Some Innovative Approaches for Public Health and Epidemiology Informatics

L. Toubiana[1], N. Griffon[1,2], Section Editors for the IMIA Yearbook Section on Public Health and Epidemiology Informatics

[1] INSERM UMRS 1142 LIMICS, Université Pierre et Marie Curie - Paris, France
[2] Département d'information et d'informatique médicale, TIBS, LITIS EA 4108, Rouen University Hospital, Normandy, Rouen, France

Summary

Objectives: Summarize excellent current research published in 2015 in the field of Public Health and Epidemiology Informatics.

Methods: The complete 2015 literature concerning public health and epidemiology informatics has been 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 allow the editorial team an enlightened selection of the best papers.

Results: Among the 1,272 references retrieved from PubMed and Web of Science, three were finally selected as best papers. The first one presents a language agnostic approach for epidemic event detection in news articles. The second paper describes a system using big health data gathered by a statewide system to forecast emergency department visits. The last paper proposes a rather original approach that uses machine learning to solve the old issue of outbreak detection and prediction.

Conclusions: The increasing availability of data, now directly from health systems, will probably lead to a boom in public health surveillance systems and in large-scale epidemiologic studies.

Keywords

Public health, epidemiology, medical informatics, International Medical Informatics Association, health information systems

Yearb Med Inform 2016:247-50
http://dx.doi.org/10.15265/IY-2016-047
Published online November 10, 2016

Introduction

Globalization, mass gathering events, and mass air travels have drastically shortened the time required for an emerging disease to spread all around the world. America and Corsica are only two examples of populations that were thought to be protected earlier but in fact are not, see Zika [1], resp. Schistosomiasis [2]. The problem of epidemics is not new however and epidemiologists and public health professionals have worked for centuries to model, predict, and contain epidemic diseases. Informatics may help: complex models, rich simulations, data provided by social networks and search engines, and health data provided by health systems themselves are now analyzed by massive computational power and become available to be used in the fight against epidemics. The literature review performed for the Public Health and Epidemiology Informatics section of the IMIA yearbook was an attempt to identify papers published in 2015 on the most interesting advances in public health and epidemiology informatics.

Useable epidemiologic signals are still needles in the haystack of big data provided by the Internet, social media, and open access databases for infectious or non-infectious diseases. Many authors propose solutions and new approaches to this problem and demonstrate their performances. Web-generated data have been recognized as a way to monitor public perception of an epidemic event, which will probably help to focus public health campaigns that may now use social media directly. Health data have become more available, but it will take time before this data is easily useable for research due to privacy and security issues. However, some efforts already exploit this data for signal detection or on a larger scale for population health monitoring.

Paper Selection

Two well-known bibliographic databases, Pubmed/Medline (from National Center for Biotechnology Information) and Web of Science® (from Thomson Reuters) were searched for public health and epidemiology papers involving computer science or big data. The search was performed at the beginning of January in 2016 and returned a total of 1,272 references for the year 2015.

The selection process was adapted from the one described by Lamy et al [3]. In order to select candidate best papers, each section editor screened 636 references on the basis of titles. This first selection step led to a short list of 132 references that were finally separately reviewed by the two section editors on the basis of abstracts (and full-texts when necessary), and divided into three categories: keep, discard, or leave pending. The two ratings were then merged, leaving 39 references that were classified as kept or pending by at least one reviewer. The two section editors jointly reviewed these 39 references and proposed a consensual list

of 14 candidate best papers. All pre-selected 14 papers were then peer-reviewed by the Yearbook editors and external reviewers (at least four reviewers per paper). Three 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.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2016 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
▪ Hu Z, Jin B, Shin AY, Zhu C, Zhao Y, Hao S, Zheng L, Fu C, Wen Q, Ji J, Li Z, Wang Y, Zheng X, Dai D, Culver DS, Alfreds ST, Rogow T, Stearns F, Sylvester KG, Widen E, Ling XB. Real-time web-based assessment of total population risk of future emergency department utilization: statewide prospective active case finding study. Interact J Med Res 2015 Jan-Mar;4(1):e2.
▪ Jafarpour N, Izadi M, Precup D, Buckeridge DL. Quantifying the determinants of outbreak detection performance through simulation and machine learning. J Biomed Inform 2015 Feb;53:180-7.
▪ Lejeune G, Brixtel R, Doucet A, Lucas N. Multilingual event extraction for epidemic detection. Artif Intell Med 2015 Oct;65(2):131-43.

Conclusion and Outlook

The best paper selection was performed independently of the 2016 Yearbook special theme – "Unintended consequences: new problems and new solutions".

Ainsworth and Buchan [4] reflected on health system learning i.e. the use of daily generated health data, to improve health outcomes, either at a public health level or for use in evidence-based medicine. They identified three major obstacles: (i) data are fragmented and sometimes managed by institutions not directly involved in patient care, (ii) the need for confidentiality of health data is not weighed rationally to the potential benefit of its secondary use, and (iii) the workforce required to analyze (big) health data (statisticians, computer specialists, clinicians, epidemiologists) is missing. However, the authors believe that recent evolutions in information technology may resolve many of these issues and enable health system learning.

The paper by Hu et al. [5] proposes such health information system. It is restricted to a specific topic: emergency department (ED) visits in the next 6 months, and its performance is limited. Nevertheless, it provides interesting information to both clinicians and public health managers. The authors use the big Maine (USA) health information exchange data in real time, which appears to be a promising approach to monitor multiple public health indicators closely.

Cooper et al. [6] proposed an outbreak detection and characterization system. The system computes the probability of any disease outbreak and estimates outbreak characteristics. The system uses input information provided by a case detection system that gathers and analyses data from ED visits. A feedback loop from the outbreak detection and characterization system to the case detection system improves the latter system efficiency, which is an important step toward a learning health system.

As Ainsworth and Buchan [4] reported, the United Kingdom National Health Service (NHS) is engaged in a learning health information system pathway: Elliot et al. [7] proposed a limited but smart evaluation of NHS daily generated data from its web service for syndromic surveillance.

Infectious diseases continue to cause high morbidity, mortality, and financial costs. Early detection of disease outbreaks is not a new issue but it remains crucial. In 2015, still many researchers [8][9][10] tried to develop the ultimate surveillance system using search engines, social network, or wiki data. While authors combined more and more data sources for their surveillance systems, the use of health system data remained limited.

Ram et al. [8] aggregated data from Google and Twitter with air pollution data to predict asthma-related ED visits. Unlike Hu et al. [5], the researchers had to be satisfied with data from one hospital. However, their prediction model provides good estimates. Santillana et al. [9] evaluated with promising results the same data sources, as well as some others, against the Center for Disease Control's (CDC) gold standard for influenza-like illness surveillance. Lejeune et al. [10] proposed an innovative way to detect epidemic events in news articles based on string character repetition. Unlike classical surveillance systems, this one is almost independent from the language of the article. This approach seems useful, necessary, and deserves to be highlighted as epidemics may spread from one country to other countries of different languages. Surveillance mostly concerns infectious disease, however, some non-infectious diseases were also studied. Wang et al. [11] presented a Google trend based method to nowcast and forecast outpatient visits, which may be useful for resource allocation. Unfortunately, the evaluation metrics used for surveillance systems differ in every article presented here, which impedes systems comparison. Some of these differences are inherent to the evaluated strategy; nevertheless, a commonly accepted evaluation framework for such evaluations would be very interesting to the scientific community. Jafarpour et al.'s work could be a first step in such direction [12]. They propose an original probabilistic model that aims at discovering either outbreaks or the determinants of outbreak detection.

Another potentially interesting use of social networks was studied by Lazard et al. [13]. The authors identified topics of public interest on Ebola. A text mining approach was used to dissect the 2,155 tweets generated during a CDC twitter chat. Eight topics were identified. They allowed a better understanding of public concerns and may be applicable for any public health event to improve public health communication, which may be achieved through social networks themselves. Lister et al. [14] related how they promoted "Family Meal" in Utah using the Internet. Instead of relying on educational material, this campaign was based on the "laugh model", which roughly consists in using business-like marketing and branding strategies – entertaining,

engaging, and sometimes irreverent social media materials – to promote public health programs. The model was very efficient with a cost of 0.2 cent per person reached. Internet activity may also allow evaluating public health actions. Barak-Cohen and Reis [15] describe such a method applied to polio vaccination compliance during a recent Israel wide vaccination campaign against polio. They demonstrated an interesting correlation between the official vaccination rate and Google trends. Given Google's data timeliness and inexpensiveness, this may be a useful complementary tool to optimize vaccination campaign.

Polo et al [16] proposed a model to improve the spatial planning of public health services. Authors worked on sterilization sites for cats and dogs. However, this model may help public health stakeholders to decide where services should be deployed in order to limit or avoid a forecasted epidemic. Model parameters may allow multiple ways of planning site implantation with respect to available resources.

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

Correspondence to:
Dr. Nicolas Griffon
Unité d'Informatique Médicale
CHU de Rouen
1 rue de Germont
76031, Rouen
France
Tel. +33 6 42 25 44 11
E-mail: nicolas.griffon@chu-rouen.fr

Dr. Laurent Toubiana, PhD
INSERM UMRS 1142 "LIMICS"
15, rue de l'École de Médecine
75006 Paris
France
Tel: +33 1 44 27 91 97
E-mail: Laurent.toubiana@inserm.fr

References

1. Petersen E, Wilson ME, Touch S, McCloskey B, Mwaba P, Bates M, et al. Rapid Spread of Zika Virus in The Americas - Implications for Public Health Preparedness for Mass Gatherings at the 2016 Brazil Olympic Games. Int J Infect Dis 2016 Mar;44:11-5.
2. Brunet J, Pfaff AW, Hansmann Y, Gregorowicz G, Pesson B, Abou-Bacar A, et al. An unusual case of hematuria in a French family returning from Corsica. Int J Infect Dis 2015 Feb;31:59-60.
3. Lamy JB, Séroussi B, Griffon N, Kerdelhué G, Jaulent MC, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54(2):135-44.
4. Ainsworth J, Buchan I. Combining Health Data Uses to Ignite Health System Learning. Methods Inf Med 2015;54:479–87.
5. Hu Z, Jin B, Shin AY, Zhu C, Zhao Y, Hao S, et al. Real-time web-based assessment of total population risk of future emergency department utilization: statewide prospective active case finding study. Interact J Med Res 2015;4:e2.
6. Cooper GF, Villamarin R, Rich Tsui F-C, Millett N, Espino JU, et al. A method for detecting and characterizing outbreaks of infectious disease from clinical reports. J Biomed Inform 2015;53:15–26.
7. Elliot AJ, Kara EO, Loveridge P, Bawa Z, Morbey RA, Moth M, et al. Internet-based remote health self-checker symptom data as an adjuvant to a national syndromic surveillance system. Epidemiol Infect 2015;143:3416–22.
8. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform 2015;19:1216–23.
9. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol 2015;11:e1004513.
10. Lejeune G, Brixtel R, Doucet A, Lucas N. Multilingual event extraction for epidemic detection. Artif Intell Med 2015;65:131–43.
11. Wang HW, Chen DR, Yu HW, Chen YM. Forecasting the Incidence of Dementia and Dementia-Related Outpatient Visits With Google Trends: Evidence From Taiwan. J Med Internet Res 2015 Nov 19;17(11):e264.
12. Jafarpour N, Izadi M, Precup D, Buckeridge DL. Quantifying the determinants of outbreak detection performance through simulation and machine learning. J Biomed Inform 2015;53:180–7.
13. Lazard AJ, Scheinfeld E, Bernhardt JM, Wilcox GB, Suran M. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. Am J Infect Control 2015;43:1109–11.
14. Lister C, Royne M, Payne HE, Cannon B, Hanson C, Barnes M. The Laugh Model: Reframing and Rebranding Public Health Through Social Media. Am J Public Health 2015 Nov;105(11):2245-51.
15. Barak-Corren Y, Reis BY. Internet activity as a proxy for vaccination compliance. Vaccine 2015 May 15;33(21):2395-8.
16. Polo G, Acosta CM, Ferreira F, Dias RA. Location-allocation and accessibility models for improving the spatial planning of public health services. PLoS One 2015;10(3):e0119190.

Appendix: Content Summaries of Best Papers for the 2016 IMIA Yearbook Section on 'Public Health and Epidemiology Informatics'

Hu Z, Jin B, Shin AY, Zhu C, Zhao Y, Hao S, Zheng L, Fu C, Wen Q, Ji J, Li Z, Wang Y, Zheng X, Dai D, Culver DS, Alfreds ST, Rogow T, Stearns F, Sylvester KG, Widen E, Ling XB

Real-time web-based assessment of total population risk of future emergency department utilization: statewide prospective active case finding study

Interact J Med Res 2015 Jan-Mar;4(1):e2

Public health stakeholders need tools to monitor population health, to anticipate health care system use, and to decide which health policy to lead. Such a tool is presented in this paper. It capitalizes on the health information exchange in Maine (USA) to create statewide cohorts of patients. These cohorts were used to train, calibrate, and test a measure of the future 6-month emergency department visit risk. After a careful selection of clinical features to be introduced in their model, the authors propose a score classifying each Maine patient according to his own risk (thresholds may be modified according to field care provider objectives). The tool also proposes subgroup analysis to precisely identify clusters justifying a specific health intervention or management.

The proposed score reaches an Area Under the Receiver-Operator Curve (AUC) of 0.732 in the prospective test cohort. High risk identified patient visit ED earlier and more frequently. The tool allows the identification of six subgroups in the high risk population. High risk patient identification may ease the resolution of gaps in health care coverage, help to focusing on specific interventions, and, finally, may avoid preventable ED visits. As authors mention, despite the huge amount of data gathered, this tool lacks some important information that would have enhanced the results. Nevertheless, it seems a promising approach for data-driven public

health policy and real-time health surveillance at the population level. This tool may allow public health stakeholders precisely targeting public health policy.

Jafarpour N, Izadi M, Precup D, Buckeridge DL

Quantifying the determinants of outbreak detection performance through simulation and machine learning

J Biomed Inform 2015 Feb;53:180-7

Many outbreak detection algorithms have been proposed for use in syndromic surveillance. Algorithms perform differently when they are applied to different data sources or used in different surveillance situations. The objective of this research is to develop and evaluate a model for quantitatively characterizing the determinants of outbreak detection performance and predicting the performance of detection methods. Authors applied structure and parameter learning algorithms to build a Bayesian network model for predicting detection performance as a function of outbreak characteristics and surveillance system parameters and evaluated the predictions of this model through 5-fold cross-validation. The model quantified the influence of different outbreak characteristics and parameters on detection performance. The alerting threshold and the peak size of the outbreak are characteristics with a strong influence on detection performance. The model suggested the important role of other algorithm features, such as adjustment for weekly patterns. This model can be used to compare the performance of detection methods under different surveillance scenarios, to gain insight to which characteristics of outbreaks and biosurveillance algorithms drive detection performance, and to guide the configuration of surveillance systems.

Lejeune G, Brixtel R, Doucet A, Lucas N

Multilingual event extraction for epidemic detection

Artif Intell Med 2015 Oct;65(2):131-43

Lejeune et al. proposed a language agnostic system aiming at detecting epidemic events as fast as possible. This system allows the extraction of epidemic events in sources written in various languages. This solution assumes that news articles, used as the input of the system, are structured in a similar way in every country and language. The system looks for specific character string repetition found in both news articles and disease knowledge bases, i.e. the Wikipedia English list of infectious diseases and the interlingual outgoing links for each disease, and the location knowledge bases, i.e. the Wikipedia list of sovereign states and the interlingual outgoing links for each state. Two parameters (for disease and location respectively) determinate the appropriate string matching ratio between the motifs found in articles and knowledge bases. The system exhibits good performance for disease extraction both on the 2,000 5-language documents corpus (Chinese, Greek, English, Russian and Polish) created by the authors (0.72 precision and 0.91 recall), and on the BEcorpus, which is not fitted for precision evaluation (0.88 recall). Performance varies according to the language (from 0.65 to 0.84 precision and from 0.87 to 1 recall for Polish and Chinese in both cases, respectively) but remains quite high. The performance of location extraction is high (0.79) but worse than when using the country that emitted the news (precision = 0.87). The main strength of this approach is the low marginal cost adding a new watched language, which only requires an easily accessible knowledge base and fitting parameters values. Nevertheless it relies on the assumption that journalistic style is consistent and that it ultimately depends on language characteristics that may not be well suited for this approach. Some enhancements are required for the tool to be able to detect multiword diseases affected by inflection on several words.