How to cite this paper:

Nur Laila Ab Ghani, Sulfeeza Mohd Drus, Noor Hafizah Hassan, & Aliza Abdul Latif. (2017). Factors of emerging infectious disease outbreak prediction using big data analytics: A systematic literature review in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference of Computing & Informatics (pp 37-42). Sintok: School of Computing.

FACTORS OF EMERGING INFECTIOUS DISEASE OUTBREAK PREDICTION USING BIG DATA ANALYTICS: A SYSTEMATIC LITERATURE REVIEW

Nur Laila Ab Ghani[1], Sulfeeza Mohd Drus[2], Noor Hafizah Hassan, and Aliza Abdul Latif[4]

[1]*Universiti Tenaga Nasional*, Laila@uniten.edu.my
[2] *Universiti Tenaga Nasional*, sulfeeza@uniten.edu.my
[3] *Universiti Tenaga Nasional*, hafizah@uniten.edu.my
[4] *Universiti Tenaga Nasional*, aliza@uniten.edu.my

ABSTRACT. Infectious disease is an illness that can be transmitted from an infected individual to another. During the pre-vaccine era, infectious disease epidemics caused major fatalities in the population. The invention of vaccines that have dramatically reduced fatalities caused by infectious disease, led to the establishment of Global Immunization Vision and Strategy initiative that aims at increasing national vaccination coverage around the world. However, the appearance of emerging infectious disease calls for an establishment of an early warning mechanisms that can predict the next outbreak. Mathematical and statistical model that has been used to predict infectious disease outbreak used single source datasets that is inadequate for public health policymaking. Literatures suggested using big data analytics to get a better and accurate model. Big data deals not only with structured data from electronic health records but also integrate unstructured data obtained from social medias and webpages. Thus, this paper aims at identifying the factors frequently used in studies on infectious disease outbreak prediction, focusing specifically on two common disease outbreak in southeast Asia: dengue fever and measles. A systematic literature review approach that search across four databases found 284 literatures, of which 10 literatures were selected in the final process. Based on the review, it seems that studies on measles outbreak employed only single source datasets of patient data retrieved from electronic health records. Further research on measles outbreak prediction should combine various types of big data to produce more accurate prediction results.

Keywords: infectious disease; dengue fever; measles; outbreak prediction; epidemics; human; data analytics

INTRODUCTION

Infectious disease refers to an illness caused by the transmission of specific infectious agents, either directly or indirectly, from an infected individual to another. All types of bacteria, viruses, parasites, fungi, and proteins that are capable of producing an infection are called as infectious agents. The term *endemic* and *epidemic* is used to differentiate between the diseases that continuously occur from the diseases that unexpectedly occur in a population. The

epidemic that involves different countries and a large population is called *pandemic*, while the epidemic that is restricted to a small population is called an *outbreak* (Baretto et al., 2006).

The earliest recorded epidemics dated as far back as the ancient Greece and Egypt periods, with the fatal incidences that greatly affected the population. Since the discovery of the first smallpox vaccine in 1796 and the development of new vaccines in 20th century, the rates of global fatalities from infectious diseases have been greatly reduced. The Global Immunization Vision and Strategy (GIVS) initiative published by the World Health Organization (WHO) and United Nations Children's Fund (UNICEF) in 2005 aims at further increasing national vaccination coverage by at least 90% and reducing global childhood morbidity and mortality by at least two thirds compared to year 2000 (WHO, UNICEF & World Bank, 2009).

However, the emergence of infectious diseases in recent years may jeopardize GIVS goals. Diseases that have newly appeared in a population or may have existed in the past but reappear are categorized as emerging infectious diseases. Dengue fever and measles is among the common emerging diseases that have caused severe outbreaks in southeast Asia (Coker et al., 2011). Therefore, an early warning mechanism needs to be established to help public health decision makers in controlling potential future outbreaks.

Mathematical and statistical model has long been used to discover patterns and predict future outbreak trends. Susceptible-Infected-Recovered (SIR) model is the simplest and most fundamental model that divided the population based on their infection status (Keeling & Danon, 2009). The susceptible class represents those that have not yet been infected and can conceive the disease. The infected class represents those that have been infected and can spread the disease. The recovered class is those that has obtained permanent immunity when they have recovered from the disease or died.

Another model, Susceptible-Infected-Susceptible (SIS) model considers that the recovered population gained no immunity and can rejoin the susceptible population (Bentil, 2013). Recent studies by Fred et al. (2014), Glover (2015) and Bier & Brak (2015) implemented the extension of the basic model, Susceptible-Exposed-Infected-Recovered (SEIR) model which includes the individuals who have been exposed but are not yet able to spread the disease. Although these models are said to be helpful in modelling disease outbreak, Huppert & Katriel (2013) and De Angelis et al. (2015) recommended further experimental studies on the mechanisms of combining datasets from multiple sources of information as single source datasets seldom gave an informative output needed for policy-making.

Datasets that comes from multiple sources and exceed the processing capability of conventional database systems are presently referred as big data (Asokan & Asokan, 2015). By definition, big data comprises of structured and unstructured data described in terms of its volume, variety, velocity, veracity, variability, and value (Gandomi & Haider, 2015). In healthcare, big data is characterized not only by volume, but also on the data types diversity, speed, and authenticity of the data (Raghupathi & Raghupathi, 2014).

According to Kambatla et al. (2014), Raghupathi & Raghupathi (2014), Amankwah-Amoah (2015) and Wang et al. (2016), healthcare big data can come from multiple sources, locations and formats as well as a mix of following data types:

1. Structured data: i.e. patients-related data from electronic medical records (EMR) and electronic health records (EHR).
2. Semi-structured data: i.e. health monitoring devices logs.
3. Unstructured data: i.e. clinical images; vital sign devices; web and social media data from Twitter, Facebook status updates, blogs, newsfeeds, webpages, and search engines queries; satellite data; emergency care data; physician notes; articles in medical journals.

In this paper, a systematic review of existing literatures was performed mainly to identify the factors that have been used in infectious disease outbreak prediction specifically dengue fever and measles. The paper has two objectives: (1) to determine the data types, sources and factors currently used in both dengue fever and measles outbreak prediction, and (2) to identify gaps and limitations for future research.

METHODS

The systematic review was performed using a basic systematic literature review (SLR) approach as described by Kitchenham & Charters (2007). The literature search was conducted using four databases: PubMed, ScienceDirect, Scopus and Google Scholar. The following keywords were used in the search process: (dengue OR measles AND outbreak AND prediction).

The main inclusion criterion for this study is to include the factors of emerging infectious disease outbreak predictions practice in previous articles. Peer-reviewed articles published from 2006-2016 are taken into consideration for the inclusion in search criteria. The detail inclusion criteria included are: (1) studies that investigates infectious disease outbreak prediction related to dengue fever or measles, (2) studies that states its data sources for the outbreak prediction, and (3) studies that uses various factors for the outbreak prediction.

Meanwhile, the articles that are excluded from our research criteria are studies that (1) focus on causal factors and treatment of dengue fever or measles, (2) discuss dengue fever or measles in general, and (3) investigate other types of infectious disease.

Figure 1 below shows the SLR flow diagram:

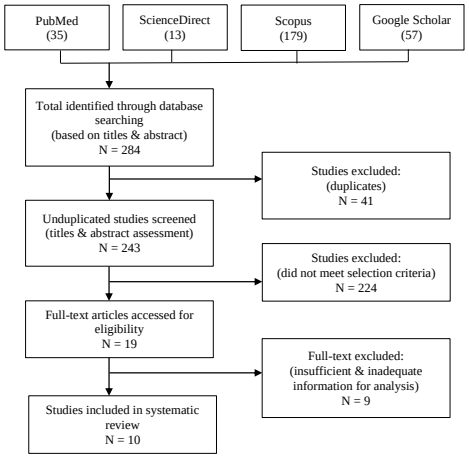

Figure 1. SLR Flow Diagram.

A total of 284 articles were identified based on the keywords search of its title and abstract: 35 articles were derived from PubMed, 13 articles were derived from ScienceDirect, 179 articles were derived from Scopus, and 57 papers were derived from Google Scholar. All articles were checked for duplication, resulting in 243 different articles. Subsequently, these articles were further screened by examining the content of their title and abstract. 224 articles that did not meet the selection criteria were excluded accordingly. The remaining 19 articles were accessed for eligibility, and 10 articles that sufficiently provide

information of its data source and used distinctive factors for prediction were selected for the review.

RESULTS & DISCUSSION

The review found 196 articles of studies associated to dengue fever and 47 articles related to measles. There are extensive research covering on the causal factors, treatment, surveillance, contingency plan, and outbreak prediction of dengue fever. In contrast for measles, limited research was carried out especially on the prediction of measles outbreak.

Table 1 and Table 2 show the comparison of data types, sources and factors used in studies related to both dengue fever and measles outbreak prediction. The studies on dengue fever outbreak prediction utilized a combination of structured and unstructured data from multiple sources. The prediction output for dengue fever can be produced by incorporating the meteorological data, laboratory data, entomological data, and socio-economic data.

Table 1. Data Types, Sources and Factors in Dengue Fever Outbreak Prediction.

Reference	Data types	Data sources	Factors
Ramachandran et al. (2016)	Structured & Unstructured	Meteorological department	• Rainfall data • Temperature data • Humidity data
		Hospital laboratory	• Number of dengue cases
Siriyasatien et al. (2016)	Structured & Unstructured	Meteorological department	• Season • Rainfall data • Temperature data • Humidity data • Wind speed
		Entomological department	• *Aedes Aegypti* infection rate
		Socio-economic department	• Population density
		Hospital laboratory	• Number of dengue cases
Ali et al. (2016)	Structured & unstructured	Meteorological department	• Rainfall data • Temperature data • Humidity data
		Hospital emergency department	• Ambulatory records • Drug sales • Emergency calls
		Socio-economic department	• Geographic information • Population density
Chan et al. (2015)	Structured & unstructured	Meteorological department	• Rainfall data • Temperature data • Humidity data
		Socio-economic department	• Population density
		Hospital laboratory	• Number of dengue cases
Buczak et al. (2014)	Structured & unstructured	Meteorological department	• Rainfall data • Temperature data • Vegetation index • Sea surface temperature • Southern Oscillation Index
		Socio-economic department	• Political stability • Sanitation

			• Water • Electricity
		Hospital laboratory	• Number of dengue cases

The studies on measles outbreak prediction, however, only utilized structured data and limited source of clinical and patient data collected through electronic health records. The prediction output for measles in existing literatures is produced based on the infected patients' information such as age, gender, vaccination status, and residence area.

Table 2. Data Types, Sources and Factors in Measles Outbreak Prediction.

Reference	Data types	Data sources	Factors
Pinchoff et al. (2015)	Structured	Electronic health records	• Hospitalization date • Vaccination status • Residence area • HIV infection status
Wood et al. (2015)	Structured	Electronic health records	• Vaccination status
Bier & Brak (2015)	Structured	Electronic health records	• Vaccination status
Assamnew (2011)	Structured	Electronic health records	• Age • Gender • Residence area • Vaccination status • Laboratory test
Salim et al. (2007)	Structured	Electronic health records	• Vaccination status • Nutritional status

CONCLUSION

This paper focuses on a systematic review of dengue fever and measles outbreak prediction specifically on the data types, sources and factors used for the prediction. The review found that limited studies have been done on measles outbreak prediction while the existing studies of measles outbreak does not use datasets from variety of sources. Since single source data is unreliable, further research should focus on the integration of structured and unstructured data for an accurate measles outbreak prediction.

REFERENCES

Ali, M. A., Ahsan, Z., Amin, M., Latif, S., Ayyaz, A., & Ayyaz, M. N. (2016). ID-Viewer: A visual analytics architecture for infectious diseases surveillance and response management in Pakistan. *Public health*, 134, 72-85.

Amankwah-Amoah, J. (2015). *Emerging economies, emerging challenges: Mobilising and capturing value from big data.* Technological Forecasting and Social Change.

Assamnew, S. (2011). Predicting the occurrence of measles outbreak in Ethiopia using data mining technology. *Doctoral dissertation*, AAU.

Asokan, G. V., & Asokan, V. (2015). Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. *Journal of Epidemiology and Global Health*, 5(4), 311-314.

Barreto, M. L., Teixeira, M. G., & Carmo, E. H. (2006). Infectious diseases epidemiology. *Journal of Epidemiology and Community Health*, 60(3), 192-195.

Bentil, I. (2013). A comparative analysis on the mathematical models of pertussis and measles. *Doctoral dissertation*, Institute of Distance Learning, Kwame Nkrumah University of Science and Technology.

Bier, M., & Brak, B. (2015). A simple model to quantitatively account for periodic outbreaks of the measles in the Dutch Bible Belt. *The European Physical Journal B*, 88(4), 1-11.

Buczak, A. L., Baugher, B., Babin, S. M., Ramac-Thomas, L. C., Guven, E., Elbert, Y., & Yoon, I. K. (2014). Prediction of high incidence of dengue in the Philippines. *PLoS Negl Trop Dis*, 8(4), e2771.

Chan, T. C., Hu, T. H., & Hwang, J. S. (2015). Daily forecast of dengue fever incidents for urban villages in a city. *International Journal of Health Geographics*, 14(1), 1.

Coker, R. J., Hunter, B. M., Rudge, J. W., Liverani, M., & Hanvoravongchai, P. (2011). Emerging infectious diseases in southeast Asia: Regional challenges to control. *The Lancet*, 377(9765), 599-609.

De Angelis, D., Presanis, A. M., Birrell, P. J., Tomba, G. S., & House, T. (2015). Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, 10, 83-87.

Fred, M. O., Sigey, J.K., Okello J.A., Okwoyo J.M. & Kang'ethe G.J. (2014). Mathematical modeling on the control of measles by vaccination: Case study of Kisii County, Kenya. The SIJ Transactions on *Computer Science Engineering & its Applications (CSEA)*, 2(3), 61-69.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144.

Glover, C. E. (2015). A mathematical model of the 2014 Ohio measles outbreak to assess the effectiveness of the public health. *Doctoral dissertation*, The Ohio State University.

Huppert, A., & Katriel, G. (2013). Mathematical modelling and prediction in infectious disease epidemiology. *Clinical Microbiology and Infection*, 19(11), 999-1005.

Keeling, M. J., & Danon, L. (2009). Mathematical modelling of infectious diseases. *British Medical Bulletin*, 92(1), 33-42.

Kitchenham, B., & Charters, S. Guidelines for performing systematic literature reviews in software engineering. In Technical report, Ver. 2.3 *EBSE Technical Report*. EBSE.

Nelson, K. E., & Williams, C. (2014). *Infectious disease epidemiology*. Jones & Bartlett Publishers.

Pinchoff, J., Chipeta, J., Banda, G. C., Miti, S., Shields, T., Curriero, F., & Moss, W. J. (2015). Spatial clustering of measles cases during endemic (1998–2002) and epidemic (2010) periods in Lusaka, Zambia. *BMC infectious diseases*, 15(1), 1.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: *Promise and potential. Health Information Science and Systems*, 2(1), 3.

Ramachandran, V. G., Roy, P., Das, S., Mogha, N. S., & Bansal, A. K. (2016). Empirical model for calculating dengue incidence using temperature, rainfall and relative humidity: A 19-year retrospective analysis in East Delhi, India. *Epidemiology and health.*

Ramadona, A. L., Lazuardi, L., Hii, Y. L., Holmner, Å., Kusnanto, H., & Rocklöv, J. (2016). Prediction of dengue outbreaks based on disease surveillance and meteorological data. *PloS One*, 11(3), e0152688.

Salim, A., Hari Basuki, N., & Syahrul, F. (2007). Indikator prediksi kejadian luar biasa (KLB) campak di Provinsi Jawa Barat. *The Indonesian Journal of Public Health*, 4(3), 112-116.

Siriyasatien, P., Phumee, A., Ongruk, P., Jampachaisri, K., & Kesorn, K. (2016). Analysis of significant factors for dengue fever incidence prediction. *BMC bioinformatics*, 17(1), 1.

Wang, Y., Kung, L., & Byrd, T. A. (2016). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*.

WHO, UNICEF & World Bank. (2009) State of the world's vaccines and immunization. 3rd ed. Geneva, *World Health Organization.*

Wood, J. G., Heywood, A. E., Menzies, R. I., McIntyre, P. B., & MacIntyre, C. R. (2015). Predicting localised measles outbreak potential in Australia. *Vaccine*, 33(9), 1176-1181.