 data in different variables. Why does "Cluster level mean female Anopheles" fall under the "Cluster" search box while "Female Anopheles count" fall under the "Household" box?[7]

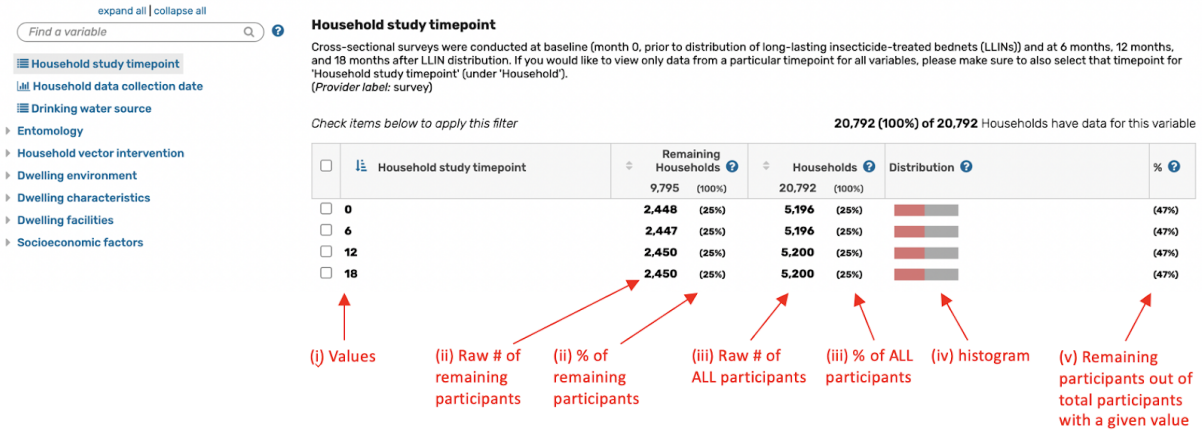
Continuous vs Categorical Variables

11. Click on the "Household" box in the search wizard. Notice that the first two variables in the variable tree have different icons in front of them. In ClinEpiDB, we indicate categorical versus continuous data using the icons to the left of the variable.

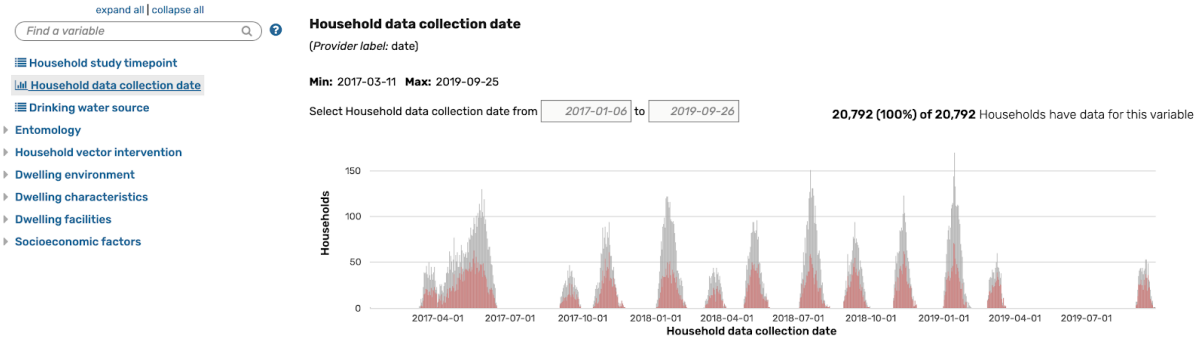
	**Categorical variable** (or continuous variable with ≤ 10 values)	Categorical variables (and continuous variables with fewer than 10 values) are displayed in a table format.
	**Continuous variable** with > 10 values	When a continuous variable has more than 10 values, the data is displayed as a histogram rather than in table format. This allows you to see the distribution of values of your selected participants.

[7] "Female Anopheles count" aggregates the data for the number of female Anopheles that were collected from each individual household, while "Cluster level mean female Anopheles" aggregates the averages for all households belonging to each cluster. Variables organized under "Cluster" were reported for the cluster as a whole, while variables organized under "Household" were reported for individual households.

12. The “Household study timepoint” variable is categorical, therefore the data contained in this variable is displayed in a table that includes:
    i. Possible values for the variable
    ii. **Remaining households** – number of households with a given value AFTER filters have been applied
    iii. **Households** – the total number of households with a given value
    iv. **Distribution** – a histogram, where the length of each red bar indicates the number of households with a given value
    v. **%** – remaining households out of total households with a given value *100



13. The “Household data collection date” variable is continuous, therefore the data contained in this variable is displayed as a histogram. The x-axis displays the values for the variable you have selected. In this case, household data was collected between March 2017 and September 2019. The y-axis displays the count of the number of households that match each value. Basic statistics about the selected data is displayed immediately above in the histogram summary information, and is calculated based on any filters you’ve applied previously.

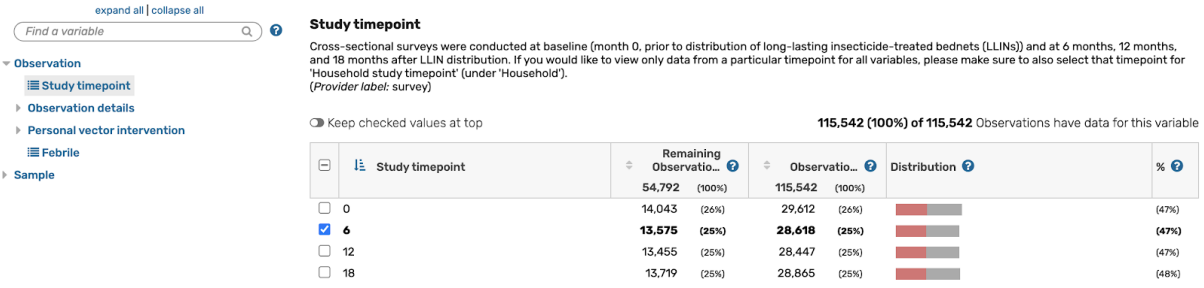


14. Notice that both the “Distribution” column for the “Household study timepoint” table and the histogram for “Household data collection date” are partially grey and partially red. Grey data represents data from ALL households in the study, while red data represents data from households that match your selection criteria. In our example, red represents the distribution of data from households in clusters that received PBO LLINs.

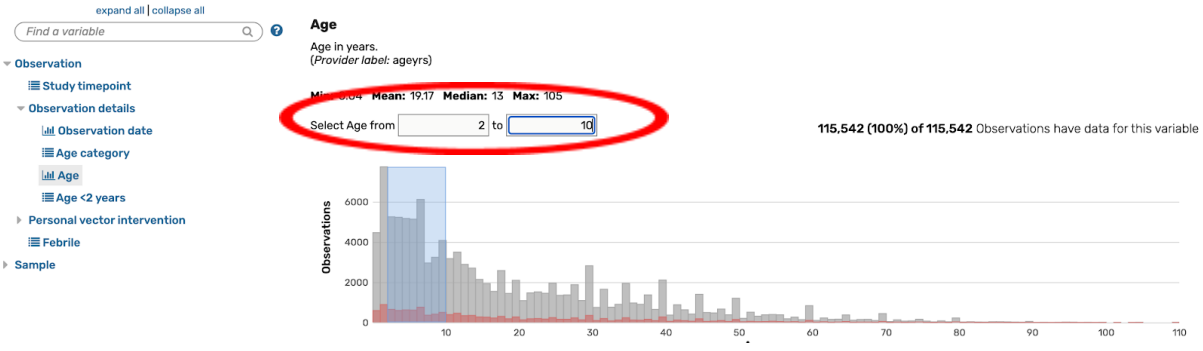
15. Click on the “Observations” box. By default, the “Study timepoint” variable is open. Select “6” from the values for “Study timepoint”, as we are interested in determining the parasite prevalence 6 months after the LLIN distribution campaign.

    - How many total participants were surveyed at the 6 month post-LLIN distribution timepoint?[8]

    - How many participants belonged to clusters that received PBO LLINs and were surveyed at the 6 month post-LLIN distribution timepoint?[9]



16. Next we want to limit our selected participants to those who are 2-10 years of age. Navigate to the “Age” variable, under “Observation details”, and input “select age from 2 to 10”. This will highlight a section of the histogram corresponding to ages 2-10 years, and will limit your selected participants to those within this age range. How many participants now meet your selection criteria (belong to clusters that received PBO LLINs, were surveyed 6 months after LLIN distribution, and were between 2-10 years of age at the 6 month survey)?[10] *Hint: if your histogram looks different from what is shown here, update “Bin width” to 1 under plot settings.*

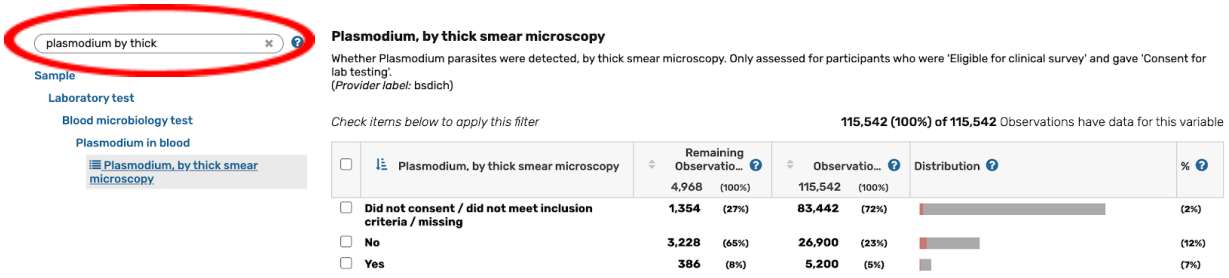


[8] 28,618 participants were surveyed at the 6-month follow-up timepoint.

[9] 13,575 of the 28,618 participants who were surveyed at the 6-month follow-up timepoint belonged to clusters that received PBO LLINs.

[10] 4,968 participants met the selection criteria.

17. Next, within our set of selected participants, we want to determine how many were positive for *Plasmodium.* We know that the name of the variable providing this information is "*Plasmodium*, by thick smear microscopy", and falls under "Observations" within the organization of the search wizard. We can quickly locate this variable by typing in the "Find a filter" box above the variable tree when the "Observations" box is selected. Click to select "Plasmodium, by thick smear microscopy".



18. Parasite prevalence can be calculated with this formula:

$$\frac{\text{\# of participants with } \textit{Plasmodium}}{\text{\# of participants tested}} \times 100 = \text{Parasite Prevalence (as \%)}$$

To calculate the parasite prevalence in our selected subset of participants, determine:

- The number of participants who tested positive for *Plasmodium* by thick smear microscopy (numerator)[11]
- The total number of participants who had a thick smear performed on their blood (denominator)[12]
- **What is the parasite prevalence 6 months after LLIN distribution in children 2-10 years of age who live in clusters that received PBO LLINs?**[13]

[11] 386 of the selected participants had *Plasmodium* detected by thick smear microscopy.

[12] 3,614 of the selected participants had a thick smear performed (386 + 3,228 = 3,614).

[13] The parasite prevalence amongst children 2-10 years of age who live in clusters that received PBO LLINs in the distribution campaign that occurred 6 months ago is 10.7% (386/3,614 * 100).

19. If we want to instead determine the parasite prevalence 6 months after LLIN distribution in children 2-10 years of age who live in clusters that predominantly received ***non-PBO LLINs,*** notice that the only difference in these two sets of participants is the type of LLIN the cluster received. To quickly determine the parasite prevalence in this second population, click back on the "Study details" box in the search wizard. Select "Cluster as treated, dichotomized", uncheck the box next to PBO LLINs and instead check the box next to non-PBO LLINs. How has the number of participants meeting this new selection criteria changed?[14] *Hint: click "Refresh counts" under the search wizard.*

20. **What is the parasite prevalence 6 months after LLIN distribution in children 2-10 years of age who live in clusters that received non-PBO LLINs? How does this compare those that received PBO-LLINs?[15]** *Hint: click on the "Observations" box in the search wizard and navigate to "Plasmodium, by thick smear microscopy". All of your previous selections (6 month study timepoint, ages 2-10 years) will still remain.*

[14] The number of selected participants increases from 4,968 to 5,197 when selections for "Cluster as-treated, dichotomized" switch from PBO LLINs to non-PBO LLINs.

[15] The parasite prevalence amongst children 2-10 years of age who live in clusters that received non-PBO LLINs in the distribution campaign that occurred 6 months ago is 14.5% (556/(556+3,288) * 100). 6 months after LLIN distribution, children that resided in areas that received LLINs with PBO had a lower prevalence of *Plasmodium* (10.7%) than children residing in areas receiving LLINs without PBO (14.5%).