Using Clinicians' Search Query Data to Monitor Influenza Epidemics

Mauricio Santillana,[1,2] Elaine O. Nsoesie,[2,3] Sumiko R. Mekaru,[2] David Scales,[2,4] and John S. Brownstein[2,3,5]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, [2]Children's Hospital Informatics Program, Boston Children's Hospital, [3]Department of Pediatrics, Harvard Medical School, Boston, and [4]Department of Internal Medicine, Cambridge Health Alliance, Massachusetts; and [5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

Search query information from a clinician's database, UpToDate, is shown to predict influenza epidemics in the United States in a timely manner. Our results show that digital disease surveillance tools based on experts' databases may be able to provide an alternative, reliable, and stable signal for accurate predictions of influenza outbreaks.

Keywords. digital disease detection; Internet-based disease surveillance; prediction of influenza.

The discovery of unusual outbreaks often depends on individual health practitioners who can promptly identify abnormal circumstances and then report those concerns to the greater community [1, 2]. Although the impact of these reports cannot be overstated, recent developments in Internet technologies have demonstrated the power of the crowd as well. For example, crowdsourcing approaches allow members of the public to complete tasks relevant to a larger goal [3]. Search activity on diseases such as influenza and dengue has been shown to correlate with traditional surveillance data in multiple instances [4–8]. Google Flu Trends (GFT) demonstrated a link between influenza-related search query data and the Centers for Disease Control and Prevention's (CDC) influenza-like Illness (ILI) index [5]. Other examples include the use of search query data from Yahoo! [9] and from Baidu [8] to track influenza epidemics. Internet search queries are available much earlier than data from validated traditional surveillance systems and have the potential to provide timely epidemiologic intelligence to inform prevention messaging and healthcare facility staffing decisions.

The potential for the public's search activity to be influenced by anxiety, fears, and rumors raises concerns regarding reliability [10–13]. Although recent revisions to GFT have shown that these concerns can be partially mitigated [13–15], shifting Internet-based surveillance from the entire public to subject-matter experts may maintain timeliness while generating a more reliable and stable signal requiring much less data. A recent small retrospective study using data on queries to a Finnish primary care guidelines database demonstrated, for example, that disease-specific queries for Lyme disease, tularemia, and other infectious diseases correlated well with concurrent confirmed cases [16].

Here, we show that UpToDate (www.uptodate.com), a physician-authored clinical decision support Internet resource that is used by 700 000 clinicians in 158 countries and almost 90% of academic medical centers in the United States, can be used for syndromic surveillance of influenza. Specifically, we use UpToDate's search query activity related to ILI to design a timely sentinel of influenza incidence in the United States.

METHODS

Data

UpToDate is a professional database utilized by healthcare practitioners for point-of-care decisions. The information provided is rigorously authored and edited by experienced physicians. Also, UpToDate topics are accessed >18 million times monthly, and studies suggest that information provided through the site helps improve healthcare outcomes in hospitals [17–19].

In collaboration with UpToDate, we obtained search volume of 23 search terms related to ILI, as well as overall search activity from November 2011 to November 2013 for US accounts only. The search terms were as follows: influenza, *Haemophilus influenzae*, flu, parainfluenza, H1N1, H7N9, H5N1, H3N2, grippe, gripe, adenovirus, rhinovirus, respiratory syncytial virus, metapneumovirus, coronavirus, *Bordetella pertussis*, *Mycoplasma pneumoniae*, pneumonia, bronchitis, H9N2, sinusitis, upper respiratory tract infection, and Tamiflu. We obtained a weekly search fraction for each search term, at any given point in time, by dividing the number of searches for a given phrase by the total number of searches in the UpToDate database, thus minimizing the effects of variation in the overall use of the UpToDate database through

Received 18 June 2014; accepted 4 August 2014; electronically published 12 August 2014.
Correspondence: Mauricio Santillana, MS, PhD, School of Engineering and Applied Sciences, Harvard University, 29 Oxford St, Cambridge, MA 02138 (msantill@fas.harvard.edu).

Clinical Infectious Diseases® 2014;59(10):1446–50

DOI: 10.1093/cid/ciu647

time. We also obtained the national ILI weekly index from the CDC for the same time period to use as a comparator (available at: http://www.cdc.gov/flu/weekly/pastreports.htm).

Analysis

We built a collection of multivariate linear models using the z scores of the aforementioned 23 search terms' weekly search fraction as explanatory variables and the CDC ILI index as our dependent variable. The multiplicative coefficients associated with each search term in each multivariate linear model were updated weekly as the CDC ILI index was updated. Our multivariate models can be expressed as

$$I(t) = \sum_{i=1}^{23} \alpha_i(t)\, Q_i(t) + e,$$

where $I(t)$ is the percentage of national ILI physician visits, $Q_i(t)$ is the search fraction associated with term i at time t, $\alpha_i(t)$ is the multiplicative coefficient associated with each term at time t, and e is the normally distributed error term.

Model selection was performed using a least absolute shrinkage and selection operator (LASSO) technique [20] at every single week incorporating new CDC ILI information as it became available. Therefore, our approach recalibrated weekly the relevance of the search activity for each individual term according to its historical prediction ability. The LASSO technique uses an optimization algorithm that favors models that minimize the mean squared error between the observations and predictions, while penalizing models containing many variables by simultaneously minimizing the sum of the absolute size of the regression coefficients.

We produced real-time estimates of ILI activity at time t, assuming that (1) we only had access to CDC-reported ILI data up to 2 weeks prior (ie, up to t–2 weeks), and (2) assuming that we had access to the real-time (time = t) number of searches in the UpToDate database. Our dynamic approach is similar to the one presented in Santillana et al [15], and inspired by data assimilation techniques widely used in weather forecasting and oceanography [21, 22] and supervised machine-learning techniques [20]. Our methodology was implemented in Matlab version R2011a. The LASSO routine was obtained from (available at: http://www.stanford.edu/~hastie/glmnet_matlab/) in November 2013 [23].

RESULTS

The training period for our first prediction comprised 26 weeks (5 November 2011–28 April 2012). Thus, our first real-time

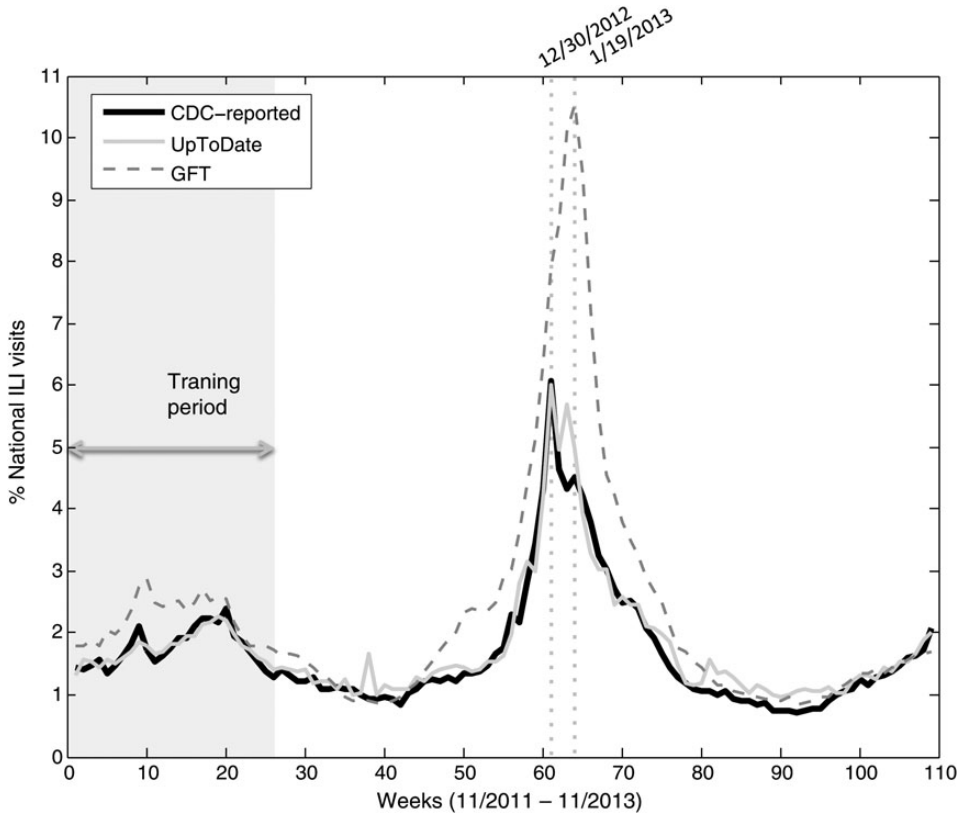

Figure 1. Performance of our methodology along with Centers for Disease Control and Prevention (CDC)–reported influenza-like illness (ILI) activity. CDC ILI is shown in black; our model, named UpToDate, is shown in light grey; and Google Flu Trends (GFT) estimates are shown with a dashed grey line for context.

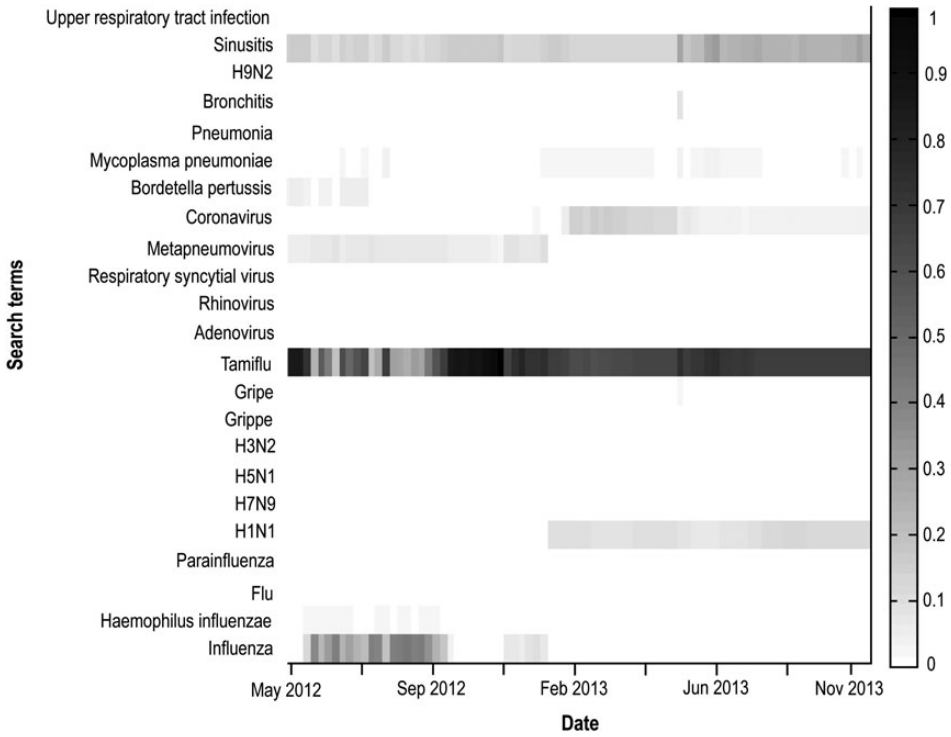

Figure 2. Heatmap representing the relevance of each search term in predicting influenza activity as a function of time (in weeks, starting in May 2012). Clinicians' Tamiflu search activity among clinicians is highly correlated with Centers for Disease Control and Prevention–reported influenza-like illness and thus is found to be the strongest predictor by our algorithm. Sinusitis, influenza, H1N1, and coronavirus display significant relevance as predictors during different time periods.

estimate of ILI was calculated for the week of 12 May 2012 (2 weeks later) using the optimal multivariate model. We produced a weekly time series consisting of real-time estimates using our approach for the subsequent weeks up to the week of 30 November 2013. Figure 1 shows our real-time estimates and the CDC-reported ILI visits. GFT estimates are included for context.

Our estimates predict very well the CDC-reported ILI visits and outperform GFT estimates during the prediction period. Moreover, our approach estimates accurately the peak of the 2012–2013 influenza season (in the week of 30 December 2012) and produces a slight overestimation of the influenza epidemic curve in the second week of January 2013 (overestimating the flu activity by approximately 25% in relative terms—ie, 5.6% of ILI as opposed to the actual 4.5%). This overestimation is minimal when compared to the GFT estimates (overestimating the influenza activity by 130% in relative terms—ie, 10.5% of ILI as opposed to the actual 4.5%).

Our methodology has strong predictive power (Pearson correlation of 0.972; a root mean square error [RMSE] of 0.2829%) during the prediction period starting in the week of 12 May 2012 and ending in the last week of November 2013. Although GFT has a very high Pearson correlation (0.9499) during this same time period, it clearly fails to produce reliable estimates for the peak of the 2012–2013 influenza season. This mismatch is better captured by the RMSE, which shows that GFT estimates are on average off by 1.4% of the national population (ie, almost 5 times larger than our RMSE).

In Figure 2 we present a heatmap representing the relevance of each search term in predicting influenza activity as a function of time, during the validation time period. The term *Tamiflu* is the strongest predictor, whereas *sinusitis*, *influenza*, *H1N1*, and *coronavirus* display relevance as predictors during different time periods.

DISCUSSION

Our findings demonstrate that combining a robust dynamic methodology and subject-matter experts' search activity more accurately predicts influenza activity than the well-established Internet-based tool Google Flu Trends. Specifically, the model presented here has numerous strengths compared to GFT. First, the model does not require expert supervision to adjust the search terms over the course of the influenza season. Our

approach can also accommodate and identify changes in clinicians' selection of search terms over time while retaining the model's predictive power as demonstrated in Figure 2. Not only does this strength address evolving medical vocabulary, it also avoids "model drift" (static models typically match the training data well; however, as time progresses its deviation from truth may cause its predictions to drift farther and farther from truth (as seen in Cook et al [10] with GFT).

The success of our approach suggests that low volumes of queries (in the order of 100–10 000 seconds) in relevant subject-matter experts' databases, such as UpToDate, provide a promising way to identify meaningful signals to track influenza activity. This will motivate the need for future research aimed at testing the accuracy of our methodology at state and city levels, and potentially in the prediction of other diseases. Moreover, our findings in combination with those shown in Jormanainen et al [16] suggest that data acquired from specialized databases may have an improved signal-to-noise ratio and may be less likely to be impacted by public disruption resulting from anxiety or media reports on increased morbidity and mortality during (novel) outbreaks of influenza.

Limitations in this data source include those inherent in most novel data sources advanced for monitoring infectious diseases. Although timely, these data sources lack the specificity observed in traditional surveillance systems, which rely on hierarchical reporting procedures. These data streams therefore supplement traditional disease surveillance provided by organizations such as the CDC. Finally, UpToDate data is not publicly available and thus not ready to be used as an alternative disease detection sentinel.

CONCLUSIONS

In this study, we demonstrate that search queries from the UpToDate database in conjunction with a dynamic multivariate methodology can be successfully utilized to obtain real-time estimates of influenza incidence in the United States before the release of official reports. Clinicians can use outcomes from the model to monitor estimated levels of influenza in the United States. We also discuss the potential usefulness and limitations of digital data sources for infectious disease surveillance based on search query data [5, 7, 8, 24–28]. Future work may include analysis of smaller geographic units.

Notes

Acknowledgments. We thank the Analytics team at UpToDate for sharing the data used in this manuscript.

Disclaimer. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Financial support. Financial support for this study was provided by the National Library of Medicine (grant number R01 LM010812-04).

Potential conflicts of interest. All authors: No potential conflicts of interest.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis **2004**; 39:227–32.
2. Cowen P, Garland T, Hugh-Jones ME, et al. Evaluation of ProMED-mail as an electronic early warning system for emerging animal diseases: 1996 to 2004. J Am Vet Med Assoc **2006**; 229:1090–9.
3. Ranard B, Ha Y, Meisel Z, et al. Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med **2014**; 29:187–203.
4. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using Internet searches for influenza surveillance. Clin Infect Dis **2008**; 47:1443–8.
5. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature **2009**; 457:1012–4.
6. Madoff LC, Fisman DN, Kass-Hout T. A new approach to monitoring dengue activity. PLoS Negl Trop Dis **2011**; 5:e1215.
7. Chan EH, Sahai V, Conrad C, Brownstein JS. Using Web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS Negl Trop Dis **2011**; 5:e1206.
8. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in China with search query from Baidu. PLoS One **2013**; 8:e64323.
9. Polgreen PM, Nelson FD, Neumann GR, Weinstein RA. Use of prediction markets to forecast infectious disease activity. Clin Infect Dis **2007**; 44:272–9.
10. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One **2011**; 6:e23610.
11. Butler D. When Google got flu wrong. Nature **2013**; 494:155–6.
12. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol **2013**; 9:e1003256.
13. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science **2014**; 343:1203–5.
14. Copeland P, Romano R, Zhang T, Hecht G, Zigmond D, Stefansen C. Google disease trends: an update. Int Soc Negl Trop Dis **2013**; 3.
15. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google Flu Trends? Am J Prev Med **2014**; 14:S0749–3797.
16. Jormanainen V, Jousimaa J, Kunnamo I, Ruutu P. Physicians' database searches as a tool for early detection of epidemics. Emerg Infect Dis **2001**; 7:474–6.
17. Bartlett JC, Marshall JG. The value of library and information services in patient care: Canadian results from an international multisite study. J Can Health Libr Assoc **2013**; 34:138–46.
18. Addison J, Whitcombe J, William Glover S. How doctors make use of online, point-of-care clinical decision support systems: a case study of UpToDate. Health Info Libr J **2013**; 30:13–22.
19. Bonis PA, Pickens GT, Rind DM, Foster DA. Association of a clinical knowledge support system with improved patient safety, reduced complications and shorter length of stay among Medicare beneficiaries in acute care hospitals in the United States. Int J Med Inform **2008**; 77:745–53.
20. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B **1996**; 58:267–88.
21. Ghil M, Malanotte-Rizzoli P. Data assimilation in meteorology and oceanography. Adv Geophys **1991**; 33:141–266.
22. Wang B, Zou X, Zhu J. Data assimilation and its applications. Proc Natl Acad Sci U S A **2000**; 97:11143–4.

23. Qian J, Hastie T, Friedman J, Tibshirani R, Simon N. Glmnet for Matlab, **2013**. Available at: http://www.stanford.edu/~hastie/glmnet_matlab/. Accessed November 2013.
24. Dugas AF, Jalalpour M, Gel Y, et al. Influenza forecasting with Google Flu Trends. PLoS One **2013**; 8:e56176.
25. Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK: Association for Computational Linguistics, **2011**: 1568–76.
26. Lamb A, Paul MJ, Dredze M. Separating fact from fear: tracking flu infections on Twitter. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [Internet]. Atlanta, GA: Association for Computational Linguistics, **2013**; 789–95. Available at: http://www.aclweb.org/anthology/N13-1097. Accessed November 2013.
27. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med **2008**; 5:e151.
28. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc **2008**; 15:150–7.