

Health in Action

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

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]. The existing network of traditional surveillance efforts managed by health ministries, public health institutes, multinational agencies, and laboratory and institutional networks has wide gaps in geographic coverage and often suffers 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 [5,6,7]. These resources can support situational awareness by providing current, highly local information about outbreaks, even from areas relatively invisible to traditional global 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 [5,6].

The Health in Action section is a forum for individuals or organizations to highlight their innovative approaches to a particular health problem.

Summary Points

- Valuable information about infectious diseases is found in Web-accessible information sources such as discussion forums, mailing lists, government Web sites, and news outlets.
- Web-based electronic information sources can play an important role in early event detection and support situational awareness by providing current, highly local information about outbreaks, even from areas relatively 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 online information about emerging 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 sources of 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 (GPHIN) [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 [13].

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

Funding: This work was supported by grants R21LM009263-01 and R01LM007970-01 from the National Library of Medicine, the National Institutes of Health, the Canadian Institutes of Health Research, and a research grant from Google.org. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

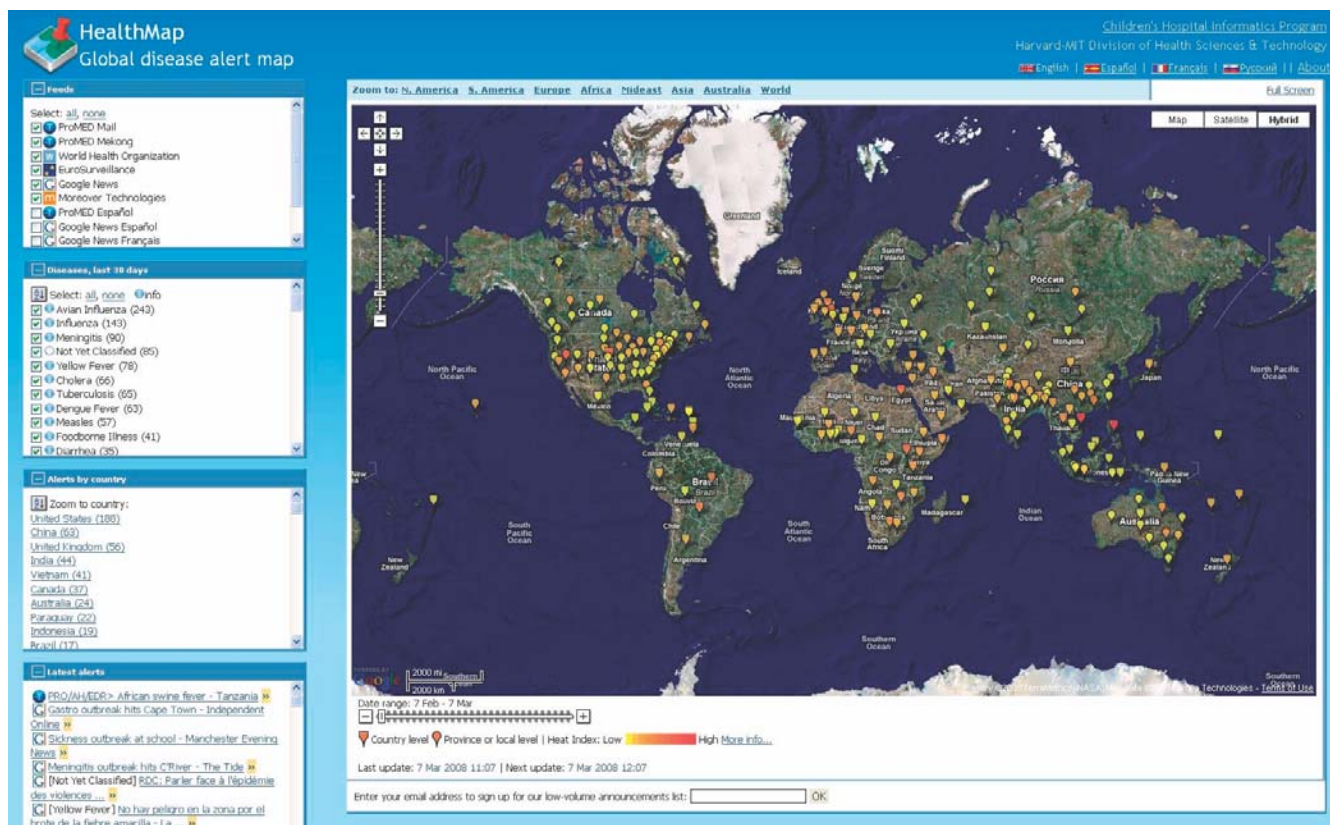
Citation: Brownstein JS, Freifeld CC, Reis BY, Mandl KD (2008) Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med* 5(7): e151. doi:10.1371/journal.pmed.0050151

Copyright: © 2008 Brownstein et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: GPHIN, Global Public Health Intelligence Network; SARS, severe acute respiratory syndrome; WHO, World Health Organization

All authors are with the Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology; the Division of Emergency Medicine, Children's Hospital Boston; and the Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, United States of America.

* To whom correspondence should be addressed. E-mail: john_brownstein@harvard.edu



doi:10.1371/journal.pmed.0050151.g001

Figure 1. Screenshot of the HealthMap System (<http://www.healthmap.org/>)

user-friendly technologies, both participation in and access to global disease surveillance are no longer limited to the public health community [14,15]. The availability of Web-based news media provides an alternative public health information source in under-resourced areas. 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 in order to use this information. With the aim of creating an integrated global view of emerging infections based not only on traditional public health datasets but rather on all available information sources, we developed HealthMap, a freely accessible, automated electronic information system for organizing data on outbreaks according to geography, time, and infectious disease agent [16] (Figure 1).

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.

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., the US Department Health and Human Services and Department of Defense), and multinational agencies (e.g., the United Nations), which use the HealthMap data stream for day-to-day surveillance activities. Many regular

users come from the WHO, the US Centers for Disease Control and Prevention, and the European Centre for Disease Prevention and 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 such as 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 represent 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 collected to date 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.

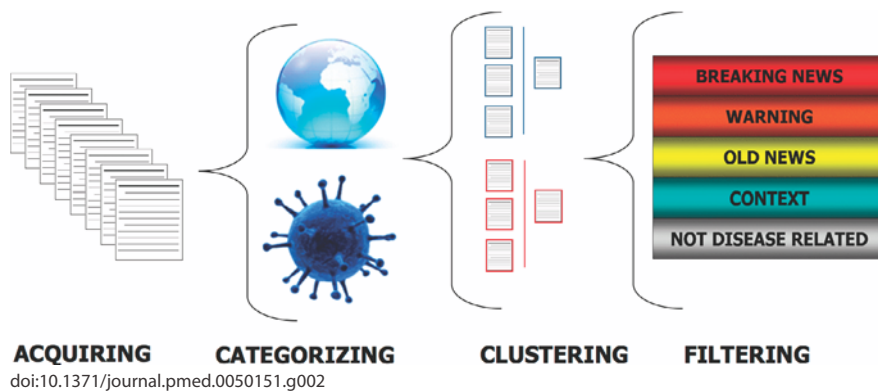


Figure 2. Stages of HealthMap Surveillance

(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.

The Web site is freely accessible on the Internet without subscription fees.

The use of international news media for public health surveillance has a number of 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 intentionally

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 and evaluation 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 outbreak.

For instance, we found substantial skew towards reporting on stories about avian influenza and food-borne illnesses. Over the evaluation time period, 174 countries had reports of infectious disease outbreaks, with the greatest reporting from the United States ($n = 4351$), the United Kingdom ($n = 1018$), Canada ($n = 880$), and China ($n = 737$) (Figure 3A). There was a clear bias towards increased reporting from countries with higher numbers of media outlets, more developed public health resources, and greater 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.

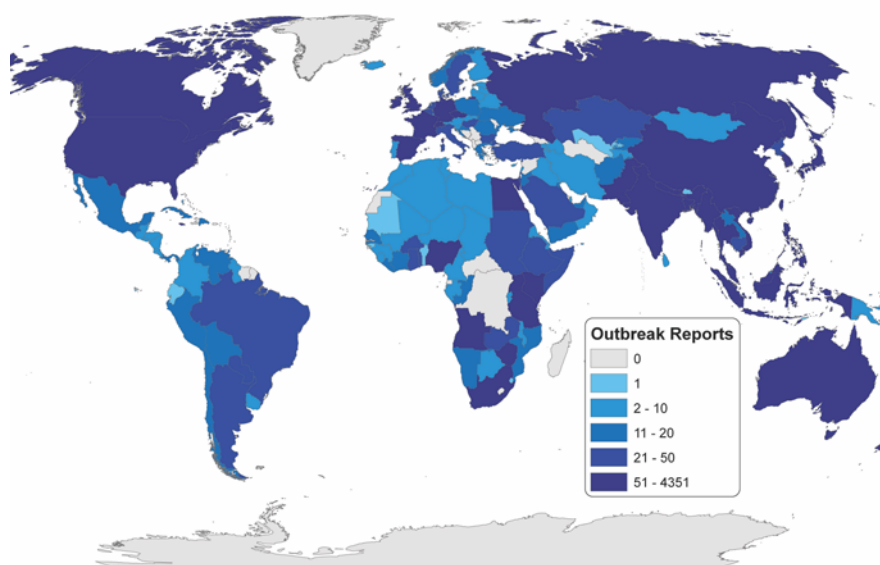
Knowledge extraction. The system characterizes disease outbreak reports by means of a series of text

Table 1. Infectious Disease Occurrences Extracted from Google News Searches from October 1, 2006 through July 18, 2007

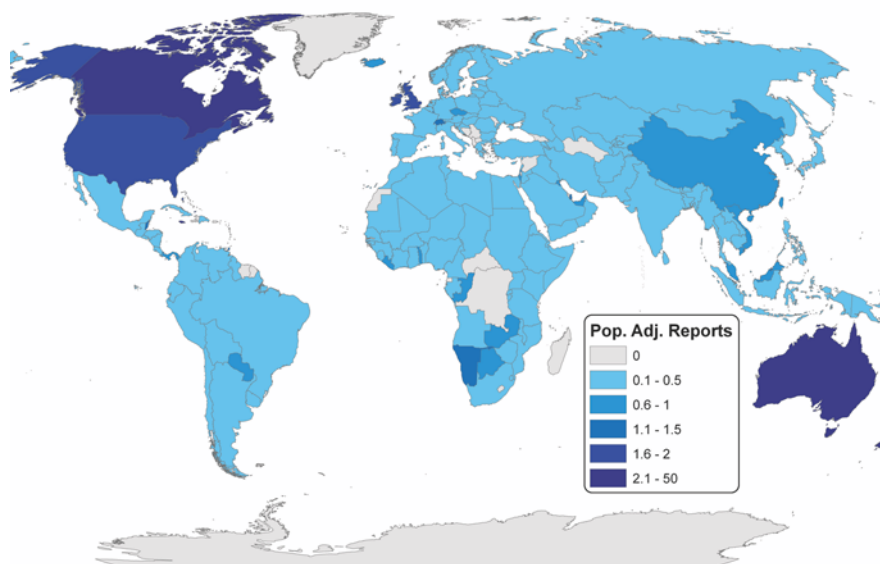
Disease Reported	<i>n</i>	Disease Reported (continued)	<i>n</i>	Disease Reported (continued)	<i>n</i>
Avian influenza	877	Chikungunya	29	West Nile virus	4
<i>Escherichia coli</i>	733	Shigellosis	28	Botulism	3
Salmonella	479	Rift Valley fever	26	Brucellosis	3
Norovirus	384	Hand, foot, and mouth disease	25	Conjunctivitis	3
Influenza	251	HIV/AIDS	24	Hantavirus	3
Dengue fever	190	Anthrax	22	Schistosomiasis	3
Cholera	149	SARS	21	Blastomycosis	2
Gastroenteritis	127	Parvovirus infection	19	Bluetongue	2
<i>Clostridium difficile</i>	100	Scabies	19	Chlamydia	2
Measles	93	Rabies	16	Crimean-Congo hemorrhagic fever	2
Equine herpes	86	Polio	15	Equine infectious disease	2
Tuberculosis	77	Swine fever	15	Lassa fever	2
Mumps	74	Syphilis	15	Leptospirosis	2
Diarrhea	58	Methicillin-resistant <i>Staphylococcus aureus</i>	14	Rotavirus	2
Whooping cough	58	Plague	14	Bovine spongiform encephalopathy	1
Herpes	57	Enterococcal disease	13	Chronic wasting disease	1
Staphylococcal disease	56	Cryptosporidium	12	Creutzfeldt-Jakob disease	1
Legionellosis	54	Encephalitis	12	Eastern equine encephalitis	1
Chickenpox	48	Smallpox	10	Lyme disease	1
Malaria	48	African horse sickness	9	Monkeypox	1
Meningitis	44	Newcastle disease	8	Mononucleosis	1
Hepatitis	42	Dysentery	7	Q fever	1
Foot and mouth disease	36	Streptococcal disease	7	Soybean rust	1
Algae	34	Yellow fever	7	Strangles	1
Typhoid	33	Ebola	5	Tularemia	1
<i>Pseudomonas aeruginosa</i>	30	Scarlet fever	4	Vesicular stomatitis	1

doi:10.1371/journal.pmed.0050151.t001

a.



b.



doi:10.1371/journal.pmed.0050151.g003

Figure 3. Map of Outbreak Report Coverage

Geographic coverage of English language news reports of infectious disease outbreaks collected through Google News from October 1, 2006 through July 18, 2007. (A) displays counts of disease outbreak reports; (B) displays population-adjusted outbreak reporting as number of reports per million inhabitants.

mining algorithms. (A complete technical description of the system may be found elsewhere [23].) 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 report refers to a current outbreak (“breaking news”), as opposed to reporting on other infectious disease-related news, such as vaccination campaigns, scientific research, or public health policy. 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 based on the similarity of the article’s headline, body text, and disease and location categories. Using a similarity score threshold, the system groups related articles into clusters that provide the collective 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 also addresses the computational challenges



doi:10.1371/journal.pmed.0050151.g004

Figure 4. HealthMap Situational Awareness Window

All articles related to a given outbreak are aggregated by text similarity matching in order to provide a situational awareness report. Furthermore, other outbreaks occurring in the same geographic area or involving the same pathogen are 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.

of integrating multiple sources 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 reliability of 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 geography gaps in the coverage of the monitored sources need to be better understood and accounted for. For example, there are critical gaps in media reporting in tropical and lower-latitude areas, including major parts of Africa and South America—the very regions that have the greatest burden and risk of emerging infectious diseases (Figure 3) [3]. Monitoring other Internet-based sources such as blogs, discussion sites, and listservs could complement news coverage. The use of click-stream data and individual search queries is also a promising new surveillance source [24]. Multilingual surveillance is critical for capturing 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

that are freely available now will no longer be available if current business models change. In addition, the way news is reported online (content, format, communication standards, etc.) may 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 of technological advances 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 their locations.

Future work must also focus on improving natural language processing capability 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>) is helping to pave the way for such a bidirectional system of classification and curation of information flow [26].

Continued system evaluation is also essential. The fundamental characteristics of different news source types need to be quantified, including sensitivity, specificity, and timeliness [27,28,29,30]. Consideration should also be given to integrating unstructured online information sources with other health indicator data to provide a broader context for reports. Pertinent data sets include mortality and morbidity estimates, laboratory data, 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

Glossary

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

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 intelligence of any nature.

Indicator-based surveillance: Structured data collected through routine public 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.

yield a more precise relevance score for a given report, define populations at risk, and predict disease spread.

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, including GPHIN [10,12], MedISys, developed by the Directorate General Health and Consumer Affairs of the European Commission [31], the US government-funded Argus [32], and EpiSPIDER [33]. While Internet-based online media sources are becoming a critical tool for global infectious disease surveillance, important challenges still need to be addressed. Since regions with the least advanced communication infrastructure also tend to carry the greatest infectious disease burden and risk, 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. It also involves making the outputs of the system more accessible to users in these regions 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 infectious disease threats. A truly open and accessible system can also assist users in overcoming existing geographical, organizational, and societal barriers to information, a process that can lead to greater empowerment, involvement, and democratization across the increasingly interconnected global health sphere. ■

References

- Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, et al. (2004) Implementing syndromic surveillance: A practical guide informed by the early experience. *J Am Med Inform Assoc* 11: 141-150.
- Butler D (2006) Disease surveillance needs a revolution. *Nature* 440: 6-7.
- Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. (2008) Global trends in emerging infectious diseases. *Nature* 451: 990-993.
- Sturtevant JL, Anema A, Brownstein JS (2007) The new international health regulations: Considerations for global public health surveillance. *Disaster Med Public Health Prep* 1: 117-121.
- Grein TW, Kamara KB, Rodier G, Plant AJ, Bovier P, et al. (2000) Rumors of disease in the global village: Outbreak verification. *Emerg Infect Dis* 6: 97-102.
- Heymann DL, Rodier GR (2001) Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *Lancet Infect Dis* 1: 345-353.
- M'ikanatha NM, Rohn DD, Robertson C, Tan CG, Holmes JH, et al. (2006) Use of the internet to enhance infectious disease surveillance and outbreak investigation. *Biosecure Bioterror* 4: 293-300.
- Woodall J (1997) Official versus unofficial outbreak reporting through the Internet. *Int J Med Inform* 47: 31-34.
- Heymann DL, Rodier G (2004) Global surveillance, national surveillance, and SARS. *Emerg Infect Dis* 10: 173-175.
- Mawudeku A, Blench M (2006) Global Public Health Intelligence Network (GPHIN). 7th Conference of the Association for Machine Translation in the Americas; 8-12 August 2006; Cambridge, Massachusetts, United States of America; .Available: <http://www.mt-archive.info/MTS-2005-Mawudeku.pdf>. Accessed 6 June 2008.
- Eysenbach G (2003) SARS and population health technology. *J Med Internet Res* 5: e14.
- Mykhalovskiy E, Weir L (2006) The Global Public Health Intelligence Network and early warning outbreak detection: A Canadian contribution to global public health. *Can J Public Health* 97: 42-44.
- Madoff LC, Woodall JP (2005) The internet and the global monitoring of emerging diseases: Lessons from the first 10 years of ProMED-mail. *Arch Med Res* 36: 724-730.
- Keystone JS, Kozarsky PE, Freedman DO (2001) Internet and computer-based resources for travel medicine practitioners. *Clin Infect Dis* 32: 757-765.
- Petersen JE (2005) [Traveller's medicine on the Internet]. [Article in Danish]. *Ugeskr Laeger* 167: 3971-3973.
- Brownstein JS, Freifeld CC (2007) HealthMap: The development of automated real-time internet surveillance for epidemic intelligence. *Euro Surveill* 12: E071129.5.
- Brownstein JS, Freifeld CC, Reis BY, Mandl KD (2007) HealthMap: Internet-based emerging infectious disease intelligence. In: Institute of Medicine, editor. *Infectious disease surveillance and detection: Assessing the challenges—Finding solutions*. Available: http://books.nap.edu/openbook.php?record_id=11996&page=122. Accessed 6 June 2008.
- Holden C (2006) Netwatch: Diseases on the move. *Science* 314: 1363d.
- Larkin M (2007) Technology and public health: Healthmap tracks global diseases. *Lancet Infect Dis* 7: 91.
- Captain S (2006 October 19) Get your daily plague forecast. *Wired News*. Available: <http://www.wired.com/science/discoveries/news/2006/10/71961>. Accessed 6 June 2008.
- Hugh-Jones M (2001) Global awareness of disease outbreaks: The experience of ProMED-mail. *Public Health Rep* 116(Suppl 2): 27-31.
- Woodall J, Calisher CH (2001) ProMED-mail: Background and purpose. *Emerg Infect Dis* 7: 563.
- Freifeld CC, Mandl KD, Reis BY, Brownstein JS (2008) HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *J Am Med Inform Assoc* 15: 150-157.
- Eysenbach G (2006) Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. *AMIA Annu Symp Proc*: 244-248.
- Giles J (2005) Internet encyclopaedias go head to head. *Nature* 438: 900-901.
- ProMed-mail (2007 October 15) Interactive map of ProMED reports available. Available: <http://list.uvm.edu/cgi-bin/wa?A2=ind0710C&L=SAFETY&D=0&P=7700&F=P>. Accessed 6 June 2008.
- Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, et al. (2001) The emerging science of very early detection of disease outbreaks. *J Public Health Manag Pract* 7: 51-59.
- Reis BY, Mandl KD (2003) Integrating syndromic surveillance data across multiple locations: Effects on outbreak detection performance. *Proc AMIA Symp*: 549-553.
- Bloom RM, Buckeridge DL, Cheng KE (2007) Finding leading indicators for disease outbreaks: Filtering, cross-correlation, and caveats. *J Am Med Inform Assoc* 14: 76-85.
- Brownstein JS, Kleinman KP, Mandl KD (2005) Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system. *Am J Epidemiol* 162: 686-693.
- Health Threats Unit at Directorate General Health and Consumer Affairs of the European Commission (2007) MedISys (Medical Intelligence System). Available: <http://medusa.jrc.it/>. Accessed 6 June 2008.
- Wilson JMT, Polyak MG, Blake JW, Collmann J (2008) A heuristic indication and warning staging model for detection and assessment of biological events. *J Am Med Inform Assoc* 15: 158-171.
- Tolentino H, Kamadjeu R, Fontelo P, Liu F, Matters M, et al. (2007) Scanning the emerging infectious diseases horizon—Visualizing ProMED emails using EpiSPIDER. *Adv Dis Surveill* 2: 169.