Exercise: Exploratory data analysis on ClinEpiDB
Urinary schistosomiasis in Mozambique

In this exercise you will perform a step-by-step **exploratory data analysis** on the ClinEpiDB platform to explore trends in urinary schistosomiasis in the **SCORE Mozambique *S. haematobium* Cluster Randomized Trial dataset.**

Step 1: Choose a question

Open a **SCORE Mozambique *S. haematobium* Cluster Randomized Trial analysis**. The **View Study Details** tab describes this cluster randomized trial conducted in urinary schistosomiasis-endemic areas of Mozambique. Mass drug administration regimens (community-wide or school-based, with or without drug holidays) against *Schistosoma haematobium* were compared over a 5 year period. The target population was 9-12 year old children, who represent the highest risk group.

Possible questions to explore in this dataset include *(choose one)*

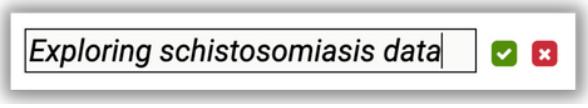
Q1: Is community-wide treatment more effective than school-based treatment in reducing prevalence of schistosomiasis over time?

Q2: Is there an association between sex and urinary schistosomiasis in this study population?

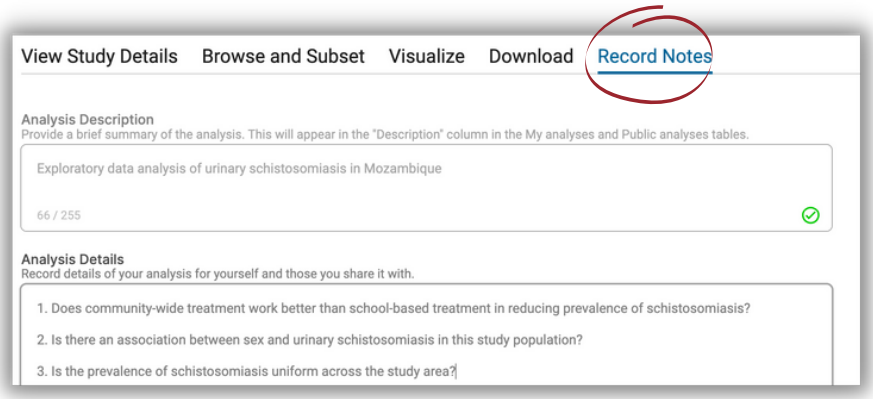
Q3: Is the prevalence of schistosomiasis uniform across the study area?

Step 2: Name and plan your analysis

Give your analysis a name at the top of the page. It may look something like this.



Use the **Notes** tab to plan the analysis and write notes that will be saved along with the analysis.



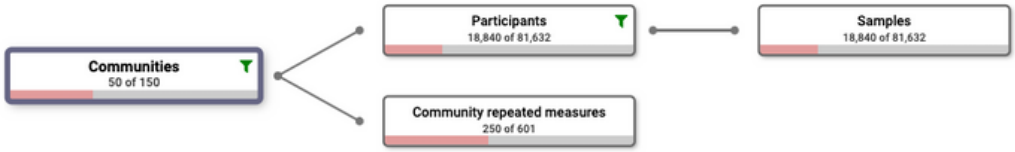
Step 3: Choose an appropriate subset of data

In the **Browse and Subset** tab, restrict the analysis to participants in the highest risk age group (**9-12 years**) and study arms which received either community-wide treatment each year (notation **cccc**), or school-based treatment each year for 4 years (**ssss**).

Which two variables would you subset?

a. _____ b. _____

c. How many participants are present in your subset? _____



Step 4: Identify variables of interest for this analysis

Browse or search the variable tree on the left and identify variables that will be of interest to your analysis. Look at the distribution of each variable and note whether it is numeric or categorical. This will help you decide what visualization tools to use in the next step. Star variables of interest to make them easier to access later. **Check out the variables below, which will be relevant based on the exploration question(s) you chose on the first page of this exercise**.

Q1, Q2, Q3: ***Schistosoma haematobium, by microscopy:*** This is the main outcome variable. It is a categorical variable with two levels - positive and negative - indicating presence or absence of *S. haematobium* infection in a urine sample taken from the participant. Look for this variable under **Sample** in the variable tree.

Q1: ***Study timepoint:*** This variable is categorical with 5 levels, one for each study year. It will be important for exploring schistosomiasis prevalence over the time period of the study. It is a featured variable.

Q1: ***Village study arm***: This variable is categorical with 6 levels, one for each of the study arms. It will allow us to visualize whether the study arms differ in how their prevalence changes over time. It is a featured variable.

Q2: ***Sex***: This variable will allow us to explore if schistosomiasis prevalence differs by sex. It's found under Participant > Demographics.

a. Is ***Sex*** a categorical or numeric variable? _____

b. How many levels does it have in this dataset? _____

Step 5: Plan visualizations to examine associations between variables

Make a list of the questions you would like to plot. What variables do you want to plot on the X-axis and, if applicable, on the Y-axis? Which of the plots in the **Visualize** tab would be appropriate for these variables? Fill in this table for the question you chose.

Question	X-axis/Main variable	Y-axis variable	Plot
Q1: Schisto prevalence over time, by study arm			
Q2: Association of sex and schistosomiasis			
Q3: Mapping schisto prevalence			

Click on the **Visualize** tab, then on **new visualization**, and select the appropriate tool and make the plot(s). Name each plot.

Step 6: Interpret the plots

What does the data say about the question you asked?

To find your analysis again later, open the **Workspace** menu in the header and click **My analyses**. You should then see a table that includes this analysis of the **SCORE Mozambique S. haematobium Cluster Randomized Trial**.

Turn to the next page for answers to this exercise!

ANSWERS

Step 3: Which two variables would you subset? a. **Age group** b. **Village study arm**

c. How many participants are present in your subset? **18,840**

Step 4: a. Is Sex a categorical or numeric variable? **Categorical**

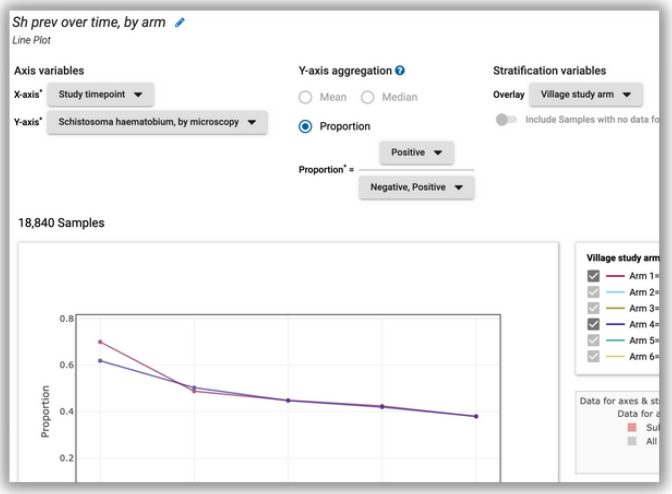
b. How many levels does it have in this dataset? **3 (Male, Female, Unknown)**

Step 5: Fill in this table for the question you chose. Your answers may look a bit different

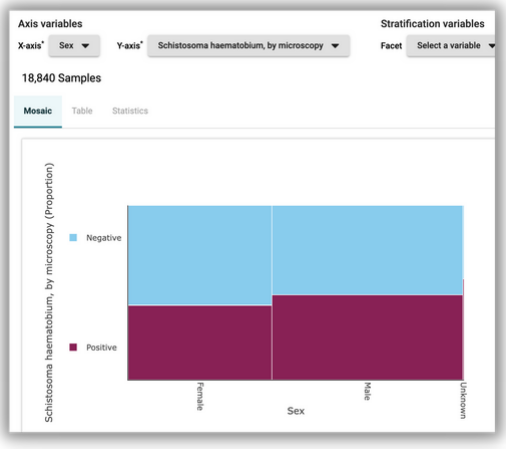
Question	X-axis/Main variable	Y-axis variable	Plot
Q1: Schisto prevalence over time, by study arm	Study timepoint	Schistosoma haematobium, by microscopy	Line Plot
Q2: Association of sex and schistosomiasis	Sex	Schistosoma haematobium, by microscopy	Mosaic Plot, RxC Table
Q3: Mapping schisto prevalence	Schistosoma haematobium, by microscopy	N/A	Geolocation Map

Your plots may look like this:

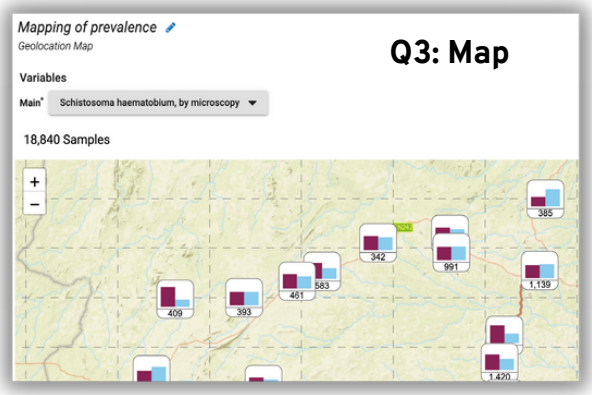
Q1: Line plot



Q2: Mosaic Plot, RxC table



Q3: Map



Step 6: Possible interpretations of the data

Q1: **Does community-wide treatment work better than school-based treatment in reducing prevalence of schistosomiasis over time?** When we look at the line plot, it appears that the community-wide treatment arm (Arm 1=cccc, in dark pink) started at about 70% prevalence in Year 1 while the school-based treatment arm (Arm 4=ssss, in dark blue) started a bit lower at ~62% prevalence. But at the end of the study, in Year 5, communities in both study arms are at 38% prevalence. Thus there does not appear to be a major difference between the two study arms. Furthermore, 4 rounds of MDA did not bring the *S. haematobium* prevalence down to zero or even close. To continue exploring, it may be worth looking at variable such as *Total population treated in village (%)* and *School-age children treated in village (%)* to see if treatment coverage was adequate. One may speculate that interventions other than MDA may be needed to further control schistosomiasis in this area of Mozambique.

Q2: **Is there an association between sex and urinary schistosomiasis in this study population?** Visually, males in this dataset appear to have slightly higher prevalence of schistosomiasis (49%) compared to females (43%). The statistics tab shows the result of a chi-square test of independence to examine the relation between sex and presence of *S. haematobium* infection. The relation between these variables was statistically significant, chi-squared (2, N = 18,840) = 68.01, p < 0. 001. But keep in mind that chi-square calculations are extremely sensitive to sample size; for a large sample size like this, almost any small difference will appear statistically significant.

Q3: **Is the prevalence of schistosomiasis uniform across the study area?** The map tool allows you to plot *S. haematobium* prevalence on a map of the study area. The map shows that even in communities that are separated by a short distance, there can be wide variation in the prevalence. It illustrates the focal nature of schistosomiasis, which can benefit from higher-resolution precision mapping.

Thank you for completing this exercise on performing an exploratory data analysis!