CSE 8803 EPI: Data Science for Epidemiology, Fall 2020

Lecturer: B. Aditya Prakash Lecture #2
Scribe: Kipp Morris August 19, 2020

1 Summary of Lecture Content

This lecture served as an introduction to epidemiology, data science, and why data science is useful in epidemiology.

We started by looking at a few different definitions of epidemiology and defining what our focus within epidemiology will be for the purposes of this class. After that, we discussed computational epidemiology and its role in the general field of epidemiology.

We moved on to discuss epidemics in history, drawing attention to the fact that they will almost certainly continue to be part of history and that that is a key reason for us to be here studying computational epidemiology.

We brought up COVID-19 and how it is affecting our lives right now. We then examined a few case studies of past epidemics and how some of the methods people used in response can be considered precursors to modern computational epidemiology.

Finally, we defined data science and established why it is important to the present and future of epidemiology.

2 What is Epidemiology?

We looked at a few different definitions of epidemiology. The following are the two that we focused on:

- Webster's dictionary: "The science which investigates the causes and controls of epidemic diseases" [1]
- CDC: "The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems" [3]

The definition from Webster's dictionary is what we immediately think of when we think of epidemiology. The CDC's definition is broader than just the study of infectious diseases and includes the idea that epidemiology is at the population-level, not at the patient-level like other disease-related sciences. Another key point of the CDC's definition is that it includes the "the application of this study to the control of health problems", emphasizing that the end goal of epidemiology is to apply any acquired knowledge to the mitigation of real world health problems.

With that said, we established that for our class we will focus on epidemiology as it pertains to the study of infectious diseases caused by microparasites.

3 Computational Epidemiology

We defined computational epidemiology as follows:

"The development of computational and mathematical methods, tools, and techniques to support epidemiology"

The key thing that makes computational epidemiology important is that the use of computational techniques makes population-level analyses feasible when they otherwise would not be. Data-based approaches to analyzing epidemiology problems can save a lot of time and resources compared to traditional laboratory/experimental approaches.

One caveat of computational epidemiology is that the models are likely to be oversimplified. However, they are still useful since traditional approaches are just not doable at the population-level [2].

4 Epidemics in History, Including the Present and Future

Epidemics actually turn out to be common, which, given that they are the 2nd leading cause of death worldwide, gives us ample motivation to study them. In both past and very recent history, there are examples of epidemics that significantly affected populations in different places in the world.

Examples of recent epidemics include:

- 1918 flu pandemic (killed more people than World War I! [4])
- SARS
- 2009 Swine flu
- 2014 Ebola
- COVID-19 (ongoing)

Examples of epidemics further in the past include:

- Plagues in Roman times
- The Plague from the Middle Ages (killed 30-60% of Europe's population)
- A plague that resulted in the fall of the Han dynasty in third century China
- A smallpox outbreak in the 1500s that caused the defeat of the Aztecs.

We looked into a few specific examples of measures that were taken to fight past epidemics:

4.1 Smallpox: Bernoulli and Jenner (1726-1800)

The concept of variolation was known for a while, but Bernoulli was the first major example of someone using a mathematical argument to argue in favor of it.

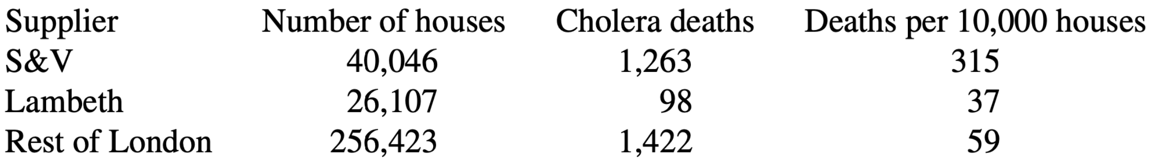
Jenner basically discovered vaccination; it turned out that infecting people with cowpox, a zoonotic (meaning it transfers from animals to humans) virus that is relatively mild in humans, provided immunity against smallpox.

4.2 Cholera: John Snow (1800-1900)

Before John Snow's work, people widely believed that "bad air" made you sick, and this belief is known as the miasma theory.

After observing the cholera outbreak, Snow did not believe that cholera was being transmitted through air; he thought it was being transmitted through the water. So he set out to collect data to support this idea.

Given the assumption that cholera is transmitted through water, Snow assumed that there would be more cases in parts of London downstream from the Thames River than from upstream. For his comparison, he selected two populations that were very similar except for the water companies that operated in the areas in order to control for other factors. The upstream area was serviced by a company named Lambeth, and the downstream area was serviced by a company named Southhall and Vaxhaull. The data, including data for the rest of London, is shown in Figure 1.



Supplier	Number of houses	Cholera deaths	Deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Figure 1: A table showing the data that John Snow collected about cholera counts from a region upstream of the Thames (Lambeth), a region downstream of the Thames (S&V), and the rest of London.

Snow then analyzed the distribution of cases in relation to water pump locations. Figure 2 shows a diagram that represents the concentrations of cholera cases on different blocks using black bars. He then used a voronoi map to get a better look at which pump each case was most likely to be related to. and he found that most of the cases were in the area of the Broad Street pump. The voronoi map is shown in Figure 3. The Broad Street pump is in the yellow region that the red arrow is pointing to, and you can see that most of the cases are concentrated in that yellow region.

After Snow convinced the authorities to replace the Broad Street pump, the epidemic was brought under control.

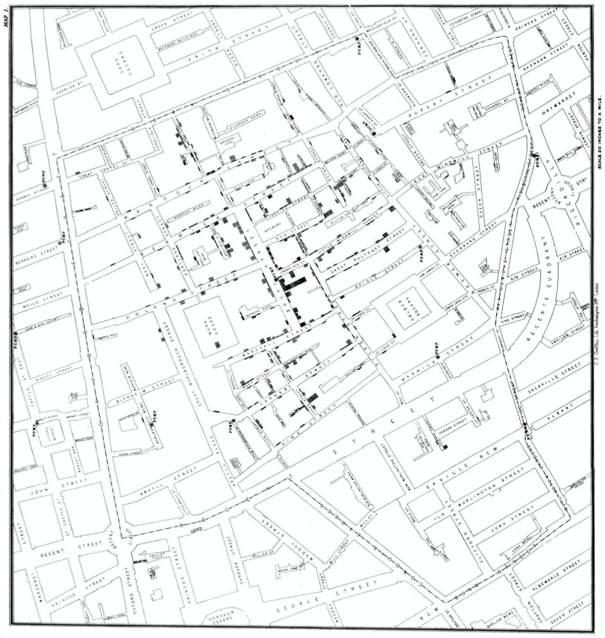


Figure 2: A map created by John Snow showing concentrations of cholera case counts on different blocks in London. Blocks with higher numbers of black bars had more cases.

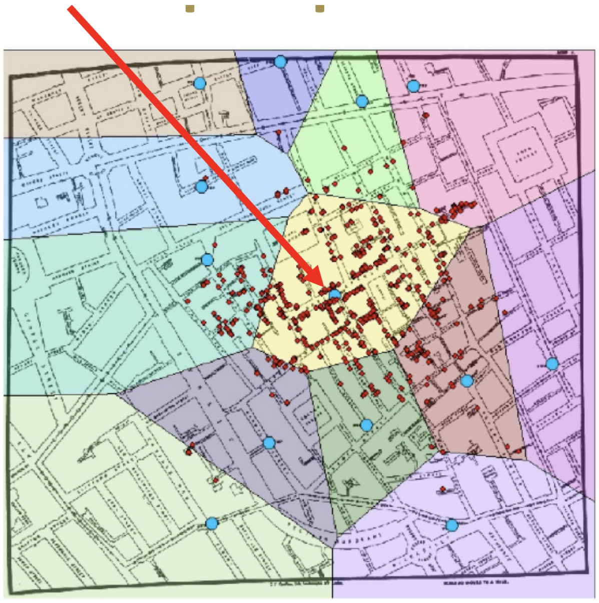


Figure 3: A voronoi map created by John Snow showing cholera cases separated into groups by the closest water pump. The yellow region, which had the most cases, was the Broad Street region where Snow convinced the authorities to replace the pump to reduce cholera cases.

4.3 Malaria: Ross, McDonald, Lotka, and McKendrick (1900-1960)

Ross found the cause of malaria by dissecting an Anopheles mosquito.

He then used a mathematical model to suggest that mosquito reduction could contain the spread of the disease.

5 Risks of Pandemics Today

Fortunately, advances in medicine, sanitation, coordination of government agencies, etc. have improved so much that a pandemic as devastating as the 1918-1919 Spanish flu is unlikely in our future, but COVID has shown us that we should still be prepared for epidemics.

However, the main problem in preparing for epidemics is a lack of information. Even now, as the COVID pandemic has been ongoing since last year, there are still many unknowns (exact source, how effective treatments are, etc.).

Urbanization, increased travel, and mis-information are all trends that are exacerbating the COVID pandemic, but we also have data science nowadays.

6 Data Science

Data science can be used to extract value and knowledge from data, and there is data everywhere in our world.

What John Snow did with cholera was data science before there was data science. However, what he did could not possibly happen at scale without computers, which we can use today for our data science. Today, we have the advantages that data is everywhere and computers are constantly becoming faster and cheaper.

Data science is not just about computing, however. It requires mathematics, statistics, domain expertise (since data science can be applied to a vast number of real-world problem domains), and communication skills (to convince people to accept the knowledge you claim to have extracted from your data).

These are some of the questions data science can address in epidemiology:

- **When** and **where** did the outbreak start? **Who** initially got infected?

  These questions are about finding out how the epidemic started. It is quite difficult to get accurate information about this question, because data has to be collected on the scene; it can't just be done from behind a desk. In addition, when the first case of an outbreak is confirmed, there are likely many other people who are infected because some people who get infected will never seek care, not all who seek care will be tested, and so on. On the flip side, there are a lot of surveillance data sources that can be used such as Internet search terms over time, restaurant reservations, parking lot video surveillance systems, etc.

- **What** can we expect as the epidemic spreads? **What** kind of people are likely to be infected? **When** will the case number peak?

  This is about attempting to plan for the future based on the current data. It could be about predicting peaks, assessing vulnerable populations/areas, predicting the end of the epidemic, etc.

- **How** to control the epidemic? **What** preventative measures can be taken?

  This goes back to the discussion of the “application” in our definition of epidemiology. The main problem is that resources such as medical equipment, the amount of economic disruption that is allowable, etc. are all limited, so we need to do our best to find out how to use them wisely.

As a final remark, another strong point of data science is that it is relatively fast, meaning that it can be used to analyze data from past epidemics in addition to acting as a fast way to learn about an ongoing pandemic.

References

[1] epidemiology. 2020. In Merriam-Webster.com. Retrieved August 24, 2020, from https://www.merriam-webster.com/dictionary/epidemiology

[2] Habtemariam, T., Ghartey-Tagoe, A., Mamo, E., & Robnett, V. (1988). Epidemiologic modelling of diseases — a case example using Schistosoma and Trypanosoma. Mathematical and Computer Modelling, 11, 244-249. doi:10.1016/0895-7177(88)90491-8

[3] Introduction to Epidemiology—Public Health 101 Series—CDC. (2018, November 15). Retrieved August 26, 2020, from https://www.cdc.gov/publichealth101/epidemiology.html

[4] Rosenberg, T. (2017, June 27). Stopping Pandemics Before They Start. Retrieved August 26, 2020, from https://www.nytimes.com/2017/06/27/opinion/stopping-pandemics-before-they-start.html