Published online: 2020-08-21

Public Health and Epidemiology Informatics: Recent Research Trends Moving toward Public Health Data Science

Sébastien Cossin[1,2], Rodolphe Thiébaut[1,2,3], 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: PubMed searches of 2019 literature concerning public health and epidemiology informatics were conducted 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 allow the Editorial Committee a curated selection of the best papers.

Results: Among the 835 references retrieved from PubMed, two were finally selected as best papers. The first best paper leverages satellite images and deep learning to identify remote rural communities in low-income countries; the second paper describes the development of a worldwide human disease surveillance system based on near real-time news data from the GDELT project. Internet data and electronic health records are still widely used to detect and monitor disease activity. Identifying and targeting specific audiences for public health interventions is a growing subject of interest.

Conclusions: The ever-increasing amount of data available offers endless opportunities to develop methods and tools that could assist public health surveillance and intervention belonging to the growing field of public health Data Science. The transition from proofs of concept to real world applications and adoption by health authorities remains a difficult leap to make.

Keywords

Public health, epidemiology, surveillance, medical informatics, International Medical Informatics Association, data science, ethics

Yearb Med Inform 2020:231-4
http://dx.doi.org/10.1055/s-0040-1702020

Introduction

The increasing digitization of health data and the recent advances in several fields of computer science such as natural language processing and deep learning offer more opportunities for applications in the domain of public health and epidemiology.

Data generated on the Internet can be used to measure the prevalence and incidence of diseases and allows the development of real-time applications to serve the early detection of epidemics [1]. Although easy and cheap to access, Internet data is often noisy and extracting good quality data for decision makers is often very challenging and requires strong and multidisciplinary expertise.

Harder to access, electronic health records (EHRs) and clinical data registries contain very high quality data generated by health professionals. Several international initiatives like the Observational Health Data Sciences and Informatics (OHDSI, https://ohdsi.org) aim at facilitating the interoperability and the exploitation of clinical data while guaranteeing data protection and ownership. Recently, the feasibility of building a cohort of hundreds of millions patients across the globe has been demonstrated [2] and the activity opens up new research perspectives at the global scale.

A promising technology for public health is the increasing use of mobile phones. The surge of computing power and the ubiquity of mobile phones around the globe make it possible for large populations to participate in public health surveillance and prevention campaigns [3]. Further research is expected to fully leverage this technology for the benefit of public health.

This synopsis looks at the literature published in 2019 in the domain of medical informatics applied to public health and epidemiology. The aim is to identify new topics and trends as compared to previous years and describe the selection process of the best papers published in 2019 based on quality and originality of articles.

Methods

A comprehensive literature search was performed using PubMed/Medline database from NCBI, National Center for Biotechnology Information. Using a large set of MeSH descriptors, the queries targeted public health or epidemiological journal articles over the year 2019 that included medical informatics topics. Returned references addressing topics of the other sections of the Yearbook, *e.g.*, those related to sensors, were excluded from our search. The study was performed at the beginning of January 2020, and the search returned a total of 835 references.

Articles were separately reviewed by the two section editors and were first classified into three categories: keep, discard, or leave pending with the BibReview tool [4]. Then, the two lists of references were merged yielding 90 references that were retained by at least one reviewer or classified as "pending" by both of them. The

two section editors jointly reviewed the 90 references and selected a consensual list of 15 candidate best papers. Two candidate best papers were removed from the list because the papers were selected by other sections and one paper was drafted to obtain a final list of 14 candidate best papers. All of these papers were then peer-reviewed by editors and external reviewers. Each paper was reviewed by at least four reviewers. Two papers were finally selected as best papers by the Yearbook Editorial Committee (Table 1). A content summary of these selected papers can be found in the appendix of this synopsis.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2020 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
▪ Bruzelius E, Le M, Kenny A, Downey J, Danieletto M, Baum A, Doupe P, Silva B, Landrigan PJ, Singh P. Satellite images and machine learning can identify remote communities to facilitate access to health services. J Am Med Inform Assoc 2019;26(8-9):806-12.
▪ Feldman J, Thomas-Bachli A, Forsyth J, Patel ZH, Khan K. Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise. J Am Med Inform Assoc 2019;26(11):1355-9.

Results

The trend towards the increase in the number of publications in infodemiology noticed in 2018 [5] continued in 2019 with new emerging use cases like the monitoring of physical activity using Twitter Data [6], the surveillance of plague outbreak with Google Trends [7], the identification of patients with diabetes or the detection of conjunctivitis epidemics worldwide based on search engine queries [8,9]. One selected best paper describes the development of a global infectious disease database using natural language processing, machine learning, and human expertise [10]. The original idea of this paper was to exploit the publicly available data of the GDELT project that monitors in near real time the world's broadcast, print, and web news. The system developed was capable of analyzing news in 65 languages to early detect onset of epidemics worldwide.

Disease surveillance systems based on social media and search queries aim to measure current disease activity, aka *nowcasting*, but are still prone to errors due to the imperfect features of the models they rely upon. Priedhorsky *et al.*, [11] proposed the metric of *deceptiveness* which quantifies the noise in the features of a model. This metric could improve in the future the measurement of disease prevalence and incidence. In order to be adopted by health authorities, there is much room for further research to improve the performance of these statistical models as accurate and reliable estimations of disease activity based on Internet data.

EHRs are still a source of high quality information for public health researchers. Post marketing drug surveillance [12, 13] and healthcare-associated outbreaks detection [14, 15] continue to be hot topics of research.

Also, as quoted by the survey paper of the Public Health and Epidemiology Informatics section of the 2020 International Medical Informatics Association (IMIA) Yearbook [16], the targeting of sub-populations for dedicated public health interventions is a growing subject of interest. Digital segmentation aims to reach audiences using digital technologies offering new opportunities to deliver appropriate prevention messages [17]. Several studies have already shown the interest of social media for public health campaigns such as smoking cessation [18]. A way to maximize their impact and efficiency could be to identify and target specific audiences. To do so, natural language processing, data mining, and machine learning have been used to classify user traits [19]. The strategy is similar to that of online targeted advertising except that the goal is to deliver dedicated public health, rather than advertising, messages. The second selected best paper applied deep learning on satellite images to identify rural and hard-to-reach remote communities in low-income countries and help community health workers deliver health services [20]. The geographical segmentation of population based on their access to healthcare is needed to organize specific healthcare delivery and to reduce inequalities.

Despite the obvious need of these new approaches, the use of phenotyping algorithms to classify individuals raises ethical issues about data privacy, confidentiality, and informed consent. Interestingly, these issues are rarely discussed when information is retrieved from social media, unlike from EHRs where an institutional review board authorization is often mandatory to carry out such analyses. A consensus has yet to emerge to handle Internet data [21]. In the meantime, public health researchers must do their utmost to protect user data and to keep confidential the models of individual prediction.

Conclusion

The huge amount of data available from Internet, from EHRs, and upcoming from mobile phones, is the fuel for a lot of research on different topics covering statistics, informatics, and epidemiology defining public health Data Science.

Acknowledgements

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

References

1. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011 May;40(5 Suppl 2):S154-8.
2. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-8.
3. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. Sci Transl Med 2015 Apr

15;7(283):283rv3.
4. 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.
5. Thiébaut R, Cossin S, Section Editors for the IMIA Yearbook Section on Public Health and Epidemiology Informatics. Artificial Intelligence for Surveillance in Public Health. Yearb Med Inform 2019 Aug;28(1):232-4.
6. Liu S, Chen B, Kuo A. Monitoring Physical Activity Levels Using Twitter Data: Infodemiology Study. J Med Internet Res 2019 Jun 3;21(6):e12394.
7. Bragazzi NL, Mahroum N. Google Trends Predicts Present and Future Plague Cases During the Plague Outbreak in Madagascar: Infodemiological Study. JMIR Public Health Surveill 2019 Mar 8;5(1):e13142.
8. Hochberg I, Daoud D, Shehadeh N, Yom-Tov E. Can internet search engine queries be used to diagnose diabetes? Analysis of archival search data. Acta Diabetol 2019;56(10):1149-54.
9. Deiner MS, McLeod SD, Wong J, Chodosh J, Lietman TM, Porco TC. Google Searches and Detection of Conjunctivitis Epidemics Worldwide. Ophthalmology 2019 Sep;126(9):1219-29.
10. Feldman J, Thomas-Bachli A, Forsyth J, Patel ZH, Khan K. Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise. J Am Med Inform Assoc 2019 Nov 1;26(11):1355-9.
11. Priedhorsky R, Daughton AR, Barnard M, O'Connell F, Osthus D. Estimating influenza incidence using search query deceptiveness and generalized ridge regression. PLoS Comput Biol 2019 Oct 1;15(10):e1007165.
12. Banerji A, Lai KH, Li Y, Saff RR, Camargo CA, Blumenthal KG, et al. Natural Language Processing Combined with ICD-9-CM Codes as a Novel Method to Study the Epidemiology of Allergic Drug Reactions. J Allergy Clin Immunol Pract 2020 Mar;8(3):1032-1038.e1.
13. Lin F-C, Huang S-T, Shang RJ, Wang C-C, Hsiao F-Y, Lin F-J, et al. A Web-Based Clinical System for Cohort Surveillance of Specific Clinical Effectiveness and Safety Outcomes: A Cohort Study of Non-Vitamin K Antagonist Oral Anticoagulants and Warfarin. JMIR Med Inform 2019 Jul 3;7(3):e13329.
14. Bush K, Barbosa H, Farooq S, Weisenthal SJ, Trayhan M, White RJ, et al. Predicting hospital-onset Clostridium difficile using patient mobility data: A network approach. Infect Control Hosp Epidemiol 2019;40(12):1380-6.
15. Sundermann AJ, Miller JK, Marsh JW, Saul MI, Shutt KA, Pacey M, et al. Automated data mining of the electronic health record for investigation of healthcare-associated outbreaks. Infect Control Hosp Epidemiol 2019;40(3):314-9.
16. Buckeridge DL. Precision, Equity, and Public Health and Epidemiology Informatics – A Scoping Review. Yearb Med Inform 2020:226-30.
17. Evans WD, Thomas CN, Favatas D, Smyser J, Briggs J. Digital Segmentation of Priority Populations in Public Health. Health Educ Behav 2019;46(2_suppl):81-9.
18. Naslund JA, Kim SJ, Aschbrenner KA, McCulloch LJ, Brunette MF, Dallery J, et al. Systematic review of social media interventions for smoking cessation. Addict Behav 2017;73:81-93.
19. Chu K-H, Colditz J, Malik M, Yates T, Primack B. Identifying Key Target Audiences for Public Health Campaigns: Leveraging Machine Learning in the Case of Hookah Tobacco Smoking. J Med Internet Res 2019 Jul 21(7):e12443.
20. Bruzelius E, Le M, Kenny A, Downey J, Danieletto M, Baum A, et al. Satellite images and machine learning can identify remote communities to facilitate access to health services. J Am Med Inform Assoc 2019 Aug 1;26(8-9):806-12.
21. Hunter RF, Gough A, O'Kane N, McKeown G, Fitzpatrick A, Walker T, et al. Ethical Issues in Social Media Research for Public Health. Am J Public Health 2018;108(3):343-8.

Correspondence to:
Sébastien Cossin
Univ. Bordeaux, Inserm
Bordeaux Population Health Research Center
UMR 1219
F-33000 Bordeaux, France
E-mail: sebastien.cossin@u-bordeaux.fr

Appendix: Content Summaries of Selected Best Papers for the 2020 IMIA Yearbook, Section 'Public Health and Epidemiology Informatics'

Bruzelius E, Le M, Kenny A, Downey J, Danieletto M, Baum A, Doupe P, Silva B, Landrigan PJ, Singh P

Satellite images and machine learning can identify remote communities to facilitate access to health services

J Am Med Inform Assoc 2019;26(8-9):806-12

In low-income countries, a promising strategy for improving care access among remote rural population is via the expansion of community health worker (CHW) programs. In settings where census data is missing and vital registration systems are weak, a persistent barrier of the expansion of CHW programs has been the difficulty to accurately enumerate population catchment areas. The authors used satellite-based neural network methods to automate the identification of communities in very rural areas.

Training data came from the publicly available SpaceNet corpus and a rural satellite image dataset specifically built for this project. External validation data was provided by a geographic information system dataset identifying all known Liberian communities within the health service catchment area of Last Mile Health, a non-profit organization. Community geolocation data was obtained by sending a team into the field with handheld GPS devices to collect community locations. Then 26,180 candidate rural images were labeled for this project and split into training and testing sets using an 80:20 ratio. The community prediction approach involved recognition of individual buildings from satellite imagery with TensorBox that output a set of coordinates describing the bounding box of each building. In a second phase, a clustering method was used to identify groups of densely connected buildings indicative of a community. The source code of their program is published.

Compared with existing health system community census data, the study method detected 75% of registered communities and identified an additional 167 building groupings that had not previously been identified. This new method for identifying communities in rural and remote settings using satellite imagery and deep learning has the potential to facilitate greater targeting of health services in low-income countries.

Feldman J, Thomas-Bachli A, Forsyth J, Patel ZH, Khan K

Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise

J Am Med Inform Assoc 2019;26(11):1355-9

Rapid onset of infectious disease epidemics can significantly reduce cases and deaths. Online media reports can facilitate timelier identification. The huge volume of media reports and the different languages make the identification of disease activity very challenging. The authors collected media records from the Global Database of Events Language and Tone (GDELT), that monitors the world's broadcast, print, and web news from nearly every country. Its global coverage and its updates every 15 minutes make it an invaluable source.

The authors used Google Translate to translate every media report they found into English. A dictionary containing a curated list of disease names was created. If an article didn't contain a disease name in its title, the article was deemed irrelevant. To distinguish articles talking about general infectious disease information and about disease activity, a supervised classification model was trained on 8,322 manually labeled articles. Finally, a user interface was built to allow clinical experts to verify articles clustered by disease, location, and time. The authors compared their GDELT-derived feed to the WHO disease Outbreak News reports from July 2017 to June 2018.

Their classification model achieved a F1 score of 0.87. On the study period, 37 outbreaks were reported by the WHO. Out of the 37 outbreaks, 89% were covered by online news outlets before the WHO reported the outbreak and the system correctly detected 94% of these events before reported by the WHO with a mean of 43.4 days earlier. Since it takes time for health authorities to investigate and confirm a disease, outbreak media reports can provide timelier information, but news reports fail often to distinguish between suspected and confirmed cases and are prone to false positive errors.

Combining natural language processing, machine learning, and human expertise, the authors created an international and near real-time event-based infectious disease activity database.