Exploring & Analyzing Enteric Epidemiology Datasets II

Exercise 1. Comparing pathogen detection methods

With continuing advances in molecular diagnostics many epidemiologic studies are beginning to adopt new culture-independent diagnostics for detecting pathogens. In ClinEpiDB we have loaded both conventional microbiologic test results as well as the TaqMan gene expression array cards (TAC) for GEMS1 and MAL-ED. The conventional diagnostics are generally presented as categorical variables either in (a) their 'raw' unprocessed form and/or as (b) transformed binary analytic variables and the TaqMan data is presented through a (c) continuous Ct value.
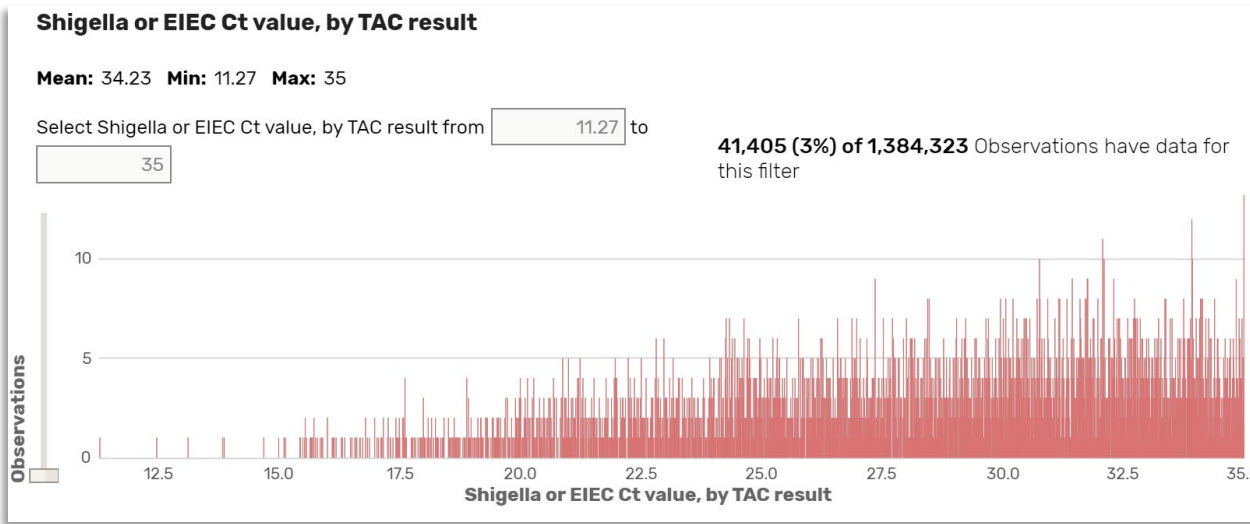
(a)

	Shigella, by bacteriology result	Remaining Observatio... ❓		Observatio... ❓		Distribution ❓	% ❓
		39,837	(100%)	39,837	(100%)		
☐	**No**	39,492	(99%)	39,492	(99%)	▬▬▬▬▬▬▬▬▬▬	(100%)
☐	**Shigella boydii GpC**	55	(0%)	55	(0%)	|	(100%)
☐	**Shigella dysenteriae GpA**	35	(0%)	35	(0%)	|	(100%)
☐	**Shigella flexneri GpB**	142	(0%)	142	(0%)	|	(100%)
☐	**Shigella sonnei GpD**	105	(0%)	105	(0%)	|	(100%)
☐	**Shigella sp non-typable**	9	(0%)	9	(0%)	|	(100%)

(b)

	Bacteria	Remaining Observations		All Observations		Distribution	%
	Shigella ✖						
	Shigella, by bacteriology	39,837	(100%)	39,837	(100%)	**1344486** Observations have no data	
☐	No	39,492	(99%)	39,492	(99%)	▬▬▬▬▬▬▬▬▬	(100%)
☐	Yes	346	(1%)	346	(1%)	|	(100%)

(c)

Shigella or EIEC Ct value, by TAC result

Mean: 34.23   **Min:** 11.27   **Max:** 35

Select Shigella or EIEC Ct value, by TAC result from [11.27] to [35]

41,405 (3%) of 1,384,323 Observations have data for this filter

Note that (a) and (b) are two different versions of the same result. The binary version (a) tells you whether or not any Shigella was found in the stool sample by conventional culture result. The categorical version (b) indicates the specific Shigella subgroup of serotype that was identified. Alternatively, detection version (c) indicate a completely separate test result; the Ct value from TaqMan gene expression array card (TAC) result. The maximum Ct value for the majority of the TAC results are truncated at 35.0; this was done in data cleaning as a means of establishing the analytical limit of detection.

Use the stool microbiology test results from the MAL-ED study to answer the following questions:

a. # of participants who had a Shigella or EIEC Ct value, by TAC result of exactly 35 _____[1]

b. # of participants who had a Shigella or EIEC Ct value, by TAC result less than 25 _____[2]

c. # of participant who had a bacteriology result indicating Shigella flexneri GpB _____[3]

d. # of participants who had Cryptosporidium identified by modified acid stain _____[4]

e. # of participants who had a positive Cryptosporidium, by ELISA result _____[5]

f. # of participants who had a Cryptosporidium Ct value less than 35 _____[6]

[Note to get less than any # you will either have to point and select very carefully on the histogram or type ##.99 into the result selection maximum.]

Notice that the conventional microbiologic results often differ from the molecular methods. Generally, molecular diagnostics may have a greater sensitivity but a lower specificity than culture-based methods. Both GEMS and MAL-ED reanalyzed primary study objectives using quantitative molecular diagnostic assays following publication of conventional method results. Without even downloading the dataset we can use ClinEpiDB to do exploratory analyses on how results differ between the conventional and diagnostic methods.
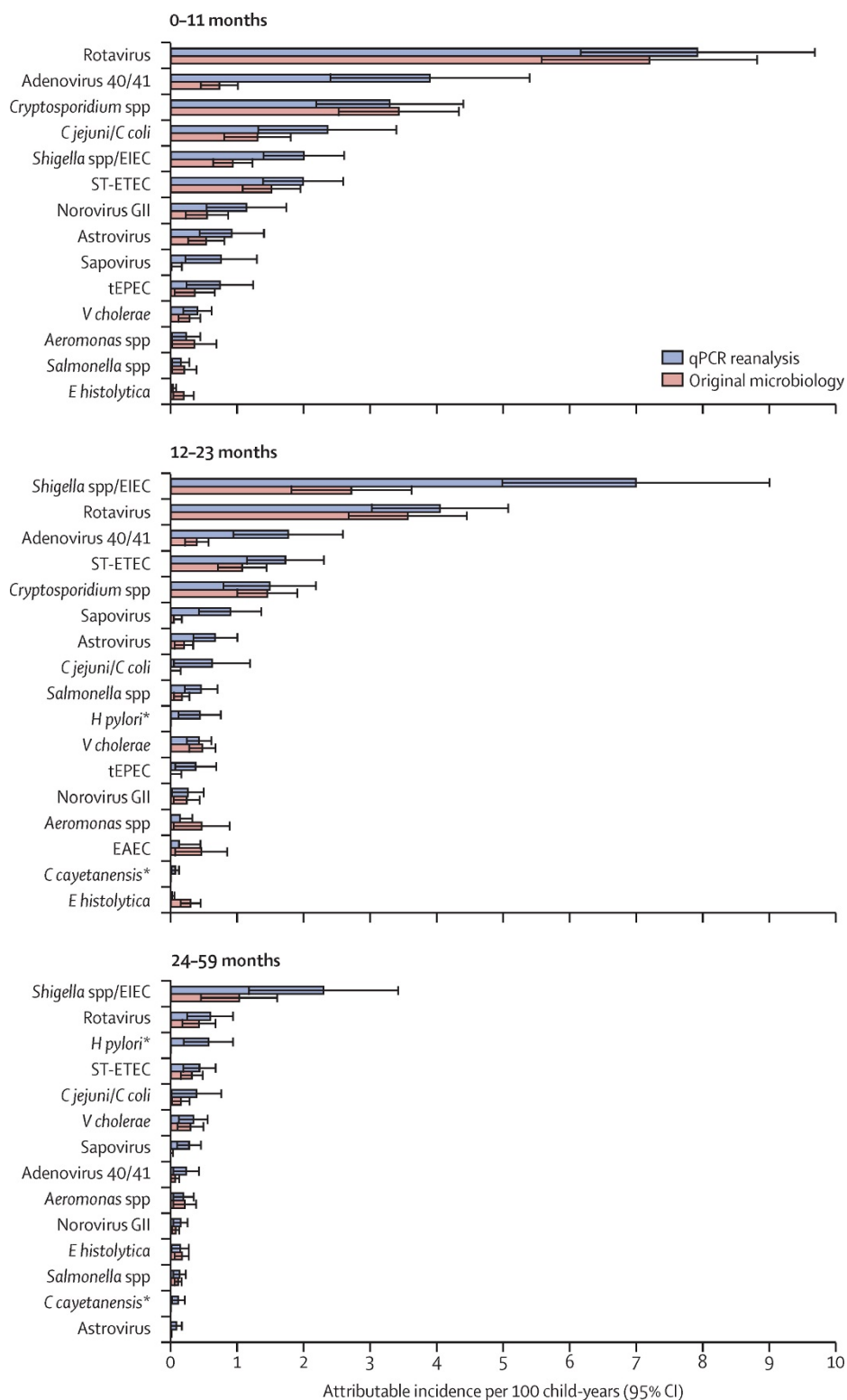
---

[1] 1,715
[2] 711
[3] 122
[4] 608
[5] 994
[6] 1,235

0–11 months

Rotavirus
Adenovirus 40/41
Cryptosporidium spp
C jejuni/C coli
Shigella spp/EIEC
ST-ETEC
Norovirus GII
Astrovirus
Sapovirus
tEPEC
V cholerae
Aeromonas spp
Salmonella spp
E histolytica

qPCR reanalysis
Original microbiology

12–23 months

Shigella spp/EIEC
Rotavirus
Adenovirus 40/41
ST-ETEC
Cryptosporidium spp
Sapovirus
Astrovirus
C jejuni/C coli
Salmonella spp
*H pylori**
V cholerae
tEPEC
Norovirus GII
Aeromonas spp
EAEC
*C cayetanensis**
E histolytica

24–59 months

Shigella spp/EIEC
Rotavirus
*H pylori**
ST-ETEC
C jejuni/C coli
V cholerae
Sapovirus
Adenovirus 40/41
Aeromonas spp
Norovirus GII
E histolytica
Salmonella spp
*C cayetanensis**
Astrovirus

Attributable incidence per 100 child-years (95% CI)

A remarkable finding from a GEMS study re-analysis of the stool samples using quantitative PCR (qPCR) produced higher estimates for Shigella, increasing the attributable incidence by about twofold for Shigella and 1.5-foldfor ST-ETEC (STh-producing enterotoxigenic E coli).[7] We can estimate an approximation of this result using ClinEpiDB.

---

[7] J. Liu, J.A. Platts-Mills, J. Juma, et al. 2016. Lancet. Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study.

Examining the difference between conventional and quantitative molecular diagnostics of ST-ETEC in GEMS

For ST-ETEC in GEMS1, how many participants overall tested positive by conventional microbiologic methods? (ETEC ST-pos, by PCR) _____ Proportion: _____ [8]

Compare this to, how many participants overall tested positive by TAC? (ETEC ST-pos Ct value, by TAC result <35) _____ Proportion: _____ [9]

Remember that if a diagnostic test is more sensitive it may increase detection in diarrheal disease cases *but also* controls. The increased detection will not result in an impact to attributable incidence *unless* it is differential in cases versus controls.

Cases: ETEC ST-pos, by PCR _____   Proportion _____ [10]

Controls: ETEC ST-pos, by PCR _____   Proportion _____ [11]

Difference in Case vs. Control Proportion: _____ [12]

Cases: ETEC ST-pos Ct value, by TAC result <35) _____   Proportion _____ [13]

Controls: ETEC ST-pos Ct value, by TAC result <35) _____   Proportion _____ [14]

Difference in Case vs. Control Proportion: _____ [15]

Which detection method had a greater difference in the proportion positive in cases vs. controls? [16]

Finally, look at what happens to the conventional test results when you vary the threshold for positivity by Ct value. The advantage of a quantitative test is that it may give you more information about pathogen load and/or disease severity. We might hypothesize that detecting a greater quantity of DNA in the stool might make it more likely to detect in the binary method. Try this for ST ETEC and Shigella. Do you notice any differences? What about for severity? We know that Shigella is associated with blood in stool; does a lower Ct value correspond to a higher likelihood of blood being detected in the stool?

[Hint: You can find the identification of blood in the stool under Sample>Stool characteristics]

While you cannot do a direct comparison to the MAL-ED study because of the difference in study design and distinct case definition for GEMS, take these next few minutes to explore MAL-ED in the

---

[8] 1,009, 4%

[9] 1,957, 8.7%

[10] 645, 6.8%

[11] 364, 2.8%

[12] 4.0%

[13] 1,127, 11.9%

[14] 830, 6.3%

[15] 5.6%

[16] TAC

same way? Do the results of the conventional detection method for ST-ETEC differ from the TAC results?

Exercise 2. Building a Search with Related Cases or Controls (15 min)

In this exercise, you will learn to do a more advanced search using our related case/control function. One of the unique assets of the GEMS1 and GEMS1A data is the matched set design and the ability to search these data easily is extremely valuable.

We will start with a very simple related case/control query. For this exercise, to start, we will modify as few filters as possible to keep things simple but you can always increase complexity as you refine your research questions. One question you might want to ask of the many case control pairs is **"How many Giardia positive cases had *at least* one matched control that was Giardia negative?"**

To explore, your first step might be to review the data and see how many Cases tested positive for *Giardia*. Then you would ask the *Giardia* status of those particular Cases matched Controls. You can do this by using the **related Case/Control** functionality.

1. Start this exercise at the Participant level in the search wizard. We describe this functionality as a 'Related Case/Control search because you are narrowing your dataset based on the relationship between your Case and Control participants. First select Cases who tested positive for *Giardia*.
2. In the Choose Case/Control box in the Search Wizard, select the 'Cases' from the drop-down.

3. Next, select from Observations that *Giardia* positive. This can be found under 'Stool microbiology test' You can navigate to this through the filter hierarchy on the left or by typing "stool" into the search box.
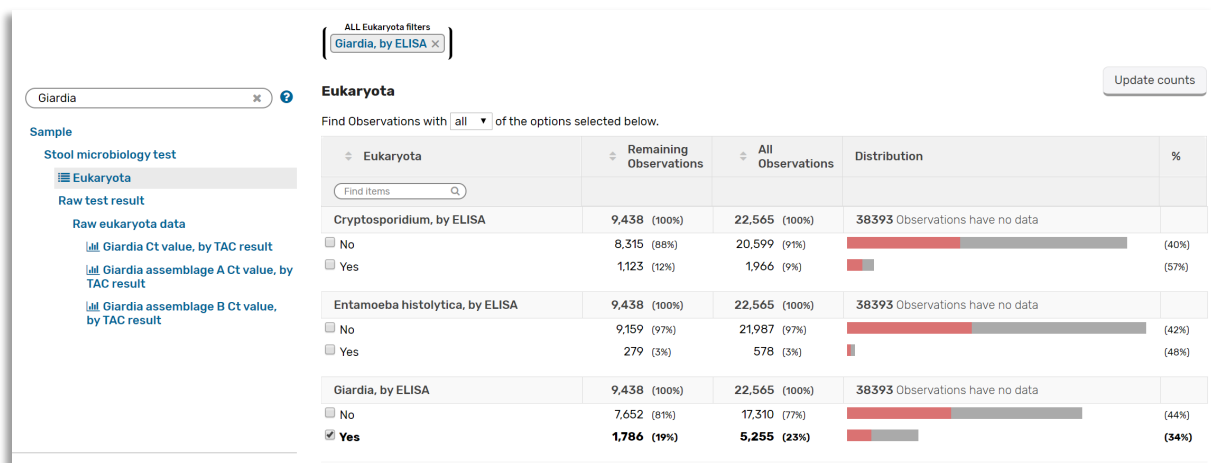
You should notice that the number of participants reduced to 1,786. This means that 1,786 Cases had a positive test for *Giardia*.

4. Now, imagine that of these *Giardia* positive Cases, how many had at least one matched Control that was Giardia negative. To do this, you would use the 'Related Case/Control box in the query wizard.

5. Click on 'Related Case/Control'
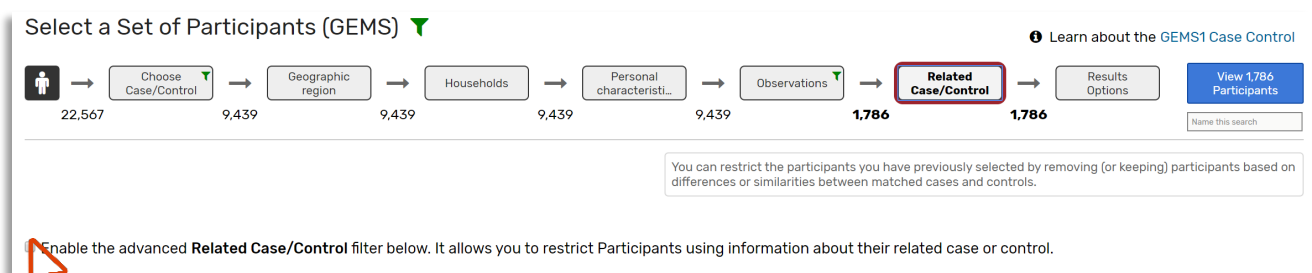
6. Then, click on the small check-box next to the sentence 'Enable the Related observations filter below. Enabling this option will allow you to restrict Observations by relating them to your choice of Related observations.'

Think carefully about what we want to do. We want to keep the selected Cases that have a Control that is Giardia negative.
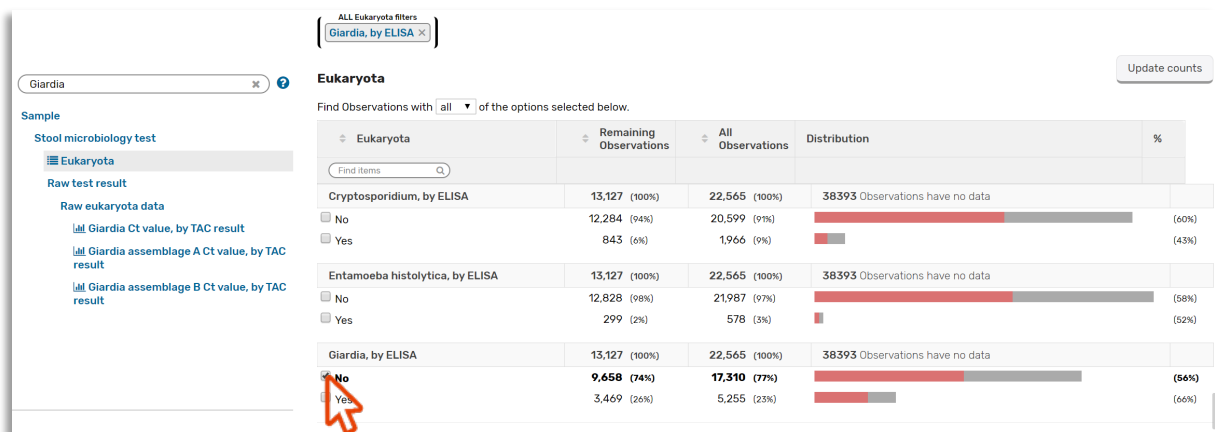
Complete the sentence like this:

7. Then, select *Giardia* negative from the Stool microbiology test variable

How does this impact the number of participants that match your search? There should only be 1,259 participants remaining.

8. Click on the 'View 1,259 Participants' box to see the table listing of these Participants. Click on the small histogram icons next to the column names to see the distribution of characteristics in this sample.

To change the default columns listed in the result click on the 'Add Columns' button located at the top right corner of your Participant result tab and add the desired columns. Try adding 'Dysentery' to the columns

9. Click on the 'Participant ID' 1010001181 to see this Participant Record page and view the data in a little more detail. See if you can find where it says they are Giardia positive.

GEMS1 CC Participant: 1010001181

Case or Control: Case
Country: The Gambia
Biological Sex: Female
Age at Eligibility Form Completion: 9 months

1 Public health and epidemiology

▼ **Observations**

Showing 2 rows

⇅ name	⇅ Observation type
1010001181_2008-02-13	60 day follow-up
1010001181_2007-12-10	Enrollment

▶ **Participant Characteristics**

▼ **Stool Microbiology Test**

Showing 62 rows

⇅ Sample ID	⇅ Observation Type	⇅ Stool Microbiology Test	⇅ Result
100015	Enrollment	Adenovirus 40/41, by ELISA	No
100015	Enrollment	Adenovirus non-40/41, by ELISA	No
100015	Enrollment	Rotavirus, by ELISA	No
100015	Enrollment	Cryptosporidium, by ELISA	No
100015	Enrollment	Norovirus, by RT-PCR	Yes