Old-school Data Science with Real Impact

Data Science 101

Stanford University, Department of Statistics

Today's Lecture

Example I

- Take a look back at one of the earliest data-science quest.
- Looking at past experiences highlights what makes for a "good" or "bad" data analysis.
- Quite surprising how many of the tools and insights from back in the time before computers we can still use today.
- We might have more data, new questions and need more tools, but there is no need to re-invent the wheel.

Example II

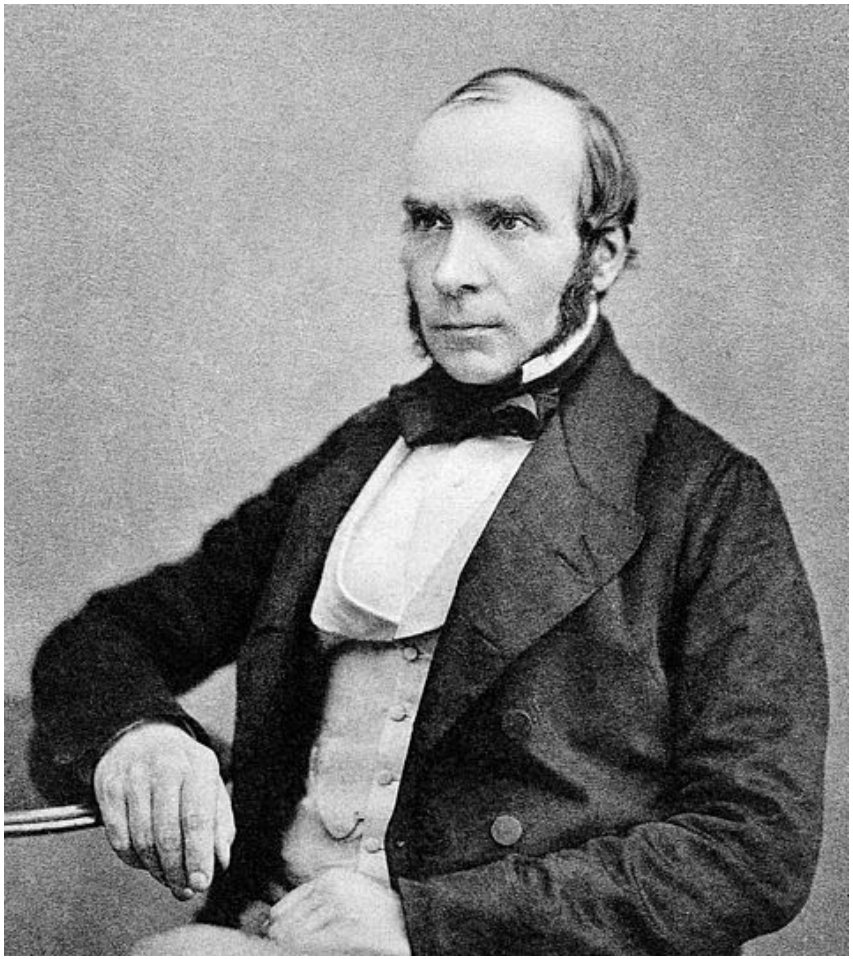
- A little bit more recent example.
- The actual code and analysis is "easy" to execute, but this does not mean that it is "easy" to come up with.

John Snow (1813–1858)

This is Jon Snow. He knew nothing about data analysis.

John Snow (1813–1858)

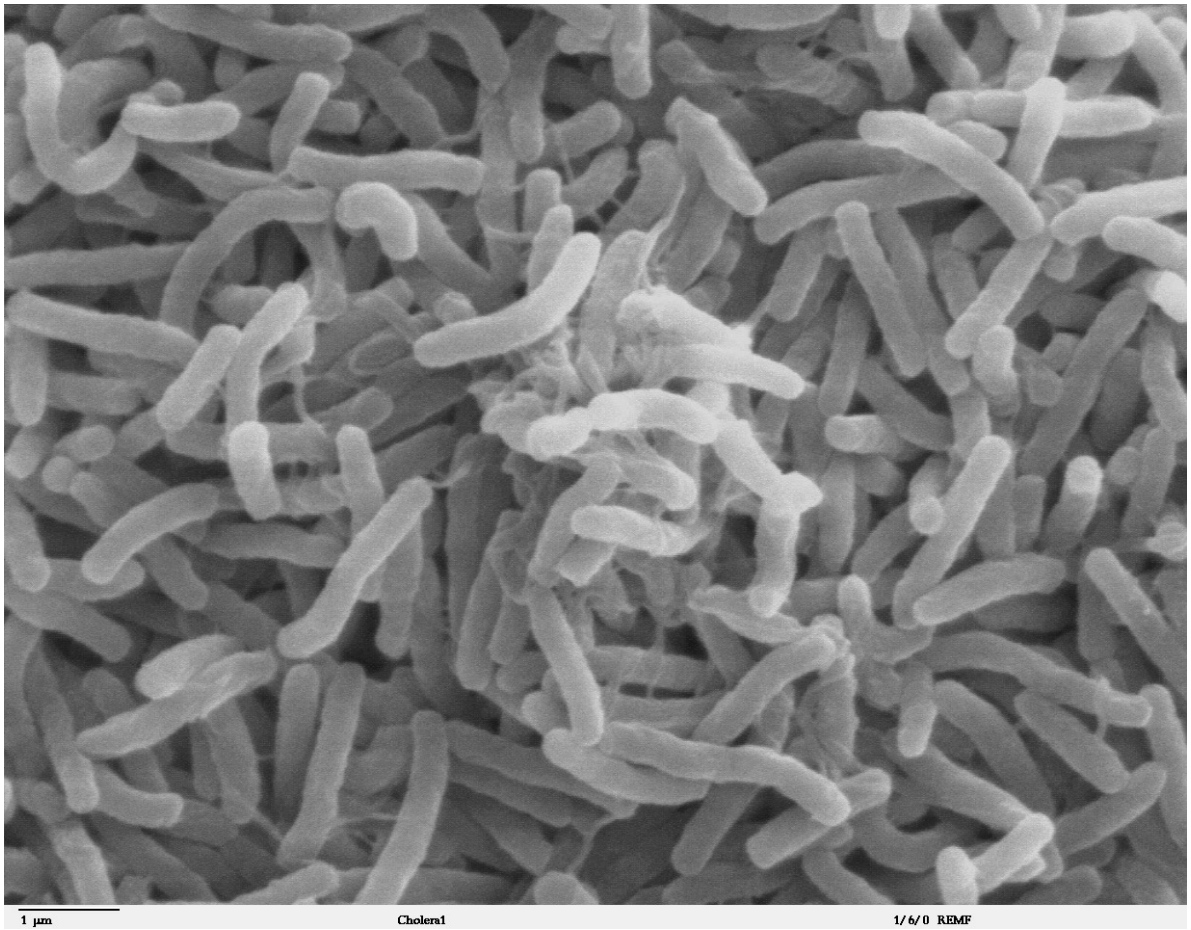
This is ~~Jon~~ John Snow. He knew ~~nothing~~ quite a bit about data analysis.



Cholera (WHO)

- Acute diarrhoeal disease that can kill within hours if left untreated
- Each year there are 1.3 to 4.0 million estimated cases of cholera, and 21,000 to 143,000 deaths worldwide due to cholera (more in the past – in Russia one million people died of cholera between 1847-1851)
- Most of those infected will have no or mild symptoms, and can be successfully treated with oral rehydration solution
- Severe cases will need rapid treatment with intravenous fluids and antibiotics
- Provision of safe water and sanitation is critical to control the transmission of cholera and other waterborne diseases
- But in the mid-1800s, little was known about Cholera, and there was controversy about what caused it and how it spread

Cholera bacteria (Vibrio cholera)



Filippo Pacini documented in 1854 that the cause of the infection was a bacterium, see here. Robert Koch also discovered the cholera bacillus about 30 years later, see here.

Controversy about what caused Cholera and how it was transmitted

- The cholera epidemics in England in the early-to-mid 1800s predated our modern understanding of the microbial etiology of disease
- The theory at the time was that "miasmas" in the air caused disease in general, and cholera in particular
- "miasmas" were associated with foul odor, and some associated them mainly with decaying animal matter
- John Snow was skeptical of the "miasma" theory
- In particular, he noticed that during cholera outbreaks, often an entire household would die of the disease while a neighboring household would be unaffected
- It was unlikely that the air they were breathing was very different, the two homes being right nearby. So the miasma theory was, in Snow's view, **inconsistent with the data**

Evidence from data

- After studying data from several cholera outbreaks, Snow proposed a theory that we now know to be correct: Cholera is caused by contaminated water; especially by water contaminated with human feces
- We'll now talk about how Snow reached his conclusions about the cause and transmission of Cholera
- He performed at least two important data analyses using data from the 1854 London Cholera epidemic
- The first was smaller scale and suggested an **association** between Cholera and drinking water
- The second was much larger scale and took the form of a "natural experiment"
- The second analysis came closer to formally establishing that Cholera is **caused** by drinking contaminated water

Outbreak in 1854

First analysis: the broad street pump

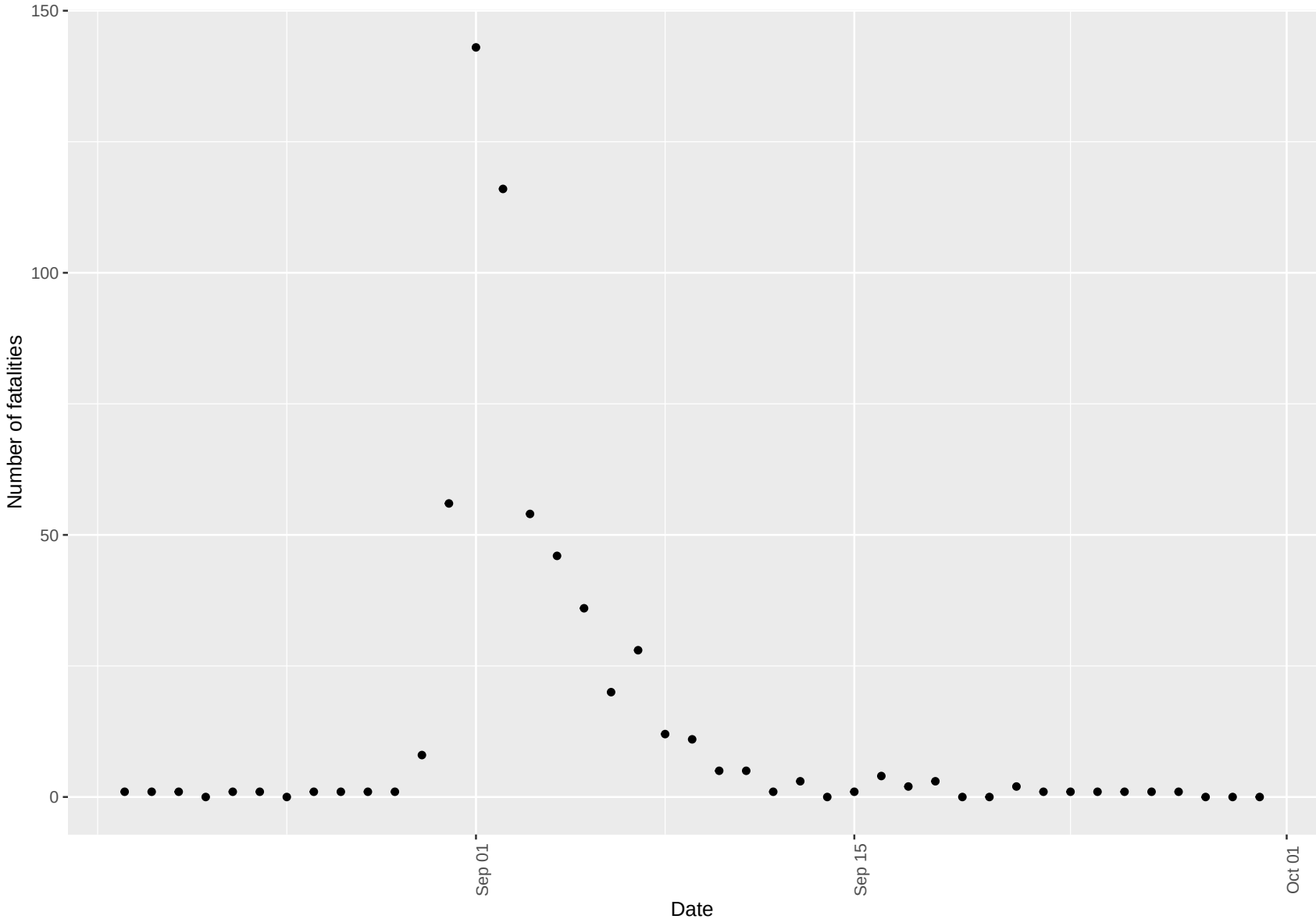
The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and the adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street joins Broad Street, there were upwards of five hundred fatal attacks of cholera in ten days. The mortality in this limited area probably equals any that was ever caused in this country, even by the plague; and it was much more sudden, as the greater number of cases terminated in a few hours. The mortality

Outbreak in 1854

TABLE I.

Date.		No. of Fatal Attacks.	Deaths.
August	19	1	1
„	20	1	0
„	21	1	2
„	22	0	0
„	23	1	0
„	24	1	2
„	25	0	0
„	26	1	0
„	27	1	1
„	28	1	0
„	29	1	1
„	30	8	2
„	31	56	3
September	1	143	70
„	2	116	127
„	3	54	76
„	4	46	71
„	5	36	45
„	6	20	37
„	7	28	32
„	8	12	30
„	9	11	24
„	10	5	18
„	11	5	15
„	12	1	6
„	13	3	13
„	14	0	6
„	15	1	8
„	16	4	6
„	17	2	5
„	18	3	2
„	19	0	3
„	20	0	0
„	21	2	0
„	22	1	2
„	23	1	3
„	24	1	0
„	25	1	0
„	26	1	2
„	27	1	0
„	28	0	2
„	29	0	1
„	30	0	0
Date unknown		45	0
	Total	616	616

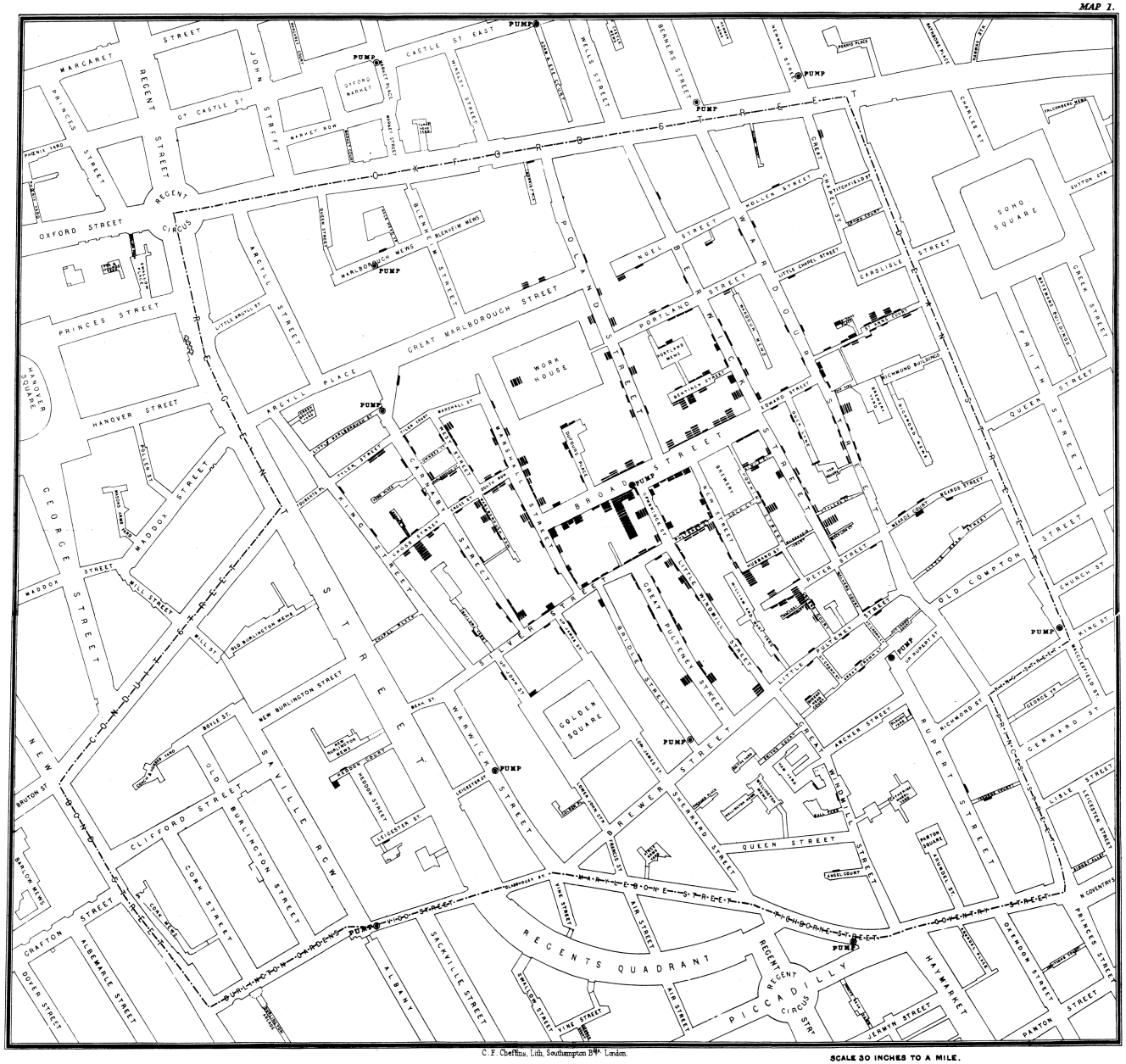
Outbreak in 1854



Snow's investigation

- "I suspected some contamination of the water of the much-frequented street-pump in Broad Street"
- "I requested permission, therefore, to take a list, at the General Register Office, of the deaths from cholera, registered during the week ending 2nd September, in the sub-districts of Golden Square, Berwick Street, and St.Ann's, Soho, which was kindly granted."
- 83 deaths in the last days of the week: only 10 in houses not near the pump
  - 5 did drink water from the Broad Street pump
  - 3 children that went to school near Broad Street
  - 2 not connected

How he formed this hypothesis



As a result of the investigation

"I had an interview with the Board of Guardians of St. James's parish, on the evening of Thursday, 7th September,and represented the above circumstances to them. In consequence of what I said, **the handle of the pump was removed on the following day**."

Further facts: the brewery

"There is a Brewery in Broad Street, near to the pump, and on perceiving that no brewer's men were registered as having died of cholera, I called on Mr. Huggins, the proprietor. He informed me that there were above seventy workmen employed in the brewery, and that none of them had suffered from cholera–at least in a severe form–only two having been indisposed, and that not seriously, at the time the disease prevailed. **The men are allowed a certain quantity of malt liquor, and Mr. Huggins believes they do not drink water at all**; and he is quite certain that the workmen never obtained water from the pump in the street. There is a deep well in the brewery, in addition to the New River water."

Further facts: the workhouse

"The Workhouse in Poland Street is more than three fourths surrounded by houses in which deaths from cholera occurred, yet out of five hundred and thirtyfive inmates only five died of cholera, the other deaths which took place being those of persons admitted after they were attacked. The workhouse **has a pump-well on the premises**, in addition to the supply from the Grand Junction Water Works, and the inmates never sent to Broad Street for water."

Further facts: the west end widow

"I was informed by this lady's son that she had not been in the neighbourhood of Broad Street for many months. A cart went from Broad Street to West End every day, and **it was the custom to take out a large bottle of the water from the pump in Broad Street, as she preferred it**. The water was taken on Thursday, 31st August, and she drank of it in the evening, and also on Friday. She was seized with cholera on the evening of the latter day, and died on Saturday, as the above quotation from the register shows. A niece, who was on a visit to this lady, also drank of the water; she returned to her residence, in a high and healthy part of Islington, was attacked with cholera, and died also."

Summary: the broad street pump study

- Snow suspected that contaminated water caused Cholera, and when the outbreak of 1854 occurred, he used it as an opportunity to assess evidence for this hypothesis
- Tools he used included
  - **Visualization** by overlaying deaths and pump locations on a map
  - **Identifying outliers** that didn't fit with the overall pattern – for example, the brewery workers – who lived near the pump and did not become infected, then gathering additional data to develop an explanation
- His analysis suggests a possible **association** between Cholera and contaminated water
- However, it is hard to completely rule out the possibility that Cholera is caused by some **third factor** that is correlated with proximity to the broad street pump
- It's also possible that the association is **spurious**

How to establish cause?

- How could Snow make a more rigorous, persuasive argument that Cholera is caused by drinking contaminated water?
- Well, in science we traditionally **conduct an experiment** to test our hypotheses
- The "canonical" experiment in this case is to
  1. Select some people **at random**
  2. **Randomly** assign some of them to a **control** group. The control group drinks **clean water**
  3. Assign the rest to a **treatment** group. The treatment group drinks **contaminated water**
  4. Compare how many people in the **treatment** group get Cholera to the number in the **control** group
- This is called a **randomized experiment** or a **randomized trial**
- Can you see any problems with this experiment?

Observational data and "natural" experiments

- Often even when it isn't feasible to do a randomized experiment, it is possible to find some data that is almost as good
- Sometimes nature **creates experiments for us**, by essentially randomizing otherwise similar people into control and treatment
- Even when this doesn't happen, we can try to **model** the process by which people are sorted into control and treatment groups
- These types of analyses are called **observational studies**, and the goal of the analysis is usually to **assess causality**
- Lucky for Snow, a natural experiment was created for him by London's water companies

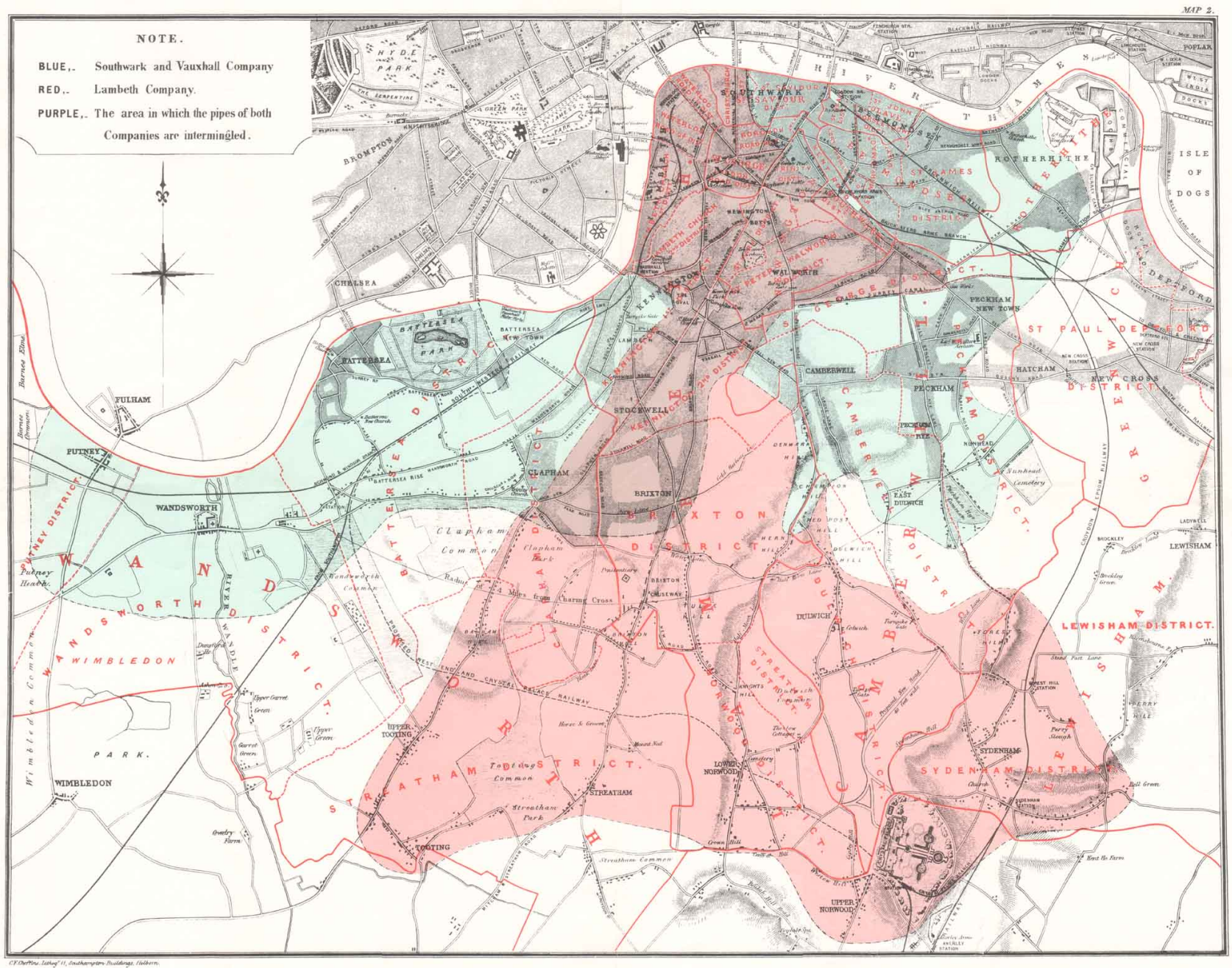
Water supplies in London

- Looking at epidemics in 1832 and 1849, Snow notices that areas of London served by different water companies have different mortality rates
- "...in every district to which the supply of the Southwark and Vauxhall, or the Lambeth Water Company extends, the cholera was more fatal than in any other district whatever."
- These companies took their water from the Thames, which might be contaminated by sewage (at the time, much of London's sewage was discharged into the Thames)

A "natural experiment"

- London was without cholera from the latter part of 1849 to August 1853.
- During this interval an important change had taken place in the water supply of several of the south districts of London. The Lambeth Company removed their water works, in 1852, from opposite Hungerford Market to Thames Ditton; thus obtaining a supply of water quite free from the sewage of London
- Lambeth and Southwark & Vauxhall Companies serve some of the same area

Map of service areas of the two companies



Two comparable groups

"In the sub-districts being supplied by both Companies, the **mixing of the supply is of the most intimate kind**. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active competition. In many cases a single house has a supply different from that on either side. Each Company **supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies.**"

A grand experiment

"The experiment, too, was on the grandest scale. No fewer than **three hundred thousand people** of both sexes, of every age and occupation, and of every rank and station, from gentle folks down to the very poor, **were divided into two groups without their choice, and, in most cases, with out their knowledge**; one group being supplied with water containing the sewage of London, and, amongst it, what ever might have come from the cholera patients, the other group having water quite free from such impurity."

Comparison of death rates by water company

This is a table of deaths during the first seven weeks of the epidemic, ending August 26, 1854, in the area served by both companies

TABLE IX.

	Number of houses.	Deaths from Cholera.	Deaths in each 10,000 houses.
Southwark and Vauxhall Company	40,046	1,263	315
Lambeth Company	26,107	98	37
Rest of London	256,423	1,422	59

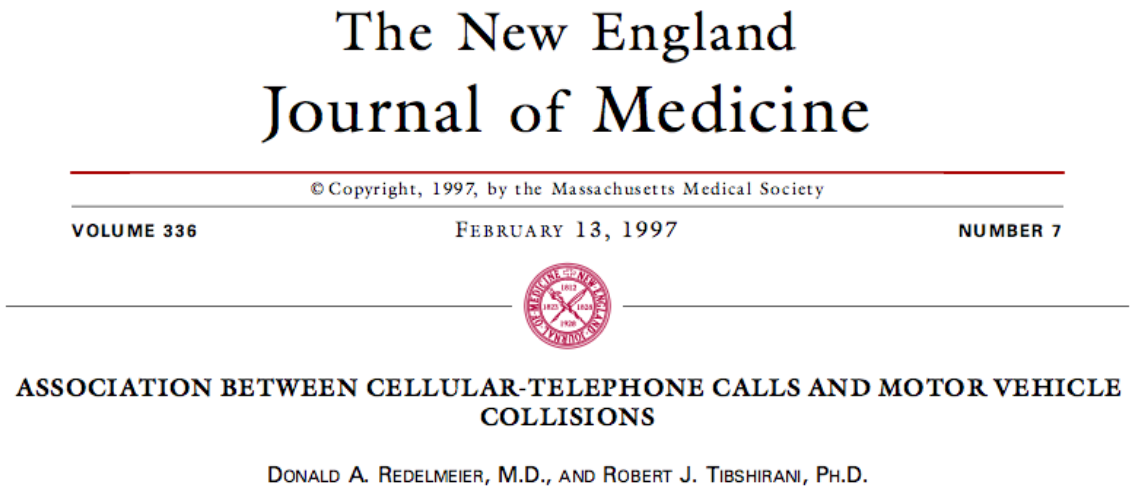
Question to think about for later: the deaths rates **seem** convincingly different in comparing S&V to Lambeth. Is it large enough that we are convinced it isn't just a coincidence? How large is "large enough"? What about Lambeth vs rest of London. Do Lambeth water customers "really" have a lower death rate, or is just "by chance" that the observed rate is lower?

What makes Snow's work so powerful

- It **documents** the sources and characteristics of the data
- It makes appropriate **controlled comparisons**
- It gets to mechanisms of **cause and effect**
- It uses information from **many different sources**
- It expresses the differences **quantitatively**
- It inspect and evaluates **alternative explanations**

A little bit more contemporary example

- 1997 study on the use of cell phones and car collisions, one of the first of its kind



The New England
Journal of Medicine

VOLUME 336 FEBRUARY 13, 1997 NUMBER 7

ASSOCIATION BETWEEN CELLULAR-TELEPHONE CALLS AND MOTOR VEHICLE COLLISIONS

DONALD A. REDELMEIER, M.D., AND ROBERT J. TIBSHIRANI, PH.D.

link to source

Background

- Motor vehicle collisions are a leading cause of accidental death in North America
- In 1997, cell phones were not widely used and many were fixed (in car) phones, but their popularity was growing rapidly
- Phone calls are a distraction: this study asked whether there is a link with collisions
- Industry sponsored studies suggested no increased risk of collisions for cell phone users, but studies with simulators indicated worse reaction time while using a cell phone

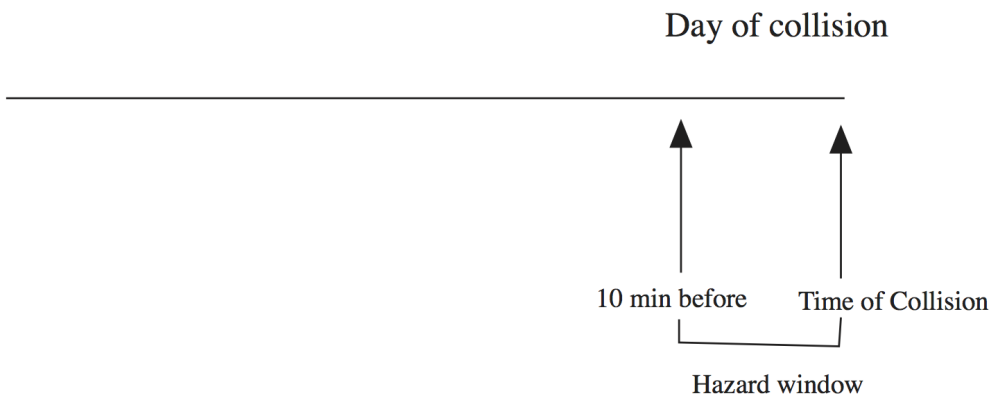
1997 Study design

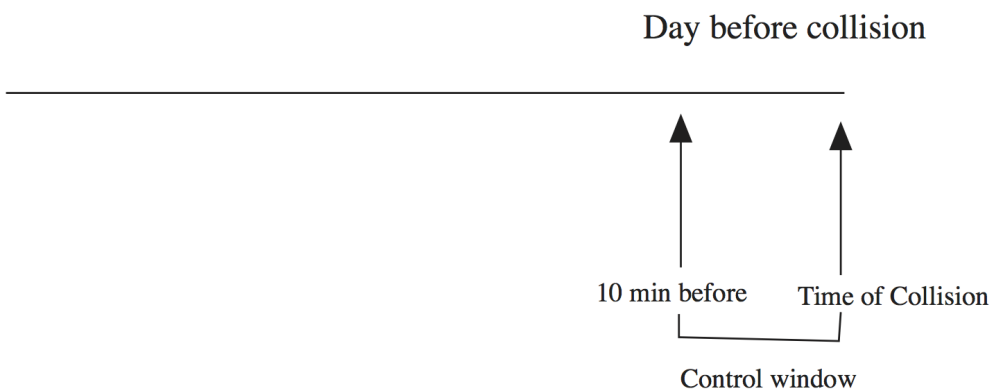
- 5890 drivers approached in collision reporting centers in Toronto, who were involved in collisions causing significant property damage but no personal injury between July 1, 1994, and August 31, 1995
- 1064 had cell phones, and 742 consented to be in the study. 699 provided accurate telephone numbers that could be linked to detailed billing records
- Some basic information (age, sex . . .) was obtained by questionnaire
- Using questionnaire, police reports, and 911 calls, they carefully deduced the actual time of collision and the subjects' phone usage in the previous week

1997 Study: case-crossover design

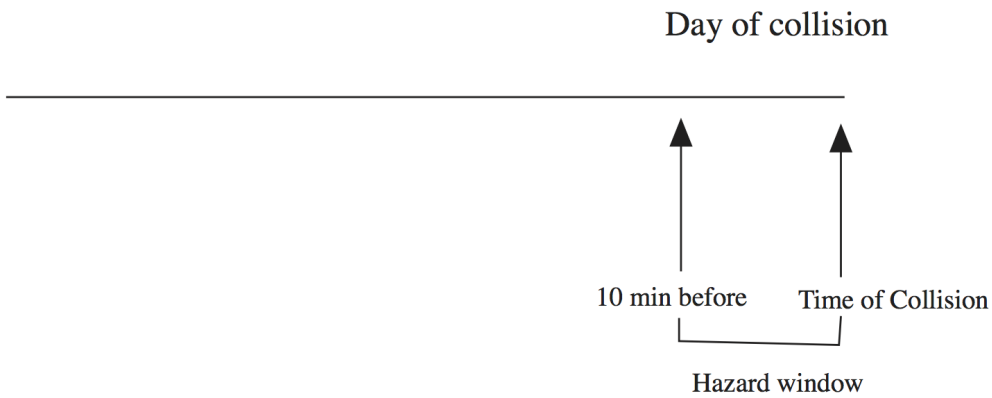
- **Self matching case-control design**: each person acts as their own control.
- Contrast a time period on the day of the collision with a comparable period on a day preceding the collision
- Case–crossover analysis would identify an increase in risk if there were more telephone calls immediately before the collision than would be expected solely as a result of chance

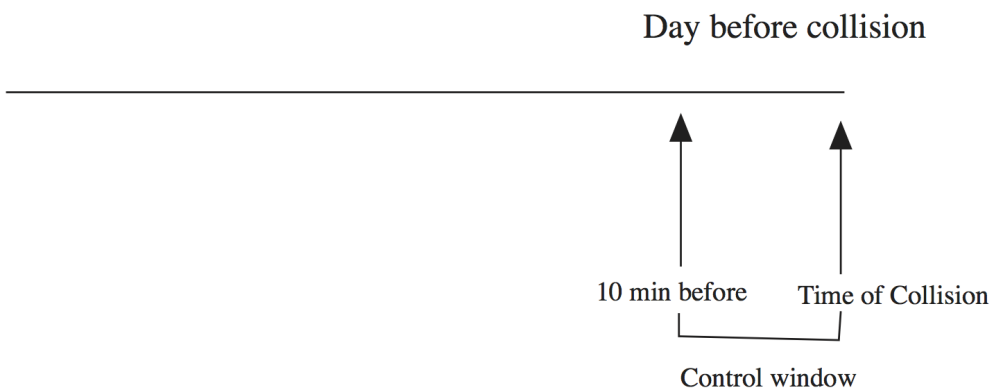
Hazard and Control Windows





Hazard and Control Windows





Can you think of any possible problems with this designation?

Data

Results	Day before/Not on	Day before/On
Day of collision: Not On	505	24
Day of collision: On	157	13

Is this evidence that using cell phones when driving is dangerous?

Hypothetical data

Results	Day before/On	Day before/Not on
Day of collision: On	13	91
Day of collision: Not on	90	505

Another alternative, hypothetical set of findings. . .

How can we compare these scenarios?

Relative risk

Ratio of

- the probability of having a collision given that you used a cell phone at some time during the preceding 10-minute interval
- the probability of having a collision given that you **did not** use a cell telephone at any time during the preceding 10-minute interval

Relative risk

Ratio of

- the probability of having a collision given that you used a cell phone at some time during the preceding 10-minute interval
- the probability of having a collision given that you **did not** use a cell telephone at any time during the preceding 10-minute interval

Recall our data:

Results	Day before/Not on	Day before/On
Day of collision: Not On	505	24
Day of collision: On	157	13

Relative risk

Ratio of

- the probability of having a collision given that you used a cell phone at some time during the preceding 10-minute interval
- the probability of having a collision given that you **did not** use a cell telephone at any time during the preceding 10-minute interval

Recall our data:

Results	Day before/Not on	Day before/On
Day of collision: Not On	N_{00}	N_{01}
Day of collision: On	N_{10}	N_{11}

Estimator of relative risk:

$$\widehat{RR} = \frac{N_{10}}{N_{01}}$$

Looking at the data

```
celldata <- read.csv("./data/celldataNew.csv")
attach(celldata)
table(dayof.on,daybef.on)
```

```
         daybef.on
dayof.on   0   1
       0 505  24
       1 157  13
```

```
rrisk.hat<-sum(dayof.on*(1-daybef.on))/
  (sum((1-dayof.on)*daybef.on))
rrisk.hat
```

```
[1] 6.541667
```

Let's look a bit more at the data

```
table(dayof.on,daybef.on,gender)
```

```
, , gender = f

        daybef.on
dayof.on   0   1
       0 147   6
       1  48   4

, , gender = m

        daybef.on
dayof.on   0   1
       0 358  18
       1 109   9
```

Let's look a bit more at the data

```
table(dayof.on,daybef.on,type.of.cell)
```

```
, , type.of.cell = handheld

         daybef.on
dayof.on   0   1
       0 104   4
       1  37   5

, , type.of.cell = handsfree

         daybef.on
dayof.on   0   1
       0 401  20
       1 120   8
```

A problem in the design

You cannot be involved in a car collision if you are not driving.

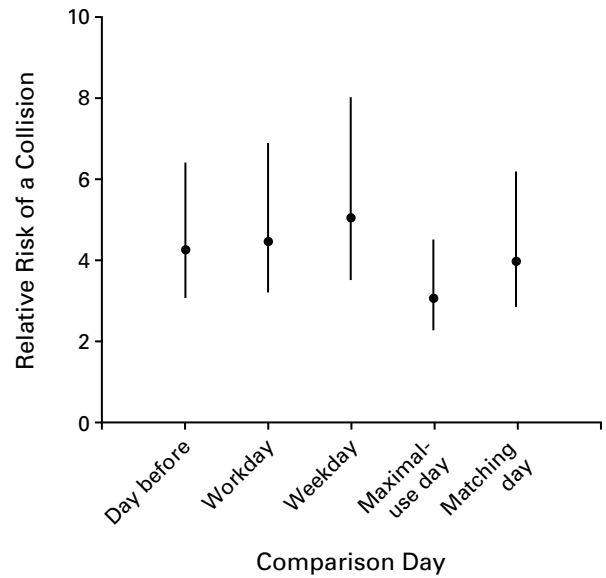
- The authors estimated that only 65% of the subjects were driving at the same time the day before the collision.
- They modified the relative risk estimate, by dividing the denominator of the ratio by 0.65

```
rrisk<-sum(dayof.on*daybef.noton)/
  (sum(dayof.noton*daybef.on)/.65)
rrisk
```

```
[1] 4.252083
```

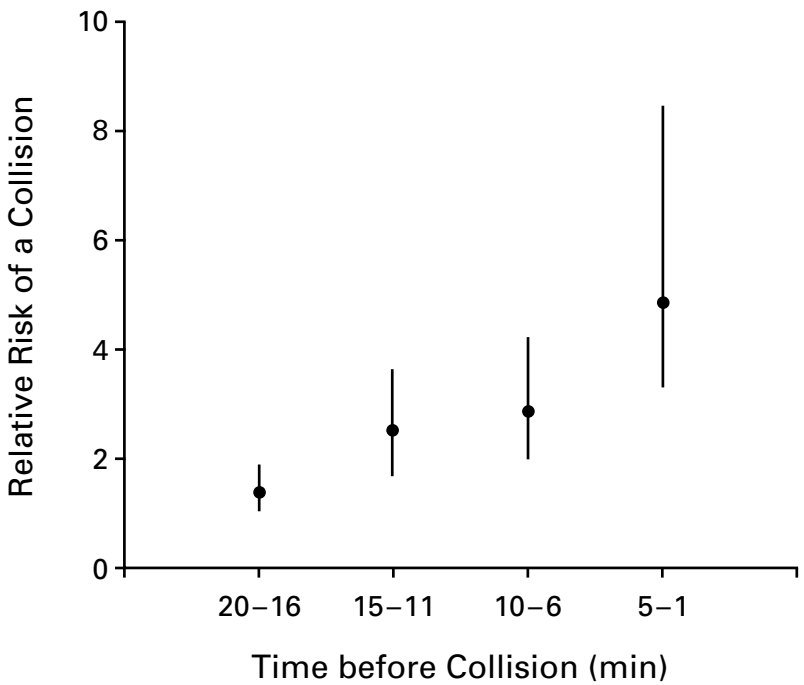
Better control windows

- What if the day before the accident was a Sunday and driving behavior is different on week-ends?
- Address this by using different comparison periods: Day before, workday, weekday, maximal-use day, matching day



Some additional comparisons

ASSOCIATION BETWEEN CELLULAR-TELEPHONE CALLS AND MOTOR VEHICLE COLLISIONS



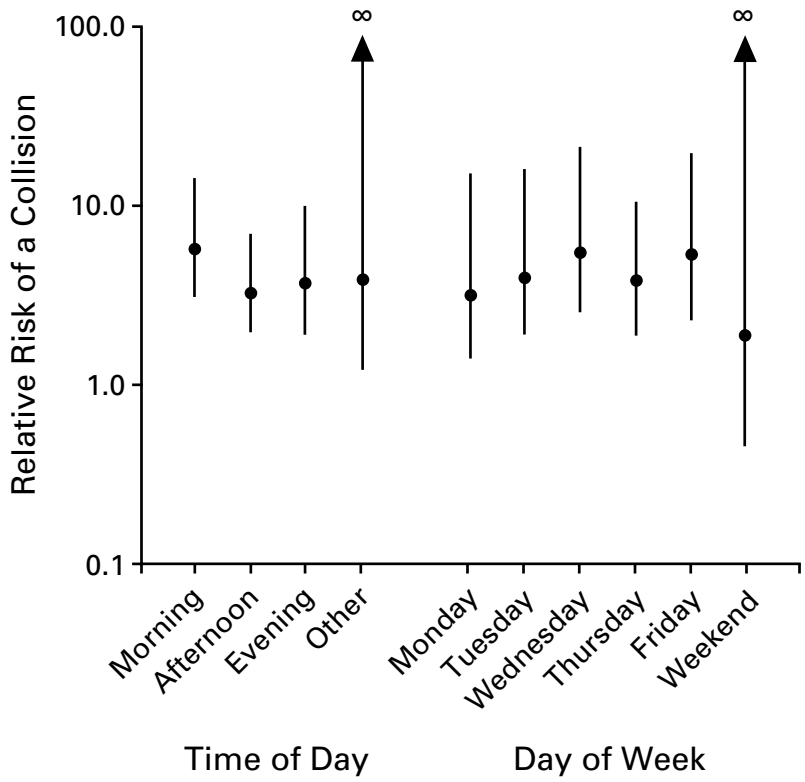


Figure 2. Time of Cellular-Telephone Call in Relation to the Relative Risk of a Collision.

Confidence of the estimates

- The values we are reporting are *estimates*: just one guess of the true relative risk based on the data we observed
- Even if there is really a difference in your risk of having a car collision depending on whether you were recently using your cell phone, the risk is just a **probability** that the crash occurs
- Therefore if we were to obtain another sample of people who had been in accidents, we would get different estimates
- This **variability due to sampling** is a critical concept in statistics
- We will learn how we can not only summarize the information in the data about a quantity of interest (like relative risk), but also how we can quantify the **strength of the evidence** that this conveys about a hypothesis

Main recommendations

In 1997:

Our study may serve as a warning: if cell phones become standard equipment in cars, collision rates may increase

But smart phones and texting were coming!