The sentence this analysis is entitled to

Nord's case fatality is 6.30% against Centre's 2.00%, and the delay gradient rests on one death among thirty cases. Four claims could be written from that, three of them are wrong, and the difference between them is the whole of this course.

Pierre Robentz Cassion

Lesson 8 of 8

What this lesson covers

- Four sentences, one set of numbers
- Why the first three fail
- The four things a claim needs
- Strength of evidence, and where routine data sits
- Write the limitation as a decision, not an apology
- The habit this course was for
- What comes next

Four sentences, one set of numbers

- *"Nord's cholera response is failing: case fatality there is three times Centre's."*
- *"Late presentation causes death — cases admitted four or more days after onset die at 3.3% against 1.9% for those...
- *"Case fatality in Nord is 6.30% (24 deaths among 381 cases with a recorded outcome) against 2.00% in Centre (8 of...
- *"The data cannot support a comparison between districts."*
- The third is the only defensible one, and it is also the only one long enough to be useful — Work out why each of the...

Why the first three fail

- "Nord's response is failing" — attributes a difference to a cause the data does not distinguish
- "Late presentation causes death" — is a stronger claim than the design supports, and it is built on two cells that...

Why the first three fail — In Python

```
import pandas as pd

cases = pd.read_csv("cholera-line-list-2024.v1.csv", parse_dates=["onset_date"])
with_onset = cases[cases["onset_date"].notna()]

delayed = with_onset.assign(delay=lambda d: (
    pd.to_datetime(d["admission_date"]) - d["onset_date"]).dt.days)

print(delayed.groupby("district")["delay"]
      .agg(median="median", same_day=lambda s: (s == 0).mean()).round(3))

band = pd.cut(delayed["delay"], [-1, 1, 3, 99], labels=["0-1", "2-3", "4+"])
print(delayed.dropna(subset=["outcome"]).groupby(band).size())
```

Why the first three fail — In R

```r
library(dplyr)

cases |>
  filter(!is.na(onset_date)) |>
  mutate(delay = as.numeric(admission_date - onset_date)) |>
  summarise(median = median(delay), same_day = mean(delay == 0), .by = district)
```

Why the first three fail

- "The data cannot support a comparison" — is the failure mode of the cautious analyst and it is the most expensive of...

The four things a claim needs

- The numerator and the denominator, as counts — "6.30%" is unauditable; "24 deaths among 381 cases with a recorded...
- The comparison it is against — A rate with no reference point is a number looking for an opinion
- The adjustment, or its absence — If you standardised, name the standard
- The limitation that could change the direction — Not a paragraph of hedging — the one or two things that, if they went...

The four things a claim needs — In Python

```
claim = {
    "numerator": 24,
    "denominator": 381,
    "denominator_definition": "cases with a recorded outcome",
    "value": "6.30%",
    "benchmark": "Sphere: below 1%",
    "adjustment": "none; crude",
    "limitation": "onset dates back-filled in this district",
}
```

The four things a claim needs — In R

```
# The same fields, in the footnote of the table.
```

Strength of evidence, and where routine data sits

Design	What it can claim	Where it appears here
Routine data, cross-section	Association, adjusted for what was measured	This whole course
Cohort	Association with the exposure preceding the outcome	The CMAM register, followed forward
Case-control	Association, efficiently for rare outcomes	Investigating the source of an outbreak
Randomised trial	Causation	Almost never in programme M&E

Strength of evidence, and where routine data sits

- Programme data lives on the top row — and the top row can carry a great deal of weight if it is honest about which row…

Write the limitation as a decision, not an apology — Example (cont.)

```
Cholera case fatality, weeks 1-16

  Overall      3.90%   38 deaths / 974 cases with an outcome
  Nord         6.30%   24 / 381
  Sud          3.11%    6 / 193
  Centre       2.00%    8 / 400

  Against the Sphere threshold of 1%, every district is above standard.

  What this supports: prioritising treatment capacity in Nord.
  What it does not: attributing the gap to clinical management or to late
  presentation. Nord's register back-fills onset dates, so its delays are not
  comparable, and the 4+ day band holds one death among 30 cases. Water
  source and crowding, the main determinants of cholera severity, are not
  recorded at all.
  What would settle it: two weeks of prospectively recorded onset dates in
```

Write the limitation as a decision, not an apology — Example (cont.)

```
Nord, and the WASH assessment already scheduled for week 18.
```

Write the limitation as a decision, not an apology

- "What would settle it" is the line that makes the limitation useful — A limitation with no route out reads as an...

The habit this course was for

- **What did I count, over what?** Numerator, denominator, and who was excluded.
- **What else could explain it?** The table from lesson 7, written before the analysis rather than after.
- **What should someone do differently because of this?** If nothing, the analysis was not worth the time, and if...

What comes next

- You can defend a rate, adjust a comparison and state what it is entitled to claim.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/public-health-epidemiology/epi-08-what-observation-can-claim