Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks

Herman Anthony Carneiro[1,2] and Eleftherios Mylonakis[1]

[1]Division of Infectious Diseases, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts; and [2]School of Medicine, Imperial College London, London, United Kingdom

Google Flu Trends can detect regional outbreaks of influenza 7–10 days before conventional Centers for Disease Control and Prevention surveillance systems. We describe the Google Trends tool, explain how the data are processed, present examples, and discuss its strengths and limitations. Google Trends shows great promise as a timely, robust, and sensitive surveillance system. It is best used for surveillance of epidemics and diseases with high prevalences and is currently better suited to track disease activity in developed countries, because to be most effective, it requires large populations of Web search users. Spikes in search volume are currently hard to interpret but have the benefit of increasing vigilance. Google should work with public health care practitioners to develop specialized tools, using Google Flu Trends as a blueprint, to track infectious diseases. Suitable Web search query proxies for diseases need to be established for specialized tools or syndromic surveillance. This unique and innovative technology takes us one step closer to true real-time outbreak surveillance.

Millions of people worldwide search online for health-related information each day [1], which makes Web search queries a valuable source of information on collective health trends [1–3]. The internet company Google recently launched an experimental tool for near real-time detection of influenza outbreaks by monitoring and analyzing health care–seeking behavior in the form of queries to its online search engine. The tool, Google Flu Trends, is a sophisticated Web-based tool for detection of regional outbreaks of influenza in the United States [4]. It is so promising that the Centers for Disease Control and Prevention (CDC) is testing it in the United States. Preliminary testing suggests that Google Flu Trends can detect regional outbreaks of influenza 7–10 days before conventional CDC surveillance [5]. The CDC uses laboratory and clinical data to publish national and regional weekly statistics, typically with a 1–2 week lag in reporting.

Real-time surveillance would alert public health care practitioners in the early phases of an outbreak, enabling them to promptly institute control measures and case finding and to ensure adequate access to treatment, thereby reducing morbidity and mortality [6, 7]. With international concerns about emerging infectious diseases, bioterrorism, and pandemics, the need for a real-time surveillance system is at an all-time high [8–10]. The data generated would also be useful for public health care practice, clinical decision making, and research [11].

The main aim of this article is to introduce the more generic Google Trends (GT) tool to health professionals, to show how they can track disease activity of interest to them. We describe GT, how the data are processed, potential uses, and the tool's strengths and limitations.

METHODS

Google Flu Trends. Google Flu Trends is available at http://www.google.com/flutrends/ (Figure 1). There is a close relationship between the number of people searching for influenza-related topics and those who have influenza symptoms. Naturally, all the people searching for influenza-related topics are not ill, but trends emerge when all influenza-related searches are added together. Google Flu Trends has strong correlations with retrospective surveillance data from the CDC and accurately estimated influenza levels 1–2 weeks earlier than published CDC reports (Figure 2).

Google Flu Trends was developed from the more generic

Received 16 February 2009; accepted 29 May 2009; electronically published 21 October 2009.

Reprints or correspondence: Dr Eleftherios Mylonakis, Harvard Medical School, Massachusetts General Hospital, Div of Infectious Diseases, 55 Fruit St, Gray Jackson 5, Rm GRJ-504, Boston, MA 02114-2696 (emylonakis@partners.org).

Clinical Infectious Diseases 2009; 49:1557–64

1058-4838/2009/4910-0017$15.00
DOI: 10.1086/630200

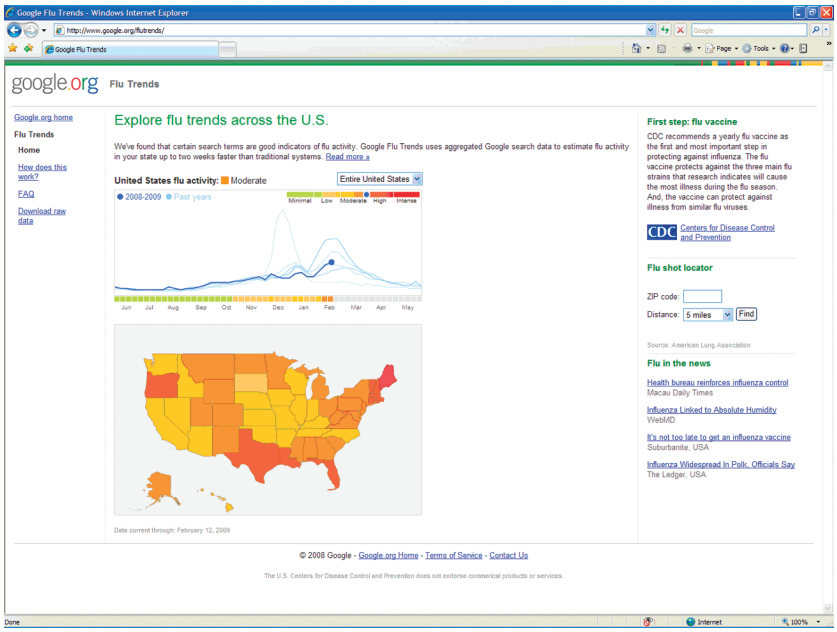

Figure 1. Google Flu Trends Web tool interface (available at http://www.google.com/flutrends/).

GT tool, which is available to all internet users at http://google.com/trends/. Users enter Web search queries to see the relative search volume of these queries—for example, queries for "flu" (Figure 3). See the Appendix (online only) for information on functionality and comparing trends. GT analyzes a fraction of the total Google Web searches over a period of time and extrapolates the data to estimate the search volume. This information is displayed in a search volume index graph, which is currently updated daily. Beneath the search volume index graph is the news reference volume graph. This graph shows the frequency with which the Web search queries appeared in Google News stories. When a spike is detected in the news reference volume graph, GT labels the search volume index graph with a headline of a relevant but randomly selected Google News story published near the time of the spike. These headlines are shown to the right of the search volume index graph. The regions, cities, and languages with the highest search volume are displayed on the bottom of the page.

Scaling the data. GT data are scaled in 2 ways: relative and fixed [12]. The difference between them is the time frame used to normalize the search volume. In relative mode, data are scaled using the average search volume over the time period selected. For example, the search volume index graph for "flu" is normalized using the extrapolated search volume for "flu" from January 2004 to the present. If the time frame is restricted—for example, restricted to 2008—then the data are scaled using the average search volume for "flu" in 2008 as the denominator.

In fixed mode, the data are normalized using the extrapolated search volume at a fixed time point (January 2004), which is when GT data start. Because the denominator does not change, users can look at different time periods and can relate them to

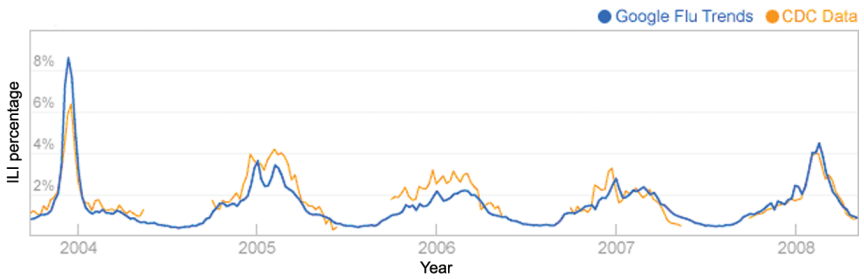

Figure 2. Correlation between Google Flu Trends and Centers for Disease Control and Prevention (CDC) surveillance data for the US Mid-Atlantic Region from 2004 through 2008. Reproduced from Ginsberg et al. [4], with permission from Nature Publishing Group. ILI, influenza-like illness.

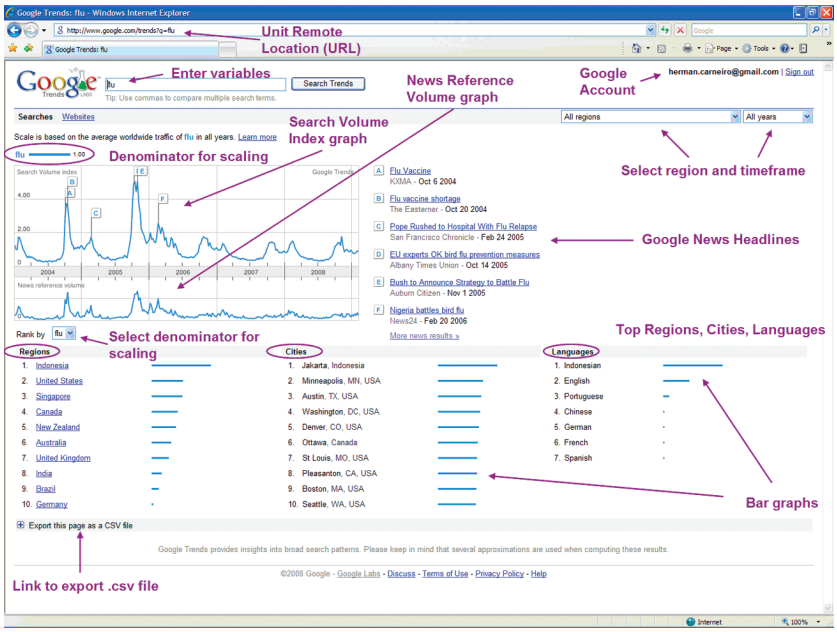

Figure 3. Google Trends output for Web search queries for the term "flu" worldwide from January 2004 to March 2009.

each other. A Web search query for the terms being investigated must have been done at the reference time point (January 2004); otherwise, no denominator exists for comparison. To view data scaled in fixed mode, the data must be downloaded in a comma-separated values (.csv) file and subsequently imported into a database or spreadsheet program. This feature is available only to users who are logged into their Google accounts. The news reference volume graph shows the raw number of Web search queries that appeared in Google News stories.

Data normalization. An increase in the volume of a Web search query increases its own average over time and thus its denominator for future comparisons. This reduces the sensitivity in detection of changes in future search volume trends. GT controls for this by dividing by an unrelated, common Web search query. For example, the search volume for the term "fever" may be normalized by dividing it by the search volume for the unrelated and common term "baseball." Normalization also compensates for population sizes, making it possible to rank cities purely on the basis of search volume trends. If, for example, the proportion of the population of New York that searches for "baseball" is the same as that of the population of Boston, the effect of the larger population of New York is factored out.

In ranking the top regions, cities, and languages, GT takes a sample of all Web search queries and determines the areas or languages from which the most searches for the entered terms originate. Internet protocol addresses from server logs are used to establish the origin of Web search queries. Language information is based on the language version of Google used for Web searches. The algorithm then calculates the ratio of a variable's search volume from each city and the total search volume from those cities. The city name and bar charts alongside it represent this ratio. When these are close together, the ranking between the cities is less meaningful.

RESULTS

Google Flu Trends uses a multitude of Web search queries that correlate well with physician visits for influenza-like symptoms to estimate current weekly levels of influenza activity at regional and state levels [4]. GT users, on the other hand, can enter only up to 5 Web search queries, which raises questions about its ability to monitor disease trends effectively. Some examples are presented below to illustrate the current issues in using GT for disease surveillance.

West Nile virus. West Nile virus (WNV) is a mosquito-borne disease that causes seasonal epidemics in the United States that peaks in the summer and continues into the fall. Its natural cycle is bird to mosquito to bird and mammal [13, 14]. From 1999 through 2001, the incidence of WNV in the United States was fairly stable. However, in 2002, WNV swept across the country. In 2001, there were 66 confirmed cases in 10 states in the United States, with a total of 10 fatalities [13]. In 2002, there were 4156 cases in 40 states, with 284 fatalities. The majority of cases occurred in the summer months, with a peak in August, when mosquitoes are most active [14] (Figure 4).

Most of the people who are infected with WNV develop no clinical illness or symptoms. Symptoms develop in 20%–40% of those infected [15–17], most of whom develop influenza-

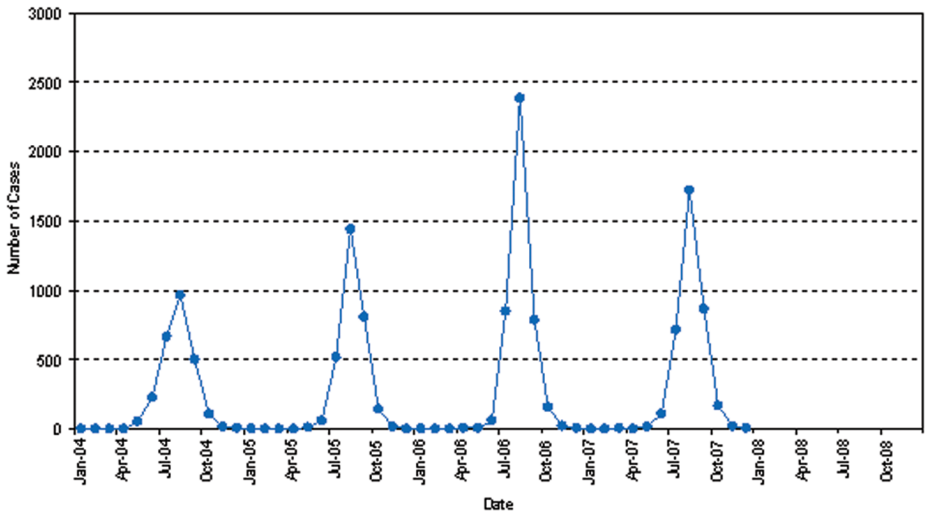

Figure 4. West Nile virus cases by month in the United States from 2004 through 2007. Data kindly provided by Centers for Disease Control and Prevention ArboNET surveillance group.

like symptoms termed "West Nile fever," characterized by fever, headache, malaise, myalgia, fatigue, skin rash, lymphadenopathy, vomiting, and diarrhea [18]. When the central nervous system is affected, clinical syndromes ranging from febrile headaches to aseptic meningitis and encephalitis occur [19].

GT data show good correlation to CDC surveillance data for WNV. The search volume index graph for WNV shows a cyclical pattern, with peaks in August each year from 2004 through 2008, which correspond well with peaks in the number of actual cases reported to the CDC (Figure 5). Figure 6 shows the search volume index graph of symptoms of West Nile fever, including fever, headache, fatigue, rash, and eye pain. Of note, rash has seasonal patterns that correspond well to the peaks in the number of WNV cases. Increases in search volume for rash, which may be a good proxy for WNV, start to increase in May, just before the increases in cases seen in June each year in the CDC data. Also of note, the top-ranked US cities on GT are in states with the highest burden of actual cases, according to CDC surveillance data for WNV [20].

Respiratory syncytial virus. There are yearly outbreaks of

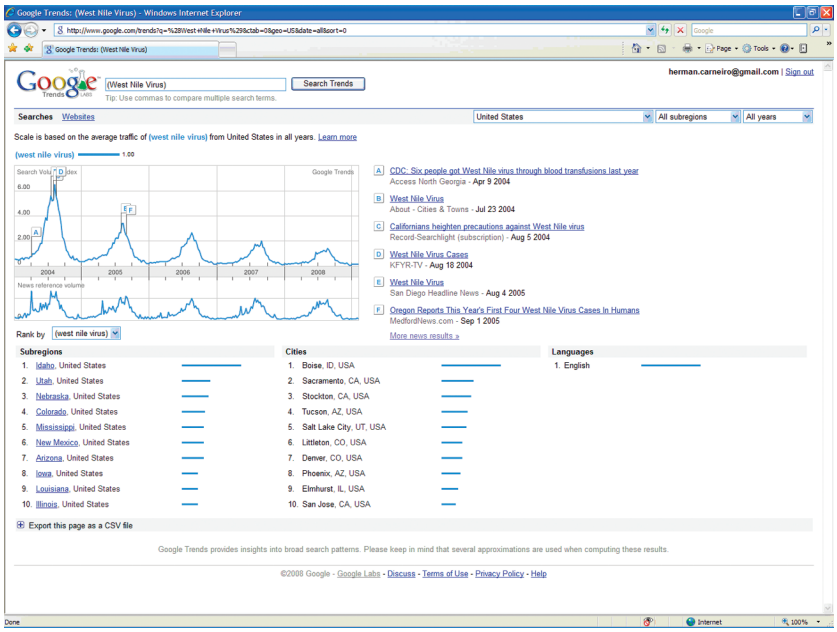

Figure 5. Google Trends output for Web search queries for the term "West Nile virus" in the United States from January 2004 to March 2009.

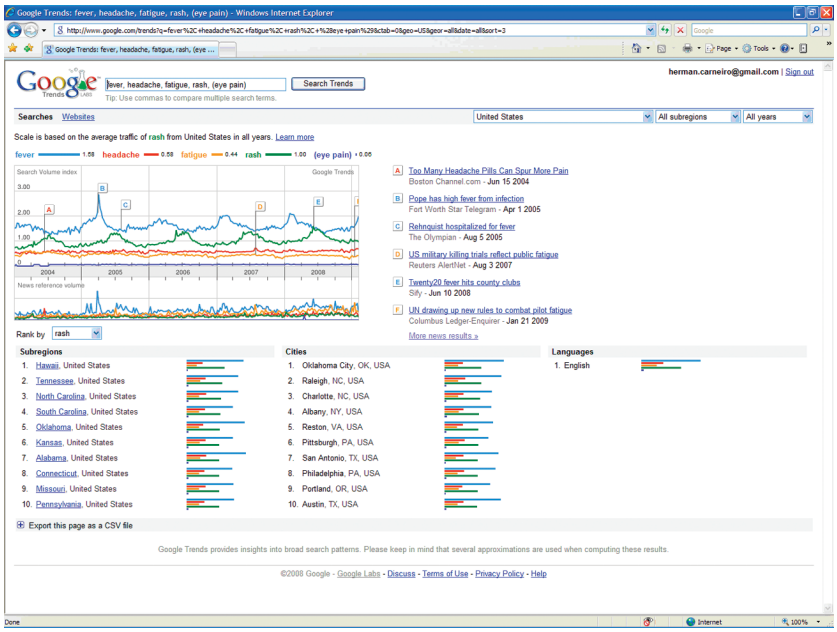

Figure 6. Google Trends output for Web search queries for symptoms of West Nile fever in the United States from January 2004 to March 2009. The search volume for "rash" is the denominator used for comparison with the search volumes for the other symptoms.

respiratory syncytial virus (RSV) in the United States that usually last 3–4 months during the fall, winter, and/or spring months [21]. The search volume index graph for the term "RSV" in the United States (Figure 7) shows spikes in search volume that correspond well to established seasonal patterns of RSV outbreaks in the United States [22, 23].

Avian influenza. Avian influenza, or "bird flu," refers to influenza A viruses passed from birds to humans. The majority of cases have resulted from contact with infected poultry, such as domestic chickens, ducks, and turkeys [24, 25]. Human-to-human transmission has been reported but is very rare. The symptoms of avian influenza depend on which virus caused the

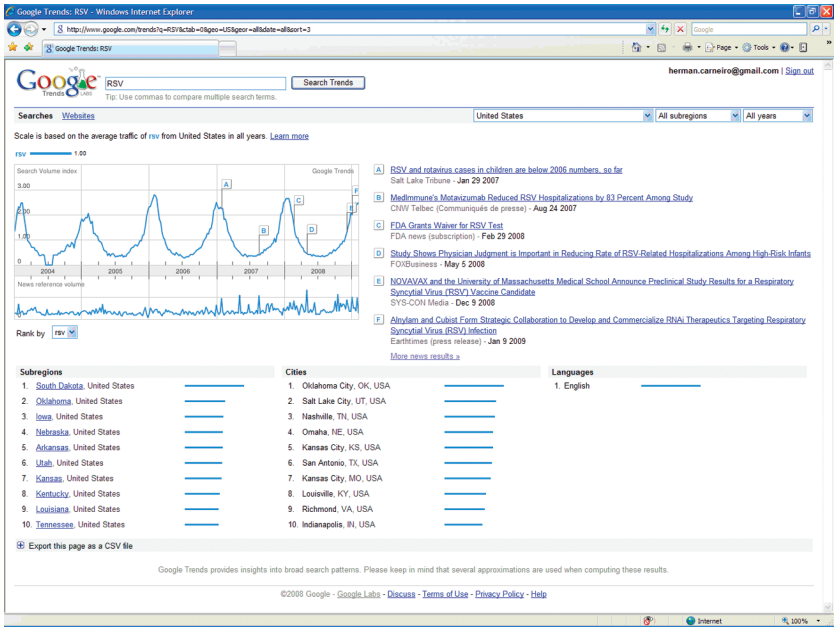

Figure 7. Google trends output for Web search queries for the term "RSV" (respiratory syncytial virus) in the United States from January 2004 to March 2009.

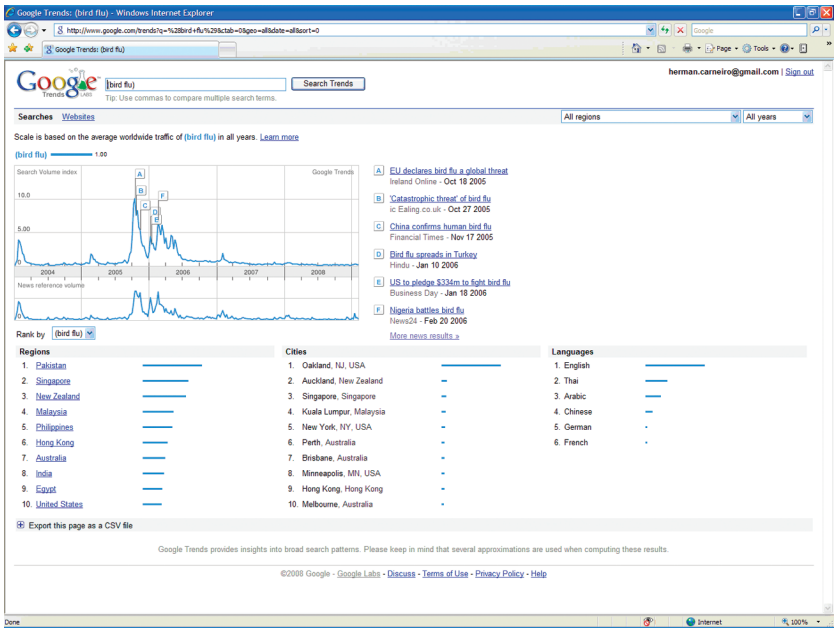

Figure 8. Google trends output for Web search queries for the term "bird flu" worldwide from January 2004 to March 2009.

infection. Symptoms depend on the infecting virus strain and range from typical influenza-like symptoms (eg, fever, headache, cough, sore throat, and myalgias) to eye infections, pneumonia, acute respiratory distress, and other severe and life-threatening complications [24]. Search volume index graphs for the term "(bird flu)" worldwide and in the United States are shown in Figures 8 and 9, respectively. The worldwide search volume index graph indicates an outbreak of avian influenza between 2005 and 2006, which spread from China to Turkey. The US search volume index graph shows a similar spike; however, there were no reported outbreaks of avian influenza in the United States in 2005 and 2006 [26]. The US spike is probably a reaction to media reports of an outbreak in Asia.

DISCUSSION AND CONCLUSIONS

GT is in the early phases of development, and its data "may contain inaccuracies for a number of reasons, including data sampling issues and approximation methods used to compute the results" [12]. Currently, GT search criteria are not standardized. Users may enter symptoms differently, depending on their level of education and their cultural and language backgrounds. For example, users may enter fever, pyrexia, chills, and rigors for the same symptom. Detailed analyses are required to find search query proxies that correlate well with specific diseases. The WNV example mentioned above shows that a Web search proxy may be used to track disease, but a rigorous statistical analysis is needed to verify this. These queries can be standardized by grouping search queries into syndromes. Syndromic surveillance has been established using the chief presenting complaints in emergency departments and medical records, with good success [9, 13, 18–28]. Clinicians can increase vigilance when spikes in search volume for syndromes increase. The creation of specialized tools for diseases or syndromic surveillance systems, however, may not be possible, for several reasons: diseases with low prevalences may not generate enough search volume, endemic diseases may have subtle changes from baselines levels that do not add value to traditional surveillance systems, and the time and resources for development of dedicated tools are likely to be limited.

To be most effective, GT requires large populations of Web search users, which means that GT is currently better suited to tracking disease activity in developed countries. Furthermore, GT is available only in a limited number of languages and region-specific versions.

In the examples of WNV and RSV given above, the data show good correlation to seasonal spikes in disease activity. This shows that GT may be able to signal disease activity while being constrained to 5 search variables. A rigorous analysis is needed to validate the search data against actual disease reports. We were unsuccessful in our attempts to contact Google for raw data to conduct such an analysis.

The example of "bird flu" showed spikes in search volume in regions where there were no actual cases of disease. This emphasizes the need for GT data to be used in conjunction with surveillance data from traditional modalities. A detailed analysis

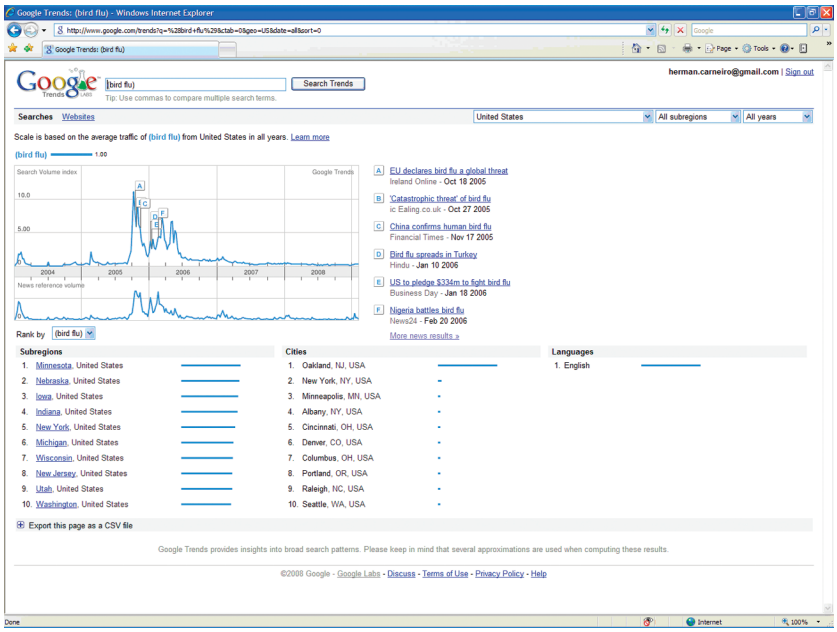

Figure 9. Google trends output for Web search queries for the term "bird flu" in the United States from January 2004 to March 2009.

is required to assess the statistical methods that would account for this phenomenon. With globalization and jet-powered airplane travel, spikes in search volume serve to increase vigilance for diseases that can spread between countries rapidly, such as severe acute respiratory syndrome (SARS) and swine influenza. SARS, a coronaviral respiratory illness, likely originated in mainland China and spread to Hong Kong, Vietnam, Singapore, Canada, Germany, and beyond, in just a few days. From November 2002 to July 2003, >8098 probable SARS cases with >774 deaths were reported in 26 countries [29]. An epidemic of swine influenza A (H1N1) recently started in Mexico, with 26 laboratory-confirmed cases and 7 confirmed deaths in the first week of the epidemic. Within 1 week, there were confirmed cases in Canada, New Zealand, the United Kingdom, Israel, and Spain, causing the World Health Organization to raise its influenza pandemic alert from phase 3 to phase 4 [30]. Of note, there has been a large spike in the search volume for "(swine flu)" on GT that started on 23 April 2009 and, as of 29 April 2009, the 7 countries with the highest number of searches were the United States, New Zealand, Canada, Singapore, the United Kingdom, Australia, and Mexico.

GT is an exciting tool with enormous potential, as shown by Google Flu Trends. It is a convenient, easy, and accessible source of search data. In its current stage of development, GT can be used in conjunction with traditional surveillance systems to improve the efficacy of disease surveillance. Google Flu Trends has detected influenza outbreaks 7–10 days before the traditional surveillance systems used by the CDC. GT also has this potential, as shown by the examples of WNV and RSV. Experts in their respective fields should work with Google to build a specialized tool for infectious diseases that are amenable to this type of surveillance.

Still in its early phases of development, GT data may contain inaccuracies, which Google is working to resolve. Research is needed to find suitable Web search query proxies that correlate well to actual cases of diseases of interest. These proxies then can be used to establish specialized tools for infectious diseases, using Google Flu Trends as a blueprint, or to setup syndromic surveillance of Web search queries. Also, there are privacy issues involved in using Google Web search data. Google stores and uses data from personal Web searches for public research, often without the consent and knowledge of Internet users. In some cases, Google search data may be traced to individuals if they are signed into their accounts when they conduct online searches. Google assures users that personal search data remain safe and private.

In conclusion, GT is currently better suited to track epidemics, diseases with high prevalences, and diseases in developed countries than other types of diseases, because it requires large populations of Web search users to be most effective. However, the world is changing, and society is becoming more dependent on the Internet, thus providing a wealth of information that reflects the "collective intelligence" of populations. Google Flu Trends, and possibly GT, make it possible to track infectious disease activity faster than by traditional surveillance systems. This unique and innovative technology takes us one step closer to true real-time outbreak surveillance.

Acknowledgments

Potential conflicts of interest. E.M. has received research funding from Astellas, has served as a consultant for Bind, and has been on the speakers' bureau for Pfizer. H.A.C.: no conflicts.

References

1. 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.
2. Eysenbach G. Infodemiology: tracking flu-related searches on the Web for syndromic surveillance. AMIA Annu Symp Proc **2006**: 244–8.
3. Polgreen PM, Chen Y, Pennock DM, Forrest ND. Using Internet searches for influenza surveillance. Clin Infect Dis **2008**; 47:1443–8.
4. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature **2009**; 457:1012–4.
5. New York Times. Google uses searches to track flu's spread. 11 November **2008**. Available at: http://www.nytimes.com/2008/11/12/technology/internet/12flu.html?_r = 1. Accessed 19 January 2009.
6. Ferguson NM, Cummings DA, Cauchemez S, et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature **2005**; 437:209–14.
7. Longini IM Jr, Nizam A, Xu S, et al. Containing pandemic influenza at the source. Science **2005**; 309:1083–7.
8. Frenk J, Gomez-Dantes O. Globalization and the challenges to health systems. Health Aff (Millwood) **2002**; 21:160–5.
9. 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.
10. National Electronic Disease Surveillance System Working Group. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine. J Public Health Manag Pract **2001**; 7:43–50.
11. Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc **2004**; 11:141–50.
12. Google Trends Labs. About Google Trends. Available at: http://www.google.com/intl/en/trends/about.html#7. Accessed 1 October 2009.
13. Centers for Disease Control and Prevention, Division of Vector-Borne Infectious Diseases. West Nile virus: statistics, surveillance, and control. 2001 West Nile virus activity in the United States. Available at: http://www.cdc.gov/ncidod/dvbid/westnile/surv&controlCaseCount01_detailed.htm. Accessed 12 January 2009.
14. Liu H, Weng Q, Gaines D. Spatio-temporal analysis of the relationship between West Nile virus dissemination and environmental variables in Indianapolis, USA. Int J Health Geogr **2008**; 7:66.
15. Kramer LD, Li J, Shi PY. West Nile virus. Lancet Neurol **2007**; 6:171–81.
16. Hubálek Z, Halouzka J, Juricová Z. West Nile fever in Czechland. Emerg Infect Dis **1999**; 5:594–5.
17. Mostashari F, Bunning ML, Kitsutani PT, et al. Epidemic West Nile encephalitis, New York, 1999: results of a household-based seroepidemiological survey. Lancet **2001**; 358:261–4.
18. Hayes EB, Gubler DJ. West Nile virus: epidemiology and clinical features of an emerging epidemic in the United States. Annu Rev Med **2006**; 57:181–94.
19. Centers for Disease Control and Prevention, Division of Vector-Borne Infectious Diseases. West Nile virus: clinical description. Available at: http://www.cdc.gov/ncidod/dvbid/westnile/clinicians/clindesc.htm. Accessed 18 January 2009.
20. Centers for Disease Control and Prevention, Division of Vector-Borne Infectious Diseases. West Nile virus: statistics, surveillance, and control. 2007 West Nile virus activity in the United States. Available at: http://www.cdc.gov/ncidod/dvbid/westnile/surv&controlCaseCount07_detailed.htm. Accessed 12 January 2009.
21. Collins PL, Crowe JE Jr. Respiratory syncytial virus and metapneumovirus. In: Knipe DM, Howley PM, eds. Fields virology. Philadelphia: Wolters Kluwer and Lippincott Williams & Wilkins, **2007**:1601–46.
22. Centers for Disease Control and Prevention. CDC features: respiratory syncytial virus (RSV) season varies by region and year. Available at: http://www.cdc.gov/Features/dsRSV/. Accessed 17 January 2009.
23. Mullins JA, Lamonte AC, Bresee JS, Anderson LJ. Substantial variability in community respiratory syncytial virus season timing. Pediatr Infect Dis J **2003**; 22:857–62.
24. Centers for Disease Control and Prevention. Key facts about avian influenza (bird flu) and avian influenza A (H5N1) virus. Available at: http://www.cdc.gov/flu/avian/gen-info/facts.htm. Accessed 17 January 2009.
25. Thomas JK, Noppenberger J. Avian influenza: a review. Am J Health Syst Pharm **2007**; 64:149–65.
26. Centers for Disease Control and Prevention. Past avian influenza outbreaks. Available at: http://www.cdc.gov/flu/avian/outbreaks/past.htm. Accessed 17 January 2009.
27. Carrico R, Goss L. Syndromic surveillance: hospital emergency department participation during the Kentucky Derby Festival. Disaster Manag Response **2005**; 3:73–9.
28. Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health **2001**; 1:9.
29. World Health Organization, Regional Office for the Western Pacific. Health topics: SARS. Available at: http://www.wpro.who.int/health_topics/sars/. Accessed 29 April 2009.
30. World Health Organization. Global alert and response (GAR). Swine influenza—update 4. Available at: http://www.who.int/csr/don/2009_04_28/en/index.html. Accessed 29 April 2009.