International Journal of Infectious Diseases 63 (2017) 77–87

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Review

Utility and potential of rapid epidemic intelligence from internet-based sources

S.J. Yan[a], A.A. Chughtai[a], C.R. Macintyre[a,b,*]

[a] School of Public Health and Community Medicine, University of New South Wales, Australia
[b] College of Public Service and Community Solutions, Arizona State University, Phoenix, USA

ARTICLE INFO

Article history:
Received 28 May 2017
Received in revised form 19 July 2017
Accepted 21 July 2017
Corresponding Editor: Eskild Petersen, Aarhus, Denmark

ABSTRACT

Objectives: Rapid epidemic detection is an important objective of surveillance to enable timely intervention, but traditional validated surveillance data may not be available in the required timeframe for acute epidemic control. Increasing volumes of data on the Internet have prompted interest in methods that could use unstructured sources to enhance traditional disease surveillance and gain rapid epidemic intelligence. We aimed to summarise Internet-based methods that use freely-accessible, unstructured data for epidemic surveillance and explore their timeliness and accuracy outcomes.
Methods: Steps outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist were used to guide a systematic review of research related to the use of informal or unstructured data by Internet-based intelligence methods for surveillance.
Results: We identified 84 articles published between 2006-2016 relating to Internet-based public health surveillance methods. Studies used search queries, social media posts and approaches derived from existing Internet-based systems for early epidemic alerts and real-time monitoring. Most studies noted improved timeliness compared to official reporting, such as in the 2014 Ebola epidemic where epidemic alerts were generated first from ProMED-mail. Internet-based methods showed variable correlation strength with official datasets, with some methods showing reasonable accuracy.
Conclusion: The proliferation of publicly available information on the Internet provided a new avenue for epidemic intelligence. Methodologies have been developed to collect Internet data and some systems are already used to enhance the timeliness of traditional surveillance systems. To improve the utility of Internet-based systems, the key attributes of timeliness and data accuracy should be included in future evaluations of surveillance systems.

Contents

* Corresponding author at: School of Public Health and Community Medicine, Samuels Building, Room 325, Faculty of Medicine, University of New South Wales, Sydney, 2052, NSW, Australia. Tel: +61 2 9385 3811; Fax: +61 2 9313 6185.
E-mail address: r.macintyre@unsw.edu.au (C.R. Macintyre).

http://dx.doi.org/10.1016/j.ijid.2017.07.020

Background

Broadly, 'intelligence' is defined as information collected, analysed and converted to gain insights. Intelligence can be described as a process or product (Hughbank, 2010). The application of intelligence principles in public health gave rise to the discipline of 'epidemic intelligence' (Bowsher et al., 2016), denoting all activities related to the early detection of potential health-related hazards (Paquet et al., 2006). Intelligence is generally formed through distinct steps such as data collection, processing, analysis, dissemination, feedback and tasking. Firstly, a specific request for intelligence is issued. Information-gathering methods are used to collect unstructured data. Following conversion to a manageable format, data can be interpreted and a final report produced. Feedback can inform a subsequent task (Hughbank, 2010). Though criticism exists in literature regarding this simplified model of intelligence, public health surveillance follows similar steps in case detection, reporting, analysis and confirmation of cases (Hulnick, 2006).

Automated intelligence methods are of growing interest due to improvements in information technology (Hughbank, 2010). Open-source intelligence (OSINT) can utilise user-generated data found in the Internet or social media (Li et al., 2016). The application of information technology for electronic data collection and interpretation have encouraged Internet-based methods that can inform rapid epidemic intelligence on public health events (Collier, 2012). The International Health Regulations issued by the World Health Organization (WHO) for health threat detection emphasized the importance of both indicator-based and event-based components of epidemic intelligence for the early detection of events (Barboza et al., 2013). Informal information sources are important, with the WHO reporting that more than 60% of initial disease epidemic reports come from unofficial sources (World Health Organization, 2016a). For the purposes of this report, a disease epidemic is "The occurrence of cases of disease in excess of what would normally be expected in a defined community, geographical area or season" (World Health Organization, 2016b). Public health surveillance, hereby referred to as 'surveillance', is defined by the WHO as: "The continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice." (World Health Organization, 2016c)

Overview of public health surveillance

Rapid epidemic detection and real-time monitoring are important objectives of syndromic surveillance to minimise the morbidity and mortality caused by infectious diseases. Different types of surveillance can be conducted according to desired objectives. Sociocultural or ethical issues are also considered, with cost-effectiveness having implications on system feasibility (McNabb et al., 2002). Active surveillance involves regular monitoring of sources, providing the most complete information but ideally requiring trained epidemiologists. In contrast, passive surveillance is less resource intensive, involving regular reports from a wide range of sources and chance detection of cases. Integrated or enhanced surveillance makes use of both active and passive systems (Nsubuga et al., 2006). Active syndromic surveillance, or reporting based on clinical case definitions and pre-diagnostic data can be conducted in settings lacking laboratory confirmation to support diagnoses (Chaudet et al., 2006). Syndromic surveillance systems have therefore been established for early epidemic detection to minimize mortality and morbidity associated with emerging disease threats (Flamand et al., 2011).

During infectious disease epidemics, validated data collected through traditional or indicator-based surveillance methods may not be available for timely use. Surveillance has traditionally involved the monitoring of public health indicators from a range of sources. Key indicators outlined by the WHO in 1968 initially included mortality, morbidity, clinical data, laboratory reports, relevant field investigations, surveys, animal or vector studies, demographic and environmental data (Declich and Carter, 1994). Other data have since been included, such as hospital statistics, disease registries, over-the-counter drug sales data, school or work absenteeism, telephone triage calls and news reports (Chan et al., 2011; Lee and Thacker, 2011). Traditional surveillance methods have disadvantages in timeliness and sensitivity, attributable to factors such as a lengthy data validation process, bureaucratic barriers, higher costs and resource requirements (Yang et al., 2013). The time lag between the onset of epidemics and released official reports may render surveillance data redundant for the purpose of early detection and response (Zhou et al., 2011).

To address these disadvantages, researchers have proposed methods utilising newer technologies like the Internet, mobile phones, improved point-of-care diagnostic tools and other event-based surveillance methods (Chunara et al., 2012a). In conjunction with developments in computing or automated technology, Internet-based methods can handle and utilise 'big data' collected from informal sources. Big datasets are terabytes to petabytes in size, requiring higher-level software tools to capture, store, manage and analyse effectively (Bello-Orgaz et al., 2016).

Internet-based sources have potential to provide more timely information for detecting infectious disease epidemics, such as by identifying events or clusters of disease-related keywords in local news reports or social media (Chunara et al., 2012a). Common informal sources found on the Internet include search queries, online news, blogs and social media (Salathe et al., 2013; Bernardo et al., 2013). Through a scoping review, Bernardo et al. traced the use of Internet-based sources for disease surveillance back to 2006, with early work focusing on influenza. Pioneering studies introduced 'infodemiology', or the study of the determinants and distribution of health information (Eysenbach, 2006) in electronic media, specifically the Internet or in a population with the aim to inform public health policy (Bernardo et al., 2013).