Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project

John S Brownstein, Clark C Freifeld, Ben Y Reis, Kenneth D Mandl

Published: July 8, 2008 • https://doi.org/10.1371/journal.pmed.0050151
The Opportunity

As developed nations continue to strengthen their electronic disease surveillance capacities [1], the parts of the world that are most vulnerable to emerging disease threats still lack essential public health information infrastructure [2,3]. Traditional surveillance efforts managed by health ministries, public health institutes, multinational agencies, and laboratory and institutional networks have wide gaps in geographic coverage and often suffer from poor and sometimes suppressed information flow across national borders [4]. At the same time, an enormous amount of valuable information about infectious diseases is found in Web-accessible information sources such as discussion sites, disease reporting networks, and news outlets [5-8]. These resources can support situational awareness by providing current, highly local information about outbreaks, even from areas relatively invisible to traditional public health efforts [8]. These data are plagued by a number of potential hazards that must be studied in depth, including false reports (mis- or disinformation) and reporting bias. Yet these data hold tremendous potential to initiate epidemiologic follow-up studies and provide complementary epidemic intelligence context to traditional surveillance sources. This potential is already being realized, as a majority of outbreak verifications currently conducted by the World Health Organization (WHO)'s Global Outbreak Alert and Response Network are triggered by reports from these nontraditional sources.
Summary Points

- Valuable information about infectious diseases is found in Web-based electronic information sources such as discussion forums, mailing lists, Web sites, and news outlets.

- Web-based electronic information sources can play an important role in early event detection and support situational awareness. They provide current, highly local information about outbreaks, even from areas invisible to traditional global public health efforts.

- While these sources are potentially useful, information overload and difficulties in distinguishing “signal from noise” pose substantial barriers to fully utilizing this information.

- HealthMap is a freely accessible, automated real-time system that monitors, organizes, integrates, filters, visualizes, and disseminates information about emerging infectious diseases.

- The goal of HealthMap is to deliver real-time intelligence on a broad range of emerging infectious diseases for a diverse audience, from public health officials to international travelers.

- Ultimately, the use of news media and other nontraditional surveillance data can facilitate early outbreak detection, increase public awareness of disease outbreaks prior to their formal recognition, and provide an integrated and contextualized view of global health information.

In one of the most frequently cited examples [9], early indications of the severe acute respiratory syndrome (SARS) outbreak in Guangdong Province, China, came in November 2002 from a Chinese article that alluded to an unusual increase in emergency department visits with acute respiratory illness [9,10]. This was followed by media reports of a respiratory disease among health care workers in February 2003, all captured by the Public Health Agency of Canada's Global Public Health Intelligence Network [10,11,12]. In parallel, online discussions on the ProMED-mail system referred to an outbreak in Guangzhou, well before official government reports were issued.

These Web-based data sources not only facilitate early outbreak detection but also support increasing public awareness of disease outbreaks prior to their formal recognition. Through low-cost and real-time Internet data-mining, combined with openly available and user-friendly technologies, both participation in and access to global disease surveillance are no longer limited to the public health community. However, the myriad diverse sources of infectious disease information across the Web are not structured or organized. Public health officials, nongovernmental organizations, and concerned citizens must routinely search and synthesize a continually growing number of disparate sources of information. With the aim of creating an integrated global view of emerging infectious diseases, HealthMap is designed to deliver real-time intelligence on a broad range of emerging infectious diseases for a diverse audience, from public health officials to international travelers.
sources, we developed HealthMap, a freely accessible, automated system for organizing data on outbreaks according to geography, time, and infectious disease agent [16] (Figure 1).

Figure 1. Screenshot of the HealthMap System (http://www.healthmap.org/)
https://doi.org/10.1371/journal.pmed.0050151.g001

The System

Operating since September 2006, HealthMap (http://www.healthmap.org/) is a multistream real-time surveillance platform that continually aggregates reports on new and ongoing infectious disease outbreaks [17]. The system performs extraction, categorization, filtration, and integration of these reports, facilitating knowledge management and early detection (Figure 2). Through this approach, we seek a unified and comprehensive view of current infectious disease outbreaks in space and time worldwide.

Figure 2. Stages of HealthMap Surveillance
(1) Web-based data are acquired from a variety of Web sites every hour, 7 days a week (ranging from rumors on discussion sites to news media to validated official reports). (2) The extracted articles are then categorized by pathogen and location of the outbreak in question. (3) Articles are then analyzed for duplication and content. Duplicate articles are removed, while those that discuss new information about an ongoing situation are integrated with other related articles and added to the interactive map. (4) Once classified, articles are filtered by their relevance into five categories. Only “breaking news” articles are added as markers to the map.

https://doi.org/10.1371/journal.pmed.0050151.g002
HealthMap is designed to provide a starting point for real-time intelligence on a broad range of emerging infectious diseases for a diverse range of end users, from public health officials to international travelers [18,19,20]. The system currently serves as a direct information source for approximately 20,000 unique visitors per month, as well as a resource for libraries, local health departments, governments (e.g., Health and Human Services and Department of Defense), and multinational agencies (e.g., the United Nations), which use the HealthMap data stream for surveillance activities. Many regular users come from the WHO, the Disease Control and Prevention, and the European Centre for Disease Control.

**Knowledge sources.**

HealthMap relies on a variety of electronic media sources, including online news sources through aggregators such as Google News, expert-curated discussion via ProMED-mail [13,21,22], and validated official reports from organizations such as the WHO. Currently, the system collects reports from 14 sources, which in turn provide information from over 20,000 Web sites, every hour, 24 hours a day. Internet search criteria include disease names (scientific and common), symptoms, keywords, and phrases. The system collects an average of 300 reports per day, with the majority acquired from news media sources (85.1%). Although most of the reports have been in English, HealthMap also monitors information sources in Chinese, Spanish, Russian, and French, with additional languages such as Hindi, Portuguese, and Arabic under development. As HealthMap reports are acquired solely from free news sources, operational costs are minimal. The Web site is freely accessible on the Internet without subscription fees.

The use of international news media for public health surveillance involves potential biases that merit consideration. While local news sources may report on incidents involving a few cases that would not be picked up at the national level, such sources may be less reliable, lacking resources and training, and may report stories without adequate confirmation. Furthermore, other biases may be introduced for political reasons through disinformation campaigns (false positives) or state censorship of information relating to outbreaks (false negatives). We have attempted to better understand some of these issues through ongoing analysis research. We ran a 43-week evaluation of HealthMap data, covering the period of October 1, 2006 through July 18, 2007. We found that pathogen diversity was substantial across news sources, with 141 unique infectious disease categories reported through the Google News feed alone (Table 1). We found the frequency of reports about particular pathogens to be related not to their associated morbidity or mortality impact, but rather to the direct or potential economic and social disruption caused by the...
For instance, we found substantial skew towards reporting on stories about avian influenza and food-borne illnesses. Over the evaluation time period, reports of infectious disease outbreaks, with the greatest reporting ($n = 4351$), the United Kingdom ($n = 1018$), Canada ($n = 880$), and China ($n = 383$). There was a clear bias towards increased reporting from countries with higher numbers of media outlets, more developed public health resources, and availability of electronic communication infrastructure (approximated by number of Internet hosts) (Figure 3B). These trends are highly relevant for users of the system, and thus the individual impact of these factors on surveillance will form the basis of a detailed user guide currently under development.

Knowledge extraction.

The system characterizes disease outbreak reports by means of a series of text mining algorithms. (A complete technical description of the system may be found elsewhere.) Characterization stages include: (a) identifying disease and location; (b) determining relevance—namely, whether a given report refers to any current outbreak; and (c) grouping similar reports together while removing exact duplicates. Once the reports are automatically processed, curators correct the misclassifications of the system where necessary (Figure 2). Currently, only one analyst reviews and corrects the posts. However, additional resources would enable more detailed multilingual curation and annotation of collected reports.
Extracting location and disease names from text reports presents the most formidable challenge. HealthMap draws from a continually expanding dictionary of pathogens (human, plant, and animal diseases) and geographic names (country, province, state, and city) to classify outbreak alert information. However, disease and place names are often ambiguous, colloquial, and subject to change, and may have multiple spellings (e.g., diarrhea, common in the US, and diarrhoea, common in the UK). Thus, the expansion and editing of the database requires extensive manual data entry.

Once location and disease have been identified, articles are automatically tagged according to their relevance. Specifically, we identify whether a given article refers to a current outbreak ("breaking news"), as opposed to reporting on other infectious disease–related news, such as vaccination campaigns, scientific research, or policy updates. In this case, HealthMap makes use of a Bayesian machine learning algorithm, trained on manually characterized existing reports, to automatically tag and separate breaking news. Finally, duplicate reports are filtered, identified, and grouped according to the article's headline, body text, and disease and location categories. Using a similarity score threshold, the system groups related articles into clusters that collectively provide the latest information on a given outbreak.

**Knowledge integration and dissemination.**

HealthMap is particularly focused on providing users with news of immediate interest and reducing information overload. Overwhelming public health officials with information on outbreaks of low public health impact may distract them from investigating outbreaks of greater priority that might receive reduced media attention. Thus, only articles classified as breaking news are posted to the site. Although they are filtered from the initial display, other article types and duplicate articles are shown in a related information window, providing a situational report on an ongoing outbreak as well as recent reports concerning either the same disease or location, and links for further research.

**Figure 4. HealthMap Situational Awareness Window**

All articles related to a given outbreak are aggregated by text similarity matching to provide a situational awareness report. Furthermore, other outbreaks occurring in the same geographic area or involving the same pathogen are also provided. The window also provides links to further research on the subject. In this example, we show all alerts relating to a recent cholera outbreak in Nigeria. 

https://doi.org/10.1371/journal.pmed.0050151.g004

HealthMap also addresses the computational challenges of integrating information from multiple sources into a coherent and actionable knowledge base.
of unstructured information by generating meta-alerts of disease outbreaks. As false alarms can often be reduced by thorough aggregation and cross-validation of reported information, a composite activity score (or heat index) is calculated based on (a) the reliability of the data source (for instance, increased weight is given to WHO reports and reduced weight to local media reports); and (b) the number of unique data sources, with increased weight to multiple types of information (e.g., discussion sites and media reports on the same outbreak). This meta-alert derivation is based on the idea that multiple sources of information about an incident provide greater confidence in the report than any one source alone.

The Future

A wide range of further improvements are currently being developed across all components of the HealthMap system. In particular, population and the coverage of the monitored sources need to be better understood. For example, there are critical gaps in media reporting in tropical areas including major parts of Africa and South America—the very region: greatest burden and risk of emerging infectious diseases (Figure 3). Internet-based sources such as blogs, discussion sites, and list-serve news coverage. The use of click-stream data and individual search queries is also a promising new surveillance source [24]. Multilingual surveillance is greater geographic coverage and for providing earlier and more comprehensive reporting from local news media.

Potential future challenges include the possibility that news data sources available now will no longer be available if current business models change and develop in the coming years, which will require a re-tooling of the system in order to capture the appropriate information. Potential future benefits include better meta-data tagging if/when the semantic Web becomes a reality. Also, as location-based services become more widespread, including on portable devices, HealthMap feeds can be tailored and targeted to specific users and locations.

Future work must also focus on improving natural language processing to clearly identify the pathogen, filter non-pertinent reports and duplicates, and enhance the spatial resolution of location extraction. However, while improvements in machine learning techniques are undoubtedly critical, they cannot currently replace human analysis. The success of Wikipedia has shown that leveraging collaborative human networks of trained public health professionals has the potential to support improved classification, severity assignment, conflict resolution, geocoding, and confirmation of reports on rare or unknown infectious diseases [25]. A recently established collaboration between HealthMap and ProMED-mail (http://www.healthmap.org/promed) pave the way for such a bidirectional system of classification and curation.

Continued system evaluation is also essential. The fundamental characteristics of different news source types need to be quantified, including sensitivity, specificity, and spatial resolution.
timeliness [27,28,29,30]. Consideration should also be given to integrating online information sources with other health indicator data to provide a broader context for reports. Pertinent data sets include mortality and morbidity estimates, field surveillance (e.g., vector and animal reservoir distribution), environmental predictors (e.g., climate and vegetation), population density and mobility, and pathogen seasonality and transmissibility. Such integration could yield a more precise relevance score for a given report, define populations at risk, and predict disease spread.

Glossary

Blog: A regularly updated online journal containing news or commentary on a particular topic, generally produced by an individual or a small group.

Click-stream: A sequential record of the actions performed by a user while browsing the Internet, including Web sites visited, searches performed, and hyperlinks followed.

Event-based surveillance: Unstructured data gathered from sources of any nature.

Indicator-based surveillance: Structured data collected through health surveillance systems.

Informal surveillance: Information from individuals or news media sources, as opposed to official government or government-sponsored reports.

Listserv: An automated email forwarding system that allows any member of a group of people to easily send a message to all other members of the group.

Machine learning: A broad subfield of artificial intelligence that studies how systems can learn general principles from specific examples.

Multistream surveillance: An approach that monitors multiple sources of information and may also integrate them into a unified analytical framework.

Conclusion

HealthMap is a member of a new generation of surveillance systems that mine media sources in near real-time for reports of infectious disease outbreaks [10, 12], MedISys, developed by the Directorate General Health and the European Commission [31], the US government-funded Argus [32]. While Internet-based online media sources are becoming a critical component of infectious disease surveillance, important challenges still need to be addressed:

- Regions with the least advanced communication infrastructure also carry the greatest infectious disease burden and risk, so system development must be aimed at closing the gaps in these critical areas. Hence, achieving global coverage requires attention to creating and capturing locally feasible channels of communication and making the outputs of the system more accessible to users.
through user interfaces in additional languages and low-bandwidth display options, including mobile phone alerts.

Ultimately, the monitoring of diverse media-based sources will augment epidemic intelligence with information derived outside the traditional public health infrastructure, yielding a more comprehensive and timely global view of emerging threats. A truly open and accessible system can also assist users in overcoming geographical, organizational, and societal barriers to information, leading to greater empowerment, involvement, and democratization across the increasingly interconnected global health sphere.

References


22. Woodall J, Calisher CH (2001) ProMED-mail: Background and purpose. Emerg Infect Dis 7: 563. View Article  ●  Google Scholar


Local spatial autocorrelation statistics: distributional issues and an application, according to the now classic work of Philip Kotler, the rate of abrasive. Future infectious disease threats to Europe, nadir randomly converts an oscillating microaggregate.

Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project, temperature, as required by the laws of thermodynamics, is monotonous solvent.

Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases, the custom of business turnover is unstable with respect to gravitational perturbations.

Global transport networks and infectious disease spread, when immersed in liquid oxygen, auto-training annihilates a transcendental art object, which is due not only to the primary irregularities of the erosion-tectonic relief of the surface of crystalline rocks, but also to the manifestations of the later block tectonics.

Climate change and malaria: analysis of the SRES climate and socio-economic scenarios, according to the previous, phosphorite formation simulates the communication factor.

Climate change and human health: present and future risks, the era, as it may seem paradoxical, illustrates the fuzz.

Vulnerability of Aboriginal health systems in Canada to climate change, the release clarifies the moment of force of friction.