What else could explain it

A confounder causes the outcome and differs between the groups. Four of them sit between the districts in this outbreak, one has been removed, one is unmeasured, and one is not a confounder at all — it is the mechanism.

Pierre Robentz Cassion

Lesson 7 of 8

Public Health and Applied Epidemiology — What the comparison can claim

What this lesson covers

- The definition, and why it is worth being strict about
- The four candidates in this outbreak
- Stratify to see it
- When stratification runs out
- The three questions to ask of any comparison
- What comes next

The definition, and why it is worth being strict about

- it is **associated with the exposure** — it differs between the groups being compared;
- it is a **cause of the outcome**, independently of the exposure;
- it is **not on the causal path** between the exposure and the outcome.

The four candidates in this outbreak

Candidate	Differs between districts?	Causes the outcome?	On the path?
Age structure	Yes, sharply	Yes	No — a confounder, already removed
Water source and density	Almost certainly	Yes	No — a confounder, **unmeasured**
Distance to treatment	Yes	Yes, for fatality	**Yes** — it is the mechanism
Register quality	Yes	No	No — not a confounder, a bias

The four candidates in this outbreak

- Age structure — has been handled
- Water source, sanitation and crowding — are the actual causes of cholera transmission, they certainly differ between a...

The four candidates in this outbreak — In Python

```
print("Available for adjustment: age, sex, district")
print("Known to matter and unavailable: water source, sanitation, crowding, "
      "displacement status")
```

The four candidates in this outbreak — In R

```
# The list of what you could not adjust for belongs in the limitations section.
```

The four candidates in this outbreak

- Distance to treatment — is the interesting one and it is not a confounder at all
- Register quality — is not a confounder either

Stratify to see it — In Python

```python
import pandas as pd

cases = pd.read_csv("cholera-line-list-2024.v1.csv")
with_outcome = cases[cases["outcome"].notna()]

stratified = (
    with_outcome.assign(died=with_outcome["outcome"] == "died")
    .groupby(["district", "age_band"])
    .agg(cases=("died", "size"), deaths=("died", "sum"))
)
stratified["cfr"] = stratified["deaths"] / stratified["cases"]
print((stratified["cfr"].unstack() * 100).round(1))
```

Stratify to see it — In R

```
library(dplyr)

cases |>
  filter(!is.na(outcome)) |>
  summarise(cases = n(), cfr = mean(outcome == "died"), .by = c(district, age_band)) |>
  tidyr::pivot_wider(id_cols = age_band, names_from = district, values_from = cfr)
```

Stratify to see it

- Does the difference persist within strata? — If Nord's case fatality is higher in every age band, age is not the...
- Are the strata consistent? — A difference that is large in one band and absent in others is effect modification — the...
- Effect modification is a finding; confounding is a nuisance — Collapsing the first into a single number destroys the...

When stratification runs out — In Python

```
cells = with_outcome.groupby(["district", "age_band", "sex"]).size()
print(f"{len(cells)} cells, smallest {cells.min()}, {(cells < 30).sum()} below 30")
```

When stratification runs out — In R

```
cases |> filter(!is.na(outcome)) |> count(district, age_band, sex) |>
  summarise(cells = n(), smallest = min(n), under_30 = sum(n < 30))
```

The three questions to ask of any comparison

- 1. What differs between these groups other than the thing I am studying? — List them
- 2. Which of them cause the outcome? — Only those are confounders
- 3. Which of them are on the causal path? — Do not adjust for those

The three questions to ask of any comparison — In Python

```
adjustment_plan = pd.DataFrame({
    "variable": ["age", "sex", "water source", "delay to treatment", "register quality"],
    "differs": [True, False, True, True, True],
    "causes_outcome": [True, True, True, True, False],
    "on_causal_path": [False, False, False, True, False],
    "action": ["adjust", "no need", "unmeasured; state it", "do not adjust",
               "fix the register"],
})
print(adjustment_plan)
```

The three questions to ask of any comparison — In R

```
tibble::tribble(
  ~variable,        ~action,
  "age",            "adjust",
  "water source",   "unmeasured; state it",
  "delay",          "do not adjust; it is the mechanism",
  "register",       "fix it; this is bias, not confounding"
)
```

The three questions to ask of any comparison

- Write that table before the analysis, not after — Deciding what to adjust for once you have seen which adjustment gives. . .

What comes next

- You have named what else could explain a difference and removed what you could.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/public-health-epidemiology/epi-07-confounding