Do disease clusters have a common cause? The first epidemiological study.

Unit 6 Lecture 1

Jonathan Auerbach
STAT 489 Pre-Cap Prof Development
jauerba@gmu.edu

November 11, 2021

How do scientists determine whether a cluster of diseases have a common cause?

- Do cancer clusters suggest a common environmental cause?
- Do clusters of mass shootings suggest violence is contagious?
- Is winning the lottery multiple times evidence of fraud?

These slides use the following R packages

Setup:

```
library("knitr")
library("HistData")
library("tidyverse")
library("ggmap")
library("sp")
theme_set(theme_bw())
```

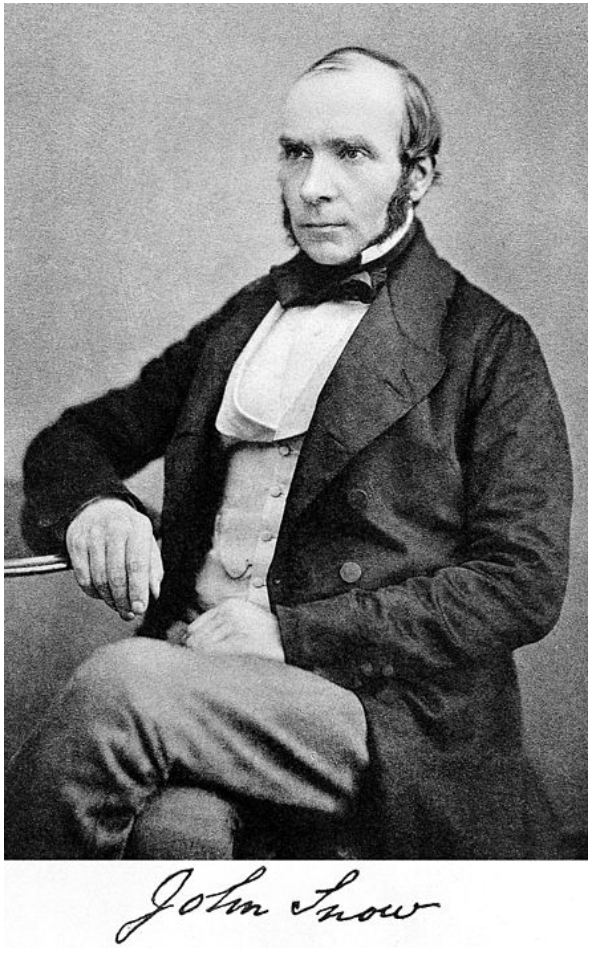
These slides use Google Maps. To obtain an API key and enable services, go to https://cloud.google.com/maps-platform/.

```
register_google(key = "[your key]")
```

How convincing is a cluster of individuals with the same disease?

- In 1854, a cholera outbreak killed more than ten thousand people.
- Scientists disagreed on whether the cause was airborne or waterborne.
- Snow recorded the location of every documented cholera case.
  - He noticed that the cases concentrated around the Broad Street (Water) Pump
  - His subsequent work proving the link between drinking water and cholera dispelled the false theory that cholera spread by air particles ("miasma theory")

John Snow (1856) and Mode of Communication of Cholera (1855)



ON THE

MODE OF COMMUNICATION OF CHOLERA.

THE existence of Asiatic Cholera cannot be distinctly traced back further than the year 1769. Previous to that time the greater part of India was unknown to European medical men; and this is probably the reason why the history of cholera does not extend to a more remote period. It has been proved by various documents, quoted by Mr. Scot,* that cholera was prevalent at Madras in the year above mentioned, and that it carried off many thousands of persons in the peninsula of India from that time to 1790. From this period we have very little account of the disease till 1814, although, of course, it might exist in many parts of Asia without coming under the notice of Europeans.

In June 1814, the cholera appeared with great severity in the 1st bat. 9th regt. N.I., on its march from Jaulnah to Trichinopoly; while another battalion, which accompanied it, did not suffer, although it had been exposed to exactly the same circumstances, with one exception. Mr. Cruikshanks, who attended the cases, made a report, which will be alluded to further on.

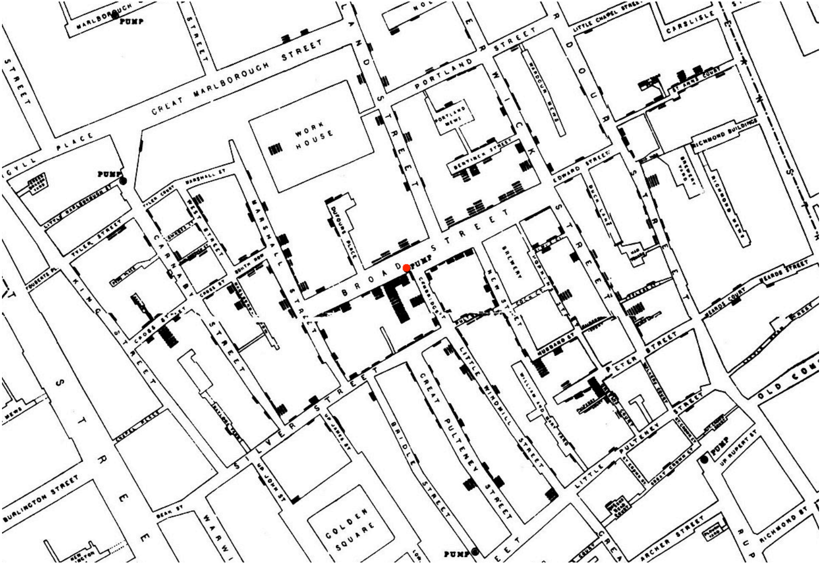
In 1817, the cholera prevailed with unusual virulence at several places in the Delta of the Ganges; and, as it

* Report on the Epidemic Cholera, 1824, p. 5.

B

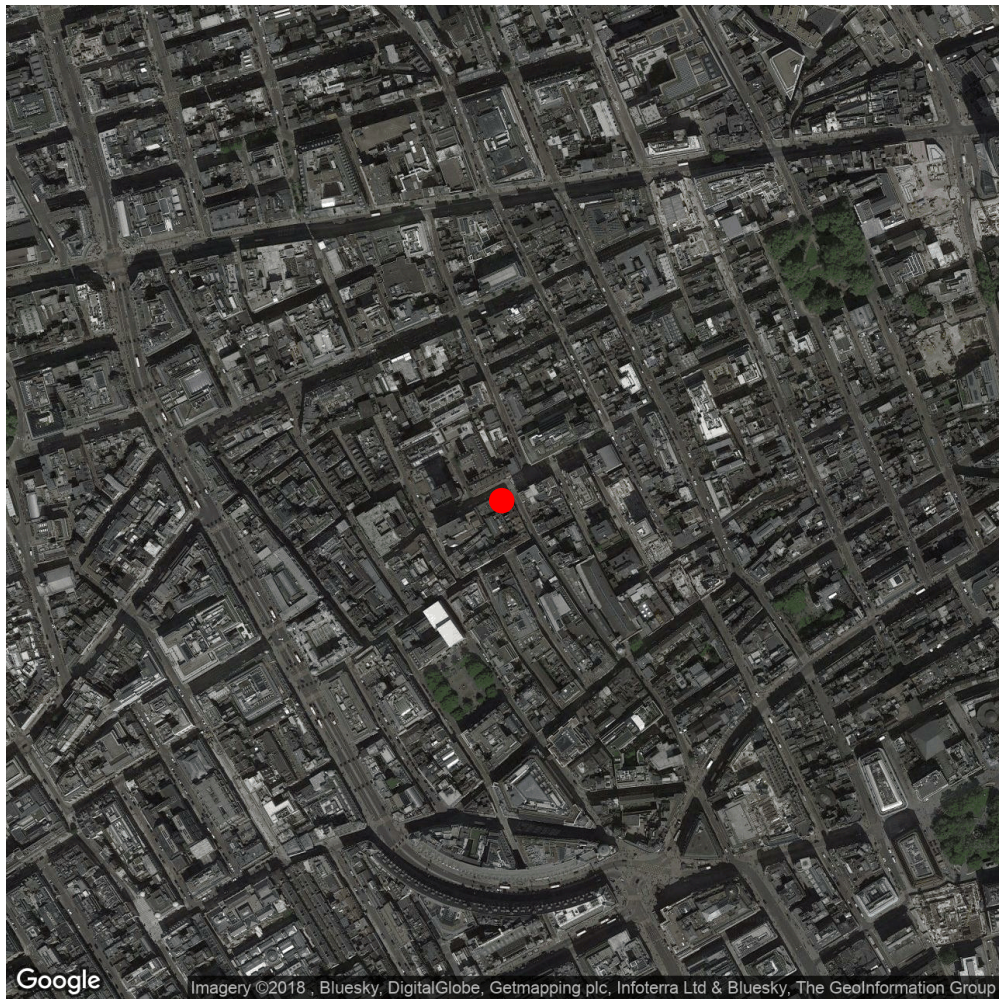
Source: https://commons.wikimedia.org/wiki/File:John_Snow.jpg

Broad Street Pump and Cholera Cases (London, 1854)



Broad Street Pump (London, today)

```
(map <- ggmap(get_map(location = "Broad St Pump, London",
                      zoom = 16, maptype = "satellite")) +
         geom_point(aes(X1, X2), color = "red",
                    data = pump))
```



Broad Street Pump and Snow Map (1855)

```
map +
  geom_path(aes(x=long, y=lat, group=group),
      color = "white", alpha = .5, size = 2,
      data = Snow_df)
```

Broad Street Pump and Snow Map (1855)

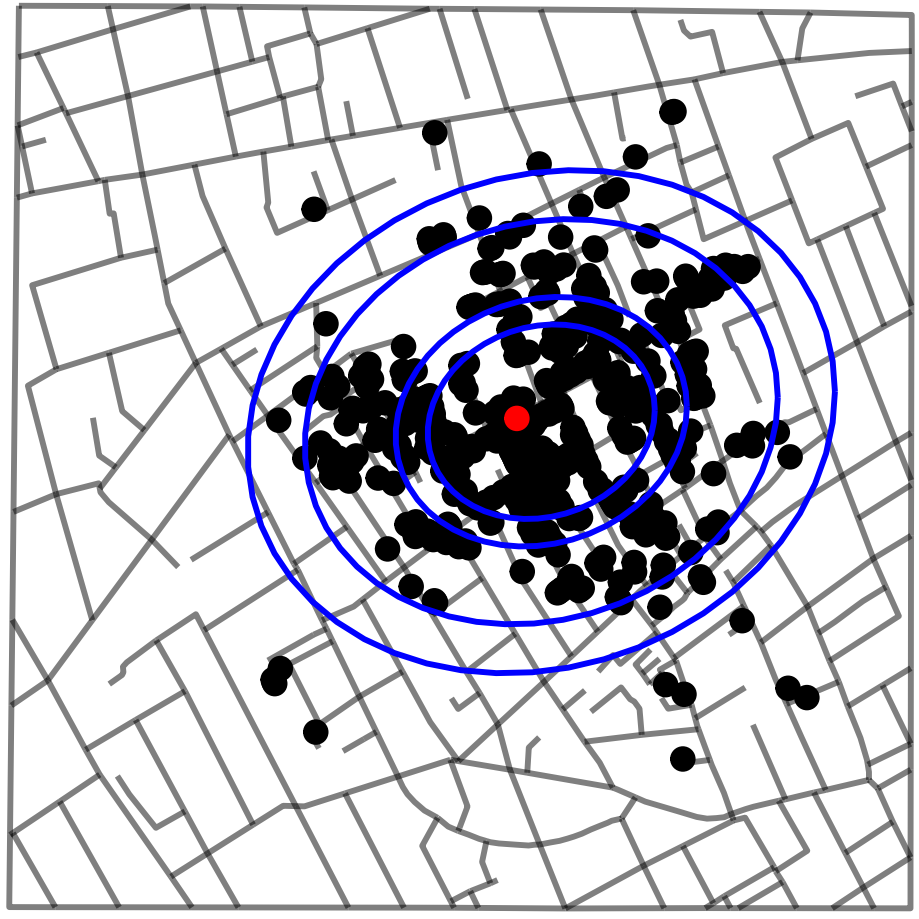
```
(map <- ggplot() + theme_nothing() +
    geom_path(aes(x=long, y=lat, group=group),
              color = "black", alpha = .5,
              data = Snow_df))
```

Cholera cases surrounding Broad Street Pump

```
(map <- map + geom_point(aes(x=long, y=lat),
                  data = Snow_deaths) +
    geom_point(aes(X1, X2), color = "red",
                  data = pump))
```

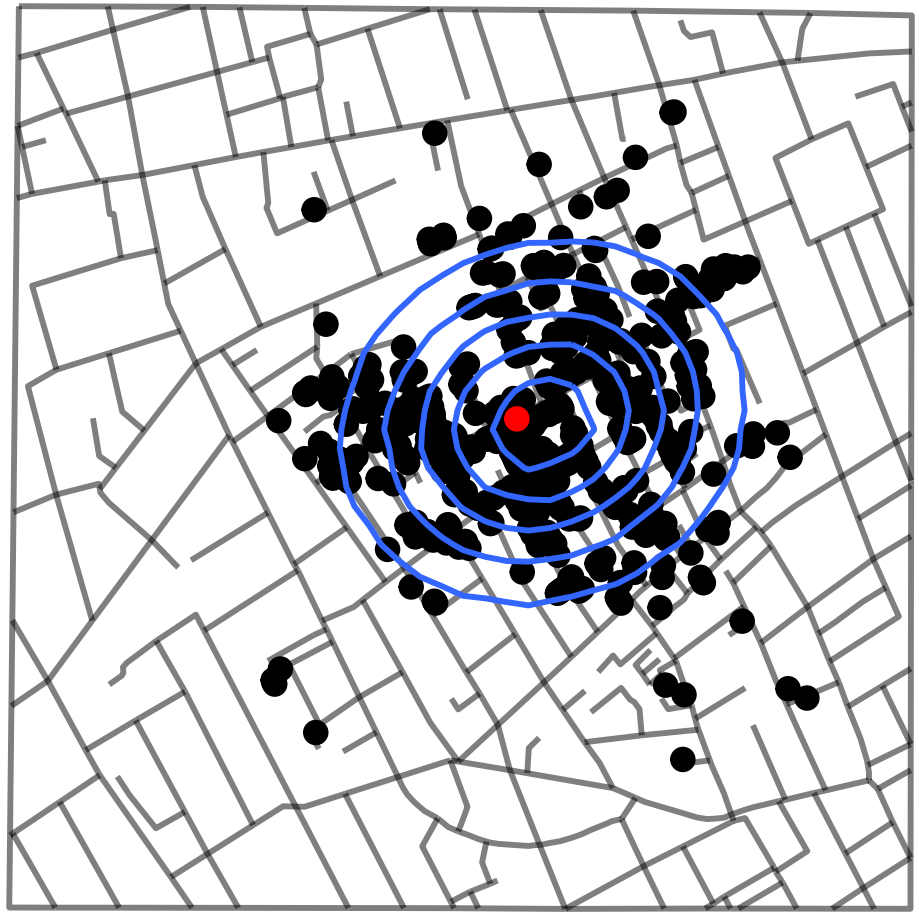
Today we might quantify clustering with ellipses

```
ellipse <- map
for(level in c(.5, .68, .95, .99)) ellipse <- ellipse +
  stat_ellipse(aes(x = long, y = lat), color = "blue",
               data = Snow_deaths, level = level)
ellipse
```



Contours of bivariate normal distribution are ellipses

```
sim <- as_tibble(MASS::mvrnorm(1e6,
       colMeans(Snow_deaths), cov(Snow_deaths)))
map + geom_density_2d(aes(x = long, y = lat),
                      data = sim, n = 50, bins = 6)
```



Is the pump at the center of the cluster?

- Confidence ellipses help determine if pump consistent with center.
  - Let $[X, Y] = \{(X_i, Y_i)\}_{i=1}^{n}$ have expected values (μ_X, μ_Y), standard deviations (σ_X, σ_Y), and correlation $\rho = \frac{\mathsf{Cov}(X,Y)}{\sigma_X \sigma_Y}$.

If $[X, Y]$ follow a bivariate normal distribution

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathsf{Normal}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right),$$

then $100 \times (1-\alpha)\%$ confidence ellipse is parameterized by equations:

$$\begin{cases} x(t) = \mu_X + v_{1X} c\sqrt{\lambda_1}\mathsf{cos}(t) + v_{2X} c\sqrt{\lambda_2}\mathsf{sin}(t) \\ y(t) = \mu_Y + v_{1Y} c\sqrt{\lambda_1}\mathsf{cos}(t) + v_{2Y} c\sqrt{\lambda_2}\mathsf{sin}(t) \end{cases}$$

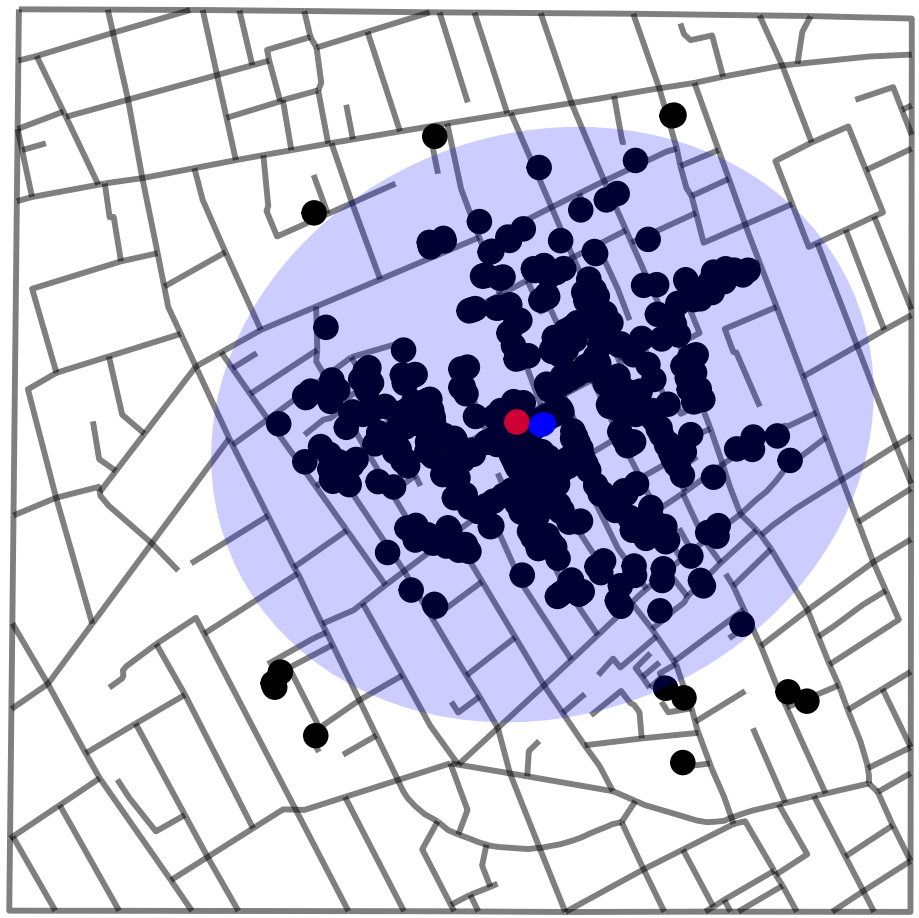
where $0 \le t \le 2\pi$; λ_1 and λ_2 are eigenvalues of covariance matrix $\Sigma = \frac{1}{n}\begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$ with eigenvectors $\begin{bmatrix} v_{1X} \\ v_{1Y} \end{bmatrix}$ and $\begin{bmatrix} v_{2X} \\ v_{2Y} \end{bmatrix}$; and c the size—providing desired coverage, e.g. $c^2 = 2\frac{(n-1)}{(n-2)}F_{1-\alpha}(2, n-2)$.

R implementation of confidence or prediction ellipse

```
n_obs <- nrow(Snow_deaths)
mu <- c(mean(Snow_deaths$long),
        mean(Snow_deaths$lat))

data_ellipse <- function(alpha, type = "c") {
  if(type == "c") {Sigma <- cov(Snow_deaths)/n_obs} else {
    Sigma <- cov(Snow_deaths)}
  V <- eigen(Sigma)$vectors; lambda <- eigen(Sigma)$values

 ellipse_alpha <- function(t)
    mu + V %*% (sqrt(
     qf(1 - alpha, 2, n_obs - 2) *
       2 * (n_obs - 1) / (n_obs - 2) * lambda) *
       c(cos(t), sin(t)))

 as_tibble(t(sapply(seq(0, 2 * pi, len = 100),
                    ellipse_alpha)))
}
```

The pump is a plausible center. Prediction ellipse (large region) and confidence ellipse (small region)

```
map + geom_polygon(aes(V1, V2),
       fill = "blue", data = data_ellipse(.01, t = "c")) +
      geom_polygon(aes(V1, V2), alpha = .2,
       fill = "blue", data = data_ellipse(.01, t = "p"))
```



What caused the Cholera Outbreak of 1854?

- Snow didn't simply describe the clustering of cases. He met with residents, studied where they got their water, and tested samples
  - He found residents more or less randomly chose where they got their water, and that infected residents frequented the Broad Street Pump— establishing the pump as the likely cause
- The epidemic ended after Snow convinced the City to remove the handle to the Broad Street pump, validating Snow's extensive work
  - The Cholera Inquiry Committee even identified patient zero: a five month old baby
- Snow's study was historic. It is considered the classic example of good epidemiology
  - Similar observational evidence links cigarette smoking to lung cancer (e.g. Cornfield et al. 1959)

Do all clusters have a cause?

- Movies popularize clusters as strong evidence of misconduct
  - e.g. Erin Brockovich (Hinkley, CA) and Lois Gibbs (Love Canal, NY)

- The CDC is skeptical in general: "the likelihood of establishing a definitive cause-and-effect relationship between the health event and an exposure is slight"
  - A 1989 national conference on disease clusters found that cluster studies rarely produce important findings
  - Goodman et al. reviewed over 500 cancer cluster investigations and found only one was able to identify a cause with certainty

Most investigated clusters have no cause (Goodman et al. 2012)

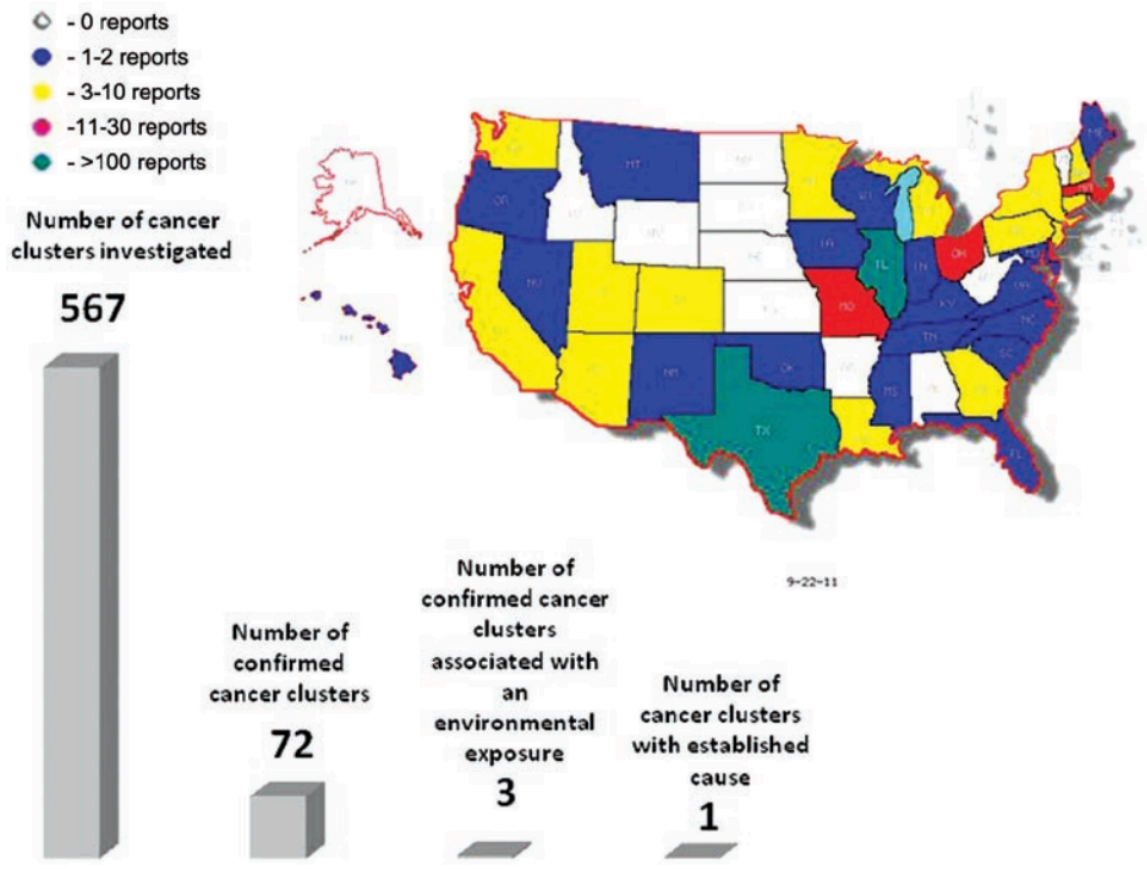


Figure 1. Numbers of publicly available cancer cluster investigation reports by state and comparison of numbers of investigated cancer clusters, confirmed cancer clusters (e.g. investigated clusters where number of cancer cases is greater than expected), clusters linked to an environmental exposure, and cancer clusters with an established cause. Although some of the cluster investigations may have been described in several reports, the numbers in this figure represent unique reported clusters. (Map generated from data in Table 1 using Map-Maker Utility, http://monarch.tamu.edu/~maps2/us_12.htm)

Hill (1965) criteria for association to be 'causal':

1. Strength - the magnitude of the association should be large

2. Consistency - the association should be observed repeatedly by different persons, in different places, under different circumstances and times

3. Specificity - association should be limited to specific circumstances and not others. e.g. exposed workers get disease, unexposed do not

4. Temporality - the proposed cause should proceed the effect

5. Biological Gradient - The response should increase with the dose. e.g. more exposure should result in more deaths

6. Plausibility - A plausible (biological) mechanism should link the proposed cause with the effect

7. Coherence - The proposed cause-and-effect relationship should not seriously conflict with generally known facts

References

1. Cornfield, Jerome, William Haenszel, E. Cuyler Hammond, Abraham Lilienfeld, Michael Shimkin, and Ernst Wynder. "Smoking and lung cancer: recent evidence and a discussion of some questions." Journal of the National Cancer institute 22.1 (1959): 173-203.

2. Friendly, Michael. "HistData: Data sets from the history of statistics and data visualization." R package version 0.7-5 (2014).

3. Friendly, Michael, Georges Monette, and John Fox. "Elliptical insights: understanding statistical methods through elliptical geometry." Statistical Science 28.1 (2013): 1-39.

4. Goodman, Michael, Joshua Naiman, Dina Goodman, and Judy LaKind. "Cancer clusters in the USA: what do the last twenty years of state and federal investigations tell us?" Critical reviews in toxicology 42.6 (2012): 474-490.

5. Hill, Austin Bradford. "The environment and disease: association or causation?." Journal of the Royal Society of Medicine 108.1 (2015): 32-37.

6. Snow, John. On the mode of communication of cholera. John Churchill,