Artificial Intelligence in Public Health and Epidemiology

Rodolphe Thiébaut[1,2,3], Frantz Thiessard[1,2], Section Editors for the IMIA Yearbook Section on Public Health and Epidemiology Informatics

[1] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, F-33000 Bordeaux, France
[2] Centre Hospitalier Universitaire de Bordeaux, Service d'Information Médicale, F-33000 Bordeaux, France
[3] Inria, SISTM, F-33400 Talence, France

Summary

Objectives: To introduce and summarize current research in the field of Public Health and Epidemiology Informatics.

Methods: The 2017 literature concerning public health and epidemiology informatics was searched in PubMed and Web of Science, and the returned references were reviewed by the two section editors to select 14 candidate best papers. These papers were then peer-reviewed by external reviewers to provide the editorial team with an enlightened vision to select the best papers.

Results: Among the 843 references retrieved from PubMed and Web of Science, two were finally selected as best papers. The first one analyzes the relationship between the disease, social/mass media, and public emotions to understand public overre-action (leading to a noticeable reduction of social and economic activities) in the context of a nation-wide outbreak of Middle East Respiratory Syndrome (MERS) in Korea in 2015. The second paper concerns a new methodology to de-identify patient notes in electronic health records based on artificial neural networks that outperformed existing methods.

Conclusions: Surveillance is still a productive topic in public health informatics but other very important topics in Public Health are appearing. For example, the use of artificial intelligence approaches is increasing.

Keywords

Public health, epidemiology; medical informatics; International Medical Informatics Association; health information systems; artificial intelligence.

Introduction

As quoted in the synopsis of the Public Health and Epidemiology Informatics section of the 2017 IMIA Yearbook [1], precision public/global health and digital epidemiology are terms that are still in use in 2018 [2,3]. The first term is about providing the right intervention to the right population at the right time [2]. The second term is about the use of digital data, especially those that were not collected on purpose, to answer epidemiologic questions [3]. Both refer to the unforeseen opportunities provided by our digital world and new technologies. Although genomics (and more broadly any "-omics") data continue to contribute, as it is the case for precision medicine, there are many other sources of information that can be used: social networks, internet search engines, cell phone data, electronic health data, and more. The challenge today is to analyze these big data in a meaningful way. One recently improved method that showed very nice success especially in image analysis is deep learning [4]. Applications of this method appear to be only limited by the quantity of information available. Predicting the unplanned readmission at the hospital within 6 months based on electronic health data [5], de-identifying electronic health records (EHRs) [6], analyzing social media [7–9] are various types of applications relevant in epidemiology and public health. But artificial intelligence covers many other techniques, such as machine learning approaches and statistical learning that offer a panel of methods which usefulness is only limited by pairing them with the right question; the two best papers of this year section are very good examples [6,7]. Naïvely mining any large dataset will not give immediate answers. Epidemiologic approaches start with clever and appropriate questions, careful collection of relevant data with the most appropriate design, and validation of the results.

Paper Selection

A comprehensive literature search was performed using two bibliographic databases, Pubmed/Medline (from NCBI, National Center for Biotechnology Information), and Web of Science® (from Thomson Reuters). The search was targeted at public health and epidemiology papers that involve computer science or the massive amount of web-generated data. References addressing the topics of the other sections of the Yearbook, such as those related to interoperability between data providers were excluded from our search. The study was performed at the beginning of January 2018, and the search over the year 2017 returned a total of 843 references.

Articles were separately reviewed by the two section editors, and were first classified into three categories: keep, discard, or leave pending. Then, the "keep" and "leave pending" lists of references built by the two section editors were merged, yielding 97 references. The two section editors jointly

reviewed the 97 references and drafted a consensual list of 14 candidate best papers. All pre-selected 14 papers were then peer-reviewed by Yearbook editors and external reviewers (at least four reviewers per paper). Two papers were finally selected as best papers (Table 1). A content summary of these selected papers can be found in the appendix of this synopsis. Lamy et al. [10] describe the whole selection process.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2018 in the section 'Public Health and Epidemiology Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Public Health and Epidemiology Informatics
▪ Choi S, Lee J, Kang MG, Min H, Chang YS, Yoon S. Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. Methods (2017) 129:50–59.
▪ Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc (2017) 24:596–606.

Outlook and Conclusion

As expected in this section of the Yearbook, the use of digital sources for infectious diseases surveillance leads to many research reports [11]. The originality here is the use of data coming from the EHR [12], Twitter [8], or the climate data produced by the US National Aeronautics and Space Administration (NASA) with software [13] dedicated to following the occurrence of infectious disease epidemics of either influenza [8,12] or malaria [13]. All these studies demonstrating the feasibility of new approaches for the surveillance of infectious diseases still need to be validated for confirming their predictive accuracy and generalizability. Interestingly, we also found studies reporting the results of the surveillance of non-infectious diseases, e.g., road traffic crashes [14] and elevated blood pressure [15]. Road traffic injuries represent a public health issue in low-income countries [16]. Therefore, improvement of surveillance systems is required. Bonnet et al. [14] have experimented a simple affordable approach based on the city's National Police road crash intervention service equipped with geotracers that geolocalized the crash sites and sent their positions by short message service (SMS) to a surveillance platform developed by using the open-source tool, Ushahidi. This system implemented in partnership with the National Police in the city of Ouagadougou required acceptance by officers and authorities. In the other study, the authors showed a new validation of the use of EHR for public health purposes. Here they reproduced the seasonability of blood pressure variations (with a peak in summer) based on the data extracted from EHRs [15].

The other topic covered by several of the papers selected by the review process concerned the analysis of online social media to better understand the attitudes and beliefs toward a given topic, such as vaccination. A social network analysis of Twitter messages ("tweets") revealed a semantic network for positive, negative, and neutral vaccine sentiment [9]. Beyond this type of analysis, it is fruitful to predict and understand the dynamics of vaccinating behavior. In another paper, Pananos et al. modeled the interaction between vaccination decisions and disease dynamics where one influences another in a nonlinear feedback loop [17]. They used the theory of critical transitions to derive indicators that may help public health officials anticipate when resistance to vaccination might develop and intensify. They applied their approach to data from tweets and Google searches around the Disneyland measles outbreak that occurred in 2015 in California [17]. One of the two best papers described below, analyzed the relationship between Middle East Respiratory Syndrome (MERS), mass media, and public emotions during an outbreak in 2015 in Korea [7].

Digital materials such as tweets are also a potential tool for communication in public health [18, 19] with hopefully an improvement of knowledge and attitudes. However, the indicators and the methods to be used for evaluating social media must be adapted to this specific context. Digital tools used in epidemiology need to be validated as any other measure [20].

Last but not least, it is important to question how interventions and new knowledge generate corresponding changes in public health performance. This is where indicators, measures are needed [21].

References

1. Thiébaut R, Thiessard F. Public Health and Epidemiology Informatics. Yearb Med Inform 2017;26:248–50.
2. Flahault A, Geissbuhler A, Guessous I, Guérin PJ, Bolon I, Salathé M, et al. Precision global health in the digital age. Swiss Med Wkly 2017;147:w14423.
3. Salathé M. Digital epidemiology: what is it, and where is it going? Life Sci Soc Policy 2018;14:1–5.
4. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
5. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: A Convolutional Net for Medical Records. IEEE J Biomed Health Inform 2017;21:22–30.
6. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017;24:596–606.
7. Choi S, Lee J, Kang MG, Min H, Chang YS, Yoon S. Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. Methods Inf Med 2017;129:50–9.
8. Kagashe I, Yan Z, Suheryani I. Enhancing seasonal influenza surveillance: Topic analysis of widely used medicinal drugs using twitter data. J Med Internet Res 2017;19:1–14.
9. Kang GJ, Ewing-Nelson SR, Mackey L, Schlitt JT, Marathe A, Abbas KM, et al. Semantic network analysis of vaccine sentiment in online social media. Vaccine 2017;35:3621–38.
10. Lamy J-B, Séroussi B, Griffon N, Kerdelhué G, Jaulent M-C, Bouaud J. Toward a Formalization of the Process to Select IMIA Yearbook Best Papers. Methods Inf Med 2015;54:135–44.
11. Bhattarai AK, Zarrin A, Lee J. Applications of information and communications technologies to public health: A scoping review using the MeSH term "public health informatics." Online J Public

Health Inform 2017;9:e192.

12. Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert ML, Chabot M, Chazard E, et al. Leveraging hospital big data to monitor flu epidemics. Comput Methods Programs Biomed 2018;154:153–60.

13. Merkord CL, Liu Y, Mihretie A, Gebrehiwot T, Awoke W, Bayabil E, et al. Integrating malaria surveillance with climate data for outbreak detection and forecasting: The EPIDEMIA system. Malar J 2017;16:1–15.

14. Bonnet E, Nikiéma A, Traoré Z, Sidbega S, Ridde V. Technological solutions for an effective health surveillance system for road traffic crashes in Burkina Faso. Glob Health Action 2017;10:1295698.

15. Amoah AO, Angell SY, Byrnes-Enoch H, Amirfar S, Maa P, Wang JJ. Bridging the gap between clinical practice and public health: Using EHR data to assess trends in the seasonality of blood-pressure control. Prev Med Reports 2017;6:369–75.

16. Lagarde E. Road Traffic Injury Is an Escalating Burden in Africa and Deserves Proportionate Research Efforts. PLoS Med 2007;4:170.

17. Pananos AD, Bury TM, Wang C, Schonfeld J, Mohanty SP, Nyhan B, et al. Critical dynamics in population vaccinating behavior. Proc Natl Acad Sci 2017;114:201704093.

18. Rabarison KM, Croston MA, Englar NK, Bish CL, Flynn SM, Johnson CC. Measuring Audience Engagement for Public Health Twitter Chats: Insights From #LiveFitNOLA. JMIR Public Health Surveill 2017;3:e34.

19. Gough A, Hunter RF, Ajao O, Jurek A, McKeown G, Hong J, et al. Tweet for Behavior Change: Using Social Media for the Dissemination of Public Health Messages. JMIR Public Health Surveill 2017;3:e14.

20. Margulis A V, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Value of Free-Text Comments for Validating Cancer Cases Using Primary-Care Data in the UK. Epidemiology 2018;29:308-13.

21. Carney TJ, Shea CM. Informatics Metrics and Measures for a Smart Public Health Systems Approach: Information Science Perspective. Comput Math Methods Med 2017;2017:1452415.

Correspondence to:
Rodolphe Thiébaut
Inserm U1219, ISPED, Univ. Bordeaux
146 rue Leo Saignat
33076 Bordeaux cedex, France
Tel: +33 5 57 57 45 21
Fax: +33 5 56 24 00 81
E-mail: rodolphe.thiebaut@u-bordeaux.fr

Appendix: Content Summaries of Selected Best Papers for IMIA Yearbook 2018, Section 'Public Health and Epidemiology Informatics'.

Choi S, Lee J, Kang MG, Min H, Chang YS, Yoon S

Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks

Methods Inf Med 2017;129:50–9

Analyzing digital media for understanding public reaction is a current hot topic in Public Health informatics. In this paper, Choi et al. studied, in the context of a nation-wide outbreak of Middle East respiratory syndrome (MERS) in Korea in 2015, the relationship between the disease, social/mass media, and public emotions. They used a sophisticated approach collecting data from 153 news media in Korea (articles and comments representing 86 millions words), generating a dictionary, and performing data analysis based on statistical learning methods (including latent Dirichlet allocation). Then, they analyzed the interplay of public reaction with the epidemics using transfer entropy. The methodological approach and the results are very interesting with the proposition of a positive feedback loop created between the mass media and public emotion variables. The first result is an objectivation of the high levels of fear and worries when mining social media. The second result is the causal interpretation starting by an overestimation of the lethal rate of MERS that led to a high number of articles in the media which triggered fear in the public. This public reaction likely motivated reporters to write poor papers leading to the positive loop.

Dernoncourt F, Lee JY, Uzuner O, Szolovits P
De-identification of patient notes with recurrent neural networks

J Am Med Inform Assoc 2017;24:596–606

The paper presents a new methodology to de-identify Electronic Health Record (EHR) based on artificial neural networks. EHRs are representing a fabulous opportunity for researchers and investigators but their use needs de-identification, that is leaving out any information about name, address, coordinates... Manual approaches are time-consuming and present a poor reproducibility. Statistical approaches have been tried and compared among which decision trees, support vector machines, conditional random fields. This last method has been compared in the present paper with a completely new approach based on artificial neural network (Long Short Term Memory Recurrent Neural Networks) through an i2b2 challenge. The artificial neural network approach out-performed the previous ones being better at incorporating context and being more flexible to variations inherent in human languages.