CSE 8803 EPI: Data Science for Epidemiology, Fall 2024

Lecturer: Prof B. Aditya Prakash
Scribe: Arjun Verma (prev. Sayan Sinha)

October 3, 2024
Lecture #13: Surveillance I

---

1 Summary

Public health surveillance is the systematic collection, analysis and interpretation of health-related data for downstream use in predicting outbreaks, disease trends, guiding public health policy and other epidemiological tasks. A variety of data sources are used in surveillance, including traditional health records, disease reporting systems, and more recent digital sources like smartphone data and internet searches. These data sets offer complementary insights into disease dynamics and can provide early warnings of outbreaks. Recent advancements, particularly driven by the COVID-19 pandemic, have expanded the use of digital data in public health surveillance. This ongoing evolution ensures that surveillance systems remain effective in addressing emerging health challenges.

2 What is Surveillance?

The CDC officially defines public health surveillance as "the ongoing, systematic collection, analysis, and interpretation of health-related data essential to planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those responsible for prevention and control" – CDC

In a broader sense, we refer to 'surveillance data-sets' as the more comprehensive data that can aid in the prediction of epidemic outcomes rather than only health-related data. Other crucial elements for the propagation of disease are included in this:

- Behavioral factors include mobility and following public health advice
- Environmental elements like the climate

We can explore the diverse data sources that fuel epidemiological surveillance. These sources vary widely, influencing their application in disease tracking. For instance, mobility data from apps like Google Maps can help build social networks among populations, aiding in disease surveillance. Meanwhile, social media trends on platforms like Google and Twitter can offer insights into individual health states through sentiment analysis. To better categorize these varied datasets, we construct a data organization structure known as a Surveillance Pyramid.

3 Surveillance pyramid

The surveillance pyramid from top to bottom as shown in Figure 1; shows the many stages of a person's disease, with the area corresponding to the population. Our proposed taxonomy of data-sets utilized in the literature to inform forecasting models is connected to each of its levels, and some of their typical examples are shown. Direct surveillance of the disease's

transmission is represented on the left side of the pyramid. The right side of the pyramid represents the proxy measures of epidemiological indicators of disease transmission. The proxy measures act as surrogates when we don't have enough actual data from the sources. These include, search trends, mobility data, satellite images etc. Each are best suited for a particular subgroup/ state of the population which is marked as levels in the pyramid.

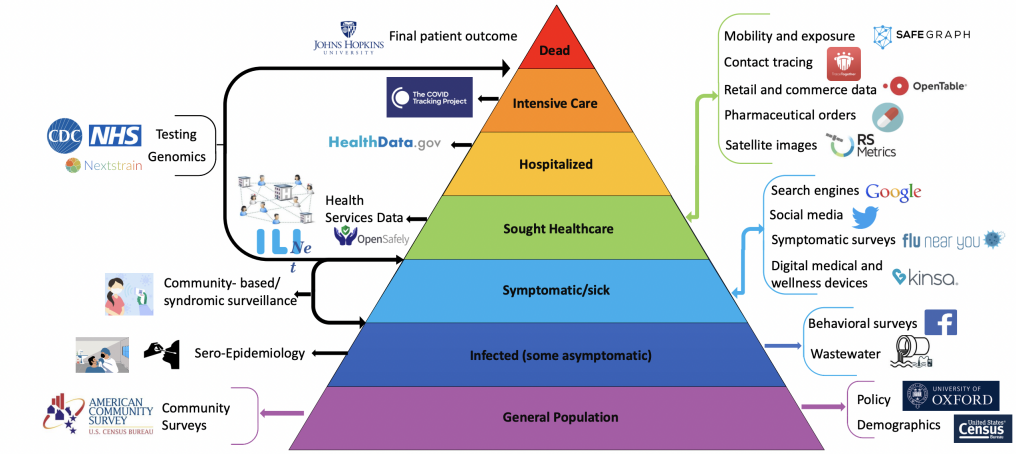


Figure 1: Surveillance pyramid describing the conceptualization of surveillance data sources. [13]

From Figure 1; At the bottom we observe the general population and corresponding health data collection methods like Community surveys and help in deciding demographic policy measures. Further we see that the population size in the pyramid becomes smaller as we go towards the end of the pyramid. The other end of the pyramid representing the deceased population due to the disease. Each of the levels in the pyramid and it's corresponding direct surveillance and epidemiological proxies are discussed as follows:

- **General population** : At the base of the pyramid lies the general population, our target for evaluating the predictive power of our model. This population will provide data for network construction and broader community-based inferences. We can gather this data through community surveys, which offer insights into the population's characteristics. Alternatively, proxy measures like census data or demographic studies can provide valuable information.

- **Infected population** : The subset of General population which have been infected by a particular disease forms the next level of the pyramid. For this set we have a lot of actual testing data coming from Sero-Epidemiology (which means lab testing, like a swab test). Moreover, we can also have data from proxy measures such as 'Behavioral surveys' over the social media platforms or sampling wastewater for disease pathogens and then relating them to infected people carrying those disease pathogens.

- **Symptomatic** : The next subset of the infected population are those who are infected and showing symptoms of the infection. Here we need not go for the Sero-test (lab test) as we can now test for the symptoms. For example, body temperature scan on a community basis can give us a good estimate of symptomatic individuals. Moreover, we can also have data from surrogate measures like search trends and symptomatic

surveys, where people who feel symptoms of any disease come and report in order to seek help from wellness advisors.

- **Sought healthcare** : Some of the symptomatic individuals go on to take healthcare advice/support from medical advisors, and they form the next level of the pyramid. Here, the actual data can be easily provided by the hospitals and medical support units. We can also have surrogate data from sources such as pharmacies, as most of the people who sought healthcare advice go on to buy medication. We can also utilize sources like Satellite imaging of hospital parking lots, where full parking lots imply full hospitals.

- **Hospitalised** : Among the symptomatic individuals who sought healthcare advice, some may need hospitalisation. Hospitalized individuals are very well documented by the hospital and thus the data is very high quality and surrogate sources aren't needed.

- **Intensive care** : Some hospitalised individuals will need further intensive care which can again be tracked down to a precise value and hence no need for surrogates.

- **Deceased** : Out of the people who received intensive care unit treatment some may end up deceased and again this can be tracked to a precise value since the previous level was up to a certain precision. Most of the epidemiological models which we build are tested at this data set only, since it is one the most validated data sets among all the data-sets discussed above.

As we delve deeper into the Surveillance Pyramid, we'll examine the specific datasets utilized at each level.

4 Datasets

To enhance epidemic forecasting, researchers have explored a diverse range of datasets, each offering unique advantages in capturing the nuances of disease spread and providing early warning signs. Recent advancements driven by the COVID-19 pandemic have led to the increased availability and exploration of novel data sources, such as smartphone data, internet search trends, and satellite imagery. In this section, we categorize and briefly describe the dataset corpora identified in our previous work.

- **Clinical Surveillance** :

    1. Line List and Testing: Classical method of surveillance in which the number of people in hospital waiting lists and testing results is taken into consideration. However, this type of surveillance isn't perfect. These datasets are sparse, and investigation and testing can be an expensive affair. Also, for specific diseases like flu and ebola, only specific people in the "risk" category are tested. Line lists are limited to people who have been specifically tested for the disease.
    2. Health Service Records: These are faster and for larger samples, estimated based on individual's symptoms and syndromes. They are not specifically carried out for certain individuals. It focuses on estimating the number of cases based on symptoms of all the people visiting a hospital, without actually testing for the

particular disease. This method is also called syndromic data surveillance, and helps as a leading indicator for an outbreak.

3. Electronic Health Records (EHR): This uses an individual's health records for the purpose of clinical investigations.

- **Digital Surveillance** : As mentioned earlier, surveillance is not only limited to disease propagation. There are other exogenous factors that might also indicate an outbreak through behavioural patterns or aid in surveillance.

  1. Online search and social media: Keywords used in search or posts on social media.
  2. Online surveys: Surveys and polls conducted online on social media platforms.
  3. Mobility and contact tracing: Tracking population mobility patterns through various networks.
  4. Retail and Commerce: Data from items being purchased from stores and online can give an idea on mobility patterns.

- **Novel data modalities** :

  1. Satellite Images: Hospital parking lot images.
  2. Genomics: Understanding how a disease would respond to certain special situations such as medications.
  3. Environmental: Understanding changes in environment that can bring about increase or decrease of an outbreak.

5 Clinical Surveillance

These data sets give firsthand information to conduct disease surveillance since they are derived from clinical information of patients (observation and treatment) by healthcare professionals and governmental agencies. The following are the various sources which contribute to Clinical Surveillance.

5.1 Line List and Testing :

The earliest datasets used in conventional epidemiology were these. Line lists are individual records that detail **who, when, and where** an infection occurred as well as **how many infected, recovered, and deceased** individuals there were. Public health organizations all across the world gather, assemble, and swiftly publish line lists because they are information of general interest. For instance, the National Health Services (NHS) in the UK [14], the Centers for Disease and Control (CDC) in the US [6], and state-level public health ministries in India [9].

During the COVID-19 pandemic, government and public health initiatives significantly enhanced testing efforts, making COVID-19 datasets more valuable compared to those for other communicable diseases. Virologically confirmed cases directly reflect the disease's spread. For example, an increase in testing results might indicate heightened efforts by local authorities and healthcare providers to mitigate the disease's spread. These datasets can be effectively utilized for epidemiological predictions. However, it's important to note

that line test datasets may not always accurately represent the true case scenario. As illustrated in Figure 2, reported COVID-19 cases often precede deaths, and underreporting can diminish the reliability of case data compared to death data.

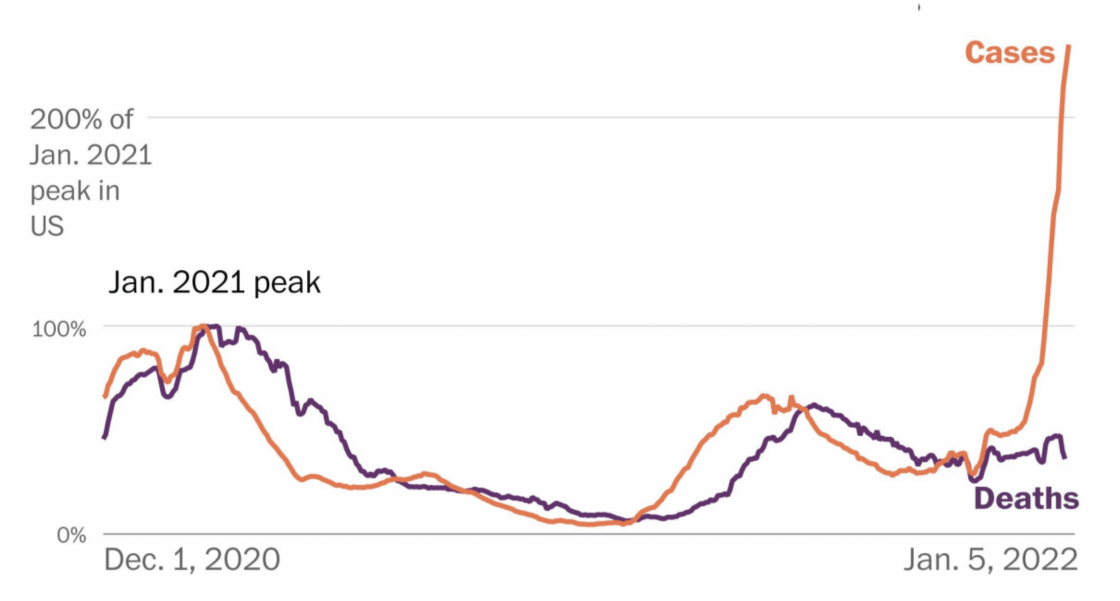


Figure 2: Time series COVID data regarding cases and deaths in the US. The disparity in cases and deaths at the end of the curve is due to the introduction of the COVID vaccine, which means that the number of cases increases a lot but people aren't dying from it.

The other negative point can be that investigation and testing are typically expensive and time-consuming. COVID 19 testing is being done widely thanks to significant government financing available even to those who are symptom-free In contrast, only those who meet stringent requirements based on risk factors and symptoms get tested for the flu and Ebola.

5.2 Health Service Records :

Health Service Records provide a cheaper and larger alternative to line lists. These databases are compiled from the service records of patients who visit healthcare facilities for medical attention. They can be separated into inpatients and outpatients (hospitalized vs non-hospitalized patients). Data is largely collected based on symptoms rather than testing a particular disease, thus this method of collecting data is also known as **syndromic surveillance**. The focus of symptomatic surveillance is on one or more symptoms rather than a condition that has been confirmed.

An example for such a dataset is the **influenza-like illness (ILI)** counts, which the CDC gathers via the US Outpatient **Influenza-like Illness Surveillance Network (ILINet)** and aggregates from healthcare providers across all US states and territories. It calculates the proportion of people seeking medical attention who have flu-like symptoms, which are described as "fever (temperature of 100°F/37.8°C or above) and a cough and/or sore throat without a known cause other than influenza."

We have discussed both traditional surveliance as in line tests and Syndromic surveillance via health service records. Among these two there is general agreement that syndromic

surveillance can help with early identification and forecasting, but no one recommends it as a substitute for traditional disease surveillance. An example of this is syndromic surveillance of the flu. Influzena affects around a million Americans annually and hundreds of thousands need hospitalisation, with thousands dying each year. Testing for this is expensive and only done in exceptional circumstances. Instead, we can use ILI as a more realistic solution.

As shown below in Figure 3, the CDC utilizes different multimodal methods of surveillance, collecting ILI both over time and over space (in different states). Because of ease of collection, ILI datasets can be multimodal and rich.

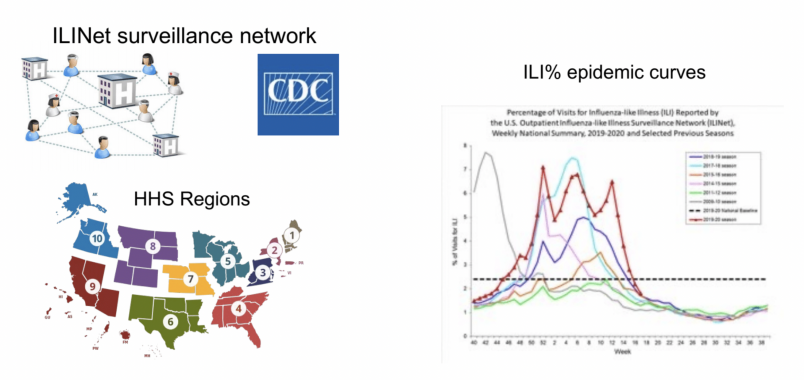


Figure 3: Various line list and testing methods, such as datasets from Hospital records, Lab surveys and Population surveys. [Source: Washington Post]

Another common challenge is people choosing not to report. An example of this can be seen below in Figure 4, where ILI numbers decrease unexpectedly on peak holidays.

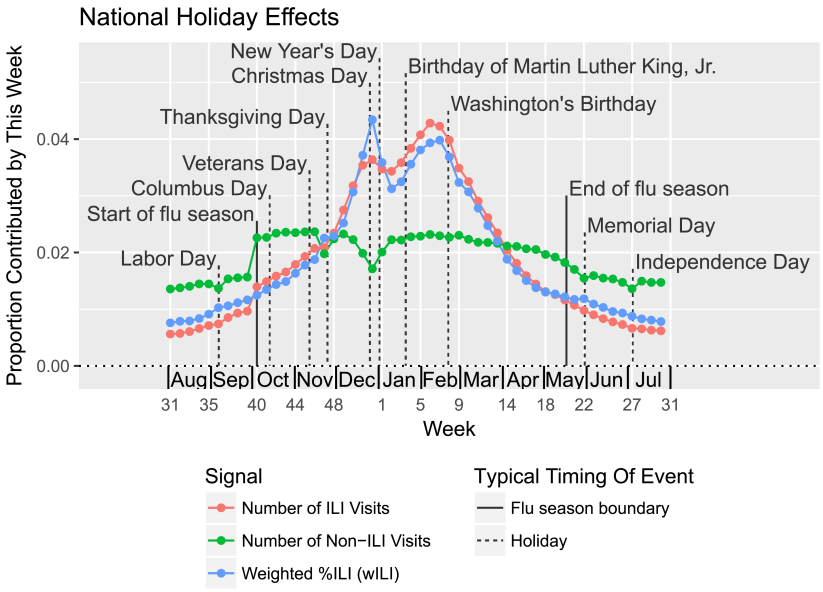
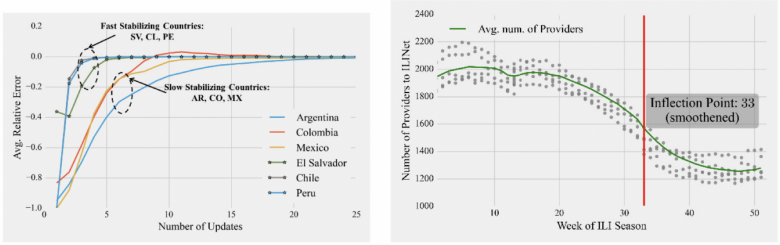


Figure 4: ILI on US holidays. People tend to report less on major holidays, so while increased mobility may increase the disease, less people report, leading to unexpected drops.)[2]

Yet another common challenge is data instability and inconsistency. Data can be revised and can take a long time to stabilize, as well as this surveillance data collection practices aren't uniform across space and time. An example of this can be seen below in figure 5.



(c) Surveillance Instability (d) Surveillance drop-off

Figure 5: Surveillance reports can be revised many weeks after the first report. While countries like Chile stabilize quickly (within 5 weeks), other countries like Argentina stabilize after many weeks (10). (d) Surveillance drops off towards the end of the season—scatter plot of number of providers reporting to CDC ILINet as a function of ILI season week. The Green Line shows the smoothed average while the red vertical line shows the smoothed inflection point of surveillance coverage.[4]

5.3 Electronic Health Records (EHR) :

Electronic health records (EHR) are more thorough datasets that include specific patient data. Although this information has been extensively used in clinical investigations, public health forecasting is still in its infancy. The most important aspect for the downside of this method is its issues with Privacy. Since the dataset is collected at individual level, it needs approvals of the concerned person which makes it complicated to use. Recent studies have used EHRs like OpenSAFELY from the NHS to study the clinical aspects related to COVID-19 [15]. These kinds of studies pave the way for further investigation of EHR databases through a transparent, secure method based on privacy. Other instances include Zhang et al. [16], who create effective interventions for preventing epidemic spread using contact networks and EHR data.

6 Digital Surveillance

Edge devices like mobile phones and smart watches have become ubiquitous in today's age. With sophisticated devices and advances in digital communication comes a huge opportunity of electronic surveillance which can provide real-time access to useful data. This data can be used to complement clinical data and provide reliable outbreak detection. However, in most cases this data must be publicly available, meaning that we can't always use the best sources (i.e. using tweets over Facebook information).

6.1 Online Search

Trends from search engines has proved to be a good surveillance method. This method uses trends from websites like Google [7], Yahoo [12] etc. to detect epidemics. Google Flu

Trends was an effort by google to predict flu outbreaks. However that failed during the H1N1 pandemic and was discontinued later.

Since then, Google has released data sets of search trends with differential privacy [1]. It consists of the top 500 symptoms since 2017 and has a county level spatial resolution as well as a daily/weekly temporal resolution.

Specialized search engines are also used to predict outbreaks. Wikipedia has been used [11] to estimate prevalence of Influenza-Like Illness in the United States in Near Real-Time. Other search engines like UpToDate are used widely by health practitioners for this use case. However, this data is only temporal and not spatial. It also needs linguistic and statistical post processing to be useful.

6.2 Social Media

From the perspective of forecasting epidemics, News, Opinions, Tweets and Blogs are all great sources for real-time electronic surveillance at scale. Famously, Twitter posts have been used for surveillance [5] by tracking the number of tweets with flu related keywords.

Health specific social media like healthMap have a database of RSS feeds with health-related content. Keller et al leveraged a webscraper [8] that collected thousands of RSS feeds on medical articles - they parsed the HTML structure of the documents to extract information such as date, headline, summary and location.

6.3 Symptomatic Surveys

Access to internet and mobile devices have made online symptomatic surveys highly accessible to the general audience. Some examples are Flu Near You in the USA and Dengue Na Web in Brazil. Facebook randomly invited users to participate in a COVID 19 survey which consisted questions about symptoms, behavior and accessibility. This was called the COVID-19 Trends and Impact Survey.

6.4 Mobility

With location tracking present on smart phones and smart watches, Mobility data becomes easily available with the aggregated movement between regions being a good indicator of how a disease might spread. Exposure measures the density of people in location of interests. It can be measured by the number of devices in a particular location. This can be obtained from sources like mobile phone records, GPS location, etc.

Google mobility uses location history with differential privacy when people use Google services that leverage GPS, which capture mobility patterns at country, state and county levels. SafeGraph leverages GPS data to measure visitor counts, dwell times, distance traveled to locations of interest and provide anonymized data for modeling mobility.

Contact tracing is tracking of patients which have been exposed to a particular disease. It is used to track the spread of infections among individuals via proximal contact. Recent advances in digital technology which have Bluetooth and GPS capabilities allowed epidemiologists to build peer to peer and centralized contact tracing.

6.5 Retail and Commerce Data

Data from retail and commerce can also be useful in predicting outbreaks as surveillance data. For instance, the OpenTable dataset tracks reservation at restaurants in North Amer-

ica. It was found that increase in restaurant table cancellations (perhaps because they were overcrowded) was associated with an increase in disease incidence, specifically influenza - like illness (ILI) as seen in Figure 6.

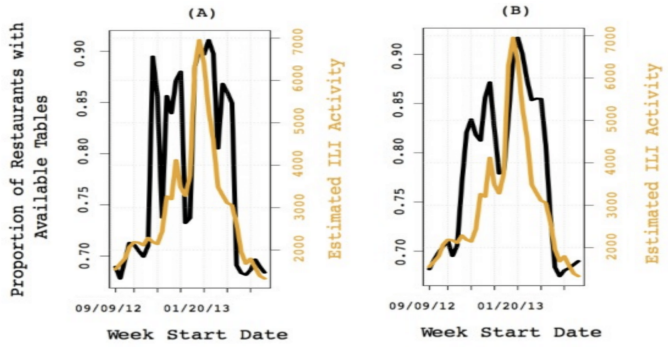


Figure 6: Number of available tables vs ILI

7 Novel data modalities

7.1 Satellite images

Images from satellites from space can also be used for surveillance. RSmetrics is a company which uses the images (like the one shown below) collected from remote sensing satellites to track COVID-19 and influenza outbreaks [13]. They published a paper with Butler et al [3] to track the number of cars in the parking lots of hospitals and other strategic locations along with temperature, humidity and other environment factors to track vector-borne illnesses like cholera, hantavirus, and malaria. An application of this type of technology can be seen below in Figure 7, which displays time series hospital car traffic data in Wuhan right before COVID.

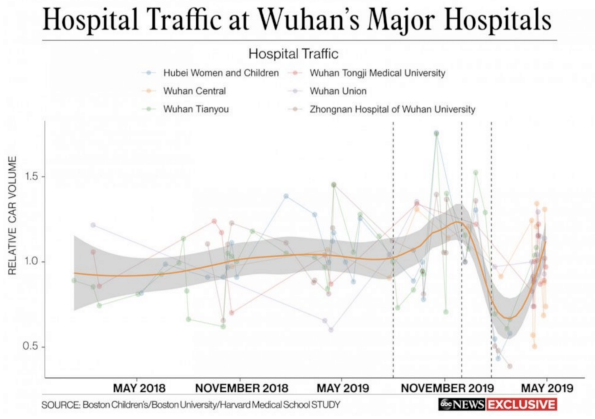


Figure 7: Hospital traffic in Wuhan: We can see a clear dip in car volume at the onset of COVID, which may be due to governmental restrictions on mobility at the start of the pandemic. This is clearly counter-intuitive, as one would expect an increase in hospital during a pandemic.

7.2 Genomics

Genomic epidemiology links pathogen genomes with the associated metadata to understand disease transmission. This is used in outbreak response to understand how the strain would respond with seasons, weather and medication. This is also useful in qualitative forecasting. Datasets like NextGen include multiple pathogen genomes and their mutations. There are several genomic repositories like GSAID, GenBank, COG-UK which are also used to perform the same task.

7.3 Environmental Sources

Changes in the environment such as temperature and humidity can lead to changes in outbreak. This is due to multiple factors, such as how the immune system is affected by changes in temperature or how weather can impact how people interact. One such example of a environmental dataset is the Microsoft Premonition Project that tracks the spread of diseases via mosquitoes, whose populations are affected by the environment. [10].

7.3.1 Meteorological

Weather and temperature are important markers for detection of onset of Influenza like Illness. This heuristics influence transmission especially in the tropical regions.

7.3.2 Zoonotic

Many Infections originate from animals and transfer over to humans by animal vectors. Identifying and tracking hotpots of wildlife where zoonotic diseases are more likely to appear is very relevant for the early detection of zoonotic diseases.

7.3.3 Wastewater data

Wastewater can be analyzed for markers of epidemic pathogens. This is a useful measure for community wise affliction of disease. The results of wastewater analysis have the potential to predict an outbreak earlier than traditional epidemiological indicators.

References

[1] S. Bavadekar, A. Dai, J. Davis, D. Desfontaines, I. Eckstein, K. Everett, A. Fabrikant, G. Flores, E. Gabrilovich, K. Gadepalli, S. Glass, R. Huang, C. Kamath, D. Kraft, A. Kumok, H. Marfatia, Y. Mayer, B. Miller, A. Pearce, I. M. Perera, V. Ramachandran, K. Raman, T. Roessler, I. Shafran, T. Shekel, C. Stanton, J. Stimes, M. Sun, G. Wellenius, and M. Zoghi. Google covid-19 search trends symptoms dataset: Anonymization process description (version 1.0), 2020.

[2] L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, and R. Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS computational biology*, 14(6):e1006134, 2018.

[3] P. Butler, N. Ramakrishnan, E. O. Nsoesie, and J. S. Brownstein. Satellite imagery analysis: What can hospital parking lots tell us about a disease outbreak? *Computer*, 47(04):94–97, apr 2014.

[4] P. Chakraborty, B. Lewis, S. Eubank, J. S. Brownstein, M. Marathe, and N. Ramakrishnan. What to know before forecasting the flu. *PLoS computational biology*, 14(10):e1005964, 2018.

[5] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 115–122, New York, NY, USA, 2010. Association for Computing Machinery.

[6] C. for Disease Control and Prevention. Centers for disease control and prevention. 2020. the national respiratory and enteric virus surveillance system (nrevss). `https://www.cdc.gov/surveillance/nrevss/index.html`. Accessed: 2010-09-30.

[7] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, Feb. 2009.

[8] M. Keller, M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein. Use of unstructured event-based reports for global infectious disease surveillance. *Emerging Infectious Diseases*, 15(5):689–695, May 2009.

[9] R. Laxminarayan, B. Wahl, S. R. Dudala, K. Gopal, C. Mohan B, S. Neelima, K. Jawahar Reddy, J. Radhakrishnan, and J. A. Lewnard. Epidemiology and transmission dynamics of covid-19 in two indian states. *Science*, 370(6517):691–697, 2020.

[10] A. Linn. Building a better mosquito trap. *International Pest Control*, 58(4):213, 2016.

[11] D. J. McIver and J. S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS Computational Biology*, 10(4):e1003581, Apr. 2014.

[12] P. M. Polgreen, F. D. Nelson, G. R. Neumann, and R. A. Weinstein. Use of prediction markets to forecast infectious disease activity. *Clinical Infectious Diseases*, 44(2):272–279, Jan. 2007.

[13] A. Rodríguez, H. Kamarthi, P. Agarwal, J. Ho, M. Patel, S. Sapre, and B. A. Prakash. Data-centric epidemic forecasting: A survey, 2022.

[14] G. UK. Coronavirus (covid-19) in the uk, 2020.

[15] E. J. Williamson, A. J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C. E. Morton, H. J. Curtis, A. Mehrkar, D. Evans, P. Inglesby, et al. Factors associated with covid-19-related death using opensafely. *Nature*, 584(7821):430–436, 2020.

[16] Y. Zhang, A. Ramanathan, A. Vullikanti, L. Pullum, and B. A. Prakash. Data-driven immunization. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 615–624. IEEE, 2017.