



Effect of interventions on diarrhea in the WASH-B Kenya 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

Open a [WASH benefits Kenya analysis](#). Click on the **View study details** tab and read the summary. In this trial, clusters of households were randomized to one of several WASH (water, sanitation and hygiene) intervention arms or to control arms. One of the primary outcomes evaluated was diarrhea in target children.

Hypothesis: Target children belonging to intervention arms have lower incidence of diarrhea as compared to target children in control arms. Interventions can succeed only if there is sufficient uptake in the target population, so we also want to evaluate the uptake of the interventions.



Step 2: Name and plan your analysis

Give your analysis a name at the top of the page.

WASH Benefits Kenya Cluster Randomized Trial

Diarrhea in target childr ☒ ☐

The **Notes** tab provides a convenient space to plan the analysis and write notes that will be saved along with the analysis.

View study details Browse and subset Visualize **Notes**

Analysis Description

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

Comparing diarrhea incidence in target children in intervention and control arms

80 / 255



Analysis Details

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

Hypothesis: We hypothesize that target children belonging to intervention arms have lower incidence of diarrhea as compared to target children in control arms.

Interventions can succeed only if there is sufficient uptake in the target population, so we also want to evaluate uptake of the interventions.

Step 3: Choose an appropriate subset of data

Click the **Browse and subset** tab.

1. If you are interested in the outcome of this trial, how would you restrict your subset to the endpoint? The featured variables list shows two timepoint variables, and we will have to choose the endpoint data for both these variables:

- Household level: *Household study timepoint*
- Participant level: *Study timepoint*

2. If you want to evaluate diarrhea specifically in target children, how would you choose this subset of participants from the variable *Target child or sibling/neighbor*?

1 a. Subset the *Household study timepoint* = 2 (Household endpoint data)

Household study timepoint
Original variable name: uptake_0:time, uptake_2:time, enrol-time, uptake_1:time
Survey data on households were collected at enrollment (year 0, pre-intervention baseline), 1 year post-intervention (midline), and 2 years post-intervention (endline). If you would like to view only data from a particular timepoint for all variables, please make sure to select that timepoint in two locations: 'Household study timepoint' and 'Study timepoint'.

☐ Keep checked values at top

	Subset of Household repeated measures	All Household repeated measures	Distribution	%
<input type="checkbox"/> 0	8,246 (39%)	8,246 (39%)		(100%)
<input type="checkbox"/> 1	6,084 (28%)	6,084 (28%)		(100%)
<input checked="" type="checkbox"/> 2	7,072 (33%)	7,072 (33%)		(100%)

1 b. Subset the *Study timepoint* = 2 (Participant endpoint data)

Study timepoint
Original variable name: anthro_1:time, anthro_2:time, diar:time, bruise:time
Survey data on participants were collected at enrollment (year 0, pre-intervention baseline), 1 year post-intervention (midline), and 2 years post-intervention (endline). Outcomes were measured after 1 year (midline) and 2 years (endline) of intervention delivery. If you would like to view only data from a particular timepoint for all variables, please make sure to select that timepoint in two locations: 'Household study timepoint' and 'Study timepoint'.

☐ Keep checked values at top

	Subset of Participant repeated measures	All Participant repeated measures	Distribution	%
<input type="checkbox"/> 0	11,731 (33%)	13,574 (33%)		(86%)
<input type="checkbox"/> 1	11,731 (33%)	13,574 (33%)		(86%)
<input checked="" type="checkbox"/> 2	11,731 (33%)	13,574 (33%)		(86%)

2. Subset the *Target child or sibling/neighbor* = Target child

Target child or sibling/neighbor
Original variable name: anthro_2::targetchild, diar::targetchild, anthro_1::targetchild
Target child participants were in utero at enrollment and were the primary study population (referred to as 'index children' in the manuscript). Older sibling or neighbor participants were less than 36m at the time of target child enrollment.

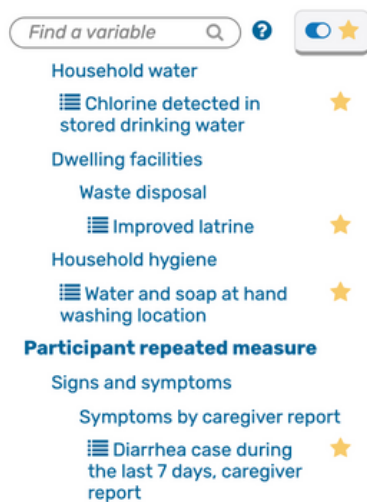
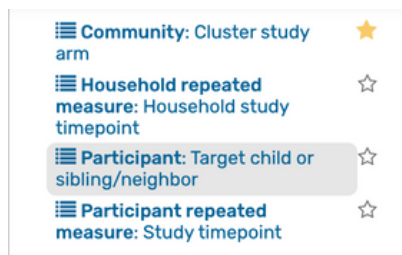
☐ Keep checked values at top

	Subset of Participants	All Participants	Distribution	%
<input checked="" type="checkbox"/> Target child	6,532 (59%)	7,417 (58%)		
<input type="checkbox"/> Sibling or neighbor child	4,603 (41%)	5,309 (42%)		

How many participants are present in your subset?
Check the dataset diagram at the top of the page.

Participants
6,532 of 13,574

Step 4: Identify variables of interest for this analysis



When identifying variables to use in an analysis, look at the distribution of the variable and note whether it is numeric or categorical, as this will help decide what visualization tools to use in the next step.

1. **Intervention:** Identify the variable that indicates whether the cluster was randomized to receive one of the interventions. This is the *Cluster study arm*, a categorical variable.

2. **Intervention uptake:** There are a number of variables under *Household repeated measures* in the variable tree that assess uptake of WASH interventions, such as *Chlorine detected in stored drinking water*, *Improved latrine*, and *Water and soap at hand washing location*. They are all categorical variables.

3. **Outcome:** What variable could we use to assess the outcome of diarrhea? A suitable variable would be *Diarrhea case during the last 7 days, caregiver report*, a categorical variable.

You can star the variables of interest to make them easier to access.

Step 5: Create visualizations to examine associations between variables

Make a list of the associations you want to plot and the appropriate tool to use for each.

Association	X axis	Y axis	Plot
Diarrhea among target children in different study arms	Cluster study arm (categorical variable with 7 levels)	Diarrhea case during the last 7 days, caregiver report (categorical variable with 2 levels)	Mosaic Plot RxC Table
Uptake of water quality intervention in different study arms	Cluster study arm (categorical variable with 7 levels)	Chlorine detected in stored drinking water (categorical variable with 2 levels)	Mosaic Plot RxC Table
Uptake of sanitation intervention in different study arms	Cluster study arm (categorical variable with 7 levels)	Improved latrine (categorical variable with 2 levels)	Mosaic Plot RxC Table
Uptake of handwashing intervention in different study arms	Cluster study arm (categorical variable with 7 levels)	Water and soap at hand washing location (categorical variable with 2 levels)	Mosaic Plot RxC Table

Click on the **Visualize** tab to select the appropriate tool and make the plots.
Name each plot.

Uptake: water quality

Mosaic Plot, RxC Table

Axis variables

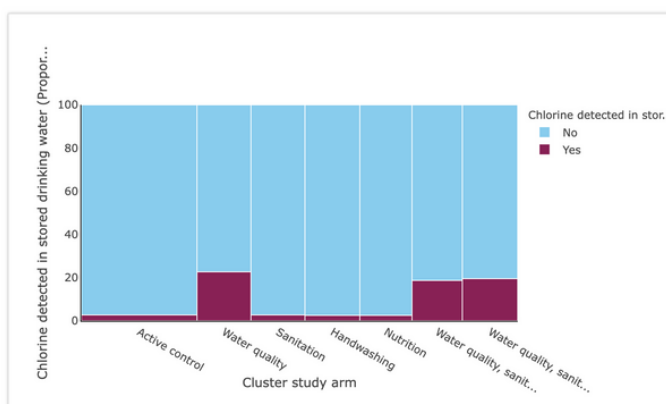
X-axis Cluster study arm

Y-axis Chlorine detected in stored drinking water

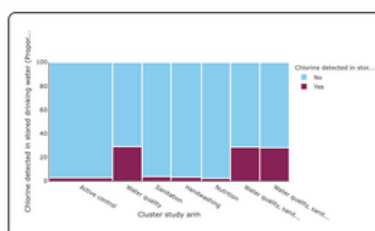
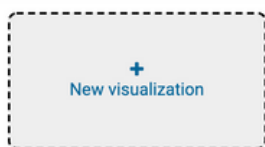
5,151 Household repeated measures

Mosaic

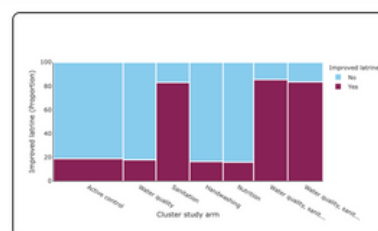
Table



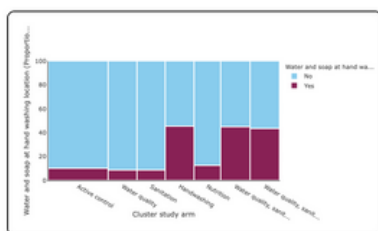
Your plots may look like this:



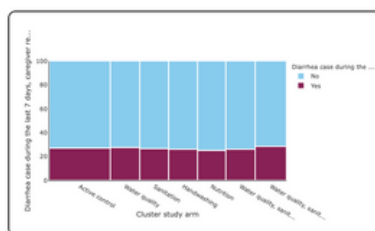
Uptake: water quality
Mosaic Plot, RxC Table



Uptake: sanitation
Mosaic Plot, RxC Table



Uptake: handwashing
Mosaic Plot, RxC Table



Diarrhea in each study arm
Mosaic Plot, RxC Table

What do each of these plots say about the uptake of the intervention across study arms? The mosaic plots indicate clear differences of uptake between intervention and control arms, although the level of uptake differs for different interventions, for instance, uptake is higher for the sanitation intervention compared to the water quality intervention.

Does the outcome of diarrhea differ across study arms? The plot indicates no noticeable difference in diarrhea among target children across study arms.

Indeed, the published results of this trial show that none of the interventions reduced diarrhea prevalence compared with the active control.

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 tutorial on performing an exploratory data analysis on clinepidb.org! Please contact help@clinepidb.org with feedback or questions.