Infection, Disease & Health (2019) 24, 44–48

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.journals.elsevier.com/infection-disease-and-health/

Discussion paper

Artificial Intelligence for infectious disease Big Data Analytics

Zoie S.Y. Wong [a,*], Jiaqi Zhou [b], Qingpeng Zhang [b]

[a] *Graduate School of Public Health, St. Luke's International University, Tokyo, 104-0045, Japan*
[b] *Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China*

Received 4 July 2018; received in revised form 3 October 2018; accepted 8 October 2018
Available online 2 November 2018

KEYWORDS
Infectious diseases modelling;
Emergency response;
Artificial Intelligence;
Machine learning

Abstract *Background:* Since the beginning of the 21st century, the amount of data obtained from public health surveillance has increased dramatically due to the advancement of information and communications technology and the data collection systems now in place.
Methods: This paper aims to highlight the opportunities gained through the use of Artificial Intelligence (AI) methods to enable reliable disease-oriented monitoring and projection in this information age.
Results and Conclusion: It is foreseeable that together with reliable data management platforms AI methods will enable analysis of massive infectious disease and surveillance data effectively to support government agencies, healthcare service providers, and medical professionals to response to disease in the future.

Highlights

- This paper provides discussion and visionary perspective on data usage and management for infectious diseases.
- It highlights the opportunity to make use of Artificial Intelligence (AI) methods to enable reliable and data-oriented disease monitoring and projection under this information age.
- It is foreseeable that together with reliable data management platforms AI methods will enable effective analysis of massive infectious disease and surveillance data to support risk and resource analysis for government agencies, healthcare service providers, and medical professionals in the future.

* Corresponding author.
E-mail address: zoiesywong@gmail.com (Z.S.Y. Wong).

https://doi.org/10.1016/j.idh.2018.10.002

Introduction

Computer scientists have coined the term "Big Data" to refer to the volume of ever-growing data, which is not only large in size and variety, but can be harnessed for important purposes. "Big Data" refers to data sets that are so massive or complex in nature that traditional data processing methods and methodology are inadequate to effectively analyze it. Big Data usually exhibits properties of high volume (in terms of the amount of data), high velocity (in terms of the speed of data in and out), high variety (in terms of the range of data types and sources), and high veracity (in terms of accuracy and correctness). "Big Data Analytics" refers to the process, methods and technologies to analyze such Big Data and uncover valuable information using non-traditional advanced methods [1,2]. It is designed in a way that advances accuracy and scalability compared to traditional standard methods, such as regression-based models and other statistical models. Among such advanced methods, Artificial Intelligence (AI) has also been identified as an especially important development regarding the roles it can play in various application domains [3], including public health related disciplines [4].

Problem statement

Thanks to the advancement of information and communication technology (ICT) [5], in the 21st Century we are able to access massive and potentially useful infectious diseases transmission related data through multiple channels, including sentinel reporting systems, national syndromic surveillance systems (usually operated by national and regional disease centers, such as [6,7]), genome databases, internet search key trends [8], Twitter data [9], outbreak investigation reports, transportation dynamics [10], vaccinology related data [11] and human dynamics information [12,13] from various transport authorities or national surveys. Much of the abovementioned data is massive in size and usually regarded as "dirty data", which implies that a large fraction of data is irrelevant to the specific topic of interest (e.g. a certain disease). As a result, data quality varies widely across disparate data sources. Current research efforts attempt to use mainstream primary disease-related data. Yet, a massive volume of big disease data remains unexplored, awaiting further technological or scientific advances in order to devise effective infectious disease management strategies.

Current literatures and the contribution of this paper

Recently the success of AI applications in mastering human cognitive intelligence problems has received wide attention from the public [3]. However, the grand challenge in infectious diseases field remains how to process the best available data in an efficient and timely manner, while maintaining a high level of coherence among the information collected [14]. It is about how to manage an ever-growing amount of diseases indicators, data and information, and how to assure their best utilization in order to aid better decision making and disease response measures [15].

This paper aims to highlight the future roles of AI methods in enabling reliable disease monitoring and projection under this Big Data information age.

Challenges to use Artificial Intelligence for infectious diseases applications

It is challenging to apply conventional statistical methods to medical data with large dimensionality (such as textual and image data). In many cases, the underlying assumptions of traditional statistical-based methods are breached when we are dealing with high dimension data with complexity. Most classic statistical methods are unable to handle Big Data problems, because emerging infectious disease outbreak related data exists in various formats and exhibits different limitations and assumptions.

Big Data, AI and machine learning methods have so far shown promising outcomes in multiple business and industrial settings to reveal hidden patterns and predict future possibilities. The latest advanced models, such as deep learning artificial neutral network approaches, have shown positive results in extracting highly nonlinear structures from a massive dataset. In a disease related context, recent research using AI methods to track down rodent reservoirs of future zoonotic diseases [16], predict Extended-spectrum β-lactamase (ESBL) producing organisms [17], and control tuberculosis (TB) and gonorrhea disease spreads [18], has been in place. Public reaction towards disease outbreaks can be difficult to predict. However, with the availability of Big Data and advent of AI methods, we are increasingly able to correlate the population behavior with disease outbreaks. For instance, using a behavioral informatics and analytics approach, dedicated researchers [19—21] investigated infectious disease outbreaks associated digital behavior patterns through web search patterns (such as Google Trend) during the outbreak period. It is envisioned that the advancement of AI methods for infectious disease Big Data analytics could improve our ability to observe public reactions on disease outbreaks dynamically and predict disease spread accurately, which can help authorities take timely response measures to infectious diseases.

Multiple sources of data

For effective data integration, management, and knowledge extraction, a comprehensive data management system is needed. Epidemic modelling methods have already played an important role in disease transmission propagation and policy advice [22]. Epidemic models typically categorise into phenomenological and mechanistic modelling studies [23], where a mechanistic model may be designed in different forms, such as a simple homogeneous compartmental model or with the consideration of host heterogeneities, multi-pathogen and multi-host situation, temporally forced model, spatial model etc. [24]. It is anticipated in the coming future that AI enhanced methods will also play a role in disease preparedness. Fig. 1 illustrates our vision to take advantage of shared data

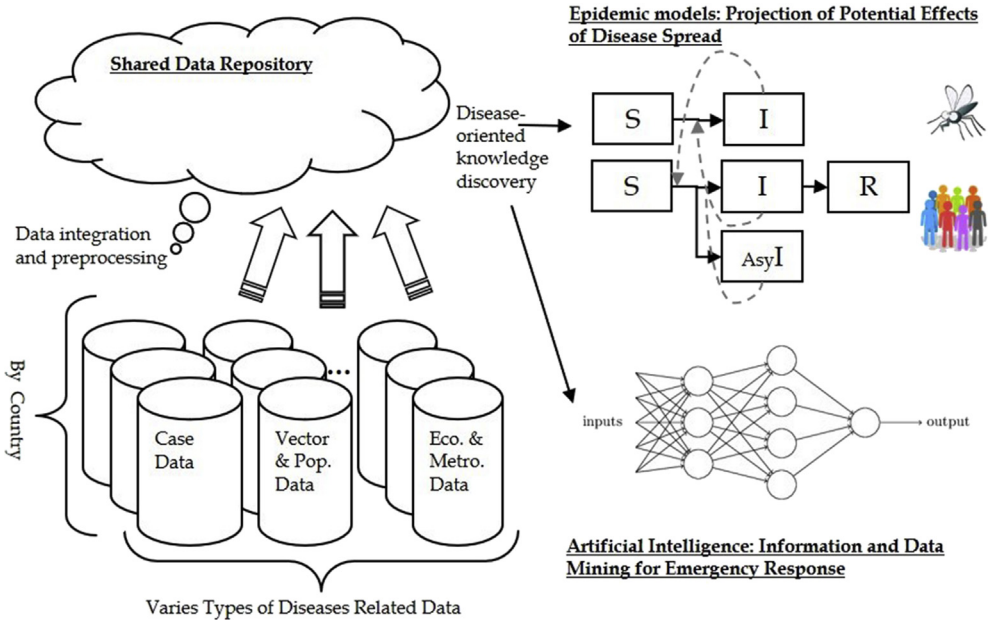

Fig. 1 An integrated big data conceptual model for infectious diseases.

repository and AI methods for an integrated disease response and preparedness approach.

Malaria elimination example

Here we use Malaria elimination to illustrate our arguments. Ongoing malaria indicators (including malaria cases (local and imported), prevalence, bed net usage, and population immunity) have been captured by various health authorities across the world. Many high quality open access malaria disease related data sources are made available in public repositories, for instance, the Malaria Atlas Project [25], the Malaria Immunology Database [26], the Mapping Malaria Risk in Africa projects [27], and PlasmoDB [28]. There are multiple data sources in play within each data source: there are usually multiple series and many of these series are further broken down into sub-series. Currently, most series are typically monitored in a univariate fashion, after which multiple detection algorithms are applied. Traditional epidemic models have shown encouraging results in devising malaria control strategies [29]. By integrating a disease simulation model, the University of Oxford team has demonstrated early success in using AI algorithms to search for the optimal Malaria intervention strategies [30]. Bent et al. [30] incorporated stochastic simulation using Gaussian Process regression to account for the spread of malaria in a Western Kenya population. The authors have acknowledged the possibility of incorporating AI methods (such as deep learning) in optimizing policy deployment. For instance, to carry out functional approximation and simulation parameters optimization tasks to learn the optimal strategies for malaria control. Bioinformatics methods using advanced data mining methods have also demonstrated great potential in frontier Malaria research using parasite genome sequences [31]. It is hoping that advanced next-generation models can further deepen our understanding of the observed disease mechanism associated with environment and be incorporated into such kinds of novel simulation exploration technique. While human populations are constantly facing challenges from emerging and reemerging diseases, a comprehensive and integrated approach for data usage with the optimum use of appropriate methods will be essential in order to incrementally build up future capability.

It is foreseen that an advanced data warehouse and knowledge discovery platform based on the properties of multiple disease data streams can be created. Currently, a wide range of malaria related datasets, such as case detection system, malariometric study, human travel patterns, population denominators, and interventions efforts, have been gathered separately. Some prototype shared data repositories to enable the discovery of, access to, and integration of datasets are now available [32]. It is essential for future research to devise a data capturing plan that comprehensively includes and integrates multiple associated data streams [9] — for data mining, knowledge discovery, surveillance, and outbreak detection.

AI methods applicable to infectious disease data

AI and machine learning methods could be summarized into two categories based on two different learning strategies: supervised learning and unsupervised learning. Supervised learning is the task of inferring a function from labeled training data, including Support Vector Machine (SVM), Decision Tree, Random Forest, Naïve Bayes (NB), Artificial Neural Network (ANN), Bootstrap Aggregating, AdaBoost, which could handle the classification and regression problem in medical data efficiently [33,34]. All these methods could be useful in improving the accuracy of diagnosis and suggesting appropriate treatment for patients. Additionally, the prediction outcomes from these methods could be

applied to warn the authorities and the public in advances and suggest appropriate prevention and control strategies. According to different aims, unsupervised learning methods such as Principal Component Analysis (PCA) could be used to reduce the dimensions of data, which would make it easier for researchers to uncover some key factors related to infectious disease [35]. Other unsupervised learning methods, such as K-means, could cluster patients into subgroups and detect abnormal patients, which could guide researchers to focus on these medical cases. In addition, topic models such as Latent Dirichlet allocation (LDA) could be used to extract the topics from medical textual record. Recently deep learning architectures have been broadly applied into prediction and classification, social network filtering, and bioinformatics and are deemed to be powerful tools for infectious disease analytics.

Conclusion

Since the beginning of the 21st century, the amount of data obtained from public health surveillance has increased dramatically with the advancement of information and communications technology, and new data collection systems now in place. It is foreseeable that machine learning and data management platforms will enable analysis of multiple infectious disease outbreaks. The resultant improvements in syndromic surveillance will be of interest for risk analysis and resource deployment by government agencies, healthcare service providers, and medical professionals.

Future research efforts should focus on investigating infectious disease Big Data integration and knowledge discovery, as well as modelling techniques for effective infectious disease monitoring using multiple data sources and trustworthy surveillance methods [36,37]. It is envisioned that the development of a Big Data Analytic approach for disease modeling involves cross-disciplinary efforts to utilize well-established theories and methodologies from computer science, information technology, disease modelling, and disease epidemiology.

In the future, can we attempt to answer ambitious question: How could we make use of recent advances in shared platforms to facilitate data usage for focused infectious disease epidemiology query? Yet, two major challenges related to Big Data Analytics in the application to diseases control need to be considered:

- Diverse data challenges: how to control, manage, and manipulate the multiple data sources relating to infectious disease epidemiology, known and uncertain disease characteristics and the recent advance of data management methods.
- Methodological challenges: how to best integrate advanced data-mining, disease modelling and syndromic surveillance methods to maximize our understanding of emerging and reemerging diseases based on the multiple data sources.

Several key limitations and concerns in utilizing AI models for enhancing decision-making relating to disease preparedness should also be taken into consideration. First, every single infectious disease exhibits unique natural characteristics (for instance: transmission route, infectivity, incubation period). Second, our level of understanding of a new emerging infectious disease may be limited at the early phase of outbreak. Third, disease-related data may reside in different formats, which would require substantial information extraction efforts. AI algorithms may require specific calibration to disease-specific scenarios. In the other words, a one-size-fit-all approach [5] may not be applicable across all diseases contexts.

Ethics

Ethics is not applicable as this is a discussion paper.

Authorship statement

The study was conceived by ZW. ZW, JZ and QZ contributed to the study design. ZW led the writing of the paper and all authors revised and refined the arguments. All authors approved the article.

Conflict of interest

All authors do not have any conflict of interest to declare.

Funding

This work was supported by Japan Society for the Promotion of Science KAKENHI (Grant number 18H03336), the Research Grants Council Theme-Based Research Scheme (Ref.: T32-102/14N) and The National Natural Science Foundation of China (NSFC) Grant Nos. 71402157 and 71672163.

Provenance and peer review

Not commissioned; externally peer reviewed.

References

[1] Schneeweiss S. Learning from big health care data. N Engl J Med 2014;370(23):2161—3.
[2] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3.
[3] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017;550(7676):354—9.
[4] IBM. Watson for oncology. 2017 [cited 2017 15 Sept]. Available from: https://www.ibm.com/watson/health/oncology-and-genomics/oncology/.
[5] Wong ZS, Nohr C, Kuziemsky CE, Leung E, Chen F. Context sensitive health informatics: delivering 21st century healthcare — building a quality-and-efficiency driven system. Stud Health Technol Inf 2017;241:1—5.
[6] CDC. Data & statistics. 2018 [cited 2018 22 Feb]. Available from: https://www.cdc.gov/datastatistics/index.html.
[7] UKgovernment. Health promotion. Infectious diseases. 2018 [cited 2018 22 Feb]. Available from: https://www.gov.uk/topic/health-protection/infectious-diseases.

[8] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012—4.
[9] Gianfredi V, Bragazzi NL, Nucci D, Martini M, Rosselli R, Minelli L, et al. Harnessing big data for communicable tropical and sub-tropical disorders: implications from a systematic review of the literature. Front Public Health 2018;6:90.
[10] Air Transport Statistics. International air transport association (IATA). 2018 [cited 2018 22 Feb]. Available from: http://www.iata.org/services/statistics/air-transport-stats/Pages/index.aspx.
[11] Bragazzi NL, Gianfredi V, Villarini M, Rosselli R, Nasr A, Hussein A, et al. Vaccines meet big data: state-of-the-art and future prospects. From the classical 3Is ("Isolate-Inactivate-Inject") vaccinology 1.0 to vaccinology 3.0, vaccinomics, and beyond: a historical overview. Front Public Health 2018;6:62.
[12] HILDA survey. 2018 [cited 2018 22 Feb]. Available from: http://melbourneinstitute.unimelb.edu.au/hilda.
[13] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 2008;5(3):e74.
[14] Hoyt RE, Yoshihashi AK. Health informatics: practical guide for healthcare and information technology professionals. 6th ed. Lulu.com; 2014.
[15] Tsui K, Wong Z, Goldsman D, Edesess M. Tracking infectious disease spread for global pandemic containment. Intell Syst IEEE 2013;28(6):60—4.
[16] Han BA, Schmidt JP, Bowden SE, Drake JM. Rodent reservoirs of future zoonotic diseases. Proc Natl Acad Sci Unit States Am 2015;112(22):7039—44.
[17] Goodman KE, Lessler J, Cosgrove SE, Harris AD, Lautenbach E, Han JH, et al. A clinical decision tree to predict whether a bacteremic patient is infected with an extended-spectrum beta-lactamase-producing organism. Clin Infect Dis Off Pub Infect Dis Soc Am 2016;63(7):896—903.
[18] Blumenthal A. Artificial Intelligence to fight the spread of infectious diseases. Phys Org 2018 20 Feb [cited 2018 25 Oct]. Available from: https://phys.org/news/2018-02-artificial-intelligence-infectious-diseases.html.
[19] Bragazzi NL, Alicino C, Trucchi C, Paganino C, Barberis I, Martini M, et al. Global reaction to the recent outbreaks of Zika virus: insights from a Big Data analysis. PloS One 2017;12(9), e0185263.
[20] Alicino C, Bragazzi NL, Faccio V, Amicizia D, Panatto D, Gasparini R, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty 2015;4:54.
[21] Mahroum N, Adawi M, Sharif K, Waknin R, Mahagna H, Bisharat B, et al. Public reaction to Chikungunya outbreaks in Italy-Insights from an extensive novel data streams-based structural equation modeling analysis. PloS One 2018;13(5), e0197337.
[22] Heesterbeek H, Anderson RM, Andreasen V, Bansal S, De Angelis D, Dye C, et al. Modeling infectious disease dynamics in the complex landscape of global health. Science (New York, NY) 2015;347(6227). aaa4339.
[23] Wong ZSY, Bui CM, Chughtai AA, Macintyre CR. A systematic review of early modelling studies of Ebola virus disease in West Africa. Epidemiol Infect 2017:1—26.
[24] Keeling MJ. In: Rohani P, editor. Modeling infectious diseases in humans and animals. Woodstock: Princeton University Press; 2008.
[25] Moyes CL, Temperley WH, Henry AJ, Burgert CR, Hay SI. Providing open access data online to advance malaria research and control. Malar J 2013;12(1):161.
[26] Deroost K, Opdenakker G, Van den Steen PE. MalarImDB: an open-access literature-based malaria immunology database. Trends Parasitol 2014;30(6):309—16.
[27] Mapping malaria risk in Africa (MARA) database. 2009. Available from: http://www.mara-database.org//login.html;jsessionid=571B54B64F54C08C6DC0306A2BF3360B.
[28] Palsnodium genomics resource. 2018. Available from: http://plasmodb.org/plasmo/.
[29] Okell LC, Drakeley CJ, Bousema T, Whitty CJ, Ghani AC. Modelling the impact of artemisinin combination therapy and long-acting treatments on malaria transmission intensity. PLoS Med 2008;5.
[30] Bent ORS, Roberts S, Walcott-Bryant A. Novel exploration techniques (NETs) for malaria policy interventions. 2018.
[31] Izak D, Klim J, Kaczanowski S. Host-parasite interactions and ecology of the malaria parasite-a bioinformatics approach. Briefings Funct Genom 2018:1—7.
[32] Wan J, Zou C, Ullah S, Lai CF, Zhou M, Wang X. Cloud-enabled wireless body area networks for pervasive healthcare. IEEE Network 2013;27(5):56—61.
[33] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23(1):89—109.
[34] Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research — commentary. Biomed Eng Online 2014;13:94.
[35] Wang Y, Fan Y, Bhatt P, Davatzikos C. High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. Neuroimage 2010;50(4):1519—35.
[36] Rolka H, Burkom H, Cooper GF, Kulldorff M, Madigan D, Wong WK. Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs. Stat Med 2007;26(8):1834—56.
[37] Tsui K-L, Wong SY, Jiang W, Lin C-J. Recent research and developments in temporal and spatiotemporal surveillance for public health. IEEE Trans Reliab 2011;60(1):49—58.