Using Internet Searches for Influenza Surveillance

Philip M. Polgreen,[1,2] Yiling Chen,[4] David M. Pennock,[5] and Forrest D. Nelson[3]

[1]Department of Internal Medicine, Carver College of Medicine, [2]Department of Epidemiology, College of Public Health, and [3]Department of Economics, Henry B. Tippie College of Business, University of Iowa, Iowa City; [4]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts; and [5]Yahoo! Research, New York, New York

The Internet is an important source of health information. Thus, the frequency of Internet searches may provide information regarding infectious disease activity. As an example, we examined the relationship between searches for influenza and actual influenza occurrence. Using search queries from the Yahoo! search engine (http://search.yahoo.com) from March 2004 through May 2008, we counted daily unique queries originating in the United States that contained influenza-related search terms. Counts were divided by the total number of searches, and the resulting daily fraction of searches was averaged over the week. We estimated linear models, using searches with 1–10-week lead times as explanatory variables to predict the percentage of cultures positive for influenza and deaths attributable to pneumonia and influenza in the United States. With use of the frequency of searches, our models predicted an increase in cultures positive for influenza 1–3 weeks in advance of when they occurred ($P < .001$), and similar models predicted an increase in mortality attributable to pneumonia and influenza up to 5 weeks in advance ($P < .001$). Search-term surveillance may provide an additional tool for disease surveillance.

The Internet has dramatically changed how people search for medical information. During the past decade, an increasing amount of information has become available on Web sites, especially about infectious diseases. For example, public health organizations at the local, state, national, and international level now routinely provide health-related information via their Web sites. These sites provide important updates about infectious disease activity and outbreaks. Also, most medical journals are available online, and to facilitate searching for journal articles, the National Library of Medicine Web site now contains >16 million citation records [1].

In addition to medical journals and public health Web sites, news Web sites supply a constant stream of updated health information. Also, several commercial firms organize medical information exclusively for clinicians, some catering specifically to infectious disease physicians and microbiologists [2]. Professional societies, such as the Infectious Diseases Society of America, the American Society of Microbiology, and the Society for Healthcare Epidemiology of America, also support Web sites with relevant scientific information, position statements, and practice guidelines. Some of these societies support electronic communities focused on infectious diseases, expanding the flow of medical information between clinicians and public health officials [3–5].

To capitalize on the dynamic nature of Web-based information, investigators have launched efforts to exploit this information for disease surveillance. For example, the Global Public Health Intelligence Network, developed by the Public Health Agency of Canada, continuously monitors media sources and Web-based information related to disease outbreaks around the world [6]. Global Public Health Intelligence Network data are not available to the general public. However, the relatively new HealthMap site (http://healthmap.org) monitors information from a variety of sources and displays results in real-time on a world map [7]. Access to this Web site is free, and it is available to the public.

An estimated 113 million people in the United States use the Internet to find health-related information [8]. Searchers include patients and their families and health care professionals [8–11]. However, the large number of health-related sites has made it difficult to find specific information that is credible and reliable. Thus, Internet search engines (e.g., Ask, Google, and Yahoo!) are now essential for Internet users to find information. In fact, most people searching for medical information use a search engine [8].

On a typical day, 8 million people search for health-related

Received 8 May 2008; accepted 11 August 2008; electronically published 27 October 2008.

Reprints or correspondence: Dr. Philip M. Polgreen, University of Iowa Dept. of Internal Medicine, 200 Hawkins Dr., Iowa City, IA 52242 (Polgreen@uiowa.edu).

Clinical Infectious Diseases 2008; 47:1443–8

1058-4838/2008/4711-0013$15.00
DOI: 10.1086/593098

information [8]. Thus, the pattern of how and when people search may provide clues or early indications about future concerns and expectations. For example, an analysis of Internet search terms related to jobs and job opportunities has produced accurate and useful statistics about the unemployment rate [12]. Similarly, searches for health-related information might also yield useful health statistics. Eysenbach [13], unable to get access to search engine query logs, demonstrated that the number of clicks on a "sponsored link" from Google Adsense, which was triggered by persons in Canada who entered "flu" or "flu symptoms" as search terms, accurately anticipated the Flu Watch reports collected by the Public Health Agency of Canada.

Thus, analyzing actual search query logs for terms related to infectious diseases may provide a unique supplement to traditional infectious disease surveillance systems. The Centers for Disease Control and Prevention influenza surveillance program identifies disease as or after it occurs and, therefore, does not provide advance warnings. Furthermore, the Centers for Disease Control and Prevention's data regarding influenza activity are no longer current when released to health care professionals. To supplement influenza surveillance, several forms of syndromic surveillance have been suggested, ranging from analysis of over-the-counter medication sales to school absentee records [14]. As another supplemental form of surveillance, we describe how Internet search query logs may help detect changes in disease activity. Using influenza as an example, we examine the temporal relationship between the search terms related to a disease and the actual cases of disease to determine if, and to what extent, an increase in search frequency matches or precedes actual disease activity.

INFLUENZA DATA

To measure influenza disease occurrence, we used 2 types of US influenza surveillance data. The first type of data were based on weekly influenza cultures [15]. Each week during the influenza season, clinical laboratories throughout the United States that are members of the World Health Organization Collaborating Laboratories or the National Respiratory and Enteric Virus Surveillance System report the total number of respiratory specimens tested and the number that were positive for influenza.

The second type of data summarize weekly mortality attributable to pneumonia and influenza [15]. These data are collected from the 122 Cities Mortality Reporting System. Each week, the participating cities report the total number of death certificates received and the number that list pneumonia or influenza as the underlying and/or contributing cause of death. On the basis of these data, we obtain national influenza mortality figures. To match the date range of our Internet-search data, both types of influenza-surveillance data that we used were collected from March 2004 through May 2008.

SEARCH DATA

Search query logs were obtained from Yahoo! for the period from March 2004 through May 2008. From the Internet protocol address associated with a search, we attempted to identify the geographic location (i.e., US Census region) from which the search was initiated. The number of unique queries that came from the United States and contained influenza-related terms was counted daily. We excluded searches from outside the United States, because the influenza season varies geographically. These daily influenza-search counts were divided by the total number of all searches originating in the United States for each day, to obtain the daily fraction of influenza-related searches. This normalization removed the possible effect of the overall increase in the number of searches. Because the influenza surveillance data were reported weekly, we used a weekly influenza-related search fraction by calculating the average of the daily fraction for each week.

We obtained 2 series of influenza-related search fraction data at the national level: (1) the fraction of US search queries that contain the terms "influenza" or "flu" but do not contain the terms "bird," "avian," or "pandemic" and (2) the fraction of US search queries that contain the terms "influenza" or "flu" but do not contain the terms "bird," "avian," "pandemic," "vaccine," "vaccination," or "shot."

By restricting these series to queries that did not contain the terms "bird," "avian," and "pandemic," we attempted to remove searches for avian influenza rather than seasonal influenza. Also, because most influenza vaccination occurs before the influenza season, we excluded all obvious vaccination-related searches.

We also classified weekly influenza-related search data into 9 US Census regions. Census-region data were normalized by total searches within that region. Because we identified the geographic location of origin from the Internet protocol address, there were cases for which we were not able to identify the exact region in which a search originated, but we were able to identify that the search came from within the United States. Therefore, the sum of the search data for the 9 US Census regions does not equal the amount of data at the national level. For each US Census region, we obtained only 1 series of data: weekly search data from the region for queries that contain the terms "influenza" or "flu" but do not contain the terms "bird," "avian," "pandemic," "vaccine," "vaccination," and "shot."

SEARCH AND CULTURE RESULTS POSITIVE FOR INFLUENZA

To define the relationship between culture-positive cases of influenza and influenza-related searches, we examined the relationship between influenza culture data and influenza-related searches at the national level. These data are presented as a time series in figure 1.

The fraction of influenza-related search queries and the rates of cultures positive for influenza have similar patterns over time, but a sharp increase in the fraction of influenza-related search queries precedes the sharp increase in the rate of cultures positive for influenza. With use of the culture data, we fitted the following linear model to test the predictability of search frequency on positive influenza culture results, including a time-trend variable:

$$c_t = \beta_0 + \beta_1 s_{t-x} + \beta_2 t + \varepsilon_t,$$

where t is a time trend (measured in weeks), c_t is the rate of cultures positive for influenza received during week t, and s_{t-x} is the search frequency in week $t-x$. To determine the appropriate lag (in weeks), we examined 11 possible values for x and compared the R^2 value for each model. The model with a search term with a 1-week lag provided the best fit. However, models with lags up to 3 weeks in advance of culture data fit similarly with regard to the R^2 value. A summary of the regression results for the search term models with lags of 0–10 weeks are presented in table 1.

The coefficient (β_2) on the time trend variable is not significantly different from zero in any of the models. However, a positive relationship was found between the fraction of influenza-related queries and rates of cultures positive for influenza 2 weeks later ($P<.001$). The large coefficient (β_1) on s_{t-2} reflects that influenza-related search frequency is measured as a fraction of all searches. The predicted values generated by the 2-week–lag model and the actual culture result data are presented in figure 2.

We also fit separate models with lags of 1–10 weeks for each of the 9 US Census regions. Results were similar to the national model, with the best fitting models predicting an increase in the number of cultures positive for influenza 1–3 weeks in advance. The average R^2 value at 2 weeks was 0.3788. However, values varied between a high of 0.5729 in the East-South-Central region and a low of 0.1656 in the Mid-Atlantic region.

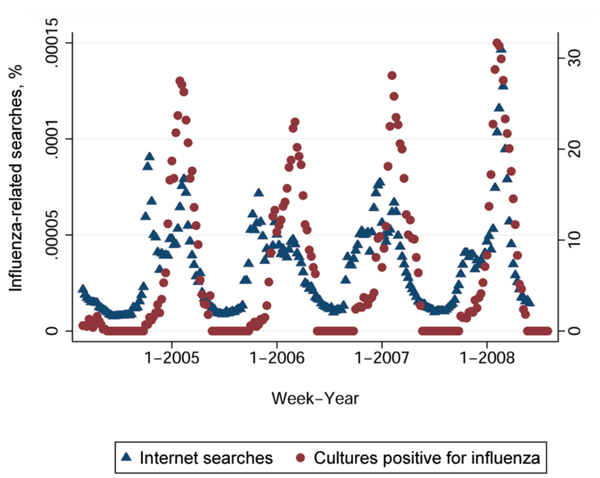

Figure 1. Fraction of Internet searches for influenza-related terms and the rate of detection of influenza in cultures, by week.

Table 1. Regression-analysis results for cultures positive for influenza.

Lag *(x)*, weeks	Coefficient (s_{t-x})	SE	t	P	R^2
0	239636.2	18301.99	13.09	<.001	0.4672
1	242579.5	18218.11	13.32	<.001	0.4723
2	239568.6	18487.33	12.96	<.001	0.4568
3	234749.1	18848.97	12.45	<.001	0.4356
4	229446.4	19225.16	11.93	<.001	0.4134
5	223257.3	19628.85	11.37	<.001	0.3890
6	215900.2	20064.8	10.76	<.001	0.3618
7	206683.5	20565.4	10.05	<.001	0.3300
8	195520.6	21118.44	9.26	<.001	0.2943
9	184502.1	21619.25	8.53	<.001	0.2610
10	173491.3	22164.1	7.83	<.001	0.2305

NOTE. In the regression analysis, t represented a time trend (measured in weeks), s_{t-x} represented the search frequency in week $t-x$, and R^2 indicated how well the model fit the culture data. SE, standard error.

SEARCH AND INFLUENZA MORTALITY RESULTS

Figure 3 presents the fraction of influenza-related searches and mortality attributable to influenza over time in the United States. To account for the relationship between searches and mortality, as described for the culture data, we fitted the following linear model to test the predictability of search frequency with regard to influenza-related mortality:

$$m_t = \beta_0 + \beta_1 s_{t-x} + \beta_2 t + \varepsilon_t,$$

where m_t is the total number of deaths attributable to pneumonia and influenza in week t, and all other variables are as defined for the culture model. A model incorporating searches at time $t-5$ (i.e., with 5 weeks of lag time) fits slightly better than do other models with a search variable ranging from time $t-0$ to time $t-10$. All of the regression-model results using searches with lags of 0–10 weeks are listed in table 2. A positive relationship exists between the fraction of influenza-related search queries and increases in mortality attributable to pneumonia and influenza 5 weeks later ($P<.001$). The large coefficient (β_1) on s_{t-5} reflects that influenza-related search frequency is measured as a fraction of all searches and, therefore, has small values, on the order of 1×10^{-6}. Figure 4 presents the predicted values from the model with a 5-week lag and the actual mortality data.

Finally, we fit models with lags of 0–10 weeks for each of

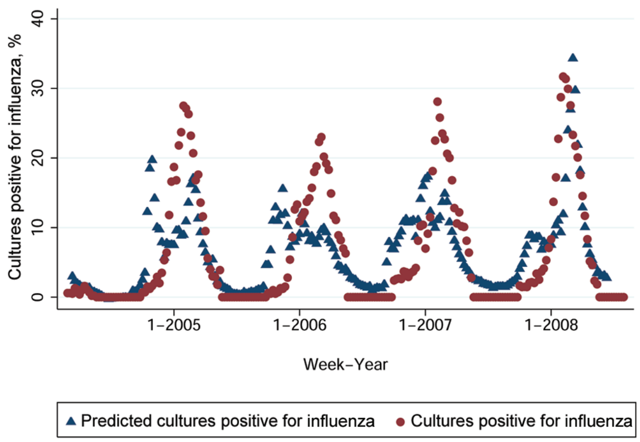

Figure 2. Predicted percentage of cultures positive for inflenza based on Internet searches with a 2-week lag and the actual percentage of cultures positive for influenza, by week.

the 9 US Census regions. Results were similar to the national model: for the best fitting models, the fraction of searches peaked 4–6 weeks before an increase in mortality attributable to influenza and pneumonia. The average R^2 at 5 weeks was 0.3041. However, values varied between a high of 0.4250 in the East-North-Central region to a low of 0.1227 in the Pacific region.

DISCUSSION

Influenza reoccurs each season in regular cycles, but the geographic location, timing, and size of each outbreak vary, complicating efforts to produce reliable and timely estimates of influenza activity. However, we found that a distinct temporal association exists between influenza-related search-term frequency and influenza disease activity. On a national level, influenza-related search-term activity seems to precede an increase in the number of cultures positive for influenza and deaths attributable to pneumonia and influenza. Furthermore, the temporal relationship between searches and cultures positive for influenza and searches and mortality corresponds to the epidemiology of influenza, because the number of deaths from pneumonia typically peaks a few weeks after a peak in the number of influenza cases.

Investigators have suggested several supplemental approaches for influenza surveillance, at prediagnosis and diagnosis stages. Prediagnosis approaches mainly include the analysis of information collected before specific influenza-related diagnoses are made, including analyses of telephone triage calls [16], purchases of over-the-counter medications for respiratory diseases [17–20], and school absenteeism [21]. In contrast, diagnosis-level approaches attempt to gather clinical data from emergency department visits [22–24] or microbiologic sources in as close to real-time as possible. The timeliness of influenza surveillance approaches has recently been thoroughly reviewed elsewhere [14]. Prediction markets have also been used to provide future estimates of influenza activity by aggregating both prediagnostic and postdiagnostic information [25]. In general, the efforts described herein provide information days to weeks in advance of traditional sources, but it is difficult to compare these approaches, because different geographic regions were studied, different statistical approaches were used, and some reports only include 1 influenza season [14]. To generalize these approaches to the national level would require merging several data sources from different geographic areas and multiple firms (in the case of pharmacy data or billing data). In contrast, search query data are efficiently collected in a standard usable form and aggregate both prediagnostic and postdiagnostic information. Although it is difficult to compare with other methods, analysis of Internet search terms seems to perform reasonably well. In addition, data are easy to collect, and unlike other nontraditional forms of surveillance, search data can easily be used to study other diseases.

If future results are consistent with these findings, search-term surveillance may provide an important and cost-effective supplement to traditional disease-surveillance systems. In the case of influenza, a few weeks of lead time could help inform epidemiological investigations and assist with both prevention and treatment efforts. Search terms classified by different geographic regions may provide even more useful information. For example, we fit linear models with data from the 9 US Census regions and found that influenza-related search terms are statistically significantly related to influenza mortality. Models performed better for some regions than others, suggesting that events in some regions may increase searches in other regions. Additional work is needed to examine the spatial relationship between Internet searches and the geographic spread

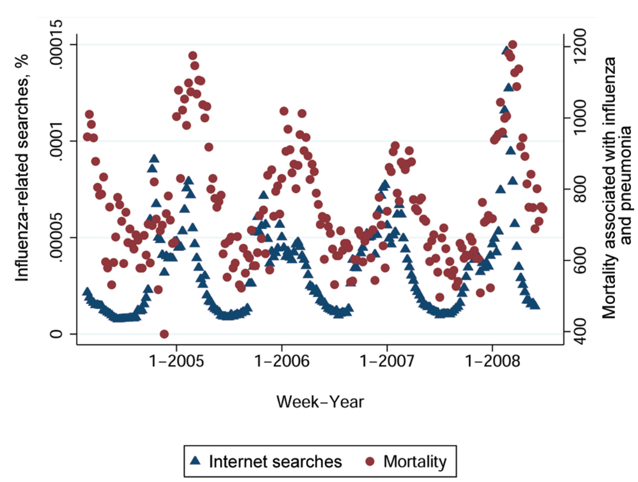

Figure 3. Fraction of Internet searches for influenza-related terms and mortality attributable to influenza and pneumonia, by week.

Table 2. Influenza mortality regression results.

Lag *(x)*, weeks	Coefficient (s_{t-x})	SE	*t*	*P*	R^2
0	3300788	436385.8	7.56	<.001	0.2075
1	3810620	415148.2	9.18	<.001	0.2787
2	4194847	394455.2	10.63	<.001	0.3418
3	4445665	378633.3	11.74	<.001	0.3882
4	4604043	367573.4	12.53	<.001	0.4198
5	4625652	368166.3	12.56	<.001	0.4229
6	4461079	379889.1	11.74	<.001	0.3919
7	4314867	390405	11.05	<.001	0.3649
8	4248610	396362.5	10.72	<.001	0.3523
9	3992864	410770.2	9.72	<.001	0.3111
10	3767351	422055.3	8.93	<.001	0.2765

NOTE. In the regression analysis, *t* represented a time trend (measured in weeks), s_{t-x} represented the search frequency in week $t-x$, and R^2 indicated how well the model fit the mortality data. SE, standard error.

of influenza. However, because culture and mortality data are not uniformly reported at the state level, our geographic analysis stopped at the US Census–region level.

Despite the promise that use of Internet search data has shown for surveillance purposes, there are several limitations. First, with only 4 years of data, the inferential conclusions that can be made from time-series analyses are limited. A second limitation is that we need to account for the possibility that some searches may be generated by news reports or a "celebrity effect," instead of actual disease activity. For example, the publication of a medical journal article about influenza may generate searches with no relationship to disease occurrence, and the same may be true if a celebrity contracts a specific disease. Cooper et al. [26], who used Yahoo! search queries, found that daily variations of search frequency regarding cancer were heavily influenced by news reports . However, Internet searches for specific cancers were still correlated with their estimated incidence and mortality. Also, a news item that causes a large increase in search volumes should be easy to identify, and its effect should be rather short lived, given the half-life of most news cycles.

The limited geographic data that was gleaned from search terms is a third limitation of Internet search data. Geographic search data are extracted from Internet protocol addresses and may not always represent actual geographic location. Privacy issues represent another significant limitation. The search data described in this article were aggregated from users across 9 US Census regions. However, the use of search data with much finer geographic information that is linked to individuals across multiple different search topics could represent a privacy concern. Thus, we envision that health investigators would only use aggregated search volumes representing larger geographic regions for surveillance purposes. Finally, access to search query logs from search engines will need to be made available to investigators. Other attempts to study actual search query data for public health reasons have been unsuccessful [13].

In addition to data from search engines, data gleaned from Web site hits or searches on specific Web sites may also provide useful information about disease activity. For example, the number of articles retrieved on the Web site Healthlink (http://healthlink.mcw.edu), a consumer health information Web site maintained at the Medical College of Wisconsin, was found to correlate with influenza activity [27]. Therefore, searches for specific diseases on high-traffic Web sites (e.g., a US state health department) may provide important time-series data, because such data capture the number and, to some extent, the geographic location (via Internet protocol address) of people who are investigating the activity of a specific disease. Searches for specific medical conditions on the National Library of Medicine's PubMed Web site may indicate changing patterns in infectious disease activity or potential adverse drug events. In addition, changes in the volume of searches on commercial Web sites (e.g., Up-To-Date [http://www.uptodate.com] and MD Consult [http://www.mdconsult.com]) may indicate gaps in clinical knowledge or the need for clinical trials. Data from such sites may be more representative of what health care professionals are searching for, as opposed to the general public.

We propose that search-term surveillance may represent a novel and inexpensive way of performing supplemental disease surveillance. The use of search-term series is not limited to influenza; it could also be used to monitor emerging and reemerging infectious diseases and to detect changes in phenomena related to chronic illnesses. Surveillance of symptom-based searches (e.g., diarrhea) may help detect illness outbreaks if search levels increase over an established baseline. Deidentified search volumes for sexually transmitted infections (e.g., syphilis) may provide public health officials an indication of disease trends in advance of official reports of disease activity. Although

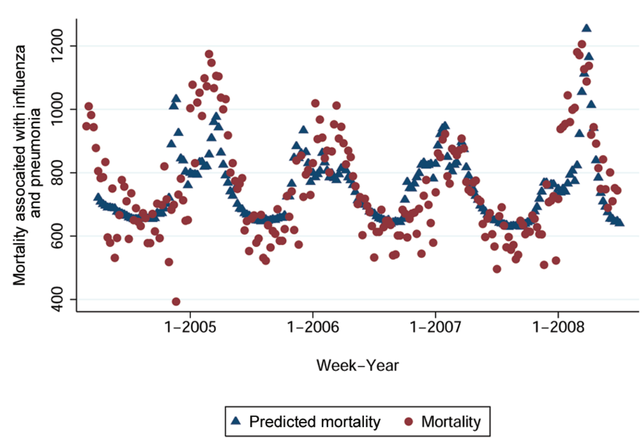

Figure 4. Predicted mortality attributable to influenza and pneumonia based on Internet searches with a 5-week lag, and the actual mortality attributable to influenza and pneumonia, by week.

search-term analysis probably provides some aggregation of news reports, it also adds a behavioral component by signaling how important topics are to individuals searching the Internet. Thus, analysis of search data may also reveal how people respond to medical news and may provide indications about their concerns and future expectations. Despite several limitations, the ability to detect trends and confirm observations from traditional surveillance approaches make this new form of surveillance a promising area of research at the interface between computer science, epidemiology, and medicine.

Acknowledgments

Financial support. Robert Wood Johnson Foundation, Pioneer Portfolio.

Potential conflicts of interest. P.M.P. has been a member of the Emerging Trends in Seasonal Influenza Advisory Panel of Roche Laboratories. D.M.P. and Y.C. are employees of Yahoo! Research. F.D.N.: no conflicts.

References

1. National Library of Medicine (NLM)/National Institutes of Health. NLM Technical Bulletin: MLA 2006, NLM online users' meeting remarks. Available at: http://www.nlm.nih.gov/pubs/techbull/ja06/ja06_mla_dg.html. Accessed 25 April 2008.
2. Burdette SD. Electronic tools for infectious diseases and microbiology. Can J Infect Dis Med Microbiol **2007**; 18:347–52.
3. Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis **2004**; 39:227–32.
4. Strausbaugh LJ, Liedtke LA. The Emerging Infections Network electronic mail conference and Web page. Clin Infect Dis **2001**; 32:270–6.
5. Dwyer V. ClinMicroNet: sharing experiences and building knowledge virtually. Clinical Microbiology Newsletter **2003**; 25:121–5.
6. 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**; 97:42–4.
7. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc **2008**; 15:150–7.
8. Pew Internet and American Life Project. Online health search 2006. Available at: http://www.pewinternet.org/PPF/r/190/report_display.asp. Accessed 25 April 2008.
9. Ybarra ML, Suman M. Help seeking behavior and the Internet: a national survey. Int J Med Inform **2006**; 75:29–41.
10. Bundorf MK, Wagner TH, Singer SJ, Baker LC. Who searches the Internet for health information? Health Serv Res **2006**; 41:819–36.
11. Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients' use of the Internet for medical information. J Gen Intern Med **2002**; 17:180–5.
12. Ettredge M, Gerdes J, Karuga G. Using web-based search data to predict macroeconomic statistics. Commun ACM **2005**; 48:87–92.
13. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc **2006**:244–8.
14. Dailey L, Watkins RE, Plant AJ. Timeliness of data sources used for influenza surveillance. J Am Med Inform Assoc **2007**; 14:626–31.
15. Centers for Disease Control and Prevention. Seasonal flu: flu activity and surveillance. Available at: http://www.cdc.gov/flu/weekly/fluactivity.htm. Accessed October 2007.
16. Espino JU, Hogan WR, Wagner MM. Telephone triage: a timely data source for surveillance influenza-like diseases. AMIA Annu Symp Proc **2003**; 215–9.
17. Hogan WR, Tsui FC, Ivanov O, et al.; Indiana-Pennsylvania-Utah Collaboration. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. J Am Med Inform Assoc **2003**; 10:555–62.
18. Welliver RC, Cherry JD, Boyer KM, et al. Sales of nonprescription cold remedies: a unique method of influenza surveillance. Pediatr Res **1979**; 13:1015–7.
19. Magruder S. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. Johns Hopkins University Applied Physics Laboratory Technical Digest **2003**; 24: 349–53.
20. Davies GR, Finch RG. Sales of over-the-counter remedies as an early warning system for winter bed crises. Clin Microbiol Infect **2003**; 9: 858–63.
21. Lenaway DD, Ambler A. Evaluation of a school-based influenza surveillance system. Public Health Rep **1995**; 110:333–7.
22. Irvin CB, Nouhan PP, Rice K. Syndromic analysis of computerized emergency department patients' chief complaints: an opportunity for bioterrorism and influenza surveillance. Ann Emerg Med **2003**; 41: 447–52.
23. Yuan CM, Love S, Wilson M. Syndromic surveillance at hospital emergency departments—southeastern Virginia. MMWR Morb Mortal Wkly Rep **2004**; 53(Suppl):56–8.
24. Suyama J, Sztajnkrycer M, Lindsell C, Otten EJ, Daniels JM, Kressel AB. Surveillance of infectious disease occurrences in the community: an analysis of symptom presentation in the emergency department. Acad Emerg Med **2003**; 10:753–63.
25. Polgreen PM, Nelson FD, Neumann GR. Use of prediction markets to forecast infectious disease activity. Clin Infect Dis **2007**; 44:272–9.
26. Cooper CP, Mallon KP, Leadbetter S, Pollack LA, Peipins LA. Cancer Internet search activity on a major search engine, United States 2001–2003. J Med Internet Res **2005**; 7:e36.
27. Johnson HA, Wagner MM, Hogan WR, et al. Analysis of Web access logs for surveillance of influenza. Stud Health Technol Inform **2004**; 107:1202–6.