Big Data for Infectious Diseases Surveillance and the Potential Contribution to the Investigation of Foodborne Disease in Canada: An Overview and Discussion Paper

May 2017

Cheryl Waldner DVM PhD
Professor, University of Saskatchewan

With contributions by:

Nathaniel Osgood, PhD
Associate Professor, University of Saskatchewan

Patrick Seitzinger
MPH thesis Candidate, School of Public Health

National Collaborating Centre for Infectious Diseases
Centre de collaboration nationale des maladies infectieuses

Big data for infectious disease surveillance and the potential contribution to the investigation of foodborne disease in Canada: An overview and discussion paper

Prepared for the National Collaborating Centre for Infectious Diseases (NCCID)

May, 2017

Cheryl Waldner DVM PhD
Professor, University of Saskatchewan

With contributions by:

Nathaniel Osgood, PhD
Associate Professor, University of Saskatchewan

Patrick Seitzinger
MPH thesis Candidate, School of Public Health

Contact us at:

National Collaborating Centre for Infectious Diseases
Rady Faculty of Health Sciences,
University of Manitoba
Tel: (204) 318-2591
Email: nccid@umanitoba.ca
www.nccid.ca

This is NCCID Project number 332

Production of this document has been made possible through a financial contribution from the Public Health Agency of Canada through funding for the National Collaborating Centre for Infectious Diseases. The views expressed herein do not necessarily represent the views of the Public Health Agency of Canada.

Big Data for Infectious Diseases Surveillance and the Potential Contribution to the Investigation of Foodborne Disease in Canada: An Overview and Discussion Paper

Executive Summary

Recognizing the potential for 'big data' as "an untapped resource of evidence that may be used to inform policy and decision making", the National Collaborating Centre for Infectious Disease commissioned a background document to initiate discussion among Canadian public health professionals. The primary objective was to examine the potential for big data to inform public health policy for infectious disease management in Canada. The first part of the report introduces big data and various sources of big data for infectious disease surveillance as described in the peer-reviewed literature and publically available technical reports to December 31, 2016. The report also explores some of the options and challenges in visualizing and analyzing the resulting information so that it can be used more effectively in decision making. Finally, this document highlights types of big data that have been applied to surveillance and investigation of foodborne illness, provides commentary from a researcher who is actively working on the application of big data to foodborne disease, and summarizes the findings of an informal 'big data' survey of public health professionals working in infectious disease and outbreak investigation in Canada.

Big data is most often characterized by the challenges to data management and analysis resulting from the size or amount of information to be examined (volume); the rate at which the data is collected, transmitted or received (velocity); and the potential range of sources, file types and data structures encountered (variety). A number of different examples of big data have been explored for application in infectious disease surveillance and outbreak investigation. The most commonly reported examples included: whole genome sequencing (WGS), news reports aggregated from the internet, records of internet search engine queries or use of web pages, stories shared on internet forums, information from social media sites, data collected both passively and actively through smartphone use, retail records from pharmacies and grocery stores, electronic health records, active internet-based crowd-sourced surveillance, health call in phone line records, absenteeism records, and earth observation data (including remote sensing).

A number of original research papers and review articles have addressed the application of these different types of big data in public health. From that work several potential advantages were identified that could motivate the inclusion of emerging big data sources in existing programs to enhance both infectious disease surveillance and investigation of disease outbreaks. These included: 1.) improving the timeliness, geographic resolution, and completeness of information, 2.) addressing coverage gaps in existing surveillance programs, 3.) increasing sensitivity of surveillance for emerging and re-emerging diseases, and 4.) providing evidence to better inform predictive models and planning for disease management. One of the highlighted examples was the unprecedented growth in opportunities presented by smartphones and wearables for understanding human behavior and the impact of human behavior on disease transmission and control opportunities. These smartphones and other similar wearable devices provide the potential for real-time and high resolution data on location, activity, and contacts; the opportunity to use context-specific and user-triggered surveys; and tools for biometric, image, video and audio data collection.

One of the specific objectives of this report was to explore areas where big data have been used to enhance enteric disease surveillance and investigation of foodborne disease outbreaks. The first and one of the most apparent impacts of big data on foodborne disease is in the area of diagnostics. Whole genome sequencing (WGS) is no longer reserved for research laboratories and the National Microbiology Laboratory, rather most provincial diagnostic laboratories now have access to the necessary technology. While pulse-field gel electrophoresis (PFGE) has been the standard for identification of closely related strains, WGS is emerging as an increasingly affordable and efficient alternative to PFGE that offers comparatively more efficient data storage and the potential for finer resolution in differentiating organisms, in addition to the prediction of strain type, virulence and antimicrobial resistance.

The second biggest area of study of big data and food borne disease investigation is the application of crowd sourced data extracted from internet forums including online restaurant review sites, such as Yelp, and from social media, with Twitter being the most frequently cited. Several larger cities in the United States have used data from either Yelp or Twitter to target restaurant inspections in some cases with very promising results regarding identifying premises with serious health code violations. One of the other objectives described for these initiatives was to decrease the time from the onset of an outbreak to when it is identified by the local health authority and investigated. The goal was to decrease the total number of people affected. However, given the unstructured, diverse and ever changing nature of the raw data, sophisticated machine-learning

algorithms are required to identify potential cases with sensitivity and specificity adequate to limit the amount of time needed for expert review.

There is also ongoing interest in the collection and analysis of retail information to identify disease outbreaks and potential sources of food-borne disease. Pharmacy-based surveillance of both prescription and non-prescription drugs has been examined both in Canada and internationally with varying levels of success for enteric illness. Data from shopper loyalty cards have also been used in a few cases in Canada to identify the most likely causes of foodborne disease outbreaks. IBM has recently been exploring the use of a spatio-temporal analytics system to compare retail scanning data to foodborne disease cases and then generate a list of the most probable suspect foods. The goal was to decrease the time to identifying the cause and decrease the total number of cases.

Several authors stressed the need for the development and application of dynamic simulation and risk models in order to best leverage the increasing volume and variety of data for effective decision support. Examples of where models derived from big data have been most effectively applied to date include the understanding of influenza, dengue, and Zika virus. However, researchers cautioned about the volume and variety model inputs that are constantly updating as new data become available, and the volume and velocity of resulting model outputs. As a result, the efficient and effective use of predictive and dynamic models in disease surveillance and outbreak management requires its own big data resources for programming and data management.

Specific skills and resources are needed for analysis and visualization if big data are going to be effectively incorporated and sustained in public health decision making. A number of barriers to effective use and visualization of the data have been identified related to: 1) the limitations of current visualization tools for displaying complex interrelated data streams that are constantly changing over time, 2) human resource limitations and training needs, and 3) information technology and software access challenges within many publicly funded agencies.

A number of other cautions were also identified regarding the application of big data in infectious disease surveillance and outbreak investigation. The first and one of the most important is the risk of false positives associated not only with the large volume of data points, but also the very large numbers of variables that might be examined for potential associations, particularly with the increasing use of unsupervised or partially

supervised machine learning tools. The second limitation, as noted above, is the need for new data management skills and in some cases an entirely new vocabulary in order to apply or at least critically evaluate big data products. The third limitation relates to ongoing issues in public health with data security, governance and privacy which become even more challenging with the need for accessing data from different sites, cloud computing, emerging data mining tools and the potential for new ways of linking de-identified data. Finally, there is the critical issue of data quality. Much of the information used is being repurposed and was originally collected for very different reasons limiting the accuracy, precision, or completeness for addressing health-related questions. Many big data sources also rely on self-reported information which is limited with respect to the accuracy of descriptions and interpretations, attribution of cause, and time and place of exposure.

Big data provides many opportunities to increase the effectiveness of disease surveillance and outbreak investigation, but it is a supplement and not a replacement for most current methods. The public health professionals who responded to our survey indicated support for the application of big data to foodborne disease outbreak investigation. The examples of success stories provided by the respondents to a great extent mirrored those identified in the published literature. However, the respondents also echoed the previously identified challenges regarding the necessity for adequate resources and training; barriers related to data management, storage, and visualization; and limitations related to privacy concerns and legal considerations.

Contents

PREAMBLE 1

OBJECTIVES AND REPORT STRATEGY 3

AN INTRODUCTION TO BIG DATA 5

POTENTIAL SOURCES OF BIG DATA FOR INFECTIOUS DISEASES 6

i. Whole Genome Sequencing / Bioinformatics 6

Options for managing WGS data and tools for analysis 7

Challenges in the analysis and interpretation of WGS data 8

Other challenges and limitations in the application of WGS 10

ii. Aggregated News Reports 11

iii. Internet Queries, Search Histories and Forum Data 13

Health seeking behavior and use of search engines or web page access history 14

Information shared on internet forums 15

iv. Social Media 18

Facebook 18

Twitter 18

v. Smartphones 21

Passive data from smartphone use 21

Smartphones as tools for active surveillance 22

vi. Pharmacy-based Surveillance 23

vii. Retail-based Surveillance – Food Sales 24

viii. Mining High Volume Electronic Health Records 25

ix. Participatory/Crowd-sourced Surveillance 27

x. Crowd-sourced Health Records 28

xi. Health Call-in Phone Lines 28

xii. Absenteeism Records 29

xiii. Earth Observation Data 29

BIG DATA FOR INFECTIOUS DISEASE MODELING 31

i. Models as Tools for Understanding Big Data 31

ii. Managing Data Input and Output from Simulation Models – Another Type of Big Data 33

VISUALIZATION OF BIG DATA FOR INFECTIOUS DISEASES 34
i. Increasing Effectiveness of Visualization and Recognizing Barriers 35
ii. Areas of Focus for Visualizing Big Data 36
iii. Specific Visualization Challenges for Big Data 37
CAUTIONS FOR THE APPLICATION OF BIG DATA TO INFECTIOUS DISEASES SURVEILLANCE AND OUTBREAK INVESTIGATIONS 38
i. Data Analysis and Risk of False Positives 38
ii. Need for New Data Management Skills and Analytical Vocabulary 39
iii. Data Security, Governance and Privacy Issues 40
iv. Data Limitations and Potential Biases 41
SUMMARY OF HOW THESE DATA SOURCES MIGHT BE EXPLORED TO PREDICT AND MITIGATE FOODBORNE DISEASE OUTBREAKS IN CANADA 43
RESEARCHER COMMENTARY – DR. NATHANIEL OSGOOD 44
SUMMARY OF COMMENTARY FROM PUBLIC HEALTH PRACTITIONERS – PATRICK SEITZINGER 46
i. Perceptions of Big Data in the Context of Foodborne Disease Outbreak Investigation 46
ii. Examples of How Big Data is Currently Being Applied in Public Health Practice and Research 47
iii. Sources of Big Data that are Currently Available to Assist in the Investigation of Foodborne Disease Outbreaks in Canada 47
iv. Sources of Big Data that would Enhance the Future Foodborne Outbreak Investigations 48
CONCLUSIONS 49
REFERENCES: 50

PREAMBLE

In September 2015, the *Canadian Communicable Disease Report* devoted a special issue to big data and some examples of how it is "transform(ing) how we identify, track and control infectious diseases" in Canada. Similarly, *The Journal of Infectious Diseases* also released a special issue in 2016 summarizing recent advances in the application of big data to a number of areas including disease surveillance, infection transmission models, and tracking attitudes and movements. Interest in this area is growing rapidly with the number of publications examining big data and infectious disease having increased exponentially since 2001 (1).

There are several factors motivating increasing interest in using big data to enhance surveillance for infectious disease and investigation of disease outbreaks including the potential to:

1. Improve timeliness, geographic resolution, and completeness of information
2. Address coverage gaps in existing surveillance programs
3. Increase sensitivity of surveillance for emerging and re-emerging diseases
4. Provide evidence to better inform predictive models and disease management

Many surveillance systems are hindered by the time lag between the disease event and reporting, as well as by limitations in the spatial resolution of the data (1). Decreasing the interval from onset of illness to reporting surveillance metrics, which are summarized on an appropriate geographic scale, would hasten the identification of both local and national outbreaks, as well as inform more timely and effective disease management. Time to detection can be an important determinant of outbreak severity. The total duration of foodborne disease outbreaks has, for example, been associated with the number of days from first reported symptoms to when an investigation was initiated in a study of regional and national outbreaks involving the BC Centres for Disease Control (2).

Big data available from a number of different technologies and platforms including smartphones, social media and the internet also offer the potential to supplement existing surveillance (3). Crowd-sourced data can provide individual level and near to real-time information (1). While the resulting case counts are not directly verifiable, passive digital footprints and information supplied by active volunteers can be used to augment physician and laboratory surveillance. Self-reported data has the potential to capture illnesses that would have been considered insufficiently serious by those affected to justify seeking physician care and

therefore would be missed by existing surveillance systems (1, 3). While the cases missed by traditional monitoring systems may be less clinically severe, they can be very important to understanding disease transmission for predictive modeling and control efforts, as well as to measuring the total burden of disease on lost productivity.

The second important motivation for exploring big data is the need to access additional information from areas not covered by current surveillance initiatives and from other "hidden populations" (1). For example, the Canadian FoodNet surveillance system for enhanced sampling of human pathogens and active sampling of retail sources focuses on three sentinel sites (http://www.phac-aspc.gc.ca/foodnetcanada/necessity-importance-eng.php): 1) the Middlesex-London Health Unit in Ontario, 2) the Fraser Health Region in British Columbia, and 3) Alberta Health Services Calgary and Central Zones in Alberta. While this program provides unique and very valuable information, efficient and effective options are also needed for enhancing surveillance and better understanding risk factors for enteric illness from regions not directly included in the sentinel program, and in populations with unique challenges and information needs, such as remote Indigenous communities and newcomers to Canada. The internet and social media has provided opportunities to gain information from "hard-to-reach" populations and to obtain insight into the experiences, risk factors, contacts and movements at a level of detail and breadth that was not previously possible (4).

Alternative and emerging data collection options might also be leveraged to improve monitoring and control of infectious disease challenges in demographic groups who are less likely to participate in traditional phone surveys, including those without land-line phones (5), (6). The Canadian Radio-television and Telecommunications commission in 2015 reported for the first time than more Canadian households rely exclusively to mobile services (20.4%) than on landline phones (14.4%), with more households having at least one mobile phone (84.9%) than a landline (78.9%) (http://www.crtc.gc.ca/eng/publications/reports/policymonitoring/2015/cmr2.htm).

The third motivation for increasing interest in big data is the concern that traditional surveillance systems are being challenged by emerging and re-emerging diseases such as MERS, SARS, Zika, Ebola, TB, and AMR among others (1). To address the need to detect emerging diseases, including those that might be associated with bioterrorism, significant resources have been invested in developing syndromic surveillance systems by a number of jurisdictions in North America and beyond. Syndromic surveillance "uses clinical features that are discernable before diagnosis is confirmed or activities prompted by the onset of symptoms as an alert of changes in disease activity" (7). Syndromic surveillance was thought to be particularly well suited to

detection of emerging diseases because of the focus on detection and reporting of symptoms which potentially increases the sensitivity of the system to include previously unrecognized diseases. Many of the tools considered under the heading of big data can provide crowd-sourcing of information on symptoms in real time, prior to clinical diagnosis, and for specific geographic regions.

Further to the potential for increased opportunities for new information sources through big data platforms, there is a recognition that big data analytics and learning inference engines can provide important advantages when compared with the statistical processes traditionally used to detect abnormal activity as part of most existing syndromic surveillance systems (8). For example, Bayes-like model building approaches include algorithms that will estimate the probability that a disease event is actually occurring given both the observed data and the prior probability of disease in the current epidemiological context.

Finally, big data could have a substantial impact on the effectiveness of predictive models for public health. Predictive models are vital tools in understanding disease transmission and planning for disease management. However, the application of big data to dynamic modeling in infectious disease is still very much in its infancy when compared to other areas such as marketing and weather forecasting (1). One of the biggest differences between prediction models for infectious disease versus meteorological events is the quality and quantity of the information available to parameterize the models (9). Infectious disease models are particularly sensitive to contact rate parameters and heterogeneity in contact within and between groups. Whereas, historically most models have been limited by assuming random mixing within groups and focussing only on mixing between larger population centres (10), there are new big data tools that can provide detailed movement information to infectious disease models and capture heterogeneity in contact within populations. This precise movement data will also allow measurement of varying contact rates between infected and susceptible individuals in response to media attention and the resulting changes in public concern (10).

OBJECTIVES AND REPORT STRATEGY

Recognizing the potential of big data as "an untapped resource of evidence that may be used to inform policy and decision making", the National Collaborating Centre for Infectious Disease commissioned a background paper to initiate discussion among Canadian public health professionals. The primary objective was to

examine the potential for big data to inform public health policy for infectious disease management in Canada. The first part of the report introduces big data and the various sources of big data for infectious disease surveillance as described in the peer-reviewed literature and publically available technical reports to December 31, 2016. Most of the papers examined have a North American focus. Examples with a Canadian context were included and highlighted where possible. The report also explores options and challenges in visualizing and analysing the resulting information so that it can be more effectively used in decision making.

While many of the publications to date on the use of big data in infectious disease deal with influenza, Ebola, and vector-borne diseases, the potential use of big data to aid in foodborne disease surveillance and outbreak investigations was identified as a potential focal point for future discussions. This report highlights the types of big data applied to the surveillance and investigation of foodborne illness as described in the literature, provides commentary from a researcher who is actively working on the application of big data to foodborne disease, and finally summarizes the findings of an informal big data survey of public health professionals working in outbreak investigation in Canada.

This report is presented as an introduction and basic overview. While every attempt was made to be systematic in searching the existing literature, this was not intended as an exhaustive exploration or summary of all facets of big data and infectious disease or big data and foodborne illness. Neither does the report intend to provide specific recommendations on next steps for adoption of big data in outbreak investigations. Any opinions expressed are those of the authors unless otherwise indicated.

The papers summarized in this report were obtained following a systematic search of PubMed, Scopus and ProQuest Public Health for the terms “big data” and (“infectious disease” or “outbreak investigation” with synonyms). The database search was followed by sequentially screening titles, abstracts and finally the full reports for relevant content. The first 200 hits on Google were also examined using the same search terms to ensure the inclusion of more recent and potentially unpublished work as well as relevant grey literature. These searches were followed by a review of the reference lists from key papers to identify important information that had been missed. We also used options in the PubMed, Scopus and ProQuest databases to identify more recent papers which had subsequently cited those selected in the initial search or that were considered similar to those identified. Supplemental searches examining “big data” and “visualization” and “disease” as well as “big data” and “foodborne illness” and synonyms were also completed.

An Introduction to Big Data

Big data have been used very successfully in business for marketing specific product suggestions to individual customers of companies such as Amazon and Netflix, and directing advertising and ranking search results for Google. Big data have also been used in a number of areas of science, for example to increase the accuracy of short and moderate term meteorological forecasts. Health care has been slower to take full advantage of the ever increasing amount of information produced by new technologies. The most notable exception for infectious disease management has been the application of whole genome sequencing and bioinformatics in virology and microbiology. Public health professionals are being challenged to more effectively incorporate big data into disease surveillance and outbreak investigations with a goal of using the increasingly available evidence to inform decisions and improve efficiency and effectiveness of public health programs. However, because epidemiologists and biostatisticians often already work with large data sets, many public health practitioners might have questions about how big data are different than the information sources that have traditionally been central to disease surveillance and investigation.

While there are a number of factors that characterize big data, the three features of big data which are most commonly reported include: data volume, velocity and variety (1) (11). Volume refers to the actual size of the data recognizing that the absolute size designating big data might vary substantially among disciplines (1). Links reported that data could be considered "big" when the size of the data becomes so substantial that it interferes with translation of the data to actionable information with traditional epidemiological and statistical software tools (11). Volume might reflect a large number of subjects or observations per subject or both. Where large databases in epidemiology have recently been measured in gigabytes (10^9 bytes), there are many types of big data that are typically reported in terabytes (10^{12} bytes), petabytes (10^{15} bytes), or even in some cases zettabytes (10^{21} bytes or 1 trillion gigabytes) and larger. Databases in some of these classes exceed the storage capacity of even the largest hard drives for personal computers.

In addition to volume, big data are also characterized by velocity and variety. Velocity can refer to the speed with which the data are collected or transmitted and received. Many types of big data are collected in near real time and big data might be streamed rather than batched before periodic transfer. Finally, big data might have a variety of both sources and organizational structures (12). For example, big data might contain structured numeric data, unstructured text, PDF files, emails, voice recordings, images, and data from mobile apps (13). Because of the size, rate of growth, and complexity, big data can require substantial effort and

computing resources to clean, merge, link, match, transform, analyze, mine and store (13). Other qualifiers for big data include: veracity (a reflection of the imperfectness, incompleteness, or unreliability of the data), validity (data correctness), volatility (at what point is the data obsolete), and variability (issues with inconsistency and unpredictability of data flows) (13), (14).

An overview of the potential sources of big data provides insight on how each might contribute to public health and particularly the investigation of foodborne illness outbreaks. Cinnamon et al. describe three types of big data (15). Directed data results when technology capable of recording data (e.g. surveillance cameras, remote satellite sensing technology) is focused on an individual or location. Automated data are passively collected through the generation of electronic records resulting from the normal operation of systems or technologies such as mobile phones, web browsers, credit card transactions or consumer loyalty programs. Finally, volunteered data are actively or passively produced by citizens through platforms such as social media and crowdsourcing applications (15). Active data result from interactive, intentional contributions to digital infrastructure for participatory platforms and citizen science (1). Other classifications proposed for big health data include: medical encounter data, participatory syndromic data and non-health digital data such as contact networks, travel patterns, vaccine acceptance, and food choices (1).

Potential Sources of Big data for Infectious Diseases

i. Whole Genome Sequencing / Bioinformatics

The analysis of whole genome sequencing (WGS) data is one of the most widely recognized applications of big data in infectious disease surveillance and investigation. WGS can be used to include or exclude cases from specific outbreak investigations, or suspected sources to inform infection control measures and public health messaging (16). WGS with appropriate bioinformatic analyses can also be used to predict the existence of undiagnosed cases and intermediates in transmission chains, and to infer the directionality of transmission (16). For example, phylogenetic analysis of *Salmonella typhimurium* DT104 WGS data in human and animal populations in the UK demonstrated that there was limited transmission between species (17,16).

Bioinformatics involves the analysis of variations in biological systems at the molecular level (18). Phylogenetic analysis and similar evolutionary analytical methods can be used in some cases to make inferences about the origin and emergence of certain pathogens, estimate potential sources of disease and identify most probable transmission chains (16). Analysis of the 'molecular clock' can sometimes be used to estimate the time to the most recent common ancestor and potential dates of transmission events. Molecular clock analysis of genome sequence data is based on the assumption that nucleotide substitutions will accumulate at a constant rate. This strategy has been used to make inferences about transmission events in MRSA outbreaks (16).

WGS is increasingly being applied in foodborne disease surveillance and has been very successful in identifying the source and limiting the size of Listeriosis outbreaks; however, this organism has a relatively smaller and more conserved genome than those for Salmonella, *E. coli* and Campylobacter. While the interpretation of WGS data from larger organisms is more challenging, WGS has successfully differentiated Salmonella strains that could not be separated using pulse field gel electrophoresis (PGFE) (19). In addition to its application in outbreak investigations, WGS has also identified sources of sporadic cases of foodborne illness (20).

Options for managing WGS data and tools for analysis

Aarestrup et al. summarized current platforms for sharing WGS data, including GenomeTrakr and COMPARE (21). GenomeTrakr was developed with the US Food and Drug administration. This program includes sequencing and real-time comparison of foodborne bacterial isolates with the average number of sequences added per month increasing rapidly from 169 in 2013 to 4,529 in 2016 (www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS). The sequence data are assembled, analyzed and stored within the National Center for Biotechnology Information (NCBI). The publicly available tools on NCBI allow for the generation of basic phylogenetic trees, but more sophisticated analysis must be done locally. All 50 states are expected to be part of GenomeTrakr by 2019 (19).

COMPARE Europe is a single site for sharing and analyzing data from bacteria, parasites, and viruses that includes single genomes and metagenomic projects (21). The developers recognized the need for easy-to-use bioinformatics tools and the difficulty of downloading large volumes of data for local analysis. COMPARE will allow researchers to bring new tools to the data and it has been configured to manage secure data sharing

(www.compare-europe.eu). More advanced analytical tools are constantly under development including the potential to combine WGS data with spatiotemporal epidemiological data within single Bayesian inference models and to construct more accurate transmission trees. Jombart et al. have reported this type of analysis using data from the 2003 SARS outbreak in Singapore (22). The Centre for Genomic Epidemiology in Denmark have also developed a number of bioinformatics tools that can be used to identify species, multilocus sequence typing (MLST), plasmids, virulence, antimicrobial resistance (AMR), serotype, and phylogenetic tools (23).

IRIDA (Integrated Rapid Infectious Disease Analysis) (http://www.irida.ca/), a free open source platform, is available for managing genomic data in Canada (https://www.slideshare.net/IRIDA_community/irida-immemxi-hsiao). IRIDA uses Galaxy to manage its workflows. Participants include the Public Health Agency of Canada (PHAC), provincial public health agencies and some universities. IRIDA contains links to other tools that facilitate identification of genomic islands which may encode AMR and virulence factors (IslandViewer) and mapping (GenGIS). An overview of phylogeography with GenGIS, which was developed at Dalhousie University, can be found at https://www.slideshare.net/beiko/gengis-presentation-at-vizbi-2016. Additional tools are needed to best identify clusters based on WGS and the epidemiologic metadata (source, where, when). EpiQuant is an example of one tool developed by PHAC to enhance application of genomic data in epidemiology (https://www.slideshare.net/IRIDA_community/hetman-immem-xi-final-march-2016).

Challenges in the analysis and interpretation of WGS data

In April 2016, there was a joint 2-day meeting to discuss the application of WGS to protect public health and enhance food safety in the US that included industry, regulatory agencies and researchers (www.uspoultry.org/foodsafety/docs/WGS_Meeting_Summary_072916-02.pdf). This meeting provided an overview of how WGS is currently being used and the limitations of this technology. Neither WGS or any individual technique alone was considered sufficient to establish the cause of an outbreak; rather many of the participants emphasized that the investigation should include a combination of epidemiology, trace-back data, environmental assessments, and microbiology. Participants also warned that while WGS has the potential to provide information currently obtained through other methods, such as speciation, serotyping, the

presence of virulence factors and resistance determinants, the identification of virulence and AMR genes will not be identical to phenotypic virulence or AMR in all species or environments (20).

Similarity between organism genomes is measured using SNP (single nucleotide polymorphism) analysis and wgMLST (whole genome multilocus sequence typing) (20). However, there are issues with the standardization and reporting of these techniques. For example, there are several options for counting SNPs. For the count of SNP differences to be interpreted correctly, details on how the SNPs were counted are required (20). Relatedness is often displayed using a phylogenetic tree. However, inferences regarding paths of relatedness between organisms can be based on a number of underlying methods (parsimony, maximum likelihood, Bayesian and distance), and the confidence in the resulting trees should be reported using appropriate statistics (20). Parsimony based algorithms identify the tree that requires the fewest SNPs (23). Maximum likelihood based methods return the tree that is most likely to have resulted in the observed data. Bayesian methods identify the tree with the highest posterior probability given the data that have been observed. Finally, distance based methods report the tree best representing the SNP differences or whose branch lengths are minimized (24). Different methods will often lead to differences in interpretation.

The details of how to make meaningful distinctions between isolates are species dependent (20). The evaluation of differences can be affected by evolutionary changes within the organism over time (SNPs predicted every >2000 generations). SNP changes can be influenced by passage through different hosts or lab media, genome size and degree of conservation within the genome, and whether or not pieces other than the core genome are considered (i.e. plasmids, phage material) (20).

For WGS to be consistently useful, investigators need to be able to rapidly query WGS databases and accurately determine the genetic distances between samples. Different measures of distance based on k-mer profiles or nucleotide sites have been evaluated for Salmonella isolates (24). The authors of the Salmonella study concluded that site based distances were superior (NUCmer and extended MLST), but the computer resources required could be a limiting factor. However, questions remain regarding how different isolates of different species should be evaluated to be considered different strains.

Finally, there is a need for research regarding how likely it is that the same strain can be found in multiple environments and how that likelihood varies by species. For example, if we find the same strain of an organism in a sample from a suspected source and a diagnostic sample, there is very limited data on the

degree of probability with which investigators can rule out other sources. It is possible that genetically identical isolates can be found in different locations (20).

Other challenges and limitations in the application of WGS

There is a need for standardization of all aspects of WGS methodology and bioinformatics tools for public health practice (23). Serious concerns have been raised by bioinformatics researchers about the need for transparency in reporting the methodology for data processing and assembling WGS (19). Many of the pipelines used to translate raw data to biological interpretation are not publically available (e.g. NCBI). The same researchers highlighted the need for assessing the potential for false positives when assessing relatedness, and the need to report confidence in reported SNPs (19). In addition to false positive risks due to factors affecting the performance of analytical tools, there is also a potential for false positives due to sequencing machine contamination (20).

Another identified limitation in WGS included data quality challenges with the more commonly used short-read platforms which are cheaper and faster as compared to the long-read technologies (19). Long-read platforms are more likely to result in closed genome sequences without gaps in the genome. Long-read technologies result in data that can be assembled de novo without needing a reference strain and can be used to detect genome rearrangements and to directly calculate PGFE patterns (20). The quality of SNP identification and genome assemblies will therefore vary with original processing method. Quality of SNP identification also varies with whether the method of assembly was de nova or was based on alignment to a reference. SNPs are identified following alignment to a reference strain, and the quality of any inferences then depend on the quality of the reference strain (20). A further caution was raised that some old reference sequences based on Sanger method can contain assembly errors (25).

WGS typically still requires culture of the suspect organisms, but the demand for culture-independent diagnosis is increasing (23). One of the options currently being considered involves direct characterization of all of the genetic materials from the specimen or a metagenomics approach. There are several challenges with attempting to extract information on specific pathogens from the total genetic materials. Metagenomics require comprehensive reference databases to interpret the resulting data. There are also ethical issues with managing the information collected on the patient's DNA that might be present in diagnostic samples (23).

As WGS data in GenomeTrakr is a public database there are concerns with confidentiality of the sample source. In highly publicized outbreak situations, while the source might be confidential it might be easily predictable by knowing date, sample type, organism, and location (19). PulseNet by comparison is not publicly accessible. Data sharing issues related to rapid placement of WGS data in public databases are further complicated by practical issues, political sensitivities, legal, ethical issues and concerns, intellectual property rights (IPR), and desire to protect publication rights (21). Options have been proposed to tag data for short periods so it can be used for public health, but not for use by other researchers for publications. There has also been a lot of discussion regarding the metadata that should be associated with the WGS information (21). Suggestions for the minimum data set have included country, year, origin, and pathogenicity or whether the sample results from an infection (21).

The potential risks to the confidentiality of patients (clients, participants) need to be balanced relative to the benefits to public good. The traditional approach has been to remove what is perceived to be identifying information. This can severely limit the usefulness of the data and does not always guarantee confidentiality given the capacity to link data across different sources. Differential privacy has been proposed as a process-driven approach to protecting patient privacy (26). For example, sociodemographic data might be slightly perturbed prior to release to reduce the risk to privacy to a predefined threshold. Others have suggested a tiered approach to sharing genomics and associated meta-data to optimize the utility of this information while protecting confidentiality (27).

ii. Aggregated News Reports

Canada is recognized internationally for the Global Public Health Intelligence Network (GPHIN) (28). GPHIN is a web-based program that scans and mines more than 30,000 global news sources in nine different languages. More than 20,000 reports are evaluated every day for evidence of public health threats. The program makes use of news aggregators linking to national and local newspapers as well as select newsletters. The stories examined are not just limited to health news, but include sports, travel and finance. Scans are repeated every 15 minutes and stories are translated and processed in less than a minute. When the system detects a signal it is manually reviewed by a multilingual and multidisciplinary team of health professionals who trigger an alert if there is sufficient evidence. The GPHIN system is acknowledged for

being the first to detect the outbreak MERS-CoV and also detected early SARS activity in China due to mentions in financial pages of increased sales of antiviral drugs (28).

One of the oldest moderated event-based monitoring systems is ProMED-mail (Program for Monitoring Emerging Diseases). ProMED-mail was started in 1994 [29] and is affiliated with the International Society for Infectious Diseases. ProMED obtains information from media reports, official government reports, online summaries, and local observers (www.promedmail.org/). This information is reviewed and investigated by expert reviewers before it is distributed by email and posted to the website. ProMED is available in Portuguese, Spanish, Russian, and French in addition to English. ProMED collaborates with HealthMap.

HealthMap is hosted at Harvard University and is based on data from ProMED-mail in addition to information from the World Health Organization, GeoSentinel, World Organization for Animal Health (OIE), Food and Agriculture Organization of the United Nations (FAO), EuroSurveillance, Google News, Moreover, Wildlife Data Integration Network, Baidu News, and SOSO Info (http://www.healthmap.org). The software relies on open source programs such as Google Maps, GoogleMapAPI for PHP, Google Translate API, and the PHP AJAX library. HealthMap monitors, summarizes, filters, and maps information collected in nine languages 24 hours a day. In one study of 111 outbreaks in six months during 2012, HealthMap reported outbreaks an average of 1.26 days ($p=0.002$) before the first formal source report. In most of the outbreaks, the information reported was the same for formal and HealthMap reports (30).

In addition to using media reports as inputs for established disease surveillance systems such as GPHIN, ProMED-mail, and HealthMap, custom queries have also been developed to extract information to address specific questions. Using media reports surrounding the 2014-15 Ebola epidemic and 2015 MERS outbreak, Chowell et al. were able to extract information from reports describing epidemiologic clusters and estimate reproductive numbers and time trends for the outbreaks that were very similar to those calculated using official surveillance data. The authors recognized that reporting of case clusters in the media was subject to several potential types of bias (31).

Recognized limitations of the data from the media reports included the potential for stories to focus on sensational survivor stories, stories with American focus, and large case clusters (31). The researchers also noted that age and sex was missing from many reports. Given the potential limitations of information reported in the media, new sources of information for platforms such as GPHIN have been considered

including internet search strategies, social media, smartphones, and other volunteer-based active crowd-sourcing systems (28).

Similar types of web-based platforms to those used for media reports can also be used to aggregate specific data reported by laboratories and health care institutions. ResistanceOpen is a web-based and mobile app for aggregating, analyzing and reporting antimicrobial resistance data (www.resistanceopen.org) (31). The publicly available data are identified using web-based queries and by revisiting URLs which had previously reported appropriate data. The data, however, then have to be manually abstracted once identified using trained curators. At the time MacFadden et al. described this system there were data from 340 locations from 41 countries including eight Canadian provinces. The outputs of the system include a navigable map allowing comparisons among regions. The authors, however, do caution that comparisons are limited by differences in reporting practices among sites and differences in the standards used to classify resistance status (31).

Most of the described search and reporting processes for media and laboratory data currently require substantial person-hours of review to assess whether or not the events and data recovered by the search engines are suitable for adding to the online repository. In the cases of ResistanceOpen manual data re-entry is also required. One of the areas identified by GPHIN for future development has been the potential for more advanced artificial intelligence capabilities to decrease the person-hours needed to review the outputs before identifying an event (28). While many new initiatives originate in academia, given the current personnel requirements for daily operation in addition to the need for ongoing updates to search algorithms and system maintenance, academia may not an ideal long-term home for these types of activities. It is difficult for researchers to justify long term investment after the original research project has been completed (33).

iii. Internet Queries, Search Histories and Forum Data

Individuals who use the internet leave traces of their activity. The traces can be classified as either health-seeking behaviors, which include web-searches for information on a particular topic, or information the user has intentionally shared on a blog or web-based forum (34). In addition to data from web-based forums dealing specifically with health issues, other nonspecific forums might also contain useful information. For

example, an outbreak of *Campybacter jejuni* following a bike race was linked to mud on participants through messages and photos on a web-based social forum (35).

Health seeking behavior and use of search engines or web page access history

In one report, search engine data was used to predict Norovirus season in Sweden (36) based on queries directed to the official government health portal. The search engine described in the study was Websök, a system developed to analyze public search queries to the official health portal for Stockholm county in Sweden. However, it is unclear as to what proportion of the population would have been sufficiently aware of the official website to have used the portal to conduct their searches. While there is some appeal in summarizing search data for very specific websites managed by government health departments, there has been considerable interest in looking at data from more widely accessed internet tools.

One of the most commonly recognized applications of big data to infectious disease surveillance was Google Flu Trends which monitored internet search patterns based on the idea that many of those using the search engine might be researching symptoms. The initial results for Google Flu Trends were promising, but in February 2013 it predicted twice as many flu-related visits to doctors' offices than were reported by the Centre for Disease Control's (CDC) sentinel clinics and hospitals (37,38). While correlation of Google Flu Trends with laboratory confirmed cases has been consistently lower than for ILI (37), the forecasts substantially improved when combined with historical traditional surveillance data (37). Based on questions surrounding the tool's performance, Google is no longer publishing updated estimates of Google Flu Trends or its counterpart Google Dengue Trends (www.google.org/flutrends/about/).

One of the reported issues was that although the search algorithm appeared to be regularly updated, it did not appear to take into account the Google business model for ranking recommended searches for its users based on the search behavior of other users (38). The ongoing changes in internet service platform would also limit the reproducibility of results over time (38). One of the other major limitations of the web-based query models was the limited ability to account for age specific differences in epidemiology [39]. Web searches, however, may capture location through IP address (12).

The overestimation of the peak in 2012-13 by Google Flu Trends reported in the US was not observed in Canada (40). Google Flu Trends estimates for Canada correlated well with ILI consultation rates within the

sentinel physician system reported to PHAC and influenza A and rhinovirus positivity from the Respiratory Virus Detection Surveillance System in all seasons from 2010 to 2014 (40). Google search data had also previously been shown to predict gastroenteritis and Norovirus outbreaks (41,42), as well as rotavirus outbreaks in young children (42).

In addition to using data from the search engines themselves, specific web pages might also be queried to look at usage statistics. Wikipedia pages are often ranked highly by search engines and it is a commonly used, first source of information from the web for many in North America. Wikipedia releases hourly search history data to interested parties at a number of sites including http://stats.grok.se/ (34). The data are summarized as article views, page views, and page count files depending on the source. In one example, Wikipedia access logs were used to accurately forecast and nowcast both influenza and dengue fever (34). However, the location information in that analysis was limited in accuracy as the country of origin information was inferred based only on the search language.

Information shared on internet forums

There are several examples in the literature where online restaurant review forums have been used to estimate the likelihood of poor inspection scores, estimate the risk of foodborne illnesses or supplement information needed for outbreak investigations. One of the most commonly reported restaurant review sites is Yelp (www.yelp.ca).

Kang et al. compared text analysis from Yelp reviews in Seattle to inspection reports. The authors reported 82% accuracy in discriminating severe offenders from places with no violations. In addition to predictive cues in the text of the reviews, researchers examined the number of reviews, review length, average rating, count of negative reviews and evidence of fake reviews. They concluded the best information was found in the text content of the reviews (43).

Harrison et al. used online Yelp reviews by restaurant patrons to identify unreported cases of foodborne illness in New York City during 2012-13. One of the important findings from this study was that only 3% of reviews describing an illness were reported to the city 311 service (44). The focus of this study was on identifying infectious outbreaks. Illnesses with a very short incubation period were excluded by scoring criteria. For example, this study would not have captured most illnesses resulting from exposure to

Staphylococcus aureus toxins in food. The analysis identified three previously unreported restaurant-linked outbreaks in a 9-month period (44). Multiple violations were detected during the inspections of all three restaurants.

When the New York project was initiated, there was concern from the health department staff about the time that would be needed to investigate the resulting cases (44). Staff were required for programming expertise, to read reviews, send emails, interview reviewers, and complete follow up inspections. In this project, staff conducted manual review of higher criteria scores. The Yelp data is available publically on the website, but Yelp provided researchers with xml format data specific for New York City to make the process more efficient. Cases were defined based on evidence in the review of the words "sick" or "vomit" or "diarrhea" or "food poisoning" with two or more persons ill and incubation time ≥10 hours (44). Researchers then created accounts so they could send private messages in Yelp to the identified cases. However, there was no guarantee the original reviewer would see the message as the system required the case to log into Yelp to view it. The need for the original reviewer to log in again increased lag to follow up and may have decreased the response rate (44).

This study concluded that analysis of restaurant reviews has the potential to identify small point source outbreaks not captured by traditional surveillance tools (44). To further improve the sensitivity of the system, the researchers proposed asking Yelp to include link to local health departments reporting website, adding data from other review websites, and increasing from weekly to daily updates (44).

There is additional evidence that reports of foodborne illness in restaurant reviews do capture foods implicated in official foodborne outbreak reports. Nsoesie et al. looked at reports of illness and foods implicated in Yelp reviews for 5824 restaurants from 29 cities (45). The distribution of implicated foods on Yelp was very similar to that identified in CDC data. However, there were important limitations in the data. Only 17% of reviews reporting an illness listed the actual date of the illness. There was also the potential for misattribution of the source of infection by the person reporting due to the lack of public recognition of differing incubation periods among various types of pathogens (45).

The overall goal of examining the online reviews is to aid traditional surveillance systems in near to real-time reporting of foodborne disease and to improve time to recognition of outbreaks of foodborne illness. For example, a recent study from British Columbia reported that the median time to initiate an outbreak investigation was 36 days (2). The Yelp reviews did capture restaurant location and analysis of the reviews

identified cases that would not be captured by routine surveillance. Only 1.6% of those reporting an illness visited their doctor, and 11% of illness reports involved more than one person (45). While the emphasis on online restaurant reviews as a tool to help mitigate foodborne illnesses does not directly take meals prepared at home into account, part of the justification for this effort comes from the finding that approximately 44% of outbreaks contained in the CDC Food dataset were suspected or confirmed to be associated with restaurants (45).

In a later study, Schomberg et al. developed a model based on 71,360 Yelp reviews of San Francisco restaurants that, in the pilot phase, predicted health code violations in 78% of the 440 restaurants receiving serious health code violations (46). This model was different from the work of Kang et al. in that the intent was to preferentially identify health code violations that increased the transmission risk for foodborne disease (43). When this same model was later applied to another 1542 San Francisco restaurants, the model had a sensitivity of 91%, specificity of 74%, area under the curve of 98% and positive predictive value of 29% (prevalence 10%) for detected serious health code violations. The same model applied to New York City data achieved an AUC of 77% and a PPV of 25% (prevalence 12%). The drop in accuracy could potentially have been due to geographic differences in language used in the restaurant reviews. Model accuracy was best when the highest ranked Yelp reviews were used. Yelp "stars" were also significantly associated with health code ratings. Not surprisingly, the word "vomiting" was the most strongly associated with low health code rating.

While the models provided some useful information, other studies have cast doubt on whether Yelp ratings provide comparable information in all types of retail food facilities. For example in another recent report, Yelp ratings from New York appeared to be correlated with sanitation in chain establishments, but not for establishments that were not part of a franchise chain (47).

Schomberg et al. commented that the use of online restaurant reviews would be expected to work better in large centers where participation in online reviewing platforms would be higher (46). The authors also recognized the potential for enteric illness symptoms from other causes to be misattributed as foodborne illness. Finally, the authors raised the question of whether reviewing customers might be able to detect some issues that the public health inspectors were missing because many customers will visit the restaurant more than the once or twice per year scheduled inspections. The authors also made the point that reviewers actually consumed the products produced by the restaurants as compared to the inspectors who made observations of the facility and processes (46).

iv. Social Media

Social media provides another valuable opportunity to leverage the "collective intelligence" of the public to enhance early detection and control of infectious disease (35). Most reported studies of social media platforms rely on mining of unstructured text for passive surveillance data reflecting symptoms or public sentiments towards vaccines or food safety (48). In a recent systematic review of the use of social media for surveillance and outbreak management, the two most commonly reported platforms were Twitter and Facebook (48). To effectively use social media for surveillance, it would be important to understand the platform preferences of the demographic of interest for a particular objective. The authors of this same systematic review noted that in the articles identified to February 2013 younger individuals were most likely to use Facebook, while Twitter use was more frequently reported by adults (48).

Facebook

Studies using Facebook were less common than those using data from Twitter (48). Most studies identified using Facebook focussed on risk behaviors for chronic disease. Facebook 'likes' were reported to predict many health outcomes and behaviors with similar accuracy to data from the Behavioral Risk Factor Surveillance System – an ongoing random digit dialing telephone survey (49). Given that Facebook 'likes' might be less subject to fluctuate over short periods than sentiments captured in tweets, it is not surprising that Facebook might be better suited for investigating more stable attitudes and behaviors that could be linked with chronic disease or ongoing exposures. Gittleman et al. further questioned the transparency regarding how categories of 'likes' were determined by Facebook for extraction and sharing, and to what extent that might impact the results of data analysis (49). There were no specific studies identified regarding whether Facebook 'likes' correlated with risk of exposure to foodborne disease.

Twitter

Twitter is an online social media platform that allows registered users to post short microblogs and to respond to messages or 'tweets' posted by others. Anyone can read these messages, regardless of whether or not they are registered on the Twitter service. Twitter data can be accessed through an open application programming interface (API) allowing third parties to stream Twitter data in real time to their own

application (33). Twitter data is unique when compared to many other sources in that tweets are limited to 140 characters. Other potential limitations include biases towards younger more technologically oriented individuals as well as the recognition that location information is limited to where Twitter users choose to send tweets from (50). Location is only recorded when and if the person tweets. Twitter has been used to examine influenza in a number of different regions as well as outbreaks of cholera, *E. coli*, and Dengue fever (48).

Some Twitter-based studies focus on detecting specific disease activity. Others focus on detecting symptoms and then interpreting the symptoms with respect to disease activity (51). In addition to measuring the intensity of disease activity, outbreaks can be identified from Twitter data through detection of spatial clusters, social network analysis methods, analysing communication pattern activity, and identifying important keywords by their spatial signature (51).

Twitter has also been used to understand human mobility patterns (50) and as a proxy for estimating contact rates and changes in contact rates during a disease outbreak. Geotagged tweets offer higher resolution location data than that possible using call data records (CDRs) obtained from cell phone carriers (50). Resolution can be within 10 meters as compared to resolution of kilometers for most CDR records. The text content of tweets might be used to infer location for tweets that are not geotagged (50). For example, in one study the 'location' field was used from the Twitter use profile together with Google Maps API (52).

For Twitter data to be most useful for surveillance there is a need to be able to map the results and geographically normalize the results by linking the Tweets to the underlying population. To complete a geographic query, the Twitter database requires a latitude and longitude centroid and radius for a buffer to be included in the search. This buffer can be overlaid and spatially joined to a layer containing census tract data to provide information for the population at risk (53).

There is also a need for advanced machine learning algorithms that can be trained to identify true cases of the target illness (measured as the recall of the algorithm – or sensitivity in an epidemiological context) (e.g. get over this flu, flu medicine) as well as correctly classifying tweets that contain reference to the original search term but do not reflect illness in the individual (measured as the precision of the algorithm or specificity in an epidemiological context) (e.g. flu shot, stomach flu) (53). One example of this type of machine learning tool is a support vector machine (SVP). The language used in tweets is dynamic and can have distinct geographic anomalies requiring manual review of unexpected results and a process to continually update the

classification algorithms (53). Some of the principles of machine learning as they apply to the use of big data in health care have been reviewed by Flahault et al. and Dinov, and will be highlighted in a later section on cautions and limitations for use of big data in infection disease surveillance (9,54).

There is growing interest in using Twitter to identify foodborne illnesses. In a 2013 scoping review, only four primary research articles were identified that investigated social media and foodborne illness or gastroenteritis (35). For example, tweets were linked to Salmonella and Norovirus outbreaks in Germany (55). Since that time there have been other examples. In a case study, Twitter data and existing outbreak detection algorithms identified an outbreak of EHEC in Germany before MedISys (based on media data) and other early warning systems (51).

One criticism of the interest in using social media to monitor infectious disease has been that until recently there were no well documented examples of applications to daily practice by health departments. In 2013, FoodBorne Chicago (www.foodbornechicago.org/) was launched by the Chicago Department of Health and its partners. The program was designed to identify complaints about foodborne illness on Twitter by searching for "food poisoning" (56). The program identified 270 tweets with 3% reported to doctor or ER in the first 10 months. Staff responded to the tweets and provided links to the FoodBorne Chicago complaint form; there were 193 complaints reported to Foodborne Chicago with 10% seeking medical care. The complaints triggered 133 unannounced inspections which comprised 7% of all complaint based inspections during that period. The result of these inspections was the identification of 20% critical violations and another 22% serious violations. Investigators noted that many of these cases would not have been included in normal surveillance numbers nor prompted inspections by the health department. Inspectors were initially concerned about being overburdened with inspections but recognized the benefits once the program was underway. The open source software is available at GitHub (56).

As an extension of this type of work, a web-based dashboard (HealthMap Foodborne Dashboard developed at Boston Children's hospital) was developed and directed to identify tweets regarding foodborne illness globally. However, this dashboard can also be customized and focussed on a specific location. The experience of the St. Louis health department was described in one study (56). Tweets were captured with the words "foodpoisoning" or "food poisoning." Replies to relevant tweets were made through the dashboard and increased filed reports above that from existing mechanisms. Seven percent of replies to tweets using the dashboard resulted in reports. Replies from the investigators to those who had tweeted regarding food poisoning included an empathy statement, confirmation of authority, and call for the individual to formally

report their complaint. Restaurants inspected following these reports were no more likely to be in violation than those inspected following public complaint reports through other mechanisms. The researchers highlighted the opportunity for timely interaction with those reporting on the Twitter platform.

Another criticism of the application of social media to detecting illness was that there has been limited evidence from controlled research studies to show that these methods work in practice (48). In 2015 in Los Angeles, Sadelik et al. deployed nEmesis, an adaptive inspection system informed by machine learning algorithm applied to tweets, and then evaluated the system using a double blind trial (58). Each location flagged by a tweet was paired with control sites based on the annual inspection schedule; the controls were matched as closely as possible on location, size, cuisine, and permit type. In this trial, the adaptive inspections triggered by the nEmesis-detected tweets identified more demerit points per inspection, as well as more venues posing a significant health risk with a grade of C or worse, than comparable restaurants scheduled for routine inspections. One of the advantages highlighted in the report was the capacity of the system to identify issues in unlicensed operations that would not have been captured by routine inspections (58).

v. Smartphones

Passive data from smartphone use

One of the most commonly reported ways of using smartphones to enhance understanding of disease transmission is to leverage call data records (CDRs). These data have been used to understand spatial transmission of a number of infectious and vector-borne diseases. For example, CDRs were used to look at movements and risk of malaria transmission in Kenya (59). A tower and subscriber code are recorded for each call or SMS based text (10). These CDRs typically include the time a communication was made, a unique identifier for the caller, the receiver's telephone number, call duration, size of data transmitted, and the geographic location of the cellular tower the call was routed through and received from for every communication (call or SMS) made by mobile phone users (15). By linking the tower locations to a map it is possible to look at phone movement between calls.

CDR data is not typically publicly available. There is a very real need to assure anonymity of callers and obtain approval of regulators; access to data is provided only through negotiated agreements (10). In

addition, a one-to-one match of disease and movement data is not feasible due to privacy issues (10). Aggregation helps protect confidentiality in urban areas and aggregated data are still adequate for evaluating longer range movements.

There are also other limitations to these data. Spatial and temporal accuracy is related to tower density and calling behavior. Therefore, the resulting information is useful for larger scale and regional travel patterns, but is not as useful for movements relevant to very local disease transmission. CDR data does not capture movements finer than can be provided by the density of the cell tower grid in a particular region. The data quality will therefore vary based on tower density and consequently between urban and rural areas (10). Finally, there can be other errors associated phone sharing or where one individual has more than one phone or SIM card (15). For example, the person might have a SIM card in their phone as well their tablet or laptop.

In addition to analysis intended to describe movements between locations, social network analysis methods are also commonly applied to CDR data to look at connections between mobile phone users (60). Modeling of infectious disease spread requires an understanding of the contact network. Each individual has a certain number of contacts which can be described as the individual nodes' degree within a network. The heterogeneity of the degree distribution of the network can be an important determinant of disease dynamics. In one study, surveillance data from mobile devices was used to construct a realistic contact network and to look at changes in that network during the Ebola outbreak (60).

Smartphones as tools for active surveillance

Smartphones can also be used as simple self-monitoring tools. As an example, EbolaTracks, an SMS based platform, was developed for active self-monitoring of persons who visited an Ebola outbreak area. Participants were provided with a thermometer and mobile phone. They were asked twice a day for 21 days after visiting the area to text message to report symptoms and their temperature (15).

There are a large number of more sophisticated features on smartphones that have not yet been tapped to their full disease surveillance potential in research published to date. Mobile phones collect, store, and can transmit GPS coordinates that can then be recorded and mapped (12). Detailed GPS and WiFi data provide untapped opportunities for capturing fine scale individual movements (10). Further, Bluetooth sensors on the phones can also be used for tracking even finer indoor physical proximity to equipment with a Bluetooth

signal or other phones to generate fine grained contact networks. Mobile phone communication histories can also be used as a supplement for epidemic contact tracing (61).

Ethica, a smartphone based application developed at the University of Saskatchewan (www.ethicadata.com), has been used to acquire, store, and analyze data on human behaviour (62,63). The system can collect data from a wide variety of phone sensors including the GPS, WiFi, accelerometer, gyroscope, ambient temperature, and light, among others, in addition to providing user-triggered and context-specific survey options. A recent study assessed the feasibility of using the Ethica app for gathering data on the occurrence of enteric illness and risk factors for foodborne disease (64). By way of user-triggered and time-context prompted minisurveys, meal descriptions and PhotoFoodDiaries, the occurrence of enteric illness and food consumption behavior in 96 university student volunteers was collected during a 10-week period. Real-time food histories collected using the app features were compared with data from subsequent retrospective web-based questionnaires. Ethica served as an effective tool for collecting data on enteric symptoms and food consumption behavior in a sentinel cohort of volunteer participants. While the app had previously been used to collect food history data (62,63), this was the first described use for self-reported symptoms of enteric illness.

vi. Pharmacy-based Surveillance

Both prescription and non-prescription pharmaceutical retail sales can be an important source of syndromic surveillance data. Studies to date for infectious disease have primarily focussed on influenza, respiratory disease and enteric disease.

Retail sales data for non-prescription drugs in Great Britain have, for example, been used for syndromic surveillance to detect spatial temporal patterns in influenza activity and to determine if changes in purchasing behavior were related to public health messaging or intensity of media attention and public concern (65). Products monitored included adult cold and flu remedies, children's cold and flu remedies, cough remedies, thermometers, anti-viral products (including hand gel and wipes), and tissues. Total weekly sales and sales of milk and bananas were used to adjust for other reasons for changes in overall volume such as market share or store hours. Sales of anti-viral products followed by sales of children's remedies were most closely

correlated with case numbers. Retail sales were not associated with media reporting or the frequency of internet searches.

The Public Health Agency of Canada (PHAC) reported a study of antiviral prescriptions in Ontario and over-the-counter products in Nova Scotia (66). For severe respiratory disease surveillance, the influenza-like illness data were available to PHAC approximately 10 days after the onset of symptoms and laboratory data were available 17 days after. However, the over-the-counter sales data were available to PHAC 48 hours after the transaction was completed and antiviral prescription data were available approximately 5 days after the onset of symptoms. Seasonal antiviral sales were closely correlated with the onset dates of confirmed influenza cases and total confirmed cases (66). There was also a significant association between OTC sales and the number of cases of respiratory syncytial virus cases and number of detections of other respiratory viruses (66). The study authors concluded that the product sales data improved both the timeliness and geographic resolution of surveillance information when compared to other data sources.

The same study also looked at sales of non-prescription OTC gastrointestinal remedies and found no association with outbreaks of gastrointestinal illness in Nova Scotia during the same period (66). However, most of the outbreaks in the time frame examined were reported in residential institutions, such as long-term care homes, which would most likely not be expected to influence OTC sales.

Previously, non-prescription sales of antidiarrheal and antinausea drugs were found to be correlated with the activity of Norovirus in a community, but not with other viral (rotavirus), bacterial and parasitic causes of diarrhea under routine conditions where no outbreak had been identified (67). Rotavirus is a more of a problem in children, where parents were thought to be more likely to seek advice from a health care worker, and therefore be less likely to use over-the-counter remedies. The utility of sales information appears to be best in outbreak situations. Earlier studies of historical water-borne outbreaks of Cryptosporidium, *E. coli* and Campylobacter did show outbreak associated increases in over-the-counter sales of medications for diarrhea and nausea (68).

vii. Retail-based Surveillance – Food Sales

Retail data on food sales has also been used to investigate foodborne disease outbreaks. For example, in 2012 supermarket loyalty cards were used to identify a frozen fruit blend as the source of a hepatitis A outbreak in

British Columbia (69). Permission was obtained from the cases who shopped at major supermarket chains to release detailed 3-month shopping histories obtained from the store loyalty card records. Additional sales data were used to estimate the proportion of all households that might have purchased the product. In a previous outbreak in 2007, a customer savings card program had been used to identify organic basil as the source of an outbreak of cyclosporiasis in British Columbia (70).

More recently, IBM has created spatio-temporal analytics system to compare retail scanning data from grocery stores to locations of foodborne disease cases. This method has been reported to generate a list of the 12 most probable suspect foods (http://barfblog.com/tags/ibm/). The algorithm is described as requiring a minimum of 10 reported cases. The method considers information such as product shelf life, probable date of consumption, and the likelihood that a particular product would contain a particular pathogen. The method was used in Norway on an outbreak involving 17 confirmed *E. coli* cases to target a list of 10 suspected food products from which contaminated sausage was implicated with laboratory testing (71). In a previous study, researchers used grocery retail scanner data from Germany with spatial data to show the method could be used to narrow the list of suspect foods and accelerate the early stages of an investigation (72,73).

viii. Mining High Volume Electronic Health Records

The increasing availability of electronic medical records has raised questions regarding the extent to which this information should be accessible for disease surveillance. Examples of electronic medical health records include records from physician offices, insurance claims, hospital discharge records, and death certificates (1), in addition to laboratory test and imaging results. In the United States, detailed insurance claim forms from physicians were used to develop a fine grained model of influenza transmission at the local level (3). Individual patient information creates the potential to better define predictions using differences in risk based on age and comorbidity (3).

In the United States, the ESPnet was initiated in 2007 and conducts real-time public health surveillance using electronic health records to support local public health surveillance and interventions in Massachusetts (74). It includes automated next-day detection of reportable infectious diseases and aggregate reporting of conditions of interest such as ILI. The actual data are retained within the sentinel practices. The participating servers are configured to accept specific queries from ESPnet system. The system is set up such that public

health personnel can submit queries without extensive programming knowledge. However, for the results to be meaningful it is important to recognize the limitations of the data and factors that can change quickly that might impact data quality.

While Canadian physicians and hospitals are converting to electronic medical record surveillance systems, currently the Canadian FluWatch sentinel practice system is not based on direct access to electronic medical records. Rather it requires manual entry of the total number of patient visits and total number of ILI cases seen by age group for one day every week. There is, however, a Canadian Primary Care Sentinel Surveillance Network (CPCSSN) which focuses on chronic disease and mental health (http://cpcssn.ca/sentinel/potential-sentinels/). Participating primary care providers (e.g. family physicians) provide access to their electronic medical record systems. The program focusses on five chronic disease and mental health conditions including: hypertension, osteoarthritis, diabetes, chronic obstructive pulmonary disease (COPD), depression and three neurologic conditions (Alzheimer's and related dementias, epilepsy, and Parkinson's disease).

While the advent of electronic records appears to be an excellent source of surveillance data, there are a number of factors that could limit if and how these data can be used to enhance understanding of infectious disease and resolve outbreaks. Privacy concerns are the most important barrier to access (1). Confidentiality can be an issue with high granularity data sets even when the data are deidentified and aggregated (3). Medical claims and pharmacy transactions have location of health care facilities or retail establishment and may not have the patient's location (12), which can be quite different particularly when seeking specialist care or in rural areas. Hospital discharge and death certificate data are typically not available in a timely enough fashion to aid health departments in detecting disease outbreaks (75). In addition, a single patient might have multiple health care providers, been seen in multiple settings in some cases across different jurisdictions, and have multiple different insurance coverages for prescription drugs and additional services, requiring linkage across a vast number of systems to be complete (74). Finally, given that the privacy and technical issues can be addressed, the cost of purchasing data sets from private insurance companies for surveillance purposes can be prohibitive (3).

Most electronic health information systems have been developed to serve acute clinical health care providers, not public health. There continue to be policy and privacy questions regarding the extent to which these records can be used for public health. There are also manpower and budgetary factors in many public health departments limiting the extent to which public health departments can access and leverage data from

electronic health records as well as some other sources of big data. Public health departments will need resources to invest in technology and training to make use of opportunities emerging from electronic medical records and other emerging sources of data (76).

Finally, electronic health records may have limited utility for some public health surveillance activities as important variables of interest are typically not recorded including environmental or behavioral risk factors (76).

ix. Participatory/Crowd-sourced Surveillance

Participatory surveillance systems for infectious disease can have a high degree of sensitivity, are timely, and are independent from health seeking behaviors (77). However, participatory or volunteer-based systems may suffer from selection biases due to who chooses to participate, difficulty in adjusting for confounders, limited specificity of syndromic definitions, and issues with inconsistent participation (77). Participation could be limited especially in rural and remote areas by lack of or unreliable access to the internet. Examples of participatory systems include Influenzanet, FluTracking, Reporta, and Flu Near You. Influenzanet has been evaluated in Europe (78) and has successfully detected changes in ILI activity before sentinel physician surveillance. The system is scalable and adding extra participants does not substantially increase the cost of surveillance. Influenzanet is also flexible and can be readily adapted to different definitions of ILI. Additional risk factors can be added to facilitate detailed individual analysis or extend the platform to provide information on additional diseases.

To date this approach has been most widely used for influenza surveillance, but there is the potential to examine other diseases (78). In France food consumption surveys through Influenzanet were used to identify the source of a Salmonella outbreak in early 2016. Other custom websites have been used to crowd source data for enteric disease. For example, the Utah department of health developed the 'I Got Sick' website: https://health.utah.gov/phaccess/public/illness_report/ .

Limitations of participatory surveillance systems include self-selection of the sample, potential for intentionally misreporting data, and the self-reporting nature of signs and symptoms that have not been validated by a physician and laboratory testing (78).

x. Crowd-sourced Health Records

There is also the potential for individuals to share a combination of self-reported health data, medical and laboratory data with web-based surveillance, patient support and research initiatives (79). There are several examples of companies that collect and share medical data. The web forum patientslikeme.com has more than 500,000 members. The site provides patients with opportunities to share their experiences and links patients to clinical research projects (https://www.patientslikeme.com/research/dataforgood).

23andMe.com provides clients in the US with the opportunity to participate in research studies. People who live in Canada were not allowed to participate in these studies at the time of this report. Currently there is also no protection in Canada to prevent individuals from being discriminated against by insurance companies or employers based on their genetic information. As of 2014, 23andMe reported that more than 1,000,000 customers had been genotyped and that more than 80% of these clients had consented to allow anonymous use of their data for research (https://mediacenter.23andme.com/en-ca/fact-sheet/).

Wearable technology is also contributing to crowdsource big data for research. For example, Fitbit has established fitabase, an online library of research to date using information collected by the device (http://www.fitabase.com/research-library/). Wearable technology can provide information on activity and sleep patterns, emotions, blood glucose concentrations, heart rate and blood pressure (80). The very large number of different devices on the market and data types generated current limit the utility of this information for decision making.

xi. Health Call-in Phone Lines

Data from health call in lines are also an important part of many syndromic surveillance systems. Health Link (811) calls are summarized electronically and are part of the Alberta Real Time Surveillance System (ARTSS). Data are extracted and uploaded every 15 min (81). When patients call in to Health Link they talk to a registered nurse who works through a series of algorithms based on their symptoms to provide suggestions for managing symptoms and to direct them as appropriate to other health care providers. As of 2010, the system was generating 100,000 to 125,000 records per year. Although the amount of data is relatively small when compared to other types of big data, similar platforms are necessary to store, manage

and report the information in real-time. While several other provinces have health call in lines, not all have a system for rapid capture the data including the reason for the calls.

Call in data can be an important piece of information for building predictive models. Calls to the National Health Services telephone services NHS Direct line in the UK regarding vomiting were examined to evaluate their relationship with Norovirus activity measured by laboratory results. The authors concluded that when 4% or more of NHS Direct calls reported vomiting for two sequential weeks in all age groups, then the call in system was providing advance warning of increased numbers of laboratory cases (82).

xii. Absenteeism Records

School absenteeism records for 250 elementary schools in City of Edmonton and surrounding county were also summarized electronically as part of the ARTSS system in Alberta (81). Data were extracted and uploaded daily and denominator data were update biannually. The reasons for absence were coded automatically to a standardized reporting list. The data were provided by schools at an individual student level with grade, age, and postal code of the student's residence in addition to the location of the school. As of 2010, there were 500,000 to 650,000 records per year. These numbers would be expected to at least double if the program was expanded to the rest of the province.

xiii. Earth Observation Data

Environmental covariates or spatial data that could be of value in predicting infectious disease include precipitation, temperature, soil type, vegetation, population density, and census data for demographic variables (14). Environmental data has been most commonly used to date in predicting vector-borne diseases such as Rift Valley Fever, WVN, dengue, Murray Valley encephalitis, and Zika (14).

Variables of interest in predicting disease with environmental risk factors include: land cover and land use, vegetation cover, permanent and transient water bodies, flooding, soil moisture, and wetlands, precipitation, temperature, elevation and soil type (83). These data belong to what is referred to as earth observation products and include observations from remote sensing from satellite imagery, direct field observations (e.g.

data from meteorological stations), and outputs from process chains such a meteorological prediction models (83).

An example of processed satellite data would include Global temporal Fourier processed images from MODIS. This data has been used to generate Middle Infrared Reflectance, day- and night-time Land Surface Temperature (LST), Normalised Difference Vegetation Index (NDVI), and Enhanced Vegetation Index (EVI) (84). A number of other satellites also provide data at varying resolutions that are available either at cost or for free (83).

Remote sensing data can be classified based on the spatial resolution or pixel size or the revisit time (83). Typically, remote sensed data reflects a snapshot in time, whereas ground trothed data maybe more likely to provide data on changes over time such as daily precipitation (83). The temporal resolution of the data used in modeling infectious disease should be examined and justified to avoid misleading conclusions (83). The uncertainty in spatial data can and should, where practical, be incorporated directly into epidemiological models using Bayesian methods (83). The prospects for increased availability of satellite-based time-series data are good. Satellites scheduled to be launched after 2015 were to have resolutions as high as 10 to 60 meters with revisit times of 5 days. Other satellites scheduled for deployment were to have revisit times of 1 to 2 days (83).

Satellite image information requires substantial preprocessing before use. The initial versions of big data informed maps created to examine risk for Dengue virus took more than a year to create. The information collected and algorithms were later used to inform maps of Zika virus. Because the Zika maps were built leveraging the previous work, they only required slightly more than a month to generate (85,14). To best model the impact of changes in earth observation data on infectious disease, epidemiologist will need to leverage time-series analytical methods and recent developments in image analysis (83). Unmanned aerial vehicles or drones may also provide new and very fine resolution sources of environmental data (83).

Big Data for Infectious Disease Modeling

i. Models as Tools for Understanding Big Data

The evaluation and application of big data to disease surveillance and management can include a description of what has happened, diagnostics of why we are seeing what we are seeing, predictions of what will happen, and comparisons of alternative management scenarios (86). Predictive models are important tools for understanding disease transmission, reasoning about what is happening and what might happen as time progresses, and then communicating that information. A number of tools have been developed in recent years to model infectious diseases, the most common of which are for influenza.

EpiDMS, described by Liu et al. is one such example of a predictive modelling tool (87). EpiDMS has components for data management, analysis, visualization and epidemic simulation. A number of other modeling systems have also been developed to leverage emerging sources of big data. Researchers at the Los Alamos National Laboratory found that models based on Wikipedia data could forecast the number of people who would become sick for with a lead time of up to 4 weeks, with the best performance reported for models of Dengue fever and influenza. The models were based on queries for Wikipedia pages in seven languages. Comparable cholera models did not perform as well as those for influenza or Dengue fever, perhaps due to limitations in internet access in the most highly affected countries (http://www.livescience.com/49019-web-data-helps-forecast-infectious-diseases.html).

Model building creates opportunities to leverage big data as evidence to inform interventions and public health policy. When one type of data alone does not provide sufficient information, then multiple data types might be combined. For example, large scale demographic, socio-economic, and environmental data have been integrated using spatial methods to predict Cryptosporidia outbreaks (88). Other examples of models using big data include DEFENDER which integrates data from Twitter and news media as well as algorithms for detecting disease outbreaks, situational awareness and forecasting (51). This model includes a nowcasting function which predicted the current, but yet unknown, case counts better than when only previous case count data were considered (51).

Timely and accurate data streams are necessary for more effective modeling of infectious disease outbreaks and for near to real-time model updates. The process of model building should be iterative with the model regularly updated to consider new information, with observation, analysis of the observations, modeling,

forecasting, updating followed by another repeat of the cycle. Models need to be refined to account for changing data sources and extraction algorithms; some of the challenges with Google Flu may in part have been due to failure to account for these factors (33). However, even with access to excellent data, predictive models become less accurate as you attempt to predict farther into the future. For example, the limit of accuracy for weather models is about two weeks (89). High resolution, hourly weather data are gathered from thousands of sites with much of the resulting data being publicly available. A variety of comparably big, although less specific, data sources, including social media and other data streams, are increasingly becoming available to health researchers. The challenge is learning to best use these data to inform model development. Tools such as ORBiT (Oak Ridge Bio-surveillance Toolkit) can, for example, be used to assist in creating parameters for epidemiologic models from big data (90).

The challenges of traditional disease surveillance data limit the capacity of modellers to provide effective decision support due to the lag from observation to reporting, and the geographic and sometimes temporal aggregation necessary due to privacy concerns (89). In addition to the potential increase in the timeliness, granularity, and variety of big data available with which to parameterize predictive models, there is also an opportunity to obtain risk perception data from non-traditional big data sources. The desire to include behavior data is in response to the increasing recognition that risk perception changes uptake of prevention measures, the likelihood of specific transmission pathways and the resulting success of intervention strategies (89).

Changes in human behavior impact the rate of disease transmission in populations. Behavioral changes in response to disease outbreaks have been called prevalence elastic behavior and can include social distancing, wearing masks, and changes in travel behaviors (91). Information from social media has been used to inform models incorporating the impact of mass media on attitudes and subsequent disease transmission (91). Policy regarding the best use of the media to inform the public and control the spread of disease can be enhanced by improving our understanding of the relationship between media coverage of outbreaks and changes in behavior.

The potential to predict disease progression is not the only goal of model building. The immediate benefits of building dynamic models are the increased understanding of factors associated with disease transmission and the identification of information gaps. However, an ultimate, although still distant, objective might be to have a system for infectious disease models comparable to that for weather forecasts, where large volumes data feed into models in real time and locally relevant predictions are readily available on smartphones (89). The

combination of big spatial data and dynamic models could facilitate real-time adaptive management of infectious disease (12). However, there is a need to carefully account for data limitations and stochastics and then communicate the uncertainty in the resulting predictions (12). Bayesian methods have been used in conjunction with various data sources to create predictive models adjusting for underreporting and observational biases in the data streams (92) and to provide estimates of the resulting uncertainty in the findings.

ii. Managing Data Input and Output from Simulation Models – Another Type of Big Data

There is also a very real need to manage the large volumes of data required for and generated by simulation models for different scenarios with different parameters, targeting different spatial scales, and different interventions (87). Model inputs and outputs are often themselves big data. Modellers need to be able to generate, search, visualize and analyze the model inputs and outputs in a timely efficient fashion. EpiDMS was introduced as an example of software designed to meet some of these needs. Examples of input data include demographic attributes, contact networks, age or sex specific contact rates, mobility patterns, and details of interventions at varying levels of spatial resolution (87). The data management system also has to accommodate the potential for parameters and underlying data to dynamically vary over time. Similarly, it is necessary to allow for variation in these parameters as part of sensitivity analysis which examines the impact of range of plausible parameter values on the model findings (87).

In addition to the input parameters, the data management system must be able to accommodate the very large volumes of complex data produced by an ensemble of stochastic realizations that could include hundreds or potentially thousands of simulations (87). Changing conditions will also prompt the need for a series of new simulations in time. For big data to be effectively used by dynamic simulation modeling to inform disease management, there is a critical need for tools to help execute large scale simulation ensembles examining a variety of scenarios and to facilitate the analysis, exploration, interpretation, and visualization of the model outcomes (87). EpiDMS has capacity to query stored data from simulations to find matches for observed disease patterns or targets for intervention.

Visualization of Big Data for Infectious Diseases

For big data to be useful in public health decision making, tools are needed to summarize and support visualization and to enable identification of patterns, trends, correlations, and outliers (93). Basic user support is provided by a number of existing platforms (90). For example, GPHIN provides alerts reported as text to users. ProMED-mail provides GIS maps and alerts prioritized by disease occurrence and scale as well as a text summary of data. HealthMap extends the GIS map reporting features with timelines and tables/graphs.

For many health ministries, influenza is a focus of surveillance and public reporting efforts. The United States CDC has invested in making flu surveillance data publically available in user friendly interactive format (www.cdc.gov/flu/weekly/fluviewinteractive.htm). However, the data is only available at the state level and there is a 2-week delay to reporting laboratory data and mortality (3). The Public Health Agency of Canada also maintains the weekly updated FluWatch report (www.canada.ca/en/public-health/services/publications/diseases-conditions/fluwatch/2016-2017/week11-march-12-18-2017.html) that contains a series of maps, tables and graphs summarizing current key surveillance indicators. Animated maps are provided showing changes in influenza activity over time.

In contrast, most of the available information on foodborne disease is contained in periodic reports or presented as text and tables (www.canada.ca/en/public-health/services/foodborne-illness-canada/surveillance-foodborne-illness-canada.html). PHAC also has an infographic summarizing recent foodborne disease information from Canada (www.healthycanadians.gc.ca/publications/eating-nutrition/foodborne-illness-infographic-maladies-origine-alimentaire-infographie/alt/pub-eng.pdf). Customized time series plots and pie charts are available for all notifiable diseases, including notifiable foodborne illnesses (diseases.canada.ca/notifiable/charts-list).

The most commonly reported visualizations options include point or choropleth maps, graphs of measures of interest over time, pie charts by demographic category, and bar charts (90). These tools are used to communicate disease chronology, geographic distribution, or to make comparisons across age or other risk factors.

i. Increasing Effectiveness of Visualization and Recognizing Barriers

In 2014, Carroll et al. published a systematic review of current practices and issues in visualization and analytics for infectious disease epidemiology (94). There were several findings related to factors that contribute to effective visualizations. There is little question that good visualizations improve comprehension of the data, enhance detection of patterns and resulting inferences from the data. Visualizations should be understandable to users across disciplines and be created in such a way to minimize the risks of information overload or misleading the reader.

Common factors that can be used to assess various types of visualization and analysis tools include: usability, learnability, memorability, error prevention and recovery, efficiency, and user satisfaction (94). Tools should provide options to identify missing data and uncertainty. To optimize tool uptake and usability it is important to understand user needs and preferences, user training, integration of the tool into work practices, the level of understanding and prior use of visualizations, user trust and organization support (94).

Carroll et al. also discussed barriers to creating and interpreting visualizations by public health professionals (94). Commonly cited barriers to tool use included varying levels of computer literacy, insufficient IT support from the organization, limited access to software, and misunderstanding about the purpose of the tool or level of difficulty in using the tool. The authors suggested that routine use of analysis and visualization tools could be improved by incorporating the tools into routine workflows. Aggregation and de-identification of data were also acknowledged as barriers to the capacity to produce meaningful recommendations from the data.

Web-based systems were recommended to decrease the costs of software implementation, improve accessibility for all team members, and enhance the capacity to disseminate findings in different environments (94). The cost of many software packages was reported as a serious issue, with many agencies having no option but to use free software resources. However, the learning curves for many free packages are steeper, the options can be more restricted, and support resources potentially limited. There is the need for high quality online documentation and easy access source code. OutbreakTools is one example of an open source R platform for outbreak data management and analysis although it is not specific to dealing with big data (22).

Confidentiality and privacy issues also limit visualization strategies. One suggested solution to working at a higher spatial resolution in some instances was synthesizing a data set that mimics the underlying characteristics (12). Other commonly reported solutions included aggregating the data to coarser spatial resolution, summarizing data to parameters commonly used in epidemiologic models, and displaying the model results rather than the raw data. Other options have also been reported for meaningful mapping and spatial epidemiologic analyses of local disease data without compromising confidentiality (95,96).

ii. Areas of Focus for Visualizing Big Data

Three areas of focus were identified by Carroll et al. for increasing the uptake and application of visualization tools: geographic information systems (GIS), molecular epi, and social network analysis (SNA) (94). GIS visualization includes simplifying, integrating, and analyzing spatial data. There are current systems with continually updated point maps of infectious disease occurrence (HealthMap) or weekly updated chloropleth maps (FluWatch and WNV). There are also static spatially continuous maps of disease risk, but there were no currently identified online options of continually updated maps demonstrating spatially continuous risk (84). Spatially continuous methods include the application of techniques such as kernel smoothing and model based risk estimates. For example, model predicted WNV infection rates in female *Cx. tarsalis* mosquitos were mapped for July and August 2005–2008 across the Canadian prairies (97).

In addition, the results of spatial or space-time cluster analysis can be mapped and areas of high risk highlighted using relative risk maps as in the example a study of *Salmonella Enteritidis* infection in the greater Toronto area (98). Greene et al. also reported the use of SatScan by the NYC Department of Health to look prospectively for space-time clusters in reportable disease data (99). This process has detected outbreaks of shigellosis. The application of cluster detection and other spatial methods in outbreak investigations were recently summarized in a systematic review (100). Other methods reported to explore spatial surveillance data included geographical analysis machines, dynamic continuous area space time analysis (DYCAST), agent-based modeling, spatial statistics, and self-organizing methods (101). Commercial simulation modeling programs such as AnyLogic have increasingly sophisticated support for incorporating GIS features into agent-based and hybrid simulation models (http://www.anylogic.com/).

Molecular data can present one of the greatest areas of challenge in producing clear and understandable summaries, figures and maps especially for those unfamiliar with the resulting information. Phylogenetic analyses are often represented as trees or dendrograms (102); however, there are a number of graphical options for depicting similarities and differences among organisms (103). The results of molecular analysis with location information can also be mapped as was done for the risk of different genotypes for human Campylobacter infections in New Zealand (104).

The results of SNA are typically represented as a series of graphs highlighting important nodes and subgroups in the network in addition to critical connections among the various nodes and subgroups (94). There are some interesting examples of visualizations where the results of SNA have been combined with molecular data and WGS. In one example, WGS was used together with contact tracing and SNA to uncover the nature of two distinct molecular lineages with a tuberculosis outbreak in high risk social network in British Columbia (102).

iii. Specific Visualization Challenges for Big Data

Ola and Sedig have reviewed several issues specific to the design of visualizations for complex big data (93). They suggest the need for a conceptual framework to make the design process manageable and result in effective visualizations (93). The design of visualizations need to accommodate big data driven tasks and the need for flexible and ongoing dynamic updates of not just the data but the content, focus and questions being addressed in the visualizations.

Interactive visualizations are necessary for big data initiatives to allow data users to simultaneously explore many data elements in order to identify patterns and test hypotheses (93). One example of such an interactive platform is the ORBiT Toolkit (90). In addition to providing tools to integrate more traditional surveillance data (emergency rooms, pharmacy sales) with social media, ORBiT presents the results from the analysis as a dynamic visual interface where the end user can interact and provide feedback.

Simple scatter plots, bar charts, chloropleth maps, heat maps might not be sufficient to examine complex relationships among multiple variables from multiple data sources. Similarly, animations may not be sufficient for demonstrating changes over time as they rely on user memory and are more suitable for presentations than for analysis (93). Similarly, multiple static independent visualizations distributed across

pages on dashboards also rely on the user to integrate the information across the various images (93). Ideally there should be visualizations which allow for multiple aspects of the data to be presented in layers within the same space (93).

While big data could provide a key for significant improvements in efficiency and resulting cost savings in health care, the true impact of big data on disease management directly depends on the availability of tools to turn large pools of data into digestible information that can be used to make timely decisions (93).

Cautions for the Application of Big Data to Infectious Diseases Surveillance and Outbreak Investigations

i. Data Analysis and Risk of False Positives

The mining of big data is typically geared towards hypothesis-generating rather than hypothesis-driven research (105). Conclusions are often based on inductive rather than deductive reasoning, the cornerstone of currently accepted health research. Validation of findings is critical to minimize the potential for false positive conclusions (106). False positive results from spurious correlations can result in significant harm, for example, if concern was raised about what is actually a safe and effective treatment or vaccine. In addition to the risk of type I statistical errors, there is also the potential for the rapid spread of false information on social media and caution is required. This risk can be managed by complementing new data sources and analytical approaches with sound statistical and hypothesis driven analysis, careful causal reasoning, and expert consultation (1).

Big data studies are also often characterized by missing values, lack of congruency, disparate information sources, measurements from different scales, heterogeneity, and cumbersome size (54). Analysis should consist of careful data exploration, classification, and pattern-tracking. The joint modeling of various data sources is complicated by the heterogeneity, noise and spurious correlations, unrecognized dependence of between data and error terms, and latency of important variables of interest (54). The results of big data analysis should be integrated with those from existing systems and be subject to ongoing validation (1).

ii. Need for New Data Management Skills and Analytical Vocabulary

Researchers and practitioners using big data will benefit from becoming acquainted with machine learning, data mining, and machine-based algorithms to search for patterns and relationships in the data. While not all researchers and practitioners will want to learn all of these techniques, it is helpful to be able to communicate effectively with colleagues in computer science. Examples of terms that are currently unfamiliar to many working in disease surveillance include: Hadoop (a programming framework that supports the processing of massive data across many computers), unsupervised learning (analyses that seek to find hidden patterns within the data), graph analytics (analyses that use graphs to understand relationships and patterns), and natural language processing (analyses enabling computers to derive meaning from human language and thus to extract knowledge from documents) (106). Advanced analytics unfamiliar to most health care researchers are needed to unravel the complex interactions and causal relationships in big data analysis (106). However, some tools such as Bayesian networks and graphical models as well as spatiotemporal analytics are already used in some areas of epidemiology.

Much of the raw data available from big data sources of interest are in the form of unstructured text, images, WGS data, and laboratory reports. Preprocessing is necessary to give structure to this data before it can be analyzed. Examples of classification strategies for complex big data include unsupervised machine learning, Bayesian belief networks, deep learning, and ensemble methods (54). Other data clustering and data classification options include Gaussian mixture modeling, random forests, and K-nearest neighbors. Support vector machines (SVM) have been used to create binary linear classifications of complex data based on 'a priori' features identified in the training data (54). More simplistic GLMs may also be appropriate in some cases (54). Principal or independent component analysis are examples of unsupervised exploration tools for quantitative data (54).

Supervised approaches use a training set that includes already classified data to make inferences about or classify new data, and unsupervised approaches identify structure in unlabeled data (54). There are also semi-supervised or partially supervised algorithms (54). Unsupervised analysis of big data provides the potential to look for common threads that might link apparently unrelated data points and identify associations that might have otherwise been missed. However, this type of analysis does not explain the "how" or "why" of these associations (107). Social network analysis can also be used to help extract meaningful information from patterns and connections in data from social media platforms and sensor based devices (54).

To manage very big data, mechanisms are required to facilitate parallel processing of data among separate but connected machines. Parallel computing involves the simultaneous execution of algorithm tasks on a cluster of machines or supercomputers (18). Hadoop was mentioned previously and is an open-sourced implementation of Google's MapReduce (54). This is also an example of a cloud computing platform that allows for centralized data storage and remote internet access to computational resources for infrastructure and software (18). Spark is another example that has been reported to be faster for some applications and to facilitate interfaces with other commonly used programs such as Java, Scala, Python and R needed for queries, streaming, machine learning and graph processing (54). The use of big data requires resources and expertise in data storage and retrieval, identification of errors, security, protocols for sharing, and data analysis (18).

iii. Data Security, Governance and Privacy Issues

The need for system security has been emphasized by many organizations (14,108). Issues of concern include, but are not limited to, vulnerabilities of datasets to cyber intrusion and the potential use of WGS data to design biological weapons, flooding of datasets with false information and hacking databases or computer systems. Security become increasingly complex as big data initiatives often involve several databases and hardware systems in different locations and under different management (109). Data integrity might also be threatened from its origin by the somewhat unlikely potential for the manipulation of crowd-sourced data generation processes to meet the goals of specific groups (38).

Big data initiatives have generated an emerging need for skills in data governance (108). These skills include project cost planning, managing confidentiality and privacy issues, and ethical challenges related to informed consent. Project planners must recognize the sometimes substantial costs associated with cleaning, standardizing, storing, transmitting and securing very large and rapidly growing data sets (108). While management costs are an important consideration, data ownership and privacy concerns often create greater challenges for data governance.

Privacy-related concerns among early adopters of emerging health technologies were summarized by Cheung et al. and could be categorized into four themes: personal data privacy, control over personal information, concerns about discrimination, and considerations for contributing personal data to science (110). It is not

clear if most members of the public are aware of the level of detail in their personal electronic data or how it might be combined with other sources of publicly available data (110). Most individuals are also not aware that HIPAA does not protect personal information that can be derived by linking publicly available databases (110).

There are several ethical challenges inherent in the use of big data and new technology (4). There are questions regarding how adequate informed consent should be defined. There are suggestions for a "critical junctures approach" where consent is obtained at the start of study and then at each point where new information is requested or an intervention is initiated. In addition to consent, there are questions regarding whether the needs of the population under study have been considered and if these needs have been balanced against the increasing access to data and opportunities for rapid development of research and innovation tools (4). There are also important issues with respect to the need to minimize the potential for intrusiveness and participant fatigue, the need for clear and appropriate expectations about responding to safety concerns, and the need to safeguard new streams of sensitive and potentially identifiable data (4).

iv. Data Limitations and Potential Biases

One of the first important limitations is the lack of basic demographic information for many data sources (1). This lack of demographic information can make it challenging to evaluate the potential for selection biases in the results. Coverage for many types of data is limited for very young children and older populations (1). Further, women may be more likely than men to participate in crowdsourcing efforts such as Influenzanet (78).

Instability in data sources is also a real threat. There is no guarantee that social media platforms, internet discussion sites, web sites with publically available data will be used at a similar level of intensity in the future. One example of this is the reference to MySpace and Yahoo discussion forums in older studies (48); both platforms have fallen out of common use in recent years. Volunteer provider platforms also suffer from changes in participation and dropout rates depending on public interest and efforts to maintain participant engagement (78). Flu Near You has, for example, seen isolated spikes of participation and high dropout rates in the United States (78).

Medical claims data may not accurately reflect infectious disease risks even in care-seeking populations due to challenges with billing practices (12). There may also be spatial and temporal errors associated with pharmacy and hospitalization data as the point of sale or where care was received does not necessarily reflect the person's residence or where they became sick (1). Similarly, where a person tweeted regarding getting sick might not reflect where they were exposed to the illness (12).

Different data sources summarized at different levels of aggregation create challenges for merging and analysis (12). Similarly, reporting and health seeking behaviors will vary across time, countries, and age groups raising questions about the appropriateness of merging data from different sources (3). Spatial data is likely to be more complete and higher quality from participants in urban versus rural areas (12) as both CDRs and mailing addresses will have associated errors. Only GPS coordinates will provide good quality location data for rural areas.

Models based primarily on individual data are prone to overfitting and atomistic fallacy (12); however, the ecological fallacy is also an important issue with studies based on aggregate data where outcomes and risk factors cannot be linked at the individual level. Even for studies where data may be available for a large portion of the population, there is a need to insist on statistical rigor with external validation, open source access to data and code, outlier impact evaluations (3), and consideration for more conservative cutoffs from measures of statistical significance (111). Validation should also include replication of study findings (111) with emphasis on the importance of reporting methods adequately to ensure that replication is possible. One of the criticisms of Google Flu was the failure to transparently report the search terms used to extract the data (38).

Finally, there is the question of whether the system is practically accurate and useful for informing management decisions. It is important to ask whether the system is capable of detecting changes in the temporal or spatial dynamics of infection or changes in age groups affected (3). There will be uncertainty about findings associated with a lack of standardization, limited geographic information, and inability to verify or clarify original information from big data sources (89).

Traditional laboratory surveillance contains a lot of information, has a high signal to noise ratio, and high specificity, but suffers from low data volume. It typically has limited complexity and requires minimal preprocessing. Whereas, while Twitter data has high data volume, it is considerably more complex to interpret, and requires significant preprocessing. It also has by comparison a limited amount of useful

information, a low signal to noise ratio, and limited specificity. The best combination of information, signal to noise ratio, specificity and data volume is obtained from hybrid systems that combine traditional surveillance with big data (3).

Summary of How these Data Sources Might be Explored to Predict and Mitigate Foodborne Disease Outbreaks in Canada

PHAC has estimated that each year there are 4.0 million occurrences of domestically acquired foodborne illness in Canada with 1.6 million (40%) of these related to 30 known pathogens (112). Of the total cases of foodborne illness, 11,600 are estimated to result in hospitalizations and 238 in death with 4000 (34%) of the hospitalizations and 105 (44%) of the deaths associated with domestically acquired illness due to the 30 known pathogens (113). In another recent Canadian study, 63.5% of foodborne outbreak investigations reported a specific food as the source of the outbreak (114). In comparison, between 2003 and 2010, a food vehicle was identified in 692/1441 (48%) of foodborne disease outbreaks in the United States (115). These finding supports the need for additional resources and tools for investigating foodborne disease outbreaks to increase the percentage of cases where the source might be identified and to improve future prevention and control efforts. As 32.2% of foodborne outbreaks were associated with a food service establishment there are also immediate options to improve control efforts with better data to target public health inspections (114).

Others have recognized that event-based surveillance for foodborne disease can be enhanced by internet-based information and social media (75). The literature described in the present overview has provided several examples of how whole genome sequencing, aggregated news reports, internet search queries (Wikipedia), data from internet forums (Yelp reviews), Twitter downloads, smartphones, pharmacy-based surveillance, retail-based surveillance of food sales, participatory/crowd-sourced data, and health call in lines have specifically contributed to surveillance for foodborne disease and the prediction and mitigation of foodborne disease outbreaks.

The next steps in this discussion paper were to obtain input from an active researcher on emerging big data opportunities and then from public health practitioners on their understanding of big data, how it is currently being used and how it might be used in the future.

Researcher Commentary – Dr. Nathaniel Osgood

To build upon the findings of this literature review, Dr. Nathaniel Osgood from the University of Saskatchewan was asked to provide a researcher's commentary on the potential for various types of big data to contribute to the investigation of disease outbreaks. As the application of WGS data to foodborne disease has recently been discussed in detail (19), WGS data are not a focus of his discussion. Dr. Osgood is an internationally recognized expert in health modeling and the acquisition of high resolution sensor-based contact data and its application to disease modeling.

As discussed elsewhere in this document, systems seeking to detect cases of foodborne illness can tap many data sources. To enhance outbreak investigations, based in part on work done at the University of Saskatchewan in the Computational Epidemiology and Public Health Informatics laboratory, we envision here a multi-level system that leverages several types of surveillance and detection at varying levels of reach and depth of information provided. The first stage would start with traditional surveillance for cases of highly credible gastrointestinal illness, but would supplement the standard data retrieved from such individuals with data drawn from credit and debit card charges that that identify purchases made at food vendors. This information would be supplemented by surveillance of local restaurant reviews and commentary on sites such as Yelp, but also surveillance of Facebook status updates and Twitter, which have been shown to offer significant volumes of relevant posts. Such information could, in specific circumstances, inform prioritized investigation of vendors.

Perhaps the most important single source of prospective data could consist of syndromic surveillance of a population subset of sentinel individuals (varying in size from 4% of the population and up) who carry an app permitting them to easily report subclinical symptomology encountered, as well as featuring (via opt-out) information on geographically specific movement patterns. The reliable elicitation of self-reported subclinical illness and food consumption is likely to be significantly enhanced by supplementing proactive self-reports (e.g., via app buttons) with use of ecological momentary assessments (EMAs) querying the individual as to their recent food consumption and experience of gastrointestinal illness. By double-checking the completeness of proactive reports, such EMAs are likely to catch many cases of missed illness and food consumption.

Individuals self-reporting gastrointestinal symptoms either proactively or in response to questions may further be asked if they would consider sharing records of their recent purchases, or be queried about their consumption of food from particular food vendors. To address privacy concerns associated with such sentinel cohorts (and thus

potentially significantly expanding the pool of sentinels), an escalation policy could be established. Within this system, only aggregated counts of reports over broad geographical areas would be provided to authorities for typical operation. Individual-level geotagged reports would be maintained in escrow until such a time as a potential outbreak is declared, in which case the records may be released to aid the developing investigation. Such release may either be made automatically or (alternatively) following an explicit authorization by that individual in light of their awareness of the stated public health emergency.

Previous investigation by the authors suggest that the larger volumes of subclinical data retrieved in this more detailed fashion are likely to confer pronounced benefits for enabling far earlier identification of the source of contamination underlying outbreaks. At the same time, recent work has suggested that a stream of aggregate (and thus de-identified) counts of subclinical illnesses – can be leveraged by machine learning techniques to achieve rapid detection of occurrence of an incipient outbreak, thus supporting early initiation of outbreak control efforts.

Looking forward, such ecological momentary assessment for eating are also likely to be more effective if triggered by online classification algorithms that infer possible occurrence of eating, as they will be likely to enquire about food consumption at a time temporally proximate to occurrences, and thereby less affected by recall bias; such an EMA is further likely to be less burdensome. Explorations by the authors suggest that the classification algorithms to identify eating behaviour might fruitfully take into account phone orientation, data from accelerometers and gyroscope, as well as location as estimated by GPS and WiFi. For cases of illness, the information that might prompt a question might include changes in mobility patterns between and within a facility, and conceivably audio cues. As smartwatches grow in their availability and sophistication, detection of both likely food consumption and illness detection may also be enhanced. For food consumption, this could include information of patterns of movement of the watch that could be suggestive of eating behaviour. Classification of cases of foodborne illness may be aided by wearable-gathered physiological data, such as heart rate, heart rate variability, and electrodermal activity information.

Once particular establishments appear suspect due to information gathered through any means, further data from such vendors could be gathered in a way that could help more quickly iterate towards confirmation or falsification of the occurrence of a problem. Point of sale data on purchases made at that location is likely to include information, the timing of that purchase, contact information for the purchaser, and possibly some indication as to the number of people associated with the purchase. Most notably for certain investigations, such

information is likely to include information on the specifics of the foods purchased, which could allow for identification of certain products of concern. The information on the consumer present in such point-of-sale data may in some cases support identification of other individuals at risk via text message or phone, so as to elicit potential additional reports of possible foodborne illness.

SUMMARY OF COMMENTARY FROM PUBLIC HEALTH PRACTITIONERS – PATRICK SEITZINGER

A pilot study was undertaken by MPH thesis candidate Patrick Seitzinger from the University of Saskatchewan to explore the current application of big data to the investigation of foodborne outbreaks in Canada. A short survey consisting of four open-ended questions was administered through FluidSurvey to public health professionals currently working in foodborne outbreak investigation, including epidemiologists, microbiologists and researchers in Canada. Questions assessed perceptions and current uses of big data, gaps in availability of data and ideas for new and innovative applications of big data in outbreak investigations. After reviewing the questionnaire and protocol, an exemption was granted by the Behavioral Ethics Committee at the University of Saskatchewan. In total, 80 public health professionals currently involved in the field of foodborne outbreak investigation were contacted through email. Responses were received from 18 participants across 6 provinces. Participation was anonymous, free and voluntary; consent was obtained from each of the participants. General themes, concepts and ideas provided by these responses are described below.

i. Perceptions of Big Data in the Context of Foodborne Disease Outbreak Investigation

Definitions of big data provided by participants revolved around the idea of an extensive and complex dataset comprised of information from multiple sources and for various purposes being utilized to identify trends, detect changes and anticipate outcomes. Generic sources of big data that were provided include search engine searches (Google), internet social media (Twitter, Facebook), and news reports. Examples of data sources that were noted as being of particular relevance to foodborne outbreak investigation included case histories, telehealth calls, pharmaceutical prescription data, food traceback results and genomic data

(derived from Pulse Field Gel Electrophoresis (PFGE) and whole genome sequencing of human, animal and food samples). The goal of applying big data to foodborne outbreak investigation was understood to be to "enhance response of [foodborne] disease outbreaks" (Participant #4) and to "extend or exceed current information technology analytic capacities and capabilities" (Participant #7).

ii. Examples of How Big Data is Currently Being Applied in Public Health Practice and Research

Of the public health officials that responded to this question, 59% (10/17) indicated having used big data directly in public health practice or research. Examples of big data that had been implemented during past outbreak investigations included shopper loyalty card information, pharmacy data, meta-genomic data as well as information from historical outbreak reports and from the Canadian Food Safety Information Network (CFSIN). These data were used to guide hypothesis generating processes and to target testing and food sampling efforts. In surveillance systems, case histories, and laboratory results have been used to monitor syndromic surveillance trends and to forecast various health outcomes. Examples of the application of big data in research were provided such as in "proof-of-principle projects to evaluate the advantages/disadvantages of bacterial whole genome sequencing vs traditional molecular epidemiology techniques" (Participant #10). While these examples illustrate a wide range of current applications of big data in outbreak detection, response and research, it is important to note that 15% (3/17) stated having used big data in only a limited manner and 24% (4/17) indicated not using big data at all.

iii. Sources of Big Data that are Currently Available to Assist in the Investigation of Foodborne Disease Outbreaks in Canada

When prompted to list specific sources of big data that are currently to public health practitioners in Canada, 35% (6/17) participants explicitly indicated uncertainty in regards to the resources that are currently available. Sources of big data that practitioners were aware of fell into two broad categories; government data (public health records, Census, PulseNet Canada, National Microbiology Laboratory, provincial Integrated Public Health Information Systems (iPHIS), Canadian Network for Public Health Intelligence (CNPHI), antimicrobial resistance surveillance, regulatory agency food product sampling and/or food

product traceback results), and data collected by industry (shopper loyalty cards, market share data, GoogleFlu, GoogleTrends, and historical pharmacy data from Rx Canada). Many of these examples came with caveats such as the lack of timeliness and difficulties in finding the appropriate data mining methods. Despite the variety of examples that were provided, a general sense of inadequacy of current resources and information was evident throughout the survey responses.

iv. Sources of Big Data that would Enhance the Future Foodborne Outbreak Investigations

When asked what sort of data sources, if made available to public health practitioners and researchers in Canada, would most enhance the process of investigating foodborne outbreaks participants brought forth a wide range of ideas and strategies for improving current systems. Participants expressed the desire to be able to access more data on product traceability, social media posts related to dining and reporting issues, quantitative data in regards to google searches, regulatory compliance testing data, surveillance data and food production statistics, real time medical test requests, and reasons for medical visits. The potential benefit of fully funded pharmacy prescription data was mentioned. In particular, providing "[g]eographic locations of diarrhoeal and vomiting prescriptions, matched with human lab isolate trends" (Participant #11) was identified as a promising strategy to identify and located outbreaks in a timelier manner. Food consumption histories, as collected from apps and food logs, were highlighted as an important source of information that would assist foodborne outbreak investigations. Grocery store sales transactions data was identified as an important part of improving knowledge of market shares as it could provide denominators in regards to the number of individuals exposed to a certain product.

Participants reiterated why many of these options may not be appropriate, feasible and/or timely. Among the most commonly listed reasons were privacy concerns, legal limitations, reluctance of companies to release data and a lack of funding to obtain and analyse such data. Barriers to the merging, analysis, storage and visualization were raised repeatedly. Additional identified challenges included gaining access to information, formatting issues, a lack of training and skills required the interpret big data and as stated by Participant #10, "[c]urrently, public health and laboratory personnel do not have the skills to analyze or interpret 'big data', and government IT groups struggle to support 'big data' infrastructure." It was bought to light that jurisdictional issues currently exist not only across organizations and geographical regions, but sometimes

within departments and groups. Lastly, the need to use appropriate levels of caution when interpreting the findings of big data was emphasized to avoid issues of ecological fallacy and inappropriate resource allocation.

In summary, public health practitioners in the field indicated support for the application of big data to foodborne outbreak investigation. Foreseeable risks as well as important implications were identified. Big data was generally perceived as a worthwhile pursuit that had the potential to "contribute to a more fulsome 'picture' of the public health issue" (Participant #7).

CONCLUSIONS

Public health practitioners and researchers should take care to avoid "big-data hubris" (33,38). Big data is a supplement not a substitute for surveillance based on traditional data collection. Public health should work towards hybrid systems that enhance the timeliness and depth of information rather than replace traditional surveillance (1). A workshop on Big Data and Analytics for Infectious Disease organized by the National Academies of Science, Engineering, and Medicine concluded that big data will not replace human decision making, but can be used to provide insights that can increase efficiency and provide focus for further research (14).

"Separating the true signal from the gigantic amount of noise is neither easy nor straightforward, but it is a challenge that must be tackled if information is ever to be translated into societal well-being." (111).

REFERENCES:

1. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big data for infectious disease surveillance and modeling. J Infect Dis. 2016 Nov 14;214(4):S375–9.

2. Fong D, Otterstatter M, Taylor M, Galanis E. Analysis of enteric disease outbreak metrics, British Columbia Centre for Disease Control. Can Commun Dis Rep. 2017 Jan 5;43(1):1–6.

3. Simonsen L, Gog JR, Olson D, Viboud C. Infectious disease surveillance in the big data era: towards faster and locally relevant systems. J Infect Dis. 2016;214(4):S380–5.

4. Pisani A, Wyman P, Mohr D, Perrino T, Gallo C, Villamar J, et al. Human subjects protection and technology in prevention science: selected opportunities and challenges. Prev Sci. 17(6):765–78.

5. Choi BCK. The past, present, and future of public health surveillance. Scientifica. 2012;2012:1–26.

6. Link MW, Battaglia MP, Frankel MR, Osborn L, Mokdad AH. Reaching the U.S. cell phone generation: Comparison of cell phone survey results with an ongoing landline telephone survey. Public Opin Q. 2007;71(5):814–39.

7. Hope K, Durrheim DN, d'Espaignet ET, Dalton C. Syndromic surveillance: is it a useful tool for local outbreak detection? J Epidemiol Community Health. 2006 May;60(5):374–5.

8. Velsko S, Bates T. A Conceptual architecture for national biosurveillance: Moving beyond situational awareness to enable digital detection of emerging threats. Health Secur. 2016;14(3):189–201.

9. Flahault A, Bar-Hen A, Paragios N. Public health and epidemiology informatics. Yearb Med Inform. 2016;10(1):240–6.

10. Wesolowski A, Buckee CO, Engø-Monsen K, Metcalf CJE. Connecting mobility to infectious diseases: The promise and limits of mobile phone data. J Infect Dis. 2016 Dec 1;214(suppl_4):S414–20.

11. Links MG. Big Data is changing the battle against infectious diseases. Can Commun Dis Rep. 2015 Sep 3;41(9):215–7.

12. Lee EC, Asher JM, Goldlust S, Kraemer JD, Lawson AB, Bansal S. Mind the scales: harnessing spatial big data for infectious disease surveillance and inference. J Infect Dis. 2016;214(4):S409–13.

13. G. V Asokan, Vanitha Asokan. Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. J Epidemiol Glob Health. 2015;5(4):311–4.

14. National Academies of Sciences, Engineering, and Medicine. Big data and analytics for infectious disease research, operations, and policy: Proceedings of a workshop. In: The National Academies of Sciences, Engineering, and Medicine (NASEM). Washington, DC: The National Academies Press; 2016.

15. Jonathan Cinnamon, Sarah K Jones, W Neil Adger. Evidence and future potential of mobile phone data for disease disaster management. Geoforum. 2016;75:253–64.

16. Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission. BMJ [Internet]. 2015 May 11;350. Available from: http://www.bmj.com/content/350/bmj.h1314.abstract

17. Mather A, Reid S, Maskell D, Parkhill J, Fookes M, Harris S, et al. Distinguishable epidemics of multidrug-resistant Salmonella typhimurium DT104 in different hosts. Science. 2013;341(6153):1514–7.

18. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: A literature review. Biomed Inform Insights. 2016 Jan 19;2016(8):1–10.

19. Oliver H, Abdo Z, Ricke S. Using WGS to protect public health and enhance food safety: Meeting Summary [Internet]. Available from: www.uspoultry.org/foodsafety/docs/WGS_Meeting_Summary_072916-02.pdf

20. Stasiewicz M, den Bakker H. Introduction to the interpretation of whole genome sequence data in food safety [Internet]. 2016. Available from: https://www.uspoultry.org/foodsafety/docs/WGS_pathogen_characterization_072916-03.pdf

21. Aarestrup F, Koompmans M. Sharing data for global infectious disease surveillance and outbreak detection. Trends Microbiol. 2016 Apr;24(4):241–5.

22. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLOS Comput Biol. 2014 Jan 23;10(1):e1003457.

23. Deng X, den Bakker H, Hendriksen R. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. Annu Rev Food Sci Technol. 2016 Feb;7:353–74.

24. Pettengill JB, Pightling AW, Baugher JD, Rand H, Strain E. Real-time pathogen detection in the era of whole-genome sequencing and big data: comparison of k-mer and site-based methods for inferring the genetic distances among tens of thousands of Salmonella samples. PLOS One. 2016;11(11):e0166162.

25. Tae H, Karunasena E, Bavarva JH, Garner HR. Updating microbial genomic sequences: improving accuracy & innovation. BioData Min. 2014;7(1):25.

26. Mehta S, Vinterbo S, Little S. Ensuring privacy in the study of pathogen genetics. Lancet Infect Dis. 2014;14(8):773–7.

27. Raza S, Luheshi L. Big data or bust: realizing the microbial genomics revolution. Microb Genomics. 2016;2(2):e000046.

28. Dion M, AbdelMalik P, Mawudeku A. Big data and the global public health intelligence network (GPHIN). Can Commun Dis Rep. 2015 Sep 3;41(9):209–14.

29. 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 Mar;92(1):7–33.

30. Bahk CY, Scales DA, Mekaru SR, Brownstein JS, Freifeld CC. Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting. BMC Infect Dis. 2015;15:135.

31. Chowell G, Cleaton JM, Viboud C. Elucidating transmission patterns from internet reports: Ebola and Middle East Respiratory Syndrome as case studies. J Infect Dis. 2016;214(4):S421–6.

32. MacFadden DR, Fisman D, Andre J, Ara Y, Majmuder MS, Bogoch II, et al. A platform for monitoring regional antimicrobial resistance, using online data sources: ResistanceOpen. J Infect Dis. 2016 Nov 14;214(4):S393–8.

33. Salathé M. Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health. J Infect Dis. 2016 Nov 14;214(4):S399–403.

34. Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y Del Valle, Reid Priedhorsky. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol. 2014;10(11):e1003892.

35. Bernardo MT, Rajic A, Young I, Robiadek K, Pham TM, Funk AJ. 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.

36. Michael Edelstein, Anders Wallensten, Inga Zetterqvist, Anette Hulth. Detecting the norovirus season in Sweden using search engine data – Meeting the needs of hospital infection control teams. PLoS ONE. 2014;9(6):e100309.

37. Davidson MW, Haim DA, Radin JM. Using networks to combine "'Big Data'"and traditional surveillance to improve influenza predictions. Sci Rep. 2015;5:8154.

38. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: Traps in big data analysis. Science. 2014;343(6176):1203–5.

39. Domnich A, Panatto D, Signori A, Lai PL, Gasparini R, Amicizia D. Age-related differences in the accuracy of web query-based predictions of influenza-like illness. Plos ONE. 2015 May 26;10(5):e0127754.

40. Martin L, Lee B, Yasui Y. Google Flu Trends in Canada: a comparison of digital disease surveillance data with physician consultations and respiratory virus surveillance data, 2010–2014. Epidemiol Infect. 2016;144(2):325–32.

41. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron A. More diseases tracked by using Google Trends. Emerg Infect Dis. 2009;15(8):1327–8.

42. Rishi Desai, Benjamin A Lopman, Yair Shimshoni, John P Harris, Manish M Patel, Umesh D Parashar. Use of internet search data to monitor impact of rotavirus vaccination in the United States. Clin Infect Dis. 2012;54(9):e115–8.

43. Kang JS, Kuznetsova P, Luca M, Choi Y. Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing [Internet]. 2013. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.431.7286&rep=rep1&type=pdf

44. Harrison C, Jorder M, Henri Stern, Stavinsky F, Reddy V, Hanson H, et al. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013. Morb Mortal Wkly Rep. 2014;63(20):441–5.

45. Nsoesie EO, Kluberg SA, Brownstein JS. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. Prev Med. 2014 Oct;67:264–9.

46. Schomberg JP, Haimson OL, Hayes GR, Anton-Culver H. Supplementing public health inspection via social media. PLoS ONE. 2016;11(3):e0152117.

47. Park H, Kim J, Almanza B. Yelp versus inspection reports: Is quality correlated with sanitation in retail facilities? J Environ Health. 2016;78(10):8–12.

48. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EH, Olsen JM, et al. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. PLoS ONE. 2015;10(10):e0139701.

49. Gittelman S, Lange V, Gotway Crawford C, Okoro C, Lieb E, Dhingra S, et al. A new source of data for public health surveillance: Facebook likes. J Med Internet Res. 2015 Apr 20;17(4):e98.

50. Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, Newth D. Understanding human mobility from Twitter. PLOS ONE. 2015;10(7):e0131469.

51. Thapen N, Simmie D, Hankin C, Gillard J. DEFENDER: Detecting and forecasting epidemics using novel data-analytics for enhanced response. PLoS ONE. 2016;11(5):e0155417.

52. Hawkins J, Tuli G, Kluberg S, Harris J, Brownstein J, Nsoesie E. A digital platform for local foodborne illness and outbreak surveillance. Online J Public Health Inform. 2016;8(1):e60.

53. Allen C, Tsou M-H, Aslam A, Nagel A, Gawron J-M. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. Plos ONE. 2016 Jul 25;11(7).

54. Ivo D Dinov. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. GigaScience. 2016;5(1):1–15.

55. Krieck M, Dreesman J, Otrusina L, Denecke K. A new age of public health: Identifying disease outbreaks by analyzing tweets. In: Proceedings of Health WebScience Workshop. Koblenz, Germany; 2011.

56. Harris JK, Mansour R, Choucair B, Olson J, Nissen C, Bhatt J. Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013-2014. Morb Mortal Wkly Rep. 2014;63(32):681–5.

57. Harris JK, Hawkins JB, Nguyen L, Nsoesie EO, Tuli G, Mansour R, et al. Using Twitter to Identify and Respond to food poisoning: The Food Safety STL Project. J Public Health Manag Pract [Internet]. 2017;Publish Ahead of Print. Available from: http://journals.lww.com/jphmp/Fulltext/publishahead/Using_Twitter_to_Identify_and_Respond_to_Food.99618.aspx

58. Sadilek A, Kautz H, DiPrete L, Labus B, Portman E, Teitel J, et al. Deploying nEmesis: Preventing foodborne illness by data mining social media. In 2016. Available from: https://www.aaai.org/ocs/index.php/IAAI/IAAI16/paper/view/11823

59. Talbot D. Big data from cheap phones [Internet]. MIT Technology Review. 2013. Available from: https://www.technologyreview.com/s/513721/big-data-from-cheap-phones/

60. Chen Y, Crespi N, Ortiz AM, Shu L. Reality mining: A prediction algorithm for disease dynamics based on mobile big data. Inf Sci. 2017 Feb 10;379:82–93.

61. Farrahi K, Emonet R, Cebrian M. Epidemic contact tracing via communication traces. Plos ONE. 2014;9(5):e95133.

62. Mohammad Hashemian, Dylan Knowles, Jonathan Calver, Weicheng Qian, Michael C Bullock, Scott Bell, et al. iEpi: An end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. In: Proceedings of the 2nd ACM international workshop on Pervasive Wireless Healthcare, June 11-11, 2012. Hilton Head, South Carolina, USA; 2012. p. 3–8.

63. Mohammad Hashemian, Dylan Knowles, Kevin G Stanley, Jonathan Calver, Nathaniel Osgood. Human network data collection in the wild: The epidemiological utility of micro-contact and location data. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, January 28-30, 2012. Miami, Florida, USA; 2012.

64. Waldner C, Martin W, Osgood N, Seitzinger P, Teyhouee A, Relf-Eckstein J-A. Exploring new technologies to support investigation of foodborne disease. 2016 Jun 14; Canadian Public Health Association.

65. Todd S, Diggle PJ, White PJ, Fearne A, Read JM. The spatiotemporal association of non-prescription retail sales with cases during the 2009 influenza pandemic in Great Britain. BMJ Open. 2014 Apr 29;4(4).

66. Muchaal P, Parker S, Meganath K, Landry L, Aramini J. Evaluation of a national pharmacy-based syndromic surveillance system. Can Commun Dis Rep. 2015 Sep 3;41(9):204–10.

67. Victoria L Edge, Frank Pollari, Lai King NG, Pascal Michel, Scott A McEwen, Jeffery B Wilson, et al. Syndromic surveillance of Norovirus using over-the-counter sales of medications related to gastrointestinal illness. Can J Infect Dis Med Microbiol. 2006;17(4):235–41.

68. Edge VL, Pollari F, Lim G, Aramini J, Sockett P, Martin SW, et al. Syndromic surveillance of gastrointestinal illness using pharmacy over-the-counter sales: A retrospective study of waterborne outbreaks in Saskatchewan and Ontario. Can J Public Health Rev Can Santee Publique. 2004;95(6):446–50.

69. Swinkels H, Kuo M, Embree G, Fraser Health Environmental Health Investigation Team, Andonov A, Henry B, et al. Hepatitis A outbreak in British Columbia, Canada: The roles of established surveillance, consumer loyalty cards and collaboration, February to May 2012. Euro Surveill [Internet]. 19(18). Available from: https://www.ncbi.nlm.nih.gov/pubmed/24832119

70. Shah L, MacDougall L, Ellis A, Ong C, Shyng S, LeBlanc L, et al. Challenges of investigating community outbreaks of cyclosporiasis, British Columbia, Canada. Emerg Infect Dis. 2009;15(8):1286–8.

71. Norström M, Kristoffersen AB, Görlach FS, Nygård K, Hopp P. An adjusted likelihood ratio approach analysing distribution of food products to assist the investigation of foodborne outbreaks. PLoS ONE. 2015 Aug 3;10(8):e0134344.

72. Kaufman J, Lessler J, Harry A, Edlund S, Hu K, Douglas J, et al. A likelihood-based approach to identifying contaminated food products using sales data: Performance and challenges. PLOS Comput Biol. 2014 Jul 3;10(7):e1003692.

73. Hu K, Edlund S, Davis M, Kaufman J. From farm to fork: How spatial-temporal data can accelerate goodborne illness investigation in a global food supply chain. SIGSPATIAL. 2016;8(1):3–11.

74. Curtis L, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood). 2014;33(7):1178–86.

75. Ford L, Miller M, Cawthorne A, Fearnley E, Kirk M. Approaches to the surveillance of foodborne disease: A review of the evidence. Foodborne Pathog Dis. 2015 Dec 11;12(12):927–36.

76. Tomines A, Readhead H, Readhead A, Teutsch S. Applications of electronic health information in public health: Uses, opportunities and barriers. Gener Evid Methods Improve Patient Outcomes [Internet]. 2013;1(2). Available from: http://repository.edm-forum.org/egems/vol1/iss2/5/

77. Wójcik O, Brownstein JS, Chunara R, Johansson MA. Public health for the people: Participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol. 2014;11:7.

78. Guerrisi, C, Turbelin C, Blanchon T, Hanslik T, Bonmarin I, Levy-Bruhl D, et al. Participatory syndromic surveillance of influenza. Eur J Infect Dis. 2016;214(4):S386–92.

79. Swan M. Scaling crowd-sourced health studies: the emergence of a new form of contract research organization. Pers Med. 2012 Mar;9(2):223–34.

80. Mezghani E, Exposito E, Drira K, Da Silveira M, Pruski C. A semantic big data platform for integrating heterogeneous wearable data in healthcare. J Med Syst. 2015 Oct;39(1):185.

81. Fan S, Blair C, Brown A, Gabos S, Honish L, Hughes T, et al. A multi-function public health surveillance system and the lessons learned in its development: The Alberta Real Time Syndromic Surveillance Net. Can J Public Health. 2010;101(6):454–8.

82. Loveridge P, Cooper D, Elliot A, Harris J, Gray J, Large S, et al. Vomiting calls to NHS Direct provide an early warning of norovirus outbreaks in hospitals. J Hosp Infect. 2010 Apr;74(4):385–93.

83. Nicholas A S Hamm, Ricardo J Soares Magalhães, Archie C A Clements. Earth observation, spatial data quality, and neglected tropical diseases. PLoS Negl Trop Dis. 2015;9(12):e0004164.

84. Hay S, George D, Moyes C, Brownstein J. Big data opportunities for global infectious disease surveillance. PLoS Med. 2013;10(4):e1001413.

85. Messina JP, Kraemer MU, Brady OJ, Pigott DM, Shearer FM, Weiss DJ, et al. Mapping global environmental suitability for Zika virus. eLIFE. 2016 Apr 19;5:e15272.

86. Hilton BN. Overview of Spatial Big Data and Analytics [Internet]. 2015 Dec 13; Fort Worth, Texas. Available from: https://www.redlands.edu/globalassets/depts/school-of-business/gisab/workshops-conferences/brian-hilton-icis_2015_bnh.pdf

87. Liu S, Poccia S, Candan KS, Chowell G, Sapino ML. epiDMS: Data management and analytics for decision-making from epidemic spread simulation ensembles. J Infect Dis. 2016 Nov 14;214(4):S427–32.

88. Lal A. Spatial modelling tools to integrate public health and environmental science, illustrated with infectious Cryptosporidiosis. Int J Environ Res Public Health. 2016 Feb;13(2):1–8.

89. Moran KR, Fairchild G, Generous N, Hickmann K, Osthus D, Priedhorsky R, et al. Epidemic forecasting is messier than weather forecasting: The role of human behavior and internet data streams in epidemic forecast. J Infect Dis. 2016;214(4):S404–8.

90. Ramanathan A, Pullum L, Steed C, Quinn S, Chennubhotla C, Parker T. Integrating heterogeneous healthcare datasets and visual analytics for disease. In: 3rd IEEE Workshop on Visual Text Analytics. 2013.

91. Mitchell L, Ross JV. A data-driven model for influenza transmission incorporating media effects. R Soc Open Sci. 2016 Oct 26;3(10):160481.

92. Yang W, Lipsitch M, Shaman J. Inference of seasonal and pandemic influenza transmission dynamics. Proc Natl Acad Sci. 2015 Mar 3;112(9):2723–8.

93. Ola O, Sedig K. Beyond simple charts: Design of visualizations for big health data. Online J Public Health Inform. 2016;8(3):e195.

94. Carroll LN, Au AP, Detwiler LT, Fu T, Painter IS, Abernethy NF. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform. 2014 Oct;51:287–98.

95. Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. Int J Health Geogr. 2008;7(1):13.

96. Zandbergen PA. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. Adv Med. 2014 Apr 29;2014(1):567049.

97. Chen C, Epp T, Jenkins E, Waldner C, Curry P, Soos C. Modeling monthly variation of Culex tarsalis (Diptera: Culicidae) abundance and West Nile Virus infection rate in the Canadian prairies. Int J Environ Res Public Health. 2013;10(7):3033–51.

98. Varga C, Pearl DL, McEwen SA, Sargeant JM, Pollari F, Guerin MT. Evaluating area-level spatial clustering of Salmonella enteritidis infections and their socioeconomic determinants in the greater Toronto area, Ontario, Canada (2007 – 2009): a retrospective population-based ecological study. BMC Public Health. 2013;13(1):1078.

99. Greene SK, Peterson ER, Kapell D, Fine AD, Kulldorff M. Daily reportable disease spatiotemporal cluster detection, New York City, New York, USA, 2014–2015. Emerg Infect Dis. 2016 Oct;22(10):1808–12.

100. Smith C, Le Comber S, Fry H, Bull M, Leach S, Hayward A. Spatial methods for infectious disease outbreak investigations: systematic literature review. Euro Surveill. 2015 Oct;20(39).

101. Musa GJ, Chiang P-H, Sylk T, Hoven CW. Use of GIS mapping as a public health tool-from cholera to cancer. Health Serv Insights. 2013 Nov 19;2013(6):111–6.

102. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011;364:730–9.

103. Cowley LA, Beckett SJ, Chase-Topping M, Perry N, Dallman TJ, Gally DL, et al. Analysis of whole genome sequencing for the Escherichia coli O157:H7 typing phages. BMC Genomics. 2015 Apr 8;16(1):271.

104. Muellner P, Shadbolt T, Collins-Emerson J, French NP. Molecular and spatial epidemiology of human Campylobacteriosis: source association and genotype-related risk factors. Epidemiol Infect. 2010;138(10):1372–83.

105. Grant E. The promise of big data [Internet]. Harvard T.H. Chan School of Public Health. 2012. Available from: https://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/

106. Krumholz HM. Big data and new knowledge in medicine: The thinking, training, And tools needed for a learning health system. Health Aff (Millwood). 2014;33(7):1163–70.

107. Austin C, Kusumoto F. The application of big data in medicine: Current implications and future directions. J Interv Card Electrophysiol. 2016;47(1):51–9.

108. Kruse C, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: A systematic review. JMIR Med Inform. 2016 Nov 21;4(4):e38.

109. American Association for the Advancement of Science in conjunction with the Federal Bureau of Investigation and the United Nations Interregional Crime and Justice Research Institute (AAAS-FBI-UNICRI). National and Transnational Security Implications of Big Data in the Life Sciences [Internet]. 2014 Nov. Available from: https://www.aaas.org/report/national-and-transnational-security-implications-big-data-life-sciences

110. Cheung C, Bietz MJ, Patrick K, Bloss CS. Privacy attitudes among early adopters of emerging health technologies. Plos ONE. 2016;11(11):e0166389.

111. Khoury M, Ioannidis J. Medicine. Big data meets public health. Science. 2014 Nov 28;346(6213):1054–5.

112. Thomas M, Murray R, Flockhart L, Pintar K, Pollari F, Fazil A, et al. Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, Circa 2006. Foodborne Pathog Dis. 2013;10(7):639–48.

113. Thomas MK, Murray R, Flockhart L, Pintar K, Fazil A, Nesbitt A, et al. Estimates of foodborne illness–related hospitalizations and deaths in Canada for 30 specified pathogens and unspecified agents. Foodborne Pathog Dis. 2015 Oct 1;12(10):820–7.

114. Belanger P, Tanguay F, Hamel M, Phypers M. An overview of foodborne outbreaks in Canada reported through Outbreak Summaries: 2008-2014. Can Commun Dis Rep. 2015 Nov 5;44(11):254–62.

115. Mecher T, Stauber C, Gould LH. Contributing factors in a successful foodborne outbreak investigation: An analysis of data collected by the Foodborne Diseases Active Surveillance Network (FoodNet), 2003-2010. [Internet]. Georgia State University; 2015. Available from: http://scholarworks.gsu.edu/iph_theses/382