A Big Data Approach to Surveillance of Disease Outbreaks

Meghana Aruru* and Saumyadipta Pyne

Abstract— Disease surveillance data collection over geographical space and time is historically a big data activity. Storage and near real-time analysis of public health data, and metadata, of large volume, velocity and variety, are increasingly supported by emerging computational platforms. Guiding principles for standardization of such data are being drafted and implemented by new international initiatives thereby ushering a new global culture of big data in health.

Index Terms—Big Data, Disease Outbreak, Epidemics and pandemics, Health Analytics

1 Introduction

In a world increasingly connected by travel and trade, the risk of emerging epidemics is rising at the rate of 1 new disease per year [1]. In the last 5 years, World Health Organization (WHO) has identified more than 1,100 epidemics including viral diseases such as polio, HIV, Marburg virus, Nipah virus, Ebola and avian flu [2].

In a globalized world, new and re-emerging pathogens can spread rapidly and infect large populations in many countries. Historic pandemics such as the "black death" and the "Spanish flu" are well studied. The 2009 H1N1 Influenza pandemic was first detected in USA but it rapidly spread around the world. The U.S. Centers for Disease Control and Prevention (CDC) estimated in a modeling study that between 151,700 and 575,000 people died from the 2009 H1N1 viral infection worldwide [3]. More recently, a devastating Ebola outbreak in several West African countries led to more than 28,000 cases and more than 11,000 deaths between 2014 and 2016.

In order to be prepared to tackle such sudden and severe stresses to their health (and other) systems, many, if not most, countries rely on disease surveillance mechanisms that are based on systematic generation and analysis of disease outbreak data. Indeed, public health decision-making and action depend critically on the availability of such data.

It is important to systematically monitor, report and respond to emerging crises in to reduce the burden of epidemics. Naturally, such data has significant volume, velocity and variety.

2 Data collection, analysis and sharing

Information is disseminated quickly through public health networks initially, and later through peer-reviewed journals and accompanying datasets. In unfolding emergencies, such timely available and readily usable information is critical for deciding the appropriate course(s) of action. Data sharing enables researchers to model critical paths and interventions towards preparedness for future events.

To begin with, acquisition of surveillance data requires establishment of a public health network of various systems including specialized clinical microbiology and pathology laboratories, emergency preparedness and response centers, and hospital reporting systems, among others. Coordination is essential between automated, semi-automated and manual data gathering, and, its subsequent dissemination for public health action.

In developed countries like the United States, such sentinel surveillance systems are put in place to identify emerging threats and monitor existing threats through a wide network of public health laboratories, hospitals, epidemiological survey units, etc. A sentinel surveillance system uses high quality data about diseases that cannot be monitored through passive surveillance systems.

Data collected through sentinel surveillance systems can be used to identify trends, detect outbreaks and monitor disease burden in communities, thus providing timely information to policy makers and public health planners [4].

Naturally, data quality and standardization are important aspects that determine the utility of collected data. In an unfolding emergency, there is a critical time window during which data must be quickly gathered, analyzed and

- *Dr. Meghana Aruru is Vice-President of Pramana Analytics. She is also an Adjunct faculty member at the MediCiti Institute of Medical Sciences. Her work focuses on health policy and communications.*
- *Prof. Saumyadipta Pyne is the Scientific Director of the Public Health Dynamics Laboratory, University of Pittsburgh, USA. He holds Adjunct Professorship at the National Institute of Medical Statistics of Indian Council of Medical Research (ICMR).*

**Correspondence: meghana@healthanalytics.net*

disseminated for rapid response. In the absence of standardization, data gathered may be unreliable or unfit for immediate analysis. One of the first aims in developing surveillance systems is to generate data that is deemed to be ready-to-use. Many agencies such as the U.S. National Notifiable Disease Surveillance System (NNDS) have detailed methods for data collection, analysis, standardization and sharing to achieve usable data [5].

3 Data Stewardship

Data stewardship involves good data management practices beyond proper collection, annotation and archival. Data quality maintained in the 'long run' leads to high quality research and publications. Several guidelines exist on preserving and maintaining data.

One such set of guidelines, established by the FAIR Data Initiative, an international consortium, is based on the following four foundational principles about a given dataset –

1. **F**indability,
2. **A**ccessibility,
3. **I**nteroperability, and
4. **R**eusability.

Together, the FAIR Data Principles serve to optimize outcomes associated with such data use [6]. (See box)

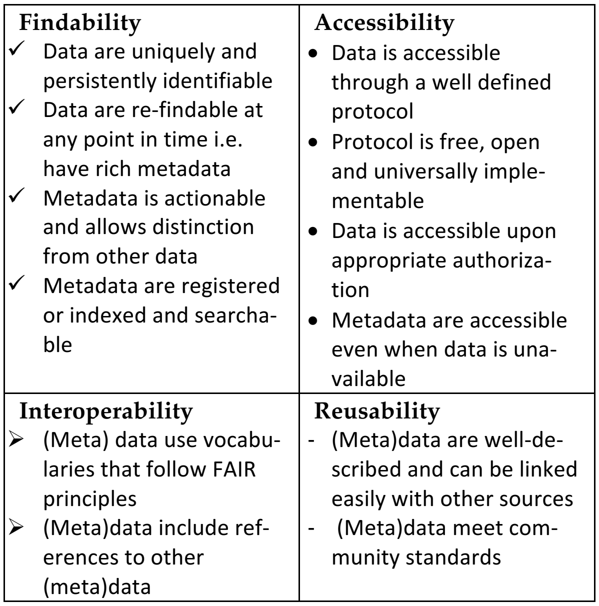
Data availability is simply not enough. Available data should be properly linked with well-established protocols on how to use them. Patterns may emerge from functionally linked datasets, calling for the subsequent steps to rationalize and conduct confirmation studies. It is therefore critical that relevant and useful metadata be systematically included and saved with all datasets for researchers to track provenance and justify the evidence from the uncovered patterns.

The FAIR guidelines promote data sharing and accessibility for researchers and scientists around the world in the interest of furthering science. Data sharing for public health response is an emerging cultural phenomenon and there are many projects that aim to make data - that are standardized - available for use by public health scientists. Many such data repositories have adopted FAIR standards, e.g., Mendeley data, Dataverse, Figshare.

In public health, one such effort is Project Tycho® at the University of Pittsburgh in USA. It "aims to advance the availability and use of public health data for science and policy" [7]. Researchers have digitized all available city and state notifiable disease data from as early as 1888 until as recent as 2011 (in the first version of the database), obtained mostly from hard copy sources. Information collected corresponding to nearly 88 million cases was stored in a database that is open to interested parties without any restriction through its online archive (http://www.tycho.pitt.edu).

This database is arguably among the earliest examples of systematic public health data integration where millions of cases were digitized from hard copies and integrated with existing data in a standardized, usable format for researchers and public health scientists to access free of cost and add to the globally growing body of knowledge on outbreaks and epidemics.

To further this contribution, researchers at Project Tycho® have collaborated with different countries in Southeast Asia to gather data on dengue surveillance and detect disease patterns at regional levels. This resource integrates data from the WHO DengueNet provided by WHO surveillance networks for its member countries.



Findability	**Accessibility**
✓ Data are uniquely and persistently identifiable ✓ Data are re-findable at any point in time i.e. have rich metadata ✓ Metadata is actionable and allows distinction from other data ✓ Metadata are registered or indexed and searchable	• Data is accessible through a well defined protocol • Protocol is free, open and universally implementable • Data is accessible upon appropriate authorization • Metadata are accessible even when data is unavailable
Interoperability	**Reusability**
➢ (Meta) data use vocabularies that follow FAIR principles ➢ (Meta)data include references to other (meta)data	- (Meta)data are well-described and can be linked easily with other sources - (Meta)data meet community standards

4 Conclusion

A Big Data approach to epidemiology would benefit from data stewardship principles to ensure transparency, reproducibility, and reusability of high volume and high velocity data. Multiple stakeholders can benefit from such data stewardship including researchers who are willing to share their data and software, organizations, funding agencies, etc., and, indeed, a rich data science community that is interested in mining both new and existing data.

Applying standardization techniques such as the FAIR guidelines aids in integration of massive data repertories in a systematic and automated manner to save time and aid pattern discovery. Big data surveillance repositories like Mendeley data, Dataverse, Figshare, Project Tycho® etc. illustrate the value of data collection, standardization and sharing to drive research progress on major public health challenges. Evidence gathered can lead to exploration of confirmatory studies as well as framing and evaluation of policies for appropriate and timely public health responses.

Importantly, it seems more certain than ever that a new and deeper Big Data "culture" - that goes well beyond the traditional tasks of data collection and analysis - involving also data curation, harmonization, standardization, annotation, and sharing of data freely - is here to stay.

References

[1] WHO | Disease outbreaks. WHO. http://www.who.int/emergencies/diseases/en/. Published 2017.

[2] Modjarrad K, Moorthy VS, Millett P, Gsell P-S, Roth C, Kieny M-P. Developing Global Norms for Sharing Data and Results during Public Health Emergencies. PLOS Med. 2016;13(1):e1001935. doi:10.1371/journal.pmed.1001935.

[3] First Global Estimates of 2009 H1N1 Pandemic Mortality Released by CDC-Led Collaboration | Spotlights (Flu) | CDC. https://www.cdc.gov/flu/spotlights/pandemic-global-estimates.htm.

[4] WHO | Sentinel Surveillance. WHO. 2014. http://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/sentinel/en/. Accessed January 3, 2018.

[5] NNDSS | Centers for Disease Control and Prevention. https://wwwn.cdc.gov/nndss/. Accessed January 3, 2018.

[6] Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 3. March 2016.

[7] van Panhuis WG, Grefenstette J, Jung SY, et al. Contagious Diseases in the United States from 1888 to the Present. N Engl J Med. 2013;369(22):2152-2158. doi:10.1056/NEJMms1215400.