CANADA COMMUNICABLE DISEASE REPORT

CCDR

40 years! (1975-2015)

September 3, 2015 • VOLUME 41•9
ISSN 1481–8531

Inside this issue: Big Data

Big Data has begun to transform how we identify, track and control infectious diseases. Big Data is produced through the analysis of large data sets to glean information that was previously unavailable. Initially pioneered in statistical analysis and computing, Big Data is now increasingly being employed in scientific research. In this issue, read how Big Data from national pharmaceutical sales can help track influenza and other infectious diseases, how scanning 3,000 news items daily from around the world in close to real-time can help identify emerging infectious diseases, how genome sequencing has increased our detective capacity to track the origins of outbreaks and how the analysis of multiple datasets can even help predict future outbreaks.

Call for papers:

Manuscript submissions are invited for two upcoming theme issues. Due dates and topics are:

Nov 12, 2015 Determinants of health, health equity and infectious diseases
Dec 3, 2015 Infectious disease as chronic disease

Interested? See Information for authors. If you have questions, contact the Editor-in-Chief: patricia.huston@phac-aspc.gc.ca

Public Health Agency of Canada | Agence de la santé publique du Canada

Canada

Canada Communicable Disease Report

The *Canada Communicable Disease Report* (CCDR) is a bilingual, peer-reviewed, open-access online scientific journal published by the Public Health Agency of Canada (PHAC). It provides timely, authoritative and practical information on infectious diseases to clinicians, public health professionals and policy-makers to inform policy, program development and practice.

Canada Communicable Disease Report
Public Health Agency of Canada
130 Colonnade Rd.
Address Locator 6503B
Ottawa, Ontario K1A 0K9
Email: CCDR-RMTC@phac-aspc.gc.ca

To promote and protect the health of Canadians through leadership, partnership, innovation and action in public health.
Public Health Agency of Canada

Published by authority of the Minister of Health.

ISSN 1481-8531
Pub.150005

This publication is also available online at http://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/15vol41/index-eng.php
Également disponible en français sous le titre: *Relevé des maladies transmissibles au Canada*

Surveillance

Evaluation of a national pharmacy-based syndromic surveillance system

Muchaal PK[1]*, Parker S[1], Meganath K[1], Landry L[1], Aramini J[2]

[1]Centre for Foodborne, Environmental and Zoonotic Infectious Diseases, Public Health Agency of Canada, Guelph, ON
[2]Intelligent Health Solutions, Fergus, ON

*Correspondence: pia.muchaal@phac-aspc.gc.ca

Abstract

Background: Traditional public health surveillance provides accurate information but is typically not timely. New early warning systems leveraging timely electronic data are emerging, but the public health value of such systems is still largely unknown.

Objective: To assess the timeliness and accuracy of pharmacy sales data for both respiratory and gastrointestinal infections and to determine its utility in supporting the surveillance of gastrointestinal illness.

Methods: To assess timeliness, a prospective and retrospective analysis of data feeds was used to compare the chronological characteristics of each data stream. To assess accuracy, Ontario antiviral prescriptions were compared to confirmed cases of influenza and cases of influenza-like-illness (ILI) from August 2009 to January 2015 and Nova Scotia sales of respiratory over-the-counter products (OTC) were compared to laboratory reports of respiratory pathogen detections from January 2014 to March 2015. Enteric outbreak data (2011-2014) from Nova Scotia were compared to sales of gastrointestinal products for the same time period. To assess utility, pharmacy sales of gastrointestinal products were monitored across Canada to detect unusual increases and reports were disseminated to the provinces and territories once a week between December 2014 and March 2015 and then a follow-up evaluation survey of stakeholders was conducted.

Results: Ontario prescriptions of antivirals between 2009 and 2015 correlated closely with the onset dates and magnitude of confirmed influenza cases. Nova Scotia sales of respiratory OTC products correlated with increases in non-influenza respiratory pathogens in the community. There were no definitive correlations identified between the occurrence of enteric outbreaks and the sales of gastrointestinal OTCs in Nova Scotia. Evaluation of national monitoring showed no significant increases in sales of gastrointestinal products that could be linked to outbreaks that included more than one province or territory.

Conclusion: Monitoring of pharmacy-based drug prescriptions and OTC sales can provide a timely and accurate complement to traditional respiratory public health surveillance activities but initial evaluation did not show that tracking gastrointestinal-related OTCs were of value in identifying an enteric disease outbreak in more than one province or territory during the study period.

Introduction

In Canada, traditional public health surveillance of infectious diseases relies heavily on the reporting of laboratory-confirmed cases. This mechanism provides robust information, but may be accompanied by an appreciable lag period between testing and reporting of illness to public health authorities. This results in a lost window of opportunity for implementing interventions. Furthermore, for infectious diseases typically associated with mild to moderate illness, treatment is often empiric (i.e., no testing is done) which limits the ability of laboratory-based surveillance to provide an accurate assessment of illness in the community.

Syndromic surveillance often involves Big Data; it is based on the use of non-specific health indicators or proxy measures (e.g., school absenteeism, drug sales, tele-health calls) to provide a provisional diagnosis (or "syndrome"). These data sources tend to be non-specific yet sensitive and rapid and can augment and complement the information provided by traditional diagnostic test-based surveillance systems (1).

Over the past 10 years, pharmacy-based surveillance has emerged as a new public health capability in Canada and internationally (2, 3, 4, 5). The Public Health Agency of Canada (the Agency) first deployed a pharmacy-based syndromic surveillance in 2004 to evaluate the feasibility of monitoring gastrointestinal over-the-counter (OTC) products as a tool in the early detection of community food and waterborne illness and outbreaks. In 2009, pharmacy-based syndromic surveillance was again used by the Agency in response to the influenza H1N1 pandemic (pH1N1) outbreak. A retrospective analysis of the second wave of pH1N1 demonstrated that pharmacy-based surveillance provided an effective mechanism to monitor and detect influenza-like activity and was faster than traditional surveillance systems (6).

Following the H1N1 pandemic, the pharmacy-based surveillance system was extended to determine utility during implementation of the Agency's *Surveillance Strategic Plan*. A number of medications were tracked, including analgesics, anti-allergy medications and prescriptions of antidepressants and cardiovascular medications.

The objective of this study was to evaluate three aspects of this system: the accuracy, timeliness and utility of pharmacy-based surveillance to support seasonal influenza and respiratory illness surveillance and detection of multijurisdictional enteric disease outbreaks (i.e., involving more than one province or territory).

Methods

Data sources

Data was obtained for respiratory surveillance on antiviral prescriptions in Ontario (2009-2015) and respiratory OTCs in Nova Scotia (2014-2015). These were compared with confirmed cases of influenza and reports of influenza-like-illness in Ontario and laboratory detections of non-influenza respiratory pathogens in Nova Scotia.

Data was obtained for gastrointestinal surveillance on OTC sales of products indicated for acute gastrointestinal illness across Canada and data from Nova Scotia was compared with local outbreak data reported to provincial health authorities.

Data collection and reporting

Pharmacy data acquisition was established through a contract with an industry partner, Rx Canada. Daily prescription sales data were sourced from 13 national pharmacy chains and four independents, representing over 3,000 stores nationwide. Over-the-counter (OTC) data were provided daily by six chain retailers representing 1,863 stores across Canada (excluding Nunavut). Both OTC and prescription datasets provided coverage for over 85% of the health regions across Canada.

Pharmacy products were categorized into syndromes by grouping products into respiratory or gastrointestinal categories. OTC sales data were standardized by store prior to aggregating by health unit, province/territory and nationally to minimize the effect of varying transmission frequency and timeliness between stores and retail chains. The proportion of daily sales of select products indicated for acute enteric illness was calculated and presented over sales of all other OTC products. The standardized seven-day moving average of gastrointestinal sub-categories divided by other OTCs was graphed on a log scale. A simple alert algorithm based on 1, 2 and 3 standard deviations above the seven-day moving average was used to detect aberrant increases in pharmacy sales.

Timeliness

Prospective and retrospective analysis of the various data feeds leveraged in this study (OTC, prescriptions, ILI, laboratory reporting) were compared to assess the timeliness of public health surveillance.

Accuracy

Accuracy of pharmacosurveillance for influenza and influenza-like illness was assessed by a retrospective analysis comparing pharmaceutical and surveillance data. Surveillance data was extracted from the FluWatch Surveillance System. Five years of prescription antiviral sales in Ontario were compared to reported cases of influenza-like-illness (ILI) between August 2009 and January 2015 and to confirmed cases of influenza between 2011 and 2015. Together with descriptive analysis, Spearman correlation coefficients (*rho*) were determined for three time frames (January to December, November to March and April to October).

To assess tracking accuracy for other respiratory infections, weekly provincial laboratory detections of respiratory viral pathogens from Nova Scotia reported between January 2014 and March 2015 were evaluated relative to sales of respiratory over-the-counter products for the same period. Respiratory OTC data were aggregated by week and one-to-five week lag variables were created. Correlations using Pearson's coefficients were calculated between respiratory OTC sales, non-influenza Respiratory Viral Detections (RVD) and tests for respiratory syncytial virus (RSV).

To assess the accuracy of enteric pharmacosurveillance, sales of gastrointestinal OTCs were compared with enteric outbreak data from Nova Scotia between 2011 and 2014 using descriptive analysis.

Utility

Pharmacy surveillance reports based on the pharmacy sales of gastrointestinal products across Canada were generated and disseminated weekly to provincial and territorial stakeholders and included province and health region-level information and trends. The usefulness of the reported information to stakeholders was evaluated using a survey developed using FluidSurvey.

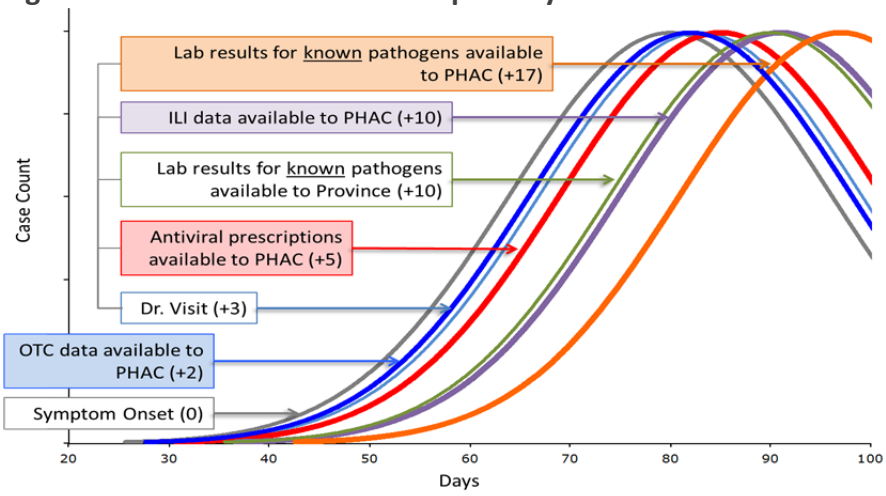
SAS 9.3 and Stata 13 were used for running analytical procedures.

Results

Timeliness

Pharmacy sales data (i.e., prescription medication, OTC products) were available in near real-time; approximately 48 hours after a completed transaction. Pharmacy sales data were available to the Agency approximately five to eight days earlier than ILI physician reports, 10-12 days prior to laboratory confirmations of influenza, and up to 17 days earlier than reports of respiratory viral detections. **Figure 1** depicts the timelines of pharmacy, clinical and laboratory data relative to the estimated onset of illness and the availability of the information for the purpose of respiratory surveillance**.**

Figure 1: Timeliness of Severe Respiratory Disease Surveillance

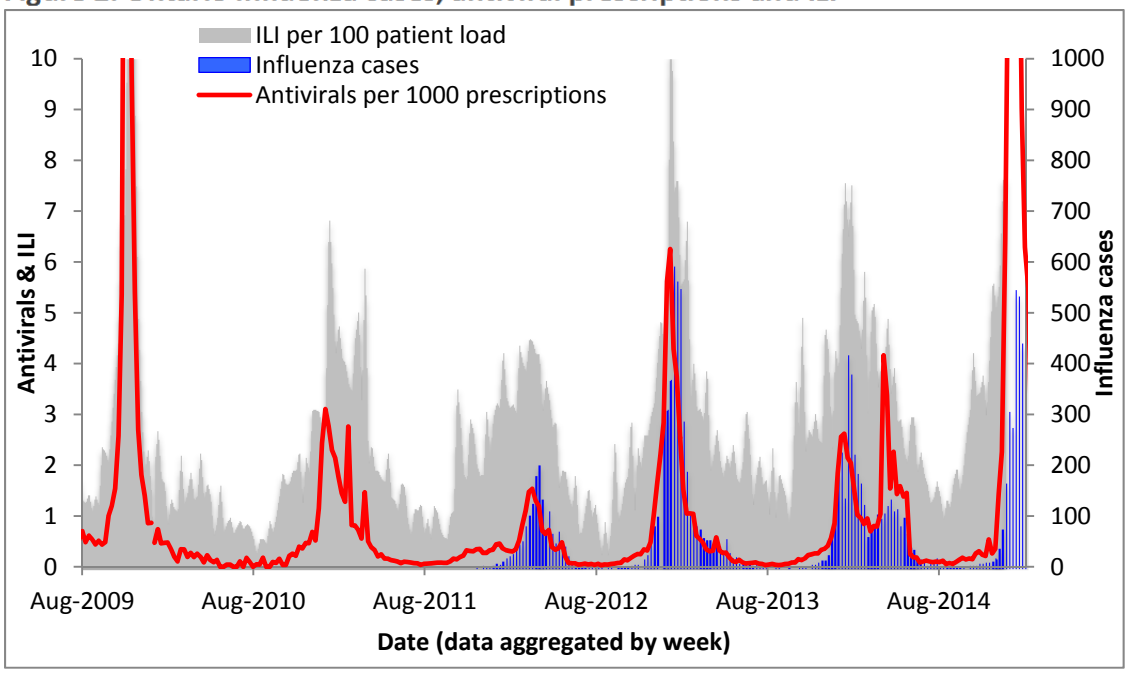


ILI: Influenza-like-illness PHAC: Public Health Agency of Canada OTC: Over-the-counter

Accuracy

Respiratory surveillance: Weekly dispensed prescriptions of the antiviral medications oseltamivir and zanamivir, as a proportion of all dispensed prescriptions, were compared with the rate of ILI (cases of ILI/total patient visits) and the number of confirmed laboratory reports of influenza in Ontario (**Figure 1**). Sales of antivirals between 2009 and 2015 correlated closely with the onset dates of confirmed influenza cases. Spearman's *rhos* were 0.90, 0.93 and 0.83 (all at $p < 0.001$) for January to December, November to March and April to October respectively. In addition, the magnitude of antiviral sales paralleled the burden of confirmed cases. The correlation coefficient between confirmed cases and ILI was not as strong comparably. For January and December the *rho* was 0.80, in November and March the *rho* was 0.70 and from April to October the *rho* was 0.60. Reports of ILI followed the same trends of confirmed cases and antivirals, (**Figure 2**) however, ILI varied considerably in late spring and summer.

Figure 2: Ontario influenza cases, antiviral prescriptions and ILI[1]



[1]ILI: Influenza-like- illness

Results of the Nova Scotia analysis demonstrated sales of respiratory OTC products associated with increases in non-influenza respiratory pathogens in the community (respiratory syncytial virus, rhinovirus, coronavirus, parainfluenza, adenovirus, human metapneumovirus). OTC sales correlated with the number of RSV tests four weeks later ($r=0.76$, $p < 0.001$). Sales of OTC products correlated with the number of other respiratory viral detections two weeks later (0.74, $p < 0.001$) (**Figure 2**).

Gastrointestinal surveillance: There were 262 outbreaks reported in Nova Scotia between 2012 and 2014. Of these, 66% of outbreaks were due to norovirus; 82% of outbreaks were primarily person-to-person transmission and the majority (81%) of outbreaks occurred in residential institutions. There were no definitive correlations detected between the occurrence of outbreaks and sales of gastrointestinal OTCs.

Utility

During the pilot phase, no significant increases in sales of gastrointestinal products were detected that could be definitively linked to any multijurisdictional enteric outbreak. Although only 40% of stakeholders responded to the evaluation survey, most (87%) indicated that the pharmacy information was important to their jurisdiction. Five jurisdictions stated that they had used the information. Of these 20% used it to initiate

an investigation, 75% to detect a health event, 20% communicated the information to clinicians and all used it for situational awareness.

Discussion

National and provincial pharmacosurveillance appeared to be more useful for early detection of respiratory illness than for provincial detection of enteric disease. Antiviral prescriptions were a clear marker of influenza activity for respiratory illness surveillance. The value of respiratory-related OTCs was less clear, but they may be useful in the surveillance of other respiratory viruses. Although the sales of gastrointestinal products have been shown to be a good marker for seasonal community norovirus infections (6), the value of pharmacosurveillance for outbreak-related enteric activity was less certain.

Strengths of this study include the national representativeness of the data and the precise documentation of pharmacosurveillance data timeliness. One potential weakness is that it included only prescribed medications and OTCs from retail pharmacies not linked to health care institutions. For example, although Nova Scotia enteric outbreak data were robust, the majority of norovirus outbreaks captured during the study period occurred in long-term care facilities. Medications provided to long-term care facilities may have originated from pharmacies not contributing sales data to the pharmacy pilot project or perhaps bulk purchases by long-term care facilities were simply not captured in the data provided.

Automated data transfer from point of sale supports real or near real-time acquisition of community-specific health data. Pharmacy prescription and OTC sales data provide greater timeliness and higher geographic resolution and national coverage than other health data currently used. Pharmacy sales data has been shown to support seasonal influenza and aid in decision-making for resource allocation. Prescription and OTC data also provide an earlier window of opportunity for intervention compared to using traditional data sources only. Furthermore, pharmacy data could be used to support surveillance and action for multiple public health purposes (e.g., mental health, physician prescription patterns and chronic illness) and by multiple agencies and different jurisdictions. As with other surveillance systems, investments in the necessary technological and analytical resources are required to conduct ongoing pharmacy-based syndromic surveillance.

Conclusion

Timely and accurate pharmacosurveillance has the potential to enhance public health capacity to detect and quantify activity at the local, multijurisdictional and national level. Further research is needed to determine under which conditions it is most useful and to compare it against other real-time surveillance strategies.

Acknowledgements

We wish to thank the provinces and territories for providing case data for the analysis and participating in the pilot study and those at Flu Watch for providing the respiratory surveillance data. The strong commitment and support of the Chain Retailers, pharmacy stores and Rx Canada was pivotal in the success of the various phases of the Pharmacy-based Syndromic Surveillance pilot.

Conflict of interest

Jeffery Aramini works for Intelligence Health Solutions, the company that provided the analysis for the data.

Funding

Support and funding of the of the pharmacosurveillance project was provided by the Public Health Agency of Canada

References

(1) Berger M, Shiau R, Weintraub JM. Review of syndromic surveillance: Implications for waterborne disease detection. J Epidemiol Community Health. 2006 Jun;60(6):543–550.

(2) Chadwick D. The Rhode Island Department of Health. The first statewide system for tracking disease using prescription data. Press Release Archives. State of Rhode Island. Department of Health. 2009. Available from: http://www.ri.gov/press/view/10017.

(3) Pavlin JA, Murdock P, Elbert E. Conducting population behavioral health surveillance by using automated diagnostic and pharmacy data systems. MMWR. 2004;53:166–172.

(4) Sugawara T, Ohkusa Y, Ibuka Y, Kawanohara H, Taniguchi K, et al. Real-time prescription surveillance and its application to monitoring seasonal influenza activity in Japan. J Med Internet Res. 2012;14(1):e14. Available from: http://www.jmir.org/2012/1/e14/.

(5) Van den Wijngaard C, van Pelt W, Nagelkerke N, Kretzschmar M, Koopmans M. Evaluation of syndromic surveillance in the Netherlands: Its added value and recommendations for implementation. Euro Surveill. 2011;16(9):19806. Available from: http://www.eurosurveillance.org/images/dynamic/EE/V16N09/art19806.pdf.

(6) Aramini J, Muchaal PK, Pollari F. Value of pharmacy-based influenza surveillance — Ontario, Canada, 2009. MMWR. 2013 May 24; 62(20);401-404.

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).

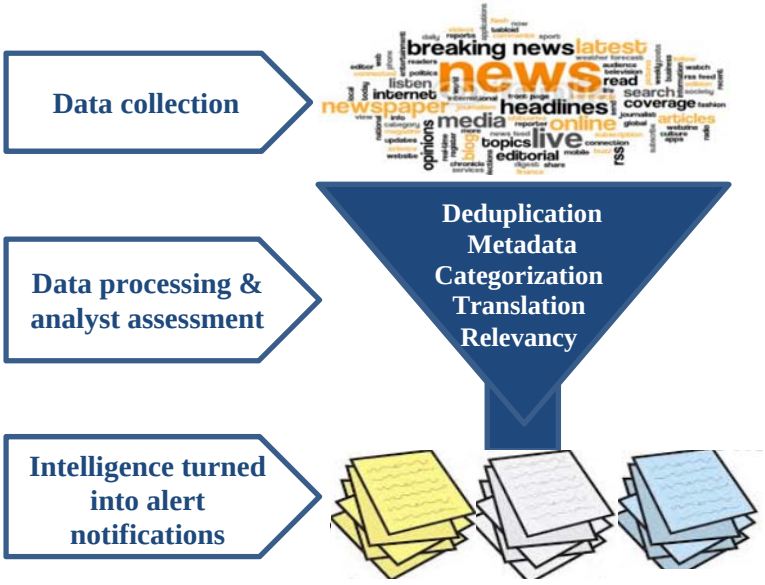
The objective of this article is to identify how the GPHIN functions within the context of Big Data, to provide recent examples of the GPHIN in action and to explore potential future directions.

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) e 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.

Commentary

Big Data is changing the battle against infectious diseases

Links MG[1,2*]

[1]Saskatoon Research Centre, Agriculture and Agri-Food Canada, Saskatoon, SK
[2]Department of Computer Science, University of Saskatchewan, Saskatoon, SK

*Correspondence: Matthew.Links@usask.ca

Abstract

Big Data has traditionally been associated with computer geeks and commercial enterprises, but it has become entrenched in many scientific disciplines including the prevention and control of infectious diseases. The use of Big Data has allowed disease trends to be identified and outbreak origins to be tracked and even predicted. Big Data is not getting smaller. The challenges we face are to hone our analytical capacity to address the huge "signal-to-noise" ratio with adequate computing power and multidisciplinary teams that can handle ever-increasing amounts of data. Big Data will also create the opportunity for future applications of *bespoke* (or personalized) treatment.

Introduction

Big Data seems like a recent development that for many is tied to phrases like "cloud computing". The term originated before the turn of the millennium when in the 1990's when John Mashey was Chief Scientist of Silicon Graphics (SGI). At the time, SGI was at the forefront of computer graphics and was struggling to deal with significantly expanding computational needs that outpaced available hardware. Mashey developed a presentation in the late '90s that laid out the looming collision between Big Data and computational performance (1).

Commonly The term "Big Data" is now used to describe situations where data volumes are characterized by properties including, but not limited to: size, rate of change over time, and the heterogeneous nature of the data itself (2). Big Data generally refers to large volumes of data that can be structured (e.g. relational databases) or not (e.g. Twitter feeds) and are mined for information. While historically Big Data originated within the fields of Computer Science, Statistics and Economics(3), it has been increasingly adopted across all scientific disciplines.

A problem is typically said to involve Big Data when the volume is so large that it hinders the ability to convert data to knowledge. In the case of infectious disease research, Big Data is having a huge impact. The ability to perform real-time disease tracking and outbreak prediction has utilized unstructured data to change how infectious diseases are managed. For example, through the use of diverse news sources GPHIN has been used for early signal of novel infections (such as SARS and MERS-CoV) that informed public health response to the outbreaks that followed (4).

Structured data is particularly useful when collating information from multiple sources based on a predictable structure of the data. In the case of public health surveillance it is now possible to look for structured data that could serve as a surrogate source of information to laboratory confirmation or physician authored case reporting. Muchaal et al. demonstrated that pharmaceutical usage is one possible source of early surrogate information (5).

How big is Big Data?

It is hard to fathom how large Big Data actually is. In a recent paper Stephens and colleagues put Big Data from a couple of disciplines into a relative context (6). While the current champions of unwieldy data size are

astronomical studies, it was suggested by Stephens et al. that genomics would be on par with astronomical data sizes by the year 2025. The scale of genomics data in 2025 may be equivalent to 8 billion of the largest iPhones available today (128GB of space in 2015). Or an iPhone's worth of data for every person on earth, each year.

Big Data and the origins of an outbreak

The application of genomics to infectious diseases can help with identifying where infectious disease outbreaks actually came from. A good example of this was the detective work undertaken to respond to a measles outbreak during the 2010 Olympics (7). Using whole genome sequencing Gardy and colleagues were able to exactly identify many of the reported cases (30 of 82). One important finding was that there was more than one type of measles virus involved in the outbreak. While traditional genotyping for the measles virus has focused on the sequences of phosphoprotein and hemmagglutinin, two specific genes used to distinguish isolates, Gardy et al. showed that there were additional variations in other measles genes that could be used for a more precise definition of the viral lineages.

What's next?

One of the newer applications of Big Data is in what Jennifer Gardy termed *bespoke* (or personalized) treatment (8). For example, whole genome approaches across thousands of isolates can identify genomic variation linked with antimicrobial phenotypes of *Mycobacterium tuberculosis* (9). In the future, this may be applied generally to identify the best treatments for bacteria with antimicrobial resistance.

How are we going to interpret it all?

Despite all the potential for advances, there are some key challenges that Big Data faces in all disciplines. As data gets bigger and bigger it becomes harder to interpret: either the integration of data is so complex as to be hard to follow without serious computing resources or the sheer scale is beyond comprehension.

With the use of unstructured text from news feeds mined for disease surveillance knowledge, Big Data is being used to find meaning in a deluge of noise. The scale of text mining scientific manuscripts published in journals, news reports describing emergent issues and 140 character tweets truly is daunting when one considers that Twitter collects ½ billion tweets per day (6). What makes matters even more challenging is that disease surveillance doesn't simply mean aggregating news feeds. Rather, the key for a meaningful Big Data strategy to disease surveillance is the identification of the potential risk. Approaches such as GPHIN (4) are thus crucial for national and international preparedness for disease outbreaks. As Big Data continues to grow at exponential rates, trying to advance our capacity to analyze it becomes an ever-changing holy grail.

All too often a Big Data-set is acquired as part of a multi-disciplinary study and handed-off to a single individual (e.g. graduate student , post-doc or fellow) in the hope that they can, alone, come to an understanding of what it all means. Having a single person responsible for increasingly complex relationships arising from overwhelming volumes of data is just not a feasible strategy. Thus the trend is to develop multidisciplinary approaches to interpretation of Big Data (4).

Conclusion

Unless we envision a computational Dark Ages it is hard to believe that Big Data will shrink. Therefore scientific disciplines have been developing the capacity to exploit ever-increasing volumes of data. Care needs to be taken not to be overwhelmed with Big Data. In fact it is the shift to multi-disciplinary analysis of Big Data that is enabling teams to track disease trends and even predict outbreaks before they occur. Big Data is positioned to move increasingly from public health into the clinical setting; *bespoke* (or personalized) treatment of infectious diseases may soon be on our doorstep.

Funding

Dr. Links' research program has, or is currently, funded by Agriculture and Agri-Food Canada (AAFC), the Government of Canada's Genomics Research Development Initiative (GRDI), the Canadian Institutes of Health Research (CIHR), the National Research Council of Canada (NRC), Saskatchewan's Agriculture Development Fund (ADF), the Government of Canada's Canadian Safety and Security Program (CSSP, formerly CRTI - Chemical, Biological, Radiological-Nuclear Research and Technology Initiative).

Conflict of interest

None.

References

(1) Mashey J, ed. Big Data and the next wave of infrastress. USENIX Annual Technical Conference; 1998; Monterey, California, USA: Usenix.
(2) Asokan GV, Asokan V. Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. J Epidemiol Glob Health. 2015 Mar 5.
(3) Diebold FX. A personal perspective on the origin(s) and development of 'Big Data': The phenomenon, the term, and and the discipline. Department of Economics, University of Pennsylvania. PIER Working Paper No. 13-003. November 2012.
(4) Dion M, AbdelMalik P, Mawudeku A. Big Data and the Global Public Health Intelligence Network (GPHIN). Can Commun Dis Rep. 2015;41:211-216.
(5) Muchaal PK, Parker S, Meganath K, Landry L, Aramini J**.** Evaluation of a national pharmacy-based syndromic surveillance system Can Commun Dis Rep. 2015;41:204-211..
(6) Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or genomical? PLoS Biol. 2015 Jul;13(7):e1002195.
(7) Gardy JL, Naus M, Amlani A, Chung W, Kim H, Tan M, et al. Whole-genome sequencing of measles virus genotypes H1 and D8 during outbreaks of infection following the 2010 Olympic Winter Games reveals viral transmission routes. J Infect Dis. 2015 Jul 6.
(8) Gardy JL. Towards genomic prediction of drug resistance in tuberculosis. Lancet Infect Dis. 2015 Jun 23.
(9) Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study. Lancet Infect Dis. 2015 Jun 23.

ID News: Big Data and predicting, preventing and controlling outbreaks

Christaki E. **New technologies in predicting, preventing and controlling emerging infectious diseases.** Virulence. 2015 Jun 11:1-8. (*Summary*)

Surveillance of emerging infectious diseases is vital for the early identification of public health threats. Emergence of novel infections is linked to human factors such as population density, travel and trade and ecological factors like climate change and agricultural practices. A wealth of new technologies is becoming increasingly available for the rapid molecular identification of pathogens but also for the more accurate monitoring of infectious disease activity. Web-based surveillance tools and epidemic intelligence methods, used by all major public health institutions, are intended to facilitate risk assessment and timely outbreak detection. This review presents new methods for regional and global infectious disease surveillance and advances in epidemic modeling aimed to predict and prevent future infectious diseases threats.

Semenza JC. **Prototype early warning systems for vector-borne diseases in Europe.** Int J Environ Res Public Health. 2015 Jun 2;12(6):6333-51. doi:10.3390/ijerph120606333. (*Summary*)

Globalization and environmental change, social and demographic determinants and health system capacity are significant drivers of infectious diseases which can also act as epidemic precursors. Thus, monitoring changes in these drivers can help anticipate, or even forecast, an upsurge of infectious diseases. The European Environment and Epidemiology (E3) Network has been built for this purpose and applied to three early warning case studies: 1) The environmental suitability of malaria transmission in Greece was mapped in order to target epidemiological and entomological surveillance and vector control activities. Malaria transmission in these areas was interrupted in 2013 through such integrated preparedness and response activities. 2) Since 2010, recurrent West Nile fever outbreaks have ensued in South/eastern Europe. Temperature deviations from a thirty year average proved to be associated with the 2010 outbreak. Drivers of subsequent outbreaks were computed through multivariate logistic regression models and included monthly temperature anomalies for July and a normalized water index. 3) Dengue is a tropical disease but sustained transmission has recently emerged in Madeira. Autochthonous transmission has also occurred repeatedly in France and in Croatia mainly due to travel importation. The risk of dengue importation into Europe in 2010 was computed with the volume of international travelers from dengue-affected areas worldwide.These prototype early warning systems indicate that monitoring drivers of infectious diseases can help predict vector-borne disease threats.

Hay SI, George DB, Moyes CL, Brownstein JS. **Big Data opportunities for global infectious disease surveillance.** PLoS Med 2013;10(4):e1001413. doi:10.1371/journal.pmed.1001413. (*Summary*)

Systems to provide static spatially continuous maps of infectious disease risk and continually updated reports of infectious disease occurrence exist but to date the two have never been combined. Novel online data sources, such as social media, combined with epidemiologically relevant environmental information are valuable new data sources that can assist the “real-time” updating of spatial maps. Advances in machine learning and the use of crowd sourcing open up the possibility of developing a continually updated atlas of infectious diseases. Freely-available dynamic infectious disease risk maps would be valuable to a wide range of health professionals from policy-makers prioritizing limited resources to individual clinicians.

ID News: Big Data and ethics

Ploug T, Holm S. **Meta consent: A flexible and autonomous way of obtaining informed consent for secondary research.** BMJ. 2015 May 7;350:h2146. doi:10.1136/bmj.h2146. (*Summary*)

A rapidly increasing capability for storing, linking and analyzing health data has led to new opportunities for research. However, it also raises new ethical and regulatory concerns. Central among these is the question of the conditions under which secondary research can use data that were collected as part of routine healthcare practice or for a specific research project. Does secondary use require renewed informed consent from the original participants? Consent to date has included: dynamic (when information about specific secondary use of health data or tissue is requested each time to each individual through a web-based platform), broad (when consent is given to future research of a particular type in addition to the current specific research project) or blanket (data could be used without further consent). We propose meta consent which means individuals can choose how they wish to provide consent for future secondary research of data collected in the past or of data that will be stored in the future, thus meta consent is both retrospective and prospective. Meta consent is a truly individual consent procedure that takes into account the differences in personal interests and levels of trust in researchers among the population. The risk of routinisation is reduced because individuals can limit the requests they receive to only those categories of research that really matter to them. Its implementation online makes meta consent easy to revoke or change.

Mittelstadt BD, Floridi L. **The ethics of Big Data: Current and foreseeable issues in biomedical contexts.** Sci Eng Ethics. 2015 May 23. [Epub ahead of print]. (*Summary*)

The capacity to collect and analyze data is growing exponentially. Referred to as 'Big Data', this scientific, social and technological trend has helped create destabilising amounts of information, which can challenge accepted social and ethical norms. As is often the case with the cutting edge of scientific and technological progress, understanding of the ethical implications of Big Data lags behind. By means of a meta-analysis of the literature, a thematic narrative is provided to guide ethicists, data scientists, regulators and other stakeholders through what is already known or hypothesised about the ethical risks of this emerging and innovative phenomenon. Five key areas of concern are identified: 1) informed consent, 2) privacy (including anonymisation and data protection), 3) ownership, 4) epistemology and objectivity and 5) 'Big Data Divides' created between those who have or lack the necessary resources to analyze increasingly large datasets. Six additional areas of concern are then suggested which, although related have not yet attracted extensive debate in the existing literature: 6) the dangers of ignoring group-level ethical harms; 7) the importance of epistemology in assessing the ethics of Big Data; 8) the changing nature of fiduciary relationships that become increasingly data saturated; 9) the need to distinguish between 'academic' and 'commercial' Big Data practices in terms of potential harm to data subjects; 10) future problems with ownership of intellectual property generated from analysis of aggregated datasets; and 11) the difficulty of providing meaningful access rights to individual data subjects that lack necessary resources. Considered together, these eleven themes provide a thorough critical framework to guide ethical assessment and governance of emerging Big Data practices.