Using Internet Searches for Influenza Surveillance

(Internet Search Term Surveillance for Flu)

Philip M. Polgreen[1,2],

Yiling Chen[3],

David M. Pennock[4],

Forrest D. Nelson[5]

[1] Department of Internal Medicine, University of Iowa Carver College of Medicine, Iowa City, IA

[2] Department of Epidemiology, University of Iowa College of Public Health, Iowa City, IA

[3] School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

[4] Yahoo! Research, New York, NY

[5] Department of Economics, Henry B. Tippie College of Business University of Iowa, Iowa City, IA

Abstract word count = 173; Text word count = 2,831

Corresponding Author Contact Information:

Philip M. Polgreen, MD

University of Iowa Department of Internal Medicine,

200 Hawkins Drive,

Iowa City, IA 52242,

Phone: 319-384-6194,

Fax: 319-353-5646

email: Philip-Polgreen@uiowa.edu

Keywords: influenza, internet, forecasting, surveillance, search

Abstract

Background: 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 examine the relationship between searches for influenza and actual influenza occurrence.

Methods: Using search queries from http://search.yahoo.com, between March 2004 and May 2008, we counted daily unique queries, originating in the U.S. and containing 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 one- to ten-week lead times as explanatory variables, to predict the percentage of positive influenza cultures and also deaths due to pneumonia and influenza in the U.S.

Results: Using the frequency of searches, our models predicted an increase in positive influenza cultures 1-3 weeks in advance ($p < 0.0001$) and similar models predicted an increase in mortality from pneumonia and influenza up to five weeks in advance ($p < 0.0001$).

Conclusions: Search-term surveillance may provide an additional tool for disease surveillance.

Introduction

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

In addition to medical journals and public health websites, news websites 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 like the Infectious Diseases Society of America, the American Society of Microbiology, and the Society for Healthcare Epidemiology of America also support websites 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 [3,4,5].

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 (GPHIN), developed by the Public Health Agency of Canada, continuously monitors media sources and web-based information related to disease outbreaks around the world [6]. GPHIN data is not available to the general public. However, a relatively new site at HealthMAP.org, monitors information from a variety of sources and displays results in real-time on a world map [7]. Access to this website is free and available to the public.

An estimated 113 million people in the U.S. use the Internet to find health-related information [8]. Searchers include patients and their families as well as healthcare providers [8, 9, 10, 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 are searching for health-related 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 clicks on a "sponsored link" on Google Adsense, triggered by Canadian searchers entering "flu" or "flu symptoms", 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 (CDC) influenza surveillance program identifies disease as, or after, it occurs, and therefore does not provide advance warnings. Furthermore, the CDC's data regarding influenza activity are no longer current when released to healthcare providers. 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 occurrence 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 two types of U.S. influenza surveillance data. The first type was based on weekly influenza cultures [15]. Each week during the influenza season, clinical laboratories throughout the U.S., which are either members of the World Health Organization (WHO) Collaborating Laboratories or National Respiratory and Enteric Virus Surveillance System (NREVSS), report the total number of respiratory specimens tested and the number that test positive for influenza.

The second type of data summarizes weekly mortality from 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 also the number that list pneumonia or influenza as the underlying and/or contributing cause of death. Based on 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 to May 2008.

Search Data

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

We have two series of influenza-related search fraction data at the national level:

1. the fraction of U.S. search queries that contain “influenza” or “flu” but do not contain “bird”, “avian”, or “pandemic”;

2. the fraction of U.S. search queries that contain “influenza” or “flu” but do not contain “bird”, “avian”, “pandemic”, “vaccine”, “vaccination”, or “shot”.

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

We also classified weekly influenza-related search data into nine U.S.-census regions. Census-region data were normalized by total searches within that region. Because we identified the geographic location of origin from the IP address, there were cases for which we were not able to identify the exact region in which a search originated but were able to identify that it came from within the U.S. Thus, the sum of the search data for the nine U.S.-census regions does not equal the data at the national level. For each census region, we obtained only one series of data: weekly search data from the region for queries that contain “influenza” or “flu”, but do not contain “bird”, “avian”, “pandemic”, “vaccine”, “vaccination”, and “shot”.

Search and Influenza Positive Culture Results

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 positive influenza cultures follow similar patterns, but a sharp increase in searches precedes the sharp increase in the rate of positive cultures. Using the culture data, we fitted the following linear model to test the predictability of search frequency on positive influenza cultures, 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 positive influenza cultures 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 eleven values for x and compared the R^2 value for each model. The model with a search term with a one-week lag fit best. However, models with lags up to 3 weeks in advance of culture data fit similarly in terms of R^2. A summary of the regression results for the 0-10-week-lag-search-term models are presented in Table 1.

The coefficient on the time trend is not significantly different from zero in any of the models. However, there is a positive relationship between the fraction of influenza related queries and positive influenza culture rates two weeks later ($p < 0.001$). The large coefficient on s_{t-2} reflects the fact that influenza-related search frequency is measured as a fraction of all searches. The predicted values from the 2 week model and the actual culture data are presented in Figure 2.

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

Search and Influenza Mortality Results

Figure 3 plots the time series of influenza-related searches and influenza mortality for the U.S. 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 on influenza mortality:

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

where m_t is the total number of deaths from pneumonia and influenza in week t, and all other variables are as defined earlier. A model incorporating searches at time t-5 fits slightly better than other models with a search variable ranging from time t to time t-10. All of the regression results using searches from 0-10 week lags are listed in Table 2. A positive relationship exists between the fraction of influenza-related search queries and pneumonia and influenza mortality 5 weeks later ($p < 0.001$). The large coefficient on s_{t-5} reflects the fact that influenza-related search frequency is measured as a fraction of all searches and thus takes on small values, on the order of 10^{-6}. Figure 4 shows the predicted values from the 5 week model and the actual mortality data.

Finally, we fit models with lags from 0-10 weeks for each of the 9 U.S. census regions. Results were similar to the national model: for the best fitting models, searches peaked 4-6 weeks before deaths from influenza and pneumonia. The average R^2 at 5 weeks was 0.3041. However, values varied from 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 varies, complicating efforts to produce reliable and timely estimates of influenza activity. However, we found that a distinct temporal association exists between search-term frequency and influenza disease activity. On a national level, influenza-related search-term activity seems to precede an increase in influenza cultures and deaths from pneumonia and influenza. Furthermore, the temporal relationship between searches and cultures and searches and mortality corresponds to the epidemiology of influenza, since deaths from pneumonia typically peak a few weeks after influenza cases peak.

Investigators have suggested several supplemental approaches for influenza surveillance, both at pre-diagnosis and diagnosis stages. Pre-diagnosis approaches mainly include the analysis of information collected before specific influenza-related diagnoses are made: telephone triage calls [16], the over-the-counter prescriptions for respiratory diseases [17, 18, 19, 20], and the analysis of absentee data [21]. In contrast, diagnosis-level approaches attempt to gather clinical data from emergency department visits [22, 23, 24] or microbiologic sources in as close to real-time as possible. The timeliness of influenza syndromic 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 pre- and post-diagnostic information [25]. In general, the efforts described above 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 one 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 is efficiently collected in a standard usable form and aggregates both pre-and post-diagnostic information. Although it is difficult to compare with other methods, search data seems to perform reasonably well. In addition, it is easy to collect and unlike other non-traditional 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 using data from the nine census regions and found that influenza related searches are statistically significantly related to influenza mortality. Models in some regions perform better than others, suggesting that information in some regions may generate searches in other regions. Further work is needed to examine the spatial relationship between searches and the geographic spread of influenza. However, because culture and mortality data are not uniformly reported at the state level, our geographic analysis stopped at the census region level.

Despite the promise of using search data for surveillance purposes, there are several limitations. First, with only four years of data, the inferential conclusions from time-series analysis are limited. A second limitation: 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. Cancer researchers, using Yahoo search queries, found that daily variations of search frequency were heavily influenced by news reports [26]. However, Internet searches for specific cancers were still correlated with their estimated incidence and mortality. Also, a news item causing a large increase in search volumes should be easy to identify and rather short lived given the half-life of most news cycles.

The limited geographic data gleaned from search terms is a third limitation of search data. Geographic search data are extracted from IP addresses and may not always represent actual geographic location. Privacy issues represent another significant limitation. The search data described in this paper was aggregated across users for 9 census regions. However, search data with much finer geographic information linked to individuals across multiple different searches for different topics could represent a privacy concern. Thus, we envision health investigators only using aggregated search volumes over 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 website hits or web searches on specific websites may also provide useful information about disease activity. For example, the number of articles retrieved on the site Healthlink, a consumer health information website maintained at the Medical College of Wisconsin, was correlated with influenza activity [27]. Thus, searches for specific diseases on high traffic websites (e.g., a state health department) may provide important time-series data as it captures the number and to some extent the geographic location (via IP address) of people investigating the activity of a specific disease. Searches for specific medical conditions on the National Library of Medicine's PubMed website may indicate changing patterns in infectious disease activity or potential adverse drug events. Furthermore, changes in volume of searches on commercial websites (e.g., Up-To-Date, MD Consult) may indicate gaps in clinical knowledge, or the need for clinical trials. Data from such sites may be more representative of what healthcare providers 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. Using search series is not limited to influenza; it could also be used to monitor emerging and re-emerging infectious diseases and to detect changes in phenomena related to chronic illnesses. Surveillance of symptom-based searches (e.g., diarrhea) may help detect outbreaks if search levels rise above an established baseline. De-identified search volumes for sexually transmitted infections (e.g., syphilis) may provide public health officials indications of disease trends in advance of official reports of disease activity. Although search probably provides some aggregation of news reports, it also adds a behavioral component by signaling how important topics are to searchers. 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.

Acknowledgements

This work was in part supported by the 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, Inc. D.M.P. is an employ of Yahoo! Research and Y.C. was an employee of Yahoo! research.

References

1. National Library of Medicine / 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 April 25, 2008.

2. Burdette SD. Electronic tools for infectious diseases and microbiology. Can J Infect Dis Med Microbiol, 2007; 18(6):347--52.

3. Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis, 2004; 39(2):227--32.

4. Strausbaugh LJ, Liedtke LA. The Emerging Infections Network electronic mail conference and web page. Clin Infect Dis, 2001; Jan 15; 32(2):270--6.

5. Dwyer V. ClinMicroNet - Sharing experiences and building knowledge virtually. Clinical Microbiology Newsletter, 2003; 25(16):121--125.

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(1):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(2):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 April 25, 2008.

9. Ybarra ML, Suman M. Help seeking behavior and the Internet: a national survey. Int J Med Inform, 2006; 75(1):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(3):180--5.

12. Ettredge M, Gerdes J, Karuga G. Using web-based search data to predict macroeconomics statistics. Commun ACM, 2005; 48(11):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, Gesteland PH, Grannis S, Overhage JM, Robinson JM, Wagner MM; 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(6):555-62.

18. Welliver RC, Cherry JD, Boyer KM, Deseda-Tous JE, Krause PJ, Dudley JP, Murray RA, Wingert W, Champion JG, Freeman G. Sales of nonprescription cold remedies: a unique method of influenza surveillance. Pediatr Res, 1979; 13(9):1015-7.

19. Magruder S. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of public health. 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(8):858-63.

21. Lenaway DD, Ambler A. Evaluation of a school-based influenza surveillance system. Public Health Rep, 1995;110(3):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(4):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(7):753-63.

25. Polgreen PM, Nelson FD, Neumann GR. Use of prediction markets to forecast infectious disease activity. Clin Infect Dis, 2007; 44(2):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(3):e36.

27. Johnson HA, Wagner MM, Hogan WR, Chapman W, Olszewski RT, Dowling J, Barnas G. Analysis of Web access logs for surveillance of influenza. Medinfo, 2004;11 (Pt 2):1202--6.

Table 1: Positive Influenza Culture Regression Results

X (Lag in weeks)	Coefficient:S_{t-x}	Std. Error	t	P > |t|	R^2
0	239636.2	18301.99	13.09	<0.001	0.4672
1	242579.5	18218.11	13.32	<0.001	0.4723
2	239568.6	18487.33	12.96	<0.001	0.4568
3	234749.1	18848.97	12.45	<0.001	0.4356
4	229446.4	19225.16	11.93	<0.001	0.4134
5	223257.3	19628.85	11.37	<0.001	0.3890
6	215900.2	20064.8	10.76	<0.001	0.3618
7	206683.5	20565.4	10.05	<0.001	0.3300
8	195520.6	21118.44	9.26	<0.001	0.2943
9	184502.1	21619.25	8.53	<0.001	0.2610
10	173491.3	22164.1	7.83	<0.001	0.2305

Table 2: Influenza Mortality Regression Results

X (Lag in weeks)	Coefficient:S_{t-x}	Std. Error	t	P > |t|	R^2
0	3300788	436385.8	7.56	<0.001	0.2075
1	3810620	415148.2	9.18	<0.001	0.2787
2	4194847	394455.2	10.63	<0.001	0.3418
3	4445665	378633.3	11.74	<0.001	0.3882
4	4604043	367573.4	12.53	<0.001	0.4198
5	4625652	368166.3	12.56	<0.001	0.4229
6	4461079	379889.1	11.74	<0.001	0.3919
7	4314867	390405	11.05	<0.001	0.3649
8	4248610	396362.5	10.72	<0.001	0.3523
9	3992864	410770.2	9.72	<0.001	0.3111
10	3767351	422055.3	8.93	<0.001	0.2765

Figure 1. Searches for Influenza and Positive Influenza Cultures by Week

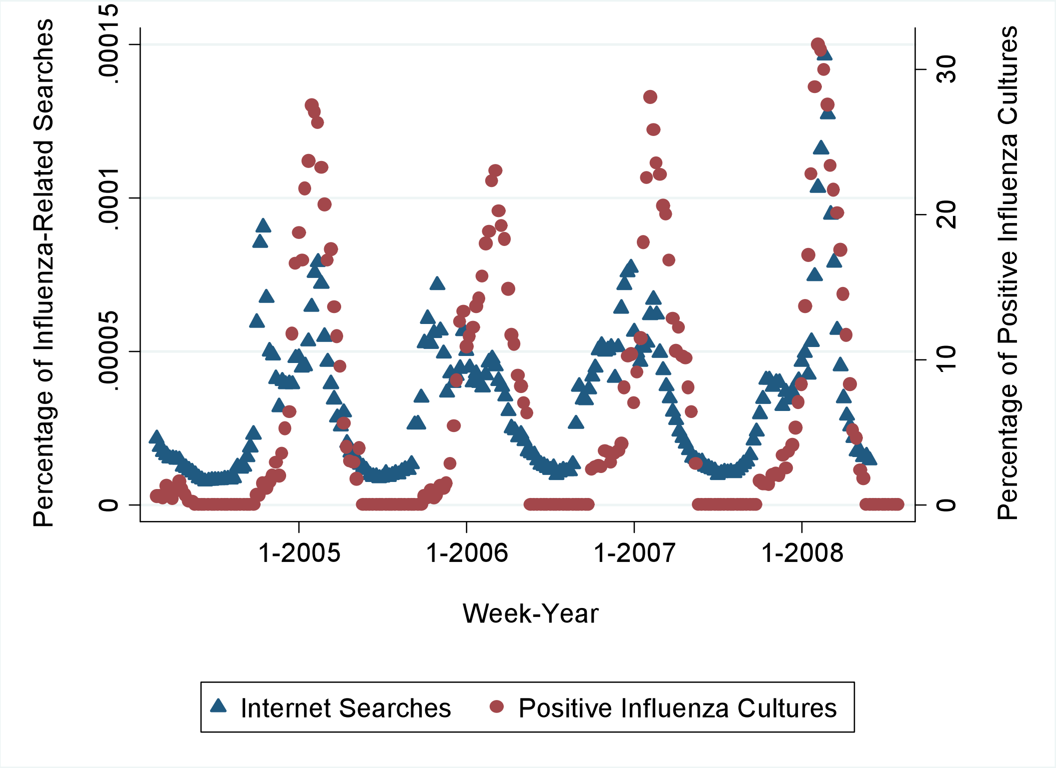


Figure 2. Predicted Values for Positive Influenza Cultures Based on Searches and Actual Values by Week

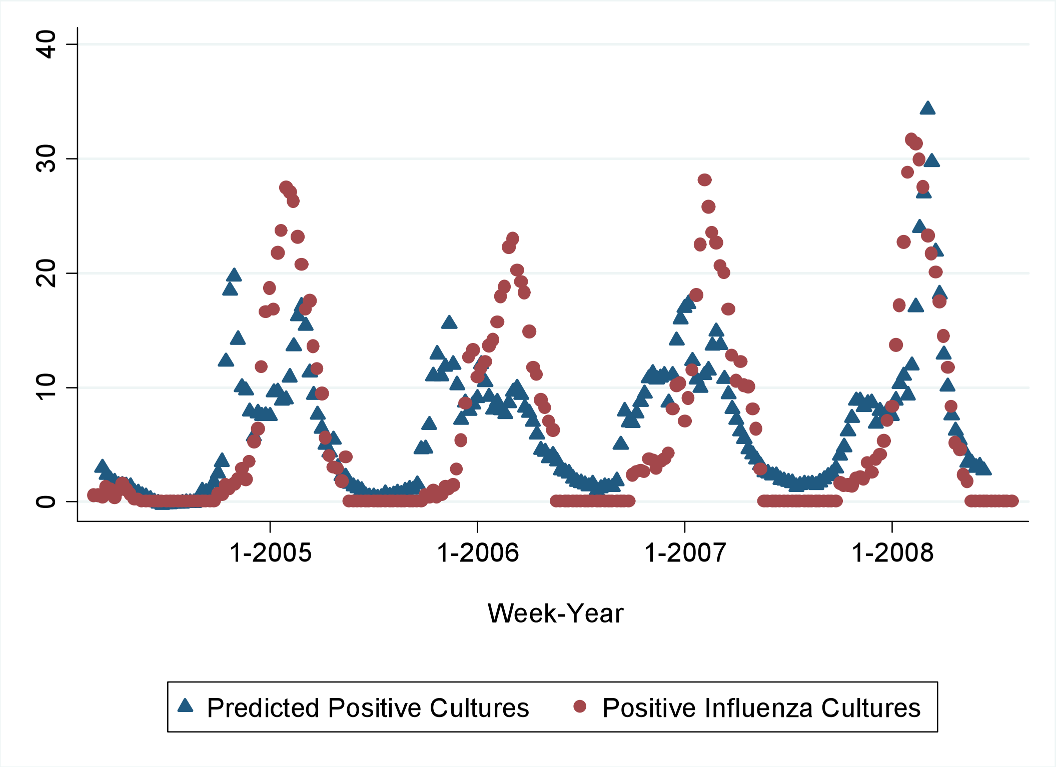


Figure 3. Searches for Influenza and Mortality from Influenza and Pneumonia by Week

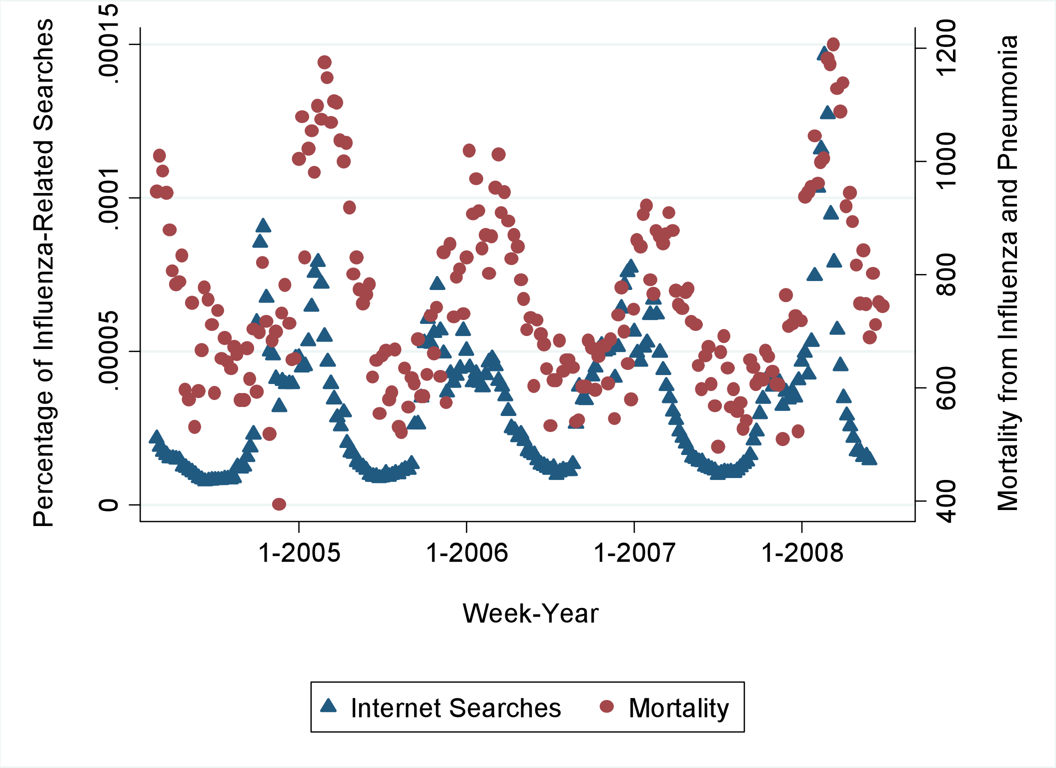
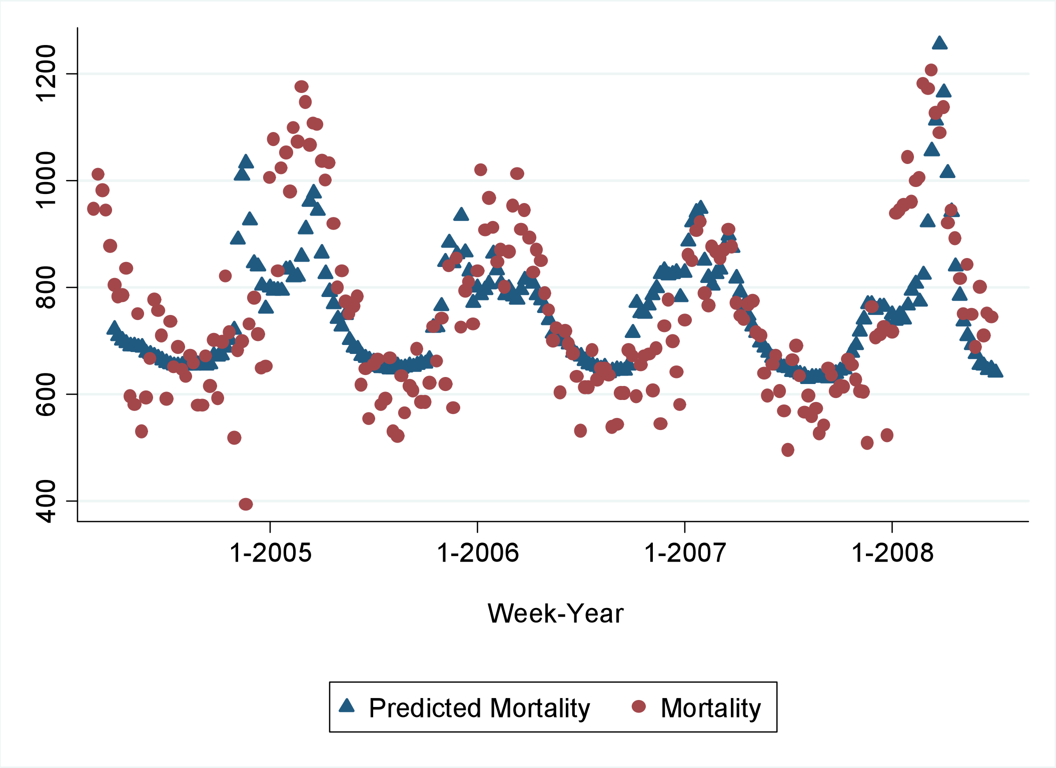
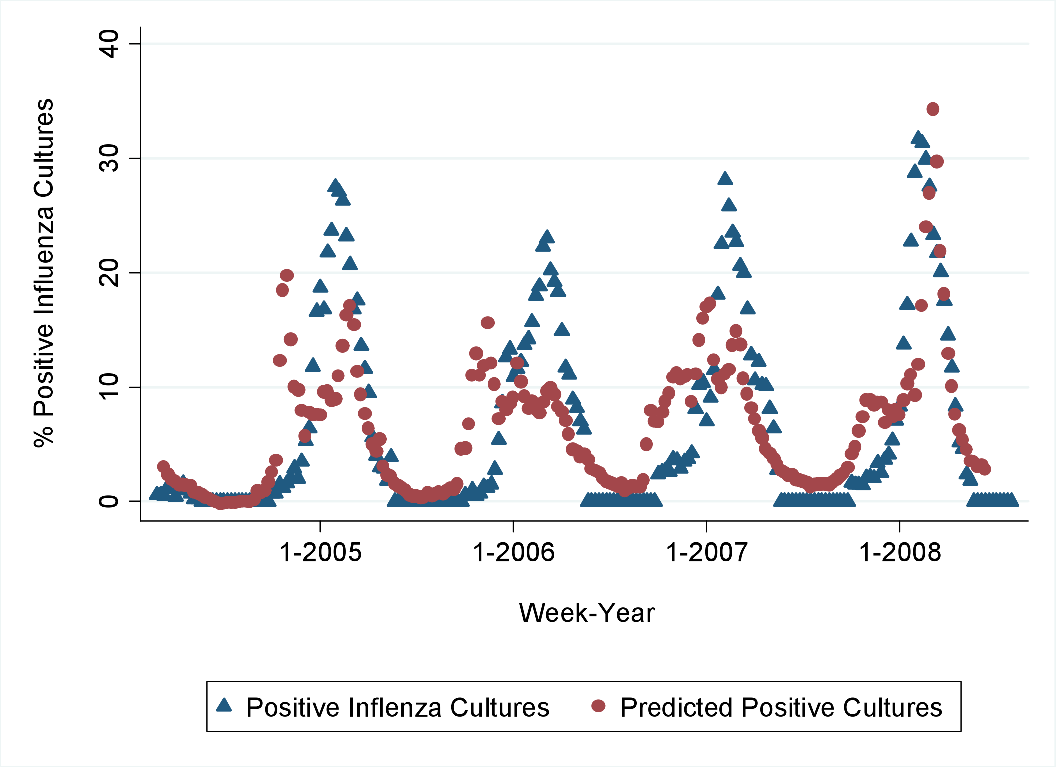


Figure 4. Predicted Values for Mortality from Influenza and Pneumonia Based on Searches and Actual Values by Week




40
30
20
10
0
% Positive Influenza Cultures
1-2005
1-2006
1-2007
1-2008
Week-Year
▲ Positive Inflenza Cultures
● Predicted Positive Cultures