THE MILBANK QUARTERLY
A MULTIDISCIPLINARY JOURNAL OF POPULATION HEALTH AND HEALTH POLICY

Review Article

Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review

EDWARD VELASCO,* TUMACHA AGHENEZA,* KERSTIN DENECKE,† GÖRAN KIRCHNER,* and TIM ECKMANNS*

*Robert Koch Institute; †L3S Research Center

Context: The exchange of health information on the Internet has been heralded as an opportunity to improve public health surveillance. In a field that has traditionally relied on an established system of mandatory and voluntary reporting of known infectious diseases by doctors and laboratories to governmental agencies, innovations in social media and so-called user-generated information could lead to faster recognition of cases of infectious disease. More direct access to such data could enable surveillance epidemiologists to detect potential public health threats such as rare, new diseases or early-level warnings for epidemics. But how useful are data from social media and the Internet, and what is the potential to enhance surveillance? The challenges of using these emerging surveillance systems for infectious disease epidemiology, including the specific resources needed, technical requirements, and acceptability to public health practitioners and policymakers, have wide-reaching implications for public health surveillance in the 21st century.

Methods: This article divides public health surveillance into indicator-based surveillance and event-based surveillance and provides an overview of each. We did an exhaustive review of published articles indexed in the databases PubMed, Scopus, and Scirus between 1990 and 2011 covering contemporary event-based systems for infectious disease surveillance.

Findings: Our literature review uncovered no event-based surveillance systems currently used in national surveillance programs. While much has been done

Address correspondence to: Edward Velasco, Robert Koch Institute, Department for Infectious Disease Epidemiology, Healthcare Associated Infections, Surveillance of Antimicrobial Resistance and Consumption, Seestrasse 13353, Berlin, Germany (email: VelascoE@rki.de).

The Milbank Quarterly, Vol. 92, No. 1, 2014 (pp. 7-33)

to develop event-based surveillance, the existing systems have limitations. Accordingly, there is a need for further development of automated technologies that monitor health-related information on the Internet, especially to handle large amounts of data and to prevent information overload. The dissemination to health authorities of new information about health events is not always efficient and could be improved. No comprehensive evaluations show whether event-based surveillance systems have been integrated into actual epidemiological work during real-time health events.

Conclusions: The acceptability of data from the Internet and social media as a regular part of public health surveillance programs varies and is related to a circular challenge: the willingness to integrate is rooted in a lack of effectiveness studies, yet such effectiveness can be proved only through a structured evaluation of integrated systems. Issues related to changing technical and social paradigms in both individual perceptions of and interactions with personal health data, as well as social media and other data from the Internet, must be further addressed before such information can be integrated into official surveillance systems.

Keywords: surveillance, health information, Internet, social media.

Recent major health events such as severe acute respiratory syndrome coronavirus (SARS-CoV) in Asia (2002-2003), pandemic H1N1/09 influenza virus worldwide (2009), and the large outbreak of *Escherichia coli* O104:H4 in Germany (2011) have prompted infectious disease scientists at government agencies, university centers, and international health agencies to invest in improving methods for conducting infectious disease surveillance.[1,2] Opportunities for improvement, however, vary and are based on the distinctive features of existing types of infectious disease surveillance, which have been developed over time to address the various critical components in public health efforts against disease. Standard infectious disease surveillance methodologies have been derived from indicator-based surveillance and event-based surveillance.

Indicator-based surveillance systems are the oldest, most common, and most widely used form of infectious disease surveillance by regional, national, and international public health agencies. These systems are designed to collect and analyze structured data based on established surveillance and monitoring protocols tailored to each disease (ie, used for calculating the incidence, seasonality, and burden of disease), in order

to gather relevant information about populations of interest to detect changes in trends or distributions in the population. Data on such indicators are reported by health care providers and diagnostic laboratories, by legal mandate or voluntary agreement, and are collected by surveillance specialists in governmental health agencies. This information then can be verified through communication between the governmental health agencies and the persons collecting the data in health care settings.

Indicator-based surveillance systems often contain reliable statistical methods that have been established to compare the observed number of cases of pathogens with an expected rate. The goal is to find increased numbers or clusters at a specific time, period, and/or location that might indicate a threat. Statistical methods set against thresholds of increased cases or clusters are crucial to finding potential health events. They are based on the relevant attributes of each infectious disease, such as epidemiological parameters like regional incidence, seasonality, and the known burden of disease. Thresholds can also be adjusted using statistical algorithms to vary sensitivity and specificity so that the detection procedure is refined to better suit the needs of the epidemiological situation for a disease or a specific area. This helps epidemiologists by giving them a greater capacity to monitor additional information that might signal threats to public health.

The ability of indicator-based surveillance systems to detect potential threats more quickly is lacking, however. Although generating signals based on statistical thresholds can provide an aggregation that will speed up a threat assessment, the data itself may not be the most recent. First, there is often a time lag between the occurrence of an event and the indicator-based surveillance. That is, data input and retrieval for indicator-based surveillance often rely on specific case definitions and reporting requirements that differ for physicians in hospital and community care and for laboratories, thereby causing delays in reporting to health agencies. Delays also may be caused by time lags between reporting procedures from the reporting bodies and the authorities who receive, store, and process the data, that is, by the structure of notification systems in official public health agencies that often trickle up from the local, state, and federal levels. Second, indicator-based systems are sometimes poorly equipped to detect new or unexpected occurrences of disease, owing to the predefined epidemiological attributes assigned to each infectious disease for which information is collected. This was true during the first cases of SARS-CoV in 2002 and pandemic H1N1/09

influenza in 2009, which at first were not detected because the existing systems could track only the clinical and epidemiological attributes for corona or influenza infections that had already been discovered and defined, but not new strains of viral infections. Incidentally, such shortfalls provided the impetus for the systemic improvement of indicator-based systems. By demonstrating the importance of detecting unknown but similar diseases, it became evident that new data sources and methods for monitoring such data were critical.[3] As a result of the SARS-CoV epidemic, for example, health agencies began to seriously consider ways to monitor symptoms and syndromes (ie, clusters of symptoms for particular diseases) in order to provide appropriate and fast detection with the most efficient use of required human resources.

Similar to indicator-based surveillance, event-based surveillance is based on the organized and rapid capture of information about events that can be a risk to public health. But rather than relying on official reports, this information is obtained directly from witnesses of real-time events or indirectly from reports transmitted through various communication channels (eg, social media or established routine alert systems) and information channels (the news media, public health networks, and nongovernmental organizations) (Table 1). Monitoring that relies on data from these Internet sources can be used to detect threats not specifically found by indicator-based surveillance, since this information relies less on data structured and filtered through the aforementioned preestablished structures for surveillance. Event-based surveillance can identify events faster than indicator-based reporting procedures can, and it can detect events that occur in populations not able to access formal channels for reporting. In addition, event-based surveillance can be used with other established indicator-based methods, thereby enhancing the combined arsenal for combatting critically prevalent pathogens with a high threat potential, such as influenza virus or *Escherichia coli*. The scientific literature recently referred to this comprehensive framework of combined activities from both indicator-based surveillance and event-based surveillance systems as "epidemic intelligence," a contemporary understanding of the 1950s term with roots in public health innovation for surveillance systems at the US Centers for Disease Control and Prevention (CDC) and the establishment of the Epidemic Intelligence Service (EIS).[4-8]

Event-based surveillance continues to offer innovation for public health surveillance, for example, by capturing information about events

TABLE 1
Indicator-Based and Event-Based Surveillance Systems

	Indicator-Based	Event-Based
Timeliness of Data Input	Information is input as soon as it is made available. Timing is set immediate/ weekly/monthly. Possible delay between identification and notification.	Information is input as soon as it occurs. Timing varies, depending on when the data are available from those who have the information. Possible delay between identification and reporting.
Reporting Structure	Clearly defined. Reporting forms. Reporting dates. Teams analyze data at regular intervals. Moderated.	Predefined or not predefined structure. Reporting forms flexible for qualitative and quantitative data. Teams analyze data at any time. Moderated or not moderated (eg, automatic).
Timeliness of Detection	Depends on the time from the occurrence of the event (ie, the onset of the disease) until a diagnosis is available that fulfills a case definition. Depends on the time it takes for reporting through the stages of a hierarchical reporting structure.	Depends on the time from the occurrence of the event (ie, onset of the disease) until the first mention occurs, which might be before diagnostic confirmation is available. Depends on the ability of the system and the time it takes to pick up a signal and to interpret it correctly.
Thresholds for Signal Generation	Statistical methods are employed to identify increased numbers (clusters) in time or in space (or combinations of both) to generate a	Signals are differentially generated (eg, human indexing in ProMED-mail) but rarely with automated statistical methods that

Continued

TABLE 1—*Continued*

	Indicator-Based	Event-Based
	signal for potential event-detection.	identify increased numbers (clusters) in time or in space (or combinations of both) to generate a signal for potential event-detection.
Trigger for Follow-up or Action	Crossing a predefined threshold leads to an in-depth analysis and further information gathering.	A confirmed event or hint at an event leads to further information gathering, verification.

that may not otherwise be detected in the routine collection of data from indicator-based surveillance. Events that may be detected in event-based surveillance include the following:

- Events, such as SARS, that are emerging or rarely occur and thus are not specifically part of the purview of standard indicator-based surveillance.
- Events that occur in real time but have not been detected by indicator-based surveillance, such as those events delayed by the required reporting procedures of notifying the designated health authority.
- Events that occur in populations that do not access health care through formal channels or in which formal, indicator-based systems do not exist, such as events that occur in populations in rural areas or countries with a less established infrastructure for surveillance.

Health information monitored via the Internet and social media is an important part of event-based surveillance and is most often the source on which many existing event-based surveillance systems focus. Existing systems for such event-based monitoring contain useful retrieval features that give epidemiologists and public health scientists involved in surveillance quick access to information compiled from many media and news sources.[9,10] Other new health information

technologies using new data sources from the Internet are important drivers of innovation in global surveillance, speeding up the collection and transmission of information to allow for better emergency preparedness or responses.[11] In research, event-based surveillance using data from the Internet, especially emails and online news sources, has been shown to identify surveillance trends comparable to those found using established indicator-based surveillance methods.[12-14] In practice, however, such systems have not yet been widely accepted and integrated into the mainstream for use by national and international health authorities.

We reviewed event-based surveillance systems that have actually been used, in order to examine the usefulness of event-based surveillance to existing surveillance efforts and its potential to improve future comprehensive infectious disease surveillance systems.

A Systematic Review of Event-Based Surveillance

We conducted a systematic review to identify all currently established event-based surveillance systems used in infectious disease surveillance and to look at the type of data collected, the mode of data acquisition used by the system, and the overall purpose and function of each system. As members of a national scientific institute, our aim was to help health policy decision makers decide whether to incorporate new methods into comprehensive programs of surveillance that already contain established indicator-based surveillance.

The previous work in this area includes a systematic review, by Bravata and colleagues, of 17,510 peer-reviewed articles and 8,088 websites on surveillance systems for the early detection of bioterrorism-related diseases, which evaluated the potential utility of existing surveillance systems for illnesses and syndromes related to bioterrorism only.[15-17] Another review of peer-reviewed articles by Vrbova and colleagues synthesized surveillance systems for emerging zoonotic diseases with selected criteria used to evaluate those systems.[18] Corley and colleagues helped US federal government agencies compile aspects and attributes associated with operational considerations in the development, testing, and validation of event-based surveillance; and Hartley and colleagues drew up an outline of technical Internet biosurveillance processes.[19,20]

Although this work is important, these reviews do not provide systematically collected details of event-based systems used in practice.

Methods

We searched for peer-reviewed articles published in the indexes Pubmed, Scopus, and Scirus between 1990 and 2011[21-23] as well as English-language studies of infectious disease surveillance (and specifically event-based surveillance) and outbreak detection in human health and medicine. We excluded articles on bioterrorism (for which there is less possibility of pathogen threat), articles on solely technical aspects of system implementation or security (eg, video surveillance), those covering sentinel surveillance systems (ie, those set up randomly, periodically, or in another unsystematic way), any surveillance not based on infectious diseases, and articles without available abstracts. We used extraction criteria to collect comparable data on each system. Appendix 1 provides a detailed overview of the search strategy and methods, and the study's complete protocol also is available.[24]

Results

Our systematic review yielded 13 event-based systems used in practice and for which complete information based on our extraction criteria was available (Tables 2 and 3).

System Category

Event-based surveillance systems can be classified as news aggregators, automatic systems, or moderated systems.[25] *News aggregators* collect articles from several sources that are commonly filtered by language or country. Although their users have easy access to many sources through a common portal, they must examine each article individually. *Automatic systems* go beyond this by adding a series of steps for analysis but differ in the levels of analysis performed, in the range of information sources, in language coverage, in the speed of delivering information, and in methods for visualization. In *moderated systems*, information is processed entirely by human analysts or is first processed automatically and then

TABLE 2
List of Event-Based Systems Identified

No.	System Name (literature reference)	Category	Country	Year Started
3.1	Argus[43,51]	Moderated	USA	2004
3.2	BioCaster[52]	Automatic	Japan	2006
3.3	EpiSPIDER[34,53]	Automatic	USA	2006
3.4	EWRS[54]	Moderated	EU	1998
3.5	GOARN[55]	Moderated	Multiple[a]	2000
3.6	GODSN[56]	Automatic	USA	2006
3.7	GPHIN[26,57]	Moderated	Canada	1997
3.8	HealthMap[58-62]	Automatic	USA	2006
3.9	InSTEDD[63]	Moderated	USA	2006
3.10	MedISys and PULS[64,65]	Automatic	EU	2004
3.11	MiTAP[66]	Automatic	USA	2001
3.12	ProMED-mail[13,67-69]	Moderated	USA	1994
3.13	Proteus-BIO[11]	Automatic	USA	2000

[a]GOARN is a WHO-coordinated network

analyzed by people. Moderated systems offer a screening for epidemiological relevance of the data found within the information before it is presented to the user.

Although each of the systems that we reviewed has different goals (mostly pertaining to various national, international, and regional audiences), they all foster the communication of health events or threats in the infectious disease community of scientists, physicians, epidemiologists, public health officials, policymakers, and politicians.

The systems overwhelmingly rely on media sources for data input, including local and national newspapers; news broadcasts; websites; news wires; or even short message service (SMS), the text messaging service component of phone, web, or mobile communication systems.[26] Some of the systems already have been incorporated into other larger systems. For example, GOARN links 110 existing networks, and GPHIN collects data already processed with ProMED-mail.[27] Surveillance scientists then review this information to assess its epidemiological significance and to support decision making. But because these data are not structured, epidemiologists must spend more time and energy determining their relevance to a particular situation of interest.

TABLE 3
Data Extraction Criteria and Data Collected

No.	Criteria	Description
1	System name	The name of the system
2	System category	The category: news aggregator, automated, or moderated systems
3	Country	Country where the system was founded
4	Year started	The year the system started operating
5	Coordinating organization	The unit that operates the system
6	Purpose	The purpose of the system
7	Geographic scope	The geographic area covered
8	Language	The number of languages the system covers or gets information from
9	Disease type	Type of diseases covered by the system; >3 as "multiple infectious diseases"
10	Accessibility	The type of access: freely accessible to the general public vs restricted access
11	Data collection and processing	The methods employed to collect the necessary data, and data analysis
12	Dissemination of data	The method for data dissemination
13	Users	The organizations or individuals using the event-based system
14	System evaluation	The existence of a previous system evaluation
15	Homepage	The web location of the system

Coordinating Organization

We identified three types of coordinating bodies for event-based systems: those based at or in cooperation with universities (Argus, BioCaster, GODSN, HealthMap, and Proteus-BIO), NGOs (GOARN, MedISys, MiTAP, and ProMED-mail) and governmental agencies (EWRS, EpiSPIDER, GPHIN and InSTEDD).

Purpose

Each of these systems has a different aim: (1) to improve early detection, (2) to enhance communication or collaboration, and (3) to supplement other existing systems. Ten of the systems are intended to

improve early detection: Argus, BioCaster, GOARN, GODSN, GPHIN, HealthMap, InSTEDD, MedISys, MiTAP, and Proteus-BIO. Two of the systems are meant to enhance communication or collaboration (EWRS and ProMED-mail), and one system supplements another (EpiSPIDER for ProMED-mail).

Geographic Scope

All the systems cover 2 or more countries, but their jurisdictions could be classified as (1) those that monitor worldwide (EpiSPIDER, GOARN, GODSN, GPHIN, HealthMap, InSTEDD, MiTAP, ProMED-mail, and Proteus-BIO); (2) those confined to a particular region, including BioCaster (mostly countries in the Asia-Pacific region), EWRS (restricted to events of interest to the European Union [EU] and the European Economic Area [EEA]), and MedISys (other regions, particularly Europe); and (3) those monitoring regions other than that where the system is based (eg, Argus, though based in the United States, does not monitor there). Most of the event-based systems are based in the United States, followed by the EU, and only 1 each is based in Canada and Japan.

Language

Five of the systems use English only (EpiSPIDER, EWRS, GODSN, InSTEDD, and Proteus-BIO), though other systems are multilingual: Argus (34 languages), BioCaster (8 languages), GPHIN (8 languages), HealthMap (5 languages), MedISys (43 languages), MiTAP (8 languages), ProMED-mail (7 languages), and GOARN (operates in English, but may also be multilingual, since it is a network collaboration between the World Health Organization [WHO] and the United Nations [UN] member states).

Disease Type

All the event-based systems that we reviewed focused on outbreaks of different and multiple infectious diseases, with some systems, such as Argus (130), BioCaster (102), and HealthMap (170), collecting information on more than 100 diseases.

Accessibility

We observed 5 levels of access: (1) freely and publicly available (HealthMap, EpiSpider, GODSN, and Proteus-BIO), (2) available with a free subscription (ProMED-mail, BioCaster, and MiTAP), (3) available with a paid subscription (GPHIN, whose subscribers include governmental organizations, NGOs, and universities), (4) access restricted to certain public health officials (EWRS, Argus, GOARN, and InSTEDD), and (5) mixed-level access (MedISys, offering free but restricted access to the public and outside the European Commission [EC] and full access to officials in the EC).

Accessibility varies from system to system, depending on both the scope of the system and the intended audience. While it is important to offer freely accessible information, some sensitive information (eg, personal data or other confidential data) is often filtered in specific ways among public health officials with specific restricted access. GPHIN has restricted access for organizations with an established public health mandate, with access varying according to factors like the organization's size and number of users. InSTEDD is one of the few systems using information to advise organizations like the UN, WHO, and CDC on strategic implementation. Such systems, like EWRS, provide, within a closed network, timely information for preparedness, early warning, and responses.

Data Collection and Processing

Each event-based system acquires data differently. Some collect information directly from sources on the Internet (eg, RSS feeds or electronic mailing lists); others collect both from formal members and informal sources; and still others collect from subscribers or members only. Ten systems collect from the Internet (Argus, BioCaster, EpiSPIDER, GODSN, GPHIN, HealthMap, InSTEDD, MedISys, MiTAP, and Proteus-BIO), and 2 systems collect from both formal members and informal sources (EWRS and GOARN). ProMED-mail is the only system obtaining firsthand information from its subscribers.

Most of the systems we studied function as news aggregators. News aggregators (eg, Google News) use RSS to collect real-time news feeds from thousands of news sources from around the world, and many systems deal with a huge amount of information each day. MediSys, for

example, monitors an average of 50,000 news articles per day from about 1,400 news portals in 43 languages. GPHIN processes from 2,000 to 3,000 news items per day, of which about a quarter are irrelevant or duplicates.[26] Many of the event-based systems utilize text-mining technology to extract only relevant data, and most have sophisticated processing systems of filtering and classifying relevant information to reduce the amount of data.

Source data (ie, event-based data retrieved from the Internet) should be reviewed for epidemiological relevance, either by human epidemiologists or automated systems. This is technologically simple but time-consuming and expensive, with human moderation having a different role in each system. The information provided through ProMED-mail, for example, is validated and confirmed by humans. EWRS utilizes an informatics tool that filters and relays information to users via a web-based system that links contact members of the EWRS network.

Human input, hypothesis generation, and review are important components of systems. InSTEDD and GPHIN incorporate human input and review, allowing users to add comments, tags, and ranks during the data-processing phases and confirmation and feedback during the dissemination phases.

Systems without human moderation often focus on data sources that already have been validated. Many systems contain new data on outbreaks or diseases, but only some are relayed as firsthand, primary information. Other data are reported as secondary sources like newspaper articles. Although this information can be useful to surveillance epidemiologists who monitor data and conduct research on a known infectious disease area, because these events already have been reported, it does not help epidemiologists interested in the early warning and alert potential for unknown or new infectious disease areas. Because MedISys offers no human mediation in collating information sources and articles, all information must be examined in order to learn more about the outbreak or event in question. Accordingly, how the information is presented is less easily adapted for use in daily practice.

Almost all the systems not relying on human moderation are automated with thresholds used to reduce noise and to present only the most relevant data. MediSys uses a scraper software, for example, that automatically generates an RSS feed from webpages and applies a text-extraction process, which then enables content analysis using analytical technology.[28] The text-extraction process uses document heuristics, an

experience-based technique for computer learning that is applied to the information to enable an intelligent decision about its relevance. The heuristics learn as their output is verified against a set threshold for the epidemiological attributes of health events that have been extracted, thus improving monitoring over time. The system aggregates the extracted events into outbreaks, across multiple documents and sources, before returning the extracted information to the system. Users of the system often prefer a more structured approach, but it may present too much information (in some cases up to 1,000 events per day). The large amount of "information noise" also may be a hindrance, since users are then required to sift through it manually. HealthMap alleviates noise by integrating data from a variety of electronic sources that already have been moderated (ProMED-mail, WHO-validated official alerts, and the Eurosurveillance RSS multinational outbreak news site), all of which are fed into a classification engine (ie, a parser), which uses the information to produce disease and location output codes. Once classified, articles are filtered into a category and stored in a database.

Dissemination of Data

Three systems are disseminated on a geographic map: BioCaster, EpiSPIDER, and HealthMap; and 4 systems are disseminated through a website or news aggregator: MedISys, MiTAP, ProMED-mail, and Proteus-BIO. We found 6 systems that were disseminated through a secured or restricted portal: Argus, EWRS, GOARN, GODSN, GPHIN, and InSTEDD.

Discussion

Our systematic literature review demonstrates the diverse attributes in current, established, event-based surveillance systems. Our review also articulates the factors that might influence the integration of such surveillance activities into official systems. The usefulness of new information sources via event-based surveillance depends on whether the information can enhance the data collection from existing surveillance methods and also on several factors related to the challenges for all systems' acquisition of data on infectious disease surveillance.[29,30]

For most epidemiologists, the process of gathering data from the Internet is complex, as it includes text mining (searching for health-related content from websites or social media), preparation (extracting and filtering relevant health-related information), and presentation of only the most relevant content (disseminating the information). In general, data are acquired and processed either automatically or by people, often relying on individual technologies for users' interaction with the data to tag (mark or catalog) the information for future use, and to comment on the information (for sharing and collaboration with other scientists), which can also be used to inform machine-learning algorithms (eg, statistical filters for data retrieval, such as Bayesian models). Systems often use automatic programming interfaces (APIs), a type of filter through which the data are passed in order to extract specific information. This is a good way of managing large sources of data from the Internet, which can be cumbersome and contain much content not related to health. These APIs process, extract, augment, and compile the epidemiological attributes in the data (ie, metadata) from multiple sources. For example, health-related attributes could include the data source, a relevant health term, the location, and the time of transmission. Natural language-processing systems extract from the feeds such relevant concepts as disease names and references to a geographic location. The information is then often assigned a dissemination format based on the information type (ie, epidemiological attribute) retrieved, through a network (eg, GPHIN), the Internet, email (ProMED-mail), or SMS (EWRS), or it is plotted on a geographic information system (GIS) published on the web (eg, HealthMap). The time needed to get from a potential data source to extracting and presenting epidemiological information that can be used as quickly as possible for preparedness or responses is critical. The results vary widely, depending on the combination of technologies used and whether or not human mediation is involved.

Even though event-based surveillance systems have been much improved, they still have limitations:

Information is not always moderated by professionals or interpreted for relevance before it is disseminated to interested surveillance epidemiologists. Information retrieved from event-based systems can originate from either official sources who can be seen as trusted health specialists or unofficial sources, such as the public, who may or may not be health specialists. Information from unofficial sources is often not prescreened by professionals, so it can cause reliability issues and necessitate

moderation. Thus, moderation affects the quality of event-based information, compared with information from indicator-based systems, which almost always is provided by official sources.

There is no standardized system for the frequency of updates, often resulting in too much information. Information from event-based information components that use news aggregators often is incomplete and may not be timely. Data may be obsolete by the time it is picked up by epidemiologists because some information may have been published by news agencies after health organizations knew about an event or problem. In existing event-based surveillance systems, the frequency of updates varies from approximately several to hundreds of notifications per day, depending on the system.

Algorithms and statistical baselines are not well developed. Until now, event-based systems have not applied algorithms and statistical baselines to information before it is presented to users, which is a standard feature of most established indicator-based systems. Event-based systems often receive a high volume of information per day, which can overwhelm epidemiologists at public health agencies who perform surveillance and may be seen as a hindrance, since users are then required to spend time moderating the retrieved information.

New information about health events or probable cases is not always disseminated efficiently. Event-based systems filter and organize information about potential events of interest before it is presented to users. Information indexed by topic or subject enables users to decide whether they need to do more research. Some systems use online watch boards offering lists or tables of information on events; others rely on SMS; and still others provide options for other notification, like the ability to subscribe to an RSS-feed or through other web capabilities, like Twitter.

The Future of Public Health Surveillance

Some studies have shown that automated methods and technologies like those used in event-based surveillance can rapidly signal the detection of infectious diseases.[31,34] In addition to speeding up detection by bypassing traditional indicator-based surveillance structures, event-based surveillance can also provide innovation in settings with weak or underdeveloped surveillance systems. In developing countries with a large disease burden, surveillance infrastructures that can use health information in the absence of traditional surveillance institutions can be critical

to prevent an outbreak or reduce its impact.[35] Recent work has begun in this area to seek out information on health threats using mobile phone technology, Internet-scanning tools, email distribution lists, or networks that complement the early warning function of routine surveillance systems.[36-40] Our research showed that the majority of event-based surveillance systems are based in North America and Europe, with fewer local, event-based systems monitoring epidemic threats in Africa, Asia, the South Pacific, and South America. Guidance and training to create such systems on the ground should be considered, as this can lead to a faster assessment of health threats and a more rapid response by local authorities.

Previous evaluations of event-based surveillance systems have been limited, so we have very few examples to draw from.[41,42] Although explored since the mid-2000s, largely in response to the SARS-CoV epidemic, event-based surveillance has yet to be fully integrated into public health surveillance systems. Evidence showing the added value to traditional infectious disease surveillance methods is sparse.[43] The development of appropriate metrics for monitoring and evaluating the quality of the data in event-based surveillance systems has become a priority but has just begun. Standard guidelines for the evaluation of surveillance systems offer much information about the attributes needed for measuring the appropriateness and effectiveness of specific systems. Most guidelines, however, rely on attribute descriptions taken from traditional or indicator-based surveillance. These have seldom been adapted to address specific concerns about the new information from event-based surveillance systems and may be inadequate.

Those standard operating procedures, tools, and guidance for event-based surveillance that do exist—as is often the case with indicator-based surveillance as well—are not universally applicable, since different regions, countries, and smaller jurisdictions must adapt the surveillance systems to their particular needs. In 2005, the WHO established international health regulations (IHR) for surveillance activities that offer the WHO's 193 member states a multilateral legal framework for surveillance, notification, and responses to disease outbreaks and other emergencies with potential international public health implications.[44-46] The new IHR require the WHO and its members to develop real-time event management systems for addressing public health risks and emergencies of international concern along with the usual epidemiological tools.

Regulating the identification of disease outbreaks and other emergencies with potential international public health implications also requires technical advice to develop adequate surveillance activities. Innovative methods for screening information will no doubt become a priority as definitions of event-based surveillance, recommendations for implementing activities, and evaluations of surveillance systems are established and grow. Event-based surveillance utilizing the fast electronic communication and news sources on the Internet have been widely successful and will likely continue to help improve event-based surveillance.[32,33,47]

The Challenges of Integrating Event-Based Surveillance

Our literature review uncovered no systems that are currently part of national programs for surveillance. Instead, they are used intermittently as complementary sources of information. We also have little information about whether or not these systems have been integrated into actual work during real-time health events. The current literature does indicate that event-based surveillance could improve official surveillance activities, but systematic evaluation within a public health agency is needed before it can be realized.[48] This is a circular dilemma, since the willingness to integrate is rooted in the lack of effectiveness studies, yet such effectiveness can be proved only by the structured evaluation of integrated systems.

The number of factors necessary for integrating such services should not be underestimated. These include time-consuming and costly collaboration with statisticians, Internet and media experts, and computer scientists to work on components of data acquisition, data processing and filtering, personalization of results, and automation for dissemination to epidemiologists. Once developed, these technical services will require staff to train and support scientific users (eg, epidemiologists) in monitoring infectious diseases, since such activities are not yet part of regular training programs for epidemiology or public health.

Another challenge is the creation of a strategy to compare and cross-verify indicator-based and event-based data, since they can differ, especially in regard to syndromes and locations, which makes it difficult to make conclusions based on specific epidemiologic attributes.[49]

Nonetheless, solutions must be found, perhaps newly elaborated epidemiologic ontologies for text mining and a related process of continuous improvement. Can all this be done, and is it worth it?

The benefits to epidemiologists clearly are the data retrieved for analysis and potential public health warnings and intervention. In particular, the data's value to the early warning and detection of outbreaks needs to be demonstrated by evaluating the content found in social media and other Internet data sources. Primary content (ie, firsthand observations) provided by the users themselves is valuable, as it would likely signal a potential health threat more quickly. Here again, a usability study over time is needed to help show how useful primary content would be. Online media, weblogs, scientific and nonscientific discussion forums, and direct electronic communication could help expand event-based surveillance activities, although they may have unforeseen social aspects affecting both the data and the development of a health threat. Learning of the existence of disease through firsthand observations, for example, besides signaling health events can also influence people's perception of what they are observing. If the perceived risk of an outbreak is increased, more firsthand reporting could overinflate the health event. Studies of human behavior and Internet interaction may also help clarify social and behavioral effects (eg, age, gender, education level, income, and personality traits like extraversion, openness, and emotional stability) on content generated by social media and the Internet.

Health authorities who intend to use content from social media and other Internet data also need to consider protection and privacy, such as legal and ethical implications related to using Internet and social media data for public health surveillance. For example, it remains unclear what data may be freely accessed and used and whether or not privacy laws and related issues will prevent the structured analysis of new data. These issues are relevant to any surveillance tool that processes Internet or social media data, especially at governmental institutions.[50]

Conclusion

Even though the importance of social media and Internet-based data to epidemiological surveillance is clear, health agencies have been reluctant to incorporate these data sources into their systems because many technical issues have not yet been addressed. The technologies used in

event-based systems must be adapted to the individual perceptions of and interactions with their own epidemiological data and to social media and other data from the Internet. Future work in this field will have wide-reaching implications for investments in systems for early warnings of and responses to health threats across the globe and for optimal public health surveillance in the 21st century.

References

1. Heymann DL. The international response to the outbreak of SARS in 2003. *Philos Trans R Soc Lond B Biol Sci.* 2004;359(1447):1127-1129.
2. Chan EH, Brewer TF, Madoff LC, et al. Global capacity for emerging infectious disease detection. *Proc Natl Acad Sci USA*. 2010 December 14;107(50):21701-21706.
3. World Health Organization. *A Guide to Establishing Event-Based Surveillance*. World Health Organization; 2008.
4. Langmuir AD. The surveillance of communicable diseases of national importance. *N Engl J Med.* 1963;268:182-192.
5. Langmuir AD. The Epidemic Intelligence Service of the Centers for Disease Control. *Public Health Rep.* September-October 1980;95(5):470-477.
6. Coulombier D. Epidemic intelligence in the European Union: strengthening the ties. *Euro Surveil.* 2008;13(6).
7. Kaiser R, Coulombier D, Baldari M, Morgan D, Paquet C. What is epidemic intelligence, and how is it being improved in Europe? *Euro Surveil.* 2006;11(5):2892.
8. Paquet C, Coulombier D, Kaiser R, Ciotti M. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro Surveil.* 2006;11(12).
9. Hartley DM, Nelson NP, Walters R, et al. The landscape of international event-based biosurveillance. *Emerging Health Threats J.* 2009;3(e3).
10. Doan S, Hung-Ngo Q, Kawazoe A, Collier N. Global health monitor—a web-based system for detecting and mapping infectious diseases. Paper presented at: International Joint Conference on Natural Language Processing; January 7, 2008; Hyderabad, India.
11. Grishman R, Huttunen S, Yangarber R. Information extraction for enhanced access to disease outbreak reports. *J Biomed Inform.* 2002;35(4):236-246.

12. Castillo-Delgado C. Trends and directions of global public health surveillance. *Epidemiologic Reviews.* 2010.
13. Madoff LC. ProMED-mail: an early warning system for emerging diseases. *Clin Infect Dis.* 2004;39(2):227-232.
14. Madoff LC, Woodall JP. The internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. *Arch Med Res.* 2005;36(6):724-730.
15. Bravata DM, McDonald KM, Smith WM, et al. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Ann Intern Med.* 2004;140(11):910-922.
16. Bravata DM, McDonald KM, Szeto H, Smith WM, Rydzak C, Owens DK. A conceptual framework for evaluating information technologies and decision support systems for bioterrorism preparedness and response. *Med Decis Making.* 2004;24(2):192-206.
17. Bravata DM, Sundaram V, McDonald KM, et al. Evaluating detection and diagnostic decision support systems for bioterrorism response. *Emerg Infect Dis.* 2004;10(1):100-108.
18. Vrbova L, Stephen C, Kasman N, et al. Systematic review of surveillance systems for emerging zoonoses. *Transbound Emerg Dis.* 2010;57(3):154-161.
19. Corley CD, Lancaster MJ, Brigantic RT, et al. Assessing the continuum of event-based biosurveillance through an operational lens. *Biosecur Bioterror.* March 2012;10(1):131-141.
20. Hartley DM, Nelson NP, Arthur RR, et al. An overview of Internet biosurveillance. *Clin Microbiol Infect.* November 2013;19(11):1006-1013.
21. PubMed. US National Library of Medicine National Institutes of Health Website. http://www.ncbi.nlm.nih.gov/pubmed. Accessed October 23, 2013.
22. Scirus. Website. www.scirus.com. Accessed October 23, 2013.
23. Rew D. SCOPUS: Another step towards seamless integration of the world's medical literature. *Eur J Surg Oncol*. January 2010;36(1): 2-3.
24. Agheneza T. *A Systematic Review of Event-Based Public Health Surveillance Systems*. Hamburg: Faculty of Life Sciences, Hamburg University of Applied Sciences; 2011.
25. Linge JP, Steinberger R, Weber TP, et al. Internet surveillance systems for early alerting of health threats. *Euro Surveill.* April 2 2009;14(13).
26. 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.

27. World Health Organization. *Independent Evaluation of the GOARN Network*. World Health Organization; 2009.
28. Steinberger R, Fuart F, Best C, et al. Text mining from the web for medical intelligence. In: Fogelman-Soulié Françoise, Domenico Perrotta, Jakub Piskorski, Ralf Steinberger, eds. *Mining Massive Data Sets for Security*. Amsterdam, The Netherlands: IOS Press; 2008:295-310.
29. Teutsch SM, Churchill RE. *Principles and Practice of Public Health Surveillance.* Oxford: Oxford University Press; 1994.
30. Teutsch SM, Thacker SB. Planning a public health surveillance system. *Epidemiol Bull.* 1995;16(1):1-6.
31. Grein TW, Kamara KB, Rodier G, et al. Rumors of disease in the global village: outbreak verification. *Emerg Infect Dis.* 2000;6(2):97-102.
32. 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(1):e11.
33. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res.* 2008;10(3):e22.
34. Keller M, Blench M, Tolentino H, et al. Use of unstructured event-based reports for global infectious disease surveillance. *Emerg Infect Dis.* 2009;15(5):689-695.
35. Aung E, Whittaker M. Preparing routine health information systems for immediate health responses to disasters. *Health Policy Plan.* August 2013;28(5):495-507.
36. Murray CJ, Lopez AD. Measuring the global burden of disease. *N Engl J Med.* August 1 2013;369(5):448-457.
37. Chretien JP, Burkom HS, Sedyaningsih ER, et al. Syndromic surveillance: adapting innovations to developing settings. *PLoS medicine.* 2008;5(3):e72.
38. Chretien JP, Lewis SH. Electronic public health surveillance in developing settings: meeting summary. *BMC Proc.* 2008 2008;2 Suppl 3:S1.
39. Robertson C, Sawford K, Daniel SL, Nelson TA, Stephen C. Mobile phone-based infectious disease surveillance system, Sri Lanka. *Emerg Infect Dis.* 2010;16(10):1524-1531.
40. Nelson A, Patel M. How mapping, SMS platforms saved lives in Haiti earthquake. *OWNI: Objet Web Non Identifié.* http://owni.eu/2011/01/20/how-mapping-sms-platforms-saved-lives-in-haiti-earthquake/. Accessed October 23, 2013.

41. Brownstein JS, Freifeld CC. Evaluation of Internet-based informal surveillance for global infectious disease intelligence. Paper presented at: 13th International Congress on Infectious Diseases. June 19, 2008; Kuala Lumpur, Malaysia.
42. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Evaluation of online media reports for global infectious disease intelligence. *Advances in Disease Surveillance.* 2007;4:3.
43. 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(30).
44. World Health Organization. *International Health Regulations*. 2nd ed. World Health Organization; 2008.
45. Merianos A, Peiris M. International health regulations. *Lancet.* 2005;366(9493):1249-1251.
46. Low CL, Chan PP, Cutter JL, Foong BH, James L, Ooi PL. International health regulations: lessons from the influenza pandemic in Singapore. *Ann Acad Med Singapore.* 2010;39(4):325-323.
47. Eysenbach G. Infodemiology: The epidemiology of (mis)information. *Am J Med.* 2002;113(9):763-765.
48. Hulth A, Andersson Y, Hedlund KO, Andersson M. Eye-opening approach to norovirus surveillance. *Emerg Infect Dis.* 2010;16(8):1319-1321.
49. Keller M, Freifeld CC, Brownstein JS. Automated vocabulary discovery for geo-parsing online epidemic intelligence. *BMC Bioinformatics.* 2009;10:385.
50. Thompson LA, Black E, Duff WP, Paradise Black N, Saliba H, Dawson K. Protected health information on social networking sites: ethical and legal considerations. *J Med Internet Res.* 2011;13(1):e8.
51. Wilson JM. Argus: a global detection and tracking system for biological events. *Advances in Disease Surveillance.* 2007;4:21.
52. Collier N, Doan S, Kawazoe A, et al. BioCaster: detecting public health rumors with a web-based text mining system. *Bioinformatics.* 2008;24(24):2940-2941.
53. Tolentino H, Kamadjeu R, Fontelo P, et al. Scanning the emerging infectious diseases horizon—visualizing ProMED emails using EpiSPIDER. *Advances in Disease Surveillance.* 2007;2:169.
54. Guglielmetti P, Coulombier D, Thinus G, Van Loock F, Schreck S. The Early Warning and Reponse System for communicable diseases in the EU: an overview from 1999 to 2005. *Euro Surveill.* 2006;11:215-220.

55. Heymann DL, Rodier GR. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *Lancet Infect Dis.* 2001;1(5):345-353.
56. Khan SA, Patel CO, Kukafka R. GODSN: Global News Driven Disease Outbreak and Surveillance. *AMIA Annu Symp Proc.* 2006:983.
57. Mawudeku A, Blench M. *Global Public Health Intelligence Network (GPHIN)*. Paper presented at: 7th Conference of the Association for Machine Translation in the Americas; August 8, 2006; Cambridge, Massachusetts.
58. Brownstein JS, Freifeld CC. HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. *Euro surveill.* 2007;12(11):E071129.
59. Chen HYP, Zeng D. Health Map. *Infectious Disease Informatics.* 2010.
60. 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(2):150-157.
61. Nelson R. HealthMap: the future of infectious diseases surveillance? *Lancet Infect Dis.* 2008;8(10):596.
62. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Medicine*. 2008;5(7):e151.
63. Kass-Hout TA, di Tada N. International System for Total Early Disease Detection (InSTEDD) platform. *Advances in Disease Surveillance.* 2008;5:108.
64. Rortais A, Belyaeva J, Germo M, van der Goot E, Linge JP. MedISys: An early-warning system for the detection of (re-) emerging food- and feed-borne hazards. *Food Research International.* 2010;43(5):1553-1556.
65. Yangarber R, Steinberger R, Best C, von Etter P, Fuart F, Horby D. Combining information retrieval and information extraction for medical intelligence. Paper presented at: Multi-source Multilingual Information Extraction and Summarization (MMIES'2007) held at RANLP'2007, pp. 41-48. 26 September 2007; Borovets, Bulgaria.
66. Damianos L, Ponte J, Wohlever S, et al. MiTAP for bio-security: a case study. *AI Magazine.* 2002;23(4):13-29.
67. Hugh-Jones M. Global awareness of disease outbreaks: the experience of ProMED-mail. *Public Health Rep.* 2001;116 Suppl 2:27-31.
68. Woodall JP. Stalking the next epidemic: ProMED tracks emerging diseases. *Public Health Rep.* 1997;112(1):78-82.

69. Woodall JP. Global surveillance of emerging diseases: the ProMED-mail perspective. *Cad Saude Publica.* 2001;17 Suppl:147-154.

Appendix 1

Methods

Search Strategy. In order to generate a list of search keywords appropriate for retrieving articles relevant to "event-based surveillance systems," we conducted a series of preliminary searches by author, as well as by the terms "event-based" and "surveillance" using the search engine PubMed. We then repeated the process to ensure that the searches were thorough. A total of 130 articles were retrieved and subjected to a rapid review of titles and abstracts to build the final keyword list for all subsequent searches. The list was reviewed and agreed upon in collaboration with surveillance epidemiologists. All subsequent searches were done separately using the PubMed, Scopus, and Scirus databases. The search was a combination of all the key terms generated by using Boolean functions of "and" and "or."

Data Abstraction and Study Selection. For abstractions of the articles retrieved from the keyword searches, we developed inclusion criteria consisting of articles with a focus on infectious diseases, surveillance, outbreaks, and those specifically describing an event-based surveillance system; only those systems covering or intersecting with human health surveillance; and only those written in English. Our exclusion criteria were topics dealing with bioterrorism, technical aspects of security (eg, video surveillance), sentinel surveillance, or any surveillance not based on human health and infectious diseases. We also excluded articles without available abstracts. Due to the large number of articles found in each of the databases, the first stage of abstraction included applying our inclusion and exclusion criteria only to titles, which was extended to abstracts in all cases in which the classification relevance from titles alone was difficult. The full study protocol is available online.[24]

Data Synthesis and Extraction. The Boolean search using PubMed produced 9,161 articles (320 articles were retained); Scopus yielded 32,824 articles (298 articles were retained); and Scirus yielded 160,637,507 articles. By default, Scirus lists all retrieved articles by relevance; we reviewed only the first 6,000 articles and found 79 articles to be

relevant. Of 827 articles, 584 remained after eliminating 243 duplicates. We carefully reviewed abstracts of the 584 remaining articles and eliminated others based on the full content of each article. The categories for classification were (1) background (ie, articles not directly describing an event-based surveillance system but rather surveillance systems in general) or (2) system (ie, articles describing at least one event-based surveillance system). Those articles categorized as system were further distinguished between those covering only one-off monitoring activities (ie, one-time collection, analysis, and interpretation of health-related data for a defined period only) vs wider surveillance (ie, continuous monitoring, systematic collection, analysis and interpretation of health-related data), and, finally, those covering either indicator-based or event-based data (Figure A1).

Search Results. The combined search terms retrieved 39,000 articles, and after applying the inclusion and exclusion criteria, including a rigorously defined synthesis and extraction methodology, 123 articles were identified as providing "Background" information only; 6 articles were identified as describing "Monitoring Systems"; 166 articles were identified as describing an "Indicator-Based System"; and 44 articles were identified as describing an "Event-Based Surveillance" system. Of those 44 articles, 18 event-based surveillance systems were identified based on reading only the abstracts. After reviewing the full texts of all 44 articles, 5 of the 18 systems that had been classified as "Event-Based Surveillance" did not contain sufficient information for assessment and thus were eliminated. A final result of 32 articles enabled us to provide full descriptions based on our rigorous categorical data extraction criteria, which resulted in full descriptions for 13 event-based surveillance systems used in practice.

FIGURE A1. Results of data abstraction and study selection.