https://doi.org/10.14745/ccdr.v41i09a02

Overview

Big Data and the Global Public Health Intelligence Network (GPHIN)

Dion M[1], AbdelMalik P[2], Mawudeku A[2*]

[1]Schulich School of Family Medicine and Dentistry, University of Western Ontario, London, ON
[2]Centre for Emergency Preparedness and Response, Public Health Agency of Canada, Ottawa, ON

[*]Correspondence: Abla.Mawudeku@phac-aspc.gc.ca

Abstract

Background: Globalization and the potential for rapid spread of emerging infectious diseases have heightened the need for ongoing surveillance and early detection. The Global Public Health Intelligence Network (GPHIN) was established to increase situational awareness and capacity for the early detection of emerging public health events.

Objective: To describe how the GPHIN has used Big Data as an effective early detection technique for infectious disease outbreaks worldwide and to identify potential future directions for the GPHIN.

Findings: Every day the GPHIN analyzes over more than 20,000 online news reports (over 30,000 sources) in nine languages worldwide. A web-based program aggregates data based on an algorithm that provides potential signals of emerging public health events which are then reviewed by a multilingual, multidisciplinary team. An alert is sent out if a potential risk is identified. This process proved useful during the Severe Acute Respiratory Syndrome (SARS) outbreak and was adopted shortly after by a number of countries to meet new International Health Regulations that require each country to have the capacity for early detection and reporting. The GPHIN identified the early SARS outbreak in China, was credited with the first alert on MERS-CoV and has played a significant role in the monitoring of the Ebola outbreak in West Africa. Future developments are being considered to advance the GPHIN's capacity in light of other Big Data sources such as social media and its analytical capacity in terms of algorithm development.

Conclusion: The GPHIN's early adoption of Big Data has increased global capacity to detect international infectious disease outbreaks and other public health events. Integration of additional Big Data sources and advances in analytical capacity could further strengthen the GPHIN's capability for timely detection and early warning.

Introduction

As globalization increases, so does the rapid spread of communicable diseases and emerging public health events. As a result, ongoing surveillance and early detection are even more important to prevent or mitigate the international spread of infectious diseases and to provide countries adequate time to prepare and respond. Big Data refers to the extremely large datasets provided by sources such as social media or newspapers which require powerful computational methods to reveal trends, patterns or the predictive likelihood of an event (1,2). Big Data has been used to optimize sales and business processes, inform trades among sports teams and to improve city planning. It is quickly becoming integral to a variety of aspects of health ranging from health care administration to Google Flu and pharmacosurveillance (3).

Canada was an early adopter of Big Data for the initial identification of emerging infections beginning in 1997 through the development of the Global Public Health Intelligence Network (GPHIN), a cooperative effort between (at the time) Health Canada and the World Health Organization (WHO) (4,5). The GPHIN continues to be maintained by the Public Health Agency of Canada (the Agency) and links a global network of public health professionals and organizations (e.g., Ministries of Health) for situational awareness and early

detection of emerging public health events. The GPHIN relies on an automated web-based system that scans newspapers and other communications worldwide for potential indicators of outbreaks (or “signals”) that are analyzed and rapidly assessed by a multilingual, multidisciplinary team at the Agency. When a risk is identified, analysts disseminate relevant information and alerts to senior officials and stakeholders for decision-making. While initially devised to identify communicable disease outbreaks, the system has also been used to monitor potential chemical and radio nuclear hazards (4,6).

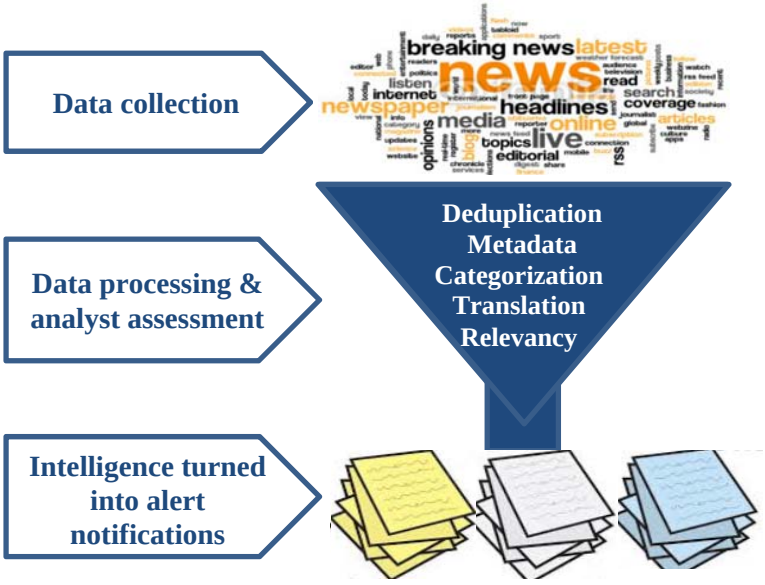
The objective of this article is to identify how the GPHIN functions within the context of Big Data, to provide recent examples of the GPHIN in action and to explore potential future directions.

The GPHIN and Big Data

Big Data has been defined by three V’s: volume, velocity and variety (7,8). Volume describes the quantity of data that is collected, velocity is the speed at which the data is collected and disseminated and variety refers to the multiplicity of sources that are used to compile the data (7).

The GPHIN’s volume and variety are exemplified through the use of search functions and news aggregators (companies that provide access to thousands of news sources whose content is automatically indexed) that gather large quantities of data sets from multiple different sources. A web-based application in the GPHIN system continuously scans and mines acquired news sources worldwide in nine languages (Arabic, English, Farsi, French, Portuguese, Russian, simplified Chinese, Spanish and traditional Chinese) (4). The quantity of data generated is dependent on the criteria, variables and algorithms outlined for the aggregators (6). These algorithms identify potential signals of emerging public health events and filter out irrelevant data considered as “noise” (**Figure 1**) (7). Every day, on average, the GPHIN processes 3,000 news reports (9). Volume increases when news sources expand coverage on emerging public health events such as the recent Ebola outbreak in West Africa.

Figure 1: The flow of information for the GPHIN process



The GPHIN has an abundant variety of data sources. The GPHIN’s news aggregators rely on a large variety of national and local newspapers and select newsletters from around the world (4,6). Local newspapers and newsletters are scanned because emerging events can be a localized phenomenon and are often reported in community newspapers and newsletters. Various sections of news publications (sports, travel and finance) are also monitored as they may signal an emerging public health event. Scanning across various languages is done in order capture public health events that are not reported in English news (10).

After further application of algorithms within the GPHIN system, approximately 60% (1,800 news reports) of the data are deemed as relevant public health events for assessment. GPHIN analysts sift through these news reports to identify and provide alerts about events with potential implications for decision-making by stakeholders. Access to the GPHIN system is provided to entities that have the responsibility to monitor, respond to and or mitigate emerging public health threats. The GPHIN includes ministries of health, other governmental departments and agencies, international and non-governmental organizations and private companies.

The capacity for velocity in the GPHIN is impressive. It operates on a near real-time, 24/7 basis (4).The GPHIN system retrieves relevant data from the news aggregators every 15 minutes and is able to complete the processing (including translation) of the data in less than one minute (9).

The GPHIN in action

Early detection

The GPHIN has proven to be an effective early detection resource for infectious disease outbreaks. Its utility was initially demonstrated during the Severe Acute Respiratory Syndrome (SARS) outbreak in 2003 when early alerts were provided in reports from Chinese newspapers. The first English report about an atypical outbreak in China was noted by a pharmaceutical company in the financial section of a newspaper that had reported increased sales of its antiviral drugs (11). This not only flagged the emergence of the outbreak but provided additional information about the local use of antiviral drugs to contain the spread of the virus.

Following the SARS outbreak, the significance of using news media to complement more traditional national public health surveillance systems was recognized by the WHO and its member states (12,13). The SARS outbreak lead to revisions of the International Health Regulations (IHRs) (14) that required countries to report and control outbreaks of potential international concern in order to strengthen global public health security. The IHRs note that the WHO may include reports from sources other than official notifications or consultations in their assessment of a potential emerging public health event (14). After the SARS outbreak, the GPHIN outputs have been used by multiple countries to expand their surveillance capacity (4,15).

Over the years, the GPHIN has continued to detect early signals of outbreaks of international concern such as the pandemic influenza H1N1 in 2009 (16). Initial Spanish language reports about the outbreak noted an unusual respiratory outbreak in the state of Veracruz, Mexico that had claimed two lives.

In April 2012, the GPHIN identified eight cases of an unknown respiratory illness and one death in Jordan. GPHIN issued an alert notifying stakeholders, including the WHO, about these cases. Following further investigation and the results of a retrospective laboratory analysis, an outbreak of Middle East Respiratory Syndrome Coronavirus (now known as MERS-CoV) was confirmed. An International Health Regulations (IHR) Notification was posted in November 2012. The GPHIN was credited with being the first to issue an alert about this new emerging illness.

Ongoing monitoring

The GPHIN has proven to be useful for both early detection and continuous monitoring. Ongoing monitoring of events is critical for situational awareness regarding the evolution of an outbreak and the response and mitigation strategies being implemented by the local, national and international communities. Examples of situational awareness of mitigation strategies include the GPHIN's ability to scan for cancellation of flights or cruises, new travel advisories, health screening procedures at border crossings or trade bans. This process has been much more efficient than individually contacting commercial transportation companies, travel agencies and airports.

For example, during pandemic Influenza H1N1, the GPHIN was used as an intelligence source by the World Trade Organization to monitor the extent and the effect of trade bans (17). Similarly, during the recent response to the Ebola outbreak in West Africa, the GPHIN provided situational awareness about the cancellation of flights, travel advisories and health screening procedures at border crossings.

Next steps

Potential new data sources

Internet, email, smart phones and social media have developed rapidly since the GPHIN was first developed. As a result, potential new sources of Big Data have emerged that can be analyzed to detect signals of early infectious disease outbreaks. Social media tools (such as Twitter and Facebook) have witnessed exponential growth over the last 10 years and these platforms create huge amounts of user-generated content and data (18).

These various social media represent potential new data sources for the GPHIN. In addition, other organizations have started to mine social media resources to improve disease surveillance (18). For example, Google Flu Trends monitors online search behaviour for early warning signs of influenza (19); researchers have used Facebook to help predict health outcomes at the local population health level (20); Twitter has been used as a large source of data to monitor health trends during an avian influenza outbreak (21); and mobile phones have been used to measure human mobility patterns in the context of malaria transmission in the developing world (22).

Social media has improved emergency response by providing real-time data capture about the health of communities (23) and the public response to an event (24). For example, the use of smartphones and Twitter in Nigeria during the Ebola outbreak in West Africa helped to identify an outbreak in a new area three days before a WHO announcement (25).

Other novel applications include crowdsourcing systems that capture voluntarily submitted symptoms from the general public through the Internet or mobile phone networks and rapidly aggregate and provide feedback about data in near real-time. This has been used by participatory infectious disease surveillance applications such as Flu Near You (26) and DoctorMe (27).

However, there are some inherent challenges in the use of social media data sources. One of the primary challenges of Big Data in general and social media content in particular, is the “signal-to-noise” ratio which can significantly increase the potential for false positives and false negatives. With the influx of discussions and tweets surrounding the Ebola outbreak in West Africa, for example, it was difficult to distinguish between actual signals of concern and the plethora of messages that would otherwise be expected during such an event. In addition, some social media, such as tweets that are limited to 140 characters, may not have enough contextual information to help discern a reliable signal (28).

Another challenge when using social media is representativeness. Not everyone has access to a smart phone and therefore data from social media platforms can only reflect the portion of the population that uses them (28). Mobile technology is expanding significantly so this may help address concerns about representativeness (29).

Finally, the use of social media poses ethical considerations associated with the rights of individuals, including privacy issues (2).

Improving data analysis

Not only might the GPHIN expand its data sources, it could also advance its data analysis capacities. Advanced computational and verification methods to improve the sensitivity and specificity of signals that are detected are being considered (30). Also up for consideration is whether better data processing could reduce reliance on a multilingual, multidisciplinary team. The GPHIN is continuously assessing and honing the aggregators and algorithms used which could potentially result in more advanced forms of artificial intelligence. Continuing to advance the GPHIN’s analytical capacity will enable the robust management, integration, analysis and interpretation of increasingly large and complex volumes of data (31).

Conclusion

Canada's Global Public Health Intelligence Network was an early adopter of Big Data and as an ongoing global resource, helps countries meet event-based surveillance capacity requirements for early detection and reporting of infectious disease outbreaks and other events of international concern. Ongoing advances in Big Data including the use of social media and smart phones, as well as advances in analytical capacity provide opportunities for the further enhancement of the GPHIN. Overall, Big Data approaches have become a vital component of local, national and international public health efforts to detect, report, and control emerging outbreaks.

Acknowledgements

We would like to acknowledge the entire GPHIN team. The success of the GPHIN is attributed to their hard work, dedication and consistent collaboration. Their support and guidance was much appreciated and contributed greatly to the development of this paper.

Conflict of interest

None.

Funding

The GPHIN is funded by the Public Health Agency of Canada.

References

(1) George G, Haas MR, Pentland A. Big Data and management. Acad Manag J. 2014;57(2):321-6.
(2) Vayena E, Salathé M, Madoff LC, Brownstein JS, Bourne PE. Ethical challenges of Big Data in public health. PLoS Comput Biol. 2015;11(2):e1003904.
(3) Muchaal P, Meganath K, Landry L, Aramini J. Evaluation of a national pharmacy-based syndromic surveillance system. Can Comm Dis Rep. 2015 41;9:204-210.
(4) Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, Mawudeku A, et al. Use of unstructured event-based reports for global infectious disease surveillance. Emerg Infect Dis. 2009 May;15(5):689-95.
(5) World Health Organization. [Internet] Epidemic intelligence - Systematic event detection. Geneva: World Health Organization; 2015. Available from: http://www.who.int/csr/alertresponse/epidemicintelligence/en/.
(6) Mykhalovskiy E, Weir L. The Global Public Health Intelligence Network and early warning outbreak detection: A Canadian contribution to global public health. Can J Public Health. 2006 Jan-Feb;97(1):42-4.
(7) McAfee A, Brynjolfsson E, Davenport TH, Patil D, Barton D. Big Data. The management revolution. Harvard Bus Rev. 2012 Oct;90(10):61-7.
(8) Hay SI, George DB, Moyes CL, Brownstein JS. Big Data opportunities for global infectious disease surveillance. PLoS Med. 2013;10(4):e1001413.
(9) Mawudeku A, Blench M, Boily L, St John R, Andraghetti R, Ruben M. The Global Public Health Intelligence Network. In: Infectious Disease Surveillance, Second Edition. New York: John Wiley and Sons; 2013. pp. 457-69.
(10) Heymann DL, Rodier G. Global surveillance, national surveillance and SARS. Emerg Infect Dis. 2004;10(2):173-5.
(11) Mawudeku A, Blench M. Global Public Health Intelligence Network (GPHIN) In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: August 2006. Cambridge, MA, USA.
(12) Davies SE. Nowhere to hide: Informal disease surveillance networks tracing state behaviour. Global Change, Peace & Security. 2012;24(1):95-107.
(13) Davies SE, Youde JR. The Politics of surveillance and response to disease outbreaks: The new frontier for states and non-state actors. Burlington VT: Ashgate Publishing, Ltd.; 2015.
(14) World Health Organization. International Health Regulations. 2008. Second Edition. Geneva: WHO; 2008.
(15) Baker MG, Fidler DP. Global public health surveillance under new international health regulations. Emerg Infect Dis. 2006 Jul;12(7):1058-65.
(16) Warren AP, Bell M, Budd L. Surveillance networks and spaces of governance: Technological openness and international cooperation during the 2009 H1N1 pandemic. Washington: Association of American Geographers; 2010.

(17) Lamy P. Report to the TPRB from the Director-General on the financial and economic crisis and trade-related development. Geneva: World Trade Organization; 2009.
(18) Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. J Med Internet Res. 2013 Jul 18;15(7):e147.
(19) Davidson MW, Haim DA, Radin JM. Using networks to combine Big Data and traditional surveillance to improve influenza predictions. Sci Rep. 2015;5:8154.
(20) Gittelman S, Lange V, Gotway Crawford CA, Okoro CA, Lieb E, Dhingra SS, et al. A new source of data for public health surveillance: Facebook likes. J Med Internet Res. 2015 Apr 20;17(4):e98.
(21) Fung IC, Wong K. Efficient use of social media during the avian influenza A (H7N9) emergency response. Western Pac Surveil Response J. 2013;4(4):1.
(22) Buckee CO, Wesolowski A, Eagle NN, Hansen E, Snow RW. Mobile phones and malaria: Modeling human and parasite travel. Travel Med Infect Dis. 2013;11(1):15-22.
(23) Merchant R, Elmer S, Lurie, N. Integrating social media into emergency-preparedness efforts. N Engl J Med. 2011 July 28;365(4):289-91.
(24) Fung IC, Fu K, Ying Y, Schaible B, Hao Y, Chan C, et al. Chinese social media reaction to the MERS-CoV and avian influenza A (H7N9) outbreaks. Infect Dis Poverty. 2013;2(1):1-12.
(25) Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control. 2015;43(6):563-71.
(26) Wojcik OP, Brownstein JS, Chunara R, Johansson MA. Public health for the people: Participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol. 2014 Jun20;11:7,7622-11-7. eCollection 2014.
(27) Susumpow P, Pansuwan P, Sajda N, Crawley AW. Participatory disease detection through digital volunteerism: How the DoctorMe application aims to capture data for faster disease detection in Thailand. Proceedings of the companion publication of the 23rd International Conference on World Wide Web Companion; International World Wide Web Conferences Steering Committee; 2014.
(28) Guy S, Ratzki-Leewing A, Bahati R, Gwadry-Sridhar F. Social media: A systematic review to understand the evidence and application in infodemiology. In: Electronic Healthcare. New York: Springer; 2012. p. 1-8.
(29) Geneva: International Telecommunications Union; 2013. Available from: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf.
(30) Kostkova P. A roadmap to integrated digital public health surveillance: The vision and the challenges. Proceedings of the 22nd International Conference on World Wide Web Companion; International World Wide Web Conferences Steering Committee; 2013.
(31) Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: A systematic review. Milbank Q. 2014;92(1):7-33.