REVIEW ARTICLE

Reviewing the Quality of "Big Data" in automatic data systems: An Example

Author:

Tom Koch, PhD
University of British Columbia
Dept. of Geography (medical)
Vancouver, BC. Canada

Alton Medical Centre
1302 Queen St. E.
Toronto, ON. Canada

http://kochworks.com
tomkoch@kochworks.com; tom.koch@geog.ubc.ca

Abstract

In recent decades there has been an extraordinary growth in and acceptance of automatic data systems that collect official and popular reports of epidemic occurrence. While different systems employ one or another proprietary algorithms to collect and parse disease reports all include, at a minimum, spatial locators, the date of a report, and the number of individual cases reported. These systems have been increasingly vital in both the study of individual epidemics and the exposition of expanding epidemics in real time. To date, however, there has been little analysis of the nature and quality of the data collected in these "big-net" programs or the degree to which redundancies and uncertainties may limit their utility. Here data on the 2009 H1N1 Type-A influenza epidemic gathered by a single system, healthmap.org, is parsed to determine where problems exist and how they might be rectified.

Keywords: Big Data, epidemic disease, H1N1 Influenza, spatial cartography, syndromic data surveillance.

Introduction

COVID-19 is the most mapped disease event in history. For the first time, details of a global pandemic have been tracked publicly in a constant progression of continually updated maps broadcast and published daily, and with attendant statistics. These have been posted first on one or another digital “dashboards” and then rapidly disseminated in magazine stories, newspaper articles and TV broadcasts. Best known, perhaps, is the Johns Hopkins University Coronavirus Center (2020) produced by its Center for System Science and Engineering. This and other sites permit the scaling of maps— global to national or regional— with an option to download the underlying data. Some dashboards permit web-generated maps to be manipulated with simple ratios and percentages, to create, for example, mortality ratios (ArcGIS.com 2020).

The basic format of these dashboards—a central map with attendant relevant statistics—has been widely copied by health professionals and researchers at varying scales. A dashboard for Milwaukee County (WI), for example, includes not only a dot map of incidence—cumulative and weekly—but a breakdown of that data by age, gender, race/ethnicity and hospital capacity reported by local health agencies (Milwaukee County 2020).

This explosion in a spatial cartography both public and rigorous has resulted in the democratization of what was, until recently, a specialized, technical area of epidemiology and medical geography. While epidemics have been mapped at least since the 1690s, and global pandemics at least since the 1830s (Koch 2011; 2017), their data was usually laboriously collected, typically by official agencies, for articles that were later published in books or professional journals. What makes the current dashboards and their mapping possible are automatic data collection programs that continually search a range of sources for data that can be immediately incorporated in an online posting (Gilbert, Degeling, Johnson 2019: Garattini, Raffle, Aishah, Kozlakis, 2019). Driven by complex proprietary algorithms, these syndromic programs are the result of the digital revolution that resulted both in modern, desktop cartography and programs (SAS, SPSS, R, etc.) facilitating statistical analytics (Kramer, Hay, Pigott, et al. 2016).

In what some geographers are called the new “digital earth” the result for spatial epidemiologists is a series of event based datasets with records of specific disease incidence by data, location (for example, Vancouver), and jurisdiction (BC, Canada). Each entry includes spatial coordinates to facilitate mapping. Some may also include, for each incident, demographic, epidemiological, genomic, migratory, or socioeconomic data (Polonsky, Baidjoe, Kamvar, et al., 2019).

A range of authors have offered reviews of these machine-based, automatic data collection systems and the tools by which their data might be analyzed (Feldman, Thomas-Bachli, Forsyth et al., 2019; Mehta, Pandit, 2018; Kraemer, Hay Pigott, Smith et al., 2016). What has been lacking to date has been a more pedestrian study of the precise nature and quality of the data returned. Are they “tidy,” or “clean,” providing pertinent and well organized materials, or fraught with messy confusions, obfuscations, or redundancies (Wickham 2014)? If the latter, how might the dataset best be cleaned?

One approach to a better understanding of these evolving, dynamic data catches would be to compare the relative merits of several distinct but competing systems. Another, chosen here, is to carry out a line by line analysis of data collected for a single disease event by a single automatic collection system. For this study data describing the first month of the 2009 H1N1 Type-A Influenza pandemic was examined. As a preliminary study this limited study will hopefully provide a basis on which other systems might be compared in the future.

This database has been used by researchers studying the 2009 pandemic whose epidemiology (its Ro rate, for example, and mortality) are now well known. Data describing the 2020 COVID-19 experience, while voluminous, is at this writing constantly evolving. Not only do incidence and mortality figures change daily but methods of diagnosis and testing vary widely across national and international regions. Complicating matters further, different testing strategies are employed by different agencies (antibody testing or nasal swabs) at different rates in different jurisdictions. It is therefore not surprising that the retraction rate for COVID-19 journal articles has been "alarming." (Ling Yeo-Teh and Luen Tang 2020). The earlier dataset, by contrast, is firmly established and the characteristics of that virus, and its resulting pandemic, well understood. This assured a stable epidemiological platform for the study.

Spatially located disease data

There is nothing particularly innovative in the use of spatially located disease data in the mapping of epidemic or endemic disease (Koch, 2011; 2017). In an 1831 unnamed authors in *The Lancet* collected data primarily from official, foreign office reports and some news studies to produce a dot map of 1200 different locations to describe the global spread of cholera from 1918-1931. In 1832 Brigham mapped the temporal progress of cholera from India to Europe and then North America across then popular global trade and travel routes (Brigham, 1832). In the 1850's John Snow famously argued the waterborne nature of cholera by mapping first a local outbreak in Soho and then, more ambitiously, the relation between water quality and the incidence of cholera in a South London epidemic (Snow, 1855; 1856). The data for both studies was primarily provided by William Farr at the General Register Office in London.

Today's spatial analytics is distinguished by the broad, public nature of the data and its method of collection. Contemporary syndromic capture systems permit the collection and almost simultaneous distribution of volumes of official and public materials as they become available. Their programs continually scan a range of potential formal (local health agencies, the CDC, PAHO, WHO) and "informal" data sources (Google News for example) for disease-related reports Relevant entries are identified by "tags," keywords loaded with each datum to identify a disease (flu, influenza, etc.), reported at a location on a given date. Depending on the algorithm they may also incorporate automatically other data necessary for analysis—population statistics, for example.

Data and Methodology

History

As early as 1948 the World Health Organization (WHO) created a network of international influenza reportage based on reports submitted by member nations. In the 1990s, the 53 nations of WHO's European region began not only sharing but then aggregating that data based on yearly incidence (Fleming, Van der Velder, Paget, 2003). That regional surveillance program became part of a Global Influenza Surveillance and Response System issuing weekly reports in participating countries. It, in turn, was the forerunner of FluNet, a contemporary, web-based online tool for global influenza surveillance (WHO. 2019).

Following the 1995 Ebola epidemic in the Democratic Republic of Congo, WHO began developing a more general, Globe Outbreak Alert and Response network (GOAR) using "systems of electronic communications supported by 151 country offices concentrated in the developing world and the participation of more than 110 existing institutes, laboratories, agencies, and surveillance systems" (Heymann, Guenael, 2004). In 2003 surveillance of SARS (Severe Acute Respiratory Syndrome) incidence demonstrated "the advantages of rapid electronic communication and new information technologies for emergency response, and the willingness of the international community to form a united front against a common threat" (Heyman and Rodier, 2004; 186).

The result has been an explosion of dedicated digital disease collection libraries. Some, like the Johns Hopkins University Coronavirus dashboard, are dedicated to a single disease event. Others lodge reports of a range of disease events within broad systems of digital data recovery and storage. The Global Database of Events, Language, and Tone (GDELT) broadly monitors globally a wide range of subjects, including disease incidence, published in international broadcast and news sites (Leetaur and Schrodt, 2013). The growth of these systems has spawned a mini-literature on what some have called “biosurveillance” (O’Shea, 2017) which, in epidemiology and public health, involves the rapid identification and study of epidemic or pandemic events (Yan, Chughtai, and MacIntyre, 2017; Lee, Asher, Goldlust et al. 2001).

Healthmap.org

Begun in 2006 at Harvard University and maintained today at Boston's Children's Hospital (Mass.), *healthmap.org* “through an automated process, updating 24/7/365, the system monitors, organizes, integrates, filters, visualizes and disseminates online information about emerging diseases in nine languages” (Frelfeld and Brownstein, 2007). Data collected includes reports by online news aggregators, eyewitness reports, expert-curated discussions and official reports published in varying media. It thus presents a “multistream real-time surveillance platform that continually aggregates reports on new and ongoing infectious disease outbreaks” (Brown, Freifeld, Reis, and Mand, 2008).

The system performs categorization, extraction, filtration, and integration of relevant data through the application of a complex set of algorithms. For each report captured the computer program employs a Parser Module using a word-level, N-gram to identify and match "disease tokens," (names: Ebola-19; SARS CoV-2, etc.) against entries in a continually expanding dictionary of pathogens. With each new entry the program simultaneously extracts locations named in the report (Boston, MA) and then matches them to locations in a dictionary of known places identified by latitude and longitude (Freifeld, Mandl, Reis and Brownstein, 2008). In theory, at least, the algorithms employed can recognize and correct for multiple definitions of a single disease and distinguish between places with similar names located in different geographies.

Because it has been used by other researchers studying the 2009 influenza pandemic (Balcan, Colizzac, Gonçalvesa et al., 2009), a dataset covering that epidemic was requested. Healthmap.org officials kindly responded with a multi-year (2008-2012) dataset of global influenza reports containing 68,720 entries for the United States of which 23,807 entries (34.64 percent) referred to the 2009-2010 epidemic in the United States. Each entry included a location (city, county, state), urban population size, diagnosis (H1N1 or other influenza), data source (typically a URL), report date, and spatial location (latitude and longitude). Separate columns in the dataset included the number of suspected and separately the number of confirmed cases reported in each entry.

Methods

The Healthmap.org dataset was opened in ESRI's ArcGIS 10.6.1. This permitted not only its mapping by latitude and longitude data included for each entry but also presented a flexible format for analysis and review. To facilitate this review, only those cases occurring during the first wave of the pandemic in the United States between May1 and August 31, 2009. The resulting 7,660 entries (32.17 percent of the total) included summary monthly reports for individual states. While potentially useful in studying broad patterns of diffusion, none identified specific incidence reports. These therefore were excluded as were stories which discussed the pandemic but DID NOT report confirmed cases. A further 22 entries reporting a total of 140 confirmed cases were clearly duplicates reporting the same number of cases at the same location on the same day. Most were news stories reporting official incidence counts at a single location but distributed by more than one news organization. Another 19 entries reporting 1,802 confirmed cases were flagged as probable duplicates reporting the same number of cases for the same location but with a date tag a day later. While these were likely duplicates this could not be confirmed and so they were retained in the database. The remaining 1,275 entries reporting confirmed cases of 2009 H1N1 Type-A influenza is mapped in **Figure 1**. This sample—16.64 percent of the total of entries collected for this period was sufficiently small to permit a line-by-line review.

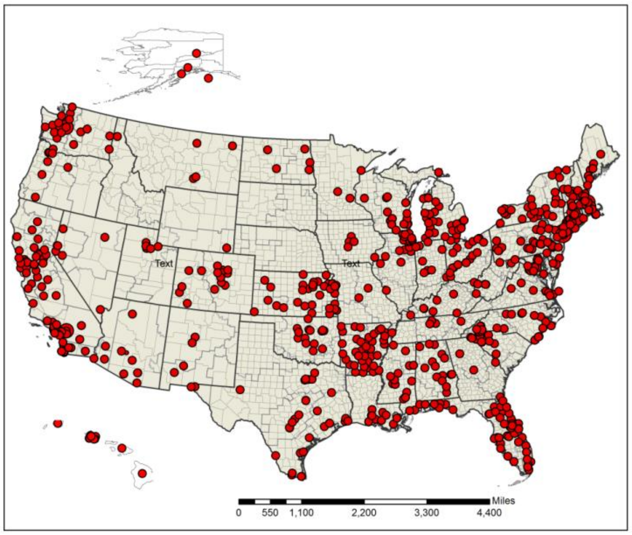

Figure 1: The location of all 1,275 entries H1N1 Influenza cases reported in the continental U.S. States from May 1-Aug. 31, 2009, during the first phase of the epidemic. Data derived from healthmap.org.

Results

Confusions

In the edited dataset of 1,275 entries, the "Location" column header included place names (for example, "Buffalo, NY, USA"), counties ("Erie County, NY, USA") and in some cases combined city-county designations ("Buffalo-Erie County, NY, USA"). In addition there were incidence reports at jails or prisons, schools, summer camps, and other institutions, all of which existed within one or another of those jurisdictions. To take one example, there were 9 separate entries totally 92 cases between May 19 and June 6 for "Rikers Island Prison**"** in New York City's borough of Queens. Rikers Island houses an average population of 10,000 inmates supervised by 9000 correction officers with a support staff of approximately 1500 persons (Goldstein 2015). That incidence data was presumably reflected in both borough and city-wide entries in the system.

Another problem was the number of multiple entries, with slightly different geographic assignments for counties, cities within counties, and cities as independent reportage sites. Each of these entries was mapped in the unchallenged database as spatially distinct. Examples of different descriptors for a single place included, for example:

- San Francisco, California.
  Lat. -122.418404; Long. 37.775002.

- San Francisco, Alameda County, California.
  Lat. -121.884399; Long. 37.599934.

- San Francisco County, California.
- Lat. -122.45108; Long. 37.766598

- San Francisco Police Department, California.
- Lat. -122.404137; Long.

Latitude and longitude locators made all these entries proximate but distinct within the greater San Francisco Bay urban area. To add to the confusion, the city and county of San Francisco are typically considered as a single entity with a 2010 population of 805,184. Alameda County, whose county seat is Oakland (2010 population was 1,510,271 persons), is in the San Francisco Bay area and adjacent to the city-county of San Francisco (U.S. Census, 2010). There were no parallel reports for this time frame from either the city of Oakland or Alameda County, however.

4.3 Location categories

The result was a confusing mixture of jurisdictional scales (one for a city and another for its encompassing county) and populations (city or region) combined with institutional reports of cases occurring within one or another reporting jurisdiction.

To rectify the problem a new column, "place-type," first was created in the ArcGIS database and then filled by the first location noun phrase in the official database. Thus "Rikers island jail, New York City, NY" became, in this new column, simply "Rikers Island jail." Other columns were created to distinguish city (Buffalo, NY), city-county (Buffalo, Erie County), and county-parish (Louisiana) designations. US territories (Puerto Rico, U.S. Virgin Islands) were separately identified. A limited set of eccentric and highly specific locations (for instance, a local church, the Kennedy Space Centre) were summarized as "miscellaneous." The San Francisco Police Department report was labeled as "miscellaneous" while both San Francisco and San Francisco County were separately described as 'city' and as 'county' data entries.

This permitted apparently distinct but spatially proximate reports to be compared. In some locations, county level data dominated while, in other involving typically larger cities, city-designated reportage outweighed incidence reportage by the county jurisdiction (for example, Chicago, IL, vs. Cook County, IL). Table 1 presents the general results of the review at this level.

Table 1: A breakdown of reports by type.

Locations	Number of	percentage	Number of	percentage	incidence	mean #
	Reports	of all reports	cases	of all cases	range	of cases
Native reserves	1	0.001	8	0.00025	8-8	8.000
police dept./academy	2	0.002	13	0.00040	1-12	6.500
Virg. Islands	8	0.007	107	0.003	1-107	13.375
Puerto Rico	12	0.010	395	0.012	1-138	32.920
jail-prison	21	0.017	207	0.006	1-47	9.860
camps (summer)	22	0.001	220	0.007	1-94	10.000
military	16	0.018	262	0.008	1-68	16.375
college-univ.	45	0.037	427	0.013	1-63	9.489
school	88	0.072	525	0.016	1-69	5.960
city-county	39	0.032	136	0.004	1-172	18.870
city	203	0.165	14,552	0.448	1-1557	71.685
county-Parish	772	0.628	15,609	0.481	1-963	20.219

Table 1: Principal categories of location designations for H1N1 Influenza reported in the U.S. between May 1 and Aug. 30, 2009. Data extracted from healthmap.org database.

4.1 Results

Because disease incidence is typically modeled, and reported, with reference to a specific urban place (Boston, New York City, Los Angeles, etc.) the assumption was that most reports would reflect that preference. And yet, only 16.5 percent of the incidence reports (45 percent of all confirmed cases) reviewed was reported solely by city name. Of the total entries examined, 62.8 percent of all reports, reflecting 48.1 percent of all cases, were at the county level. Each county was located spatially at its centroid.

At issue was more than a tidy map. Population size and distance between cities are critical components of a range of disease models including those employing distance decay or gravity formulae. Boston, MA, reported 955 cases of H1N1 Influenza between May 2 and the end of August 2009. The vast majority of these were presented in two reports dated respectively June 29 (475 cases) and July 2 (474 cases). Boston is the principal city in Suffolk County, MA, however, which reported a total of 620 cases between June 2 and August 24 with a far more even distribution across the study's time period. Boston's 2010 population in 2010 was reported as 617, 594 (US Census Factfinder, 2012) while that of Suffolk County, an administrative state subdivision including Boston and proximate towns, was 732,864 persons. The difference in reportage dates and total number of confirmed cases within different official populations would affect any attempt at modeling the epidemic's engagement in that area.

In a similar vein, 277 cases were reported among the 9,818,605 persons counted in the 2010 Census for Los Angeles County. By comparison, 107 cases were reported during the study period for the Los Angeles City population of 3,792,621 persons. Of those totals, 102 were in news reports for the city and county respectively.

While not considered in this study, reports of confirmed cases in specific locations provided potentially important insights into the dynamics of epidemic expansion. This became evident during the COVID-19 pandemic in which outbreaks occurred in specific situations reflecting high densities of closely congregated communities: assisted living and nursing homes; jails and prisons; religious institutions, etc. In late July and August, 2020, outbreaks forced the closure of U.S. summer camps in Arkansas and Missouri (Lee 2020). Report of camp outbreaks in the 2009 epidemic made this predictable. Knowing where an infectious disease is likely to expand, once introduced, and then threaten broader community engagement will be a critical component of future modeling of local disease dynamics.

5. Discussion

Healthmap.org was used, here, as a convenient and publicly accessible example of an evolving class of spatially grounded, automatically collected digital data systems. Some of the problems identified may be unique to it. But others likely are common to all systems of syndromic surveillance involving public and public resources. The results offer lessons for disease researchers as well as for those who maintain these systems as they evolve.

First, researchers modeling disease events based on these systems rarely describe the means which data has been reviewed and cleaned. They thus presumably accept the database without careful review. This study suggests studies based on syndromic data require before analysis a careful review of the nature of the data received and the locational categories it includes.

The prominence of county-level reportage was a surprise. In the future researchers using systems like healthmap.org, or similar systems, might be advised to work at that resolution despite a desire for the greatest specificity of city-name designations. Whatever their choice, consistency of resolution and scale and a transparency in their selection for any study must be a priority.

Automated data collection systems remain an evolving work in progress. Those designing the algorithms driving syndromic systems may wish to include directions that would better distinguish between different jurisdictions and those reports focused on individual place categories (jails, schools, etc.). Not only would this help separate possibly redundant reports bout would provide an easy mechanism by which investigators could choose data appropriate to the nature of the modeling they wish to pursue at different scales of address.

Acknowledgements**:** The author wishes to acknowledge the assistance of researchers at healthmap.org who provided a complete database of 2009 H1N1 Type-A influenza reports culled from their evolving automatic data retrieval program.

References

1. Khoury MJ, Cordero JF, Greenberg F, James LM, Erickson JD. A population study of the VACTERL association: evidence for its etiologic heterogeneity. Pediatrics 1983; 71:815-20.
2. ArcGIS.com. Coronavirus COVID-19 Cases. Esri.com, 2020. https://www.arcgis.com/home/item.html?id=bbb2e4f589ba40d692fab712ae37b9ac# (Accessed July 5, 2020).
3. Balcan D., Colizzac V, Gonçalvesa B, Hu H, Ramascob J, Vespignani A. Multiscale mobility networks and the spatial spreading of infectious diseases. *PNAS* 2009. 106 (51): 21484–21489. http://www.pnas.org/content/106/51/21484.full.pdf. Accessed May 15, 2018.
4. Brigham H. (1832). *A Treatise on Epidemic Cholera: Including an Historical Account of Its Origin and Press, to the Present Period*. Hartford, CT: H. and F. J. Huntington. https://archive.org/details/treatiseonepidem00brig/page/n12 .
5. Brown JS, Freifeld CC, Reis BY, and MAND KD. Surveillance *Sans Frontières*: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Medicine 2008; 5 (7): 1019-1024. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050151 .
6. Chiara GC, Raffle J, Aisyah DN, Sartain F, Kozlakidis Z. Big Data Analytics, Infectious Diseases and Associated Ethical Impacts. *Philos & Technol* 2019; 32 (1): 69-85. https://doi.org/10.1007/s13347-017-0278-y.
7. Feldman J, Thomas-Bachli A, Forsyth J, Hasnain Z, et al. Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise. *Journal of the American Medical Informatics Association* 2019; 36 (11), 1355–1359. doi: 10.1093/jamia/ocz112.
8. Fleming DM, Van der Velden J, Paget WJ. The evolution of influenza surveillance in Europe and prospects for the next ten years. *Vaccine* 2003; 21 (16): 1749-1753. https://doi.org/10.1016/S0264-410X(03)00066-5 PMID: 12686088 .
9. Frelfeld C, Brownstein J. *About Healthmap. Boston: Boston Children's Hospital, 2007.* https://healthmap.org/about/ .
10. Lancet. History of the rise, progress, ravages, etc. of the blue cholera of India. *Lancet* 1831; 17; 429: 241-284,
11. Lazaro G.L, Yourish K. 2020. See how the Coronavirus Death Toll Grew Across the U.S. *New York Times* (April 7), 2020. https://www.nytimes.com/interactive/2020/04/06/us/coronavirus-deaths-united-states.html (Accessed July 5, 2020).
12. Gilbert G L, Degeling C, and Johnson J. Communicable Disease Surveillance Ethics in the Age of Big Data and New Technology. *Asian Bioethics Review* 2019; 11: 173–187 https://doi.org/10.1007/s41649-019-00087-1
13. Heymann DL, Guenael RG. *The Brown Journal of World Affairs* 2004; 10 (2): 185-197. https://www.jstor.org/stable/24590530.
14. Johns Hopkins University. School of Medicine Coronavirus Centre. Baltimore, MD, 2020. https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 .
15. Koch T. *Disease Maps: Epidemics on the Ground*. Chicago, IL. University of Chicago Press, 2011.
16. Koch, T. *Cartographies of Disease: Maps, Mapping, and Medicine*. Redlands, CA: Esri Press, 2017: Chapter 14.

17. Kraemer M, Hay SI, Pigott DM, Smith DL, et al. Progress and Challenges in Infectious Disease Cartography. *Trends in Parasitology* 2016; 32(1): 19-29. https://www.sciencedirect.com/science/article/abs/pii/S147149221500207X .
18. Lee A. Summer camps close after Covid-19 outbreaks among campers and staff. CNN News, 2020 (July 8). https://www.cnn.com/2020/07/08/us/missouri-arkansas-summer-camp-covid-19-trnd/index.html
19. Lee EC, Asher JM, Goldlust S, Kraemer JD, et al. Mind the scales: harnessing spatial big data for infectious disease surveillance and inference. *J Infect Disease 2016; 214* (S4): S409–S413. https://arxiv.org/pdf/1605.08740.pdf.
20. Leetaru K, Schrodt P A. GDELT: Global Data on Events, Location and Tone 1979–2012. Paper presented at *The International Studies Association* meetings, San Francisco, 2013. http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf.
21. Ling Yeo-Teh N, Tang B. L. An alarming retraction rate for scientific publications on Coronavirus Disease 2019 (COVID-19). *Accountability in Research Policies and Quality Assurance*, 2020. DOI: 10.1080/08989621.2020.1782203.
22. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics* 2018; 114: 57-65. https://www.sciencedirect.com/science/article/abs/pii/S1386505618302466 .
23. Milwaukee County. 2020. Milwaukee County Covid-19 Dashboard. https://www.arcgis.com/apps/opsdashboard/index.html#/018eedbe075046779b8062b5fe1055bf (Accessed July 5, 2020).
24. O'Shea J.. Digital Disease Detection: A Systematic REview of Event-based Internet Biosurveillance Systems. *Int J. Med Informatics* 2017; 101: 14-22. Doi:10.1016/j.ijmedinf.2017.01.019.
25. Polonsky JA, Baidjoe A., Kamvar ZN, Cori A., et al. Outbreak analytics: a developing data science for informing the response to emerging pathogens. *Phil. Trans. R. Soc. B* 2019; 374: 20180276: 1-11. http://dx.doi.org/10.1098/rstb.2018.0276.
26. Snow J. On *the Mode of Communication of Cholera, Second Edition*. London: Churchill, 1855.
27. Snow J. Cholera and the Water Supply of the South Districts of London in 1854. *Journal of Public Health* 1856; 2: 239-257.
28. U.S. Census. Quick Facts: San Francisco County. Population, 2010. https://www.census.gov/quickfacts/fact/table/sanfranciscocountycalifornia,CA,US/PST045218
29. U.S. Census Annual Estimates of the Resident Population for Incorporated Places Over 50,000, Ranked by July 1, 2012 Population: April 1, 2010 to July 1, 2012 - United States -- Places of 50,000+ Population 2012 Population Estimates., 012. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk.
30. Wickham H. Tidy Data. *Journal of Statistical Software* . 2014; 59 (10): 1-22. https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf.
31. WHO. 2019. *Influenza: Flunet*. Geneva: World Health Organization. https://www.who.int/influenza/gisrs_laboratory/flunet/en/ .
32. Yan SJ, Chughtai AA, Macintyre, CR. 2017. Utility and potential of rapid epidemic intelligence from internet-based sources. *Int J Infect Dis* 2017; 63: 77–87.