CSE 8803 EPI: Data Science for Epidemiology, Fall 2024

Lecturer: B. Aditya Prakash
Scribe: Faiz Merchant, Samarth Kamat

August 22, 2024
Lecture 2 : Introduction

1 Summary

In Lecture 2, we discussed the fundamentals of epidemiology, tracing its historical roots and exploring its modern applications through computational methods. We defined epidemiology and moved into the broader scope of the subject. Historical context was provided around significant pandemics such as the 1918 flu, and how these events shaped modern epidemiological practices (as well as their application to the context of the COVID-19 pandemic). Finally, we introduced computational epidemiology, which involves the development of mathematical and computational tools to support the field, particularly in understanding the behavior of complex biological systems.

2 Epidemiology: An Overview

Epidemiology, derived from the Greek words *epi* (meaning "on" or "upon"), *demos* (meaning "people"), and *logos* (meaning "study"), is the scientific study of what happens to people within populations. This field primarily investigates the causes and control of epidemic diseases, although its scope extends beyond just infectious diseases to include non-communicable diseases and other health-related events [1].

2.1 Core Components of Epidemiology

2.1.1 Distribution

This component focuses on understanding how health-related events, such as diseases, are spread within a population. It is concerned with the patterns of health problems and their occurrences in various demographic groups. Distribution is typically analyzed on a large scale, providing a population-wide view of health trends and potential risks. The focus is not just on individual cases but on how the disease affects the population as a whole.

2.1.2 Determinants

These are the causes and factors that influence the occurrence and spread of diseases within populations. Determinants include biological factors, environmental conditions, social behaviors, and economic factors that contribute to the onset and progression of diseases. Identifying these determinants is crucial for understanding why certain populations are more vulnerable to specific health problems and for developing strategies to mitigate these risks.

2.1.3 Application

This involves the practical application of the knowledge gained from studying the distribution and determinants of diseases. The ultimate goal of epidemiology is to apply this

understanding to take immediate and effective actions to reduce harm and control the spread of diseases. This can include public health interventions, policy changes, and the implementation of preventive measures to improve health outcomes across populations.

2.2 Focus of This Course

The course primarily focuses on infectious diseases caused by microparasites - organisms that cause diseases through direct transmission. These include viruses, bacteria, and other pathogens that can spread rapidly through populations, leading to epidemics. The study of microparasite diseases is critical for understanding and controlling outbreaks, which can have devastating effects on public health.

3 Historical Foundations of Epidemiology

3.1 Early Epidemics and Responses

Epidemics have been a recurring challenge throughout human history, with recorded instances dating back to ancient civilizations. The Roman Empire, the Han Empire in China, and the Aztec civilization all faced devastating plagues that significantly impacted their populations. During these early times, responses to epidemics were often based on superstition and limited medical knowledge, relying heavily on magic, prayers, and rudimentary practices such as quarantining and social distancing.

In the 8th century, India and China developed early forms of variolation, a practice used to control smallpox by exposing individuals to a less virulent form of the disease. This practice was an early precursor to vaccination and marked a significant step towards the scientific understanding and control of infectious diseases.

3.2 Precursors to Modern Epidemiology

3.2.1 Smallpox: Bernoulli and Jenner

One of the earliest mathematical approaches to epidemiology was developed by Daniel Bernoulli in the 18th century. Bernoulli applied mathematical models to argue in favor of variolation, demonstrating that exposing individuals to cowpox (a milder disease) could effectively build immunity against smallpox. This approach laid the groundwork for Edward Jenner's later work, which led to the development of the first vaccine. Jenner famously inoculated a young boy with material from a dairymaid's cowpox lesions, which successfully protected him from smallpox. This was a groundbreaking advancement in medical science and a critical milestone in the history of epidemiology.

3.2.2 Cholera: John Snow

Another pivotal figure in the history of epidemiology is John Snow, often referred to as the father of modern epidemiology. In 1854, during a cholera outbreak in the Soho district of London, Snow challenged the prevailing miasma theory, which held that diseases were caused by "bad air." Instead, Snow proposed that cholera was waterborne. He meticulously mapped the locations of cholera cases and identified a cluster of infections near the Broad Street pump. Snow's analysis showed that neighborhoods drawing water from upstream sources (less contaminated by sewage) had significantly lower death rates compared to those

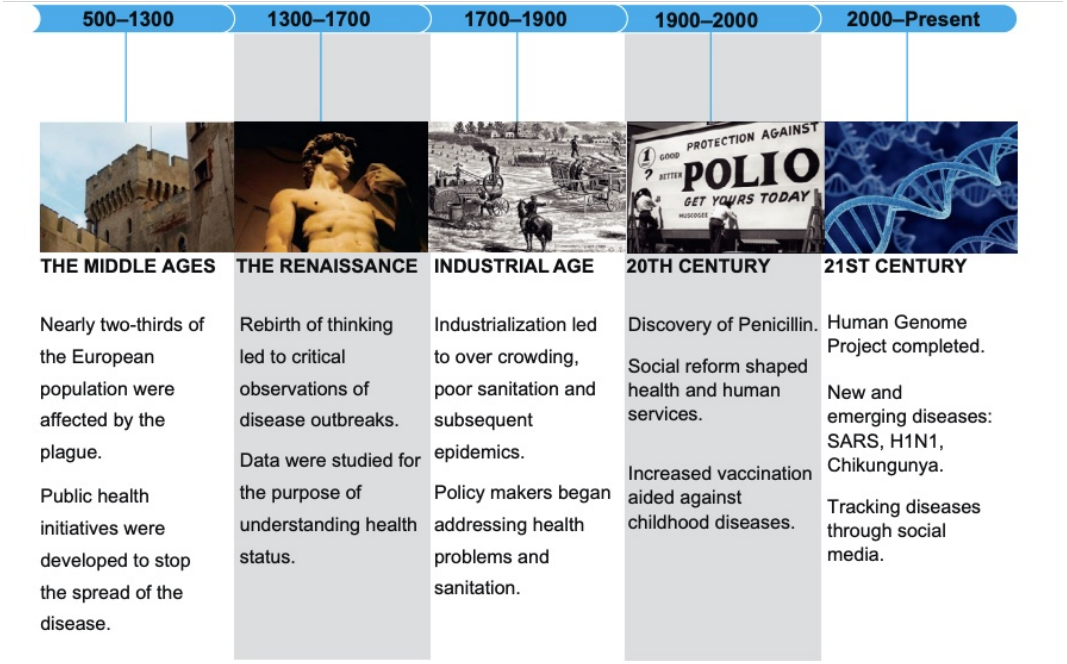

Figure 1: History of Infectious Diseases [1]

drawing from downstream sources. His recommendation to remove the handle of the Broad Street pump led to a dramatic decrease in cholera cases, providing compelling evidence for the germ theory of disease.

3.2.3 Malaria: Ross, McDonald, Lotka, and McKendrick

Malaria, an ancient and widespread disease, has also played a significant role in the development of epidemiological methods. Sir Ronald Ross, along with colleagues such as McDonald, Lotka, and McKendrick, made substantial contributions to understanding the transmission of malaria. Ross discovered that mosquitoes were the vector responsible for transmitting the malaria parasite. He developed a mathematical model to show that reducing the mosquito population would proportionally reduce malaria transmission. This work provided a scientific basis for vector control strategies, which remain a cornerstone of malaria prevention efforts to this day.

3.3 Epidemics in the Modern Era

The 20th and 21st centuries have witnessed several large-scale pandemics, including the 1918 influenza pandemic, the SARS outbreak in 2002-2003, the 2009 swine flu pandemic. These events have underscored the importance of epidemiology in understanding and controlling disease spread. We further examined the ongoing challenges in epidemiology, especially in the context of the COVID-19 pandemic. Issues like misinformation, the mobility of populations, and the complexity of modern societies were discussed as factors that complicate pandemic responses. The importance of data science in epidemiology was emphasized, par-

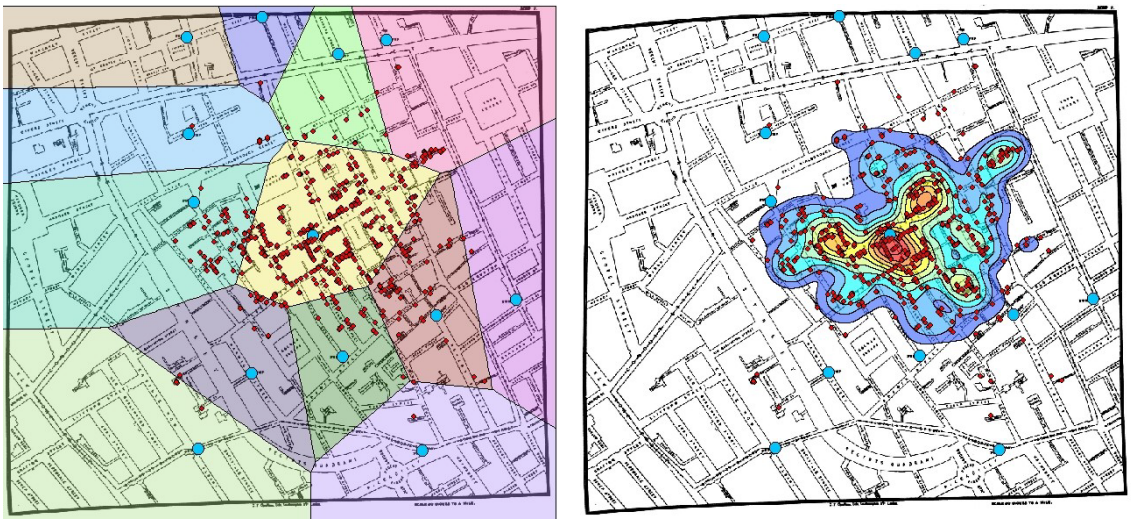

Figure 2: Voronoi map of wells [3]

ticularly in handling large, complex datasets and deriving actionable insights from them [2].

4 Epidemiological Problems and Data Science

The intersection of epidemiology and data science has become increasingly important, particularly in the context of modern pandemics. The challenges presented by these global health crises underscore the need for innovative approaches to data collection, analysis, and interpretation.

4.1 Challenges in Modern Pandemics

COVID-19 highlighted several key challenges that epidemiologists face today:

- **Lack of Information:** During the early stages of an outbreak, there is often a significant lack of information. This includes uncertainty about the source of the disease, the efficacy of treatments, the extent of infections, and the fatality rates. The lag between the initial spread of the infection and the availability of accurate data can hinder timely responses.
- **Inadequate Vaccination and Therapeutics:** The emergence of novel pathogens often outpaces the development of vaccines and therapeutics. This inadequacy leads to a critical need for efficient allocation of limited resources.
- **Logistical Complexities:** The logistical challenges of responding to a pandemic are immense. These include the distribution of medical resources, the management of healthcare facilities, and the coordination of public health interventions across different regions.
- **Modern Trends:** The modern world is characterized by highly mobile and densely populated areas, which are often filled with misinformation and conflicting interests. These factors complicate the response to pandemics, as they increase the potential for

rapid disease spread and make it more difficult to implement effective control measures [4].

4.2 Data Science in Epidemiology

Despite the challenges, the vast amount of data generated during a pandemic presents an opportunity for data science to play a pivotal role in understanding and controlling the spread of disease. The key to leveraging this data lies in the application of data mining principles, which can transform raw data into actionable insights.

4.2.1 Data Mining Principles

Data mining in epidemiology involves discovering patterns and models in large datasets that are:

- **Valid:** The findings must hold true when applied to new data with a certain level of certainty.
- **Useful:** The results should be actionable and able to inform public health decisions.
- **Unexpected:** Non-obvious insights that are not immediately apparent from the data.
- **Understandable:** The patterns and models discovered should be interpretable by humans, enabling clear communication of the findings.

The tasks involved in data mining can be categorized into:

- **Descriptive Tasks:** These tasks involve finding human-interpretable patterns that describe the data, such as clustering.
- **Predictive Tasks:** These tasks use variables from the data to predict unknown or future values of other variables, such as classification.

4.3 Epidemiology and Data Science Methodology

The methodology of integrating data science into epidemiology can be exemplified by John Snow's approach to understanding the cholera outbreak in London. His use of simple statistics, spatial analysis, and consideration of both positive and negative counterexamples laid the foundation for modern epidemiological methods.

Today, data science enables epidemiologists to:

- **Manipulate Data:** Large datasets require manipulation through computational techniques to make them suitable for analysis.
- **Analyze Data:** Statistical methods, machine learning, and other computational approaches are applied to extract meaningful insights from the data.
- **Communicate Results:** The findings must be communicated effectively to inform public health policies and interventions. This is critical, as the ability to communicate complex data-driven insights can influence public behavior and policy decisions.

4.4 Key Questions in Epidemiological Surveillance

Epidemiologists rely on data science to answer critical questions during an outbreak:

- **When, Where, and Who?** Determining the origin of an outbreak, identifying initial cases, and tracking the spread of the disease requires accurate and timely data from health agencies. However, real-time surveillance is challenging due to delayed or erroneous reports. To mitigate this, open-source surrogates such as internet searches, social media posts, and surveillance of hospital parking lots can provide more immediate, albeit raw, data.
- **What and When?** Forecasting the spread of disease, identifying what it is (and how it acts), and predicting when the outbreak will peak are essential for effective public health response. These forecasts depend on the integration of current data with predictive models.
- **How to Control?** Developing and implementing control measures, such as social distancing or vaccination strategies, requires careful consideration of available resources. Data science tools can help optimize the allocation of limited medical resources, such as determining which populations should be prioritized for immunization.

4.5 Real-Time Study and Data-Driven Interventions

Data science plays a crucial role in the real-time study of pandemics. Before an epidemic, modeling can help determine the necessary interventions and assess their feasibility. During an outbreak, real-time data analysis can inform adjustments to strategies as new information becomes available.

The integration of data science into epidemiology allows for a more dynamic and responsive approach to managing public health crises. By leveraging computational tools, epidemiologists can not only understand the current state of an epidemic but also predict its future trajectory and recommend effective interventions to mitigate its impact.

5 Conclusion

In conclusion, the study of epidemiology, particularly in the context of modern pandemics, highlights the crucial role that data science plays in understanding and managing public health crises. The increasing complexity of epidemic models, combined with the vast and often noisy data available, necessitates the use of advanced data science and statistical techniques. These tools enable the integration of diverse data sources, allowing for more accurate forecasting, real-time surveillance, and the optimization of intervention strategies. As we continue to face global health challenges, the intersection of epidemiology and data science will remain vital in developing effective public health responses and improving outcomes at both local and global scales. The big picture reveals that combining disciplines from machine learning and statistics to computational systems and social sciences not only advances scientific knowledge but also contributes to societal well-being, with the potential for both profit and progress.

References

[1] S. Pyne, A. K. S. Vullikanti, and M. V. Marathe. Chapter 8 - big data applications in health sciences and epidemiology. In V. Govindaraju, V. V. Raghavan, and C. Rao, editors, *Big Data Analytics*, volume 33 of *Handbook of Statistics*, pages 171–202. Elsevier, 2015.

[2] T. Rosenberg. Stopping pandemics before they start. *New York Times*, 2017.

[3] K. Rowell. ghost-map. https://healthdataviz.com/newsletters/ghost-map/, 2017.

[4] E. Yong. The next plague is coming. is america ready? *The Atlantic*, 2018.