Pathogens associated with diarrhea in the GEMS study

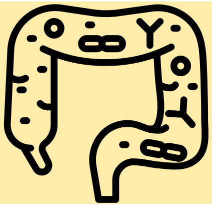
An exploratory data analysis exercise on ClinEpiDB

In this exercise you will perform a step-by-step **exploratory data analysis** on the ClinEpiDB platform to explore **pathogens associated with diarrhea in the GEMS1 Case Control study**.

Step 1: Read the study page and formulate a hypothesis

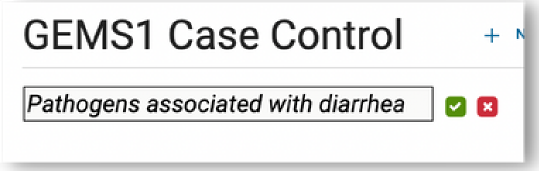
Go to the GEMS1 Case Control study page. Click on the **View study details** tab and read the summary and description of this observational study conducted in 7 sites in Asia and Africa. This is a case control study of children under 5 years of age where cases had moderate-to-severe diarrhea and controls were diarrhea-free. Stool samples from cases and controls were compared to identify the etiology (causes) of diarrhea.

Hypothesis: In infants under 1 year of age in Kenya, rotavirus, Cryptosporidium and Giardia infection are associated with moderate-to-severe diarrhea.



Step 2: Name and plan your analysis

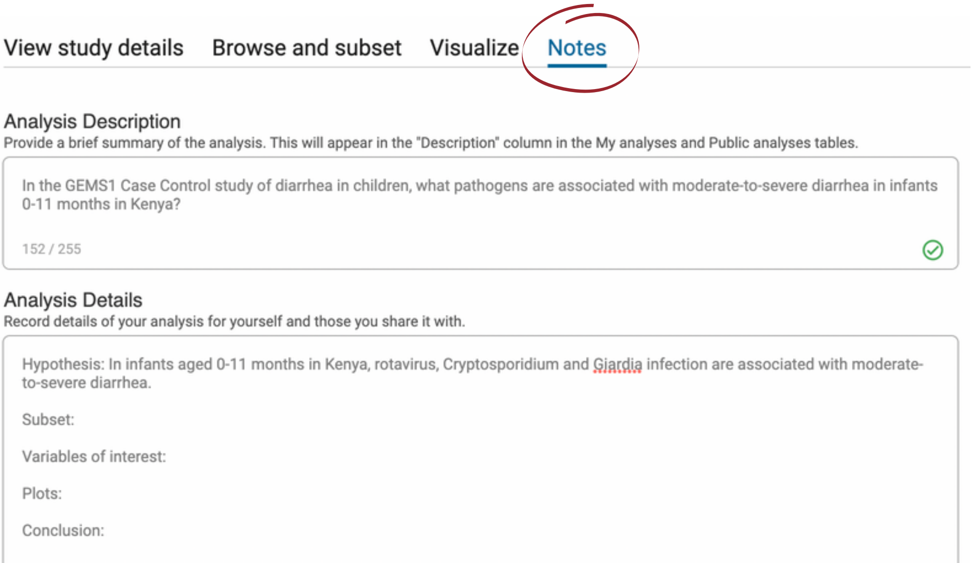
Give your analysis a name at the top of the page. It may look something like this.



GEMS1 Case Control

Pathogens associated with diarrhea

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



Analysis Description

Provide a brief summary of the analysis. This will appear in the "Description" column in the My analyses and Public analyses tables.

In the GEMS1 Case Control study of diarrhea in children, what pathogens are associated with moderate-to-severe diarrhea in infants 0-11 months in Kenya?

152 / 255

Analysis Details

Record details of your analysis for yourself and those you share it with.

Hypothesis: In infants aged 0-11 months in Kenya, rotavirus, Cryptosporidium and Giardia infection are associated with moderate-to-severe diarrhea.

Subset:

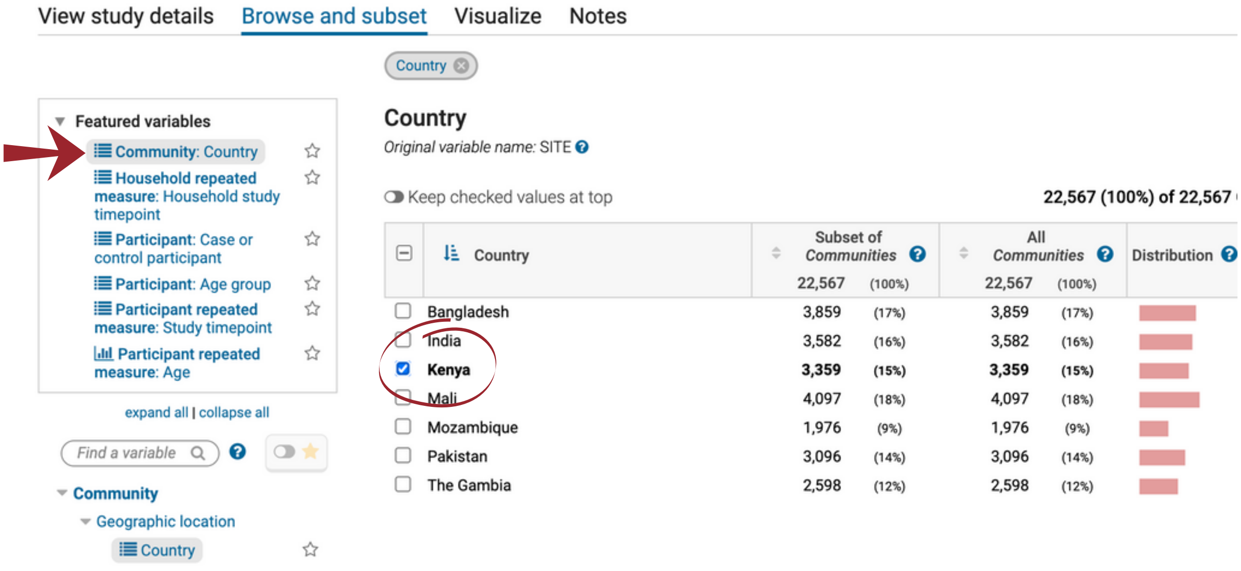
Variables of interest:

Plots:

Conclusion:

Step 3: Choose an appropriate subset of data

Click the **Browse and subset** tab. If you want to restrict your analysis to participants under 1 years of age, and to participants in Kenya, how would you choose an appropriate subset of data?



View study details | Browse and subset | Visualize | Notes

Country ⊗

▾ Featured variables
- Community: Country ☆
- Household repeated measure: Household study timepoint ☆
- Participant: Case or control participant ☆
- Participant: Age group ☆
- Participant repeated measure: Study timepoint ☆
- Participant repeated measure: Age ☆

expand all | collapse all

Find a variable

▾ Community
▾ Geographic location
Country ☆

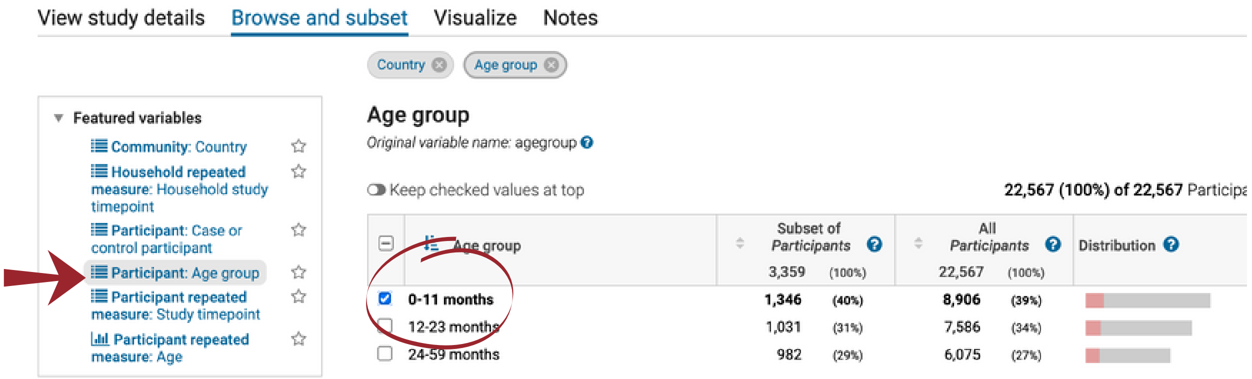
Country

Original variable name: SITE

Keep checked values at top

22,567 (100%) of 22,567

⊟	Country	Subset of *Communities*		All *Communities*		Distribution
		22,567	(100%)	22,567	(100%)	
☐	Bangladesh	3,859	(17%)	3,859	(17%)	
☐	India	3,582	(16%)	3,582	(16%)	
☑	**Kenya**	**3,359**	**(15%)**	**3,359**	**(15%)**	
☐	Mali	4,097	(18%)	4,097	(18%)	
☐	Mozambique	1,976	(9%)	1,976	(9%)	
☐	Pakistan	3,096	(14%)	3,096	(14%)	
☐	The Gambia	2,598	(12%)	2,598	(12%)	



View study details | Browse and subset | Visualize | Notes

Country ⊗ Age group ⊗

▾ Featured variables
- Community: Country ☆
- Household repeated measure: Household study timepoint ☆
- Participant: Case or control participant ☆
- Participant: Age group ☆
- Participant repeated measure: Study timepoint ☆
- Participant repeated measure: Age ☆

Age group

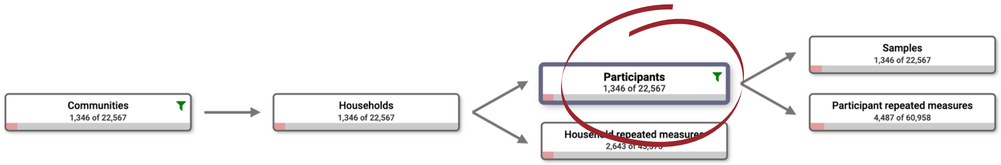
Original variable name: agegroup

Keep checked values at top

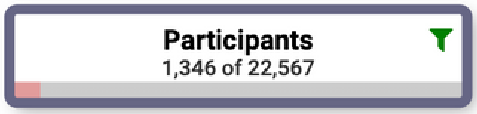
22,567 (100%) of 22,567 Participa

⊟	Age group	Subset of *Participants*		All *Participants*		Distribution
		3,359	(100%)	22,567	(100%)	
☑	**0-11 months**	**1,346**	**(40%)**	**8,906**	**(39%)**	
☐	12-23 months	1,031	(31%)	7,586	(34%)	
☐	24-59 months	982	(29%)	6,075	(27%)	

How many participants are present in your subset?

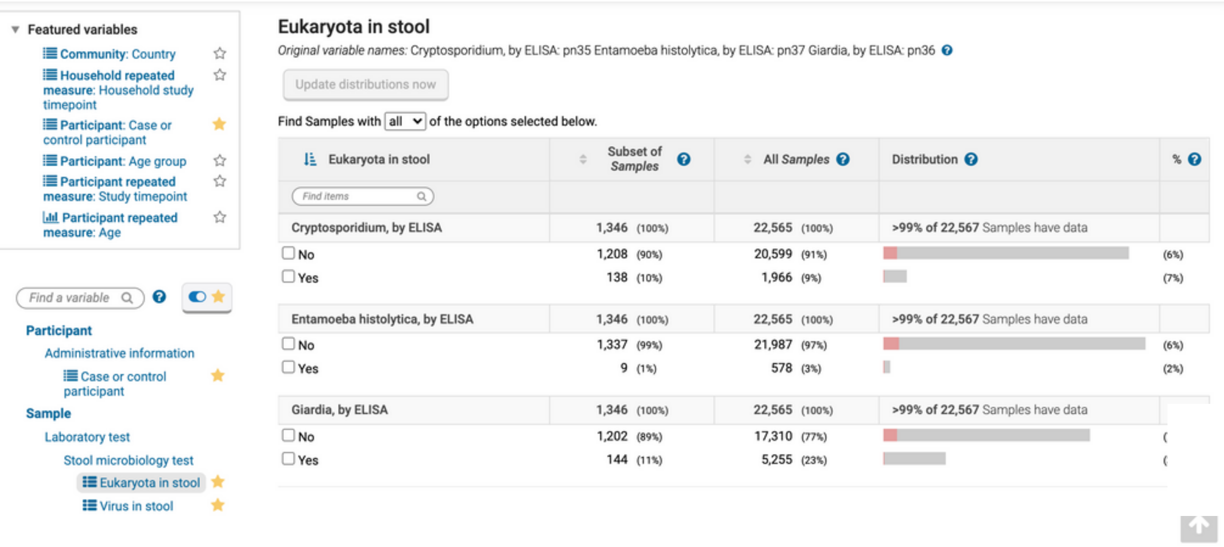


Looking at the dataset diagram at the top of the page will indicate that the subset includes **1346 participants** out of the 22,567 participants in the study.



Step 4: Identify variables of interest for this analysis

Browse or search through the variable tree on the left and identify variables that will be interest in your analysis. Look at the distribution of each variable and note whether it is numeric or categorical, as this will help you decide what visualization tools to use in the next step. Star the variables of interest to make them easier to access.


Featured variables
Community: Country ☆
Household repeated measure: Household study timepoint ☆
Participant: Case or control participant ★
Participant: Age group ☆
Participant repeated measure: Study timepoint ☆
Participant repeated measure: Age ☆
Find a variable
Participant
Administrative information
Case or control participant ★
Sample
Laboratory test
Stool microbiology test
Eukaryota in stool ★
Virus in stool ★
Update distributions now
Find items

Eukaryota in stool

Original variable names: Cryptosporidium, by ELISA: pn35 Entamoeba histolytica, by ELISA: pn37 Giardia, by ELISA: pn36

Find Samples with all of the options selected below.

Eukaryota in stool	Subset of *Samples*	All *Samples*	Distribution	%
Cryptosporidium, by ELISA	1,346 (100%)	22,565 (100%)	>99% of 22,567 Samples have data	
□ No	1,208 (90%)	20,599 (91%)		(6%)
□ Yes	138 (10%)	1,966 (9%)		(7%)
Entamoeba histolytica, by ELISA	1,346 (100%)	22,565 (100%)	>99% of 22,567 Samples have data	
□ No	1,337 (99%)	21,987 (97%)		(6%)
□ Yes	9 (1%)	578 (3%)		(2%)
Giardia, by ELISA	1,346 (100%)	22,565 (100%)	>99% of 22,567 Samples have data	
□ No	1,202 (89%)	17,310 (77%)		(
□ Yes	144 (11%)	5,255 (23%)		(

Case or Control status: We want to compare pathogen prevalence between cases and controls, so we will need the variable *Case or control participant*, a categorical variable.

Pathogens: Pathogens are detected by tests performed on samples (stool sample in the case of diarrhea), so look under the **Sample** category in the variable tree on the left. There are a number of pathogen test results under Sample > Laboratory test > Stool microbiology test. Two of the pathogens we are interested in, Cryptosporidium and Giardia, are Eukaryotes and the third, rotavirus, is a virus. The variables needed to test our hypothesis are *Cryptosporidium, by ELISA* and *Giardia, by ELISA* under **Eukaryota in stool** and *Rotavirus, by ELISA* under **Virus in stool** . They are all categorical variables.

Step 5: Create visualizations to examine associations between variables

Make a list of the associations you would like to plot. What variables do you want to plot on the X-axis and on the Y-axis? What sort of plot would be appropriate for these variables?

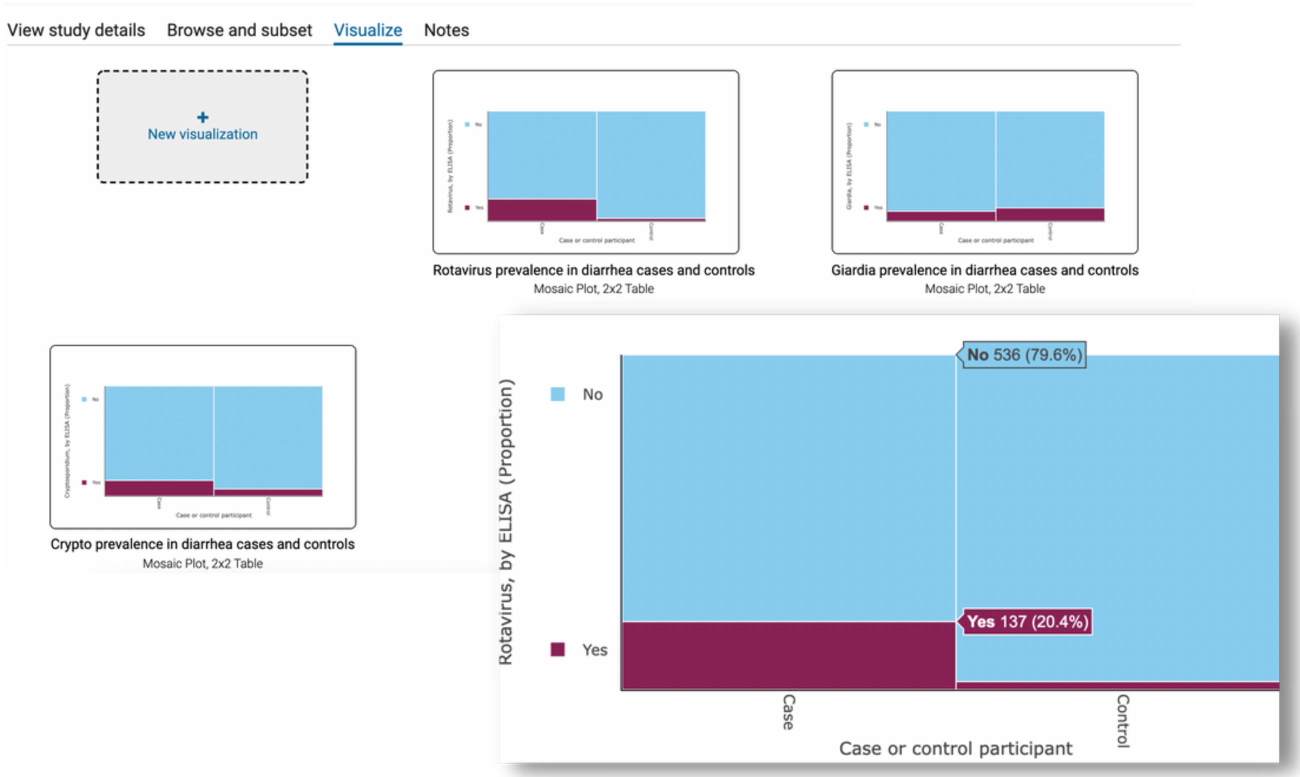
Association	X axis	Y axis	Plot
Rotavirus in cases and controls			
Cryptosporidium in cases and controls			
Giardia in cases and controls			

Your plan may look like this. Both the X-axis and Y-axis variables for each association are binary categorical variables, so a 2x2 table would be appropriate to explore this assocation.

Association	X axis	Y axis	Plot
Rotavirus in cases and controls	*Case or control participant* (categorical variable with 2 levels)	*Rotavirus, by ELISA* (categorical variable with 2 levels)	Mosaic Plot 2x2 Table
Cryptosporidium in cases and controls	*Case or control participant (categorical variable with 2 levels)*	*Cryptosporidium, by ELISA (categorical variable with 2 levels)*	Mosaic Plot 2x2 Table
Giardia in cases and controls	*Case or control participant (categorical variable with 2 levels)*	*Giardia, by ELISA (categorical variable with 2 levels)*	Mosaic Plot 2x2Table

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

Your plots may look like this:


View study details
Browse and subset
Visualize
Notes
+ New visualization
Rotavirus prevalence in diarrhea cases and controls
Mosaic Plot, 2x2 Table
Giardia prevalence in diarrhea cases and controls
Mosaic Plot, 2x2 Table
Crypto prevalence in diarrhea cases and controls
Mosaic Plot, 2x2 Table
Rotavirus, by ELISA (Proportion)
No
Yes
No 536 (79.6%)
Yes 137 (20.4%)
Case
Control
Case or control participant

Interpret the plots. What does the data say about your hypothesis?

The 2x2 mosaic plots indicate the following about participants in Kenya under 1 year of age-
a) **Rotavirus**: 20.4% of diarrhea cases have rotavirus as compared to 2.4% of controls, so rotavirus infection is associated with moderate-to-severe diarrhea, supporting our hypothesis.
b) **Cryptosporidium**: 14.1% of diarrhea cases have rotavirus as compared to 6.4% of controls, so Cryptosporidium infection is associated with moderate-to-severe diarrhea, supporting our hypothesis.
c) **Giardia**: 9.2% of diarrhea cases have rotavirus as compared to 12.2% of controls, so Giardia infection is NOT associated with moderate-to-severe diarrhea, contradicting our hypothesis.

This interpretation of our exploratory data analysis is supported by the published results of the GEMS1 study showing that rotavirus and Cryptosporidium are associated with moderate-to-severe diarrhea in children while Giardia is associated with asymptomatic colonization.

If you check the dropdown menu Workspace > My analyses in the header at the top of the page, you will see that this analysis automatically appears in the **My analyses** table.

Thank you for completing this exercise on performing an exploratory data analysis on clinepidb.org! Please contact **help@clinepidb.org** with feedback or questions.