The curve is not the outbreak

975 cases, 47 without an onset date, and one district whose register is back-filled from the admission book — so its median onset-to-admission delay reads as zero while its case fatality is the highest of the three.

Pierre Robentz Cassion

Lesson 3 of 8

Public Health and Applied Epidemiology — An outbreak, one case at a time

What this lesson covers

- What the curve is for
- Draw it
- What the missing dates do
- The district whose curve is not a curve
- Reading the shape
- Plot cases, not rates, on the curve
- What comes next

What the curve is for

- An epidemic curve is cases by date of onset.

Draw it — 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()]
print(f"{len(with_onset)} of {len(cases)} cases have an onset date")

curve = (
    with_onset.assign(week=with_onset["onset_date"].dt.to_period("W").dt.start_time)
    .groupby("week").size()
)
print(curve)
print(f"peak week {curve.idxmax().date()} with {curve.max()} cases")
```

Draw it — In R

```
library(dplyr)

cases <- readr::read_csv("cholera-line-list-2024.v1.csv")

curve <- cases |>
  filter(!is.na(onset_date)) |>
  mutate(week = lubridate::floor_date(onset_date, "week")) |>
  count(week)
```

What the missing dates do — In Python

```
missing = cases[cases["onset_date"].isna()]
print(missing.groupby("district").size())
print(f"{len(missing) / len(cases):.1%} of all cases")
```

What the missing dates do — In R

```
cases |> filter(is.na(onset_date)) |> count(district)
```

What the missing dates do

- Exclude them — The default and usually right
- Impute from admission date — Tempting and it shifts the whole curve later by the median delay, which flattens the rise...
- Plot them as a separate band — Honest and rarely done: a stacked bar with "onset unknown" in a distinct shade, placed...

What the missing dates do — In Python

```
imputed = cases.copy()
fallback = pd.to_datetime(imputed["admission_date"], errors="coerce")
imputed["onset_or_admission"] = imputed["onset_date"].fillna(fallback)

both = pd.DataFrame({
    "excluded": curve,
    "imputed": (imputed.assign(
        week=imputed["onset_or_admission"].dt.to_period("W").dt.start_time)
        .groupby("week").size()),
}).fillna(0).astype(int)
print(both)
```

What the missing dates do — In R

```
cases |>
  mutate(onset_or_admission = coalesce(onset_date, admission_date),
         week = lubridate::floor_date(onset_or_admission, "week")) |>
  count(week)
```

What the missing dates do

- Whichever you choose, state the denominator on the chart — "n = 928 of 975 cases with a recorded onset date" is six...

The district whose curve is not a curve — In Python

```
delay = (
    pd.to_datetime(cases["admission_date"], errors="coerce") - cases["onset_date"]
).dt.days

by_district = cases.assign(delay=delay).dropna(subset=["delay"]).groupby("district")
print(by_district["delay"].median())
print((by_district["delay"].apply(lambda s: (s == 0).mean()) * 100).round(0))
```

The district whose curve is not a curve — In R

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

The district whose curve is not a curve

District	Median delay	Same-day
Nord	**0 days**	**80%**
Centre	2 days	13%
Sud	2 days	20%

The district whose curve is not a curve

- Nord's cases appear to reach treatment the same day they fall ill, four times as often as anywhere else — That is not a...
- Nord's epidemic curve is an admission curve — It is shifted right by however long its cases actually took to arrive, so...
- Nord's delay statistic is unusable — and it is the statistic the next lesson uses to explain case fatality — where Nord...

The district whose curve is not a curve — In Python

```
print("Nord: median delay 0 days, case fatality 6.30%")
print("Centre: median delay 2 days, case fatality 2.00%")
```

The district whose curve is not a curve — In R

```
# The contradiction is the finding: the fastest-looking district dies most.
```

The district whose curve is not a curve

- A contradiction between two indicators from the same register is a data question before it is an epidemiological one —. . .

Reading the shape

- **A single sharp peak** suggests a point source — one contaminated water point, one event — with cases appearing within...
- **A series of peaks at roughly one incubation period apart** suggests person-to-person propagation.
- **A long plateau** suggests continuing common-source exposure, which for cholera usually means the water supply has not...

Plot cases, not rates, on the curve

- A last practical point.

What comes next

- The curve says when.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/public-health-epidemiology/epi-03-epidemic-curves