Exercise: Exploratory data analysis on ClinEpiDB
Malaria in Uganda

In this exercise, we will explore data from the **PRISM ICEMR Cohort** from Uganda. We will ask questions about the prevalence of malaria and *Plasmodium* infection in the study population and the association between fever, age, and disease.

Question 1: What's the prevalence of malaria in this study cohort?

The PRISM ICEMR cohort study enrolled one adult caretaker from a household and all children within the household and followed them over time. Participants were encouraged to come into the clinic any time they became ill for an unscheduled visits (passive case detection), and they also had regularly scheduled visits (active case detection). We want to figure out what the burden of disease looks like in this population.

1. Find the variable ***Observation type.*** Select the values *Enrollment* and *Scheduled visit*
   a. Why do we want to exclude unscheduled visits (passive case detection) if we are trying to estimate the burden of malaria?
2. Star the variable ***Observation type*** to make it easy to find later ☆ → ★
3. Now search for the term *malaria diagnosis*

   There are two variables that seem to fit, ***Malaria diagnosis and parasite status*** and ***Malaria diagnosis****.* The first variable is a derived variable combining blood smear results, LAMP results, and whether or not the participant had a fever. The second variable ***Malaria diagnosis*** is *Yes* when the participant had a blood smear positive for *Plasmodium* as well as symptoms of disease (either a reported fever or elevated temperature).
   a. What percent of the time were participants found to have malaria at enrollment or scheduled visits?
   b. At all visits?

We now know that the prevalence of malaria is around 2% in this population. In the next section, we ask: What percent of participants were infected with *Plasmodium,* even if they didn't have symptoms of malaria? In other words, what is the prevalence of *Plasmodium* infection in this cohort?

Question 2: What is the prevalence of *Plasmodium* infection in this cohort?

1. Look at ***Malaria diagnosis and parasite status.*** Complete the table with what percent of the time participants in the subset were found to have:
2. Star both ***Malaria diagnosis*** variables for later

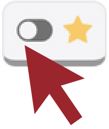
Symptomatic malaria	
Blood smear positive, no malaria	
Blood smear negative, LAMP positive	
Total	

3. We now know that the prevalence of Plasmodium infection was ~29% in this study. This study population is made up mainly of children. How might you expect the results to change if the population was mostly composed of adults? (*if you don't work on malaria, take a peek at the answer)*

Question 3: How are fever, age, and malaria related in this cohort?

Fever is a classic symptom not only of malaria, but many infectious diseases. Here, let's explore the association of temperature and age in malaria cases and then ask whether fever in this cohort is always associated with malaria.

1. Find and star the variables ***Age*** and ***Temperature***
2. Go to the **Visualize** tab, click to add a **+ New visualization**
   a. Both ***Age*** and ***Temperature*** are continuous variables. Which plot type will let us graph them against each other?
3. Click on **Scatter plot**
4. Open the **X-axis** drop down menu and turn on the toggle to shorten the list of variables to your starred variables



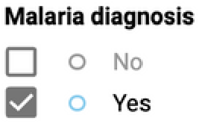
5. Set the X-axis to ***Age*** and Y-axis to ***Temperature***

a. What age range has more episodes of fever (temperature >38 C)? *(check one)*

□ **Children (0-10)** □ **Adults (>18)**

6. Add an **Overlay** with the ***Malaria diagnosis*** variable

7. In the legend for ***Malaria diagnosis***, uncheck the box next to the value *No* in order to better see temperature values when participants were diagnosed with malaria



Malaria diagnosis

□ ○ No

☑ ○ Yes

a. What is the highest temperature recorded in an adult (*hint: move your mouse over points on the plot)?* In a child?

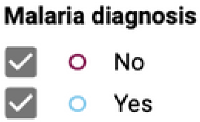
The plot suggests that symptomatic infection in children is more likely to be associated with very high temperatures than it is in adults. Now let's consider whether fever in this cohort is always associated with malaria. For this segment, we will focus on fever in children <= 10 years old.

8. Go to the **Browse and Subset** tab and subset on age <= 10 years

9. Go back to your scatter plot in the **Visualize** tab

10. Adjust the X-axis range to go from 0-10

11. Adjust the legend for ***Malaria diagnosis*** to look at both values *Yes* and *No*



Malaria diagnosis

☑ ○ No

☑ ○ Yes

a. Are there children with a temperature >38 C who don't have malaria?

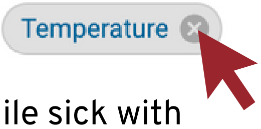
b. Are the number of episodes of non-malarial fever more or less than the number of episodes of malarial fever?

12. Confirm your estimate above by returning to the **Browse and Subset** tab and selecting all ***Temperatures*** >= 38. Then go to the **Visualize** tab and make a new **Bar plot** of ***Malaria diagnosis.***

a. What percent of measured fevers were non-malarial?

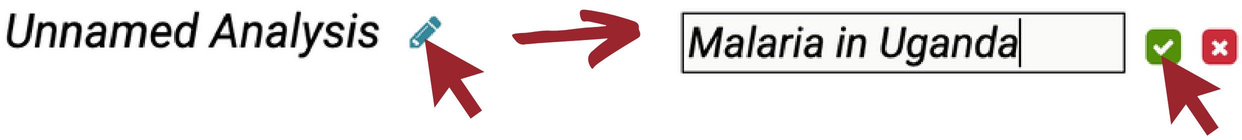
b. What subset of data are we looking at? *hint: click the* **Show all filters** *button and mouse over the variables to see how the data were filtered*

13. Remove the ***Temperature*** filter by clicking the **x** next to it



We see that while children tend to have higher temperatures than adults while sick with malaria, many cases of fever in children in this region cannot be attributed to *Plasmodium* infection.

14. Give your analysis a name so you can find it easily later



Bonus Question: Is there an association between house design and risk of *Plasmodium* infection in children?

The PRISM ICEMR Cohort study collected data on participants' homes, such as housing structure, access to water, etc. The variable ***Dwelling type*** indicates if a home is considered modern or traditional based on the wall and roof type, presence of eaves, and types of airbricks. We hypothesize that children who live in traditional houses are more likely to be exposed to mosquitos and therefore get infected with *Plasmodium* and develop malaria.

1. We want to look at the association between ***Dwelling type*** and ***Malaria diagnosis and parasite status***
    a. Is ***Dwelling type*** a categorical or continuous variable?
    b. Is ***Malaria diagnosis and parasite status*** categorical or continuous?
2. Go to the **Visualize** tab and click the **+ New visualization** button
    a. Which visualizations let you look at 2 categorical variables?
3. Open a **Mosaic plot (RxC table).** Set the X-axis to ***Dwelling type*** and the Y-axis to ***Malaria diagnosis and parasite status***
    a. Are there more *modern* or *traditional* homes in this cohort?
    b. Is there more *symptomatic malaria* (pink) in participants from *modern* or *traditional* homes?
    c. What about asymptomatic *Plasmodium* infection (yellow - *Blood smear positive, no malaria*, green- *Blood smear negative, LAMP positive*)?
4. Traditional dwellings seem to be correlated with higher prevalence of symptomatic malaria and asymptomatic *Plasmodium* infection.
    a. What potential confounders might you want to explore further?
    b. How can you start those explorations by modifying this plot?

Turn to the next page for answers to this exercise!

ANSWERS

Q1.1a. Why do we want to exclude unscheduled visits (passive case detection) if we are trying to estimate the burden of malaria? **Including unscheduled visits may introduce selection bias, since people who have malaria are more likely to come into the clinic unscheduled than people without malaria**

Q1.3a. What percent of the time were participants found to have malaria at enrollment or scheduled visits? **2%**

Q1.3b. At all visits? **12%**

Q2.1. Look at ***Malaria diagnosis and parasite status.*** What percent of the time were participants in the subset found to have:

Symptomatic malaria	**2%**
Blood smear positive, no malaria	**11%**
Blood smear negative, LAMP positive	**16%**
Total	**29%**

Q2.3. We now know that ~29% of participants were infected with *Plasmodium* at any given time. This study population is made up mainly of children. How might you expect the results to change if the population was mostly composed of adults? **Repeated exposure to *Plasmodium* provides some protection against developing malaria, so in an adult population in this area we might expect higher prevalence of microscopic (blood smear positive) or submicroscopic (LAMP positive) infection without malaria and lower prevalence of symptomatic malaria.**

Q3.2a. Both ***Age*** and ***Temperature*** are continuous variables. Which plot type will let us graph them against each other? **Scatter plot**

Q3.5a. What age range has more episodes of fever (temperature >38 C)? **Children (0-10)**

Q3.7a. What is the highest temperature recorded in an adult (hint: mouse over points on the plot)? **39.9 C** In a child? **41 C**

Q3.11a. Are there children with a temperature >38 C who don't have malaria? **Yes**

Q3.11b. Are the number of episodes of non-malarial fever more or less than the number of episodes of malarial fever? **Looks like less non-malarial fever than malarial fever**

Q3.12a. What percent of measured fevers were non-malarial? **41.76%**

Q3.12b. What subset of data are we looking at? *hint: click the **Show all filters** button and mouse over the variables to see how the data were filtered* **Observations where the participant was aged 0-10 years old, had a temperature >=38 C, and were seen at enrollment or a scheduled clinic visit (active case detection)**

QBonus.1a. Is ***Dwelling type*** a categorical or continuous variable? **Categorical**

QBonus.1b. Is ***Malaria diagnosis and parasite status*** categorical or continuous? **Categorical**

QBonus.2a. Which visualizations let you look at 2 categorical variables? **Bar plot (if you make one categorical variable an overlay or facet), Mosaic plot (2x2 table - must have 2 binary variables), Mosaic plot (RxC table)**

QBonus.3a. Are there more modern or traditional homes in this cohort? **Traditional**

QBonus.3b. Is there more symptomatic malaria (pink) in participants from modern or traditional homes? **Traditional**

QBonus.3c. What about asymptomatic Plasmodium infection? **Traditional**

QBonus.4a. What potential confounders might you want to explore further? **This study enrolled participants in different sites, so explore variables like *Sub-county in Uganda,* socioeconomic status-related variables, etc.**

QBonus.4b. How can you start those explorations by modifying this plot? **Add a facet! Try *Sub-county in Uganda***

Thank you for completing this exercise on exploratory data analysis on ClinEpiDB!