CSE 8803 EPI: Data Science for Epidemiology, Fall 2022

Lecturer: Prof B. Aditya Prakash
Scribe: Sayan Sinha

October 5, 2023
Lecture #13: Surveillance I

1 Summary

Public health surveillance is the ongoing systematic collection, analysis and interpretation of health-related data. It is extremely important as it captures the progression of a disease outbreak, as well as possible future pandemics. Planning, implementing, and evaluating public health practices require the ongoing, systematic collection, analysis, and interpretation of health-related data. This involves data collection, analysis, interpretation, dissemination and timely distribution of the analyzed data to those in charge of prevention and control. In order to improve epidemic forecasting, a variety of data-sets are in use, with varying advantages and precision including capturing complementing aspects that better characterize the dynamics of disease dissemination to early-stage indications of disease outbreaks. New data sources like smartphones, internet search engines, and satellite photos are now being explored and made more widely available as a result of recent initiatives, which were particularly sparked by the COVID-19 epidemic. The lecture categorizes and briefly describes the data-sets which are currently being used in public health surveillance.

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 now look at the various data sources which are are used in epidemiological surveillance! These data source come from diverse backgrounds and thus their application in the field of disease surveillance varies! For example, Mobility data from an app like Google maps, may help in disease surveillance by constructing a network among the concerned population, whereas social media trends like google and twitter trends may help in disease surveillance by determining the health-state of a person through sentiment analysis. In order to better classify these various data-sets; we study them by constructing a data-source pyramid called 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.

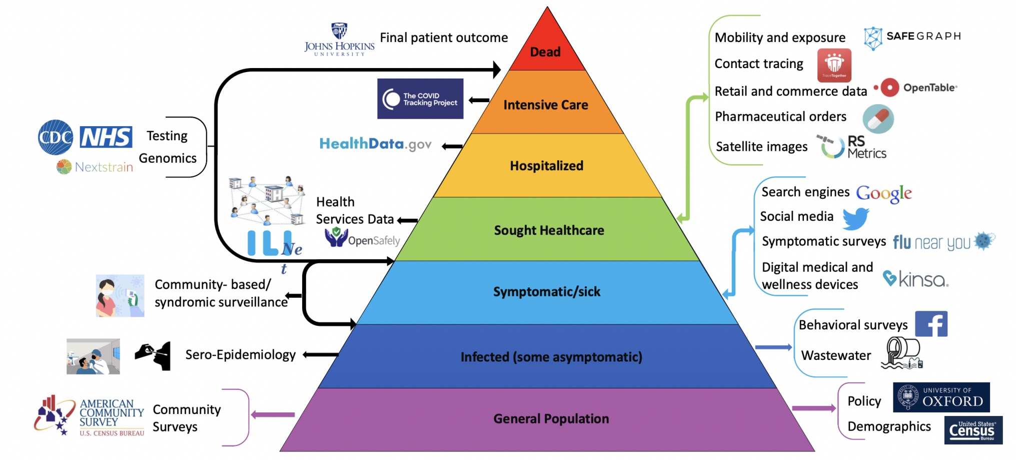


Figure 1: Surveillance pyramid describing the conceptualization of surveillance data sources. [13]

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** : The bottom of the pyramid represents the general population, this is the population on which we wish to study the prediction capabilities of our model. This would essentially provide the data for network construction and other general population based inferences! The dataset can be retrieved from community based surveys, which have the know about of the population in the community or else it can be determine through proxy measures such as census studies or other demographic studies!
- **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 that 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 provide us the need full data of the 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. While, we can also have surrogate data from sources such as Medicine shops, as all the people who sought healthcare advice may go on to buy some medication. Or from sources like Satellite imaging of the parking lots of the hospitals, the more number of vehicles in the parking lot, more the number of people seeking healthcare advice!

- **Hospitalised** : Among the symptomatic individuals who sought healthcare advice, some may need hospitalisation. And when they are being admitted to the hospitals they are always marked on the data entry of the corresponding hospital. This data-set is almost complete and requires little to no surrogate source as such!

- **Intensive care** : Some of the Hospitalised individuals will need further Intensive care which can again be tracked down to a precise value and hence no need of surrogates as such.

- **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. Also to note, that most of the epidemiological models which we build are tested at this data set only, since it is one the most validated data set among all the data-sets discussed above.

In discussing the Surveillance pyramid we see various datasets being used at each level, thus we now discuss the various data-sets of the pyramid in much detail.

4 Datasets

In order to improve epidemic forecasting, a variety of datasets have been used, with advantages ranging from capturing complementing aspects that better characterize the dynamics of disease dissemination to early-stage indications of disease outbreaks. New data sources like smartphones, internet search engines, and satellite photos are now being explored and made more widely available as a result of recent initiatives, which were particularly sparked by the COVID-19 epidemic. We categorize and briefly describe the dataset corpora we discovered in earlier work in this section. The categorisation of the list of the datasets can be done as follows:

- **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.
    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.
    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 ones. 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].

In the case of Covid-19, due to the efforts of government and public to enhance testing these datasets have become a lot more meaningfully than they are for the other communicable diseases. Since the overall number of cases and the negative cases can be utilized to comprehend societal and policy implications, virologically confirmed cases are direct indicators of the disease's spread. For instance, an increase in testing results reporting may be a reflection of efforts made by local government and healthcare professionals to slow the spread of disease. This way these datasets can be extensively used in epidemiological predictions. But on the other hand, there are several downsides of using these datasets. Such as, not all the dataset of line test may reflect the real situation of the case. As can be seen in Figure 3, the reported covid cases lead the number of deaths, along with that Underreporting makes number of cases a less reliable data set than the number of deaths. Hence, we should always test the quality of these datasets before implementing them!

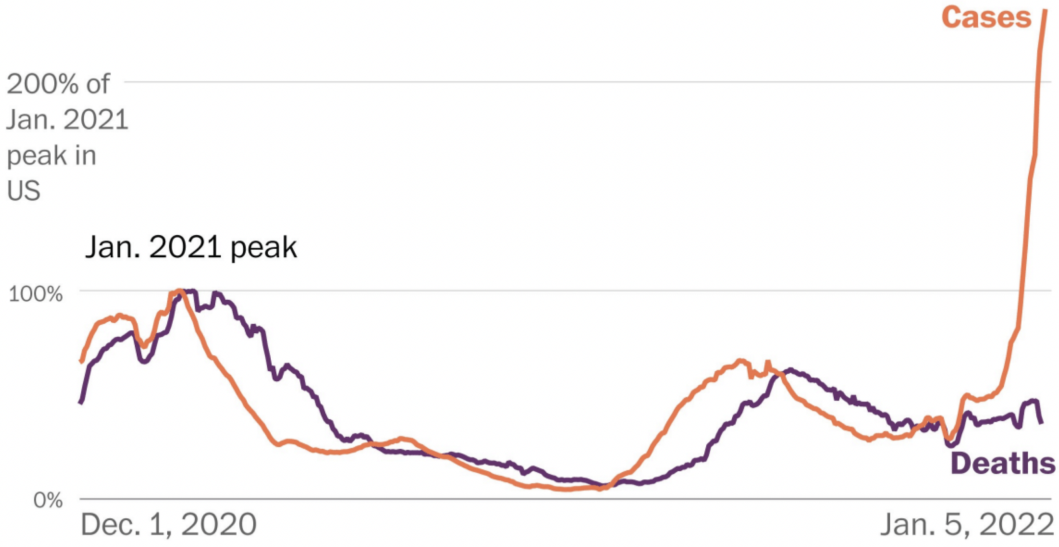


Figure 2: The various line list and testing methods, such as datset from Hospital records, Lab surveys and Population surveys. This figure does not come down even after flu season ends because of surveillance data mixing flu with covid symptoms.

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 :

A faster and larger dataset than line testing is collected by Health service providers. 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 (patients who are not hospitalized) (hospitalized patients). They are largely collected based on symptoms rather than testing a particular disease, the healthcare providers would in general note down the symptoms of any incoming or outgoing patient, thus they are also called **syndromic surveillance**. The focus of symptomatic surveillance is on one or more symptoms rather than a condition that has been identified by a doctor or tested and proven.

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, are a prominent outpatient-based indicator. 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 surveillance 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 for this can be Syndromic surveillance of Flu (which is also mentioned in the previous paragraph). Flu in US affects around a Million people each year and hundred’s of thousand need hospitalisation, with thousands dyeing each year! Testing for this is expensive (specially at such a large scale) and only done in exceptional circumstances (e.g., serious cases, hospitalized); instead we use ILI datasets with the help of cheap adn fast syndromic surveillance as mentioned in the previous paragraph.

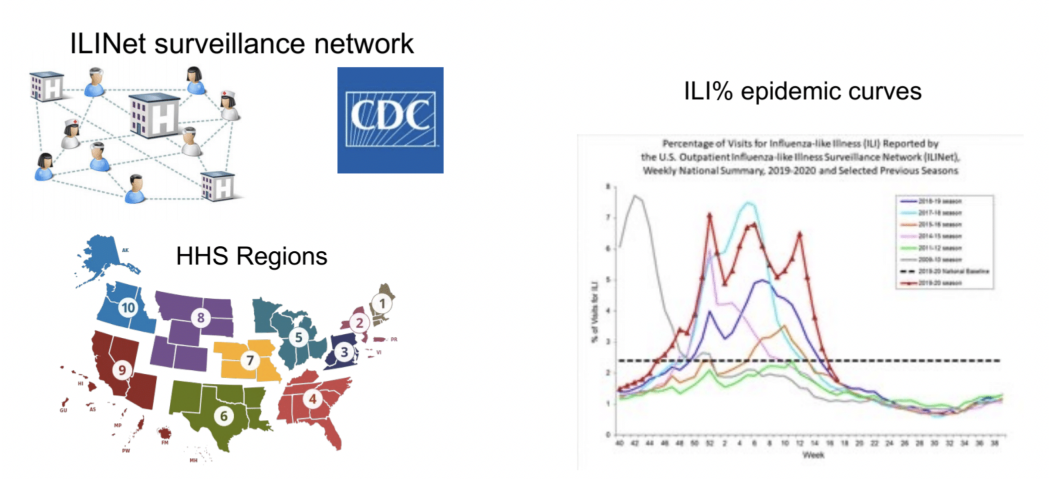


Figure 3: The various line list and testing methods, such as datset from Hospital records, Lab surveys and Population surveys. [Source: Washington Post]

As shown in Figure 4, the CDC divides US into various HHS regions based on geography. Data from a sample of healthcare providers was gathered by the CDC and provided at the state and HHS region levels. Here we also see the ILI epidemic curves as a Epiweek time-series for the various flu season. Some of the challenges which we might occur while using Syndromic survelliance datasets are, biased and incomplete datasets. Delay in reporting of cases (either by individuals or by testing platforms)

Other challenges may include The data can be revised several times, as we are not being provided the full data. The CDC gives out data till the time, and always revise the data according to the new situation. Thus, it takes several weeks to reach an stable value with this method! Other than that Surveillance data collection practices are not uniform in time, thought the epidemic! Like, when the peak passes away people change their attitude towards the disease and thus it becomes more biased towards the ends (as shown in Figure

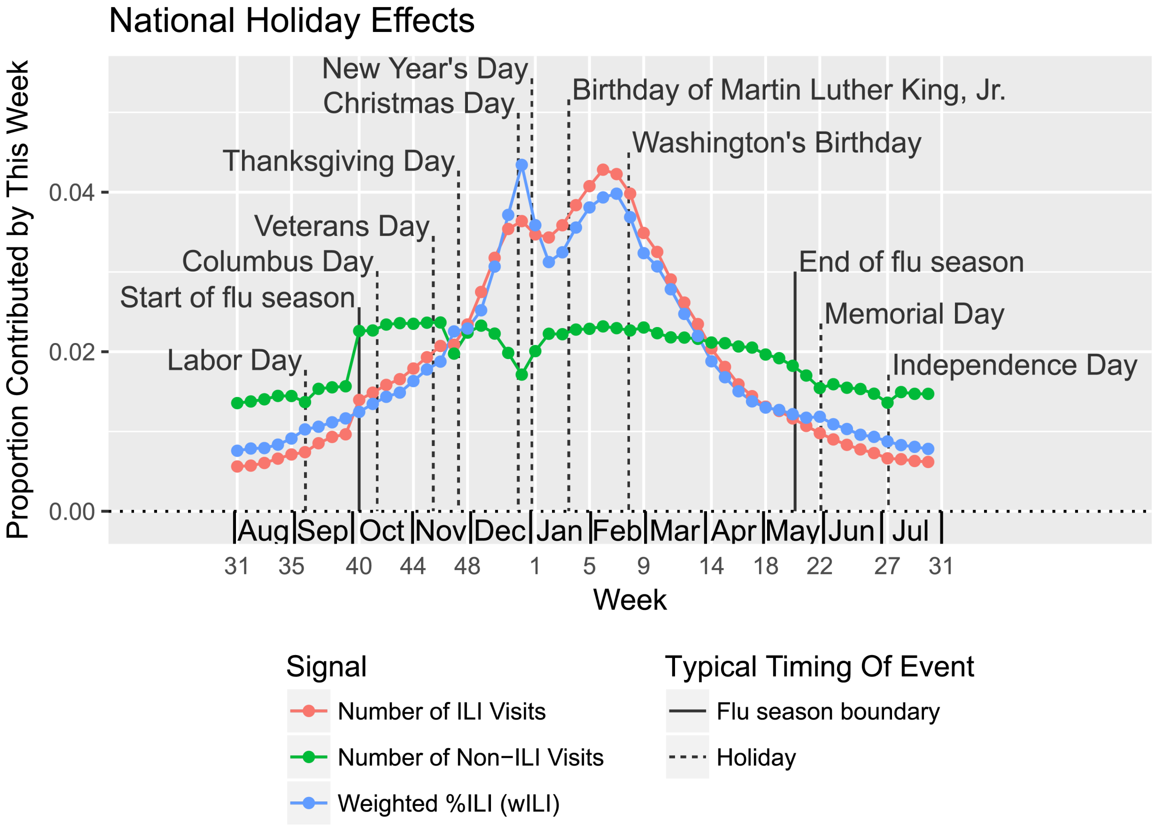


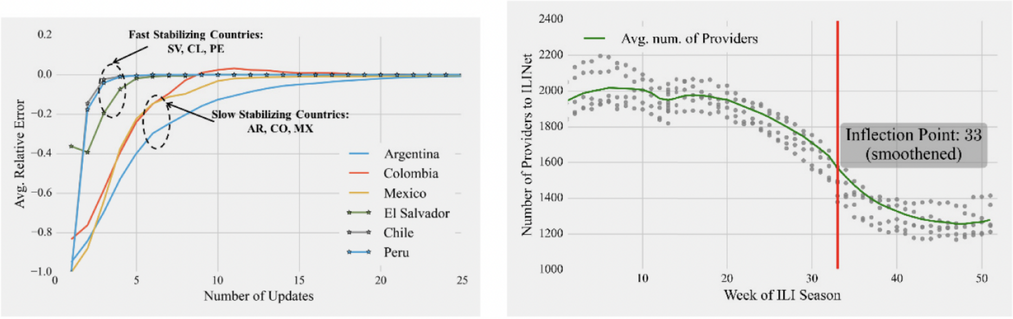
Figure 4: On average, wILI(weighted ILI) is higher on holidays than expected based on neighboring weeks. Weekly trends in wILI values, as expressed by the contribution of a each week to a sum of wILI values from seasons 2003/2004 to 2015/2016, excluding 2008/2009 and 2009/2010 (which include portions of the 2009 influenza pandemic), show spikes and bumps upward on and around major holidays. (U.S. federal holidays are indicated with event lines.)[2]

6) and hence we require to take into that too while modelling the disease and the number of cases!

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.

And the most important aspect for the downside of this dataset is the issues with Privacy. Since, the dataset is collected at individual level, hence 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.



(c) Surveillance Instability

(d) Surveillance drop-off

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

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.

These digital data sources include data created for sharing like twitter or not like Google searches.

6.1 Online Search

Trends from search engines has proved to be a good surveillance method. It uses the trends from website 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 the, Google has released data sets of search trends with differential privacy [1]. It consists of top 500 symptoms since 2017 and has a county level spatial resolution and 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 which is a database used by over 700,00 health practitioners around the world and filtered the relevant search query volumes for relevant terms. However, this data is only temporal not spatial. It also need linguistic and statistical post processing to be useful.

6.2 Social Media

From the perspective of forecasting epidemics, News, Opinions, Tweets, Blogs are a great source 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 has 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 and scalable to the general audience. The surveys can easily reach the public and they can fill out the questionnaire about the symptoms which them or their familiy members are experiencing. 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 and the aggregated movement between regions is good indicator of how the 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.

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 the patients which have been exposed to a particular disease. It is used to track spread of infections among individuals via proximal contact. The recent advances in digital technology which can leverage Bluetooth and GPS to build peer to peer or centralized contact tracing. Recent works shows that WiFi logs can also be used in contact tracing.

6.5 Retail and Commerce Data

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

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

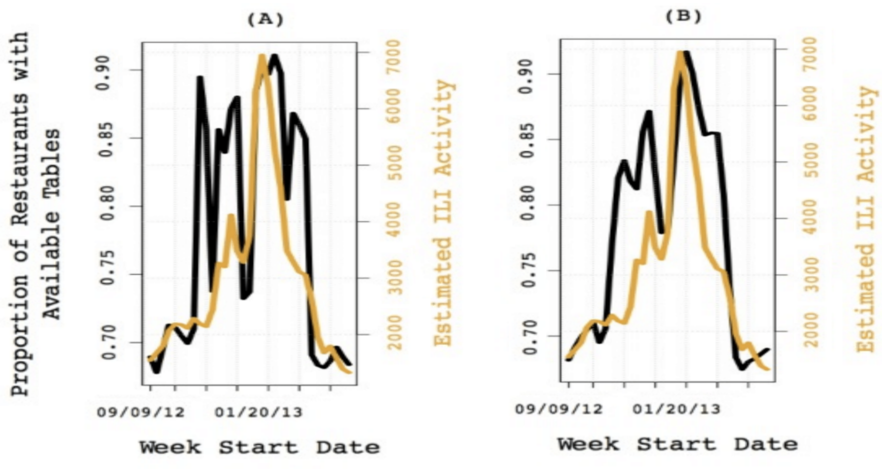


Figure 6: Number of available tables vs ILI

with temperature, humidity and other environment factors to track vector-borne illnesses like cholera, hantavirus, and malaria.

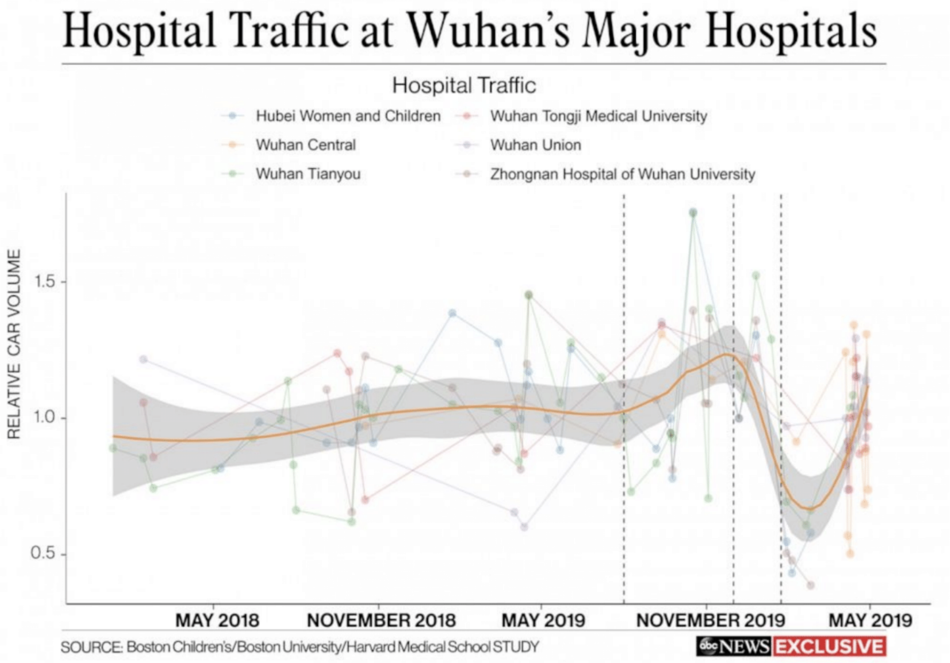


Figure 7: Hospital traffic in Wuhan: Perhaps they were already getting relevant symptoms but they were not aware its due to covid.

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 have the 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. Historical evidence of such changes might be evident that could aid in annual prediction changes (eg. Flu). Moreover, there might be other environmental factors, like the presence of pathogens. One such example of a dataset is the Microsoft Premonition Project that tracks the spread of diseases via mosquitos [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. It is important to identify and track hotpots of wildlife where zoonotic diseases are more likely to appear is very relevant for early detection like Bats for Covid-19.

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.