12-3-2014

The potentials and challenges of big data in public health

Rena N. Vithiatharan
Edith Cowan University

DOI: 10.4225/75/579824f531b45
3rd Australian eHealth Informatics and Security Conference. Held on the 1-3 December, 2014 at Edith Cowan University, Joondalup Campus, Perth, Western Australia.

THE POTENTIALS AND CHALLENGES OF BIG DATA IN PUBLIC HEALTH

Rena N Vithiatharan
Edith Cowan University
rvithiat@our.ecu.edu.au

Abstract

The potential to use big data sources for public health increases with the broadening availability of data and improved methods of analysis. Whilst there are some well-known examples of the opportunistic use of big data, such as GoogleFlu, public health has not yet realised the full potential of such data sources. A literature review was undertaken to identify the potential of such data collections to impact public health, and to identify what challenges are currently limiting this potential. The potential include improved real-time analysis, research and development and genome studies. However, challenges listed are poor universal standardisation and classification, privacy and security, as well as current inadequate platforms and tools for analysis. Without such reviews, limited understanding will hinder the rate of advance in utilising such data to improve the health status of population in public health.

Keywords
Big Data, Public Health, Healthcare, Analytics.

INTRODUCTION

Big Data

Big data is a large and complex digital dataset (Pope, Halford, Tinati & Weal, 2014). True to its name, Big data is measured in petabytes (1015) or exabyte (1018). The cumulative volume and detail of data captured by organisations, the increase of audio-visual media, social media, and the Internet increases exponential growth in data. This is likely to continue in the future. Likewise in the field of health, Harper (2013) suggests that technology growth, health care reforms and patient-centred care has caused data to experience exponential growth. Data mining or data being retrieved occurs in everyday individual transactions of diverse sources. Some examples are credit card or loyalty card transactions, mobile phone patterns, social network websites usage and internet usage patterns. Unstructured data is collected by data mining of virtual activities from Facebook, Twitter, text, video and audio material and geo-location. A secondary analysis of data pulled out enables raw data to be made into information for informed decision-making through insights taken from the analysis (Kum & Ahalt, 2013). Harper (2013) believes that big data has a valuable function at the level for the global economy.

Public Health

The World Health Organisation (WHO, 2014) defines public health as all publically or privately organised measures that promote good health, prevent diseases, and prolong life among population groups as a whole. Public health focuses on total systems and the entire populations, aiming to provide condition where people can live healthily. It is not targeted at individuals or diseases in particular. The World Health Organisation (2014) further categorises public health into three main functions. Firstly, public health assesses and monitors the health of the population to identify health problems and priorities. Secondly, public health formulates policies that are intended to address local, national and global health issues. Lastly, public health functions to ensure the public enjoys equal access to appropriate, cost effective healthcare and services.

Big Data in Public Health

There is an expectation of big data to be able to provide strong input that can be accessed, analysed and put into action. If used creatively and effectively, big data can improve the efficiency and quality of public health. Literature suggests that big data can potentially provide strategic agility in the public health domain (Jalali, Oabode & Bell, 2012). Strategic agility refers to capacity in capitalising prospects and avoiding threats with speed and assurance. Strategic agility is the ability to continuously adjust and adapt the direction of core functions and changing circumstances to create new services, models and innovative ways to enhance health.

Sensitive perception, swift decision making, the fluidity to configure systems and manage resources are other key enabling capabilities in strategic agility.

This paper is a literature review of current views of the potentials and challenges in using big data in the public health domain. The paper will also discuss the methods that will overcome the challenges faced, and the prospects that big data holds for public health.

CURRENT USES OF BIG DATA IN PUBLIC HEALTH

GoogleFlu is a widely known example where big data is used for public health. GoogleFlu predicts trends in the outbreaks of flu by tabulating internet searches related to the word 'flu' in different regions (Williams, 2013). The prediction is based on the correlation that locations with an increased number of flu related searched are experiencing an increased number of flu cases.

Similarly, the Centre of Disease Control and Prevention (CDC) in the United States use reality data mining techniques based on purchasing patterns at pharmacies, commuting traffic as well as school and work attendances (Yokeo et al., 2004). This information allows swift response to the predictions of outbreaks. Public health measures in having adequate medications and support in hospitals can then be taken. The economic impact on sick patients, residents and employees can be anticipated and contingency plans put in place.

The social media has also been a great resource for big data in which innovations and research can be done to analyse infectious and chronic disease trends in populations. Given the vast amount of people who are connected online, at one given point of time, someone is publically sharing volumes of personal and community health information. Social media allows for reciprocity; the more one person shares, the more others share in return (Kass-Hout & Alhinnawi, 2013). An example is a study by Salathe and Kandelwal (cited in Kass- Hout & Alhinnawi, 2013) where the researchers assessed vaccination sentiments on the social media during the H1N1 pandemic. Anti-vaccination sentiment trends could be seen more rapidly in some parts of the network and negative sentiments spread more effectively than positive sentiments. An algorithmic analysis was done and compared to the CDC data on vaccinations according to geographical locations. A correlation was found between higher negative sentiments on social media and lower vaccination rates according to geographical location (Kass-Hour & Alhinnawi, 2013).

POTENTIAL

Real-Time Analysis

Pope, Halford, Tunati and Weal (2014) highlight that big change is in the scale and volume of the data that can potentially be used for research purposes and in policy development. One of the major potential of big data in the area of public health is in being able to gain dynamic and real time data in healthcare.

Hay, George, Moyers and Brownstein (2013) concur as big data is able to offer an improved gauge of new diseases with the understanding of their natural and geographical setting. Such systems do not only provide static spatial continuous images of an infectious disease at risk, but can continuously update reports of infectious disease occurrences. Hay, George, Mayer and Brownstein (2013) further add that currently only 2% of world infectious diseases are mapped out comprehensively.

The combination of social media, epidemiology and applicable ecological information are valuable sources that can create the prospects of developing a continually updated atlas of infectious diseases. This can also be represented in flexible graph based data (Bromley et al., 2014). Such free dynamic real time infectious disease maps become a valuable source for health care professionals and policy makers. Public health professionals are able to prioritise where limited resources can be spent aptly with such information. The understanding of such epidemiology improves the understanding of disease factors while increasing life expectancy and bringing changes to the leading causes of death of the population. This knowledge leads to precautionary and prevention measures that can be taken by governments to protect the health of a population through environmental interventions, policies and promoting health literacy through clinical and community interventions (Remington & Brownson, 2011).

Another value of dynamic real time information as suggested by Bromley et al. (2014) lies in the ability to detect disease outbreaks at an earlier stage. This leads to more effective and swift responses, prevention and treatment. The analysis of disease pattern in real time, coupled with faster development can turn large data into actionable information.

Research and Development

Harper (2013) claims that big data has the ability to replace and support human decisions with formulated algorithms in order to manage population health and make better decisions. Big data can provide strong analytics power to gain deeper understanding and insight into the factors that impact and influence health. It also can improve statistical validity that may provide better analysis and insight compared to traditional health research methods. This enables the building of an evidence based best practice and learning in the health care system. Thus, data mined from electronic health records when matched with individual life trends will enable increase in patient care, spotting of potential health risks, predicting public health trends and ultimately reduce health care costs. Hence, the application of big data is not only beneficial to public health, but there are economic possibilities as well.

Claney et al. (2014) further reiterate that preventable errors that are highlighted from big data will save costs and money in health care. However, Ragupathi and Ragupathi (2014) note that better statistical tools and algorithms are needed for good predictive modelling. Through this, structured and unstructured data can be combined for the benefit of evidence based medicine.

Genome Studies

Another potential area of big data is in genome study for cancer (More, 2013; Issa, Byers & Dakshanamurthy, 2014). Knowledge on genome profiles gives information on the metabolism of medications within our bodies. Genomics analytics executes gene-sequencing that can contribute to a part of regular medical care and decision making (Ragupathi & Ragupathi, 2014). Issa, Byers and Dakshanamurthy (2014) predict that big data can provide insights into the genetic changes upon exposure to stressors. This contributes to the pharmacology development and side effect studies. Genomic application in the area of public health is valuable for the future. Evaluations of preventive programs and services can be conducted. Once again, to achieve this, centralisation and standardisation is essential to be able to do data mining. From such information and studies, the public health domain is also able to create individualised health plans and programs. Ragupathi and Ragupathi (2014) claim big data has benefits in managing health at an individual and population levels. Mare (2013) finds that the ability to get the most out of the data lies in interpreting them in the light of prior knowledge. There is different software available for this task, and data can be stored in the digital cloud to save money and hardware space.

Kum, Krishnamurthy, Machanavajjahala and Ahalt (2014) also suggest a concept of social genome. This is the footprint of society in general. Information from Facebook pages, for example can offer the behavioural patterns of users. In the context of mining thousand others of such social network platforms, it will give a lead to more accurate information on human behaviour. Such information can open the way to better informed and effective policy decisions and management of social programs. Through social genome, public health professionals are able to look at the long terms effects of how society behaves and evolves. Extraction insights does become challenging due to useless and erroneous information (Ola & Sedig, 2014). An example of social genome is the Google flu project (Google.org flu trends, 2011). This project combines information on physician visits with individual search queries on Google. Given its success, Google now partners with CDC in providing real time information on flu outbreaks.

CHALLENGES

Raguphati and Ragupathi (2014) define big data in terms of 4Vs; volume, variety, velocity and veracity. Volume refers to the large quantity of records and information. The size of existing data is growing at an accelerating rate. It is estimated to grow to 40 zettabytes (10^{21}) by the year 2020, representing more data (Ragupathi & Ragupathi, 2014). Sources with bigger amounts of data pool to increase the volume that can be analysed. This data is valuable to use. Variety refers to the different forms of sources and types of data. Additional sources of data are being added continually. Currently, a large number of smart phones transmit a range of information to the network infrastructure. As new applications are introduced, it is also the genesis of new data formats. Velocity is the rapid speed at which data is growing and the analyses of the streaming data. Current sources of data from social and mobile applications leave previous methods of batch data processing non-viable. The data is now streamed into the servers in real time, continuously providing useful information with minimum delays. Finally, veracity is the data assurance received, ensuring that the data is error free. Big data veracity refers to excluding biases, erroneous noise and abnormality in data. Hence, the data can be stored, mined and analysed meaningfully.

Standardisation and Classification

One of the challenges of big data is that it is not necessarily open and accessible data. This can be due to several reasons. Firstly there are non-standard computation facilities for storage, management and analysis of data. Thus, information that is available may not be linked, deeming it unusable in an unstandardized format which reflects poor integration within key technologies. (Pope, Halford, Tinati and Weal, 2014; Clansy et al., 2014; Ragupathi & Ragupathi,2014).Pope Halford, Tinati and Weal (2014) further claim that detailed data is required in order to make good sense of the information. The semantic interoperability becomes a linking factor unless careful understanding of metadata occurs. Haper (2013) adds that standardisation leads to a universal exchange, enabling the sharing of a resource of data storage. Currently data that is stored digitally may not be able to be analysed.

Classification matters arise too. Pope, Halford, Tinati and Weal (2014) claim data is normally socially constructed. Hence, it cannot be used effectively without the information of where and how the data was collected in order to analyse and categorise them. This raises the question of whether big data is open data and useable. Data that is not infused with meaning may not be valid. Harper's (2013) study shows that just having data is not the sole solution. Although we have a lot of data, we have very little information and knowledge. Ola and Sedig (2014) add that as much as data is essential, the ability to analyse and effectively use the data is critical. Ola and Sedig (2014) find that since data is collected from heterogeneous sources, it may be unreliable and volatile, which may affect the effective use of such data. Interoperability and interfacing of different systems and practice managements may not be possible without incurring disruption, high costs and be time consuming.

Privacy and Security

Privacy is another issue of concern. Although there are attempts to de-identify data, it does not mean that the data is fully protected (Kum & Ahalt, 2013; Kum, Krishnamurthy, Machanavajjahala & Ahalt, 2014). Kum and Ahalt (2013) assert that linkage attacks using quasi-identifiers can re-identify individuals. There is a direct relationship between data usability and privacy. Several matters will have to be addressed to gain the full potential of big data. Policies linked to privacy, security and intellectual property will need to be addressed in a big data world. Organisations need right technology, workflow structure, talent and incentives to maximise the use of big data. Organisations will need to integrate information from multiple data sources to access information.

Williams and Hossack (2013) point out that implementation of policies and legislations are rushed which result in poor implementation. Health organisations are unaware of the full potential risks of data breaches and its impact on privacy. This jeopardises a patient's trust, autonomy and peace of mind which may result in the patient's unwillingness to share information.

Consumers tend to be wary of privacy issues as well and this can be capitalised by agencies and organisations. An example cited in Pope, Halford, Tinati and Weal (2014) states that Tesco in the United Kingdom advertise that they protect the information of their 16million loyalty card members. This has become a selling factor for them. Other issues that arise are data ownership and consent for example. Since data mined is considered to be secondary data, answers to questions that arise from consumer consent and data ownership are still vague. This will contribute to gaps of information

Platform and Tools for Analysis

Mare (2013), Kum, Krishnamuthy, Machanavajjahala and Ahalt (2014) with Ragupathi and Ragupathi (2014) suggest an engineering problem of computational platform and software tools that are able to handle large amounts of data. Big data needs data and tools that communicate with one another. The big data pressure is for engineering tools with stability and longevity. Clansy et al. (2014) also highlights a lack of leadership in big data. An example of a tool given by Mare (2013) is DNA nexus, which is an analysis platform for non-specialist users, which makes it user friendly to public health professionals to understand data that is mined.

Predictive analytics which is a result of the analysis of big data enables decision making by using data from past performance and aggregation of information to predict a population's behaviour and risk factors (Early, 2014). However, predictive analytics does not take into account the unknown variables of the future, and may give an inaccurate representation of information. Such predictions and information should be used wisely and responsibly.

THE WAY AHEAD; HARNESSING CHALLENGES INTO POTENTIALS

Clansy et al. (2014) concurs with Kum, Krishnamurthy, Machanavajjahala & Ahalt (2014) in affirming that leadership in big data and public health informatics is important to develop an action plan to shape policies. Health care professionals need to be educated on the importance of standardised data and greater advocacy and adoption of standard codes and identifiers should be in place. The views of stakeholders need to be heard and represented in a collective voice. Current data evaluation criteria should also be assessed and updated. In addition to these, bold participation in standards development is also required.

Kum and Ahalt (2013) highlight the necessity to train scientists in spotting potential harm and differentiation between required and extraneous data. Security can also be monitored by limited access through user authentication. This also gains information accountability. Jalali, Oabode and Bell (2012) suggest using Virtual Private Cloud (VPC) windows for security as it decreases isolation and increases flexibility. Cloud computing uses a network of internet remote servers hosted to keep, manage, and process data. Mare (2013) affirms that storing data in clouds save money and hardware space. This is a solution that benefits public health. It enables the sharing of information between public health and health care organisations for real time analyses and shared decision making. The issues surrounding privacy will in turn need to be addressed. Harper (2013) with Bromley et al. (2014) note the need for adequate and proper tools to abstract, collate and synthesise data from heterogeneous sources.

Hay, George, Moyers and Brownstein (2013) remind that the challenge in harnessing the potentials is to be able to provide sustainability and to engage the wider audience. Kim (2014) describes a project done in South Korea. The Health Avatar project aimed to develop a software platform for personalised management of health information in Korea. Through this system, all health data of an individual is conceptualised as a person's health avatar. Through the Health Avatar project, smartphones of individuals become enabled to carry their personal health record. As smart phones are used, the physical activity patterns are also monitored and that information is stored. Privacy is maintained by controlled access through authentication.

CONCLUSION

This paper reviews the dimensions of potentials that big data can contribute to the public health domain. Potential include real-time data analysis, a more rigorous research and development arena as well as valuable information from genomic studies. However there are areas of concern and challenges in applying big data to public health. These challenges include the current lack of universal standardisation and classification that may render big data to be of poor use. There are also privacy and security concerns where individuals will lose the right to their private information, leading to a future with no secrets. Another challenge is the need for platforms and powerful tool to analyse the large and rapidly growing amount of data.

However, these challenges can be overcome with good leadership, training, specialisation, advocacy and contemporary policies to support the development of public health informatics. The use of big data in health care differs from its use in other industries such as marketing or product development. These differences can be seen in the need for regulations, ethical standards, privacy boundaries and some form of standardisation in the diversity of data sources and in their differing goals. There is a need for more skilled health care informatics professionals and leaders to deal with big data confidently and to address the challenges that arise from the use of big data. As in any science, data is evidence and knowledge. The ability to harmonise and analyse data enables the sharing of knowledge across disciplines and to gain insight into the complex and challenging field of public health. Big data, used wisely, is a Pandora's Box for health care.

REFERENCES

Bromley, D., Rysary, S.J., Su, R., Toofany, R.D., Svhmidlin, T &Daggett, V. (2014). DIVE: a data intensive visualization engine. *Bioinformatics, 30* (4), 593-595.

Clansy, T., Bowles, K., Gelias, L., Androwich, I., Delaney, C., Matney, S., Sensmeier, J., Warren, J., Welton, J. & Westra, B. (2014). A call to action: Engage in big data science. *American Academy of Nursing on Policy, 62,* 64-65.

Early,S. (2014). Big data and predictive analytics: What's new? *IT Professional, 16*(1), 13-15.

Google.org flu trends. (2011). Retrieved from the Google.org site: http://www.google.org/flutrends/intl/en_us/

Harper, E. (2013). The economic value of health care data. *Nurse Administration Quarterly, 37*(2), 105-108.

Hay, S.I., George, D.B., Moyers, C.L. & Brownstein, J.S. (2013). Big data opportunities for global infectious disease surveillance. *PLOS Medicine, 10*(4), e1001413-e100141x

Issa, N.T., Byers, S.W., & Dakshanamurthy, S. (2014). Big data: The next frontier for innovation in therapeutics and healthcare. *Expert Review in Clinical Pharmacology, 7*(3), 293-298.

Jalali, A., Olabode, O.A. & Bell, C.M. (2012). Leveraging cloud computing to address public health disparities: An analysis of the SPHPS. *Online Journal of Public Health Informatics, 4*(3), e13-e20.

Kass-Hout, T & Alhinnawi, H. (2013). Social media in public health. *British Medial Bulletin,* 1-20. doi:10.1093/bmb/ldt028

Kim, J.H. (2014). Health avatar: An informatics platform for personal and private big data. *Healthcare Informatics Research, 20*(1), 1-2.

Kum,H. & Ahalt, S. (2013). Privacy-by-design: Understanding data access models for secondary data. *AMIA Summits on Translational Science proceedings, 2153-4063*, 126-130.

Kum, H., Krishnamurthy, A., Machanavajjahala, A., Ahalt, S. (2014). Social genome: Putting big data to work for population informatics. *Computer, 18*, 56-63.

Mare, V. (2013). The big challenge of big data. *Nature, 498*, 255-259.

Ola, O. &Sedig, K. (2014). The challenge of big data in public health: An opportunity for visual analytics. *Online Journal of Public Health Informatics, 5*(3), e223-e245.

Pope, C., Halford, S., Tinati, R. & Weal, M. (2014). What's the big fuss about big data? *Journal of Health Services Research and Policy, 19*(2), 67-68.

Ragupathi, W. & Ragupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information, Service and Systems, 2*(3), 1-10.

Remington, R. & Brownson, R. (2011). Fifty years of progress in chroming disease epidemiology and control. *Morbidity and Mortality Weekly Report, 60*(4), 70-77.

Williams, P.A.H. (2013). Creating context: Making sense of geo-location and social media data for health. In H. Grain (ed.) *Studies in Health Technology and Informatics*, pp. 148-154. IOS Press: Amsterdam.

Williams, P.A.H. & Hossack, E. (2013). It will never happen to us: The likelihood and impact of privacy breaches on health data in Australia. In H. Grain (ed.) *Studies in Health Technology and Informatics,* pp. 155-168. IOS Press: Amsterdam.

World Health Organisation[WHO]. (2014). *Trade, foreign policy, diplomacy and health.* Retrieved from http://www.who.int/trade/glossary/story076/en/

Yokoe, D., Coon, S., Dokholyan, R., Iannuzzi, M., Jones, T., Meredith, S., Moore, M., Phillips, L., Ray, W., Schech, S., Shatin, D. & Platt, R. (2004). Pharmacy data for tuberculosis surveillance and assessment of patient management. *Emerging Infectious Diseases, 10*(8), 1426- 1431.