

ADVERTISEMENT



Call for Papers: Biomaterials

[Learn more >](#)

[plos.org](#)

[create account](#)

[sign in](#)



[Browse](#)

[Pul](#)

OPEN ACCESS

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

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

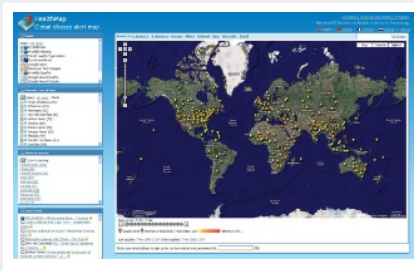
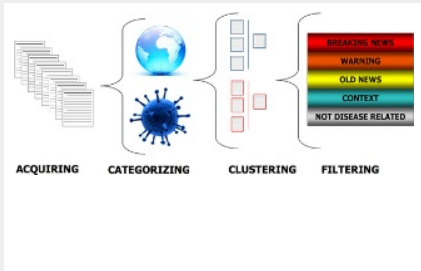
Article	Authors	Metrics	Comments
----------------	----------------	----------------	-----------------

Reader Comments (0)

Media Coverage

Figures

Figures

Disease Reported	n
Acute diarrhoea	937
Cholera	122
Schistosoma	479
Malaria	244
Chikungunya	252
Dengue fever	158
Cholera (WHO)	146
Cholera	127
Chikungunya	102
Protein	93
Escherichia coli	86
Trachoma	77
Measles	74
Hand, foot and mouth disease	58
Measles	57
Enterovirus	54
Chikungunya	48
Hand, foot and mouth disease	48
Measles	42
Hand, foot and mouth disease	36
Measles	34
Hand, foot and mouth disease	33
Hand, foot and mouth disease	28
Hand, foot and mouth disease	28

Citation: Brownstein JS, Freifeld CC, Reis BY, Mandl KD (2008) Surveillances of Emerging Infectious Diseases in the World: Frontières: Internet-Based Emerging Infectious Disease Intelligence. HealthMap Project. PLoS Med 5(7): e151. <https://doi.org/10.1371/journal.pmed.0050151>

Published: July 8, 2008

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.

Funding: This work was supported by grants R21 LM009263-01 and R01 LM009263-01 from the National Library of Medicine, the National Institutes of Health, the Canadian Institutes of Health Research, and a research grant from the Bill and Melinda Gates Foundation. 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.

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

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 infectious diseases still lack essential public health information infrastructure [2,3]. The traditional surveillance efforts managed by health ministries, public health agencies, multinational agencies, and laboratory and institutional networks have limited geographic coverage and often suffers from poor and sometimes slow information flow across national borders [4]. At the same time, an ever increasing amount of valuable information about infectious diseases is found in Web-accessible sources such as discussion sites, disease reporting networks, and social media. These resources can support situational awareness by providing complementary information about outbreaks, even from areas relatively invisible to traditional public health efforts [8]. These data are plagued by a number of potential biases that must be studied in depth, including false reports (mis- or disinformation), reporting bias. Yet these data hold tremendous potential to initiate epidemiological investigations and provide complementary epidemic intelligence context to traditional surveillance sources. This potential is already being realized, as a majority of outbreaks currently conducted by the World Health Organization (WHO)'s Global Outbreak Response Network are triggered by reports from these nontraditional

Summary Points

- Valuable information about infectious diseases is found in 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 by providing current, highly local information about outbreaks, even from areas previously invisible to traditional global public health efforts.
- While these sources are potentially useful, information overload and difficulties in distinguishing “signal from noise” pose substantial challenges to fully utilizing this information.
- HealthMap is a freely accessible, automated real-time system that organizes, integrates, filters, visualizes, and disseminates information about emerging diseases.
- The goal of HealthMap is to deliver real-time intelligence on emerging infectious diseases for a diverse audience, from local health officials to international travelers.
- Ultimately, the use of news media and other nontraditional information sources for disease surveillance data can facilitate early outbreak detection, increase awareness of disease outbreaks prior to their formal recognition, and provide an integrated and contextualized view of global health events.

In one of the most frequently cited examples [9], early indications of a severe acute respiratory syndrome (SARS) outbreak in Guangdong Province, China, were first reported in early 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 SARS outbreak among health care workers in February 2003, and the World Health Organization's Global Public Health Intelligence Network [10,11,12]. In parallel, online discussions on the ProMED-mail system identified the SARS outbreak in Guangzhou, well before official government reports were published.

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 widely available and user-friendly technologies, both participation in and the effectiveness of disease surveillance are no longer limited to the public health community. The availability of Web-based news media provides an alternative public information source in under-resourced areas. However, the myriad diverse sources of disease information across the Web are not structured or organized in a consistent manner. Health officials, nongovernmental organizations, and concerned citizens routinely collect and synthesize a continually growing number of disparate sources of disease information. With the aim of creating an integrated global view of emerging infectious diseases based not only on traditional public health datasets but rather on a

sources, we developed HealthMap, a freely accessible, automated system for organizing data on outbreaks according to geography, time, and disease agent [16] (Figure 1).

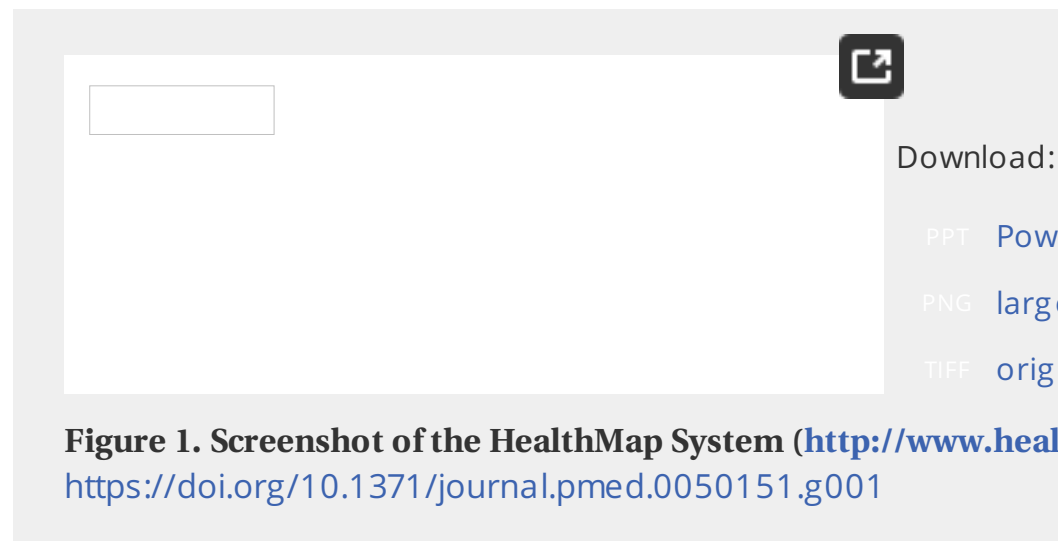


Figure 1. Screenshot of the HealthMap System (<http://www.healthmap.org>) (<https://doi.org/10.1371/journal.pmed.0050151.g001>)

The System

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

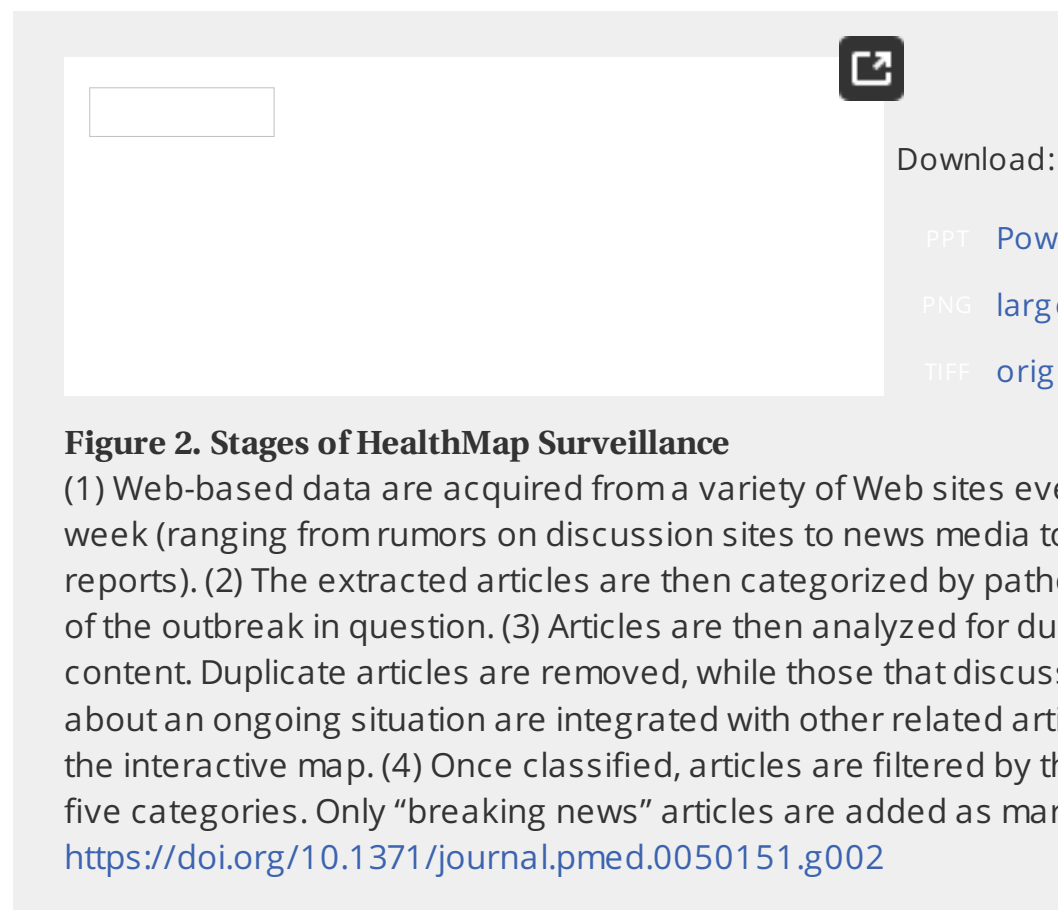


Figure 2. Stages of HealthMap Surveillance

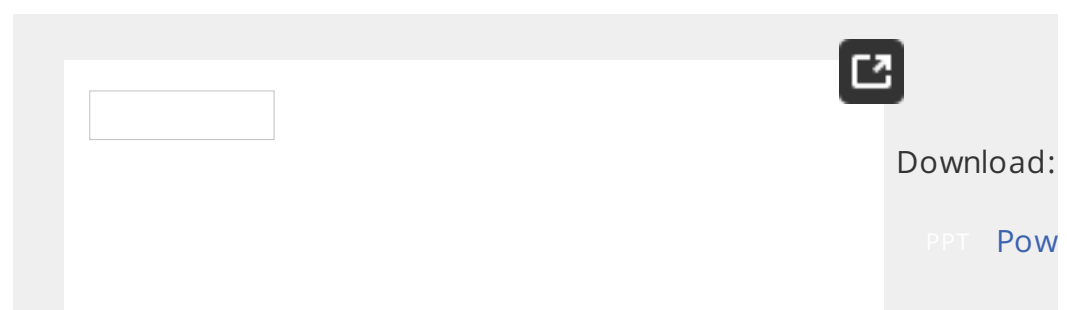
(1) Web-based data are acquired from a variety of Web sites every week (ranging from rumors on discussion sites to news media to official reports). (2) The extracted articles are then categorized by pathogen and location of the outbreak in question. (3) Articles are then analyzed for duplicate content. Duplicate articles are removed, while those that discuss a new or ongoing situation are integrated with other related articles on the interactive map. (4) Once classified, articles are filtered by time and severity into five categories. Only "breaking news" articles are added as markers on the map. (<https://doi.org/10.1371/journal.pmed.0050151.g002>)

HealthMap is designed to provide a starting point for real-time intel range of emerging infectious diseases for a diverse range of end health officials to international travelers [18,19,20]. The system cur direct information source for approximately 20,000 unique visitors resource for libraries, local health departments, governments (e.g., Health and Human Services and Department of Defense), and multi (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 through aggregators such as Google News, expert-curated discussion mail [13,21,22], and validated official reports from organizations such as the CDC. Currently, the system collects reports from 14 sources, which in turn collect information from over 20,000 Web sites, every hour, 24 hours a day. Search criteria include disease names (scientific and common), symptoms, and key phrases. The system collects an average of 300 reports per day, with 85.1% acquired from news media sources. Although most of the reports have been in English, HealthMap also monitors information sources in Russian, and French, with additional languages such as Hindi, Portuguese, and Spanish under development. As HealthMap reports are acquired solely from news media, 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 has several potential biases that merit consideration. While local news sources may report incidents involving a few cases that would not be picked up at the national level, news sources may be less reliable, lacking resources and training, and may report without adequate confirmation. Furthermore, other biases may be introduced for political reasons through disinformation campaigns (false positive reports), censorship of information relating to outbreaks (false negatives). We aim to better understand some of these issues through ongoing analysis and empirical research. We ran a 43-week evaluation of HealthMap data, covering the period from July 1, 2006 through July 18 2007. We found that pathogen diversity was reported by news sources, with 141 unique infectious disease categories reported by the Google News feed alone (Table 1). We found the frequency of reports of pathogens to be related not to their associated morbidity or mortality but to the direct or potential economic and social disruption caused by the



Download:
PPT Pow

Table 1.

Infectious Disease Occurrences Extracted from Google News Search
October 1, 2006 through July 18, 2007

<https://doi.org/10.1371/journal.pmed.0050151.t001>

For instance, we found substantial skew towards reporting on stories of influenza and food-borne illnesses. Over the evaluation time period, we collected 4351 reports of infectious disease outbreaks, with the greatest reporting occurring in the United States ($n = 4351$), the United Kingdom ($n = 1018$), Canada ($n = 880$), and China ($n = 378$) (Figure 3A). There was a clear bias towards increased reporting from countries with a high number of media outlets, more developed public health resources, and greater availability of electronic communication infrastructure (approximate number of Internet hosts) (Figure 3B). These trends are highly relevant for use in the design of a user guide thus the individual impact of these factors on surveillance will form part of a user guide currently under development.

**Knowledge extraction.**

The system characterizes disease outbreak reports by means of a set of algorithms. (A complete technical description of the system may be found in the supplementary materials.) Characterization stages include: (a) identifying disease and location relevance—namely, whether a given report refers to any current outbreak; (b) grouping similar reports together while removing exact duplicates. The reports are automatically processed, curators correct the misclassifications of reports that are necessary (Figure 2). Currently, only one analyst reviews and corrects reports; additional resources would enable more detailed multilingual curation of the collected reports.

Extracting location and disease names from text reports presents a challenge. HealthMap draws from a continually expanding dictionary (human, plant, and animal diseases) and geographic names (county, city) to classify outbreak alert information. However, disease and place names are often ambiguous, colloquial, and subject to change, and may have multiple meanings (e.g., diarrhea, common in the US, and diarrhoea, common in the UK). The editing of the database requires extensive manual data entry.

Once location and disease have been identified, articles are automatically ranked according to their relevance. Specifically, we identify whether a given article is reporting on a current outbreak (“breaking news”), as opposed to reporting on other related news, such as vaccination campaigns, scientific research, etc. In this case, HealthMap makes use of a Bayesian machine learning algorithm trained on manually characterized existing reports, to automatically tag and score news. Finally, duplicate reports are filtered, identified, and grouped based on the article's headline, body text, and disease and location categories. Above a certain score threshold, the system groups related articles into clusters that provide a collective information on a given outbreak.

Knowledge integration and dissemination.

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

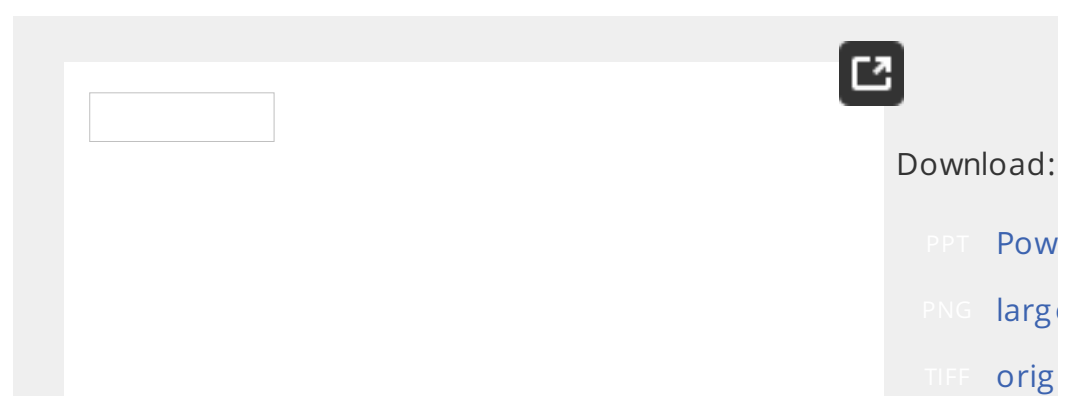


Figure 4. HealthMap Situational Awareness Window

All articles related to a given outbreak are aggregated by text similarity in order to provide a situational awareness report. Furthermore, articles occurring in the same geographic area or involving the same pathogen are provided. The window also provides links to further research on the topic. For example, we show all alerts relating to a recent cholera outbreak at <https://doi.org/10.1371/journal.pmed.0050151.g004>

HealthMap also addresses the computational challenges of integrat

of unstructured information by generating meta-alerts of disease alarms can often be reduced by thorough aggregation and cross-verification of 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 official reports and reduced weight to local media reports); and (b) the number of unique sources of information (e.g., discussion sites 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 as components of the HealthMap system. In particular, population and geographic 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 regions with the greatest burden and risk of emerging infectious diseases (Figure 3). Internet-based sources such as blogs, discussion sites, and listserve posts are also a promising new surveillance source [24]. Multilingual surveillance is needed for greater geographic coverage and for providing earlier and more complete 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 for the way news is reported online (content, format, communication structure) 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 and advances include better meta-data tagging if/when the semantic Web becomes widespread. Also, as location-based services become more widespread, including mobile devices, HealthMap feeds can be tailored and targeted to specific geographic locations.

Future work must also focus on improving natural language processing to clearly identify the pathogen, filter non-pertinent reports and duplicate reports, and improve 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 networks of trained public health professionals has the potential to improve classification, severity assignment, conflict resolution, geocoding, and verification of reports on rare or unknown infectious diseases [25]. A recently established partnership between HealthMap and ProMED-mail (<http://www.healthmap.org/promed>) will pave the way for such a bidirectional system of classification and data flow [26].

Continued system evaluation is also essential. The fundamental characteristics of different news source types need to be quantified, including sensitiv-

timeliness [27,28,29,30]. Consideration should also be given to integrating online information sources with other health indicator data to provide reports. Pertinent data sets include mortality and morbidity estimates, field surveillance (e.g., vector and animal reservoir distribution), environmental data (e.g., climate and vegetation), population density and mobility, and transmission and transmissibility. Such integration could yield a more precise regional 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, opposed to official government or government-sponsored reports.

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

Machine learning: A broad subfield of artificial intelligence that studies systems that 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 system.

Conclusion

HealthMap is a member of a new generation of surveillance systems that aggregate data from multiple sources in near real-time for reports of infectious disease outbreaks [10,12], MedISys, developed by the Directorate General Health and Consumer Protection of 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 in regions with the least advanced communication infrastructure also bear the greatest infectious disease burden and risk, system development requires closing the gaps in these critical areas. Hence, achieving global coverage requires attention to creating and capturing locally feasible channels of communication. This involves making the outputs of the system more accessible to users.

through user interfaces in additional languages and low-bandwidth including mobile phone alerts.

Ultimately, the monitoring of diverse media-based sources will augment intelligence with information derived outside the traditional public health 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, and to greater empowerment, involvement, and democratization across the interconnected global health sphere.

References

1. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani L (2004) Implementing syndromic surveillance: A practical guide informed by our experience. *J Am Med Inform Assoc* 11: 141–150.
[View Article](#) • [Google Scholar](#)
2. Butler D (2006) Disease surveillance needs a revolution. *Nature* 441: 106–107.
[View Article](#) • [Google Scholar](#)
3. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. (2008) Global trends in emerging infectious diseases. *Nature* 451: 990–993.
[View Article](#) • [Google Scholar](#)
4. Sturtevant JL, Anema A, Brownstein JS (2007) The new international regulations: Considerations for global public health surveillance. *Public Health Prep* 1: 117–121.
[View Article](#) • [Google Scholar](#)
5. Grein TW, Kamara KB, Rodier G, Plant AJ, Bovier P, et al. (2000) The global village: Outbreak verification. *Emerg Infect Dis* 6: 99–100.
[View Article](#) • [Google Scholar](#)
6. Heymann DL, Rodier GR (2001) Hot spots in a wired world: Wireless emerging and re-emerging infectious diseases. *Lancet Infect Dis* 1: 10–15.
[View Article](#) • [Google Scholar](#)
7. M'ikanatha NM, Rohn DD, Robertson C, Tan CG, Holmes JH, et al. (2001) The internet to enhance infectious disease surveillance and outbreak response. *Biosecur Bioterror* 4: 293–300.
[View Article](#) • [Google Scholar](#)
8. Woodall J (1997) Official versus unofficial outbreak reporting.

- 9.** Heymann DL, Rodier G (2004) Global surveillance, national s
Emerg Infect Dis 10: 173–175.
[View Article](#) • [Google Scholar](#)
- 10.** Mawudeku A, Blench M (2006) Global Public Health Intelliger
7th Conference of the Association for Machine Translation in
August 2006; Cambridge, Massachusetts, United States of A
<http://www.mt-archive.info/MTS-2005-Mawudeku.pdf>. Acces
- 11.** Eysenbach G (2003) SARS and population health technology
e14.
[View Article](#) • [Google Scholar](#)
- 12.** Mykhalovskiy E, Weir L (2006) The Global Public Health Intell
early warning outbreak detection: A Canadian contribution t
Can J Public Health 97: 42–44.
[View Article](#) • [Google Scholar](#)
- 13.** Madoff LC, Woodall JP (2005) The internet and the global mo
diseases: Lessons from the first 10 years of ProMED-mail. Ar
730.
[View Article](#) • [Google Scholar](#)
- 14.** Keystone JS, Kozarsky PE, Freedman DO (2001) Internet and
resources for travel medicine practitioners. Clin Infect Dis 32
[View Article](#) • [Google Scholar](#)
- 15.** Petersen JE (2005) [Traveller's medicine on the Internet]. [Art
Laeger 167: 3971–3973.
[View Article](#) • [Google Scholar](#)
- 16.** Brownstein JS, Freifeld CC (2007) HealthMap: The developme
time internet surveillance for epidemic intelligence. Euro Sur
[View Article](#) • [Google Scholar](#)
- 17.** Brownstein JS, Freifeld CC, Reis BY, Mandl KD (2007) HealthM
emerging infectious disease intelligence. Infectious disease
detection: Assessing the challenges—Finding solutions. Ava
http://books.nap.edu/openbook.php?record_id=11996&pa
June 2008.

- 18.** Holden C (2006) Netwatch: Diseases on the move. *Science* 311: 1207–1210.
[View Article](#) • [Google Scholar](#)
- 19.** Larkin M (2007) Technology and public health: Healthmap tracks disease outbreaks. *Lancet Infect Dis* 7: 91.
[View Article](#) • [Google Scholar](#)
- 20.** Captain S (2006 October 19) Get your daily plague forecast. <http://www.wired.com/science/discoveries/news/2006/10/plague-forecast/>. Accessed 6 June 2008.
- 21.** Hugh-Jones M (2001) Global awareness of disease outbreaks: A survey of ProMED-mail. *Public Health Rep* 116(Suppl 2): 27–31.
[View Article](#) • [Google Scholar](#)
- 22.** Woodall J, Calisher CH (2001) ProMED-mail: Background and use. *Lancet Infect Dis* 1: 563.
[View Article](#) • [Google Scholar](#)
- 23.** Freifeld CC, Mandl KD, Reis BY, Brownstein JS (2008) HealthMap: Disease monitoring through automated classification and visualization of media reports. *J Am Med Inform Assoc* 15: 150–157.
[View Article](#) • [Google Scholar](#)
- 24.** Eysenbach G (2006) Infodemiology: Tracking flu-related searches for syndromic surveillance. *AMIA Annu Symp Proc*. pp. 244–248.
- 25.** Giles J (2005) Internet encyclopaedias go head to head. *Nat Rev* 5: 277–279.
[View Article](#) • [Google Scholar](#)
- 26.** ProMed-mail (2007 October 15) Interactive map of ProMED reports. Available: <http://list.uvm.edu/cgi-bin/wa?A2=ind0710C&L=SAFETY&D=0&P=7700&F=P>. Accessed 6 June 2008.
- 27.** Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, et al. (2007) The science of very early detection of disease outbreaks. *J Public Health* 7: 51–59.
[View Article](#) • [Google Scholar](#)
- 28.** Reis BY, Mandl KD (2003) Integrating syndromic surveillance data with geographic locations: Effects on outbreak detection performance. *Proc Am Stat Soc* 107: 553.
- 29.** Bloom RM, Buckeridge DL, Cheng KE (2007) Finding leading indicators of disease outbreaks. *PLoS Med* 4: 193–200.
[View Article](#) • [Google Scholar](#)

outbreaks: Filtering, cross-correlation, and caveats. *J Am Me* 85.

[View Article](#) • [Google Scholar](#)

30. Brownstein JS, Kleinman KP, Mandl KD (2005) Identifying ped influenza vaccination using a real-time regional surveillance. *Epidemiol* 162: 686–693.

[View Article](#) • [Google Scholar](#)

31. Health Threats Unit at Directorate General Health and Consumer Protection, European Commission (2007) MedISys (Medical Intelligence System). <http://medusa.jrc.it/>. Accessed 6 June 2008.

32. Wilson JMt, Polyak MG, Blake JW, Collmann J (2008) A heuristic warning staging model for detection and assessment of bio threats. *Med Inform Assoc* 15: 158–171.

[View Article](#) • [Google Scholar](#)

33. Tolentino H, Kamadjeu R, Fontelo P, Liu F, Matters M, et al. (2008) Emerging infectious diseases horizon—Visualizing ProMED EpiSPIDER. *Adv Dis Surveill* 2: 169.

[View Article](#) • [Google Scholar](#)



[Privacy Policy](#) | [Terms of Use](#) | [Advertise](#) | [Media Inquiries](#)

Publications

PLOS Biology

PLOS Medicine

PLOS Computational Biology

PLOS Currents

PLOS Genetics

PLOS Pathogens

PLOS ONE

PLOS Neglected Tropical Diseases

plos.org

Blogs

Collections

Send us feedback

Contact

LOCKSS

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.