Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

[a]Department of Statistics, Harvard University, Cambridge, MA 02138; [b]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [c]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with GOogle search data), that uses publicly available online search data. In addition to having a rigorous statistical foundation, ARGO outperforms all previously available Google-search–based tracking models, including the latest version of Google Flu Trends, even though it uses only low-quality search data as input from publicly available Google Trends and Google Correlate websites. ARGO not only incorporates the seasonality in influenza epidemics but also captures changes in people's online search behavior over time. ARGO is also flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

digital disease detection | seasonal influenza | big data | influenza-like illnesses activity real-time estimation | autoregressive exogenous model

Big data sets are constantly generated nowadays as the activities of millions of users are collected from Internet-based services. Numerous studies have suggested great potential of these big data sets to detect/manage epidemic outbreaks [influenza (1–6), Ebola (7), dengue (8)], predict changes in stock prices (9, 10) and housing prices (11), etc. In 2009, Google Flu Trends (GFT), a digital disease detection system that uses the volume of selected Google search terms to estimate current influenza-like illnesses (ILI) activity, was identified by many as a good example of how big data would transform traditional statistical predictive analysis (12). However, significant discrepancies between GFT's flu estimates and those measured by the Centers for Disease Control (CDC) in subsequent years led to considerable doubt about the value of digital disease detection systems (13). Although multiple articles have identified methodological flaws in GFT's original algorithm (14–16) and have led to incremental improvements (14, 16) (see also googleresearch.blogspot.com/2014/10/google-flu-trends-gets-brand-new-engine.html), a statistical framework that is theoretically sound and capable of accurate estimation is still lacking. Here we present such a framework that culminates in a method that outperforms all existing methodologies for tracking influenza activity using internet search data.

Influenza outbreaks cause up to 500,000 deaths a year worldwide, and an estimated 3,000–50,000 deaths a year in the United States (17). Our ability to effectively prepare for and respond to these outbreaks heavily relies on the availability of accurate real-time estimates of their activity. Existing methods to predict the timing, duration, and magnitude of flu outbreaks remain limited (18). Well-established clinical methods to track flu activity, such as the CDC's ILINet, report the percentage of patients seeking medical attention with ILI symptoms (www.cdc.gov/flu/). Although CDC's %ILI is only a proxy of the flu activity in the population, it can help officials allocate resources in preparation for potential surges of patient visits to hospital facilities. See refs. 19–21 for further discussion.

CDC's ILI reports have a delay of 1–3wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity (18, 21–25). In recent years, methods that harness Internet-based information have also been proposed, such as Google (1), Yahoo (2), and Baidu (3) Internet searches, Twitter posts (4), Wikipedia article views (5), clinicians' queries (6), and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) (26), Flutracking (Australia) (27), and Flu Near You (United States) (28). Among them, GFT has received the most attention and has inspired subsequent digital disease detection systems (3, 8, 29–32). Interestingly, Google has never made their raw data public, thus making it impossible to reproduce the exact results of GFT.

We highlight three limitations of the original GFT algorithm, previously identified in refs. 15 and 16. First, it was shown that a static approach, which does not take advantage of newly available CDC's ILI activity reports as the flu season evolves, produced model drift, leading to inaccurate estimates. Second, the idea of aggregating the multiple query terms (the independent variables in the GFT model) into a single variable did not allow for changes in people's Internet search behavior over time (and thus changes in query terms' abilities to track flu) to be appropriately captured. Third, GFT ignored the intrinsic time series properties, such as seasonality of the historical ILI activity, thus overlooking potentially crucial information that could help produce accurate real-time ILI activity estimates.

Significance

Big data generated from the Internet have great potential in tracking and predicting massive social activities. In this article, we focus on tracking influenza epidemics. We propose a model that utilizes publicly available Google search data to estimate current influenza-like illness activity level. Our model outperforms all available Google-search–based real-time tracking models for influenza epidemics at the national level of the United States, including Google Flu Trends. Our model is flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for estimation and prediction at multiple temporal and spatial resolutions for other social events.

Author contributions: M.S. and S.C.K. designed research; S.Y., M.S., and S.C.K. performed research; S.Y. analyzed data; and S.Y., M.S., and S.C.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: kou@stat.harvard.edu or msantill@fas.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1515373112/-/DCSupplemental.

Our Contribution

The methodology presented here produces robust and highly accurate ILI activity level estimates by addressing the three aforementioned shortcomings of the multiple GFT engines. In addition, we provide a theoretical framework that, for the first time to our knowledge, justifies the prevailing use of linear models in the digital disease detection literature by incorporating causality arguments through a hidden Markov model. This theoretical framework contains, as a special case, the model developed in ref. 16. Our model not only achieves the goal of (*i*) dynamically incorporating new information from CDC reports as it becomes available and (*ii*) automatically selecting the most useful Google search queries for estimation as in ref. 16, but also largely improves estimation by (*iii*) including the long-term cyclic information (seasonality) from past flu seasons on record as input variables and (*iv*) using a 2-y moving window (which immediately precedes the desired date of estimation) for the training period to capture the most recent changes in people's search patterns and time series behavior (33). Our methodology efficiently builds a prediction model from individual search frequency as well as the past records of ILI activity. It uses both sources of information more efficiently than simply combining GFT with autoregressive terms as suggested in ref. 15, because GFT is not optimally aggregated to provide additional information on top of time series information. Furthermore, we provide a quantitative efficiency metric that measures the statistical significance of the improvement of our methodology over other alternatives. For example, our method is twice as accurate as the method that combines GFT with autoregressive terms. Finally, even though we use as input only the publicly available, low-quality data from the Google Correlate and Google Trends websites, our method has significant improvement over the latest version of GFT.

We name our model ARGO, which stands for AutoRegression with GOogle search data. Statistically speaking, ARGO is an autoregressive model with Google search queries as exogenous variables; ARGO also employs L_1 (and potentially L_2) regularization to achieve automatic selection of the most relevant information.

Results

Retrospective estimates of influenza activity (ILI activity level, as reported by the CDC) were produced using our model, ARGO,

Table 1. Comparison of different models for the estimation of influenza epidemics

	Whole period (Mar 29, 2009 to Jul 11, 2015)	Off-season flu H1N1	Regular flu seasons (week 40 to week 20 next year)				
			2010–2011	2011–2012	2012–2013	2013–2014	2014–15
RMSE							
ARGO	**0.608**	0.640	**0.596**	**0.807**	**0.687**	**0.306**	**0.438**
GFT (Oct 2014)	2.216	0.773	1.110	3.023	4.451	0.986	0.700
Ref. 16	0.915	0.833	0.881	2.027	1.090	0.446	0.663
GFT+AR(3)	0.912	**0.580**	0.602	1.382	1.279	0.993	0.906
AR(3)	0.957	0.813	0.794	1.051	1.191	0.969	0.928
Naive	1 (0.348)	1 (0.600)	1 (0.339)	1 (0.163)	1 (0.499)	1 (0.350)	1 (0.465)
MAE							
ARGO	**0.649**	0.584	**0.574**	**0.748**	**0.650**	**0.391**	**0.530**
GFT (Oct 2014)	1.834	0.777	1.260	3.277	5.028	0.891	0.770
Ref. 16	1.052	0.719	1.010	2.211	1.029	0.610	0.820
GFT+AR(3)	0.888	**0.570**	0.613	1.308	1.016	1.034	0.839
AR(3)	0.925	0.777	0.787	0.951	0.988	0.917	0.934
Naive	1 (0.201)	1 (0.425)	1 (0.259)	1 (0.135)	1 (0.325)	1 (0.212)	1 (0.295)
MAPE							
ARGO	**0.787**	**0.620**	**0.663**	**0.770**	**0.719**	**0.453**	**0.620**
GFT (Oct 2014)	1.937	0.721	1.394	3.442	5.419	0.892	0.895
Ref. 16	1.381	0.765	1.380	2.306	1.251	0.754	0.958
GFT+AR(3)	1.037	0.683	0.698	1.407	0.986	1.062	0.828
AR(3)	1.003	0.894	0.814	0.947	0.939	0.891	0.916
Naive	1 (0.090)	1 (0.139)	1 (0.105)	1 (0.081)	1 (0.110)	1 (0.084)	1 (0.097)
Correlation							
ARGO	**0.986**	0.985	**0.989**	**0.928**	**0.968**	**0.993**	**0.993**
GFT (Oct 2014)	0.875	**0.989**	0.968	0.833	0.926	0.969	0.986
Ref. 16	0.971	0.967	0.983	0.927	0.956	0.985	0.984
GFT+AR(3)	0.967	0.986	0.985	0.879	0.929	0.945	0.957
AR(3)	0.964	0.968	0.971	0.877	0.903	0.927	0.945
Naive	0.961	0.951	0.954	0.887	0.924	0.923	0.937
Correlation of increment							
ARGO	**0.758**	0.806	**0.810**	0.286	**0.527**	**0.938**	0.912
GFT (Oct 2014)	0.706	**0.863**	0.702	0.484	0.502	0.847	**0.918**
Ref. 16	0.690	0.776	0.693	**0.510**	0.367	0.915	0.889
GFT+AR(3)	0.512	0.708	0.708	0.165	0.141	0.534	0.587
AR(3)	0.385	0.585	0.569	0.077	0.011	0.404	0.493
Naive	0.436	0.602	0.570	0.095	0.134	0.406	0.514

GFT+AR(3) stands for the model $p_t = \mu + \alpha_1 p_{t-1} + \alpha_2 p_{t-2} + \alpha_3 p_{t-3} + \beta \mathrm{GFT}(t)$, where the GFT estimate is treated as an exogenous variable. Boldface highlights the best performance for each metric in each study period. RMSE, MAE, and MAPE are relative to the error of naive method; that is, the number reported is the ratio of error of a given method to that of the naive method. The absolute error of the naive method is reported in parentheses. All comparisons are based on the original scale of ILI activity level.

for the time period of March 29, 2009 through July 11, 2015, assuming we had access only to the historical CDC's ILI reports up to the previous week of estimation. We compared ARGO's estimates with the ground truth: the CDC-reported weighted ILI activity level, published typically with 1- or 2-wk delay, by calculating a collection of accuracy metrics described in *Materials and Methods*. These metrics include the root-mean-squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), correlation with estimation target, and correlation of increment with estimation target. For comparison, we calculated these accuracy metrics for (*i*) GFT estimates (accessed on July 11, 2015), (*ii*) estimates produced using the method of Santillana et al. (6, 16), (*iii*) estimates produced by combining GFT with a lag-3 autoregressive model, AR(3), as suggested in ref. 15, (*iv*) estimates produced with an AR(3) autoregressive model (4, 15), and (*v*) a naive method that simply uses the value of the prior week's CDC ILI activity level as the estimate for the current one. For fair comparison, all benchmark models (*ii*–*iv*) are dynamically trained with a 2-y moving window.

Table 1 summarizes these accuracy metrics for all estimation methods for multiple time periods. The "Whole period" column shows that ARGO's estimates outperform all other alternatives, in every accuracy metric for the whole time period. The other columns of Table 1 show the performance of all of the methods for the 2009 off-season H1N1 flu outbreak, and each regular flu season since 2010. Fig. 1 displays the

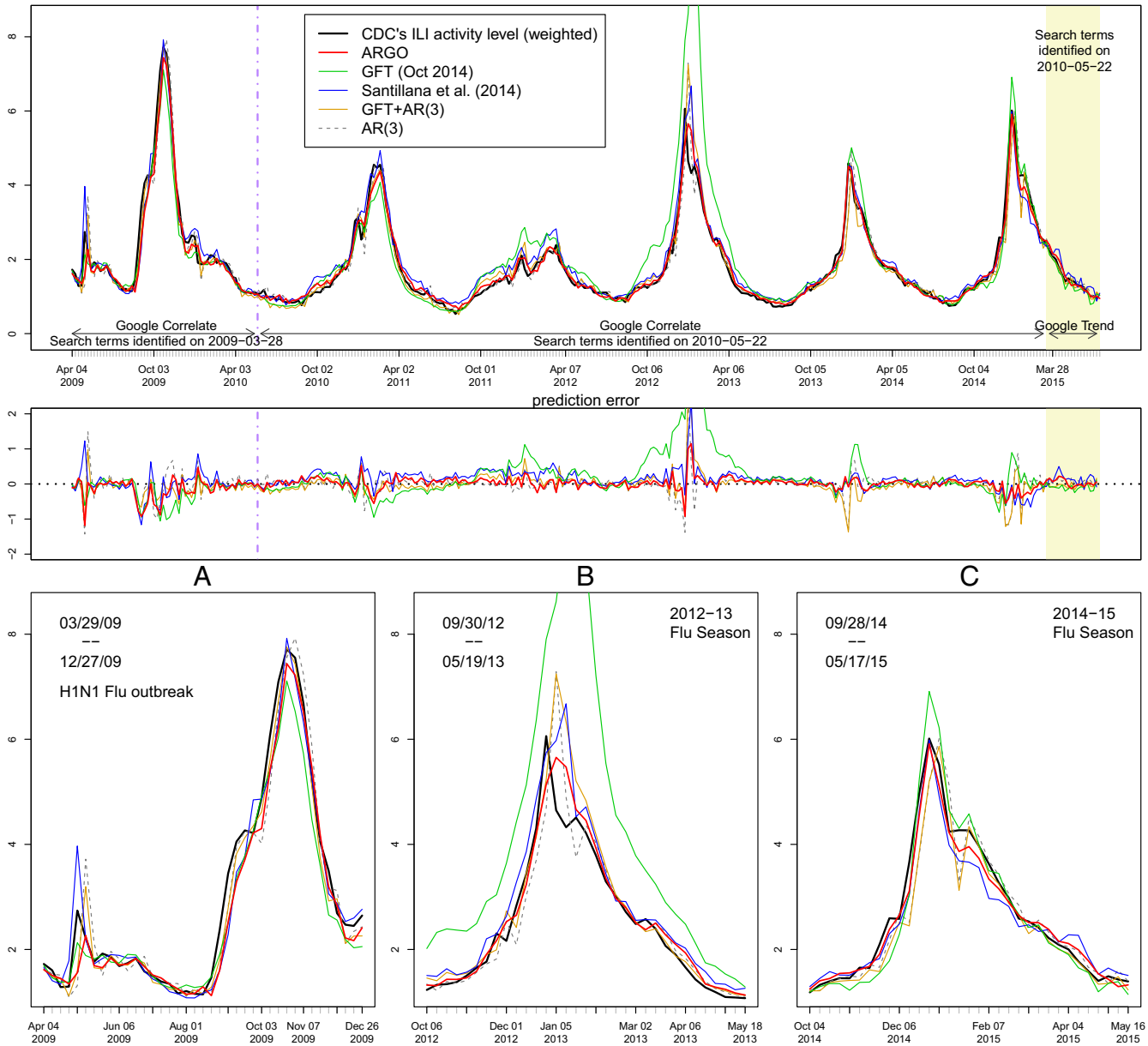

Fig. 1. Estimation results. (*Top*) The estimated ILI activity level from ARGO (thick red), contrasting with the true CDC's ILI activity level (thick black) as well as the estimates from GFT (green), method of ref. 16 (blue), GFT plus AR(3) model (dark yellow), and AR(3) model (dashed gray). The two background shades, white and yellow, reflect two data sources, Google Correlate and Google Trends, respectively. The dash-dotted purple vertical line separates Google Correlate data with search terms identified on March 28, 2009 and May 22,2010. (*Middle*) The estimation error, defined as estimated value minus the CDC's ILI activity level. (*Bottom*) Zoomed-in plots for estimation results in different study periods. (*A*) The H1N1 flu outbreak period. (*B*) The 2012–2013 regular flu season. (*C*) The 2014–2015 regular flu season. A regular flu season is defined as week 40 of one year to week 20 of the following year.

estimates against the observed CDC-reported ILI activity level.

Close inspection shows that, in the post-2009 regular flu seasons, ARGO uniformly outperformed all other alternative estimation methods in terms of RMSE, MAE, MAPE, and correlation. ARGO avoids the notorious overshooting problem of GFT, as seen in Fig. 1. During the 2009 off-season H1N1 flu outbreak, ARGO had the smallest MAPE. In terms of RMSE and MAE, ARGO (relative RMSE = 0.640, relative MAE = 0.584) had the second best performance, underperforming slightly only the GFT+AR(3) model (relative RMSE = 0.580, relative MAE = 0.570). In terms of correlation, ARGO (r = 98.5%) had similar performance to the (potentially in-sample data of) GFT (r = 98.9%) (14) and GFT+AR(3) models (r = 98.6%) and outperformed all of the other alternatives.

To assess the statistical significance of the improved prediction power of ARGO, we constructed a 95% confidence interval for the relative efficiency of ARGO compared with other benchmark methods. The relative efficiency of method 1 to method 2 is the ratio of the true mean-squared error of method 2 to that of method 1 (34), which can be estimated by its observed value (see Eq. **4**); its confidence interval can be constructed by stationary bootstrap of the error residual time series (35). Table 2 shows that ARGO is estimated to be at least twice as efficient as any other alternative, and the improvement in accuracy is highly statistically significant.

It is well known that CDC reports undergo revisions, weeks after their initial publication, that respond to internal consistency checks and lead to more accurate estimates of patients with ILI symptoms seeking medical attention. Thus, the available historical CDC information, in a given week, is not necessarily as accurate as it will be. We tested the effect of using (potentially inaccurate) unrevised information by obtaining the historical unrevised and revised reports, and the dates when the reports were revised, from the CDC website for the time period of our study. We used only the information that would have been available to us, at the time of estimation, and produced a time series of estimates for the whole time period described before. We compared our estimates to all other methods and found that ARGO still outperformed them all. Moreover, the values of all five accuracy metrics for ARGO essentially did not change, suggesting a desirable robustness to revisions in CDC's ILI activity reports. The results are shown in Table S1.

We faced an additional challenge in producing real-time estimates for the latest portion of the 2014–2015 flu season. At the time of writing this article, the only data available to us for the week of March 28, 2015 and later came from the Google Trends website. The information from Google Trends has even lower quality than from Google Correlate and changes every week. These undesired changes affected the quality of our estimates. To assess the stability of ARGO in the presence of these variations in the data, we obtained the search frequencies of the same query terms from Google Trends website on 25 different days during the month of April 2015 and produced a set of 25 historical estimates using ARGO. The results of the accuracy metrics associated to these estimates are shown in Table S2. This table shows that, despite the observed variation in the Google Trends data, ARGO is threefold more stable than the method of ref. 16, and still outperforms on average any other method.

Table 2. Estimate of relative efficiency of ARGO compared with other models with 95% confidence interval (CI)

	Point estimate	95% CI
GFT (Oct 2014)	12.85	[5.18, 91.82]
Ref. 16	2.02	[1.36, 2.83]
GFT+AR(3)	2.17	[1.23, 4.53]
AR(3)	2.40	[1.56, 3.69]

Relative efficiency being larger than 1 suggests increased predictive power of ARGO compared with the alternative method. The estimates and the bootstrap confidence intervals are constructed based on data from March 29, 2009 to May 17, 2015.

Discussion

Strength of ARGO. The results presented here demonstrate the superiority of our approach in terms of both accuracy and robustness, compared with all existing flu tracking models based on Google searches. The value of these results is even higher given the fact that they were produced with low-quality input variables. It is highly likely that our methodology would lead to even more accurate results if we were given access to the input variables that Google uses to calculate their estimates.

The combination of seasonal flu information with dynamic reweighting of search information appears to be a key factor in the enhanced accuracy of ARGO. The level of ILI activity last week typically has a significant effect on the current level of ILI activity, and ILI activity half a year ago and/or 1 y ago could provide further information, as shown in Fig. S1, which reflects a strong temporal autocorrelation. The integration of time series information leads to a smooth and continuous estimation curve and prevents undesired spikes. However, simply adding GFT to an autoregressive model is suboptimal compared with ARGO, because simply treating GFT as an individual variable does not allow adjustment for time series information at the resolution of individual query terms, and many terms included in GFT may no longer provide extra information once time series information is incorporated. In fact, once the time series information is included, fewer Google search query terms remain significant. For example, among 100 Google Correlate query terms, ARGO selected 14 terms, on average, each week, whereas the method of ref. 16 and GFT (1) selected 38 and 45 terms, respectively, each week on average. The combination of ARGO's smoothness and sparsity lead to a substantial reduction on the estimation error, as observed in Tables 1 and 2, where ARGO shows improved performance in all evaluation metrics over the whole time period and is twice as efficient as GFT+AR(3).

Our methodology allows us to transparently understand how Google search information and historical flu information complement one another. Time series models tend to be slow in response to sudden observed changes in CDC's ILI activity level. The AR(3) model shows this "delaying" effect, despite its seemingly good correlation. Google searches, on the other hand, are better at detecting sudden ILI activity changes, but are also very sensitive to public's overreaction.

To investigate further the responsiveness (comovement) of ARGO toward the change in ILI activity, we calculated the correlation of increment between each estimation model and CDC's ILI activity level. The correlation of increment between two time series a_t and b_t is defined as $\mathrm{Corr}(a_t - a_{t-1}, b_t - b_{t-1})$, which measures how well a_t captures the changes in b_t. Table 1 shows that ARGO has similar capability to that of GFT and the method of ref. 16 in capturing the changes in ILI level, and outperforms the time series model AR(3) uniformly.

Time series information (seasonality) tends to pull ARGO's estimate toward the historical level. This was evident at the onset of the off-season H1N1 flu outbreak (week ending at May 2, 2009), which resulted in ARGO's underestimation. ARGO self-corrected its performance the following week by shifting a portion of model weights from the time series domain to the Google searches domain. Inversely, at the height of 2012–2013 season, ARGO, GFT, and the method of ref. 16 all missed the peak due to an unprecedented surge of search activity. ARGO achieved the

fastest self-correction by redistributing the weights not only across Google terms but also across time series terms, missing the peak by only 1 wk, as opposed to 2 wk for ref. 16 and about 4 wk for GFT. It is important to note that although we have used CDC's ILI as our gold standard for influenza activity in the US population, and data from Google Correlate/Trends as our independent variables, our methodology can be immediately adapted to any other suitable ILI gold standard and/or set of independent variables.

Limitations and Next Steps. Although ARGO displays a clear superiority over previous methods, it is not fail-proof. Because it relies on the public's search behavior, any abrupt changes to the inner works of the search engine or any changes in the way health-related search information is displayed to users will affect the accuracy of our methodology (36, 37). We expect that ARGO will be fast at correcting itself if any such change takes place in the future. As in any predictive method, the quality of past performance does not guarantee the quality of future performance. In this article, we fixed the search query terms after 2010 so as to directly compare our results with GFT, which has kept the same query terms since 2010; future application of ARGO may update search terms more frequently. ARGO can be easily generalized to any temporal and spatial scales for a variety of diseases or social events amenable to be tracked by Internet searches or services (3, 4, 8, 9, 29, 30, 38, 39). Further improvements in influenza prediction may come from combining multiple predictors constructed from disparate data sources (40). After the initial submission of this article in May 2015, Google announced that GFT would be discontinued and that their raw data would be made accessible to selected scientific teams. This announcement happened soon after the GFT team published a manuscript that proposed a new time series-based method for the (now discontinued) GFT engine (41). This new development makes our contribution timely and useful in providing a transparent method for disease tracking in the future.

Materials and Methods

All data used in this article are publicly available. Therefore, IRB approval is not needed.

Google Data. To avoid forward-looking information in our out-of-sample predictions, and to make the search term selection in our approach consistent with the main revision to GFT (14) immediately after the H1N1 pandemic, we obtained the highest-correlated terms to the CDC's ILI using Google Correlate (www.google.com/trends/correlate) for two different time periods. For the first time period (pre-H1N1 period), we inserted only CDC's ILI data from January 2004 to March 28, 2009 into Google Correlate, and used the resulting most highly correlated search terms as independent variables for our out-of-sample predictions for the time period April 4, 2009 through May 22, 2010. For the second time period (post-H1N1), we inserted only CDC's ILI data from January 2004 to May 22, 2010 into Google Correlate to select new search terms, as done in ref. 14. These last search terms were used as independent variables for all subsequent predictions presented in this work. Tables S3 and S4 show all query terms identified. For the pre-H1N1 period (the first time period), the terms from Google Correlate include spurious (or overfitted) terms like "march vacation" or "basketball standings," as discussed in ref. 15. However, Fig. S1 shows that these spurious terms were often not selected by ARGO, i.e., ARGO would give them zero weights, demonstrating its robustness. For the post-H1N1 time period, the updated query terms from Google Correlate include mostly flu-related terms (see Table S4). This suggests that spurious terms were "filtered out" by including off-season flu data. For the time period of March 28, 2015 up to the date of submission of this article, we acquired search frequencies for this set of query terms from Google Trends (www.google.com/trends; date of access: July 11, 2015) as Google Correlate only provides data up to March 28, 2015 at the time of writing this article.

Google Correlate standardizes the search volume of each query to have mean zero and SD 1 across time and contains data only from 2004 to March 2015. To make Google Correlate data compatible with Google Trends data, we linearly transformed the Google Correlate data to the same scale of 0–100 in our analysis. We used Google Correlate data up to its last available date, and then switched to Google Trends data afterward. This is indicated in Fig. 1 by different shades of the background. We used the latest version of GFT (fourth version, revised in October 2014) weekly estimates of ILI activity level as one of our comparison methods. GFT is available at www.google.org/flutrends/about (date of access: July 11, 2015).

CDC's Data. We use the weighted version of CDC's ILI activity level as the estimation target (available at gis.cdc.gov/grasp/fluview/fluportaldashboard.html; date of access: July 11, 2015). The weekly revisions of CDC's ILI are available at the CDC website for all recorded seasons (from week 40 of a given year to week 20 of the subsequent year). For example, ILI report revision at week 50 of season 2012–2013 is available at www.cdc.gov/flu/weekly/weeklyarchives2012-2013/data/senAllregt50.htm; ILI report revision at week 9 of season 2014–2015 is available at www.cdc.gov/flu/weekly/weeklyarchives2014-2015/data/senAllregt09.html.

Formulation of Our Model. Our model ARGO is motivated by a hidden Markov model. The logit-transformed CDC-reported ILI activity level $\{y_t\}$ is the intrinsic time series of interest. We impose an autoregressive model with lag N on it, which implies that the collection of vectors $\{y_{(t-N+1):t}\}_{t\geq N}$ is a Markov chain (this captures the clinical fact that flu lasts for a period, but not indefinitely). The vector of log-transformed normalized volume of Google search queries at time t, $\boldsymbol{X}_t$, depends only on the ILI activity at the same time, y_t (this follows the intuition that flu occurrence causes people to search flu-related information online). The Markovian property on block $y_{(t-N+1):t}$ leads to the (vector) hidden Markov model structure.

$$\begin{array}{ccccccc} y_{1:N} & \rightarrow & y_{2:(N+1)} & \rightarrow & \cdots & \rightarrow & y_{(T-N+1):T} \\ \downarrow & & \downarrow & & & & \downarrow \\ \boldsymbol{X}_N & & \boldsymbol{X}_{N+1} & & & & \boldsymbol{X}_T \end{array} \qquad [1]$$

Our formal mathematical assumptions are

(assumption 1) $y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \epsilon_t, \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

(assumption 2) $\boldsymbol{X}_t | y_t \sim \mathcal{N}_K(\boldsymbol{\mu}_x + y_t \boldsymbol{\beta}, \boldsymbol{Q})$

(assumption 3) conditional on y_t, $\boldsymbol{X}_t$ is independent of $\{y_l, \boldsymbol{X}_l : l \neq t\}$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)^\top, \boldsymbol{\mu}_x = (\mu_{x_1}, \mu_{x_2}, \ldots, \mu_{x_K})^\top$, and $\boldsymbol{Q}$ is the covariance matrix. To make the variables more normal, we transform the original ILI activity level p_t from $[0,1]$ to $\mathbb{R}$ using the logit function, obtaining the y_t, and transform the Google search volumes from $[0,100]$ to $\mathbb{R}$ using the log function, obtaining $\boldsymbol{X}_t$. The log function is appropriate because Google search frequencies usually have an exponential growth rate near peaks and are artificially scaled to $[0,100]$ by dividing the running maximum. Because Google Trends is in integer scale from 0 to 100, we add a small number $\delta = 0.5$ before the transformation to avoid taking the log of 0. The predictive distribution $f(y_t | y_{1:(t-1)}, \boldsymbol{X}_{1:t})$ is normal with mean linear in $y_{(t-N):(t-1)}$ and $\boldsymbol{X}_t$ and constant variance (see *Supporting Information*). This observation leads to Eq. **2**, which defines the ARGO model.

The ARGO Model. Let $y_t = \text{logit}(p_t)$ be the logit-transformed CDC's (weighted) ILI activity level p_t at time t, and $X_{i,t}$ the log-transformed Google search frequency of term i at time t. Our ARGO model is given by

$$y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \qquad [2]$$

where $\boldsymbol{X}_t$ can be thought of as the exogenous variables to time series $\{y_t\}$.

Parameter Estimation of ARGO Model. We chose $N = 52$ (weeks) to capture the within-year seasonality in ILI activity, and $K = 100$ (Google search terms) following the data availability from Google Correlate. Because we have more independent variables than the number of observations, the usual maximum likelihood estimate (ordinary least squares) method will fail. Therefore, we impose regularities for parameter estimation. In general we have three kinds of penalties, L_1 penalty (42), L_2 penalty (43), and a linear combination of L_1 and L_2 penalties (44). All parameters are dynamically trained every week with a 2-y (104-wk) rolling window.

In a given week, the goal is to find parameters μ_y, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{52})$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{100})$ that minimize

$$\sum_t \left(y_t - \mu_y - \sum_{j=1}^{52} \alpha_j y_{t-j} - \sum_{i=1}^{100} \beta_i X_{i,t} \right)^2 + \lambda_\alpha \|\boldsymbol{\alpha}\|_1 + \eta_\alpha \|\boldsymbol{\alpha}\|_2^2 + \lambda_\beta \|\boldsymbol{\beta}\|_1 + \eta_\beta \|\boldsymbol{\beta}\|_2^2 \quad [3]$$

where $\lambda_\alpha, \lambda_\beta, \eta_\alpha$, and η_β are hyperparameters. Ideally, we would like to use cross-validation to select all four hyperparameters. However, because we have only 104 training data points at a given week due to the 2-y moving window, the cross-validation result is highly noisy. Thus, we need to prespecify some of the hyperparameters. For model simplicity and sparsity, combining with the evidence seen from cross-validation, we set $\eta_\alpha = \eta_\beta = 0$, leading to L_1 penalization on both autoregressive and Google search terms. With the remaining λ_α and λ_β, the cross-validation results still have considerable variance. By the same sparsity and simplicity consideration, we further constrained $\lambda_\alpha = \lambda_\beta$. Therefore, the ARGO model we finally propose is Eq. **3** with constraint $\eta_\alpha = \eta_\beta = 0$ and $\lambda_\alpha = \lambda_\beta$. A detailed discussion of our specification of the hyperparameters is provided in *Supporting Information* (see Table S5).

Accuracy Metrics. The RMSE, MAE, and MAPE of estimator $\hat{p}$ to the target ILI activity level p are defined, respectively, as $\mathrm{RMSE}(\hat{p}_t, p_t) = [(1/n)\sum_{t=1}^{n}(\hat{p}_t - p_t)^2]^{1/2}$, $\mathrm{MAE}(\hat{p}_t, p_t) = (1/n)\sum_{t=1}^{n}|\hat{p}_t - p_t|$, and $\mathrm{MAPE}(\hat{p}_t, p_t) = (1/n)\sum_{t=1}^{n}|\hat{p}_t - p_t|/p_t$.

The correlation of estimator $\hat{p}$ to the target ILI activity level p is their sample correlation coefficient. The correlation of increment between $\hat{p}_t$ and p_t is defined as

$$\text{Corr. of increment}(\hat{p}_t, p_t) = \mathrm{Corr}(\hat{p}_t - \hat{p}_{t-1}, p_t - p_{t-1}).$$

The relative efficiency of estimator $\hat{p}^{(1)}$ to estimator $\hat{p}^{(2)}$ is $e(\hat{p}^{(1)}, \hat{p}^{(2)}) = \mathrm{MSE}^{(2)}_{\mathrm{true}}/\mathrm{MSE}^{(1)}_{\mathrm{true}}$, where $\mathrm{MSE}^{(i)}_{\mathrm{true}} = \mathbb{E}[(\hat{p}^{(i)} - p)^2]$, which can be estimated by

$$\hat{e}\left(\hat{p}^{(1)}, \hat{p}^{(2)}\right) = \frac{\mathrm{MSE}^{(2)}_{\mathrm{obs}}}{\mathrm{MSE}^{(1)}_{\mathrm{obs}}} \quad \text{where} \quad \mathrm{MSE}^{(i)}_{\mathrm{obs}} = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{p}^{(i)}_t - p_t\right)^2. \quad [4]$$

The 95% confidence interval can be constructed by the time series stationary bootstrap method (35), where the replicated time series of the error residual is generated using geometrically distributed random blocks with mean length 52 (which corresponds to 1 y). We obtain the basic bootstrap confidence interval for $\log\{e(\hat{p}^{(1)}, \hat{p}^{(2)})\}$ and then recover the original scale by exponentiation. The nonparametric bootstrap confidence interval takes the autocorrelation and cross-correlation of the errors into account, and is insensitive to the mean block length.

ACKNOWLEDGMENTS. S.C.K.'s research is supported in part by National Science Foundation Grant DMS-1510446.

1. Ginsberg J, et al. (2009) Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.
2. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA (2008) Using Internet searches for influenza surveillance. *Clin Infect Dis* 47(11):1443–1448.
3. Yuan Q, et al. (2013) Monitoring influenza epidemics in china with search query from baidu. *PLoS One* 8(5):e64323.
4. Paul MJ, Dredze M, Broniatowski D (2014) Twitter improves influenza forecasting. *PLOS Curr Outbreaks* 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.
5. McIver DJ, Brownstein JS (2014) Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLOS Comput Biol* 10(4):e1003581.
6. Santillana M, Nsoesie EO, Mekaru SR, Scales D, Brownstein JS (2014) Using clinicians' search query data to monitor influenza epidemics. *Clin Infect Dis* 59(10):1446–1450.
7. Wesolowski A, et al. (2014) Commentary: Containing the Ebola outbreak–the potential and challenge of mobile network data. *PLOS Curr Outbreaks* 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e.
8. Chan EH, Sahai V, Conrad C, Brownstein JS (2011) Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis* 5(5):e1206.
9. Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google trends. *Sci Rep* 3:1684.
10. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J Comput Sci* 2(1):1–8.
11. Wu L, Brynjolfsson E (2015) The future of prediction: How Google searches foreshadow housing prices and sales. *Economic Analysis of the Digital Economy*, eds Goldfarb A, Greenstein SM, Tucker CE (Univ Chicago Press, Chicago), pp 89–118.
12. Helft M (November 11, 2008) Google uses searches to track flu's spread. *NY Times*. Available at www.nytimes.com/2008/11/12/technology/internet/12flu.html?_r=0#. Accessed July 11, 2015.
13. Butler D (2013) When Google got flu wrong. *Nature* 494(7436):155–156.
14. Cook S, Conrad C, Fowlkes AL, Mohebbi MH (2011) Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS One* 6(8):e23610.
15. Lazer D, Kennedy R, King G, Vespignani A (2014) Big data. The parable of Google Flu: Traps in big data analysis. *Science* 343(6176):1203–1205.
16. Santillana M, Zhang DW, Althouse BM, Ayers JW (2014) What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am J Prev Med* 47(3):341–347.
17. World Health Organization (2014) Influenza (seasonal) (World Health Org, Geneva), Fact Sheet 211.
18. Shaman J, Karspeck A (2012) Forecasting seasonal outbreaks of influenza. *Proc Natl Acad Sci USA* 109(50):20425–20430.
19. Lipsitch M, Finelli L, Heffernan RT, Leung GM, Redd SC; 2009 H1n1 Surveillance Group (2011) Improving the evidence base for decision making during a pandemic: The example of 2009 influenza A/H1N1. *Biosecur Bioterror* 9(2):89–115.
20. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV (2014) A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza Other Respi Viruses* 8(3):309–316.
21. Chretien JP, George D, Shaman J, Chitale RA, McKenzie FE (2014) Influenza forecasting in human populations: A scoping review. *PLoS One* 9(4):e94130.
22. Nsoesie E, Mararthe M, Brownstein J (2013) Forecasting peaks of seasonal influenza epidemics. *PLoS Curr* 5:5.
23. Soebiyanto RP, Adimi F, Kiang RK (2010) Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PLoS One* 5(3):e9450.
24. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M (2013) Real-time influenza forecasts during the 2012–2013 season. *Nat Commun* 4(2837):2837.
25. Yang W, Lipsitch M, Shaman J (2015) Inference of seasonal and pandemic influenza transmission dynamics. *Proc Natl Acad Sci USA* 112(9):2723–2728.
26. Paolotti D, et al. (2014) Web-based participatory surveillance of infectious diseases: The Influenzanet participatory surveillance experience. *Clin Microbiol Infect* 20(1):17–21.
27. Dalton C, et al. (2009) Flutracking: A weekly australian community online survey of influenza-like illness in 2006, 2007 and 2008. *Commun Dis Intell Q Rep* 33(3):316–322.
28. Smolinski MS, et al. (2015) Flu near you: Crowdsourced symptom reporting spanning two influenza seasons. *Am J Public Health* 105(10):2124–2130.
29. Althouse BM, Ng YY, Cummings DA (2011) Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis* 5(8):e1258.
30. Ocampo AJ, Chunara R, Brownstein JS (2013) Using search queries for malaria surveillance, Thailand. *Malar J* 12(1):390.
31. Scarpino SV, Dimitrov NB, Meyers LA (2012) Optimizing provider recruitment for influenza surveillance networks. *PLOS Comput Biol* 8(4):e1002472.
32. Davidson MW, Haim DA, Radin JM (2015) Using networks to combine "big data" and traditional surveillance to improve influenza predictions. *Sci Rep* 5:8154.
33. Burkom HS, Murphy SP, Shmueli G (2007) Automated time series forecasting for biosurveillance. *Stat Med* 26(22):4202–4218.
34. Everitt BS, Skrondal A (2002) *The Cambridge Dictionary of Statistics* (Cambridge Univ Press, Cambridge, UK).
35. Politis DN, Romano JP (1994) The stationary bootstrap. *J Am Stat Assoc* 89(428):1303–1313.
36. Tsukayama H (October 13, 2014) Google is testing live-video medical advice. *Washington Post*. Available at https://www.washingtonpost.com/news/the-switch/wp/2014/10/13/google-is-testing-live-video-medical-advice/. Accessed April 20, 2015.
37. Gianatasio D (November 10, 2014) How this agency cleverly stopped people from googling their medical symptoms: The right ads at the right time. *Adweek*. Available at www.adweek.com/adfreak/how-agency-cleverly-stopped-people-googling-their-medical-symptoms-161331. Accessed April 20, 2015.
38. Yang AC, Tsai SJ, Huang NE, Peng CK (2011) Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004–2009. *J Affect Disord* 132(1-2):179–184.
39. Cavazos-Rehg PA, et al. (2015) Monitoring of non-cigarette tobacco use using Google Trends. *Tob Control* 24(3):249–255.
40. Santillana M, et al. (2015) Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol* 11(10):e1004513.
41. Lampos V, Miller AC, Crossan S, Stefansen C (2015) Advances in nowcasting influenza-like illness rates using search query logs. *Sci Rep* 5:12760.
42. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc, B* 58(1):267–288.
43. Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67.
44. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 67(2):301–320.