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*

Early detection is the first line of defense against any epidemic, including seasonal and pandemic influenza. One way to improve early detection is to monitor health-seeking behavior. Online web search queries, a new form of health-seeking behavior, are submitted by millions of users around the world each day. We present a method of analyzing 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 influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to utilize search queries for influenza surveillance in areas with a large population of web search users.

Epidemics of seasonal influenza are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. Influenza-related health care expenses and lost productivity cost society between $71 billion and $167 billion annually[1]. In addition to seasonal influenza, a new strain of influenza virus against which no prior immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. A surveillance system which quickly and accurately detects influenza activity is therefore an important line of defense against an influenza epidemic.

Traditional surveillance systems, including those employed by the U.S. Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS), rely on both virologic and clinical data. A network of sentinel laboratories performs virologic testing, by counting and classifying influenza viruses collected from patients, while a network of sentinel physicians reports the fraction of patients presenting with an influenza-like illness (ILI). 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[5] and over-the-counter drug sales[6]. Different measures of internet activity have also been used to survey influenza. About 90 million American adults are believed to search online for information about specific diseases or medical problems each year[7], making web search queries a uniquely valuable source of information about health trends. Previous attempts at using online activity for influenza surveillance have counted online search queries submitted to a Swedish medical website[8], visitors to certain pages on a U.S. health website[9], and user clicks on a search keyword advertisement in Canada[10].

Our proposed surveillance system builds on these earlier attempts by utilizing considerably more data: hundreds of billions of individual searches from five years of Google web search logs. This enables us to create more comprehensive models for use in influenza surveillance, with regional and state-level estimates of influenza-like illness (ILI) activity in the United States. Widespread global usage of online search engines may eventually enable models to be developed in international settings.

In this paper, we demonstrate that Google web search queries can provide accurate influenza surveillance in the United States, with a reporting lag of about one day. We present a method of identifying a set of search queries related to influenza-like illness (ILI) which are used to model weekly regional influenza activity.

Spatiotemporal patterns in online web search queries

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. Billions of queries occurred infrequently and were excluded. Using the internet protocol (IP) 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. Separate aggregate weekly counts were kept for every query in each city and 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. A query fraction for the search query **q** is equivalent to the probability that a random search query submitted from a particular region at a particular time is exactly **q**. 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[11].

A model for influenza-like illness

We sought to develop a simple model which estimates the probability that a random physician visit in a particular region is related to an influenza-like illness (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 process 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: $\text{logit}(I(t)) = \alpha \times \text{logit}(Q(t)) + \epsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t, α is the multiplicative coefficient, and ϵ is the error term. logit(p) is simply $ln(p/(1-p))$.

Publically available historical data from the CDC's U.S. Influenza Sentinel Provider Surveillance Network[12] was used to help build our models. For each of the nine surveillance regions of the United States (Figure 2), 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 modeling and analysis.

An automated process was used to select a group of ILI-related search queries, by measuring how effectively our linear 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 across the nine regions, to identify the search queries which could most accurately model the CDC ILI visit percentage in each region. Despite the large overlap in timing of the seasonal epidemic between regions, the reported regional ILI percentages often varied widely (Figure 3). Our approach rewarded queries which exhibited similar regional variations: 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.

Results

The automated query selection process tested 50 million candidate search queries, and produced a list of the highest scoring queries, sorted by mean Z-transformed correlation across the nine regions. We noted that the 53 highest scoring search queries appeared to be related to influenza-like illnesses. They describe symptoms, treatments, medications, and other diseases that an average person might associate with influenza. The next highest scoring query, “high school basketball”, was the highest scoring off-topic query on the list: basketball season tends to coincide with influenza season in the United States. Though other influenza-related queries scored well and appeared near the top of the list, we wished to minimize manual intervention in the query selection process and thus retained only the top 53 queries, summing them to obtain our estimate of the ILI-related query fraction in each region (additional details in Methods).

Using this ILI-related query fraction as the explanatory variable, we fit nine final predictive models, one per region. Despite using only a single variable, the models were able to obtain a good fit with CDC-reported ILI percentages, with a mean correlation of 0.90 (min=0.76, max=0.95, n=9 regions) (Figure 4). The models were then validated on 42 points per region of previously untested data. Estimates generated by the models for these 42 points obtained a mean correlation of 0.97 (min=0.92, max=0.98, n=9 regions) with the CDC-observed ILI percentages.

Throughout the 2007-2008 influenza season, we used preliminary versions of our models to generate ILI estimates, and shared our results each week with the Epidemiology and Prevention Branch of Influenza Division at CDC to evaluate timeliness and accuracy. Figure 5 illustrates the most recently available data at different points throughout the 2007-2008 influenza season, comparing our model's latest ILI estimates against the latest CDC reports. Across the nine regions, we were able to consistently estimate the current ILI percentage one to two weeks ahead of the publication of ILI percentages by the CDC's U.S. Influenza Sentinel Provider Surveillance Network.

CDC does not make weekly ILI percentages for each state publically available; therefore, state-level ILI estimates cannot be directly generated using our methodology. However, because localized influenza surveillance is particularly useful for public health planning, we used our regional ILI models to estimate the ILI percentages for individual states, by applying the appropriate regional coefficients to the ILI-related query fraction from each specific state. Such estimates are accurate if the relationship between online health-seeking behavior and ILI percentage varies little across a multi-state region.

Using this approach, our ILI estimates for Utah matched the state-reported ILI percentage with a correlation of 0.85 (Figure 6). While a better fit could be obtained by training a model directly on the state-reported ILI percentages, this indirect methodology can be applied to all fifty states regardless of the availability of state-level ILI percentages.

Discussion

Google web search queries can be used to reliably and accurately estimate influenza-like illness percentages 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 one to two weeks ahead of the traditional CDC ILI surveillance reports. The early detection provided by this approach may become an important line of defense against future influenza epidemics in the United States, and perhaps eventually in international settings, including those which lack the infrastructure required for traditional influenza surveillance.

Up-to-date influenza estimates may enable public health officials and health professionals to better respond to seasonal epidemics. If a particular region experiences an early, sharp increase in ILI physician visits, it may be possible to focus additional resources on that region to identify the etiology 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 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. We intend to update our models each year with the latest sentinel provider ILI data, obtaining a better fit and adjusting as each population's online health-seeking behavior evolves over time.

In the event that a pandemic-causing strain of influenza emerges, accurate and early detection of ILI percentages in each region or state may enable public health officials to mount a more effective early response. Though 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 models. Alternatively, panic and concern among healthy individuals may cause a surge in the ILI-related query fraction, resulting in unreasonably exaggerated estimates of the ongoing ILI percentage.

By way of comparison, internet-based surveillance systems such as GPHIN[13] and HealthMap[14] 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.

The search queries used in our model are not, of course, exclusively submitted by users who are experiencing influenza-like symptoms. Our system has no ability to determine that any individual search user is ill, as the correlations we observe are only meaningful across large populations. And 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 some popular cold or flu remedy, could cause such a false alert.

Even without specific knowledge of which search queries are being used[15], media coverage about our system may noticeably change the health-seeking behavior of Google search users. It is difficult to predict the extent to which this might occur.

Some element of selection bias may remain in this work. In particular, the query selection process requires some manual determination of ILI-relatedness, which could be skewed to obtain more favorable results.

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.

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.

Conclusion

Search engine queries can be utilized to rapidly survey influenza activity in large user populations. Harnessing the collective intelligence of millions of users, Google web search logs can provide one of the most timely, broad reaching syndromic surveillance systems available today. We demonstrated that the relative frequency of ILI-related queries is highly correlated with the percentage of ILI physician visits, and that we can accurately estimate weekly regional ILI percentages in the United States. While traditional systems require 1-2 weeks to gather and process surveillance data, our estimates are current each day. Because search queries are an indirect signal of disease activity, the potential for false alerts and the need for laboratory confirmation of disease etiology must be acknowledged. 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 influenza-like illness throughout the 2008-2009 influenza season in the United States. We plan to make results freely available without restriction, so that users are able to view the current estimated influenza burden for all regions of the United States, with a map indicating the overall national situation and regional graphs to illustrate recent trends. Information about this system is available online at http://www.google.org/flu.

Methods

Privacy. At Google, we recognize that privacy is important. None of the queries in our project's database can be associated with a particular individual. Our project's database retains no information about the identity, IP address, or specific physical location of any user. Furthermore, any original web search logs older than 18 months have been anonymized in accordance with Google's Privacy Policy (http://www.google.com/privacypolicy.html).

Query selection. In the query selection process, we fit models using all weeks between September 28, 2003 and March 11, 2007 (inclusive) for which CDC reported a non-zero ILI percentage, yielding 128 training points for each region (each week is one data point). Using linear regression with 4-fold cross validation, we fit models to four 96-point subsets of the 128 points in each region. Each model was validated by measuring the correlation between the model's estimates for the 32 held-out points and CDC's reported regional ILI percentage at those points.

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[16] 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[17] to efficiently divide the work among hundreds of machines. 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 determine how many search queries to select, we visually inspected the list of highest scoring queries, sorted by mean Z-transformed correlation, and noted the first query on the list which seemed clearly unrelated to influenza. We discarded this query, and all queries in the sorted list which appeared below this query, even if they seemed to be related to influenza.

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.

Final validation of regional models. We fit nine final models, one per region, now using all 128 training points from the query selection process. We validated the accuracy of these models by measuring their performance on 42 additional weeks of previously untested data, from the most recently available time period (March 18, 2007 through May 11, 2008).

State-level model validation. To evaluate our ability to generate state-level ILI estimates using a model which was fit over a larger geographic area, we compared our estimates against weekly

ILI percentages provided by the state of Utah. We applied the coefficients from our Mountain Region model to the online search query fractions observed in Utah between September 28, 2003 and January 6, 2008, inclusive, resulting in 139 non-zero validation points.

1. World Health Organization. Influenza fact sheet. *http://www.who.int/mediacentre/factsheets/2003/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/* (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. *Proc AMIA Symp* 215–219 (2003).

6. Magruder, S. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of public health. *Johns Hopkins University APL Technical Digest* **24**, 349–353 (2003).

7. Fox, S. Online health search. *Pew Internet & American Life Project* (2006).

8. Hulth, A. Web queries for influenza monitoring. *ECAIDE* (2007).

9. Johnson, H. *et al.* Analysis of web access logs for surveillance of influenza. *MEDINFO* 1202–1206 (2004).

10. Eysenbach, G. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. *AMI: Symposium Proceedings* 244–248 (2006).

11. Millions of search query time series can be viewed using Google Trends, a free tool available at http://www.google.com/trends .

12. http://www.cdc.gov/flu/weekly .

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

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

15. 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 .

16. David, F. Moments of the z and f distributions. *Biometrika* **36**, 394–403 (1949).

17. Dean, J. & Ghemawat, S. Mapreduce: Simplified data processing on large clusters. *OSDI: Sixth Symposium on Operating System Design and Implementation* (2004).

Acknowledgements We thank Lyn Finelli at the CDC Influenza Division for her ongoing support and comments on this manuscript. We are grateful to Dr. Robert Rolfs and Lisa Wyman at the Utah Department of Health and Monica Patton at the CDC Influenza Division for providing ILI data. We thank Vikram Sahai for his contributions to data collection and processing, and Craig Nevill-Manning, Alex Roetter, and Kataneh Sarvian from Google for their support and comments on this manuscript.

Author Information Correspondence and requests for materials should be addressed to J.G. (email: flutrends-support@google.com).

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 The nine influenza surveillance regions of the United States, which are equivalent to census regions. CDC reports ILI physician visit percentages for each region on a weekly basis.

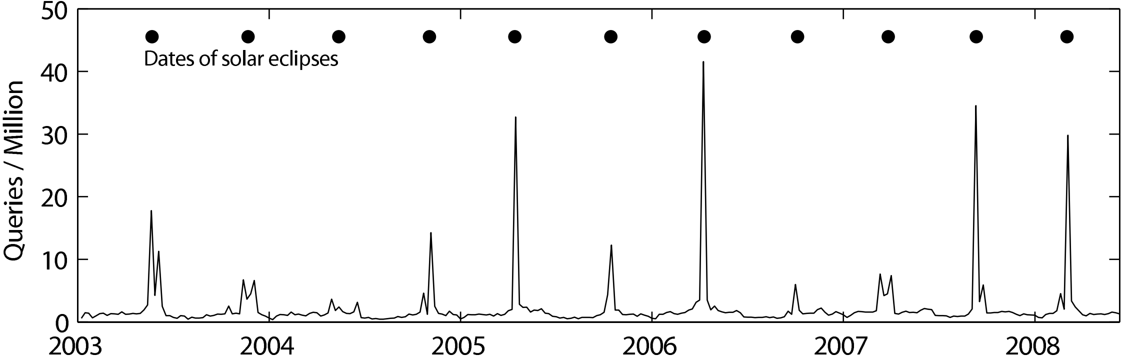
Figure 3 Regional variations in weekly ILI percentages for the South-Atlantic and Pacific Regions (source: CDC Influenza Sentinel Provider Network).

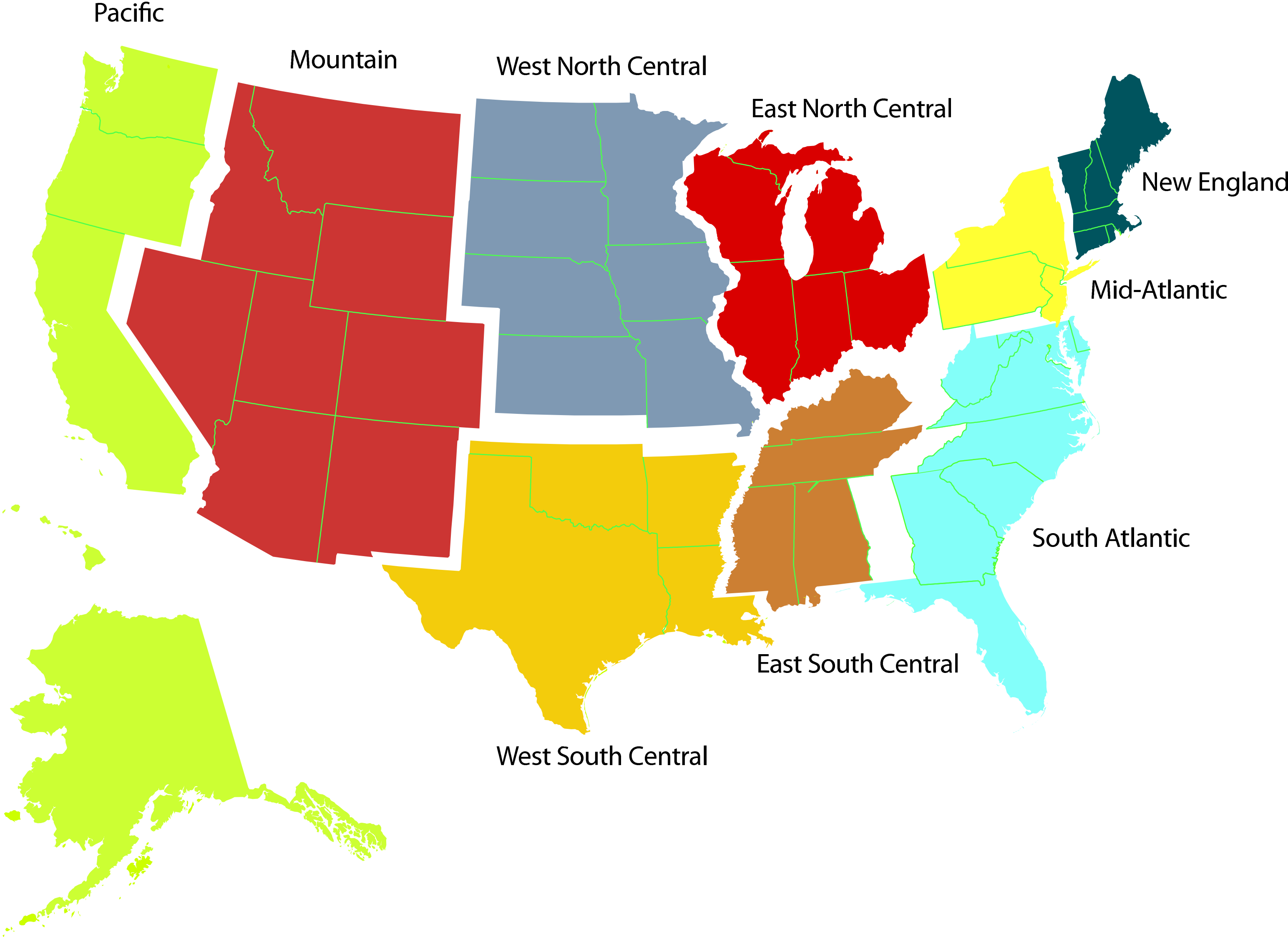
Figure 4 A comparison of model estimates for the Mid-Atlantic Region against CDC-reported ILI percentages, including all points over which the model was fit and validated. A correlation of 0.88 was obtained over 128 points to which the model was fit, while a correlation of 0.94 was obtained over 42 validation points.

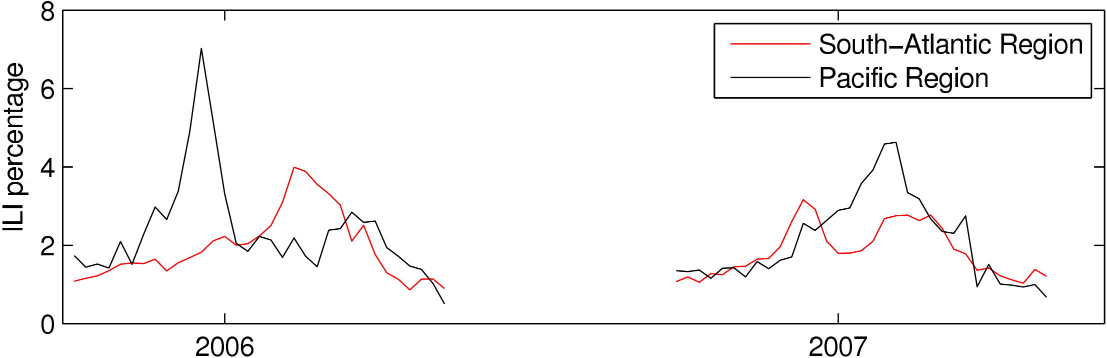
Figure 5 ILI percentages estimated by our model (black) and provided by CDC (red) in the Mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5, we detected a sharply increasing ILI percentage in the Mid-Atlantic region; similarly, on March 3, our models 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.

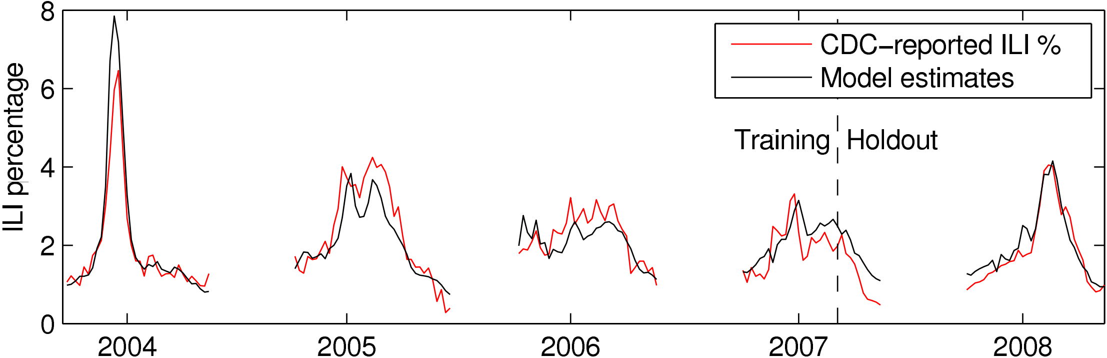
Figure 6 Weekly ILI estimates for Utah, 2003-2007. Estimates are generated using search query data from Utah with model coefficients from the larger Mountain Region.

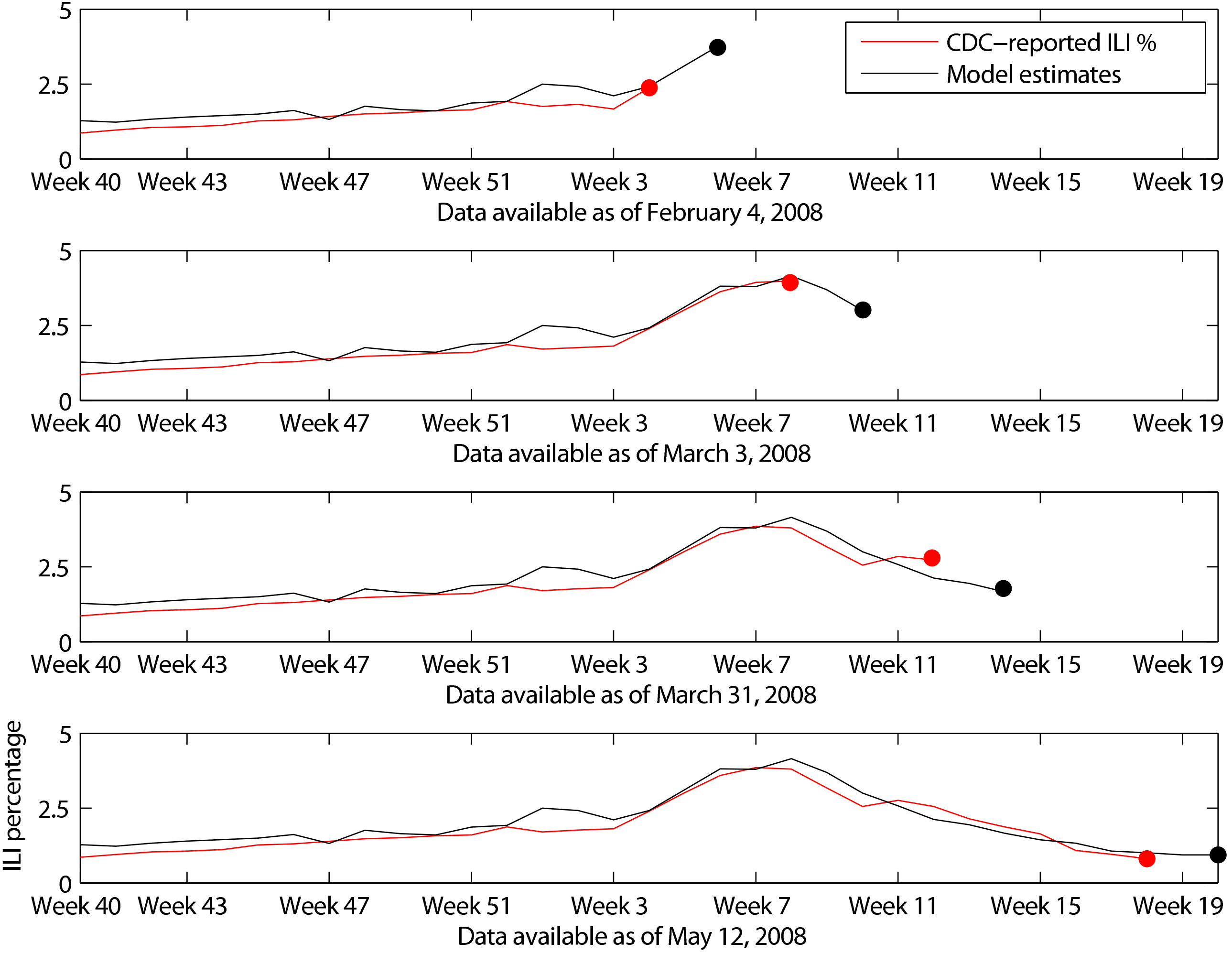
Across 139 points, model estimates obtained a correlation of 0.85 against the state-reported ILI percentages.


50
40
30
20
10
0
Queries / Million
Dates of solar eclipses
2003
2004
2005
2006
2007
2008

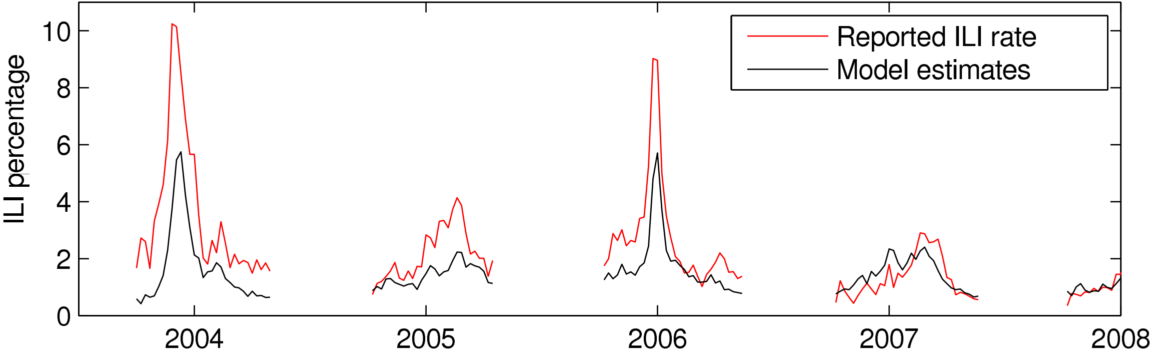

Pacific
Mountain
West North Central
East North Central
New England
Mid-Atlantic
South Atlantic
East South Central
West South Central


ILI percentage
0
2
4
6
8
2006
2007
South−Atlantic Region
Pacific Region


CDC−reported ILI %
Model estimates
Training
Holdout
ILI percentage
0
2
4
6
8
2004
2005
2006
2007
2008



5
2.5
0
CDC−reported ILI %
Model estimates
Week 40
Week 43
Week 47
Week 51
Week 3
Week 7
Week 11
Week 15
Week 19
Data available as of February 4, 2008
Data available as of March 3, 2008
Data available as of March 31, 2008
ILI percentage
Data available as of May 12, 2008


Reported ILI rate
Model estimates
ILI percentage
0
2
4
6
8
10
2004
2005
2006
2007
2008