doi: 10.1038/nature07634

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]*Google Inc.*

[2]*Centers for Disease Control and Prevention*

Supplementary Figures and Legends

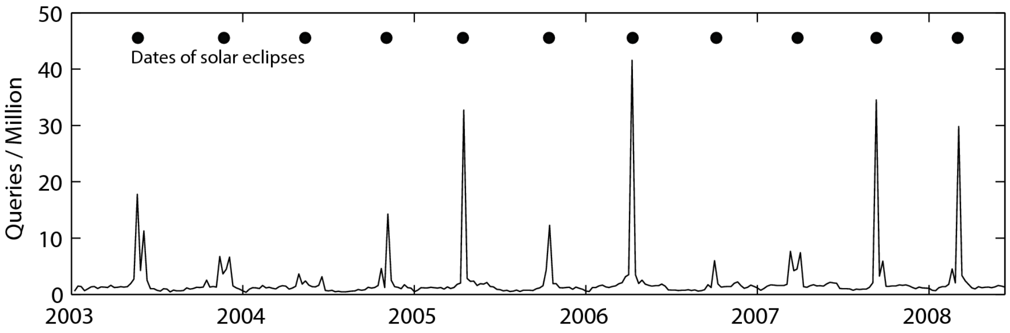

Figure 1: Weekly frequency of the search query “solar eclipse” in the United States from January 2003 to May 2008 and occurrences of solar eclipses, indicated by black dots.

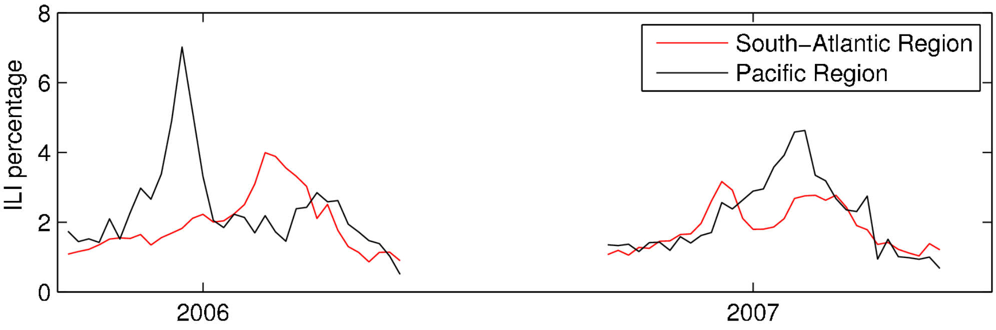

Figure 2: Regional variations in weekly ILI percentages for the South-Atlantic and Pacific Regions (source: CDC Influenza Sentinel Provider Network).

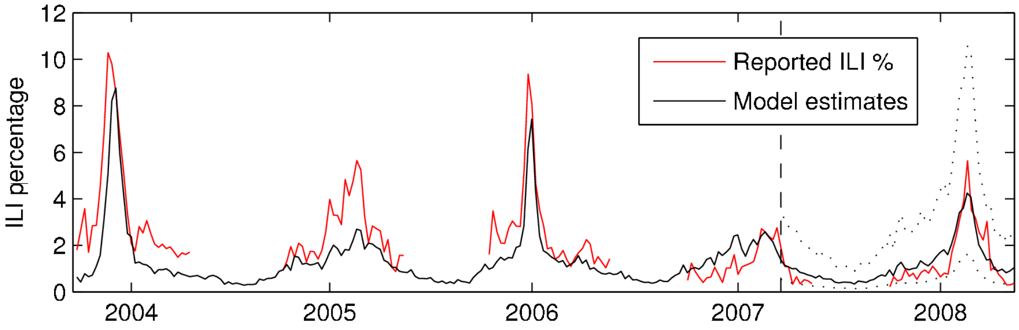

Figure 3: Weekly ILI estimates for Utah, 2003-2007. Estimates are generated using search query data from Utah as input to our final model, trained on regional data. Across 42 validation points, model estimates obtained a correlation of 0.90 against the state-reported ILI percentages. 95% prediction intervals are indicated.

Supplementary Methods

An example query-fraction time series. Supplementary Figure 1 shows one example of a query fraction time series, for the search query “solar eclipse” in the United States. Note that a spike in query volume coincided with each occurrence of a solar eclipse.

There appears to be a moderate (though somewhat inconsistent) relationship between the magnitude of the query volume spike and the visibility of the eclipse from North America. Millions of search query time series can be viewed using Google Trends, a free tool available at http://www.google.com/trends .

Number of variables used in the model. We note that each search query selected as being ILI-related doesn't add another variable to the model but increases the number of queries considered when calculating the overall fraction of ILI-related queries, which is the single variable in the model. With only a single variable, including more search queries does not increase our risk of overfitting, but rather allows us to effectively survey more users; indeed, we hope to later develop models which include many thousands of related search query terms (especially those which are rare queries, misspellings, etc.) without sacrificing the quality of our fit.

Because each of the 45 selected search queries are highly correlated with CDC-provided ILI data, the queries are also highly correlated with each other. So much so that we decided to lump the counts for these queries together as a single variable. We could, in later work, attempt to build a multivariate model which combines the selected queries from our current model with other, more orthogonal queries; however, our preliminary efforts in this direction have not yet proved to be effective. We attempted a very large stepwise forward selection approach to model building, in which we considered different queries as different variables in the model, but found that many of the queries that predict well are very highly correlated. Consequently, we ran into nonsingularity issues when estimating such models. We found that simply summing these highly correlated queries into a single variable led to a model that not only fit well but also was very simple. Although we initially believed that we would find orthogonal sets of queries that would contribute as distinct independent variables in our model, this ended up not being the case.

Region-specific coefficients. Earlier versions of our model permitted separate multiplicative coefficients for each surveillance region of the United States, rather than a single coefficient used across all regions. These models were able to obtain a similarly good fit with CDC-reported ILI percentages, with a mean correlation of 0.90 (min=0.79, max=0.95, n=9 regions). Estimates generated by the per-region models for 42 validation points obtained a mean correlation of 0.97 (min=0.91, max=0.98, n=9 regions) with the CDC-observed ILI percentages. Allowing region-specific coefficients does not improve the performance of our model.

U.S. Influenza Sentinel Provider Surveillance Network. The CDC's U.S. Influenza Sentinel Provider Surveillance Network is a network of sentinel physicians which reports the fraction of patients presenting with an influenza-like illness (ILI). Each year, more than 16 million physician visits are surveyed, with 1500 physicians reporting regularly. Because many different pathogens can cause influenza-like symptoms, and since most Google search users have no idea which pathogen may be causing their symptoms, we felt that ILI data was a natural fit for the patterns one could reasonably expect to find in search query data. In future work, we hope to examine whether data from virologic surveillance networks, which perform virologic testing by counting and classifying influenza viruses collected from patients, can also be modelled using search query data.

Supplementary Discussion

By way of comparison, internet-based surveillance systems such as GPHIN[14] and HealthMap[15] harvest newspaper articles and other web pages to detect disease outbreaks. Such systems track cases of H5N1 as they emerge around the world, as even a single case of H5N1 can attract media coverage. Our approach cannot be used to detect small numbers of influenza cases, but can detect and quantify ILI activity without any news articles being published.

We intend to update our model each year with the latest sentinel provider ILI data, obtaining a better fit and adjusting as online health-seeking behaviour evolves over time.

A number of influenza-related newspaper headline events have occurred in recent years. We were originally concerned that widespread media coverage of avian influenza cases would hurt our correlations and cause unusual spikes in our estimates. While undoubtedly some search queries were affected by such events, the queries used in our model have proven resilient to news-induced spikes, and we haven't seen major outliers in our estimates as a result. The queries which are most susceptible to news-induced spikes, by design, are not selected using our approach.

While we would like to present the full list of search queries which we found to be ILI-related, we feel that presenting this information to a wide audience could make these queries less useful for influenza surveillance. Upon hearing that Google is using specific queries for influenza surveillance, users may be inclined to submit some of the queries out of curiosity, leading to erroneous future estimates of the ILI percentage.

Media coverage about our system may noticeably change the health-seeking behaviour of Google search users, even without specific knowledge of which search queries are being used. It is difficult to predict the extent to which this might occur.

We hope to extend this system to enhance global influenza surveillance, especially in areas which currently lack the necessary resources, including laboratory diagnostic capacity. Though it may be possible for this approach to be applied to any country with a large population of web search users, we cannot currently provide accurate estimates for large parts of the developing world. Even within the developed world, small countries and less common languages may be challenging to accurately survey.

Our system has limited utility in areas which currently lack widespread internet access, and we'll be most successful in countries where a large number of users regularly perform online searches. It seems possible that the rapid spread of information technology may reach certain areas before the necessary infrastructure for traditional surveillance is developed, in which case our system could provide preliminary surveillance. As we expand into other countries, we would certainly expect to find different patterns in internet usage and search habits, and would therefore benefit from finding new search query terms in each language and country which most accurately correlate with ILI or other influenza datasets.

This system may be capable of providing ILI estimates for large cities and metropolitan areas with high internet penetration, providing even more local influenza surveillance. We hope to explore this topic as well.

This approach may not easily extend to any other communicable diseases. In the developed world, a patient experiencing obviously severe or alarming symptoms may be unlikely to consult a search engine, especially if a physician or emergency room is nearby. Millions suffer from influenza each year, while most disease outbreaks tend to involve significantly fewer cases and therefore may be impossible to detect in a large population of search engine users. Our attempts to reliably detect smaller outbreaks of other diseases (including enterics and arboviruses) using search queries have not yet succeeded, and may benefit from including a broader range of search queries through the use of synonyms, translations, and related terms. We hope to explore whether endemic diseases, including breast cancer and sexually transmitted diseases, can be surveyed using our system.

Supplementary Notes

15. Mawudeku, A. & Blench, M. Global public health intelligence network (GPHIN). *7th Conference of the Association for Machine Translation in the Americas* (2006).

16. Brownstein, J. S., Freifeld, C. C., Reis, B. Y. & Mandl, K. D. Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Medicine* **5**, 1019–1024 (2008).