Retrospective Studies: Case-Control Designs
MATH 348, Vassar College, Spring 2024

Numb3rs S3E9, https://www.hulu.com/numb3rs

Prof. Lee Kennedy-Shaffer
April 15, 2024

Outline

Cancer Cluster

We notice in a school that there are many children with pediatric cancers.

What are some relevant questions of scientific interest?

Two Key Questions

1. How do we know if this is a cluster?

Two Key Questions

1. How do we know if this is a cluster?
2. How do we identify a cause?

Outline

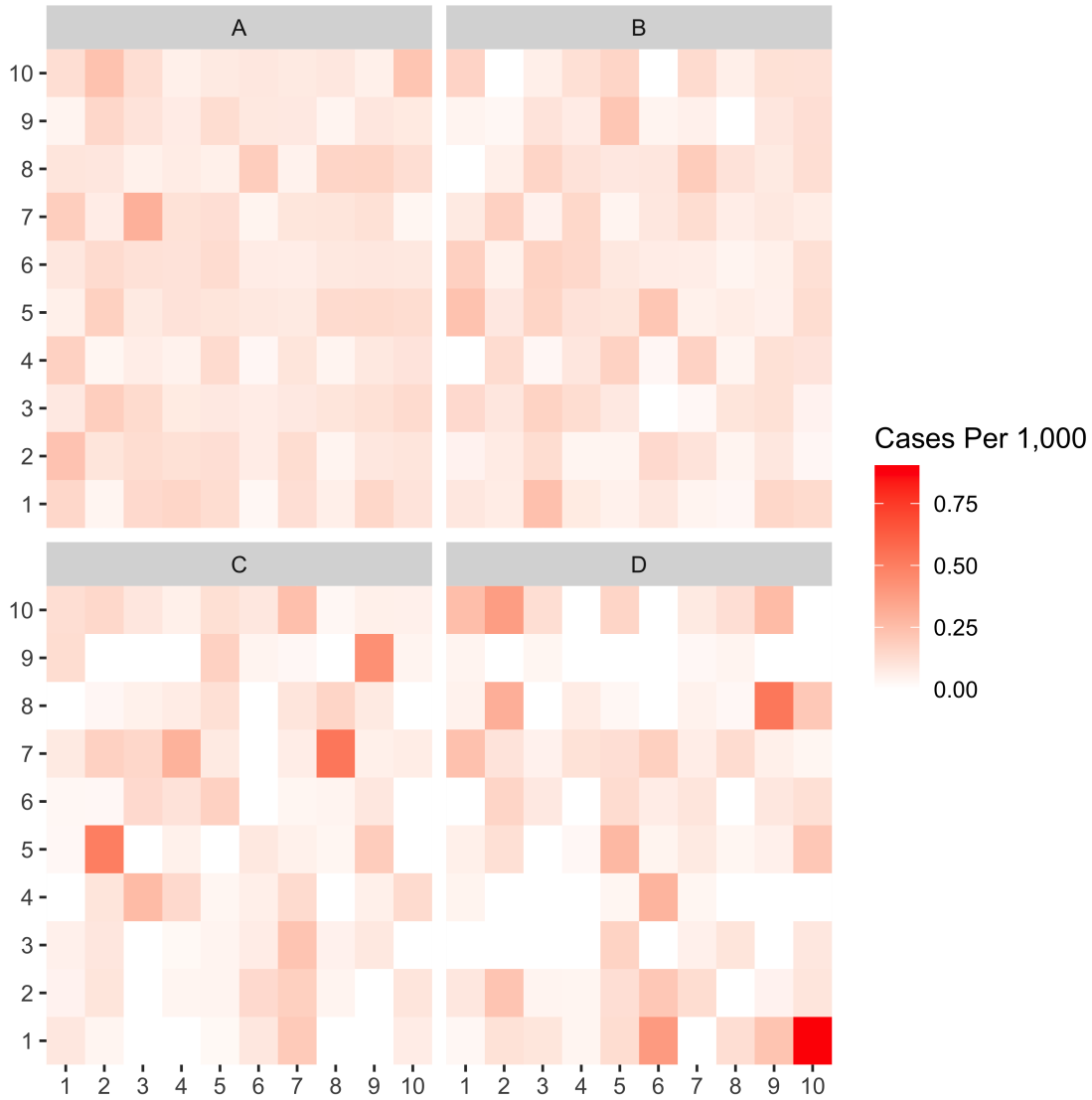
How Many is Too Many?

How Many is Too Many?

Different Question: How many is abnormal/surprising?

What statistical methods do we have to assess something like this?

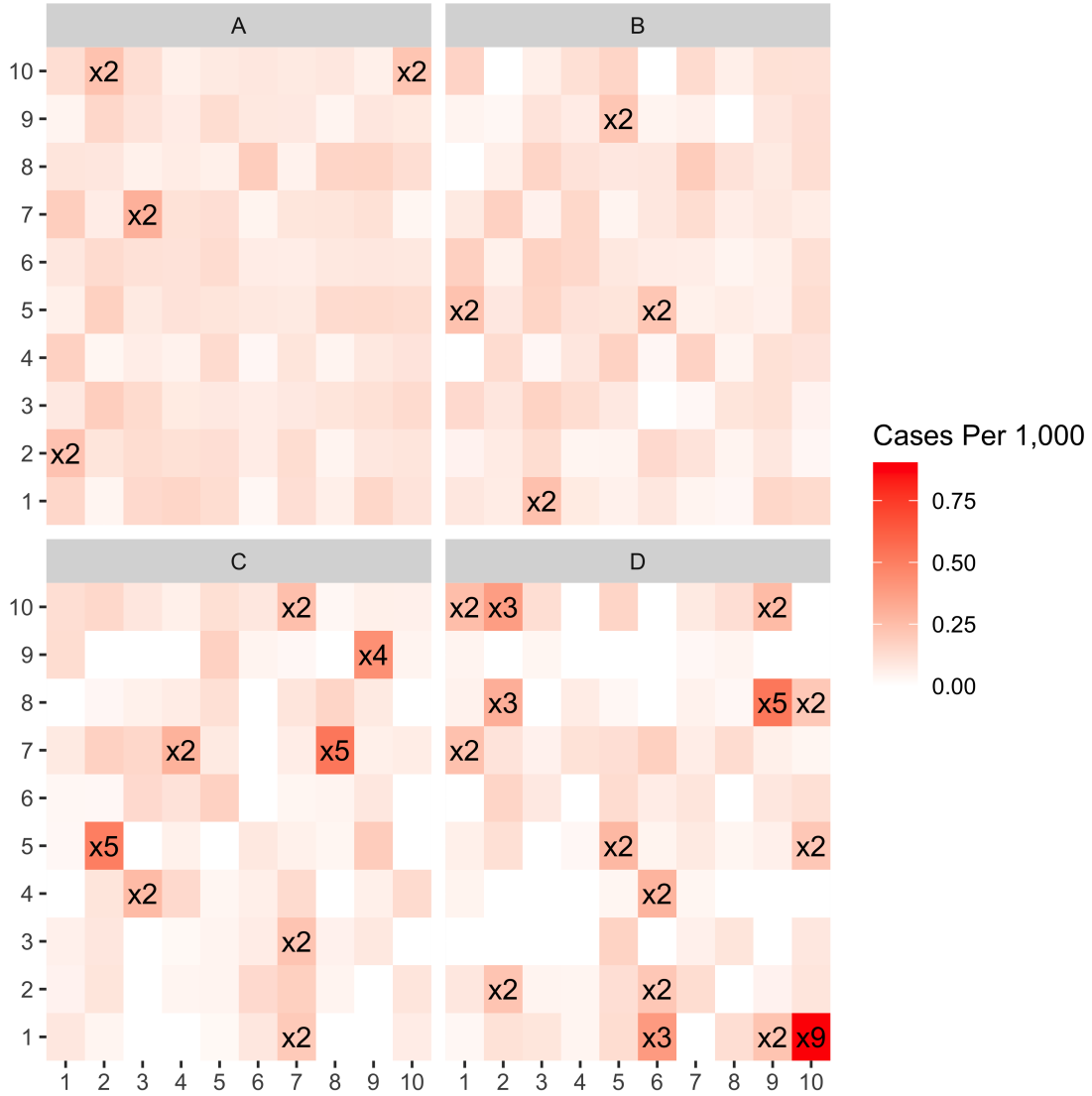
Clustering



100 communities with an overall rate of outcomes of 0.1 per 1,000

Which of these scenarios exhibit unusual clustering?

Clustering



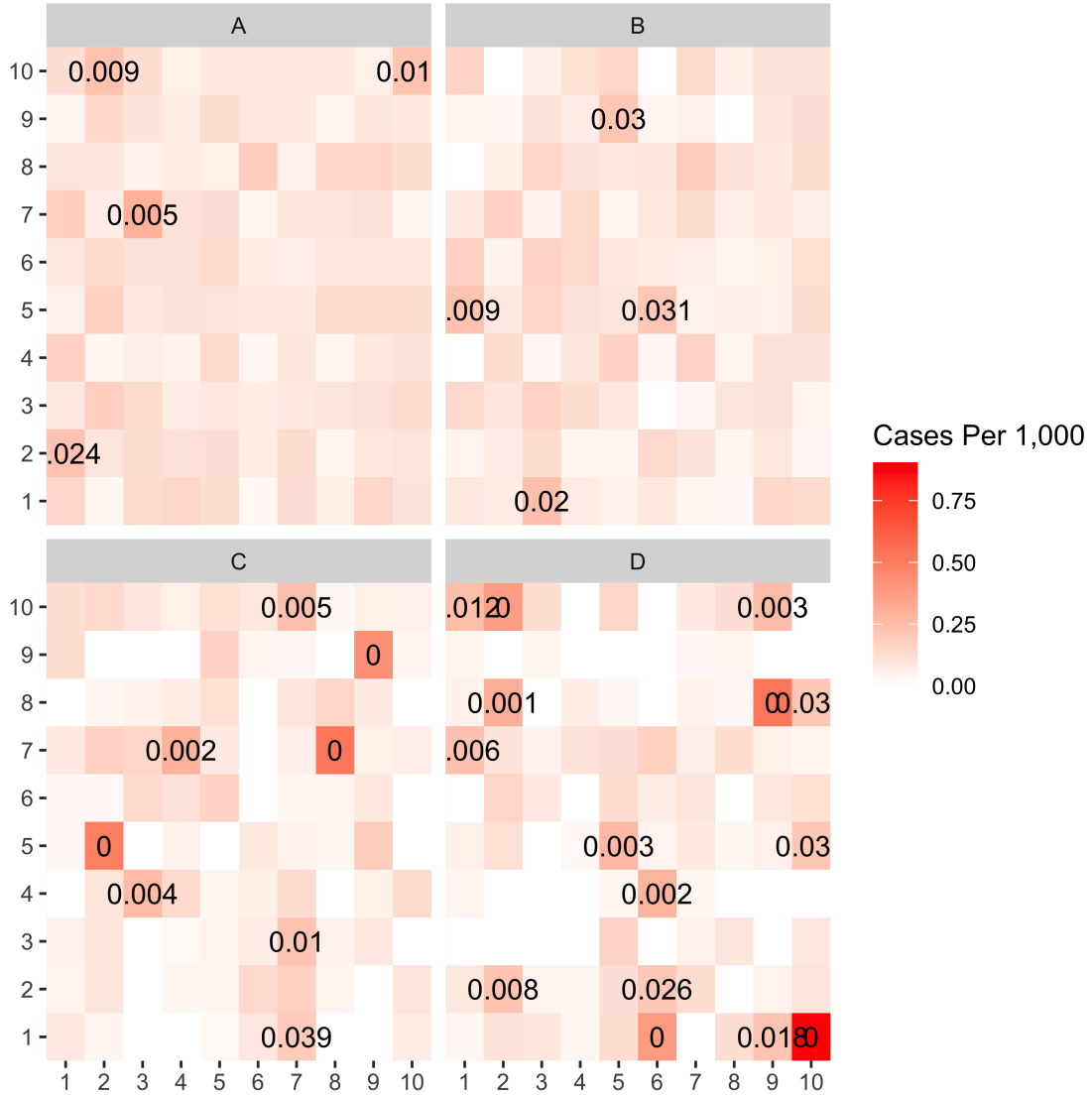
Identifying by multiplier over base rate can yield clusters in many cases

Hypothesis Testing on Clusters

We can put this into a formal hypothesis testing framework: given the overall background rate, how surprising is it for a cluster to have some number of cases?

What assumptions are required for this test?

Clustering



P-values for hypothesis tests comparing to a Poisson distribution with a single common base rate

Balancing Type I and Type II Errors

Errors

- What is the danger of a Type I Error (setting a high significance level)?
- What is the danger of a Type II Error (low power)?

Texas Sharpshooter Fallacy

In a lot of cases, there are **many possible ways of defining clusters**: people are in towns, counties, schools, workplaces, etc.

Texas Sharpshooter Fallacy

In a lot of cases, there are **many possible ways of defining clusters**: people are in towns, counties, schools, workplaces, etc.

If we keep looking until we find a definition that gives us a significant level, we are committing a fallacy (or doing a type of **multiple testing**). This is bad! We'll find spurious results.

Triangulating Evidence

Avoiding Fallacies

Ideally we would pre-specify our analyses, like we did in other studies. But if we don't even know where to look until we start seeing outcomes, that's not possible!

Instead, we need to rely on **multiple threads of evidence**.

Outline

Identifying Exposures

Once we have identified a cluster, we want to turn to question 2:

How do we identify a cause?

Identifying Exposures

Once we have identified a cluster, we want to turn to question 2:

How do we identify a cause?

We can generate possible causes with background information and scientific knowledge. Now we want to confirm a causal relationship.

Question

Can we do any randomized trials?

Observational Studies

What confounding factors might arise if we have a geographically-defined cluster?

Observational Studies

Feasibility

How long do we need to wait to see the effects?

Will we ever have enough outcomes to detect an effect?

Case-Control Design

1. Identify cases with the outcome of interest

Case-Control Design

1. Identify cases with the outcome of interest
2. Determine the **population of interest**: which people would be included in the cases if they had the outcome of interest? May be defined geographically, by demographic factors, by workplace/school/etc., or by some other factor(s)

Case-Control Design

1. Identify cases with the outcome of interest
2. Determine the **population of interest**: which people would be included in the cases if they had the outcome of interest? May be defined geographically, by demographic factors, by workplace/school/etc., or by some other factor(s)
3. Identify a sample of the population of interest who do **not** have the outcome of interest (controls)

Case-Control Design

1. Identify cases with the outcome of interest
2. Determine the **population of interest**: which people would be included in the cases if they had the outcome of interest? May be defined geographically, by demographic factors, by workplace/school/etc., or by some other factor(s)
3. Identify a sample of the population of interest who do **not** have the outcome of interest (controls)
4. Survey the cases and controls to determine the exposure status of all individuals

Case-Control Design

1. Identify cases with the outcome of interest
2. Determine the **population of interest**: which people would be included in the cases if they had the outcome of interest? May be defined geographically, by demographic factors, by workplace/school/etc., or by some other factor(s)
3. Identify a sample of the population of interest who do **not** have the outcome of interest (controls)
4. Survey the cases and controls to determine the exposure status of all individuals
5. Compare the rates of exposure in the cases group to the rate in the controls group

Case-Control Design

Key Challenge

The control group must be at risk of the outcome of interest AND be chosen independently of their exposure status.

We need cases and controls who both **could** have the exposure of interest.

School Cancer Cluster Example

Suppose in our example we identify a toxic contaminant in one area that sends children to the school. How would we define our **population of interest** and **sample controls**?

Results from Case-Control Study

We get a 2-by-2 table:

		Outcome Status	
		Case ($D+$)	Control ($D-$)
Exposure Status	Exposed ($E+$)	A	B
	Unexposed ($E-$)	C	D

Odds and Odds Ratios

Recall: Odds

The odds of an event Y are equal to $P(Y)/[1 - P(Y)] = P(Y)/P(Y^C)$.

Odds and Odds Ratios

Recall: Odds

The odds of an event Y are equal to $P(Y)/[1-P(Y)] = P(Y)/P(Y^C)$.

Definition: Odds Ratio

The odds ratio of an event Y with respect to another event X are:

$$OR_{Y|X} = \frac{Odds(Y|X)}{Odds(Y|X^C)}$$
$$= \frac{P(Y|X)/[1-P(Y|X)]}{P(Y|X^C)/[1-P(Y|X^C]} = \frac{P(Y|X)/P(Y^C|X)}{P(Y|X^C)/P(Y^C|X^C)}.$$

Inverting the Odds Ratio

From the case-control study, we can get the probability of the exposure given one's outcome status: $P(E+|D+)$ and $P(E+|D-)$.

What we really want is how the exposure affects the probability of getting the outcome: $P(D+|E+)$ vs. $P(D+|E-)$.

Variance from Case-Control Study

We can get a useful approximation of the variance as well:

Bias: Handling Confounding

Residual Confounding

We chose our populations to be as similar as possible. But there could still be confounding variables that cause both the outcome and the exposure.

Bias: Handling Confounding

Residual Confounding

We chose our populations to be as similar as possible. But there could still be confounding variables that cause both the outcome and the exposure.

Methods for handling confounding:

1. Adjust in regression
2. Stratify or match

Regression Adjustment

If we're adjusting for confounders $\boldsymbol{L}$ in a regression, we can use logistic regression:

$$\text{logit}(P[D = 1|E, \boldsymbol{L}]) = \beta_0 + \beta_1 E + \boldsymbol{\beta L}$$

Then $\hat{\beta}_1$ is our estimate of the odds ratio between exposure and outcome, adjusted for the confounders.

Regression Adjustment

If we're adjusting for confounders $\boldsymbol{L}$ in a regression, we can use logistic regression:

$$\text{logit}(P[D = 1|E, \boldsymbol{L}]) = \beta_0 + \beta_1 E + \boldsymbol{\beta L}$$

Then $\hat{\beta}_1$ is our estimate of the odds ratio between exposure and outcome, adjusted for the confounders.

1. Can still invert the odds ratio. Why?
2. Warning: Simpson's Paradox

Matched Case-Control Study

In a matched case-control study, for each case, we choose one (or more) controls who are the same on as many relevant dimensions as possible, except outcome status. The choices again have to be **independent of exposure status**.

Matched Case-Control Study

		Cases	
		Exposed	Unexposed
Controls	Exposed	A' (Concordant)	B' (Discordant)
	Unexposed	C' (Discordant)	D' (Concordant)

Matched Case-Control Study

		Cases	
		Exposed	Unexposed
Controls	Exposed	A' (Concordant)	B' (Discordant)
	Unexposed	C' (Discordant)	D' (Concordant)

Question

Which cells give us information about the odds ratio of interest?

Interpreting Results

From any case-control study, we estimate an **odds ratio for the outcome comparing the different levels of the exposure**.

Note: the odds ratio is **not** a relative risk or rate ratio. But if the outcome is **rare**, they are approximately equal.

Interpreting Results

From any case-control study, we estimate an **odds ratio for the outcome comparing the different levels of the exposure**.

Note: the odds ratio is **not** a relative risk or rate ratio. But if the outcome is **rare**, they are approximately equal.

We **cannot** estimate the rate of the outcome in the population as a whole. *Why?*

Case-Control: Advantages and Limitations

Advantages:

- Relatively easy to conduct
- Get more cases this way
- Odds ratio is a relatively standard measure of comparison

Case-Control: Advantages and Limitations

Advantages:

- Relatively easy to conduct
- Get more cases this way
- Odds ratio is a relatively standard measure of comparison

Limitations:

- Standard observational study bias concerns
- Retrospective study: risk of recall bias
- Can be difficult to obtain cases truly independent of exposure
- Cannot estimate probabilities

Outline

For Wednesday

What to Think About

- What sampling, surveying, or research designs are used in the movie and the real court cases?
- What additional studies might would be helpful in arguing the case?
- What role can/should statistics and research studies have in the legal system?
- What are the limitations, downsides, or inconsistencies between statistical goals/research studies and legal principles?

Similar Movies and Cases

1. *A Civil Action* (1998) about Woburn, MA water contamination case. See also: *A Civil Action* book by Jonathan Harr, Lagakos et al. 1986 key study, Carleton Woburn Trial Collection, BU SPH Woburn Collection
2. *Dark Waters* (2019) about DuPont Teflon factory contamination in West Virginia. See also: *HuffPo* Article on Parkersburg, *Intercept* Article on Teflon and Dupont, *New York Times* Article on the Lawsuit, Scientific Studies on PFOA/PFAS