Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance

Gunther Eysenbach MD MPH [1) 2)]
1) Centre for Global eHealth Innovation, University Health Network, Toronto M5G2C4, Canada, and 2) Department of Health Policy, Management and Evaluation, University of Toronto, Canada.
Email: geysenba@uhnres.utoronto.ca

Abstract

Background: Syndromic surveillance uses health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response.

Objective: While most syndromic surveillance systems rely on data from clinical encounters with health professionals, I started to explore in 2004 whether analysis of trends in Internet searches can be useful to predict outbreaks such as influenza epidemics and prospectively gathered data on Internet search trends for this purpose.

Results: There is an excellent correlation between the number of clicks on a keyword-triggered link in Google with epidemiological data from the flu season 2004/2005 in Canada (Pearson correlation coefficient of current week clicks with the following week influenza cases *r*=.91). The "Google ad sentinel method" proved to be more timely, more accurate and – with a total cost of Can$365.64 for the entire flu-season – considerably cheaper than the traditional method of reports on influenza-like illnesses observed in clinics by sentinel physicians.

Conclusion: Systematically collecting and analyzing health information demand data from the Internet has considerable potential to be used for syndromic surveillance. Tracking web searches on the Internet has the potential to predict population-based events relevant for public health purposes, such as real outbreaks, but may also be confounded by "epidemics of fear". Data from such "infodemiology studies" should also include longitudinal data on health information supply.

> "A new research discipline and methodology has emerged—the study of the determinants and distribution of health information (…): Information epidemiology, or infodemiology [1]

Introduction

An increasing proportion of people in industrialized countries is using the Internet to seek health information [2;3]. An interesting question is whether tracking health information seeking behaviour of populations over time can be used for public health purposes, particularly syndromic surveillance. The CDC defines syndromic surveillance as "surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response." While most syndromic surveillance systems rely on data from clinical encounters with health professionals, monitoring for example sick-leave prescriptions, house calls, hospital- or pharmacy-based data [4;5], there have also been previous experiments with unconventional methods to use preclinical "health information seeking" data for syndromic surveillance, for example monitoring calls to a "NurseLine" such as NHS Direct [6-8]. However, there does not seem to be any prior evaluation of the use of Internet search data for syndromic surveillance, and most evaluations of surveillance systems for detecting bioterrorism and emerging infections have been described as "insufficient to characterize the timeliness or sensitivity and specificity" [4]

I explored whether an automated analysis of trends in Internet searches could be useful to predict outbreaks such as influenza epidemics. To do so, I first had to develop a method for gathering data from Google. I then developed a model for predicting an influenza outbreak on the basis of changes in Internet search activity for flu-related information. The model was evaluated against a traditional surveillance method which utilizes "sentinel physicians", who manually report encounters with sick patients to a public health agency.

Methods

I aimed to correlate data from the Canadian flu season 2004/2005 over a period of 33 weeks from week 41/2004 (Oct 3-9) to week 20/2005 (May 15-21) with Internet data for flu-related searches.

One methodological problem lies in the difficulties to obtain unbiased search data, in particular as a major search engine such as Google is reluctant to share search data, even when governments demand to see such data. This was illustrated in a recent case when Google refused to comply with a subpoena issued in 2005, which included a request for 1 million random Web addresses and records of all Google searches from any one-week period, which the U.S. government contended it needed to determine how often pornography shows up in online searches [9].

One possible method to gather search data has been described previously[10]: it consists of "screenscraping" of publicly accessible searches on the MetaSpy website, which displays - in real time - search terms currently entered by people into the Metacrawler search engine. However, this method has a considerable drawback for the context of syndromic surveillance, which is that the origin of the searchers remains unknown.

I therefore developed another method to gauge the prevalence of certain search terms over time. The "trick" used here is that – while Google normally does not provide detailed log files for "all searches" conducted on its sites – it does provide rather detailed statistics for advertisers who "buy" (or rather bid for) certain keywords in the context of its keyword-triggered advertising program Google Adsense. (Note that keywords do not have fixed prices, rather advertisers bid for certain keywords, with advertisements with the highest combination of click-through-rate and bid appearing more often and ranking higher.) One advantage of this method is also that advertisements can be restricted to certain countries or even US states, thus – while the exact location of the searcher remains unknown – an analysis can be restricted to a certain geographic area.

To obtain statistics on the prevalence of searches on a certain topic I therefore created a "campaign" using a keyword-triggered "sponsored link" in Google Adsense, which appeared for Canadian searchers only, who entered "flu" or "flu symptoms" into Google. The ad read "Do you have the flu? Fever, Chest discomfort, Weakness, Aches, Headache, Cough." and contained a link to a generic patient education website.

Several metrics are provided by Google Adsense for each "campaign" (a campaign consists of one or several keywords triggering one or several different sponsored links): Number of ad views (or "impressions") (as each keyword match triggers an advertisement, the number of impressions should be roughly proportional to the number of searches containing the keyword), as well as clicks on the advertisement and click-through-rate (impressions divided by clicks).

I aggregated daily statistics on impressions and clicks provided by Google to match the time periods of the weekly national FluWatch reports.

FluWatch reports are published by the Public Health Agency Canada and provide the traditional disease surveillance metrics, including the number of influenza lab tests for Influenza A or B conducted in sentinel laboratories ("lab tests"), the number of lab tests testing positive for influenza A or B ("cases"), and the number of cases of influenza like illness reported by sentinel physicians ("ILI-SPR"), which were used as the "gold standard" against which the keyword-method was evaluated.

ILI-SPR data are collected by the Public Health Agency of Canada from about 200 participating sentinel physicians who record for 1-clinic day each week the number of patients presenting with influenza-like illness (ILI). ILI is defined as a patient with acute onset of respiratory illness with fever and cough and one or more of: sore throat, arthralgia, myalgia or prostration.

I correlated ILI-SPR – the traditional measure used for flu-surveillance – as well as different novel measures obtained from the Google campaign (clicks, ad views, click-through-rate) with the number of lab tests and cases indicating the incidence of influenza to compare the predictive value of these measures. I also experimented with different multivariate models to predict flu-cases with Google ad measures, adjusting for other factors such as the position of the ad within Google.

All analyses were conducted using SAS Release 8.02 (SAS Institute, Cary, NC, USA).

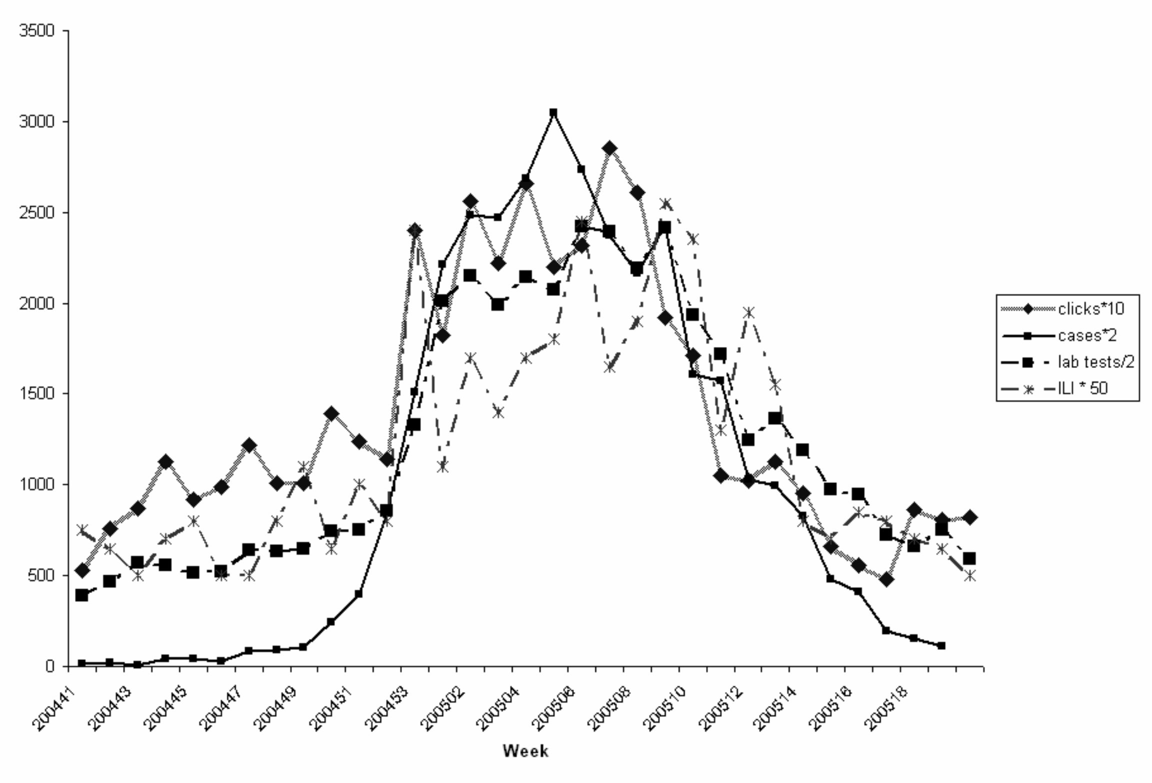

Figure 1. Normalized data from Fluwatch (influenza cases, lab tests, ILI reports from sentinel physicians) and Google (number of clicks on an keyword-triggered influenza link).

Results

Over the flu-season period, the Google campaign received a total of 54,507 impressions and 4,582 clicks (Figure 1). Among all the ad campaign measures, the number of clicks on the ad was found to have the best correlation with traditional surveillance measures, which is why I show only correlation data for clicks.

In general, clicks correlated better with flu events than ILI reports from sentinel physicians (Table 1). Internet clicks also were a *timelier* marker than ILI-SPR, in that they performed better to predict the flu events of the *following* week, whereas correlation coefficients in the ILI-SPR method were better for the *current* week than for the following week. All correlations were significant on a *P*<.001 level.

Trivariate linear regression analysis adjusting for the ad position within Google did not improve the fit substantially, as most ads appeared close to the top anyway (data not shown).

Would a threshold of 150 clicks per week have been used to trigger a flu-outbreak alert, all 11 weeks with 524 flu-cases or more following the query sampling week could be predicted with 100% specificity and sensitivity.

The costs of the Google sentinel method were negligible compared to traditional methods: Google charges $0.08 per click-through, thus the campaign cost only Can$365.64 for the entire flu-season.

Table 1. Pearson correlation coefficients of ad clicks and influenza like illness reports from sentinel physicians (ILI-SPR) as measures for predicting influenza incidence data from the current or following week (all *P*<.001). (* see Figure 2, ** see Figure 3)

	Clicks	*ILI-SPR*
Same week		
ILI-SPR	.73	—
Lab tests	.85	.83
Cases	.88	.80
Following week		
ILI-SPR	.81	.71
Lab tests	.90	.82
Cases	.91 *	.75 **

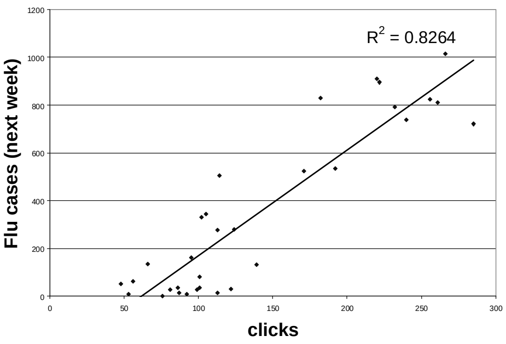

Figure 2 (above). Scatter plot (with regression line) showing the excellent correlation between "clicks" and flu cases in the following week (*)

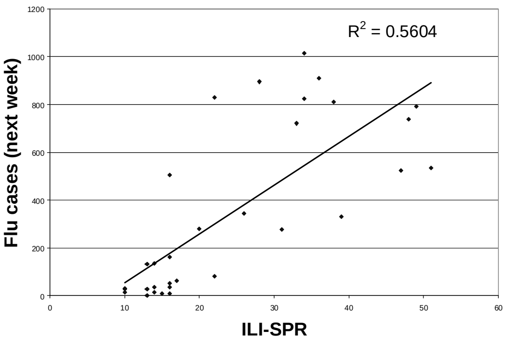

Figure 3 (above). Scatter plot (with regression line) showing the considerably worse correlation between sentinel physicians' ILI reports and flu cases in the following week (**)

Discussion

Tracking demand for health information on the Internet using keyword-triggered ads is a novel and promising method for public health surveillance. While further studies are needed to explore the accuracy and use of these "infodemiology" methods, in this analysis search engine clicks were surprisingly a *better* and *timelier* predictor for flu-cases than ILI reported by sentinel physicians, suggesting that for conditions where consumers consult the Internet first before they visit a physician, systematic text mining of web-searches may be a valuable addition complementing traditional surveillance approaches.

Future systems can be refined by adjusting the results for baseline searches. For example, there was a decrease in searches and clicks just around Christmas, partly reflecting a general decrease in Internet use during the holidays (week 52/2004), despite an increase of influenza cases in that week.

Identifying the geographic origin of searches is possible, as Google Adsense allows targeting ads down to the country or US state level; further analysis on the origin of searchers could be done analyzing IP addresses.

A recent CDC study conducted with Yahoo! suggested that Internet searches for specific cancers correlated with their estimated incidence (Spearman rank correlation, $\rho = 0.50$, $P = .015$), estimated mortality ($\rho = 0.66$, $P = .001$), and volume of related news coverage ($\rho = 0.88$, $P < .001$).[11] The authors concluded that "media coverage appears to play a powerful role in prompting online searches for cancer information".

Thus, as has been noted previously elsewhere, like other early warning systems which are based on consumer behaviour, search data will be confounded by media reports and "epidemics of fear" [12]. On the other hand, even crude (unadjusted) surges in increased search activity on a health topic not triggered by a real pandemic are still important measures for policy makers as they may (even in the absence of a "true" epidemic) nevertheless warrant a public health response to satisfy the information demand, thus even alerts triggered by infodemiology systems in the absence of an outbreak are not necessarily to be considered "false positives".

In order to reduce the confounding effect of media reports one could attempt to control for these. In our ongoing research on "infodemiology" measures we aim to include not only the demand side (queries) into our models but also the supply side (information on news websites), to explore whether adjustment for the frequency of disease keywords in media reports improves the predictive value of infodemiology methods. A system such as the Canadian Global Public Intelligence Network [13;14], run by the Public Health Agency of Canada, is already conducting a similar function by monitoring global media sources such as news wires and web sites. The system was credited with an early detection of the SARS outbreak[12], although no formal evaluation seems to have been published in the peer-reviewed literature.

Conclusion

"Information epidemiology", or "infodemiology"[1] is broadly speaking an emerging set of methods which studies the determinants and distribution of health information for public health purposes. This study illustrates that the development of "infodemiology" metrics based on automated tracking and analysis of the distribution and determinants of health information (both supply and need) in a population and/or information space is possible and can provide important clues and evidence for public health policy and practice. In a broader sense, an "infodemiology" science is needed to develop a methodology and real-time measures (indices) to understand patterns and trends for general health information, e.g. to understand "(mis)information" outbreaks, to study and quantify knowledge translation gaps (for example, the dissemination of a new treatment), and to understand the predictive value of what people are looking for (demand) for syndromic surveillance and early detection of emerging diseases.

The Internet has made measurable what was previously immeasurable: The distribution of health information in a population, tracking health information trends over time, and identifying gaps between information supply and demand. "Infodemiology" research may lead to reliable and meaningful indicators to track health information demand and supply trends, will foster our understanding on how to maximize the use of the Internet to improve public health, and may provide the possibility to use some of these metrics as early warning systems for infectious disease outbreaks, bioterrorism, or emerging diseases.

References

1. Eysenbach G. Infodemiology: The Epidemiology of (Mis)information. *Am J Med* 2002;**113**:763-5.
2. Baker L, Wagner TH, Singer S, Bundorf MK. Use of the Internet and e-mail for health care information: results from a national survey. *JAMA* 2003;**289**:2400-6.
3. Eysenbach G,.Köhler C. Health-Related Searches on the Internet. *JAMA* 2004;**291**:2946.
4. Bravata DM, McDonald KM, Smith WM, Rydzak C, Szeto H, Buckeridge DL *et al.* Systematic Review: Surveillance Systems for Early Detection of Bioterrorism-Related Diseases. *Ann.Intern.Med.* 2004;**140**:910-22.
5. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F *et al.* Implementing syndromic surveillance: a practical guide informed by the early experience. *J Am Med Inform Assoc* 2004;**11**:141-50.
6. Cooper DL, Smith G, Baker M, Chinemana F, Verlander N, Gerard E *et al*. National symptom surveillance using calls to a telephone health advice service--United Kingdom, December 2001-February 2003. *MMWR Morb.Mortal.Wkly.Rep.* 2004;**53 Suppl**:179-83.
7. Cooper DL, Smith GE, O'Brien SJ, Hollyoak VA, Baker M. What can analysis of calls to NHS direct tell us about the epidemiology of gastrointestinal infections in the community? *J Infect.* 2003;**46**:101-5.
8. Cooper DL, Smith GE, Hollyoak VA, Joseph CA, Johnson L, Chaloner R. Use of NHS Direct calls for surveillance of influenza--a second year's experience. *Commun.Dis.Public Health* 2002;**5**:127-31.
9. Gonzales v. Google, Inc.: Motion to Compel. http://www.webcitation.org/query?id=1142396222389927 . 2006.
10. Eysenbach G,.Kohler C. What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the Internet. *Proc AMIA Annu Fall Symp* 2003;225-9.
11. 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.
12. Eysenbach G. SARS and population health technology. *J Med Internet Res* 2003;**5**:e14.
13. 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.
14. Anonymous. Global Public Health Intelligence Network (GPHIN). http://www.webcitation.org/query?id=1142400987000608 . 1-11-2004. 15-3-2006.