REVIEW 10.1111/1469-0691.12273

# An overview of Internet biosurveillance

**D. M. Hartley[1,2,*], N. P. Nelson[3,*], R. R. Arthur[4], P. Barboza[5], N. Collier[6,7], N. Lightfoot[8], J. P. Linge[9], E. van der Goot[9], A. Mawudeku[10], L. C. Madoff[11], L. Vaillant[5], R. Walters[12], R. Yangarber[13], J. Mantero[14], C. D. Corley[15] and J. S. Brownstein[16]**

*1) Imaging Science and Information Systems Center, Georgetown University School of Medicine, 2) Department of Microbiology and Immunology, Georgetown University Medical Center, 3) Department of Pediatrics, Georgetown University Medical Center, Washington, DC, 4) Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, USA, 5) International Department, French Institute for Public Health Surveillance (InVS), Saint Maurice, France, 6) The National Institute of Informatics, Tokyo, Japan, 7) The European Bioinformatics Institute, Hinxton, Cambridge, UK, 8) Connecting Organizations for Regional Disease Surveillance (CORDS), Lyon, France, 9) Joint Research Centre (JRC) of the European Commission, Ispra, Italy, 10) Public Health Agency of Canada (PHAC), Ottawa, ON, Canada, 11) University of Massachusetts Medical School, Worcester, MA, 12) Pacific Northwest National Laboratory, Richland, WA, USA, 13) Department of Computer Science, University of Helsinki, Helsinki, Finland, 14) Surveillance and Response Support Unit, European Centre for Disease Prevention and Control, Stockholm, Sweden, 15) Pacific Northwest National Laboratory, Richland, WA and 16) Harvard-MIT Division of Health Sciences and Technology, Children's Hospital Boston, Harvard Medical School, Boston, MA, USA*

## Abstract

Internet biosurveillance utilizes unstructured data from diverse web-based sources to provide early warning and situational awareness of public health threats. The scope of source coverage ranges from local media in the vernacular to international media in widely read languages. Internet biosurveillance is a timely modality that is available to government and public health officials, healthcare workers, and the public and private sector, serving as a real-time complementary approach to traditional indicator-based public health disease surveillance methods. Internet biosurveillance also supports the broader activity of epidemic intelligence. This overview covers the current state of the field of Internet biosurveillance, and provides a perspective on the future of the field.

**Keywords:** Digital disease detection, digital epidemiology, electronic surveillance, epidemic intelligence , event-based surveillance, Internet-based surveillance, participatory epidemiology, web-based text mining

**Article published online:** 24 May 2013
*Clin Microbiol Infect* 2013; **19:** 1006–1013
10.1111/1469-0691.12273

**Corresponding author:** D. M. Hartley, Imaging Science and Information Systems Center, Georgetown University Medical Center, 2115 Wisconsin Avenue NW, Suite 603, Washington, DC 20057, USA
**E-mail: hartley@isis.georgetown.edu**

*These authors contributed equally to this study.

## Introduction

Internet biosurveillance, or digital disease detection [1], utilizes unstructured data from diverse web-based sources to provide early warning and situational awareness of human, animal and plant infectious diseases, as well as chemical, radiological and nuclear threats [2]. The discipline emerged in the mid-1990s, relying primarily on text media for its information, and has evolved into a globally recognized field [3,4]. With the increasing volume of information and new media types available via the Internet, the field has grown to include social media, participatory sources, and non-text-based sources. The scope of source coverage ranges from local media in the vernacular to international media in widely read languages. Online official reporting sources are typically used to supplement and verify such informal Internet sources.

Internet biosurveillance is a timely modality that is available to government and public health officials, healthcare workers, and the public and private sector, serving as a real-time complementary approach to traditional indicator-based public health disease surveillance methods [5,6]. Internet biosurveillance also supports the broader activity of epidemic intelligence (EI). This review covers the current state of the field, and provides a perspective on its future.

## Methods

This is not a 'systematic review'; rather, this article outlines a general process of Internet biosurveillance according to established best practices, and discusses common technologies employed in extant systems. Each step of the process is collectively described, drawing upon personal experiences of system builders and practitioners, as well as published studies. The authors contributing to this article are either affiliated with Internet biosurveillance systems, are end-users of Internet biosurveillance systems, and/or have published recently in the field. Authors from the following active Internet biosurveillance systems are represented: BioCaster [7], the Global Public Health Intelligence Network (GPHIN) [8], HealthMap [9], the Medical Information System (MedISys) (Steinberger *et al.*, IDRC, 2008, Short and Extended Abstracts, pp. 612–614, http://publications.jrc.ec.europa.eu/repository/handle/111111111/13078 (accessed 9 February 2013)), the Program for Monitoring Emerging Diseases (ProMED-mail) [10], and the Pattern Understanding and Learning System (PULS) [11].

## Results

The process of Internet biosurveillance varies, but, in general, includes: (i) the collection and storage of data from the Internet; (ii) processing those data to produce information; (iii) assembling that information into analyses; and (iv) dissemination of analyses to end-users (Fig. 1). Each part of the process can entail many technical steps, which are described below. Information vetting can occur through fully automated, human-moderated or partially moderated approaches throughout the process. Multilingual data are managed via human linguists, machine translation, and natural language-processing technology.

### Collection and storage

*Data sources.* Internet biosurveillance systems rely on data from a variety of sources. Publicly available, informal sources include text-based news sites (e.g. *New York Times* and *Thanh Nien News*) and social media sources (e.g. Twitter [12], Facebook, and blogs); more recently, sources that utilize public input (e.g. FluTrackers, Flu Near You, and crowdsourcing platforms [13]) have gained popularity and credibility. Information from these sources is often available in real time as an event is developing. This information is validated and supplemented by official, publically available information sources (e.g. public health agencies, ministries of health, the WHO, the World Organization for Animal Health, and the Food and Agriculture Organization). Systems also may utilize sources with paid content (e.g. newswires and news aggregators). Audio and video sources provide non-text-based information. Sources range widely in geographical coverage, from local to international, and cover all languages with publicly available media.

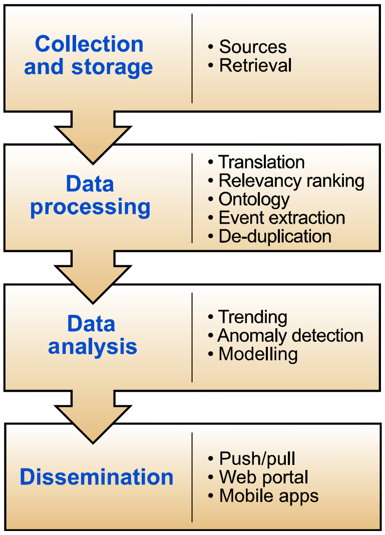

FIG. 1. The general process of Internet-based biosurveillance. Human input from information technology, public health and other experts can occur at any step.

*Data retrieval.* Data are retrieved from the Internet via two predominant modalities: media aggregators and system-specific web monitoring. As an example of the latter, Internet biosurveillance systems monitor the web by scraping (that is, specific web pages are accessed and stored) or crawling (that is, in addition to storing one specific web page, links on that page and links of links are accessed and stored).

Systems re-visit a list of predefined sites at regular intervals (typically, once to several times each day) in order to process data in a timely manner for early alerting. For paid or access-limited content, items might be accessed via a secure connection. News items from online news sites and social media are converted to a common format after retrieval, to enable searching and content mining. Public health agencies and ministries of health often provide their own feeds with official information. Feeds from aggregator news sites (e.g. Google and Yahoo) can be used to provide additional coverage. Content is extracted from the HTML code, with proper removal of advertisements and any other irrelevant text.

Social media data stem mostly from Twitter [14] and Facebook, which can be retrieved via their application programming interface. Access may be limited to a certain volume, and is subject to change according to the provider's Terms of Service. As some social media users are unaware that they publish their opinions worldwide, privacy issues arise under some jurisdictions, even with the publicly available data. Participatory data can be included via dedicated apps (e.g. iPhone and Android) or websites where users can leave comments (e.g. http://www.flutrackers.com/; http://www.healthmap.org/outbreaksnearme/) [15,16].

### Data processing

Once data are retrieved from the Internet, they must be processed to make them amenable for analysis. We emphasize that, because different types of users have different needs, there is no single, overarching goal for the data-processing step. Nevertheless, the following categories represent important steps in biosurveillance data processing: translation, relevancy ranking, ontology, event extraction, and de-duplication.

*Translation.* Although Arabic, Chinese, English, French, Spanish and Portuguese dominate the world's online news media, news of an outbreak event can appear in any language, and is often reported first in a local language. Systems have choices to make regarding the approach to translation. For example, they can build customized pipelines for a few languages, or they can translate each source language into a common target language. The decision is influenced by factors such as the availability of resources in each language, the time available to maintain each resource, and the translation quality required. For example, BioCaster employs full text translation first and uses only English language selection algorithms, whereas MedISys and HealthMap are language-specific in terms of the keywords employed to search Internet data. GPHIN employs both language-specific keywords and algorithms to extract relevant data from the Internet and news aggregator databases [17], whereas PULS employs language-specific linguistic analysis and ontologies and inference rules to extract relevant data.

*Relevancy ranking.* The next stage in processing is to assess the relevancy of the report according to some measure of the user's interest. Defining the user's interest as a set of guidelines, a decision tree or as a collection of examples is a crucial stage in system building, and provides a reference standard against which to evaluate various algorithms. Once this has been done, various approaches can be implemented, including supervised classifiers such as Naïve Bayes or Support Vector Machines with learn-to-rank, and Boolean keyword searches, which include logical operators such as AND and OR [18]. These techniques are language-specific, but it is also possible to deploy automated methods that are language-independent, such as clustering followed by automated labelling.

*Ontology.* Ontologies have proven useful in many domains (e.g. the life sciences) for structuring relationships between concepts. Biosurveillance requires a conceptual knowledge of diseases, microorganisms, signs and symptoms, and geography. A number of ontological resources have been developed or re-used for public health, although these are not generally as well known as those in experimental biology or clinical fields, such as the Unified Medical Language System. Among those developed specifically for public health are GIDEON (commercial, openly available), BioCaster (open source), and GPHIN (non-commercial, limited access). Such ontologies provide knowledge needed by Internet biosurveillance systems to make intelligent judgements about the terms appearing in news reports. For example, a mention of *Yersinia pestis* may imply that the disease under consideration is bubonic plague. However, not all ambiguities can be resolved with the static knowledge contained in an ontology. One of the most practical problems is toponym disambiguation (i.e. place names). For example, a mention of a disease outbreak in 'Cambridge' might resolve to any of several places worldwide, including the UK or the USA.

*Event extraction.* Once a set of topics of potential interest has been identified, specific biological events are extracted from the data. This can be accomplished in different ways. As one example, simple keyword recognition algorithms are often used to categorize incoming news items. In this approach, an article is categorized according to predefined keywords (see example in Table 1). Boolean combinations (e.g. AND, OR, NOT) and proximity searches (i.e. search for articles where two or more separately matching term occurrences are within

**TABLE 1. Examples of multilingual keywords used for identification of dengue fever in MedISys**

| Keywords |
|---|
| dengue |
| Denguefieber |
| лихорадк%+денге |
| 登革熱 |
| تب+دن |
| δάγκειος+πυρετός |
| Ιός+δάγγειου |
| knokkelkoorts |
| febre+hemorrágica |
| fiebre+hemorrágica |
| hemorrhagic+fever |
| … |

a specified word or character distance) can then be applied [19].

More detailed aspects of an outbreak can be extracted by event meta-data extraction, in which the aspects of interest are known and defined *a priori*. Examples of commonly detected aspects include the name of the disease, the species affected, the date of the outbreak, the numbers of cases and deaths, and the location of the outbreak. Event meta-data extraction uses the extensively researched technology known as information extraction, which is the basis of PULS and BioCaster [11]. Less common aspects include distal indicators of political and social response, such as ward closures or the deployment of international organizations to the affected region. Often, the techniques used are linguistic patterns developed with specific rule systems, but supervised, semi-supervised and unsupervised machine-learning approaches have also been evaluated [20].

*De-duplication.* Effective de-duplication is essential for events with wide coverage, so that nearly identical stories appearing in many sources do not overwhelm the user. De-duplication may involve the detection of reports that are identical in content, which are handled in practice with clustering techniques as outlined above. Reports may also be identical in the aspects of the outbreak that they report. De-duplicating these reports in practice is challenging, and can require deeper-meaning analysis. Nevertheless, there are often subtle but important aspects of an event that may not be easily captured, such as the revision of victim numbers, the change in a patient's condition, or a comparison between a novel and a known agent. De-duplication should ideally be sensitive to these grey areas, and pass forward such articles for human analysis.

### Data analysis

At this stage of the process, a biosurveillance system will have produced a structured collection of events that are potentially relevant to end-users. However, only a subset of these may be highly useful, given a particular user's interests. For example, a case of seasonal influenza in a celebrity, although widely reported, may be less relevant than a few reports of a cluster of novel influenza among farmers. Given the conflict between the volume of data to be analysed and the limited ability of humans to review large amounts of information quickly, it is often desirable to process the articles through an automated trend and anomaly detection capability in order to increase throughput and timeliness. The objective is to infer which events are more urgent or unusual in a timely manner, so that the user can investigate further and potentially initiate risk analysis. The challenge is to model what is already known (i.e. what is normal or expected), and to decide whether the current event is significantly at variance as early as possible. We focus on two complementary classes of approach in this section: trend analysis and anomaly detection.

*Trend analysis.* The temporal nature of Internet biosurveillance data produces longitudinal patterns and trends. Precursors and indicators of outbreaks can be tracked over time to show the precedence of an event before symptoms or the populace pass thresholds for warning. Timelines can also be used to track classifiers, keywords, locations, or terms, and indicate temporal traces of events for significance against predefined baselines. Visualizing topical trends and shifts over time based on such lexicons can facilitate the detection of unexpected disease events. Standard time-series algorithms and other signal-processing techniques are often used to model these temporal trends [21–23].

*Anomaly detection.* Anomaly detection attempts to put the features of the event into context in order to determine some level of significance. Context is usually considered to be spatial and/or temporal or a mixture of the two, and can be based on simple event counts of a particular disease type or on multiple features of the event. However, in situations where terminology begins to specialize or diverge (e.g. 'mad cow' to 'bovine spongiform encephalopathy', or 'swine flu' to 'H1N1'), the anomaly detection can be attenuated.

### Dissemination

Achieving the ultimate public health goals of biosurveillance systems—to facilitate early outbreak detection, thereby allowing timely interventions, limiting the severity and extent of spread—depends on the clear and rapid distribution of information. Internet-based biosurveillance systems use different means of disseminating information, depending on user needs and resources and the nature of the information.

Most systems use a combination of actively 'pushing' material to users and allowing users to 'pull' material when desired. ProMED-mail, one of the earliest Internet-based biosurveillance systems, uses mailing lists (e-mail) and listserv software, where users can subscribe to specific resources (e.g. animal or plant diseases). GPHIN uses a pushing function to send alerts about events that have been identified as significant to subscribers. Some services (e.g. HealthMap) allow users to specify parameters for pushed information, such as specific diseases, categories of disease, and geographical locations. SMS text messages, mobile telephone networks and social networks (e.g. Twitter) actively send information to anyone subscribing to a feed.

In addition, most Internet biosurveillance systems have a dedicated website where users may query and filter material on demand. Although they are passive, websites allow users to obtain specific information when it is needed, and they usually provide the capacity to search for specific data (e.g. specific disease categories, locations, or time periods). Geographical mapping, which is automatically generated and displayed by several current systems, allows users to visualize clustering of events over time and space. More recently, smartphone apps have been developed that allow a combination of active and passive dissemination of information (and also allow users to report data back to the system).

With the rationale that it is not always possible to predict who will need a specific piece of information, many systems make their data available freely to anyone. Other systems make their information available to selected groups or individuals. Selectivity of dissemination may be based on the need to restrict access to confidential information (e.g. the Epi-X system of the US CDC, which is available only to vetted public health officials), or a paid subscription model may be used in order to recoup the costs of creating and maintaining the system.

**Illustration of Internet biosurveillance: Madeira Island dengue fever outbreak, October 2012**

To illustrate how an event is detected and observed to evolve through the lens of an Internet biosurveillance system, consider the October 2012 dengue fever outbreak in the Autonomous Region (island) of Madeira, a Portuguese territory located approximately 1000 km from the mainland [24]. It was the first dengue outbreak in Europe since 1928. With the keyword-based approach outlined in Table 1, MedISys [25] identified several Portuguese media articles on 5 September 2012, reporting that 'the mosquito *Aedes aegypti* struck again in force on Madeira' and 'left pharmacies without repellents and ointments' (peak A in Fig. 2) [3,26].

The data showed a sudden increase in dengue fever reporting in the Portuguese press, and MedISys issued an alert on Wednesday 3 October 2012 (peak B in Fig. 2). In more than 40 news articles, two confirmed and 22 suspected cases of dengue were reported. The story was run in newspapers in other European Union (EU) countries (Spain, Finland, etc.) on 4 October (peak C). On 5 October, 34 cases were reported as confirmed. The story was reported in the French and Belgian press on 10 October and in the UK press on 12 October, following a Reuters news wire story. An update from the Portuguese health authorities (Direcção-Geral da Saúde) was broadly discussed in the news on 8 November (peak D), and 517 confirmed cases were mentioned. The publication of the European Centre for Disease Prevention and Control (ECDC) Rapid Risk Assessment (RRA) update on 20 November met wide coverage, with over 80 articles being published within and outside the EU on 21 November (peak E).

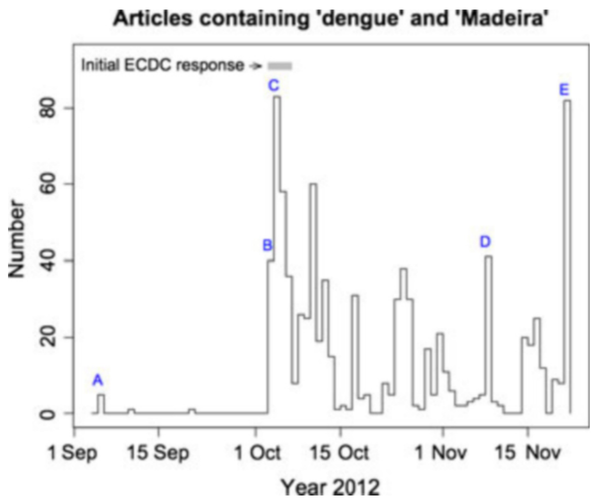

FIG. 2. Media reports on dengue fever on Madeira (number of articles per day, from 5 September to 21 November, 2012). The grey bar denotes the initial European Centre for Disease Prevention and Control (ECDC) response to the first alert, issued on 3 October 2012 (described in the text).

Internet biosurveillance played an important role in triggering an early public health response to this event (the grey bar in Fig. 2). On 3 October, the ECDC noticed a MedISys automated alert, and immediately began the process of verification by contacting the national health authorities of Portugal and gathering additional information from external experts in order to finalize an RRA for the EU population. Following this action, on 4 October, preliminary information about the outbreak was confidentially shared by the Portuguese health authorities with the EU/European Economic Area member states through the Early Warning and Reporting System (EWRS). The EWRS is the EU official communication restricted web-platform, and enables national authorities to exchange information on confirmed communicable disease events of potential international concern [27].

Early in the outbreak (near peak C in Fig. 2) on 6 October, the first ECDC RRA was internally finalized, and it was shared a few days later (10 October) with the EU/European Economic Area national health authorities through the EWRS. On 11 October, as agreed with the Portuguese authorities, the ECDC RRA was also made available online for the general public on the ECDC website [28]. In this outbreak, Internet biosurveillance played an important role in making international public health agencies aware of a potential outbreak earlier than would have been the case otherwise. This resulted in an early warning about the risk of infection in travellers returning from Madeira, where tourism is an important part of the economy. It also highlighted the risk of importation of dengue virus to continental Europe via air and sea cargo at the onset of the outbreak [29].

## Discussion

Outbreak data for human, animal and plant disease, available through informal media channels via the Internet, have been demonstrated to provide detection of anomalous disease events prior to official reporting [30–32]. In general, Internet media have the advantage of being timely, comprehensive, and available in any language from local and international sources. Such information can help to focus traditional surveillance efforts, and provides key data that can be used for a range of important public health purposes [33]. The value and pertinence of Internet biosurveillance have been demonstrated [34–36], and the approach has been integrated into the revised International Health Regulations [37]. Internet biosurveillance therefore contributes to early warning and situational awareness, and aims to trigger public health responses to mitigate outbreaks of infectious disease.

### Biosurveillance as an input to EI

Internet biosurveillance has influenced the way in which EI is gathered. To meet its objective of early warning, EI typically combines one or more Internet biosurveillance systems that are complementary to one another, in order to gain a broad view of topics and regions of interest. EI is widely used by national and trans-national public health organizations (e.g. the US CDC, the ECDC, the Public Health Agency of Canada, the French Institute for Public Health Surveillance (InVS), and the WHO) to strengthen their early detection functions [38–40]. The scope of EI and its final objective are broad, and vary according to the mandate and objectives of the implementing institution. For example, EI can be adapted to specific goals, including the early detection of public health emergencies, of specific infectious diseases only [1], and of public health events during mass gatherings [41]. Nevertheless, core functions and EI can be defined as the process of early detection, collection, verification, analysis and organization of information in relation to public health events [42,43]. EI processes integrate both formal and informal sources of information (e.g. Internet biosurveillance and traditional public health surveillance).

From the end-user perspective, the first EI step is the detection of pertinent raw signals. Official sources of health information (e.g. ministries of health, and surveillance networks) are typically easily identified, and their content is meant to support public health analysis. However, access to these may be difficult and constrained (for example, the information may be available only in the national language, and access to the information may be restricted), and their frequency of publication may not be appropriate for early disease detection. Therefore, informal sources (e.g. Internet media, discussion forums, and social networks) often represent the main source of signals. To collect and process large volumes of such material requires the use of Internet biosurveillance systems.

From the many raw signals observed from Internet biosurveillance systems, EI teams select information according to selection criteria defined by their public health institution. Following this, signals are verified; it is this verification phase that discriminates biosurveillance from EI. Verification consists of confirming and supplementing available information from additional and reliable sources, which are mainly networks of public health experts such as public health institutes, international institutions such as the WHO, World Organization for Animal Health, and ECDC, regional networks such as EpiSouth, laboratories, and non-governmental organizations.

Once verified, events are analysed to assess potential public health significance and potential national and/or international implications. Each is considered within its context and in the light of available scientific knowledge regarding spread, severity, and the efficacy of appropriate control measures [44]. Finally, following this analysis, the detected health threats are communicated to alert health authorities and to inform the public health community.

### Needs for future research

Above, we have described the current state of the field of Internet biosurveillance, from data collection to data utilization for EI. Internet technology has significantly advanced the disease surveillance landscape; however, gaps in biosurveillance processes exist, and many challenges lie ahead in the field; some of those are described below.

*Real-time signal detection.* Sifting through the vast array of multimedia information on the Internet in real time is challenging. The noise of non-specific reports and misinformation complicates signal detection. Moreover, identifying anomalous activity without an established multi-year baseline of reporting for a given disease in a particular region is an obstacle. Anomaly detection is a capability in some biosurveillance systems at present, but there is a need for more robust anomaly detection approaches, including better entity extraction, visual analytical modalities, clustering methods, etc. [45]. Moreover, more work is needed on capturing and analysing the data from multilingual sources through linguistic algorithms or automated translation.

*Data analysis.* Internet biosurveillance data typically cannot be analysed with traditional epidemiological approaches, owing to a lack of timely data verification and validation. For example, recognizing false-positive and false-negative events is problematic, owing to the lack of official comparison data or delays in diagnostic testing [33]. Frequencies of reports or events are often used for anomaly detection. However, identifying a common denominator (e.g. reports, events, articles, and sources) for analysis, and assigning a weight to sources based on accuracy, scope, and publication frequency, are not well established.

*Collaboration, networking, and participatory epidemiology.* Public self-reporting of events is increasingly recognized as benefiting disease detection. Extracting the data from participatory platforms (e.g. FluNearYou, Twitter, and Facebook) and utilizing it for early detection and surveillance is a critical area of current focus. For example, DIZIE, a project developed at the National Institute of Informatics in Tokyo, Japan, is used to visualize the extent to which Twitter data can detect/track infectious disease outbreaks [46]. More work is needed in this area, as health information sharing on social networking platforms has become prolific [47]. Users and public health experts can utilize this data in real time to track and assess disease situations [48].

Platforms with user-customizable features based on their specific needs and interests may make participatory modalities more attractive to a wider range of users. Also, more interactive functions for users (e.g. scoring option and comment field), may facilitate user interactions and information dissemination. An example of sharing and networking is the fully functional system for early alerting and reporting of potential chemical, biological, radiological, and nuclear events that has been developed by the Global Health Security Action Group through an extensive collaboration between the Joint research Centre of the European Commission and a team of risk assessment specialists from the G7+ Mexico countries [49].

## Transparency Declaration

The authors declare that they have no conflicts of interest.

## References

1. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—harnessing the Web for public health surveillance. *N Engl J Med* 2009; 360: 2153–2155, 2157.
2. Walters RA, Harlan PA, Nelson NP, Hartley DM. Data sources for biosurveillance. In: Voeller JG, ed. *Wiley handbook of science and technology for homeland security*, vol. 4. Hoboken: Wiley, 2010; 2431–2447.
3. Hartley DM, Nelson N, Walters R *et al.* The landscape of international event-based biosurveillance. *Emerg Health Threats J* 2010; 3: e3. doi 10.3134/ehtj.10.003.
4. Chunara R, Freifeld CC, Brownstein JS. New technologies for reporting real-time emergent infections. *Parasitology* 2012; 5: 1–9.
5. European Centre for Prevention and Disease Control. *Framework for a strategy for infectious disease surveillance in Europe*, 2006. Available from: http://www.ecdc.europa.eu/en/activities/surveillance/documents/0806_framework_surveillance_strategy_in_europe.pdf (last accessed 17 January 2013).
6. The White House. *Homeland security presidential directive 21 (HSPD-21)*. Public Health and Medical Preparedness. October 18, 2007. Available from: http://www.fas.org/irp/offdocs/nspd/hspd-21.htm (last accessed 8 February 2013).
7. Collier N, Doan S, Kawazoe A *et al.* BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics* 2008; 24: 2940–2941.
8. Mykhalovskiy E, Weir L. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. *Can J Public Health* 2006; 97: 42–44.
9. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J Am Med Inform Assoc* 2008; 15: 150–157.
10. Cowen P, Garland T, Hugh-Jones ME *et al.* Evaluation of ProMED-mail as an electronic early warning system for emerging animal diseases: 1996 to 2004. *J Am Vet Med Assoc* 2006; 229: 1090–1099.
11. Yangarber R, Jokipii L, Rauramo A, Huttunen S. Extracting information about outbreaks of infectious epidemics. In: Proceedings of the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing: HLT/EMNLP-2005. Vancouver, Canada, 2005; pp. 22–23. Available at: http://aclweb.org/anthology/H/H05/H05-2012.pdf [accessed 19/06/2013].
12. St Louis C, Zorlu G. Can Twitter predict disease outbreaks? *BMJ* 2012; 344: e2353.
13. Morse SS. Public health surveillance and infectious disease detection. *Biosecur Bioterror* 2012; 10: 6–16.
14. Collier N, Son NT, Nguyen NM. OMG U got flu? Analysis of shared health messages for bio-surveillance. *J Biomed Semantics* 2011; 2(suppl 5): S9. doi: 10.1186/2041-1480-2-S5-S9.
15. Salathé M, Bengtsson L, Bodnar TJ *et al.* Digital epidemiology. *PLoS Comput Biol* 2012; 8: e1002616.
16. Freifeld CC, Chunara R, Mekaru SR *et al.* Participatory epidemiology: use of mobile phones for community-based health reporting. *PLoS Med* 2010; 7: e1000376.
17. Mawudeku A, Lemay R, Werker D, Andraghetti R, St John R. The Global Public Health Intelligence Network. In: M'ikanatha N, Lynfield R, Van Beneden CA, de Valk H, eds. *Infectious disease surveillance*. Oxford: Blackwell Publishing, 2008; 304–317.
18. Torii M, Yin L, Nguyena T *et al.* An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. *Int J Med Inform* 2011; 80: 56–66.
19. Mantero J, Belyaeva J. How to maximise event-based surveillance web systems: the example of ECDC/JRC collaboration to improve the performance of MedISys. JRC European Commission Publication Repository, 2011. Available from: http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/16206/1/lb-na-24763-en-c.pdf (last accessed 27 March 2013).
20. Keller M, Freifeld CC, Brownstein JS. Automated vocabulary discovery for geo-parsing online epidemic intelligence. *BMC Bioinformatics* 2009; 10: 385. doi:10.1186/1471-2105-10-385.
21. Box GEP, Jenkins GM. *Time series analysis, forecasting and control*. San Francisco, CA: Holden-Day, 1970.

22. Wiener N. *Extrapolation, interpolation, and smoothing of stationary time series*. Cambridge, MA: MIT Press, 1964.
23. Gilbert P. *Dynamic system estimation (time series package)*. Available from: http://cran.r-project.org/web/packages/dse/index.html (last accessed 27 March 2013).
24. Sousa CA, Clairouin M, Seixas G *et al.* Ongoing outbreak of dengue type 1 in the Autonomous Region of Madeira, Portugal: preliminary report. *Euro Surveill* 2012; 17: pii=20333. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20333.
25. Jens L, Belyaeva J, Steinberger R *et al.* MedISys: Medical Information System. In: Asimakopoulou E, Bessis N, eds. *Advanced ICTs for disaster management and threat detection: collaborative and distributed frameworks*. Hershey, PA: IGI Global, 2010; 131–142.
26. Jens L, Steinberger R, Weber T *et al.* Internet surveillance systems for early alerting of health threats. *Euro Surveill* 2009; 14: 1–2.
27. Guglielmetti P, Coulombier D, Thinus G, Van Loock F, Schreck S. The Early Warning and Response System for communicable diseases in the EU: an overview from 1999 to 2005. *Euro Surveill* 2006; 11: pii=666. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=666 (last accessed 27 March 2013).
28. ECDC Rapid Risk Assessment. Autochthonous dengue cases in Madeira, Portugal—10 October 2012. Available from: http://ecdc.europa.eu/en/publications/Publications/Dengue-Madeira-Portugal-risk-assessment.pdf (last accessed 15 March 2013).
29. ECDC. Mission Report 'Dengue outbreak in Madeira, Portugal: October–November 2012'. Available from: http://ecdc.europa.eu/en/publications/Publications/dengue-outbreak-madeira-mission-report-nov-2012.pdf (last accessed 6 May 2013).
30. Mondor L, Brownstein JS, Chan E *et al.* Timeliness of nongovernmental versus governmental global outbreak communications. *Emerg Infect Dis* 2012; 18: 1184–1187.
31. Chan EH, Brewer TF, Madoff LC *et al.* Global capacity for emerging infectious disease detection. *Proc Natl Acad Sci USA* 2010; 107: 21701–21706.
32. Woodall J. Official versus unofficial outbreak reporting through the Internet. *Int J Med Inform* 1997; 47: 31–34.
33. Nelson NP. Advantages and challenges of using Internet media for disease detection and tracking. *Phytopathology* 2012; 102: 161–162.
34. Heymann DL, Rodier GR. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *Lancet Infect Dis* 2001; 1: 345–353.
35. Wilson K, Brownstein JS. Early detection of disease outbreaks using the Internet. *CMAJ* 2009; 180: 829–831.
36. Hoen AG, Keller M, Verma AD, Buckeridge DL, Brownstein JS. Electronic event-based surveillance for monitoring dengue, Latin America. *Emerg Infect Dis* 2012; 18: 1147–1150.
37. World Health Assembly. Revision of the International Health Regulations. World Health Assembly Resolution 58.3. 23 May 2005. Available from: http://who.int/csr/ihr/IHRWHA58_3-en.pdf (last accessed 27 March 2013).
38. Kaiser R, Coulombier D, Baldari M, Morgan D, Paquet C. What is epidemic intelligence, and how is it being improved in Europe? *Euro Surveill* 2006; 11: E060202.4. Available from: http://www.eurosurveillance.org/ew/2006/060202.asp#4 (last accessed 27 March 2013).
39. Blench M. Global Public Health Intelligence Network (GPHIN). In: *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*. Waikiki, HI: Elsevier, 2008; 1–5
40. ECDC Epidemic Intelligence Group. *ECDC Epidemic Intelligence e-tutorial, a tool to learn more about the detection and assessment of public health threats*. 2011. Available from: http://external.ecdc.europa.eu/EI_Tutorial/course.htm (last accessed 27 March 2013).
41. Khan K, Freifeld CC, Wang J *et al.* Preparing for infectious disease threats at mass gatherings: the case of the Vancouver 2010 Olympic Winter Games. *CMAJ* 2010; 182: 579–583.
42. Paquet C, Coulombier D, Kaiser R, Ciotti M. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro Surveill* 2006; 11: 212–214. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=665 (last accessed 27 March 2013).
43. Rotureau B, Barboza P, Tarantola A, Paquet C. International epidemic intelligence at the Institut de Veille Sanitaire, France. *Emerg Infect Dis* 2007; 13: 1590–1592.
44. World Health Organization. *Rapid risk assessment of acute public health events*. Available from: http://whqlibdoc.who.int/hq/2012/WHO_HSE_GAR_ARO_2012.1_eng.pdf (last accessed 27 March 2013).
45. Nelson NP, Brownstein JS, Hartley DM. Event-based biosurveillance of respiratory disease in Mexico, 2007–2009: connection to the 2009 influenza A(H1N1) pandemic? *Euro Surveill* 2010; 15: pii=19626. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19626 (last accessed 27 March 2013).
46. Collier N, Doan S. Syndromic classification of Twitter messages. In: Akan O, Bellavista P, Cao J, Dressler F, Ferrari D, Gerla M, Kobayashi H, Palazzo S, Sahni S, Shen X, Stan M, Xiaohua J, Zomaya A, Coulson G, eds, *Lecture Notes of the Institute for Computer Science, vol. 91, Social Informatics and Telecommunications Engineering*. Berlin: Springer, 2012; 186–195.
47. Fox S. The Social Life of Health Information, 2011. Report of the Pew Research Center, Washington, DC, 2011. Available from: http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx (last accessed 13 March 2013).
48. Anonymous. *Bulletin: London 2012*. HPA, ECDC, WHO, 2012. Available from: http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317135289768 (last accessed 13 March 2013).
49. Barboza P, Vaillant L, Mawudeku A *et al.* Evaluation of epidemic intelligence systems integrated in the Early Alerting and Reporting project for the detection of A/H5N1 influenza events. *PLoS ONE* 2013; 8: e57252.