

Where programme data comes from

The shape of exports from CommCare, KoboToolbox and DHIS2, why each one fights you in a different way, and the questions to ask before you write a line of code.

Pierre Robentz Cassion

Lesson 1 of 8

Data Analysis Foundations for M&E

What this lesson covers

- The three exports you will meet
- The unit of observation decides everything
- What a real export looks like
- Mid-upper arm circumference, in one paragraph
- The three questions to ask the person who sent you the file
- What comes next

The three exports you will meet

- CommCare — exports one row per form submission, with repeat groups flattened into numbered columns (`child_1_muac,...`
- KoboToolbox — exports close to the XLSForm structure, with repeats in separate sheets joined by `_index` and...
- DHIS2 — exports aggregate data — already summed by period and org unit — so the question is rarely "how do I clean..."

The unit of observation decides everything

Export	One row is	The trap
CommCare form dump	One form submission	A form may cover several children
KoboToolbox main sheet	One interview	Members live in the repeat sheet
KoboToolbox repeat sheet	One member	No household context until you join
DHIS2 pivot	One org unit $\times$ period $\times$ data element	Individuals are gone for good

The unit of observation decides everything

Ask the question out loud before you open the file, and write the answer at the top of your script as a comment. Most of the disputes later in this course trace back to two people holding different answers to it.

What a real export looks like — Example

```
child_id,commune,screening_date,age_months,sex,muac_mm,oedema,outcome  
CH00854,Terre-Neuve,2024-01-15,22,m,133,false,no-action  
CH01439,Verrettes,2024-01-15,16,f,143,false,no-action
```

Mid-upper arm circumference, in one paragraph

- Below 115 mm — severe acute malnutrition (SAM)
- 115 mm to under 125 mm — moderate acute malnutrition (MAM)
- 125 mm and above — no acute malnutrition by this measure

The three questions to ask the person who sent you the file

- **What does a blank mean here?** Not measured, measured as zero, refused, or the tablet crashed? These are four. . .
- **Has anything already been cleaned or dropped?** If someone removed the rows they thought were wrong, your row count. . .
- **What period does this cover, and by which date?** Date of screening, date of form submission and date of sync can. . .

What comes next

- The next lesson sets up a working environment in Python and in R — deliberately one that works offline, because much of this audience is on a connection that cannot be relied on to install a package at the moment it is needed.

Read the full lesson, with runnable code

[https://data-analysis.cassion.dev/courses/data-analysis-foundations/
foundations-01-programme-data](https://data-analysis.cassion.dev/courses/data-analysis-foundations/foundations-01-programme-data)