Artificial Intelligence for Surveillance in Public Health

Rodolphe Thiébaut[1, 2, 3], Sébastien Cossin[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, Bordeaux, France
[2] Centre Hospitalier Universitaire de Bordeaux, Service d'Information Médicale, Bordeaux, France
[3] Inria, SISTM, Talence, France

Summary

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

Methods: The 2018 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 15 candidate best papers. These papers were then peer-reviewed by external reviewers to give the editorial team an enlightened selection of the best papers.

Results: Among the 805 references retrieved from PubMed and Web of Science, three were finally selected as best papers. All three papers are about surveillance using digital tools. One study is about the surveillance of flu, another about emerging animal infectious diseases and the last one is about foodborne illness. The sources of information are Google news, Twitter, and Yelp restaurant reviews. Machine learning approaches are most often used to detect signals.

Conclusions: Surveillance is a central topic in public health informatics with the growing use of machine learning approaches in regards of the size and complexity of data. The evaluation of the approaches developed remains a serious challenge.

Keywords

Public health, epidemiology, surveillance, medical informatics, International Medical Informatics Association, artificial intelligence

Yearb Med Inform 2019:232-5

http://dx.doi.org/10.1055/s-0039-1677939

Introduction

As compared to 2017 literature analyzed in the Public Health and Epidemiology Informatics section of the International Medical Informatics Association (IMIA) Yearbook [1], in addition to Precision Public Health or Digital epidemiology, a new term has appeared in 2018: infodemiology and infoveillance [2-4]. A large number of the papers published in Public Health informatics is about the epidemiological surveillance based on the new data generated in the current digital era. The papers include the analysis of the massive data from social media (leading to a so-called social sensor) or electronic health records (EHRs). The availability of this data has led to new opportunities to perform passive surveillance. However, this data requires organization to allow an architecture that makes it valuable.

The use of web-based data requires Natural Language Processing (NLP) approaches to extract the information. Electronic health records may also benefit from NLP but the key element is most often the integration of a large volume of structured and unstructured clinical data in a data warehouse. Several solutions are now widely used to construct a data warehouse such as i2b2 [5] or Labkey [6]. Once the architecture for collecting data is ready, signals may be detected by machine learning approaches from standard statistical methods to neural networks. At this stage, the work is far from finished. Actually, a crucial step is the evaluation of the system proposed to perform surveillance. This evaluation is not straightforward as it requests a good reference (*i.e.*, a gold standard) which is rarely perfect and multi-sources of information are often used. Furthermore, the algorithm is most often dynamical because the system learns with the new real-time data collected, which may require repeated evaluation. Hence, as underlined in a systematic review [7], the transfer from research to practice is not obvious, especially because of the challenges underlined above. Public health surveillance is therefore a natural application for artificial intelligence techniques but more work remains to be done by epidemiological scientists to evaluate digital surveillance systems.

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 topics of 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 2019, and the search over the year 2018 returned a total of 805 references.

Articles were separately reviewed by the two section editors, and were first classified into three categories: keep, discard, or leave

pending. Then, the two lists of references were merged, yielding 74 references that were retained by at least one reviewer or classified as "pending" by both of them. The two section editors jointly reviewed the 74 references and drafted an agreed upon list of 15 candidate best papers. All pre-selected 15 papers were then peer-reviewed by both section editors and external reviewers (at least four reviewers per paper). Three papers [8-10] were finally selected as best papers (see Table 1). A content summary of these selected papers can be found in the appendix of this synopsis. The whole selection process has been described by Lamy *et al*. [11].

Outlook and Conclusion

As in 2017, papers published in 2018 and selected by the review were mainly on public health surveillance using EHRs and social media, mainly Twitter. Hospital databases are clearly a source for surveillance which is increasingly considered [12-14]. A good review [15] shows pilot studies of public health surveillance of chronic diseases and risk factors performed in several states of the United States. The topics were type 2 diabetes (based on hemoglobin A1c), pediatric asthma, amyotrophic lateral sclerosis, obesity, and smoking. All these studies constituted a valuable proof of concept but the challenges remain in the definition of the algorithms and their standardization across states and countries. Several papers were about surveillance of seasonal influenza using hospital databases [12, 13, 16]. When several sources were evaluated, it appeared that coupling standard surveillance systems with EHRs constituted the best approach at least for the surveillance of influenza in the United States [16]. It was better than influenza-related search engine and Twitter flu activity social media data [16]. In France, even the single EHR source gave results closer to the reference surveillance system ("Sentinelles" network) than Google data at the national and regional scale [12].

It is interesting to see the spectrum of the outcomes that can be followed using hospital databases. Other published applications of the use of EHRs were the surveillance of antibiotic consumption [17] or hospital acquired infections [14]. Surveillance systems are also proposed to alert health care professionals on a real time basis. The impact of these surveillance techniques should be evaluated in various dimensions. How much is it informative? How valid is the alert? How does it change the practices of healthcare professionals? How does it improve the patient's condition? Hence, the evaluation of a system used in an emergency department did not show any significant impact on the final clinical outcome: the incidence of death [18].

Social media constitutes a source of information for the surveillance of various public health outcomes on a real-time basis such as influenza [9], but also foodborne illnesses [10], or heat alerts [19], which may allow investigation or action in case of alerts, for instance in the context of mass gathering setting [19]. The exploitation of these data needs specific approaches to define the outcomes of interest according to the information available [9, 10] and to extract the signal using various machine learning techniques [12]. Then, the approaches are evaluated by comparison with reference surveillance systems that are often weak gold-standards. The metrics used for these comparisons are usually correlation coefficients [4, 9, 12, 13]. Other metrics such as sensitivity, specificity or accuracy, precision, and recall could be advantageously used in this context to provide a better evaluation of the validity of the surveillance tools [8].

In conclusion, although fantastic opportunities are expected from these new information sources, a lot of work should be done to exploit and validate them.

Acknowledgements
We would like to thank the external reviewers for their participation in the selection process of the Public Health and Epidemiology Informatics section of the IMIA Yearbook.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2019 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
▪ Arsevska E, Valentin S, Rabatel J, de Goër de Hervé J, Falala S, Lancelot R, Roche M. Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. PLoS One 2018 Aug 3;13(8):e0199960.
▪ Effland T, Lawson A, Balter S, Devinney K, Reddy V, Waechter H, Gravano L, Hsu D. Discovering foodborne illness in online restaurant reviews. J Am Med Inform Assoc 2018 Dec 1;25(12):1586-92.
▪ Wakamiya S, Kawai Y, Aramaki E. Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study. JMIR Public Health Surveill 2018 Sep 25;4(3):e65.

References

1. Thiébaut R, Thiessard F. Public Health and Epidemiology Informatics. Yearb Med Inform 2017;26(1):248–50.
2. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009;11:e11.
3. Sciascia S, Radin M, Unlu O, Erkan D, Roccatello D. Infodemiology of antiphospholipid syndrome: Merging informatics and epidemiology. Eur J Rheumatol 2018;5:92–5.
4. Mavragani A, Sampri A, Sypsa K, Tsagarakis KP. Integrating Smart Health in the US Health Care System: Infodemiology Study of Asthma Monitoring in the Google Era. JMIR Public Heal Surveill 2018;4:e24.
5. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc 2006:1040. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17238659 [Accessed May 31, 2019].
6. Nelson EK, Piehler B, Eckels J, Rauch A, Bellew M, Hussey P, et al. LabKey Server: An open source platform for scientific data integration, analysis and collaboration. BMC Bioinformatics 2011;12:71.
7. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM, et al. Using Social Media for Actionable Disease Surveillance and

Outbreak Management: A Systematic Literature Review. PLoS One 2015;10:e0139701.
8. Arsevska E, Valentin S, Rabatel J, de Goër de Hervé J, Falala S, et al. Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. PLoS One 2018;13:e0199960.
9. Wakamiya S, Kawai Y, Aramaki E. Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study. JMIR Public Heal Surveill 2018;4:e65.
10. Effland T, Lawson A, Balter S, Devinney K, Reddy V, Waechter H, et al. Discovering foodborne illness in online restaurant reviews. J Am Med Inform Assoc 2018;25:1586–92.
11. 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.
12. Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, et al. Machine Learning Methods: Comparison Study. JMIR Public Heal Surveill 2018;4:e11361.
13. 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.
14. Ehrentraut C, Ekholm M, Tanushi H, Tiedemann J, Dalianis H. Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. Health Informatics J 2018;24:24–42.
15. Namulanda G, Qualters J, Vaidyanathan A, Roberts E, Richardson M, Fraser A, et al. Electronic health record case studies to advance environmental public health tracking. J Biomed Inform 2018;79:98–104.
16. Ertem Z, Raymond D, Meyers LA. Optimal multisource forecasting of seasonal influenza. PLoS Comput Biol 2018;14:e1006236.
17. Schweickert B, Feig M, Schneider M, Willrich N, Behnke M, Peña Diaz LA, et al. Antibiotic consumption in Germany: first data of a newly implemented web-based tool for local and national surveillance. J Antimicrob Chemother 2018;73:3505–15.
18. Austrian JS, Jamin CT, Doty GR, Blecker S. Impact of an emergency department electronic sepsis surveillance system on patient mortality and length of stay. J Am Med Informatics Assoc 2018;25:523–9.
19. Khan Y, Leung GJ, Belanger P, Gournis E, Buckeridge DL, Liu L, et al. Comparing Twitter data to routine data sources in public health surveillance for the 2015 Pan/Parapan American Games: an ecological study. Can J Public Health 2018;109:419–26.

Correspondence to:
Rodolphe Thiébaut
Univ. Bordeaux, Inserm
Bordeaux Population Health
Research Center, UMR 1219
F-33000 Bordeaux, France
E-mail: rodolphe.thiebaut@u-bordeaux.fr

Appendix: Content Summaries of Best Papers for the Public Health and Epidemiology Informatics Section of the 2019 IMIA Yearbook

Arsevska E, Valentin S, Rabatel J, de Goër de Hervé J, Falala S, Lancelot R, Roche M

Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System

PLoS One 2018 Aug 3;13(8):e0199960

Animal infection outbreaks are a major public health threat. In this paper, the researchers developed a platform to detect automatically animal infection outbreaks from online news sources, the Platform for Automated extraction of Disease Information from the web (PADI-web). Information is retrieved from Google News in the English language and a free news aggregator of about 4,500 news sites. Five types of information are extracted in each news article with natural language processing techniques and machine learning: the number of cases, the location, the date, the disease name, and the affected host (cattle, pig...). An automatic rule discovery module based on the frequent item discovery algorithm, a data mining technique, discovers rules that are later used as features for a machine learning classifier. As an example of rule automatically discovered, when the word « killed » is found in one of three words preceding a number then this number is likely to be the number of cases. A support vector machine classifier combines different rules to predict the class (number of cases, date…) of a term in the news article. The authors achieved good performances in these information extraction tasks ranging from 80 % to 95 % as measured by the F-score, the harmonic mean of precision and recall.

Although many limitations hamper the feasibility of an exhaustive surveillance worldwide, this platform detected in 2016 several animal infectious outbreaks days before it was notified by the World Organization for Animal Health. Once again, free online content on the Internet contains key information for public health surveillance that can be extracted with artificial intelligence methods. This paper is a very good example of a full implementation and evaluation of such a web-based system.

Effland T, Lawson A, Balter S, Devinney K, Reddy V, Waechter H, Gravano L, Hsu D

Discovering foodborne illness in online restaurant reviews

J Am Med Inform Assoc 2018 Dec 1;25(12):1586-92

A foodborne illness is an infectious intestinal disease caused by a pathogen (bacteria, virus, or parasite) that enters the body through food or drink consumption. Common symptoms include abdominal cramps, nausea, and vomiting. In this paper, the authors used data from consumer reviews obtained from the Yelp website, a social networking site that records user's rate and review on restaurants. In collaboration with the New York City (NYC) Department of Health and Mental Hygiene (DOHMH), the authors developed a system to identify restaurant reviews on Yelp indicating foodborne illness. The classification algorithm uses a bag-of-words approach with a logistic regression classifier. The system is parameterized to favor recall over precision to reduce the risk of missing true positives. Yelp reviews of NYC restaurants are pulled every day and DOHMH epidemiologists validate each signal in a user interface. If a signal is validated, a Yelp message is sent to the author of the review to gather more information for further investigation. The system identified 10 outbreaks and 8,523 reports of foodborne illness associated with NYC restaurants since July 2012. Interestingly, only 3% of the illness incidents had been reported to DOHMH, which highlights the value of social media as an important source for foodborne illness surveillance. A very interesting aspect of the work presented is the use of biased adjusted data for improving the performance of the classifier in a continuous approach where validated data coming from previous alerts are used to further improve the algorithm. Obviously, the true performance of the system is depending on the quality of the information available on the social media and any manipulation of these data by the provider (in regards of the allegations reported in the news about Yelp) could have an influence on the results but this is not discussed in the paper.

Wakamiya S, Kawai Y, Aramaki E

Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study

JMIR Public Health Surveill 2018 Sep 25;4(3):e65

Social media data such as Twitter can be used to detect Influenza outbreak. It requires natural language processing (NLP) techniques and a classifier to perform tweet classification. However, differentiating direct and indirect information is crucial for Public Health surveillance that needs a good approximation of the number of cases. For example, "I got the flu today" is a direct information (D) tweet whereas "there is a major outbreak in Okinawa" provides indirect information (I). The authors developed a specific module to handle this task by applying a binary classifier based on support vector machine under the bag-of-words representation of tweets. A location module tries to infer the user location using the user profile, the GPS coordinates when available, and the content of tweets. Furthermore, the authors introduced the concept of "trapped sensors" to better predict the number of cases per area. The idea is that after the onset of an epidemic, people become deactivated by the event and they do not report having flu. Such deactivated people are called "trapped sensors". Statistical models that take into account these "trapped sensors" to predict the number of cases showed better correlation with the reference (information from Health Authorities) than over models. This paper explains the many pitfalls that exist while using Twitter data for flu monitoring and proposes a new approach to make good approximation of Influenza cases.