

# Detecting influenza epidemics using search engine query data

Jeremy Ginsberg<sup>1</sup>, Matthew H. Mohebbi<sup>1</sup>, Rajan S. Patel<sup>1</sup>, Lynnette Brammer<sup>2</sup>, Mark S. Smolinski<sup>1</sup> & Larry Brilliant<sup>1</sup>

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year<sup>1</sup>. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities<sup>2</sup>. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza<sup>3,4</sup>. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

Traditional surveillance systems, including those used by the US Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS), rely on both virological and clinical data, including influenza-like illness (ILI) physician visits. The CDC publishes national and regional data from these surveillance systems on a weekly basis, typically with a 1–2-week reporting lag.

In an attempt to provide faster detection, innovative surveillance systems have been created to monitor indirect signals of influenza activity, such as call volume to telephone triage advice lines<sup>5</sup> and over-the-counter drug sales<sup>6</sup>. About 90 million American adults are believed to search online for information about specific diseases or medical problems each year<sup>7</sup>, making web search queries a uniquely valuable source of information about health trends. Previous attempts at using online activity for influenza surveillance have counted search queries submitted to a Swedish medical website (A. Hulth, G. Rydevik and A. Linde, manuscript in preparation), visitors to certain pages on a US health website<sup>8</sup>, and user clicks on a search keyword advertisement in Canada<sup>9</sup>. A set of Yahoo search queries containing the words 'flu' or 'influenza' were found to correlate with virological and mortality surveillance data over multiple years<sup>10</sup>.

Our proposed system builds on this earlier work by using an automated method of discovering influenza-related search queries. By processing hundreds of billions of individual searches from 5 years of Google web search logs, our system generates more comprehensive models for use in influenza surveillance, with regional and state-level estimates of ILI activity in the United States. Widespread global usage of online search engines may eventually enable models to be developed in international settings.

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

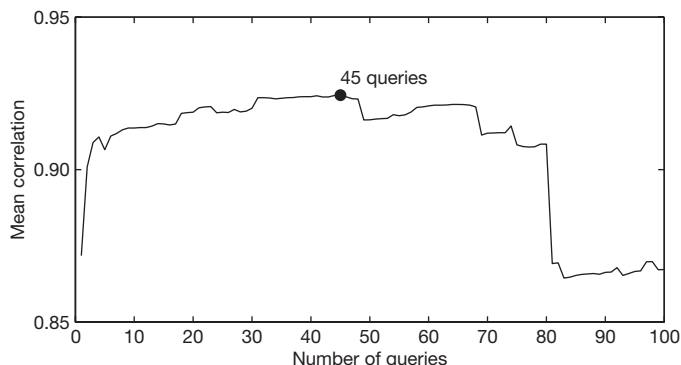
We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query:  $\text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \varepsilon$ , where  $I(t)$  is the percentage of ILI physician visits,  $Q(t)$  is the ILI-related query fraction at time  $t$ ,  $\alpha$  is the multiplicative coefficient, and  $\varepsilon$  is the error term.  $\text{logit}(p)$  is simply  $\ln(p/(1-p))$ .

Publicly available historical data from the CDC's US Influenza Sentinel Provider Surveillance Network (<http://www.cdc.gov/flu/weekly>) was used to help build our models. For each of the nine surveillance regions of the United States, the CDC reported the average percentage of all outpatient visits to sentinel providers that were ILI-related on a weekly basis. No data were provided for weeks outside of the annual influenza season, and we excluded such dates from model fitting, although our model was used to generate unvalidated ILI estimates for these weeks.

We designed an automated method of selecting ILI-related search queries, requiring no previous knowledge about influenza. We measured how effectively our model would fit the CDC ILI data in each region if we used only a single query as the explanatory variable,  $Q(t)$ . Each of the 50 million candidate queries in our database was separately tested in this manner, to identify the search queries which could most accurately model the CDC ILI visit percentage in each region. Our approach rewarded queries that showed regional variations similar to the regional variations in CDC ILI data: the chance that a random search query can fit the ILI percentage in all nine regions is considerably less than the chance that a random search query can fit a single location (Supplementary Fig. 2).

The automated query selection process produced a list of the highest scoring search queries, sorted by mean  $Z$ -transformed correlation across the nine regions. To decide which queries would be included in the ILI-related query fraction,  $Q(t)$ , we considered different sets of  $n$  top-scoring queries. We measured the performance of these models based on the sum of the queries in each set, and picked  $n$  such that we obtained the best fit against out-of-sample ILI data across the nine regions (Fig. 1).

<sup>1</sup>Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA. <sup>2</sup>Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA.



**Figure 1 | An evaluation of how many top-scoring queries to include in the ILI-related query fraction.** Maximal performance at estimating out-of-sample points during cross-validation was obtained by summing the top 45 search queries. A steep drop in model performance occurs after adding query 81, which is 'oscar nominations'.

Combining the  $n = 45$  highest-scoring queries was found to obtain the best fit. These 45 search queries, although selected automatically, appeared to be consistently related to ILIs. Other search queries in the top 100, not included in our model, included topics like 'high school basketball', which tend to coincide with influenza season in the United States (Table 1).

Using this ILI-related query fraction as the explanatory variable, we fit a final linear model to weekly ILI percentages between 2003 and 2007 for all nine regions together, thus obtaining a single, region-independent coefficient. The model was able to obtain a good fit with CDC-reported ILI percentages, with a mean correlation of 0.90 (min = 0.80, max = 0.96,  $n = 9$  regions; Fig. 2).

The final model was validated on 42 points per region of previously untested data from 2007 to 2008, which were excluded from all previous steps. Estimates generated for these 42 points obtained a mean correlation of 0.97 (min = 0.92, max = 0.99,  $n = 9$  regions) with the CDC-observed ILI percentages.

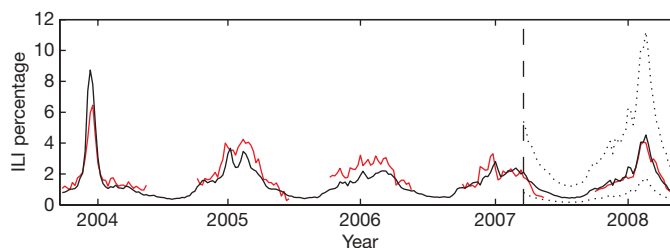
Throughout the 2007–08 influenza season we used preliminary versions of our model to generate ILI estimates, and shared our results each week with the Epidemiology and Prevention Branch of Influenza Division at the CDC to evaluate timeliness and accuracy. Figure 3 illustrates data available at different points throughout the season. Across the nine regions, we were able to estimate consistently the current ILI percentage 1–2 weeks ahead of the publication of reports by the CDC's US Influenza Sentinel Provider Surveillance Network.

Because localized influenza surveillance is particularly useful for public health planning, we sought to validate further our model

**Table 1 | Topics found in search queries which were found to be most correlated with CDC ILI data**

Search query topic	Top 45 queries		Next 55 queries	
	<i>n</i>	Weighted	<i>n</i>	Weighted
Influenza complication	11	18.15	5	3.40
Cold/flu remedy	8	5.05	6	5.03
General influenza symptoms	5	2.60	1	0.07
Term for influenza	4	3.74	6	0.30
Specific influenza symptom	4	2.54	6	3.74
Symptoms of an influenza complication	4	2.21	2	0.92
Antibiotic medication	3	6.23	3	3.17
General influenza remedies	2	0.18	1	0.32
Symptoms of a related disease	2	1.66	2	0.77
Antiviral medication	1	0.39	1	0.74
Related disease	1	6.66	3	3.77
Unrelated to influenza	0	0.00	19	28.37
Total	45	49.40	55	50.60

The top 45 queries were used in our final model; the next 55 queries are presented for comparison purposes. The number of queries in each topic is indicated, as well as query-volume-weighted counts, reflecting the relative frequency of queries in each topic.



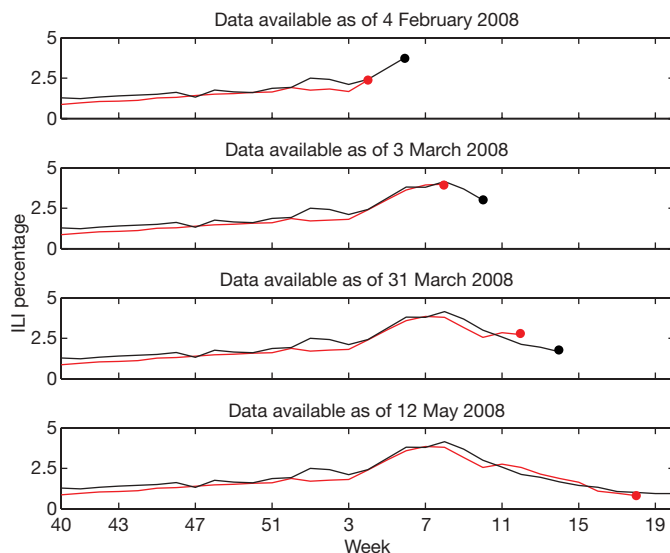
**Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated.** A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

against weekly ILI percentages for individual states. The CDC does not make state-level data publicly available, but we validated our model against state-reported ILI percentages provided by the state of Utah, and obtained a correlation of 0.90 across 42 validation points (Supplementary Fig. 3).

Google web search queries can be used to estimate ILI percentages accurately in each of the nine public health regions of the United States. Because search queries can be processed quickly, the resulting ILI estimates were consistently 1–2 weeks ahead of CDC ILI surveillance reports. The early detection provided by this approach may become an important line of defence against future influenza epidemics in the United States, and perhaps eventually in international settings.

Up-to-date influenza estimates may enable public health officials and health professionals to respond better to seasonal epidemics. If a region experiences an early, sharp increase in ILI physician visits, it may be possible to focus additional resources on that region to identify the aetiology of the outbreak, providing extra vaccine capacity or raising local media awareness as necessary.

This system is not designed to be a replacement for traditional surveillance networks or supplant the need for laboratory-based diagnoses and surveillance. Notable increases in ILI-related search activity



**Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007–2008 influenza season.** During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3 March our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by CDC ILI data.

may indicate a need for public health inquiry to identify the pathogen or pathogens involved. Demographic data, often provided by traditional surveillance, cannot be obtained using search queries.

In the event that a pandemic-causing strain of influenza emerges, accurate and early detection of ILI percentages may enable public health officials to mount a more effective early response. Although we cannot be certain how search engine users will behave in such a scenario, affected individuals may submit the same ILI-related search queries used in our model. Alternatively, panic and concern among healthy individuals may cause a surge in the ILI-related query fraction and exaggerated estimates of the ongoing ILI percentage.

The search queries in our model are not, of course, exclusively submitted by users who are experiencing influenza-like symptoms, and the correlations we observe are only meaningful across large populations. Despite strong historical correlations, our system remains susceptible to false alerts caused by a sudden increase in ILI-related queries. An unusual event, such as a drug recall for a popular cold or flu remedy, could cause such a false alert.

Harnessing the collective intelligence of millions of users, Google web search logs can provide one of the most timely, broad-reaching influenza monitoring systems available today. Whereas traditional systems require 1–2 weeks to gather and process surveillance data, our estimates are current each day. As with other syndromic surveillance systems, the data are most useful as a means to spur further investigation and collection of direct measures of disease activity. This system will be used to track the spread of ILI throughout the 2008–09 influenza season in the United States. Results are freely available online at <http://www.google.org/flutrends>.

## METHODS SUMMARY

**Privacy.** None of the queries in the Google database for this project can be associated with a particular individual. The database retains no information about the identity, internet protocol (IP) address, or specific physical location of any user. Furthermore, any original web search logs older than 9 months are being made anonymous in accordance with Google's privacy policy (<http://www.google.com/privacypolicy.html>).

**Search query database.** For the purposes of our database, a search query is a complete, exact sequence of terms issued by a Google search user; we don't combine linguistic variations, synonyms, cross-language translations, misspellings, or subsequences, although we hope to explore these options in future work. For example, we tallied the search query 'indications of flu' separately from the search queries 'flu indications' and 'indications of the flu'.

Our database of queries contains 50 million of the most common search queries on all possible topics, without pre-filtering. Billions of queries occurred infrequently and were excluded. Using the internet protocol address associated with each search query, the general physical location from which the query

originated can often be identified, including the nearest major city if within the United States.

**Model data.** In the query selection process, we fit per-query models using all weeks between 28 September 2003 and 11 March 2007 (inclusive) for which the CDC reported a non-zero ILI percentage, yielding 128 training points for each region (each week is one data point). Forty-two additional weeks of data (18 March 2007 through to 11 May 2008) were reserved for final validation. Search query data before 2003 was not available for this project.

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Received 14 August; accepted 13 November 2008.**

**Published online 19 November 2008.**

1. World Health Organization. Influenza fact sheet. (<http://www.who.int/mediacentre/factsheets/fs211/en/>) (2003).
2. World Health Organization. WHO consultation on priority public health interventions before and during an influenza pandemic. ([http://www.who.int/csr/disease/avian\\_influenza/consultation/en/](http://www.who.int/csr/disease/avian_influenza/consultation/en/)) (2004).
3. Ferguson, N. M. *et al.* Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **437**, 209–214 (2005).
4. Longini, I. M. *et al.* Containing pandemic influenza at the source. *Science* **309**, 1083–1087 (2005).
5. Espino, J., Hogan, W. & Wagner, M. Telephone triage: A timely data source for surveillance of influenza-like diseases. *AMIA Annu. Symp. Proc.* 215–219 (2003).
6. Magruder, S. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL Tech. Digest* **24**, 349–353 (2003).
7. Fox, S. Online Health Search 2006. *Pew Internet & American Life Project* ([http://www.pewinternet.org/pdfs/pip\\_online\\_health\\_2006.pdf](http://www.pewinternet.org/pdfs/pip_online_health_2006.pdf)) (2006).
8. Johnson, H. *et al.* Analysis of Web access logs for surveillance of influenza. *Stud. Health Technol. Inform.* **107**, 1202–1206 (2004).
9. Eysenbach, G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA Symp. Proc.* 244–248 (2006).
10. Polgreen, P. M., Chen, Y., Pennock, D. M. & Forrest, N. D. Using internet searches for influenza surveillance. *Clin. Infect. Dis.* **47**, 1443–1448 (2008).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank L. Finelli for providing background knowledge, helping us validate results and comments on this manuscript. We are grateful to R. Rolfs, L. Wyman and M. Patton for providing ILI data. We thank V. Sahai for his contributions to data collection and processing, and C. Nevill-Manning, A. Roetter and K. Sarvian for their comments on this manuscript.

**Author Contributions** J.G. and M.H.M. conceived, designed and implemented the system. J.G., M.H.M. and R.S.P. analysed the results and wrote the paper. L.B. contributed data. All authors edited and commented on the paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to J.G. or M.H.M. ([flutrends-support@google.com](mailto:flutrends-support@google.com)).

## METHODS

**Automated query selection process.** Using linear regression with fourfold cross validation, we fit models to four 96-point subsets of the 128 points in each region. Each per-query model was validated by measuring the correlation between the model's estimates for the 32 held-out points and the CDC's reported regional ILI percentage at those points. Temporal lags were considered, but ultimately not used in our modelling process.

Each candidate search query was evaluated nine times, once per region, using the search data originating from a particular region to explain the ILI percentage in that region. With four cross-validation folds per region, we obtained 36 different correlations between the candidate model's estimates and the observed ILI percentages. To combine these into a single measure of the candidate query's performance, we applied the Fisher Z-transformation<sup>11</sup> to each correlation, and took the mean of the 36 Z-transformed correlations.

**Computation and pre-filtering.** In total, we fit 450 million different models to test each of the candidate queries. We used a distributed computing framework<sup>12</sup> to divide the work among hundreds of machines efficiently. The amount of computation required could have been reduced by making assumptions about which queries might be correlated with ILI. For example, we could have attempted to eliminate non-influenza-related queries before fitting any models. However, we were concerned that aggressive filtering might accidentally eliminate valuable data. Furthermore, if the highest-scoring queries seemed entirely unrelated to influenza, it would provide evidence that our query selection approach was invalid.

**Constructing the ILI-related query fraction.** We concluded the query selection process by choosing to keep the search queries whose models obtained the highest mean Z-transformed correlations across regions: these queries were deemed to be 'ILI-related'.

To combine the selected search queries into a single aggregate variable, we summed the query fractions on a regional basis, yielding our estimate of the ILI-related query fraction,  $Q(t)$ , in each region. Note that the same set of queries was selected for each region.

**Fitting and validating a final model.** We fit one final univariate model, used for making estimates in any region or state based on the ILI-related query fraction from that region or state. We regressed over 1,152 points, combining all 128 training points used in the query selection process from each of the nine regions. We validated the accuracy of this final model by measuring its performance on 42 additional weeks of previously untested data in each region, from the most recently available time period (18 March 2007 through to 11 May 2008). These 42 points represent approximately 25% of the total data available for the project, the first 75% of which was used for query selection and model fitting.

**State-level model validation.** To evaluate the accuracy of state-level ILI estimates generated using our final model, we compared our estimates against weekly ILI percentages provided by the state of Utah. Because the model was fit using regional data through 11 March 2007, we validated our Utah ILI estimates using 42 weeks of previously untested data, from the most recently available time period (18 March 2007 through to 11 May 2008).

11. David, F. The moments of the z and F distributions. *Biometrika* **36**, 394–403 (1949).
12. Dean, J. & Ghemawat, S. Mapreduce: Simplified data processing on large clusters. *Sixth Symp. Oper. Syst. Des. Implement.*, (2004).