

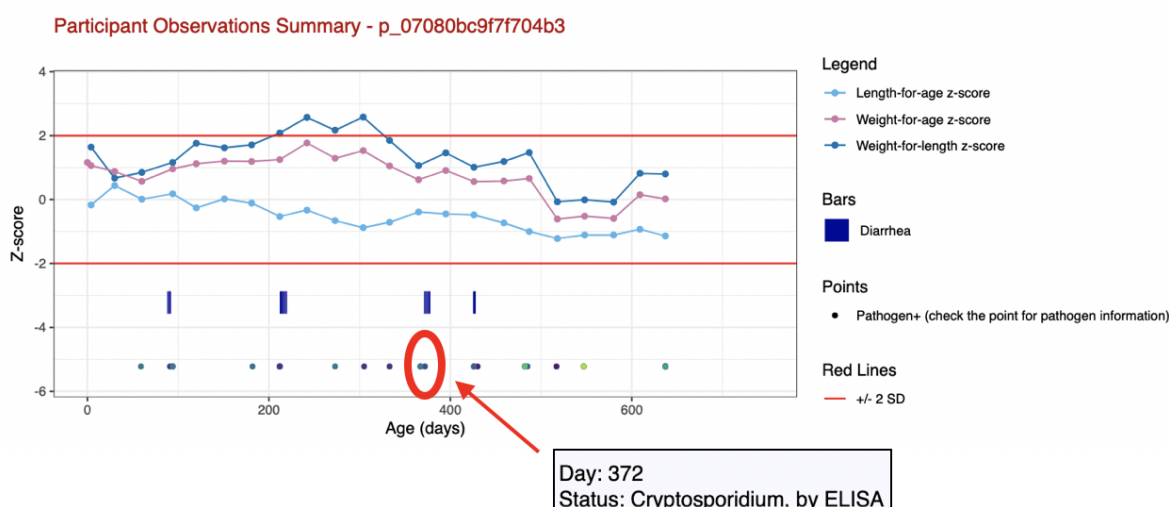
# Complex Epidemiology Queries

## Guided Walk-Through

### Building a Search with Related Observations

In this exercise, we will go through an example together to perform a more advanced search using the “Related Observations” function. Several studies on ClinEpiDB included extensive collection of longitudinal information. The ability to search these data easily is extremely valuable. Integrating the “Related Observations” function into your searches allows you to examine data over time.

Before we start, consider the graph below. This graph plots height-for-age z-scores (light blue), weight-for-height z-scores (dark blue), and weight-for-age z-scores (pink), measured over time, for one of the participants in the MAL-ED study.



During follow-up, this participant had four episodes of diarrhea (the blue hash bars on the graph indicate diarrheal episodes). The thickness of each blue hash bar is proportional to the duration that diarrheal episode lasted. There are two diarrheal episodes that appear to be greater than a few days in length.

You can also see that this participant had multiple instances of positive laboratory test results for microbiology, indicated by the colored dots at the bottom of the graph (different colored dots represent the detection of different pathogens). One question you might want to ask is, “**Did participants tend to test positive for a certain type of pathogen before an episode of diarrhea that was five or more days in duration?**”

To explore this question, your first step might be to explore the data and see how many participants tested positive for *Cryptosporidium* within the 14 days before a diarrheal episode that was five or more days in duration. Then you might want to repeat the same question for any *Campylobacter*.

**You can do this by using the Related Observations functionality!** We will start with a very simple related observations query. For this exercise, to start, we will modify as few filters as possible to keep things straightforward.

## Question 1: How many participants in the MAL-ED study tested positive for *Cryptosporidium* within the 14 days before a diarrheal episode that was five or more days in duration?

1. Start a participant-level search of the MAL-ED data. To answer this question, first select participants who tested positive for *Cryptosporidium* at any point during the study.
  - The cleaned, analytic versions of all microbiology detection results of various organisms in stool samples can be found under “Bacteria in stool”, “Virus in stool”, and “Eukaryota in stool”.
  - Navigate to “Eukaryota in stool” and then select participants who had *Cryptosporidium* detected at any time in the 60 months of follow-up.

The screenshot shows the MAL-ED data search interface. The top navigation bar includes filters for Geographic Region (2,145), Household (2,145), Participant (2,145), Observation (2,145), and Related Observation (1,007). The 'Observation' filter is highlighted with a red box. Below the navigation bar, the 'Eukaryota in stool' section is active. It displays a table of observations with columns for 'Remaining Observations', 'All Observations', 'Distribution', and '%'. The table shows results for 'Chilomastix mesnili, by microscopy', 'Cryptosporidium, by ELISA', and 'Cyclospora, by microscopy'. Under 'Cryptosporidium, by ELISA', the 'Yes' option is selected, showing 1,796 observations (5% of the total).

Observation	Remaining Observations	All Observations	Distribution	%
Chilomastix mesnili, by microscopy	44,628 (100%)	44,628 (100%)	2% of 1,848,829 Observations have data	
<input type="checkbox"/> No	44,538 (>99%)	44,538 (>99%)		(100%)
<input type="checkbox"/> Yes	90 (< 1%)	90 (< 1%)		(100%)
Cryptosporidium, by ELISA	39,228 (100%)	39,228 (100%)	2% of 1,848,829 Observations have data	
<input type="checkbox"/> No	37,432 (95%)	37,432 (95%)		(100%)
<input checked="" type="checkbox"/> Yes	1,796 (5%)	1,796 (5%)		(100%)
Cyclospora, by microscopy	42,284 (100%)	42,284 (100%)	2% of 1,848,829 Observations have data	

2. Now, click on the “Related Observations” box in the Search Wizard. “Related Observations” are used to narrow the dataset based on the relationship between two variables at two different points in time. Click on the small check box next to: “Enable the advanced Related observations filter below.” It allows you to restrict Observations by relating them to your choice of Related observations.”

The screenshot shows the MAL-ED data search interface with the 'Related Observation' filter selected. Below the filter, a message states: 'Your Related Observation filters reduce 1,007 Participants to 7'. A checkbox labeled 'Enable the advanced Related Observation filter below. It allows you to restrict Observation by relating them to your choice of Related Observation.' is checked. Below this, a filter configuration is shown: 'Remove Observation that are 0 to 10 days before the Related Observation specified below'.

Filter selected observations based on a time-dependent relationship to a second set of observations defined below. For example, keep only diarrheal observations that were followed by a shigella-positive lab test 0-5 days later.

☒ Enable the advanced Related Observation filter below. It allows you to restrict Observation by relating them to your choice of Related Observation.

Remove Observation that are 0 to 10 days before the Related Observation specified below

3. After you have enabled the Related Observations filter, notice that the variables are now accessible. Take a look at the various variables and categories that appear in the variable hierarchy on the left side of the page. Do you notice anything familiar about these variables?<sup>1</sup>
4. Now we will want to specify that we are selecting diarrheal episodes of at least 5 days in length that occurred at any point during the study for our Related Observation. Use the “find a filter” search box to find variables related to “Diarrheal duration” and type “5” into the appropriate window to select episodes lasting at least 5 days.
  - For now, do not worry about properly completing the sentence above the variable selection panels. We will update this in the next step.

2,145 → 2,145 → 2,145 → 2,145 → 1,007

**Your Related Observation filters reduce 1,007 Participants to 930**

☒ Enable the advanced Related Observation filter below. It allows you to restrict Observation by relating them to your choice of Related Observation.

Remove Observation that are 0 to 10 days before the Related Observation specified below

Diarrheal episode duration (days)

diarrheal duration

**Diarrheal episode duration (days)**

Duration of the diarrheal episode, in days. Diarrhea episodes were separated by two diarrhea-free days. Days with missing data were classified as diarrhea-free (if there were two missing days before a day with diarrhea, the episode started on day with diarrhea). Days with Lactulose:Mannitol (L:M)-induced diarrhea were classified as diarrhea-free. Surveillance was conducted twice weekly for up to 3 years of age to assess illness on all days since the last surveillance visit. (Provider label: dur)

Min: 1 Mean: 8.58 Median: 6 Max: 113

Select Diarrheal episode duration (days) from 5 to 113

49,306 (3%) of 1,848,829 Related Observations have data for this variable

Related Observations

Diarrheal episode duration (days)

<sup>1</sup> The variables and categories available on the Related Observations step in the Search Wizard are duplicates of those that were available in the Observations step in the Search Wizard. The Related Observations step allows us to take the timing of different episodes, results, or data collection into account so we can select for data that occurred before (or after) something else had occurred.

6. We now want to account for the appropriate timing of the *Cryptosporidium* infections in relation to the 5+ day diarrheal episodes. We need to identify participants who tested positive for *Cryptosporidium* within the 2 weeks **BEFORE** a diarrheal episode lasting at least 5 days.

Completing the sentence that we skipped over in step 4 allows us to properly account for the timing of our selections in relation to one another.

In the Observations step in the Search Wizard, we previously selected participants who had *Cryptosporidium* detected at any point in the study period.

In the Related Observations step, we now want to narrow down the selected participants further, and keep only the selected participants who also had a diarrheal episode lasting at least 5 days. However, we don't want to keep all selected participants who **EVER** had a diarrheal episode lasting at least 5 days at any point in the 5 years of follow-up. The timing of the *Cryptosporidium* test in relation to particular diarrheal episodes is important.

Complete the 4 boxes in the sentence below to properly account for the timing of *Cryptosporidium*-positive laboratory results in relation to diarrheal episodes lasting at least 5 days.<sup>2</sup> *HINT: In the sentence, "Observations" refers to the observations we selected in the previous step (positive *Cryptosporidium* test results). "Related observations" refers to the observations we are selecting now (diarrheal episodes that lasted at least 5 days).*

<div style="border: 1px solid black; padding: 2px; display: inline-block;">Keep / Remove</div>	Observation that are	<div style="border: 1px solid black; padding: 2px; display: inline-block;">#</div>	to	<div style="border: 1px solid black; padding: 2px; display: inline-block;">#</div>	days	<div style="border: 1px solid black; padding: 2px; display: inline-block;">Before / After</div>	the Related Observation specified below
--	----------------------	--	----	--	------	---	---

7. How many participants in the MAL-ED study tested positive for *Cryptosporidium* within the 14 days before a diarrheal episode that was five or more days in duration?<sup>3</sup>

---

<sup>2</sup> The sentence should be completed to read: "**Keep** observations that are **0 to 14** days **before** the Related Observation specified below."

<sup>3</sup> 211 participants had a 5+ day diarrheal episode that was preceded by *Cryptosporidium* detection within the previous 2 weeks.



## Your **Related Observation** filters reduce 1,007 Participants to 211

Filter selected observations based on a time-dependent relationship to a second set of observations defined below. For example, keep only diarrheal observations that were followed by a shigella-positive lab test 0-5 days later.

☒ Enable the advanced **Related Observation** filter below. It allows you to restrict **Observation** by relating them to your choice of **Related Observation**.

Keep  to  days before the **Related Observation** specified below

Diarrheal episode duration (days) X

diarrheal duration X ?

### Diarrheal episodes

Diarrheal episode duration (days)

### Diarrheal episode severity

- ☒ Persistent diarrheal episode (more than 14 days)
- ☒ Prolonged diarrheal episode (7 to 14 days)
- ☒ Acute diarrheal episode (less than 7 days)
- ☒ Diarrhea duration categorization

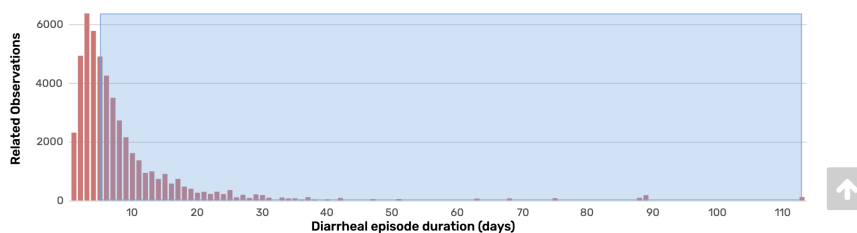
### Diarrheal episode duration (days)

Duration of the diarrheal episode, in days. Diarrhea episodes were separated by two diarrhea-free days. Days with missing data were classified as diarrhea-free (if there were two missing days before a day with diarrhea, the episode started on day with diarrhea). Days with Lactulose:Mannitol (L:M)-induced diarrhea were classified as diarrhea-free. Surveillance was conducted twice weekly for up to 3 years of age to assess illness on all days since the last surveillance visit. (Provider label: dur)

Min: 1 Mean: 8.58 Median: 6 Max: 113

Select Diarrheal episode duration (days) from  to


49,306 (3%) of 1,848,829 Related Observations have data for this variable



## Question 2: Is there a different way we could have approached Question 1? Does this give you the same result?

What would happen if we instead selected diarrheal episodes lasting at least 5 days for our Observation and *Cryptosporidium* detection for our Related Observation? Would we get the same results? Lets try!

1. First, let's clear all of the selections we made previously. Click on the green filter icon next to "Select a set of participants (MAL-ED 0-60m)". In the popup window, click "Remove all".

Select a Set of Participants (MAL-ED 0-60m)  [Study Details >](#)

**Summary:** The Etiology Center Longitudinal Study (MAL-ED) was a multi-center longitudinal study of children in 10 countries for up to 60 months of age. The study collected various data types including anthropometric measurements, clinical symptoms, survey results, lab results, etc.

**Find a variable**

**Active Filters**

**Observation**  
ALL Eukaryota in stool filters  
*Cryptosporidium*, by ELISA  
Yes

**Related Observation**  
Use related observations Yes  
Days between observations: {'min': '0', 'max': '14'}  
Choose a direction before  
Keep or Remove observations from previous step matching these criteria keep

**Remove all**

**Related Observation** [View 211 Participants](#)

**Your Related Observations**  
211

2. Now, click on the "Observations" step in the Search Wizard and select participants who had a diarrheal episode of at least 5 days in length at any point during the study.

**Search Wizard**

**Geographic Region** → **Household** → **Participant** → **Observation** → **Related Observation**

2,145 → 2,145 → 2,145 → 2,145 → 1,067

**Your Observation filters reduce 2,145 Participants to 1,067**

Filter based on observation variables (anthropometric measurements, clinical symptoms, survey results, lab results, etc.). Data was collected at multiple timepoints so may require a multi-step strategy to return all data of interest.

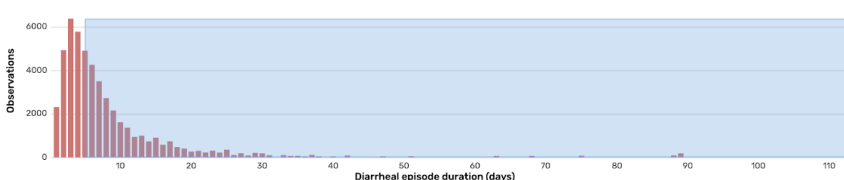
**Diarrheal episode duration (days)**

Duration of the diarrheal episode, in days. Diarrhea episodes were separated by two diarrhea-free days. Days with missing data were classified as diarrhea-free (if there were two missing days before a day with diarrhea, the episode started on day with diarrhea). Days with Lactulose-Mannitol (LM)-induced diarrhea were classified as diarrhea-free. Surveillance was conducted twice weekly for up to 3 years of age to assess illness on all days since the last surveillance visit. (Provider label: dur)

**Min: 1 Mean: 8.58 Median: 6 Max: 113**

Select Diarrheal episode duration (days) from  to

**49,306 (3%) of 1,848,829 Observations have data for this variable**



8. Next, click on the "Related Observations" box in the Search Wizard. Click on the box to "Enable the advanced Related observations filter below." Select *Cryptosporidium*-positive observations for our Related Observation.

9. Complete the 4 boxes in the sentence below to properly account for the timing of *Cryptosporidium*-positive laboratory results in relation to diarrheal episodes lasting at least 5 days.<sup>4</sup> *HINT: In the sentence, “Observations” now refers to the observations we selected in the previous step (diarrheal episodes that lasted at least 5 days). “Related observations” now refers to the observations we are selecting now (positive *Cryptosporidium* test results).*

Observation that are  to  days  the Related Observation specified below

10. How many participants in the MAL-ED study tested positive for *Cryptosporidium* within the 14 days before a diarrheal episode that was five or more days in duration? Using this approach, do we return the same number of selected participants as with the approach we took to answer Question 1?<sup>5</sup>



**Your Related Observation filters reduce 1,067 Participants to 211**

Filter selected observations based on a time-dependent relationship to a second set of observations defined below. For example, keep only diarrheal observations that were followed by a shigella-positive lab test 0-5 days later.

☒ Enable the advanced **Related Observation** filter below. It allows you to restrict **Observation** by relating them to your choice of **Related Observation**.

Keep  Observation that are  to  days  the Related Observation specified below  
 ALL Eukaryota in stool filters

cryptosporidium elisa

Laboratory test

Stool microbiology test

Eukaryota in stool

Stool microbiology test procedures

Stool test count

Stool eukaryote testing

Protozoa ELISA tests

Protozoa ELISA testing completed

Cryptosporidium ELISA performed

Raw test result for stool

Raw eukaryota data for stool

Cryptosporidium, by ELISA result

Cryptosporidium, by modified acid stain microscopy result

Test result summary for stool

Any parasites, by any method

Eukaryota in stool

Detection of the following eukaryotes in stool samples. Work up for detection of enteric pathogens was performed on diarrheal stool samples collected until 24 months of age (with a possible extension to 36 months of age) and first monthly asymptomatic stool samples collected at 1-12, 15, 18, 21, and 24 months of age (with the option to test additional monthly stool samples to 36 months of age).

Find Related Observations with  of the options selected below.

Eukaryota in stool	Remaining Related Observations	All Related Observations	Distribution	%
Chilomastix mesnili, by microscopy	44,628 (100%)	44,628 (100%)	2% of 1,848,829 Related Observations have data	
<input type="checkbox"/> No	44,538 (>99%)	44,538 (>99%)		(100%)
<input type="checkbox"/> Yes	90 (< 1%)	90 (< 1%)		(100%)
Cryptosporidium, by ELISA	39,228 (100%)	39,228 (100%)	2% of 1,848,829 Related Observations have data	
<input type="checkbox"/> No	37,432 (95%)	37,432 (95%)		(100%)
<input checked="" type="checkbox"/> Yes	1,796 (5%)	1,796 (5%)		(100%)
Cyclospora, by microscopy	42,284 (100%)	42,284 (100%)	2% of 1,848,829 Related Observations have data	
<input type="checkbox"/> No	42,265 (>99%)	42,265 (>99%)		(100%)
<input type="checkbox"/> Yes	19 (< 1%)	19 (< 1%)		(100%)

<sup>4</sup> The sentence should be completed to read: “Keep observations that are 0 to 14 days after the Related Observation specified below.”

<sup>5</sup> The two different approaches both returned 211 participants who met our selection criteria.

## Multi-step strategies and combining searches

In the last search, we identified MAL-ED participants who had 1 or more episodes of diarrhea that were at least 5 days in duration and that were preceded by a detection of *Cryptosporidium* in the last 2 weeks. How many of these participants also had a reduced length-for-age z-score at 2 years of age?

To answer this question, you will need to employ a multi-step strategy. We will combine the results of our previous search, which identified all MAL-ED participants who had a *Cryptosporidium* infection followed by an episode of diarrhea lasting at least 5 days (**search 1**), with a new search to identify all MAL-ED participants who had a length-for-age z-score of less than -2 at 2 years of age (**search 2**).

**Question 3: How many participants in the MAL-ED study tested positive for *Cryptosporidium* within the 14 days before a diarrheal episode that was five or more days in duration and also had a length-for-age z-score of less than -2 at 2 years of age?**

1. Click on “View 211 participants” button on the Search Wizard to navigate to the Results Table.



2. From the Results Table of the 211 MAL-ED participants who were *Cryptosporidium*-positive within 14 days before a diarrheal episode lasting at least 5 days, navigate to the search strategy panel above the table and click the “Add a step” button. Why do we need to add a step to our search strategy instead of just applying additional filters to our original search?<sup>6</sup>

**211 MAL-ED 0-60m Participants** [Revise this search](#)

MAL-ED 0-60m Participant Results [Analyze Results](#)

Rows per page: 20

Participant Id	Sex	Total illness surveillance days	Total diarrheal episodes	Total ALRI episodes	Age last exclusively breastfed (days)	
p_010e6df71b4eafe0	Male	1121	23	3	3	6
p_011acd8b695eafb3	Female	733	7	0	3	2
p_03dfdf53c758051f	Male	602	4	0	89	6

<sup>6</sup> In search 1, we identified all MAL-ED participants who had a *Cryptosporidium* infection followed by an episode of diarrhea lasting at least 5 days. These *Cryptosporidium* infections could have occurred at **ANY** time during follow-up. For search 2, we are being asked to identify participants who had a reduced length-for-age z-score at exactly 2 years of age. If we apply a filter to select “Age” of 2 years to search 1, we will only be selecting *Cryptosporidium* infections that occurred when the participant was 2 years old. Participants who had the *Cryptosporidium* detections occurring before and after 2 years of age would be removed from our selections.





3. In the “Add a step” popup window, click to select “Combine with other MAL-ED 0-60m Participants” and then choose how to combine this new search (search 2) with the original search (search 1).
  - Since we want to identify all participants who had a length-for-age z-score less than -2 at 2 years of age (search 2) who **ALSO** were *Campylobacter*-positive within 14 days before a diarrheal episode lasting at least 5 days (search 1), we want to find the intersection of the two searches. Go ahead and select “1 Intersect 2.”
  - Finally, choose “A new search” to combine a new subset of MAL-ED participants with the subset of participants identified in the original search (search 1).
  - Click “Select a set of participants (MAL-ED 0-60m)”.


4. The “Add a step” popup window will now contain a new Search Wizard. Use this Search Wizard to obtain a list of all MAL-ED participants who had a length-for-age z-score lower than -2 at 2 years of age (search 2).

- For search 2, we want to identify the list of participants who had a length-for-age z-score less than -2 at 2 years of age. The first step in this search is to select participants who had a length-for-age z-score less than -2 at any point during the study. Z-scores are sorted under the “Observations” step in the Search Wizard. Remember that you can use “Find a filter” to quickly search for variables of interest.

Add a step to your search strategy ?

**Search for Participants by Select a Set of Participants (MAL-ED 0-60m)**

The results will be  intersected with  the results of Step 1.

  
2,145

→

Geographic Region

  
2,145

→

Household


  
2,145

→

Participant

  
2,145

→

Observation 


  
1,081

→

Related Observation

Combine 1,081 Participants

  
Name this search

length for age z 

?

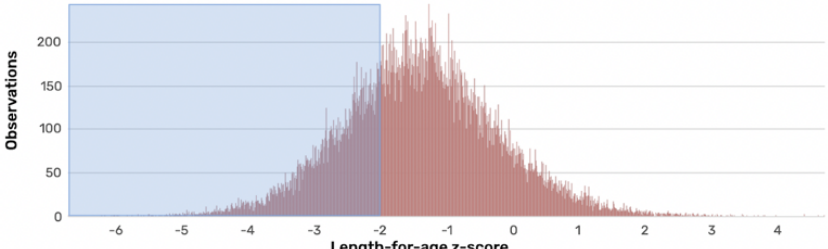
**Length-for-age z-score**

Length-for-age z-score. Anthropometry was assessed on a monthly basis up to and including 36m of age, and subsequently every 3 months up to and including 60m of age. (Provider label: zlen)

**Min:** -6.72 **Mean:** -1.4 **Median:** -1.42 **Max:** 4.71

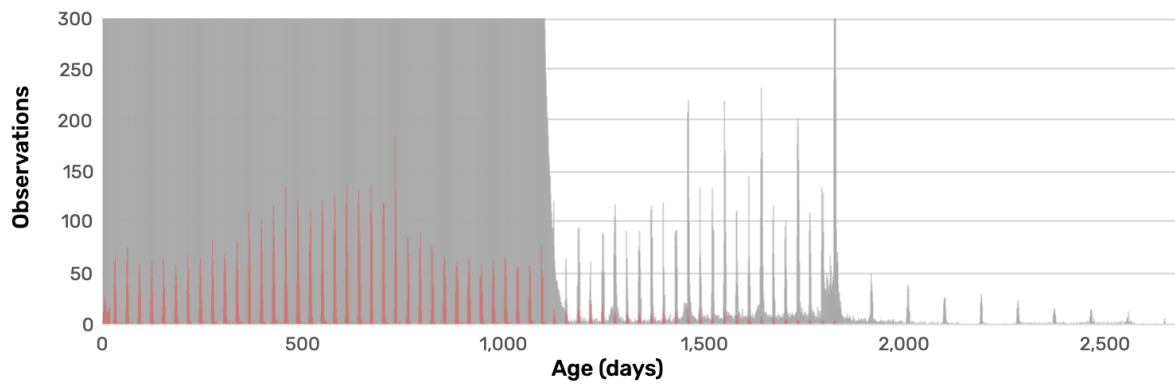
Select Length-for-age z-score from  to

**51,903 (3%) of 1,848,829** Observations have data for this variable



- We still need to limit the selected length-for-age z-scores to those that were collected when the participant was 2 years of age. Navigate to the “Age (days)” variable in the “Observations” step of the Search Wizard. Use the “Plot settings” under the graph to zoom in on the data; set the y-axis to range from 0 to 300. For this variable, notice that there are a series of red peaks on a grey background. Why do you think the data looks like this?<sup>7</sup>

<sup>7</sup> From the information on the MAL-ED Data Set page, we can determine that anthropometry variables such as length-for-age z-scores were measured monthly. The peaks represent every month of anthropometry data collection, with some variation around the exact day of age when the data was collected.



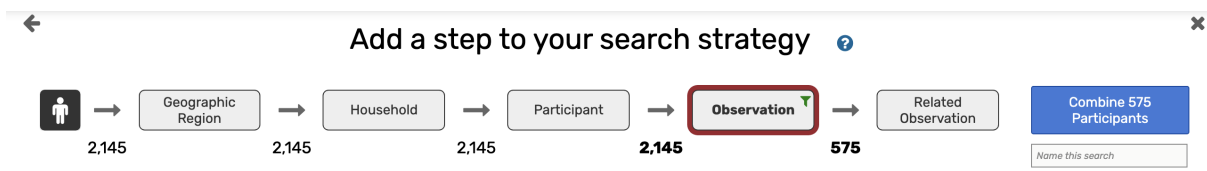
▼ ⚙ Plot Settings

**y-axis**

Scale counts: ☒ linear ☐ log<sub>10</sub>

❗ Range:  to  (1 - 2145)

7. Select data that was collected at approximately 2 years of age. There is variation on exactly when the 2 year timepoint of anthropometry data was collected, so be sure to account for this variation by choosing 710 to 750 days for the “Age (days)” range.



Your **Observation** filters reduce 2,145 Participants to 575

Filter based on observation variables (anthropometric measurements, clinical symptoms, survey results, lab results, etc.). Data was collected at multiple timepoints so may require a multi-step strategy to return all data of interest.

Length-for-age z-score × Age (days) ×

- expand all | collapse all
- Find a variable
- Age (days)**
- Diarrhea, ALRI, fever, or vomiting present
- Sample
- Observation details
- Anthropometry
- Length- or height-for-age z-score
  - Length-for-age z-score
  - Height-for-age z-score
  - Weight-for-age z-score
  - Weight-for-length or -height z-score
  - Weight-for-length z-score
  - Weight-for-height z-score
  - Head circumference-for-

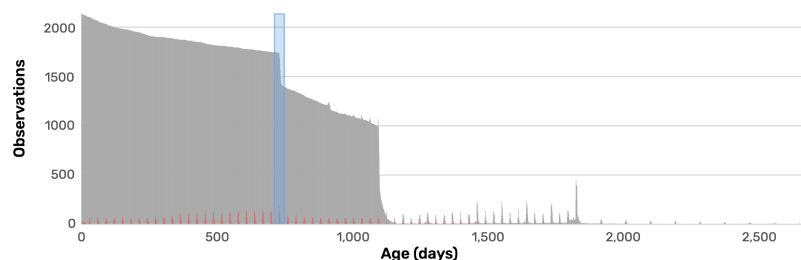
### Age (days)

(Provider label: age, agedays)

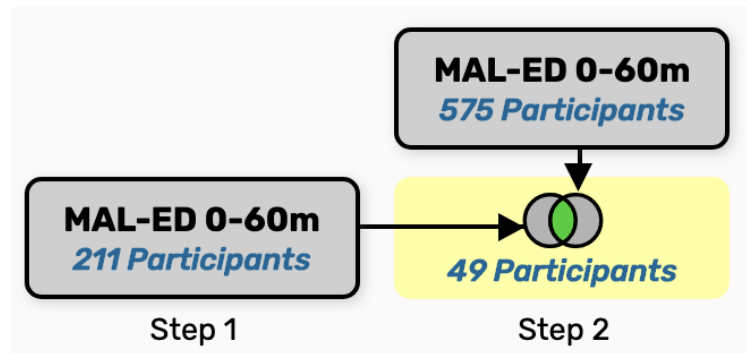
Min: 0 Mean: 600.71 Median: 580 Max: 1,832

Select Age (days) from  to

1,848,829 (100%) of 1,848,829 Observations have data for this variable



8. Search 2 determined that there were 575 MAL-ED participants who had a length-for-age z-score less than -2 at 2 years of age. Click on the blue “Combine 575 Participants” box in the Search Wizard to combine the results of search 2 with search 1.
9. How many participants in the MAL-ED study tested positive for *Cryptosporidium* within the 14 days before a diarrheal episode at least five days in duration and also had a length-for-age z-score of less than -2 at 2 years of age?<sup>8</sup>



---

<sup>8</sup> There were 49 participants that were returned in both search 1 and search 2.