MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Analysis of Web Access Logs for Surveillance of Influenza

Heather A. Johnson[a], Michael M. Wagner[a], William R. Hogan[a], Wendy Chapman[a], Robert T Olszewski[a], John Dowling[a], Gary Barnas[b]

[a]RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, PA, USA
[b]Division of General Internal Medicine, Medical College of Wisconsin, WI, USA

Abstract

The purpose of this study was to determine whether the level of influenza in a population correlates with the number of times that internet users access information about influenza on health-related Web sites. We obtained Web access logs from the Healthlink Web site. They contain information about the user and the information the user accessed, and are maintained electronically by most Web sites, including Healthlink. We developed weekly counts of the number of accesses of selected influenza-related articles on the Healthlink Web site and measured their correlation with traditional influenza surveillance data from the Centers for Disease Control and Prevention (CDC) using the cross-correlation function (CCF). We defined timeliness as the time lag at which the correlation was a maximum. There was a moderately strong correlation between the frequency of influenza-related article accesses and the CDC's traditional surveillance data, but the results on timeliness were inconclusive. With improvements in methods for performing spatial analysis of the data and the continuing increase in Web searching behavior among Americans, Web article access has the potential to become a useful data source for public health early warning systems.

Keywords:

Internet, Patients, Public Health Informatics, Influenza, Disease Outbreaks, Signal Processing, Computer Assisted.

Introduction

With ongoing awareness of the threat of bioterrorism, researchers are looking at a variety of ways to improve the timeliness of detection of disease outbreaks. One such way is to use data that are collected routinely for other purposes to determine which types of outbreaks perturb these data and how early relative to expected time of detection by current methods [1]. Examples of such data include sales of over-the-counter drugs [2,3], patient chief complaints [4,5], electronic-laboratory reporting data [6], and telephone triage data [7]. Although Web access logs (data recorded by a Web server about what information was accessed and by whom) and Web queries (free-text search terms that users type into search engines) are thought to have excellent potential as an early signal of some disease outbreaks [8], to our knowledge there have been no published studies of the potential of such data for use in public health surveillance. The potential of data collected by Web sites for outbreak detection is suggested by surveys of internet utilization. Recent reports vary on the number of individuals that look for health information online, however, studies estimate that anywhere from 16.1% to 40% of Americans already look for health information online [9,10]. Surveys by the Pew Internet & American Life Project report that the "Internet population" – defined as those Americans with online access who identify themselves as computer users – has stabilized at about 60% of Americans since 2001, but the number of experienced users has grown significantly [9]. This fact suggests that with time, the proportion of the sick seeking information online can be expected to increase, possibly to high levels.

To determine the usefulness of Web access log data in detection of outbreaks, we analyzed a year of Web access logs provided by Healthlink, a consumer health information Web site. The Web access logs contain both the identity of articles that persons searching the Web retrieved and the free-text search terms that they entered. The present study is limited to an analysis of the articles retrieved; the results about free-text search terms will be reported separately. Specifically, the purpose of this study was to determine whether a correlation exists between the frequency of access of articles on influenza-related topics and actual influenza activity as measured by standard indicators, and whether increases in the frequency of influenza-article access precede increases in standard indicators of influenza activity.

Methods

Gold Standard Determination of Influenza Activity

The standard indicators of influenza activity that we used were obtained from datasets published by the Centers for Disease Control and Prevention (CDC). The CDC collects and makes available data on the weekly influenza activity in the United States during the influenza season (October through May), and we obtained these datasets from the CDC Web site at http://www.cdc.gov/ncidod/diseases/flu/weeklychoice.htm and used them to create the following two gold standards for the study:

1. The weekly regional number of influenza-like-illness[1] (ILI) cases from the U.S. Influenza Sentinel Physicians Surveillance Network (the *ILI gold standard*)

1. The CDC defines influenza-like illness as fever (temperature of >100°F) plus either a cough or a sore throat.

2. The weekly regional number of positive influenza tests (the *positive cultures gold standard*)

Although the CDC offers the influenza surveillance data broken down by regions of the United States, we used the national counts because the finest grained spatial resolution achievable at present from the Web access log data is national, as we will discuss.

Healthlink Web Logs

We obtained 12 Web access logs from Healthlink, a consumer health information Web site developed and maintained by the Office of Clinical Informatics at the Medical College of Wisconsin as a service to their patients and community. Each Web access log contained data for one month of the 2001 calendar year. The number of records in each dataset ranged from 328,580 to 526,260 with a total of 4,980,990 records for 2001.

Each record in the dataset contained eight elements:

1. IP address – the IP address of the computer making the HTTP request.
2. RFC – a field that is used to identify the person requesting information from the Web site. (This field was null in our Web access logs.)
3. Auth– a field included to list the authenticated user, if required for Web site access. (This field was also null in our Web access logs.)
4. Timestamp – formatted as DD/Mon/YYYY HH:MM:SS
5. Action – the action requested by the user (e.g., request for a specific article or for a search performed by Healthlink's search function)
6. Status – a code indicating whether the user experienced a success, redirect, failure, or server error when they tried to access the Web page.
7. Transfer Volume – this indicates how many bytes were transferred to the user.
8. Referring URL – the URL of any Web site that led the user to the Healthlink Web site.

When an http request is received by Healthlink (e.g., the access of a particular article or a search of the system for a subject) the above-listed information is recorded.

Parsing of Web Access Logs

Because the Web access logs had complex structure (see Figure 1), we developed a Perl script to parse the logs to extract the above fields.

Removal of non-US records

Healthlink receives queries from many countries. Because influenza season varies geographically (and is inverted in the southern hemisphere), which would tend to cancel out any seasonal increase, we limited the analysis to the smallest geographic area possible which was the United States.

To remove requests made by users not in the United States, we used GeoIP Country, an open source Perl module developed by MaxMind™ to analyze the IP address of each record and assign (when possible) the country of the user. We studied only those records that GeoIP country assigned as US or did not assign to any country. The rationale for including records to which GeoIP assigned no country was that the United States has 63% of all IP addresses [12]. Only 3.3% of the records in the Web access logs had no country assignment, so this decision likely did not change the results.

```
208.26.241.42 - - [01/Jan/2001:00:00:10 -0600] "GET
/article/916871128.html HTTP/1.0" 200 6921 "
http://healthlink.mcw.edu/breast-cancer/"
205.252.144.28 - - [01/Jan/2001:00:00:13 -0600] "GET
/content/topic/Arthritis HTTP/1.1" 200 14466 "
http://search.netscape.com/search.tmpl?cp=
srpnext10&search=arthritis&jstart=21"
202.167.121.197 - - [01/Jan/2001:00:00:38 -0600] "GET
/article/901225644.html HTTP/1.1" 200 6849 "
http://healthlink.mcw.edu/content/search.cgi?Castlema
n+disease"
63.97.116.101 - - [01/Jan/2001:00:00:49 -0600] "GET
/article/954382477.html HTTP/1.0" 200 11853 "-"
208.26.241.42 - - [01/Jan/2001:00:01:37 -0600] "GET
/article/908679020.html HTTP/1.0" 200 7566 "
http://healthlink.mcw.edu/breast-cancer/"
216.166.191.181 - - [01/Jan/2001:00:01:43 -0600] "GET
/article/946490368.html HTTP/1.1" 200 8418 "
http://search.metacrawler.com/crawler?general=
Diarrhea+Chronic&method=0&domainLimit=0&rpp=20&mrr=
0&timeout=0&hpe=10&format=regular&power=0&target=
&directhit_attrib=rs&refer=related"
216.166.191.181 - - [01/Jan/2001:00:01:44 -0600] "GET
/article/946490368.html HTTP/1.1" 200 8418 "
http://search.metacrawler.com/crawler?general=
Diarrhea+Chronic&method=0&domainLimit=0&rpp=20&mrr=
0&timeout=0&hpe=10&format=regular&power=0&target=
&directhit_attrib=rs&refer=related"
216.166.191.181 - - [01/Jan/2001:00:01:44 -0600] "GET
/article/946490368.html HTTP/1.1" 200 8418 "
http://search.metacrawler.com/crawler?general=
Diarrhea+Chronic&method=0&domainLimit=0&rpp=20&mrr=
0&timeout=0&hpe=10&format=regular&power=0&target=
&directhit_attrib=rs&refer=related"
```

Figure 1 - Sample of Healthlink Web Access log data

Time Series of Accesses of Influenza-related Articles

Identification of relevant Healthlink Web articles

Author HJ reviewed the titles of all 1504 articles that were available on Healthlink during 2001 and identified 21 articles that users with influenza might consult for information about their illness. From this set, we excluded four articles that were released in 2001 because Healthlink notifies users of the existence of new articles by e-mail and this marketing action produces a spike in accesses of those articles for several days after they are released. The result of this procedure was a set of 17 influenza-related articles that we also assigned into two groups – Diagnosis/Treatment (11 articles) or Prevention/Vaccination (6 articles) – based on their content.

Development of Time Series for each article

We developed a Perl script to identify accesses of the 17 influenza-related articles in the Web access logs. The script parsed the action field of each record to obtain the article number and then summed the accesses to produce daily access counts for each article. We then summed the daily access counts into weekly counts to match the weekly aggregation found in the CDC's influenza surveillance data.

Development of Article Total Time Series

We aggregated the weekly counts of all 17 articles to create the *Article Total* time series. In like manner, we created time series for the Diagnosis/Treatment and the Prevention/Vaccination groups of articles.

Cross Correlation Analysis

We used the cross-correlation function (CCF) to measure the correlation between article access counts and influenza activity [11]. The CCF not only measures the degree of correlation but also finds the time lag between two time series that maximizes the correlation. This time lag has been used by researchers to characterize the timeliness of one signal of disease activity relative to the other [7,11]. We performed the analysis for those time periods where the Healthlink data and influenza surveillance data overlapped:[1] namely, portions of the 2000-2001 and 2001-2002 influenza seasons. In this study, a negative timeliness result means that fluctuations in the Web access data precede fluctuations in the gold standard data; conversely, a positive result means the Web access data follow the gold standard data.

Results

In 2001, the 17 influenza-related articles were accessed 10,077 times, averaging 592 times per article (max=1063, min=263). Articles in the Diagnosis/Treatment and Prevention/Vaccination groups were accessed 5,265 and 4,812 times, respectively.

Many (13/17) of the articles appeared to have a seasonal temporal pattern similar to the one seen in the CDC influenza surveillance data. This trend is illustrated in the time series from the following articles in the Diagnosis/Treatment group, "Kicking the Flu" (Figure 2) and "Persistent Winter Cough" (Figure 3). Other articles, such as "Requirements for Back-to-School Immunizations" from the Prevention/Vaccination group, clearly had a temporal pattern, but it was different from the CDC data, as it began to increase in late summer prior to the start of school (Figure 4). Other articles (4/17) did not have a temporal pattern.

Article Accesses vs. Influenza Surveillance Data

The correlation between the Article Total time series and the ILI gold standard was 0.78 for weeks 1 – 20 of 2001 and 0.76 for weeks 40 – 52 (Figure 5). The timeliness of this time series was zero for both time periods, meaning neither time series (Article Total nor ILI gold standard) preceded the other. When compared to the positive cultures gold standard the correlation was 0.67 and 0.80. Although the Article Total time series lagged three weeks behind the positive cultures gold standard during weeks 1-20, it preceded that gold standard by two weeks during weeks 40-52.

The time series for the Diagnosis/Treatment group had a stronger correlation with both gold standards than the time series for the Prevention/Vaccination group (Table 1).

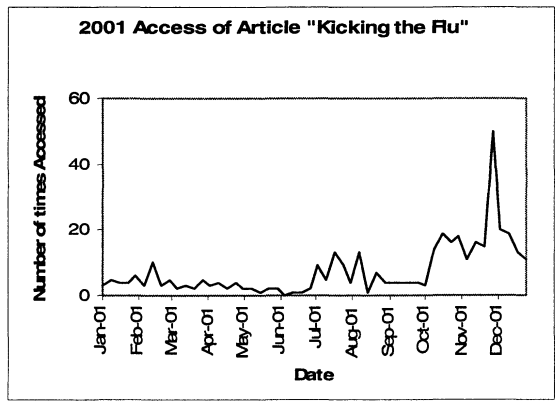

Figure 2 – 'Kicking the Flu' article accesses

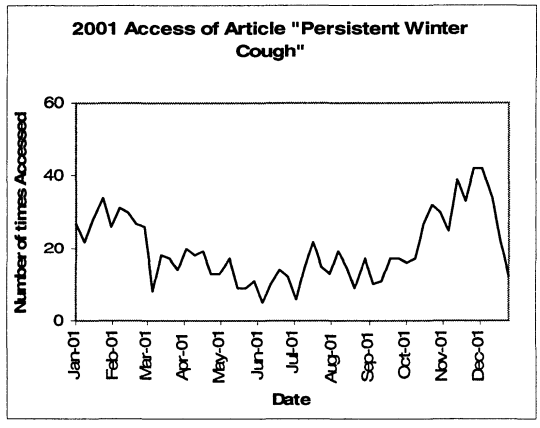

Figure 3 – 'Persistent Winter Cough' article accesses

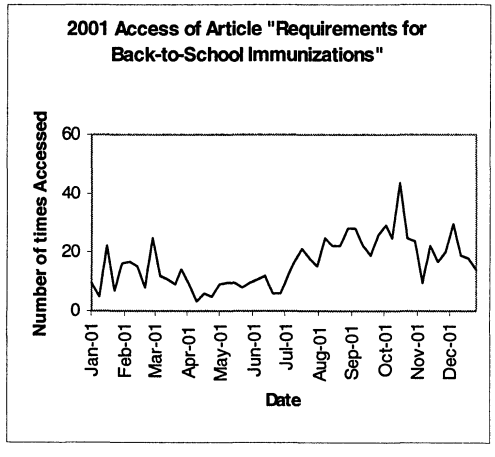

Figure 4 – 'Requirements for Back-to-School Immunizations' article accesses

1. The CDC does not provide weekly ILI and positive culture data for weeks 21 through 39 of the year.

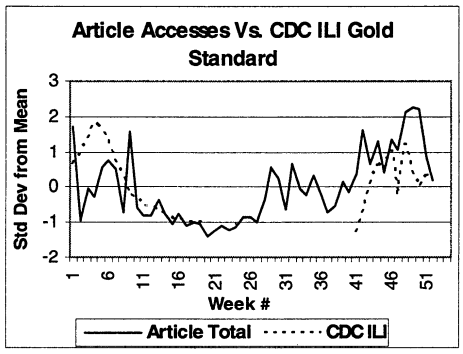

Figure 5. Total Article Accesses vs. ILI gold standard

When compared to the ILI gold standard, the Diagnosis/Treatment group had a correlation of 0.78 and 0.76, while the Prevention/Vaccination group had a correlation of 0.75 and 0.71 respectively. The correlation with the positive cultures gold standard was 0.79 and 0.80 for the Diagnosis/Treatment group and 0.75 and 0.72 for the Prevention/Vaccination group. Timeliness varied between both gold standards and the different time periods (Table 1) and did not show a consistent trend.

The article that had the best correlation during both time periods was "Influenza," for which the correlation with the positive cultures gold standard was 0.81 and 0.83, respectively. The timeliest article with a correlation of 0.70 or higher was "Influenza (Flu)," which had a correlation of 0.73 for both time periods. Despite having the same correlation in both time periods, "Influenza (Flu)" preceded the positive cultures gold standard by four weeks during the second time period (weeks 40-52), but lagged it by two weeks in the first time period.

Discussion

This first study of the correlation between Web access log data and influenza surveillance data found a moderately strong correlation between Web accesses and disease activity, suggesting that there is a signal of influenza in the Web access data. The results about timeliness were variable and we cannot draw strong conclusions from them. Nevertheless, the fact that signal can be discerned from analysis of just one Web site is encouraging. This study was limited to data from a single Web site for portions of two separate influenza seasons and should be repeated using multiple complete seasons of data and other Web access log collections such as PubMed, which differ along the dimensions of size, purpose, and popularity of the Web site from the Healthlink logs.

A key contribution of this study is that it reveals a number of issues that are particular to the use of Web data in biosurveillance that had not been identified by previous research on other types of routinely collected data.

A key contribution of this study is that it reveals a number of issues that are particular to the use of Web data in biosurveillance that had not been identified by previous research on other types of routinely collected data.

First, it is difficult to assign a spatial location to the Web-site user. In other surveillance data, such as retail data or clinical data, the address of the pharmacy or hospital—or home zip code of a patient—is available in association with the datum. For Web data, at best an IP address of the user is available (and at worst the only IP address available is that of a search engine). Without spatial information, the ability to detect local outbreaks is degraded. There are several methods available for addressing this problem of determining locations based on IP addresses in Web access logs. We used one of them in this study. Other methods include open source modules for searching public databases of IP addresses and proprietary products such as GeoLyzer from Geobytes [13], whose application searches their own proprietary database of IP addresses. Currently there is little data available about which of these methods is the most reliable. Therefore, identifying a reliable and feasible (both technically and legally) method for determining the location of users is an important research topic that should be pursued to enable the use of Web access data for public health surveillance.

Table 1: Cross-Correlation Function Results for Diagnosis/ Treatment and Prevention/Vaccination Articles

Diagnosis Articles			
Weeks	Gold Standard	Correlation	Timeliness
1-20	Influenza-like illnesses	0.7782	3
	Positive cultures	0.7853	2
40-52	Influenza-like illnesses	0.759	0
	Positive cultures	0.795	-2
Prevention Articles			
Weeks	Gold Standard	Correlation	Timeliness
1-20	Influenza-like illnesses	0.7476	0
	Positive cultures	0.7496	-1
40-52	Influenza-like illnesses	0.7101	2
	Positive cultures	0.7155	0

Second, accidental repeated access of the same article by the same user introduces additional false counts. Such a case might occur if a user repeatedly selects a link to an article when the page is slow to display in the browser. Each of these occurrences is stored as a separate request in the Web access log and may artificially inflate article access counts, with the potential to result subsequently in a false alarm. It may be possible to solve this problem by ignoring multiple requests from the same IP address with in a certain period of time (e.g., a minute or a hour).

Third, it is important to develop techniques to adjust for any marketing or promotional activities that would create a spike in the Web activity that we are monitoring that is not due to disease, but could trigger a detection algorithm. There were four articles in the selection of 21 that showed this effect and we exclude them from the study for this reason. Such spikes may appear when an article is first released, especially if it is promoted via the use of an e-mail newsletter.

A final issue was the format in which the data were delivered. The plain text, non-delimited log files were difficult to parse into intelligible fields without error. A standard format will be needed if this method is to be applied on a large scale involving multiple Web sites and search engines.

Future work will include reanalysis of the data when we obtain Web access logs for the 2002 calendar year and the planned analysis of the correlation of queries about symptoms related to respiratory illness with influenza activity. The study of the queries themselves will also be informative about whether the articles themselves or the queries are a stronger and earlier signal of influenza activity.

Conclusion

Web site usage may provide early indication about disease outbreaks both from the articles accessed by users and the search queries. This study of article access data suggests that a moderately strong signal of influenza activity can be found even in data from a single Web site. The study is inconclusive about whether the signal is earlier than conventional influenza surveillance data. With more sophisticated methods for performing spatial analysis on the data, the use of more complete Web access data from multiple Web sites, and continued increase in the use of the Web by the sick, detection through monitoring of Web article access may become a useful public health surveillance tool.

Acknowledgements

We thank the Office of Clinical Informatics at the Medical College of Wisconsin. This work was supported by Pennsylvania Department of Health Award number ME-01-737, Grant F 30602-01-2-0550 from the Defense Advanced Research Projects Agency, Contract 290-00-0009 from the Agency for Healthcare Research and Quality, and Grant T15 LM/DE07059 from the National Library of Medicine.

References

[1] Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA, McGinnis LF, Deerfield DW, Druzdzel MJ, Fridsma DB. The emerging science of very early detection of disease outbreaks. *J Public Health Manag Pract* 2001:7 (6):51-9.

[2] Wagner MM, Robinson JM, Tsui FC, Espino JU, Hogan WR. Design of a National Retail Data Monitor for public health surveillance. *J Am Med Inform Assoc* 2003:10:409-418.

[3] Hogan WR, Tsui F-C, Ivanov O, Gesteland P, Grannis S, Overhag JM, Robinson JM, Wagner MM. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. *J Am Med Inform Assoc* 2003:10:555-562.

[4] Espino JU, Wagner MM. Accuracy of ICD-9-coded chief complaints and diagnoses for the detection of acute respiratory illness. *Proc AMIA Symp* 2001:164-8.

[5] Ivanov O, Wagner MM, Chapman WW, Olszewski RT. Accuracy of three classifies of acute gastrointestinal syndrome for syndromic surveillance. *Proc AMIA Symp* 2002:345-9.

[6] Panackal AA, M'ikanatha NM, Tsui FC, McMahon J, Wagner MM, Dixon BW, Zubieta J, Phelan M, Mirza S, Morgan J, Jernigan D, Pasculle AW, Rankin JT Jr, Hajjeh RA, Harrison LH. Automatic electronic laboratory-based reporting of notifiable infectious diseases at a large health system. *Emerg Infect Dis* 2002:8 (7):685-91.

[7] Espino JU, Hogan WR, Wagner MM. Telephone Triage: A Timely Data Source for Surveillance of Influenza-like Diseases. *Proc AMIA Symp* 2003:215-9.

[8] Zeng X, Wagner MM. Modeling the Effects of Epidemics on Routinely Collected Data. *J Am Med Inform Assoc* 2002:9:S17-S22.

[9] Fox S, Fallows D. Pew Internet Health Resources. Available at: http://www.pewinternet.org/reports/toc.asp?Report=95. Accessed: July 21, 2003.

[10]Tu HT, Hargraves HL. Seeking Health Care Information: Most Consumers Still on the Sidelines. Available at: http://hschange.org/CONTENT/. Accessed: August 7, 2003.

[11]Tsui FC, Wagner MM, Dato V, Chang CC. Value of ICD-9 coded chief complaints for detection of epidemics. *J Am Med Inform Assoc* 2002:9:S41-S47.

[12]MaxMind™. Technical Details. Available at: http://www.maxmind.com/app/techinfo. Accessed: September 15, 2003.

[13]Geobytes. GeoLyzer Overview. Available at: http://www.geobytes.com/Solutions.htm#GeoLyzer. Accessed: February 18, 2004.

Address for correspondence

Correspondence and reprints: Heather A. Johnson, The RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, Suite 550, 100 Technology Drive, Pittsburgh, PA 15219; e-mail: hjohnson@cbmi.pitt.edu.