Clinical Infectious Diseases Advance Access published August 12, 2014

Using Clinician's Search Query Data to Monitor Influenza Epidemics

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

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

[2]Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA

[3]Department of Internal Medicine, Cambridge Health Alliance, Cambridge, MA

[4]Department of Pediatrics, Harvard Medical School, Boston, MA

[5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, H3A 1A2, Canada

Corresponding author: Mauricio Santillana, 29 Oxford St, Cambridge, MA 02138. Phone: 617-495-2891, Email: msantill@fas.harvard.edu.

Alternative author: John Brownstein, One Autumn Street Boston MA 02215. Phone: 617-355-6998, Email: john.brownstein@childrens.harvard.edu

Abstract:

Search query information from a clinician's database, UpToDate, is shown to timely predict influenza epidemics in the United States. 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 flu outbreaks.

1. Introduction

The discovery of unusual outbreaks often depend on individual health practitioners who can promptly identify abnormal circumstances and then report those concerns to the greater community (1,2). While 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). While 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[1] (www.uptodate.com), a physician-authored clinical decision support Internet resource, can be used for syndromic surveillance of influenza. Specifically, we use UpToDate's search query activity related to influenza-like illness to design a timely sentinel of influenza incidence in the US.

2. Methods

2.1 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 more than 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 influenza-like illness, as well as overall search activity from November 2011 to November 2013 for United States accounts only. The search terms were: *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,*

[1] UpToDate is used by 700,000 clinicians in 158 countries and almost 90% of academic medical centers in the United States

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 time. We also obtained the national influenza-like illness weekly index from the Centers of Disease Control and Prevention (CDC) for the same time period to use as a comparator (http://www.cdc.gov/flu/weekly/pastreports.htm).

2.2 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:

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

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 (a) we only had access to CDC-reported ILI data up to two weeks prior, *i.e.* up to t-2 weeks, and (b) 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 (http://www.stanford.edu/~hastie/glmnet_matlab/) in November 2013 (23).

3. Results

Our first training period contained 26 weeks (the weeks from November 5th, 2011 through April 28th, 2012), for our first prediction. Thus, our first real-time estimate of ILI was calculated for the week of May 12th, 2012 (two 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 November 30th, 2013.

Figure 1 shows our real-time estimates and the CDC ILI reported 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 Dec 30th, 2012) and produces a slight over-estimation of the influenza epidemic curve in the second week of January 2013 (over-estimating the flu activity by approximately 25% in relative terms, i.e., 5.6% of ILI as opposed to the actual 4.5%). This over-estimation is minimal when compared to the GFT estimates (over-estimating the flu activity by 130% in relative terms, i.e. 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 May 12th, 2012 and ending in the last week of November 2013. While 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, *i.e.* 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, while *sinusitis, influenza*, *h1n1,* and *coronavirus*

display relevance as predictors during different time periods.

4. Discussion

Our findings demonstrate that combining a robust dynamic methodology and subject matter experts' search activity more accurately predicts flu activity than the well-established general public 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 clinician's 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*, but it also avoids "model drift" (static models typically match the training data well, but as time progresses its deviation from truth may cause its predictions to drift farther and farther from truth as seen in Cook et al. 2011 (10) with GFT).

The success of our approach suggests that low volumes of queries (in the order of 100s to 10000s) in relevant subject matter experts' databases, such as UpToDate, provide a promising way to identify meaningful signals to track flu 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 (16) suggests 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.

5. 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 US 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

Funding: Financial support for this study was provided by research grants from the National Library of Medicine grant R01 LM010812-04. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: None

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

References

1. Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis. 2004 Jul 15;39:227–32.

2. Cowen P, Garland T, Hugh-Jones ME, Shimshony A, Handysides S, Kaye D, 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 Oct 1;229:1090–9.

3. Ranard B, Ha Y, Meisel Z, Asch D, Hill S, Becker L, et al. Crowdsourcing—Harnessing the Masses to Advance Health and Medicine, a Systematic Review. J Gen Intern Med. 2014 Jan 1;29(1):187–203.

4. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using Internet Searches for Influenza Surveillance. Clin Infect Dis. 2008 Dec 1;47(11):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 Feb 19;457:1012–4.

6. Madoff LC, Fisman DN, Kass-Hout T. A New Approach to Monitoring Dengue Activity. PLoS Negl Trop Dis. 2011 May 31;5(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 May;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 Aug 19;6(8):e23610.

11. Butler D. When Google got flu wrong. Nature. 2013 Feb 14;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 Oct 17;9(10):e1003256.

13. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014 Mar 14;343(6176):1203–5.

14. Copeland P, Romano R, Zhang T, Hecht G, Zigmond D, Stefansen C. GOOGLE DISEASE TRENDS: AN UPDATE. International Society of Neglected Tropical Diseases 2013. 2013. p. 3.

15. M Santillana, D W Zhang, B M Althouse, J W Ayers. What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends? Am J Prev Med. 2014;14(00238-4):S0749–3797.

16. Vesa Jormanainen, Jukkapekka Jousimaa, Ilkka Kunnamo, Petri Ruutu. Physicians' Database Searches as a Tool for Early Detection of Epidemics. Emerg Infect Dis. 2001;7(3):474–6.

17. Bartlett JC, Marshall JG. The Value of Library and Information Services in Patient Care: Canadian Results From an International Multisite Study1. J Can Health Libr Assoc. 2013 Nov 29;34(03):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 Inf Libr J. 2013;30(1):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 Inf. 2008 Nov 1;77(11):745–53.

20. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B. 1996;58(1):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. PNAS. 97(21):11143–4.

23. Qian J, Hastie T, Friedman J, Tibshirani R, Simon N. Glmnet for Matlab [Internet]. 2013. Available from: http://www.stanford.edu/~hastie/glmnet_matlab/

24. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, 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. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, United Kingdom: Association for Computational Linguistics; 2011. p. 1568–76.

26. Lamb A, Paul MJ, Dredze M. Separating Fact from Fear: Tracking Flu Infections on Twitter. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [Internet].

Atlanta, Georgia: Association for Computational Linguistics; 2013. p. 789–95. Available from: http://www.aclweb.org/anthology/N13-1097

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 Jul 8;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 Inf Assoc. 2008 Mar;15:150–7.

Figure Legends:

Figure 1. Performance of our methodology along with the CDC reported ILI activity. CDC-ILI is shown in black, our model, named UpToDate, is shown in blue, and Google Flu Trends estimates are shown in red 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). Clinician's *Tamiflu* search activity amongst clinician's is highly correlated with CDC-reported ILI and thus it is found to be the strongest predictor by our algorithm. *Sinusitis, Influenza*, *h1n1,* and *coronavirus* display significant relevance as predictors during different time periods.