Demo of Exploratory Data Analysis on ClinEpiDB

This demo will cover creating a hypothesis-based exploratory analysis, including:

1. Exploring data
2. Applying filters to subset data of interest
3. Visualizing data
4. Adding notes to explain the analysis
5. Sharing the analysis with others

Let's say I work for the ministry of health, and I am interested in reducing malaria prevalence in my country. I am curious about whether bed nets with PBO can help reduce the burden of malaria in my country.

- View study details tab- learn more about the study
- ***Browse and subset tab***
    - Dataset diagram
        - Shape of the data
            - Longitudinal data but no repeated measures!
            - Individuals were not followed over time, at each timepoint a random sample of the population was surveyed
        - Sample size
    - **Explore what variables exist:**
        - Featured vars
            - Explain the 2 study arm variables
            - Explain why timepoint variable is both under households & participants
        - Variable tree- organized by categories, shows what kind of data is collected
        - Star variables
            - Cluster as treated, dichotomized
            - Cluster as treated
            - Age (years)
            - Plasmodium, by thick smear microscopy
            - Household study timepoint
            - Study timepoint
            - Household ITNs (a measure of ownership)
            - ITN last night (a measure of usage)
            - Click over to starred view
    - **Apply filters to create a subset of interest**
        - Since we are interested in an as-treated analysis, we want to REMOVE clusters of households that received a mixture of both PBO and non-PBO LLINs
            - Cluster as treated, dichotomized → select PBO LLINs and non-PBO LLINs (This removed 3 clusters and ~3000 participants)
            - Since we are interested in participants who underwent microscopy testing for Plasmodium, we want to REMOVE those who did not consent to testing, did not meet the study's inclusion criteria for testing, or who have missing data
                - Plasmodium, by thick smear microscopy → select No and Yes (This will remove samples from participants who were not tested)

        - **See how your filters have impacted the distribution of other variables**
            - Look at age in years → the distribution of ages in our subset now ranges from 2-10 years. This makes sense because the study documentation indicates that only kids aged 2-10 years were tested by microscopy to find Plasmodium parasites in their blood
- ***Visualize*** to ask simple questions, looking at variable associations
    - First look to see if there is any difference in uptake between the study arms
        - Line plot (call it LLIN ownership)
            - X-axis: *Household study timepoint*
            - Y-axis: *Household ITNs*
                - Set up the proportion to calculate prevalence: Yes/(Yes+No)
            - Notice that ownership goes up to >95% after the distribution campaign and stays high over the course of the study
            - Overlay: *Cluster as treated, dichotomized*
            - Notice that there is no difference in ownership between the 2 study arms
        - Duplicate Line plot (call it LLIN usage)
            - X-axis: *Household study timepoint*
            - Y-axis: *ITN last night*
            - Overlay: *Cluster as treated, dichotomized*
            - Notice that ownership goes up to >85% after the distribution campaign and stays high over the course of the study, and that there is no major difference in ownership between the 2 study arms
    - Since there is no difference in uptake between the study arms, it is appropriate to see if there is any difference in outcomes
        - Duplicate Line plot (call it parasite prevalence)
            - X-axis: *Study timepoint*
            - Y-axis: *plasmodium by microscopy*
            - Overlay: *Cluster as treated, dichotomized*
            - Notice that in both PBO-treated and non-PBO LLIN study arms, parasite prevalence decreased from baseline and remained lower through 18 months of study. But the changes from baseline for both arms were greater than the difference between study arms.
    - I'm curious to see if there is a difference between the 2 brands of PBO bednets used in the study
        - Duplicate Line plot (call it PBO LLIN brand comparison)
            - X-axis: *Study timepoint*
            - Y-axis: *plasmodium by microscopy*
            - Overlay: *Cluster as treated*
            - Click off the non-PBO LLINs (Olyset & PermaNet 2.0)
            - Olyset Plus PBO LLINs appear to be much more effective at preventing malaria than PermaNet 3.0 PBO LLINs. I may want to download the data to do a robust statistical analysis or design a new trial comparing these two types of LLINs directly
- ***Add notes***
- ***Share the analysis with others:*** get link and email to the workshop participants

With that, we now have an exercise so you can try to navigate ClinEpiDB for yourself, looking at a different malaria study.