Interactive III. Exploring and analyzing malaria epidemiology datasets

Exercise 1: How does house design relate to malaria incidence in children in Uganda? (40 min)

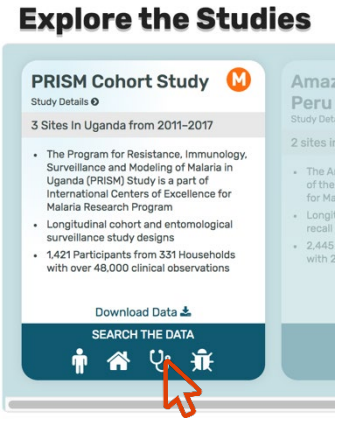
In this exercise we will use the PRISM cohort study to determine if there is an association between house design and risk of *Plasmodium* infection. The PRISM study collected variables such as wall, floor, and roof type, the type of eaves used, and the presence of airbricks. You will explore the data using the observation search. You will also use the Shiny Apps to visualize the data while trying to answer the question of how house design relates to malaria in children in Uganda.

Start an Observations search

Take a look at the PRISM Cohort Study card. Notice the 4 icons located under the words 'Search the Data'. Each icon starts a different type of search, Participant, Household, Observation, or Light trap. Within each search, you can subset the data based on any of the variables that were collected in the study, the only difference is in the data tables you get in the end. A Participant search will return one row of data per study participant, an Observation search will return one row of data per observation, etc.

We want to determine how house design relates to malaria incidence in the PRISM cohort study. This study collected data from all participants every three months and from a subset of children every month. Participants were also encouraged to come to the clinic any time they were sick, so each participant has multiple observations. *What type of search do you think we want to do based on this information? Why?*[1]

Start an Observations search by clicking the 'Observation' icon on either the study card or the drop-down menu under "Search a Study". This will take you to a page with the Search Wizard at the top.



[1] An Observation search will return one row of data per observation and allow us to capture data from every individual malaria episode.

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 subsetted data.

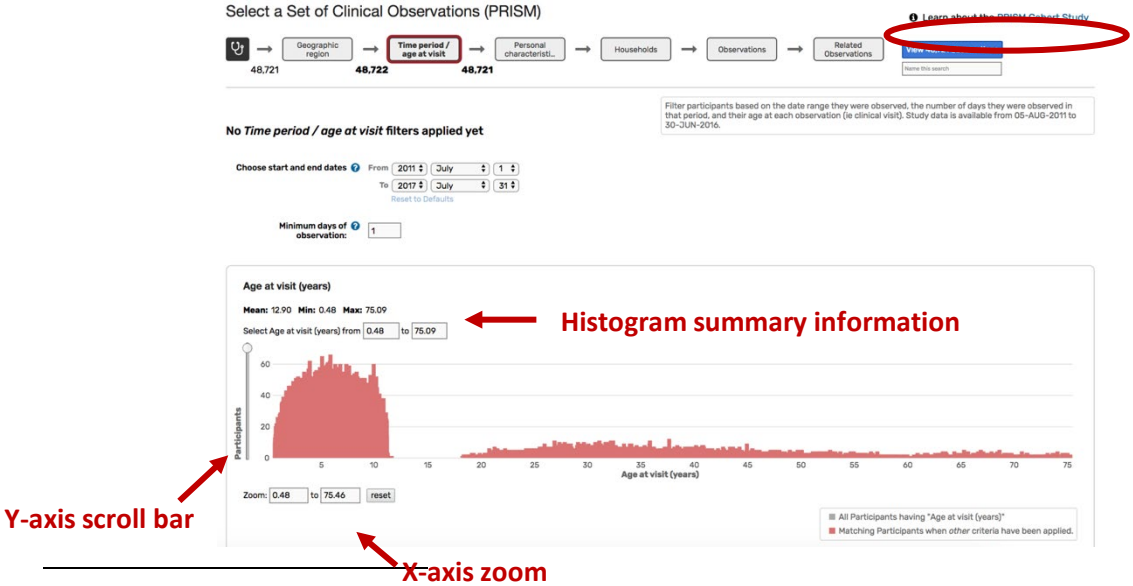
Notice that there is a number below the black square Observation icon. This number represents the total number of Observations that are included in this dataset.

Finding and Examining Data Variables – Determining what house design and malaria-related variables were collected

We are interested in looking at how house design relates to malaria in children. *Where might you expect to find age-related variables in the Search Wizard? How about house-related variables? And malaria variables?*[2]

1. Click on "Time period/age at visit" in the Search Wizard. We know that data can be either categorical or numerical. For numerical data in our database that has greater than 10 values, the data is displayed as a histogram rather than in table format. Notice that the histogram on the right has an x-axis range of zero to above 75. *Can you find the maximum Age at visit (years)? How about the average?* [3]

   The data for this variable has a very interesting distribution pattern. *Where can you find more information on how a study was conducted?* [4]

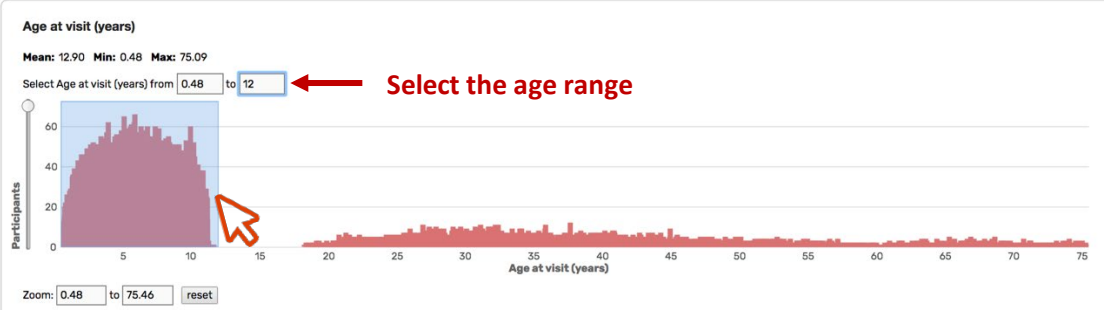


[2] Time period/age at visit and Personal characteristics have variables related to age, Households has variables related to house design, and Observations has variables related to malaria episodes

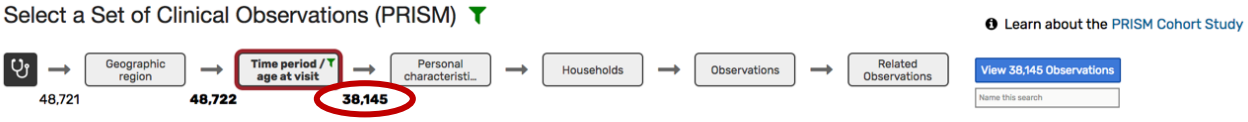
[3] The max age at visit is 75.09, the average is 12.90.

[4] The dataset page includes details about the methodology and links to publications. This can be accessed via the link "Learn about the PRISM Cohort Study" in the Search Wizard, by clicking on the body of a study card on the homepage, or by clicking on the study name under the "Search a Study" drop-down menu. In this case, the study enrolled children from age 6 months to 10 years and one primary caregiver per household.

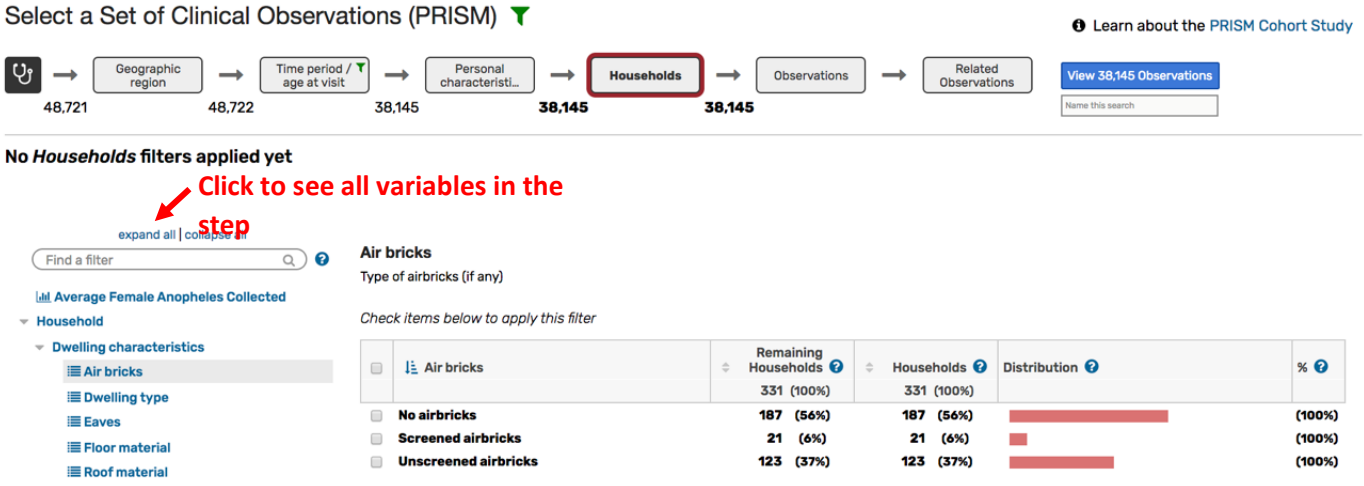
2. Select the children you want to subset the data on by either typing the age range you want into the boxes or by clicking and dragging on the graph to select the range of interest with your mouse.



3. Notice that the numbers in the Search wizard have updated based on the selection we applied in "Time period/age at visit". *How many observations for children occurred in this study?* [5]



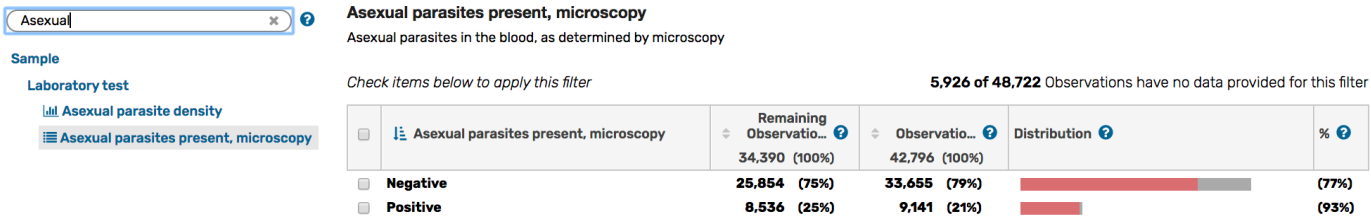
4. Next, take a look at variables that relate to house design by clicking on "Households" in the search wizard. Examine the available categories in blue text on the left-hand side of the page. Clicking a category reveals the variables within. Clicking the 'expand all' link at the top of the list will reveal all the variables in each of the subcategories. Scroll down and read through them. *How many households have open eaves? How many households have mud walls?*[6]



5. Now let's look at malaria-related variables. Click on "Observations" in the Search Wizard and look through the available variables.

[5] 38,145 observations of children
[6] 116 households have open eaves, 202 households have mud walls

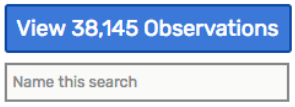
6. Can you find the information about the presence of asexual parasites? (*Hint: to quickly find variables, start typing the name in the "Find a filter" search box. This will expose the variables and subcategories of interest. Try typing "Asexual" and see what comes up*)

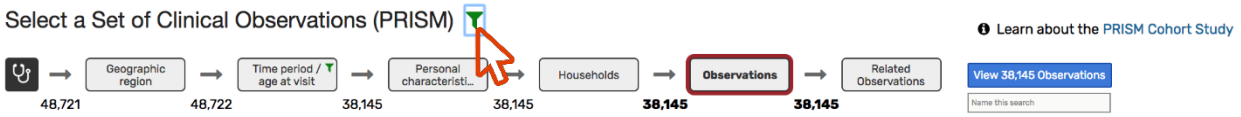


Note that the columns "Remaining observations" and "Observations" show different numbers. "Observations" indicates the total data, while "Remaining observations" indicates the selected data, which in this case is only from children. In the Distribution graphic, grey indicates the data that does not meet the selection criteria, while pink reflects the selected data. *How many times were children found to be infected with Plasmodium by microscopy*? *How about adults?*[7]

Results Page – Is a child more likely to be infected or uninfected upon observation?

Now let's look at features of the results page. Notice that the blue button at the far right-hand side of the wizard now says, "View 38,145 Observations". Note: If your button has some other number here, you might have applied additional selections. To see all the variables that you have subsetted on, click on the green filter icon at the top of the Search Wizard.

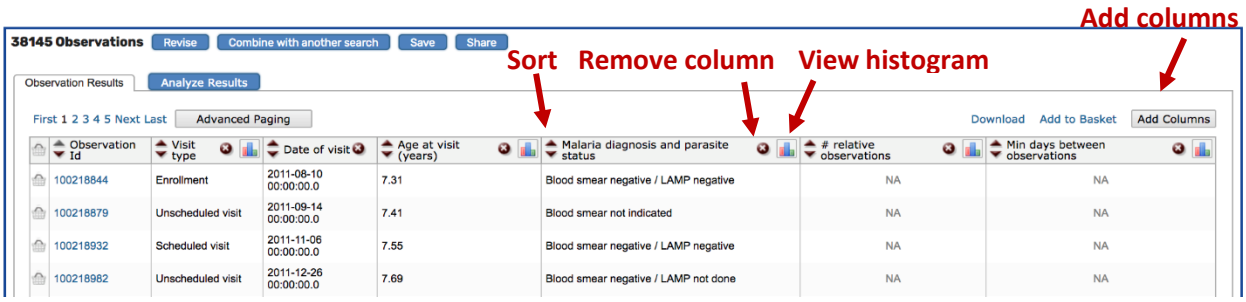




This will activate a pop-up that lists out all of the active filters you have. You can remove a filter by clicking the 'x', or you can edit it by clicking the blue link.

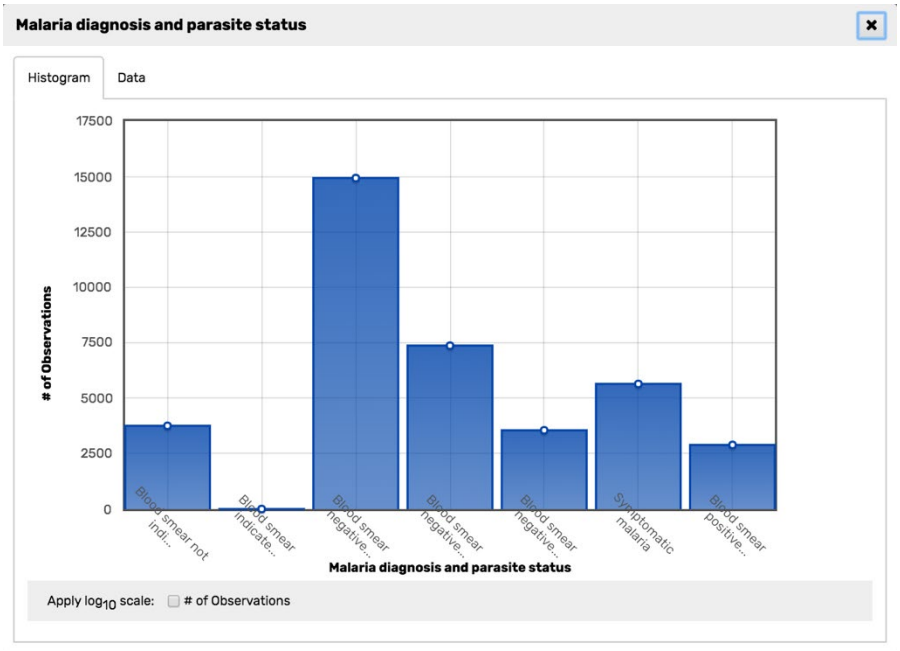
1. Once you have selected only children (<12 years old), click the View Observations button in the Search Wizard. This will load the results page. Explore the features of this page by hovering your mouse over some of the links and buttons. By default, the results table displays 20 rows at a time. If you click on 'Advanced Paging' you can display more than 20 rows per page if you wish.

[7] 8,536 child observations were positive for *Plasmodium* by microscopy, or 25% of all child observations. Since adults have been removed by our search you can calculate this by subtracting remaining observations from observations. 9,141-8,536 = 605



The top of each column in the data table has options for sorting the column in either descending or ascending order and a text description of the column contents. Clicking the 'x' circle will remove the column from the table and the small histogram icon will reveal the distribution of data for that column. Use the "Add Columns" button in the upper right to look at other variables.

2. Try clicking on the small histogram icon for "Malaria diagnosis and parasite status". Examine the pop-up box that displays. The y-axis displays the count of Observations and the x-axis displays the selected variable. *Which bin on the x-axis has the largest number of observations?* [8] *(Hint: Switch from the histogram tab to the Data tab to see the full names of each condition along with tabulated values)*



3. Close the histogram and data table and play around with other columns.

Using the Shiny Apps to Visualize Data

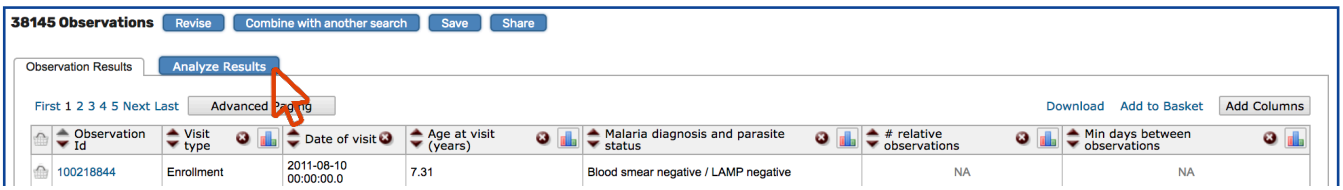
We have 3 different Shiny Apps that are available for visualizing data – Data distributions, Contingency tables, and Data summaries. Each app graphs the data in a different way. We want to know how house design is associated with malaria, so let's start by using Data distributions to look at how malaria diagnosis varies based on the house characteristic of "airbricks" and by site.

[8] Blood smear negative/LAMP negative is the most common observation

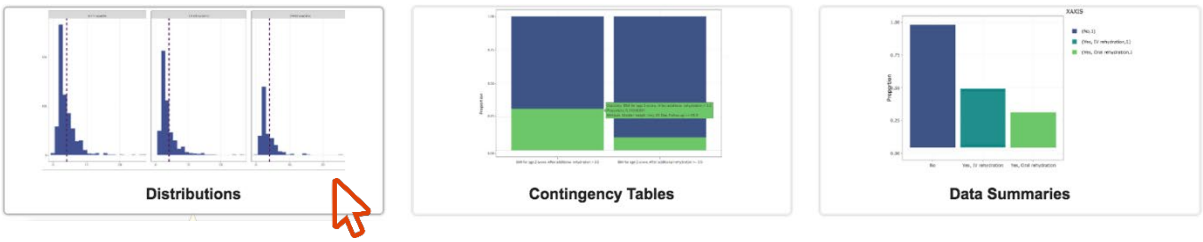
An important warning about these applications; they are intended for data exploration only! You may choose to examine any number of variables or relationships between variables using these apps but all results should be considered hypothesis generating. More complete, rigorous statistical analyses with proper consideration of study design, possible sources of bias and confounders is strongly recommended before any analyses are considered complete.

Data Distributions – How does malaria diagnosis in children vary based on the type of airbricks used?

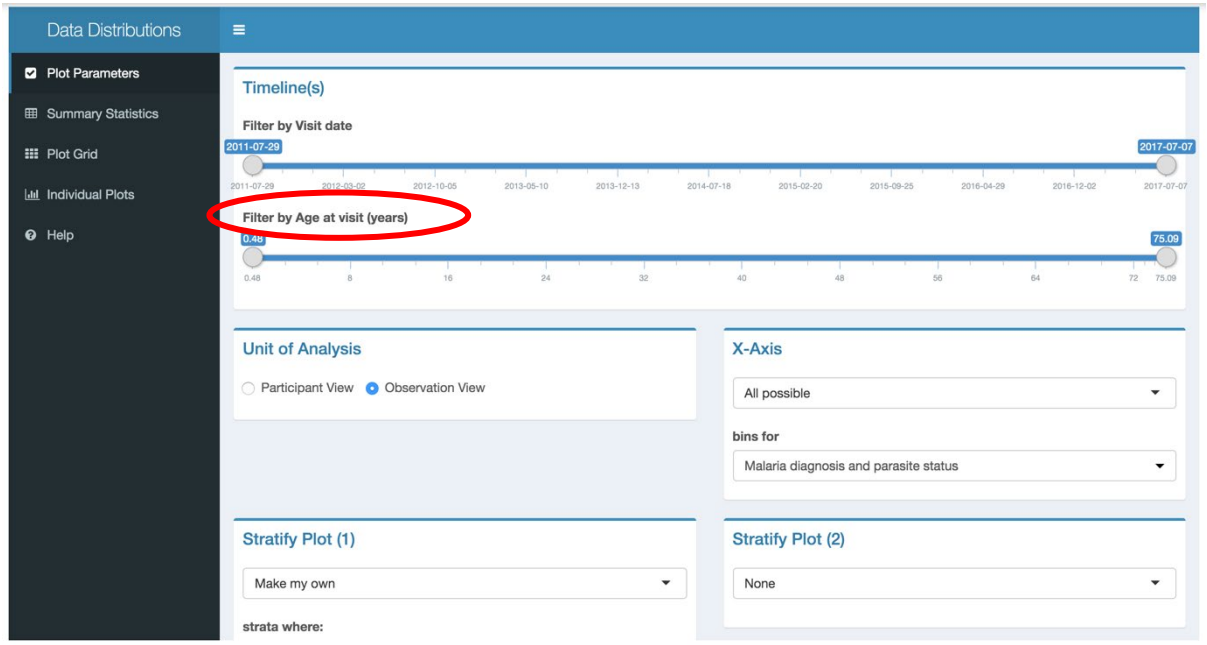
1. Click on the "Analyze Results" button from the results page



2. Select the "Distributions" analysis tool



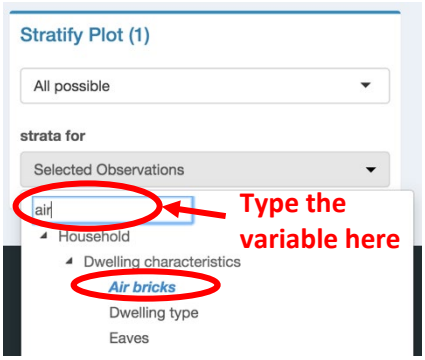
3. When you open a Shiny App, it will have parameters you can edit. By default, the Shiny Apps work on all the data, not just the data you selected in the Search Wizard. However, you can choose to stratify your data based on your selected results. Here, we are interested in looking at how the Malaria diagnosis and parasite status varies for children based on the presence of air bricks in their home and which site they are from, which will require stratifying by 2 different variables. Instead of using one stratification parameter for choosing only the selected results, we will use the Filter by Age at visit (years) parameter to select for children under 12.



4. Set the “Age at visit (years)” to select for only children 12 and under

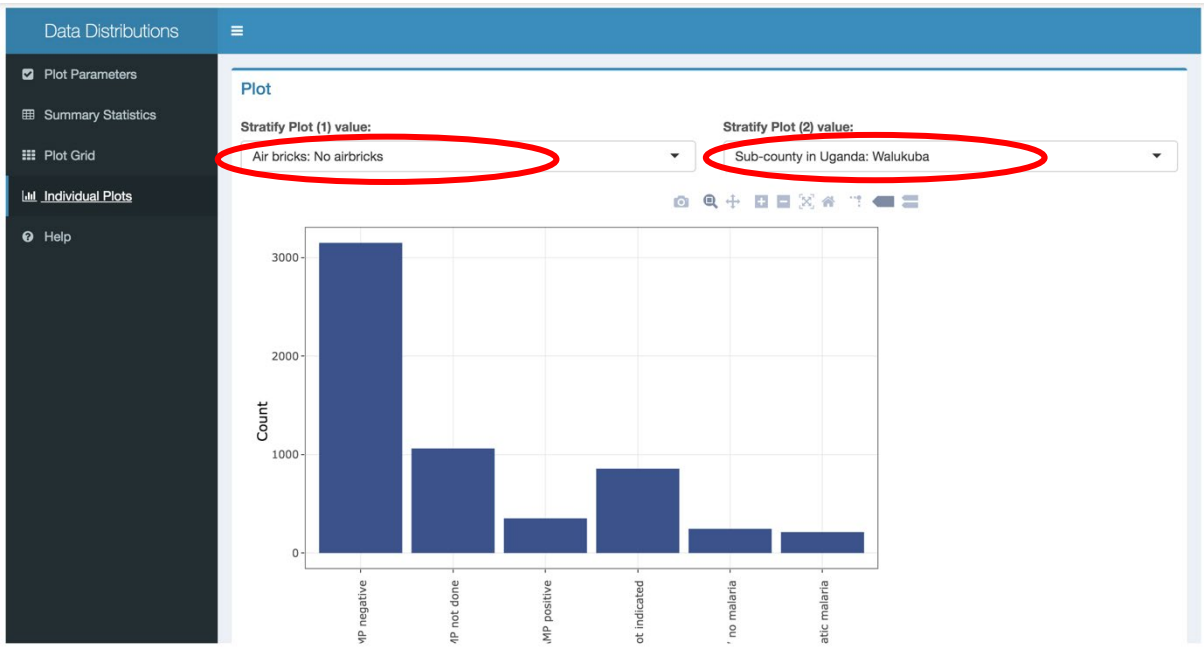
5. In this analysis, we want to see each observation of malaria diagnosis and parasite status tallied, so we will keep the Unit of Analysis on “Observation View” and the X-Axis as “All possible” bins for “Malaria diagnosis and parasite status”. *If we want to see how malaria diagnosis and parasite status varies based on air bricks, how will we stratify the data?*[9]

6. Set up Stratify Plot (1). Notice that as you change “Make my own” to “All possible”, the interface updates. When you click the drop-down menu to select the variable you see a filter tree of the possible variables. You can either click through the categories to find your variable or start typing it in the box at the top.



7. The PRISM study chose 3 sites with different transmission intensities. A relationship between house design and geographic region might therefore confound results, so let’s set up the second strata for “Sub-county in Uganda”.

8. Take a look at Individual Plots and use the drop-down menus at the top to select which distribution plot you want to look at. Hover over the bars to see the specific count. *How many observations of symptomatic malaria are associated with no airbricks in Nagongera? With unscreened airbricks?*[10] *What can you say about the use of airbricks across the different sites?*



By looking at the Plot Grid, we can see that the most common type of airbrick varies by site. This Shiny App lets us see how the data breaks down by different variables. Take a moment to try stratifying by other variables that you are interested in. Next, we’ll use the Contingency Table app to generate 2 x 2 contingency tables and calculate some basic statistics.

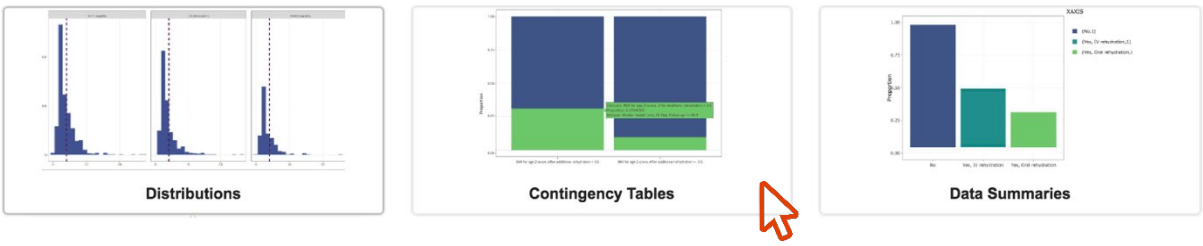
[9] Let’s look at “All possible” strata for “Air bricks”

[10] 2,159 observations of symptomatic malaria with no airbricks in Nagongera, 759 observations of symptomatic malaria with unscreened airbricks in Nagongera

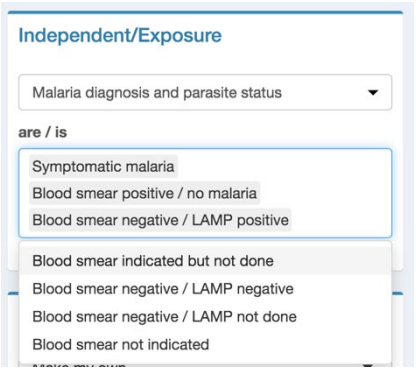
Contingency Tables – What's the risk of infection for a child living in a modern vs. traditional dwelling?

The contingency tables app allows you to look at the association between two variables and calculates a p-value, odds ratio (OR), and risk ratio (RR) with 95% confidence intervals.

1. Click on the "Analyze Results" button and select the Contingency Tables app.



2. We want to look at the relative risk of infection for a child living in modern vs. traditional dwellings and use the variable "Malaria diagnosis and parasite status". However, a 2x2 contingency table only allows two values per variable, and there are more than two potential values for "Malaria diagnosis and parasite status". Can we still use this variable?



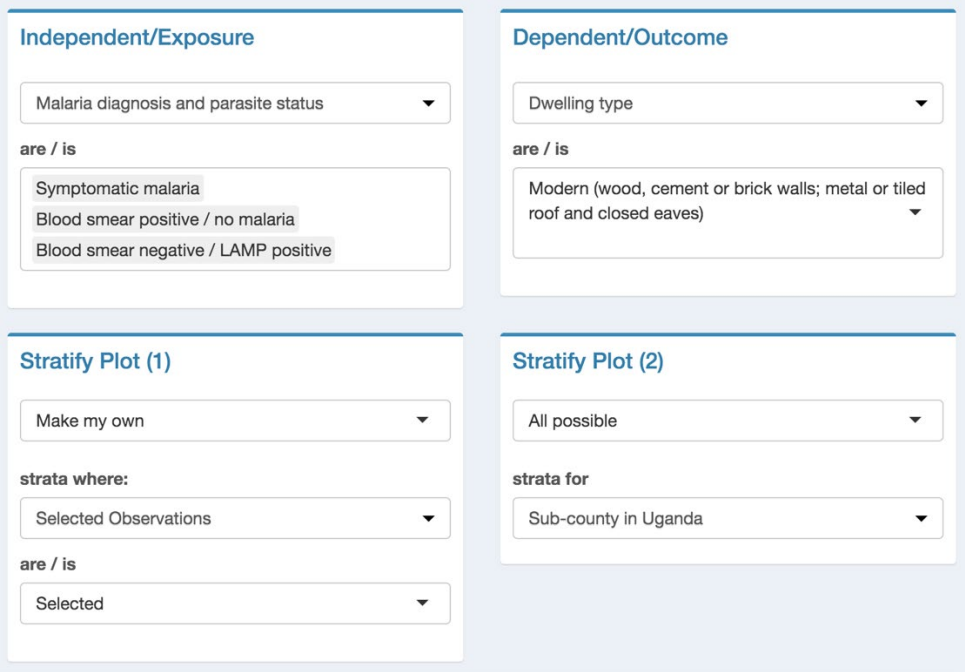
Yes, we can still use the variable, but we have to group the values. Here, we are interested in all observations where a child was found to be infected compared to all observations where the child was not found to be infected. We will therefore select "Symptomatic malaria", "Blood smear positive/no malaria", and "Blood smear negative/LAMP positive" from the drop-down menu. The remaining values ("Blood smear indicated but not done", "Blood smear negative/LAMP negative", "Blood smear negative/LAMP not done", and "Blood smear not indication") will automatically become the second group.

3. The PRISM cohort study derived the variable of "Dwelling type" based on the wall and roof material and eaves. Let's select this as our second variable and look at modern homes.

4. Stratify Plot (1) is set to stratify based on "Selected Observations" by default. *If we didn't stratify the data this way, how might we be sure to only use the same data as what we selected in the Search Wizard? Why won't this always work?*[11]

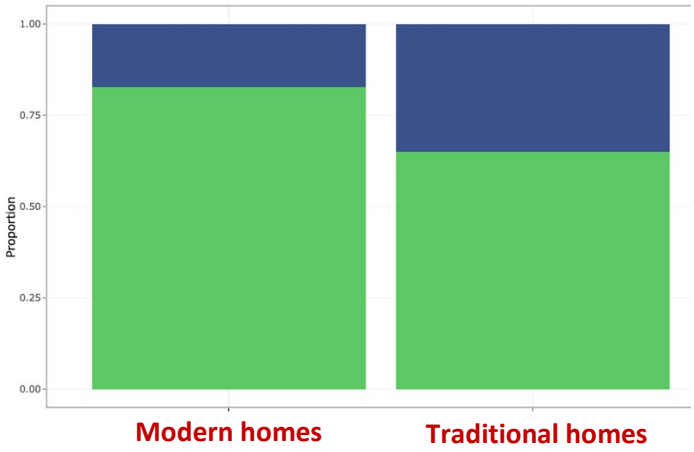
5. Set Stratify Plot (2) to stratify based on Sub-County in Uganda. Your selections should look like this:

[11] Since we only selected for children in the Search Wizard, we could use the "Filter by Age at visit (years)" parameter at the top of the Contingency Tables app. However, if we had applied multiple selection criteria, we would have to stratify based on "Selected Observations" in order to restrict our analysis to just the data of interest.



6. Look at the "Summary Statistics". *Is a modern home protective against Plasmodium infection? What's the odds ratio and confidence interval for a child in Walukuba getting infected if they live in a modern home compared to a traditional home?*[12]

7. Check out the plots as well. In this plot of Kihihi, blue shows observations of infection, while green shows observations without infection. You can see there are significantly fewer infections (blue) for children living in modern homes compared to traditional homes.



8. What other variables are you interested in? Take a look at some of them.

Do you think house design alters children's risk of *Plasmodium* infection? What other variables would you want to control for? Are there any intervention recommendations you would make based on this data? Why or why not?

[12] A modern home is protective against *Plasmodium* infection. The odds ratio is 0.5099, with a 95% confidence interval from 0.4281-0.6074.

Exercise 2: How does poverty impact the use of bednets in Rourkela and Chennai? (40 min)

In this exercise we will use the India ICEMR longitudinal cohort study to determine if there is an association between poverty and bednet usage. You will access the data and explore it using the participant search. You will also use the Shiny Apps to visualize the data while trying to answer the question of whether poverty is associated with bednet usage.

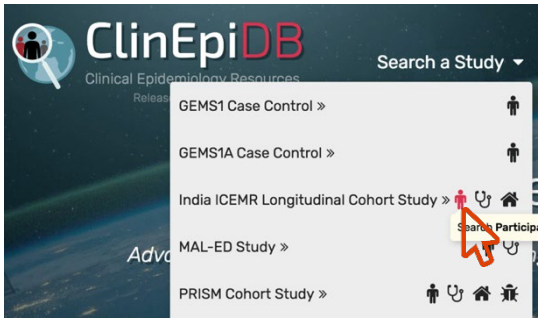
We want to determine how poverty relates to bednet usage in the India ICEMR cohort study. This study collected data from all participants every three months, but questions about bednet usage were only asked once, at the beginning of the study. *What type of search do you think we want to do based on this information? Why?*[13]

Finding and Examining Data Variables – Determining what variables were collected

1. Start a Participant search by clicking the 'Participant' icon on either the study card or the drop-down menu under "Search a Study".

Study card

Drop-down menu

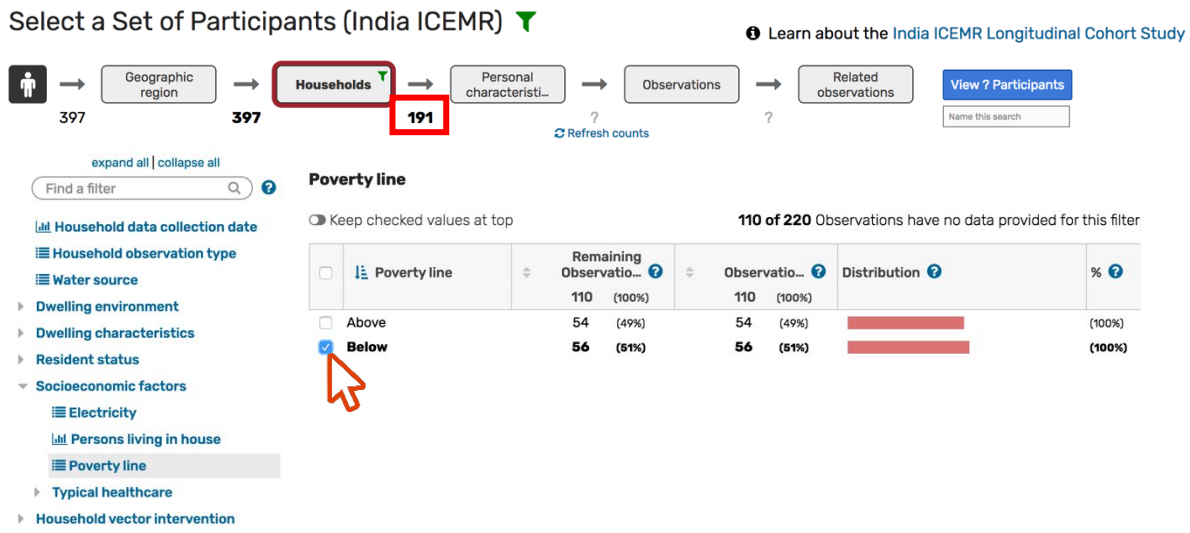


2. We are interested in looking at how poverty relates to bednet usage. Spend some time looking at what variables were collected. To do this, click on the different grey buttons in the search wizard, then click "expand all" to see all variables within that step. You can also start typing into the "Find a filter" box. Try typing "poverty line" and click on the result. *What percent of observations are recorded as "Above" the poverty line?* [14]

3. Keep looking through the variables available under different steps. Search for "bednet" under the Household step. Then search for bednet under the Observations step. *How are these variables different?* [15]

4. Go back to the Households step and find the "Poverty line" variable. Select "Below".

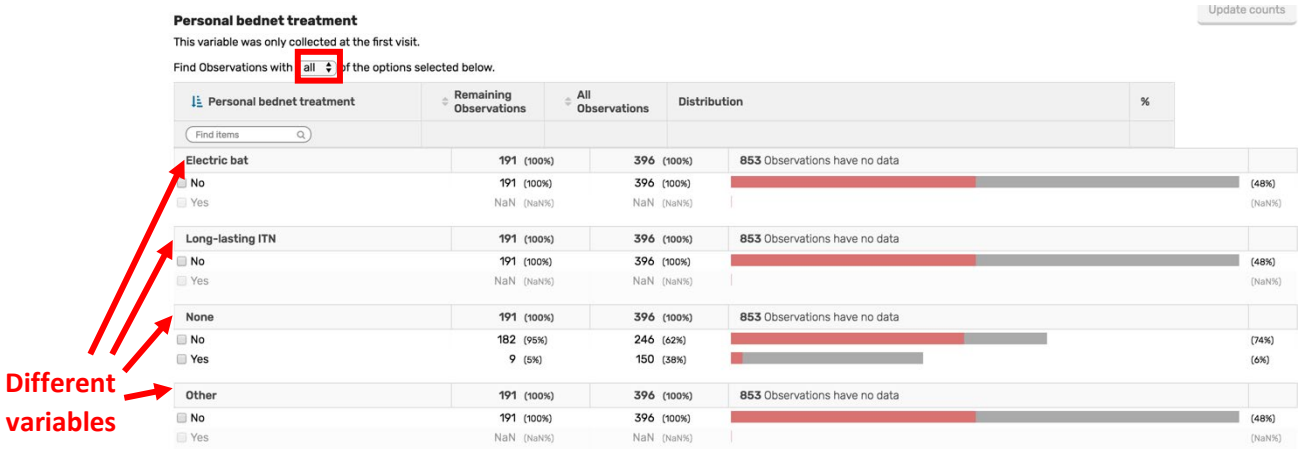
[13] A Participant search will return one row of data per participant and allow us to capture data from every participant on bednet usage.
[14] 49% of observations are above the poverty line.
[15] The variables under Household collect the number of each type of bednet that belong to a household. The variables under Observations collect what type of bednet an individual participant reported using at the first visit.



Notice that the numbers in the Search Wizard have updated based on the data we selected. *How many participants are now selected? Why doesn't the number match what you see on the poverty line table?* [16]

5. Now let's look at the impact of a household being below the poverty line on which type of bednet a participant uses. *Which variable do we want to look at and what step is it under?* [17]

6. Go to the Observations step and find the variable "Personal bednet treatment". This type of variable is called a "multifilter" since it includes multiple variables (Electric bat, Long-lasting ITN, etc.) that all appear on one screen. The parameter box at the top allows you to choose whether you want to find participants with "ALL" the values you select below or "ANY" of the values you select below. *How could you identify participants who reported that they did not use an untreated net or a regular ITN that was treated more than 6 months ago?* [18]

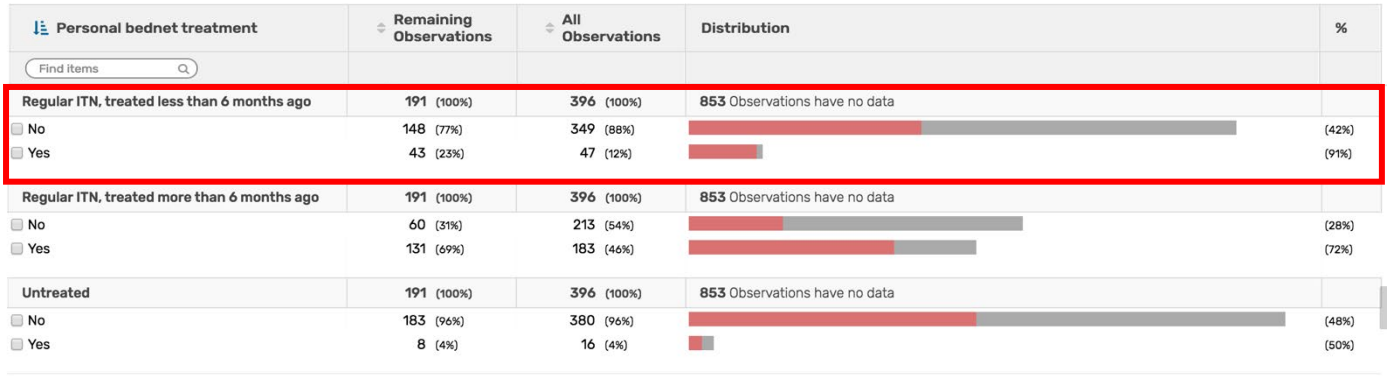


[16] 191 participants are selected. The numbers in the Search Wizard are always based on the type of search conducted. In this case, we did a participant search, so it always shows participant numbers. The poverty line table provides numbers based on the number of household observations. Since multiple participants can belong to one household, you can select 56 household observations where the household falls below the poverty line, which corresponds to 191 participants who belong to those households.

[17] Personal bednet treatment under the Observations step. This variable was collected only at the first visit and records which types of bednets each participant uses.

[18] Select "All" for the top parameter, then select no for "Untreated" and no for "Regular ITN, treated more than 6 months ago".

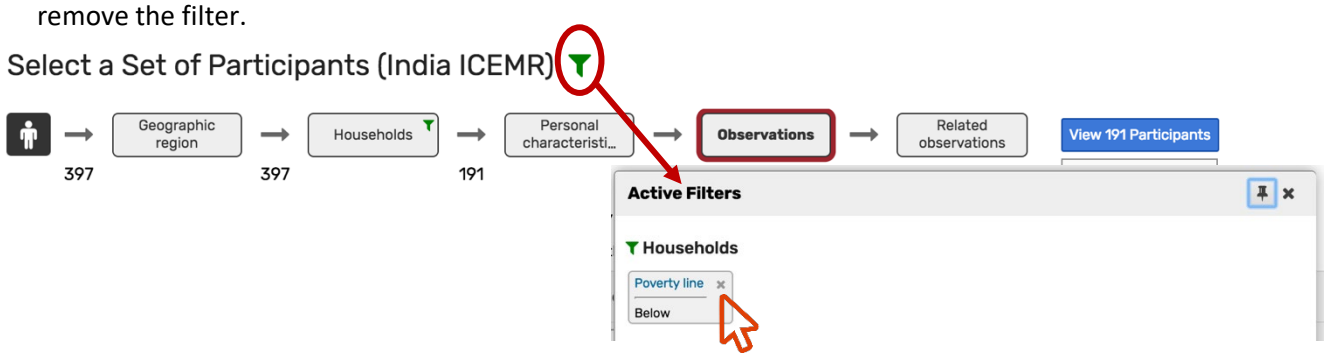
7. Note that the columns “Remaining observations” and “Observations” show different numbers. “Observations” indicates the total data, while “Remaining observations” indicates the selected data, or participants from households below the poverty line. In the Distribution graphic, grey indicates the data that does not meet the selection criteria, while pink reflects the selected data. The percentage on the far right tells you what percent of participants meet your selection criteria, or in this case, what percent of participants came from households below the poverty line. *What percent of participants who reported using a regular ITN, treated less than 6 months ago came from households below the poverty line?* [19]



Using the Shiny Apps to visualize data – Is a participant below the poverty line more likely to use a net?

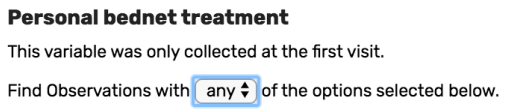
In the last section, we noticed that 91% of participants who used a regular ITN that was treated less than 6 months ago were also below the poverty line. On the same page we could also see that 72% of participants who used a regular ITN treated more than 6 months ago were below the poverty line, and that 50% of participants who used untreated nets were below the poverty line. We hypothesize that participants below the poverty line are more likely to use a net than participants above the poverty line. We can test this hypothesis using one of the shiny apps, but first we need to reconfigure what data is selected.

1. To remove specific data selections, click on the green filter icon at the top of the Search Wizard. This will activate a pop-up that lists out all of the active filters you have. Now click the ‘x’ to remove the filter.

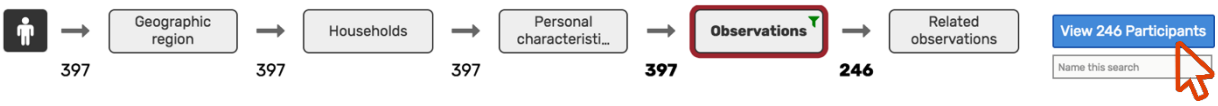


[19] 91%. To find the answer, look at the row where data for “Regular ITN, treated less than 6 months” is “Yes”. We can see that there are 43 Remaining observations (below poverty line) out of 47 total observations. If we look all the way to the right, we can see the percent written for us so we don’t have to calculate it.

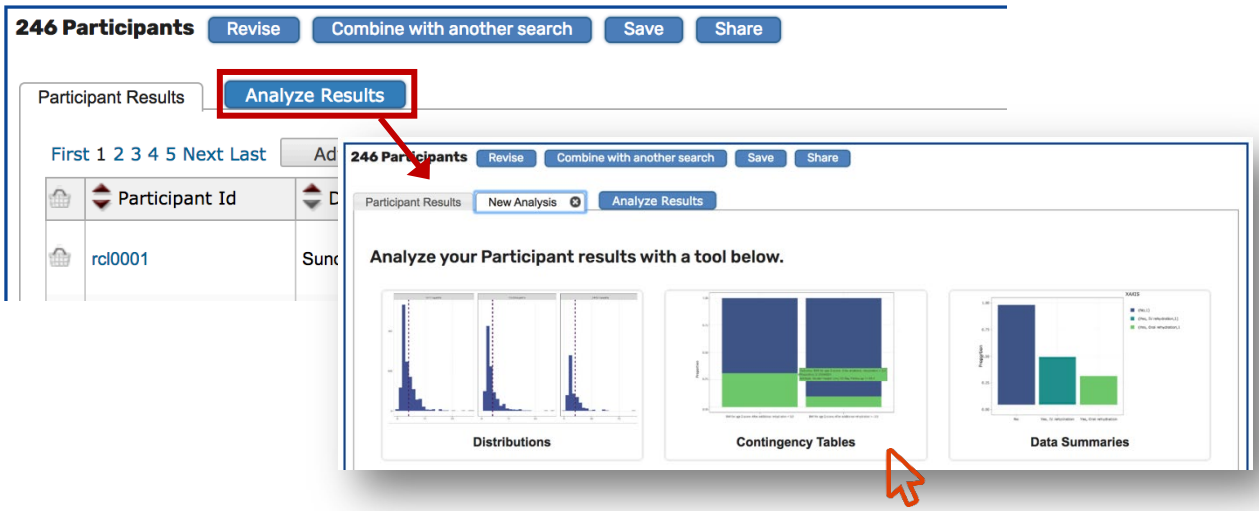
2. Now select participants who use an untreated net or regular ITN. Make sure to set the top parameter to “any”:



3. The blue button at the end of the Search Wizard should now say “View 246 Participants”. Click on that button to go to the results page.



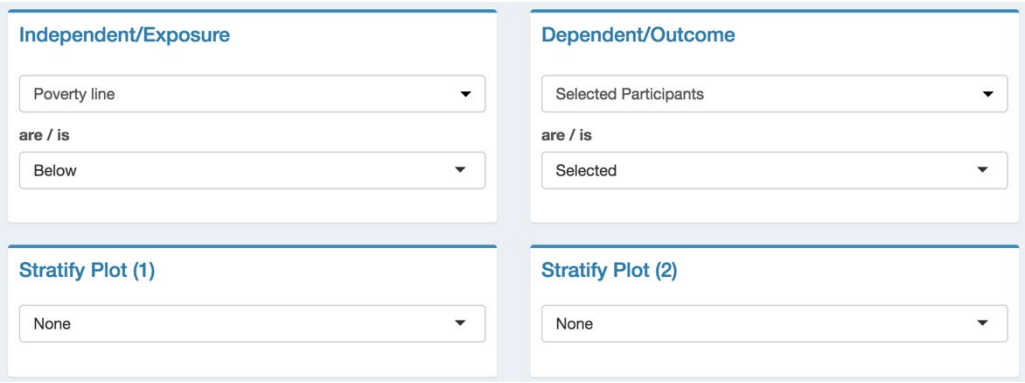
4. To access the Shiny Apps, click on “Analyze Results”. We want to make a contingency table to look at the association between use of any bednet and poverty, so click on “Contingency Tables”



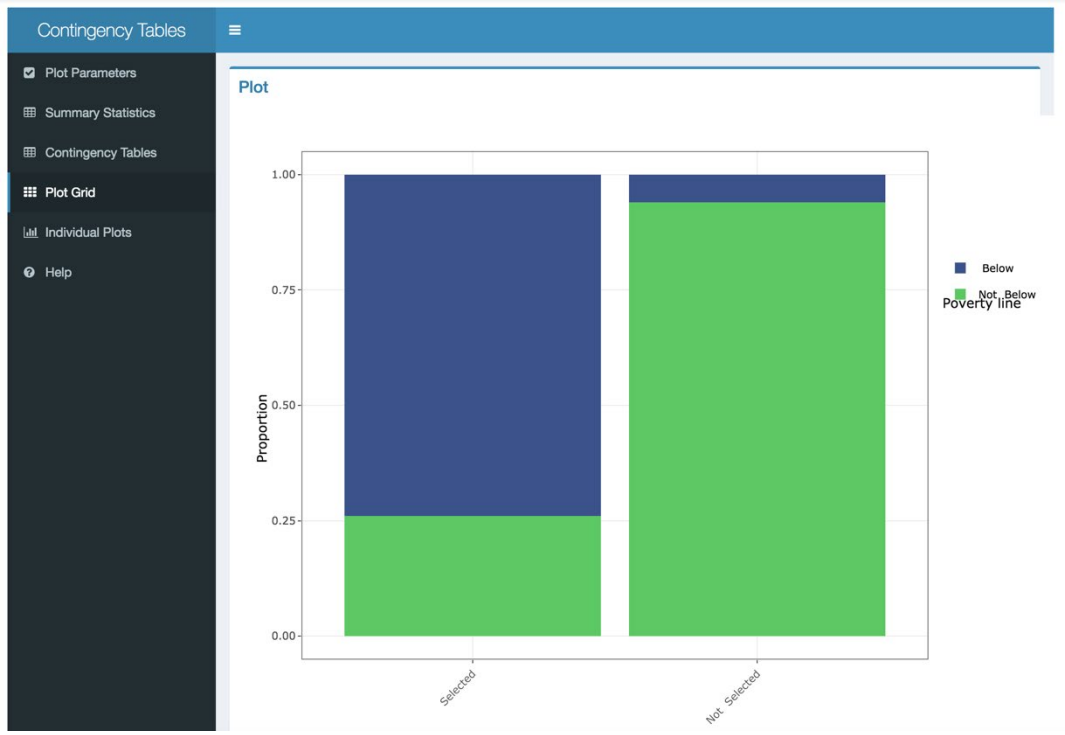
5. By default, the Shiny Apps work on all the data, not just the data you selected in the Search Wizard. However, you can choose to use your selected data as a variable which compares the data you selected (participants who report using a bednet) to the rest of the data (participants who do not report using a bednet). *Which variable will you enter as your independent/exposure variable? Which variable will you enter as your dependent/outcome variable?* [20]

6. Set up your parameters as shown below. Click on the arrow to open the drop-down menu. You can either click through all of the variables by clicking on the arrows next to the labels or type the name of the variable in the box at the top.

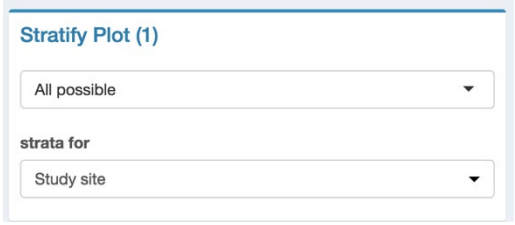
[20] For this example, let’s use “Poverty line” as the exposure variable and “Selected participants” as the outcome variable. Since we’ve selected participants who use bednets, we are really setting the outcome variable to bednet users.



7. Explore the different menu options on the left-hand side. Look at the Plot Grid, then at the Summary Statistics. *What is the p-value for this analysis? What factors might you want to control for?* [21]

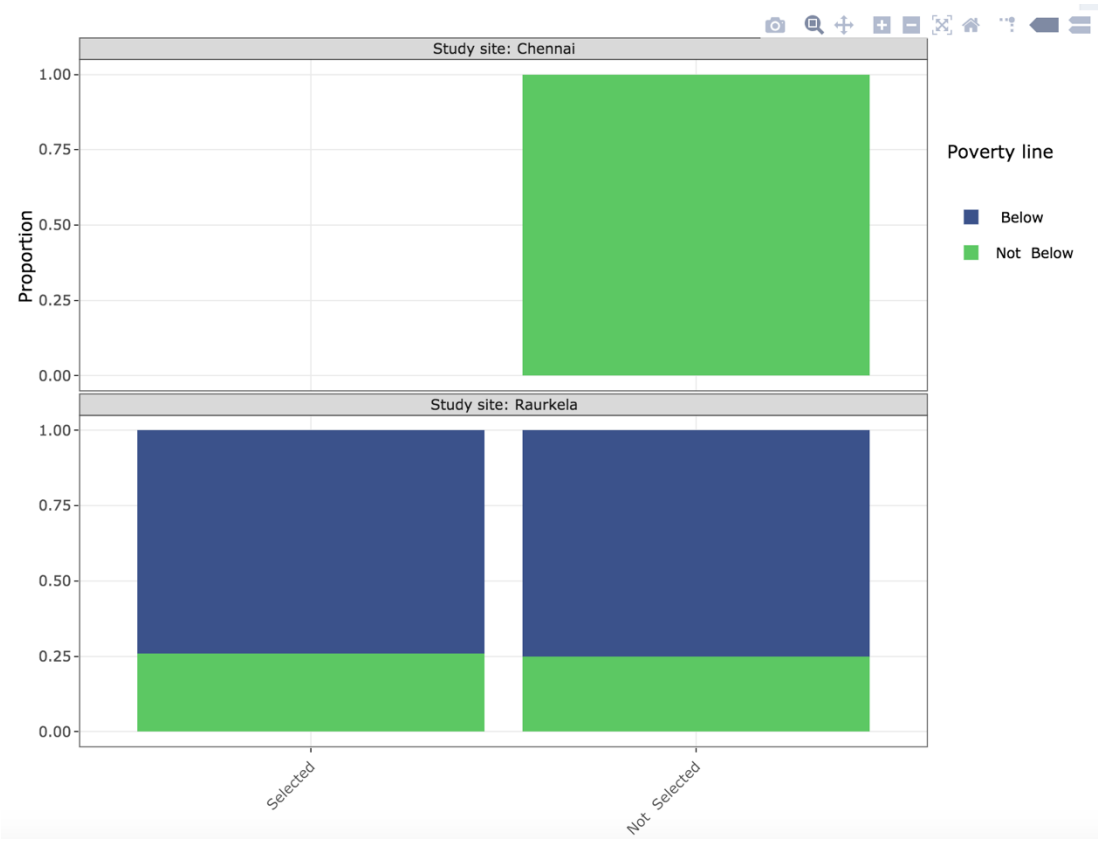


8. The cohort study was conducted in Rourkela and Chennai. A relationship between poverty and geographic region or bednet usage and geographic region might therefore confound results, so let's stratify our results based on study site. Click back to "Plot parameters" and update "Stratify Plot (1)".



9. Now look at the Plot grid again. We can see that the association we saw previously was completely dependent on geographic region. What other potential confounders do you want to explore?

[21] $p < 0.0001$. One obvious potential confounder is geographic region.



An important warning about these applications; they are intended for data exploration only! You may choose to examine any number of variables or relationships between variables using these apps but all results should be considered hypothesis generating. More complete, rigorous statistical analyses with proper consideration of study design, possible sources of bias and confounders is strongly recommended before any analyses are considered complete.

You have completed the exercise for this section! Great job! Please let us know if you have any questions or comments via the 'Contact Us' link located in the bottom right corner of any of our web pages. You can also email us at help@clinepidb.org.