scientific reports

Check for updates

OPEN

Use Internet search data to accurately track state level influenza epidemics

Shihao Yang[1,4], Shaoyang Ning[2,4] & S. C. Kou[3]✉

For epidemics control and prevention, timely insights of potential hot spots are invaluable. Alternative to traditional epidemic surveillance, which often lags behind real time by weeks, big data from the Internet provide important information of the current epidemic trends. Here we present a methodology, ARGOX (Augmented Regression with GOogle data CROSS space), for accurate real-time tracking of state-level influenza epidemics in the United States. ARGOX combines Internet search data at the national, regional and state levels with traditional influenza surveillance data from the Centers for Disease Control and Prevention, and accounts for both the spatial correlation structure of state-level influenza activities and the evolution of people's Internet search pattern. ARGOX achieves on average 28% error reduction over the best alternative for real-time state-level influenza estimation for 2014 to 2020. ARGOX is robust and reliable and can be potentially applied to track county- and city-level influenza activity and other infectious diseases.

Each year in the United States (US) alone, the seasonal influenza (flu) epidemics may claim up to 61,000 deaths[1]. Quick responses and preventive actions to changes in flu epidemics rely on timely and accurate information on the current flu severity. In particular, due to the geographically varying timing and intensity of disease epidemics, most public health decisions and executive orders for disease control and prevention are made at the state or local level. Accurate *real-time* flu tracking at the state/local level is thus indispensable. Traditional flu surveillance, such as those conducted by the US Centers for Disease Control and Prevention (CDC), however, often lags behind real time by up to two weeks. Here we propose a statistically principled, self-coherent framework ARGOX (Augmented Regression with GOogle data CROSS space) for real-time, accurate flu estimation at the state level. ARGOX efficiently combines publicly available Internet search data with traditional flu surveillance data and coherently utilizes the data from multiple geographical resolutions (national, regional, and state levels).

For the last two decades, tracking of flu activities in the US mainly relies on traditional surveillance systems, such as the US Outpatient Influenza-like Illness Surveillance Network (ILINet) by the CDC. Through the ILINet, thousands of healthcare providers across the US report their numbers of outpatients with Influenza-like Illness (ILI) to CDC on a weekly basis. CDC then aggregates the data and publishes the ILI percentages (%ILI, i.e., the percentages of outpatients with ILI) in its weekly reports at the national and regional levels (there are ten Health and Human Services (HHS) regions in the US, each consisting of multiple states). Starting from 2017, the state-level %ILI reports became available for selected states, and in late 2018 the state-level %ILI reports became available for all states except Florida. Owing to the time for administrative processing and aggregation, CDC's flu reports typically lag behind real time for up to 2 weeks and are also subject to subsequent revisions. Such delay and inaccuracy are far from optimal for public health decision making, especially in the face of epidemic outbreaks or pandemics.

Big data from the Internet offer the potential of real-time tracking of public health or social events. In fact, valuable insights have been gained from the Internet data about current social and economical status of a nation, including epidemic outbreaks[2,3] and macro economic indices[4,5]. Furthermore, real-time data from the Internet could also offer insights at the regional, state, or local level. Examples include foreshadowing state-wise housing price index in the US[6], estimating New York City flu activity[7], estimating real-time county-level unreported COVID-19 severity in the US[8] among others. For epidemic surveillance, such real-time digital data at local level can be potentially used to provide insights for early epidemic hot-spot detection and timely public health resource allocation (e.g. vaccine campaigns) as well as to gather information on the overall disease prevalence.

[1]Georgia Institute of Technology, H. Milton Stewart School of Industrial and Systems Engineering, Atlanta, GA 30332, USA. [2]Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, USA. [3]Department of Statistics, Harvard University, Cambridge, MA 02138, USA. [4]These authors contributed equally: Shihao Yang and Shaoyang Ning. ✉email: kou@stat.harvard.edu

Various models have been proposed to utilize Internet data, especially Internet search volume data, to provide real-time estimation of the current flu activity at the national level. Google Flu Trends (GFT), as one of the early examples, uses the search frequency of selected query terms from Google to estimate the real-time %ILI[2]. Recent models on combining CDC's surveillance data with Internet-derived data appear to work well at the national level[9,10]. Other methods, primarily targeting national flu epidemics, were also developed based on traditional epidemiology data and mechanistic models, such as susceptible-infectious-recovered-susceptible model with ensemble adjustment Kalman filter (SIRS-EAKF)[7,11–14].

Compared to estimation at the national level, %ILI estimation at the regional or state level is much more challenging, as documented by FluSight, the CDC-sponsored Flu Prediction Initiative[15]. Due to factors like geographical proximity, transportation connectivity, and public health communication, the state-wise epidemic spread exhibits strong spatial structure. However, many digital flu estimation methods[12,16,17], including GFT, ignore such spatial structure and apply the same national-level method to regional, and/or state-level flu estimation. A few attempts have been made to incorporate the geographical dependence structure. For example, Ref.[18] studied the estimation of ILI activity in the boroughs and neighborhoods of New York City using a traditional epidemiological mechanistic SIRS-network model without Internet search data, where the dynamic system is multivariate with explicit parameters to characterize traffic between locales, and concluded that the spatial network is helpful at the borough scale but not at the neighborhood scale; Ref.[19] utilized an ordinary-least-squares-based network model to improve upon the output of GFT, where a weighted average of GFT from all regions is produced as an network-enhanced final estimate for each individual region; Ref.[20] employs a multi-task nonlinear regression method for regional %ILI estimation, where a Multi-Task Gaussian Process is proposed to regress each region's %ILI on the corresponding Google search data; Ref.[21] uses a network approach for %ILI estimation in a few selected states, where they first built a stand-alone state %ILI prediction based on the ARGO method[9], and then obtained a multiple linear regression prediction for a given state's %ILI from other states' %ILI, and finally a winner-takes-all approach was adopted for each state separately to select one of the two approaches; Ref.[22] shows that careful spatial structure modeling can lead to much improved accuracy in %ILI estimation at the regional level. An ensemble approach has also been proposed to utilize the output of a variety of available models to achieve better accuracy[23].

Nevertheless, at the state level, no existing methods provide real-time flu tracking with satisfactory accuracy and reliability. (i) There are no unified approaches to combine multi-resolution and cross-state information effectively to provide national, regional and state-level estimates within the same framework. (ii) Few existing models can outperform a naive estimation method, which, for each state, without any modeling effort, simply uses CDC's reported %ILI from the previous week as the %ILI estimate for the current week (see Fig. 1 for an illustration). This would be particularly worrisome for public health officials who rely on accurate flu estimation at the local level to make informed decisions.

In this article we introduce ARGOX, a unified spatial-temporal statistical framework that combines multi-resolution, multi-source information to provide real-time state-level %ILI estimates while maintaining coherency with %ILI estimation at the regional and national levels (in a cascading fashion). To illustrate the underlying idea of ARGOX, let us take estimating the %ILI in California as an example. The real-time Google search volumes for flu-related terms like "flu symptoms" or "flu duration" from California reflect its current state-level flu intensity to some extent. In addition, California's flu epidemics could be highly correlated with flu epidemics of nearby states such as Oregon and Nevada, as well as with geographically distant but transportation-wise well-connected states such as Illinois. California's current flu situation may also depend heavily on the recent trends of flu epidemics, in particular, the overall national and Pacific-west regional flu trends. Taken these considerations together, ARGOX operates in two steps: at the first step, it extracts Google search information of most relevant query terms at three geographical resolutions—national, regional, and state levels; at the second step, the cross-time, cross-resolution, cross-state information mentioned above, together with Internet-extracted information, is integrated through careful modeling of their temporal-spatial dependence structure, which yields significant enhancement in the estimation accuracy.

ARGOX was inspired in part by Refs.[9] and[22], which studied the %ILI estimation at the national and regional levels respectively. Although the methods introduced in Refs.[9] and[22] worked well for flu-tracking at the national or regional level, these methods cannot be directly applied to accurately track state-level %ILI for a number of reasons, which are specifically solved by ARGOX. In particular, ARGOX addresses the following issues: (i) how to *simultaneously* provide accurate, real-time flu tracking at the higher-resolution level for all 51 US states (district/city), as opposed to only at the national or regional level, (ii) how to effectively combine multi-resolution information from the national, regional and state levels for state %ILI estimation, i.e., how to leverage the information from the national and regional levels, in addition to the information at a particular state, for the %ILI estimation at a given state; (iii) how to solve the challenge of declining quality of Internet search data at higher geographical resolution, since compared to the Internet search data at the national level, the state-level Internet search data are of much inferior quality; (iv) how to determine when to borrow information from other states for the %ILI estimation at a given state and when not to borrow, since the states have varying degree of connections—for a state well connected with others, borrowing information probably would help its %ILI estimation, but for a state not (geographically or epidemically) well connected with others, using information from other states might hurt (as opposed to help) its %ILI estimation; and (v) how to model the correlation structure of %ILI across the "well-connected" states to effectively borrow such cross-state information to improve prediction accuracy. ARGOX, therefore, significantly advances accurate flu tracking from the national and regional levels to the state level, which could help public health officials make much more informed decisions.

Through the ARGOX framework, the state-level flu activity estimates are produced in a unified and coherent way with the national and regional estimates. ARGOX achieves on average 28% mean squared error (MSE) reduction compared to the best alternative and shows strong advantages over all benchmark methods, including

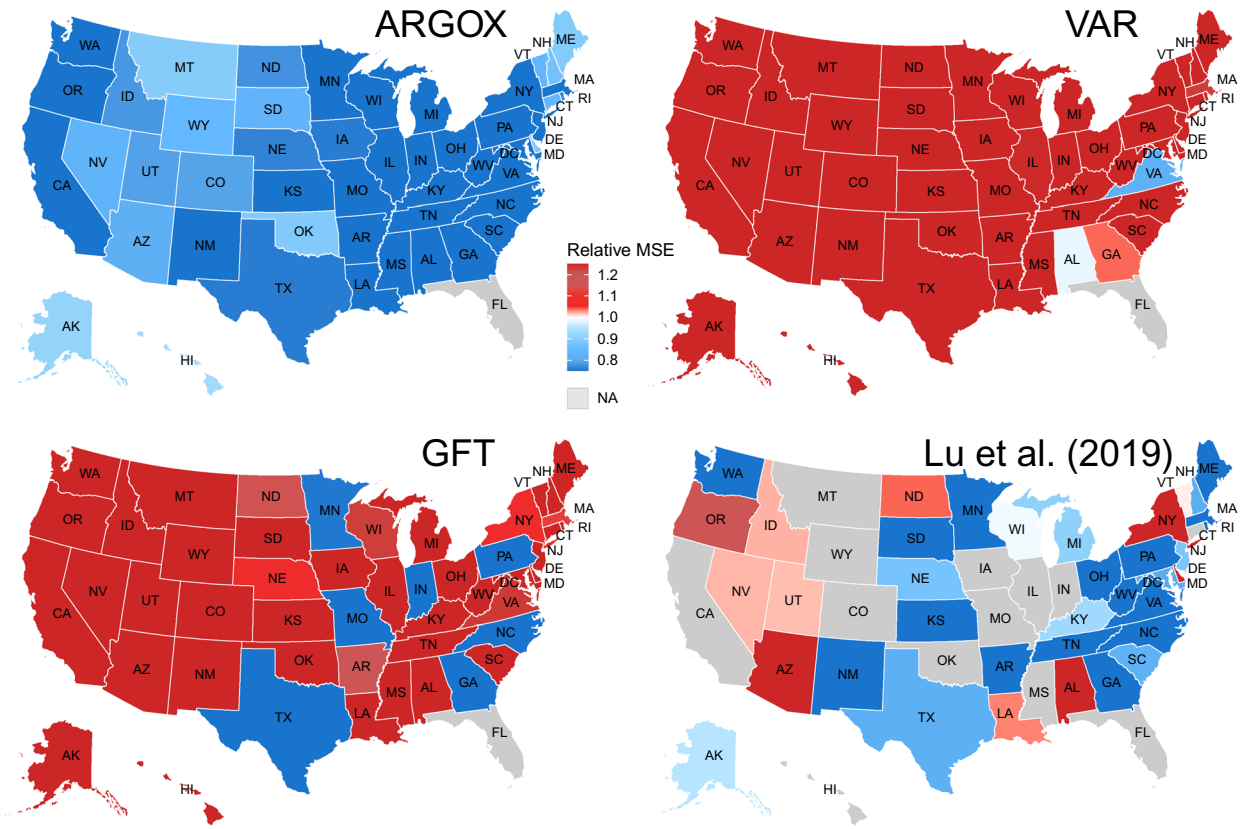

Figure 1. State-by-state heatmap of relative Mean Squared Error of ARGOX, VAR, GFT, and Lu et al.[21] to the naive method. The relative MSE is the ratio of the MSE of a given method to that of the naive method. Blue color means smaller MSE (i.e., better performance) than the naive method; red color means larger MSE (i.e., worse performance) than the naive method; grey color means result not available. ARGOX with all blue colors uniformly dominates the naive method, while mixed colors in the rest of the plots show that VAR, GFT, and Lu et al.[21] were worse than the naive method in a large proportion of states. ARGOX and VAR are evaluated for the whole period of Oct 11, 2014 to March 21, 2020; GFT is evaluated for the period of Oct 11, 2014 to August 15, 2015 due to GFT data availability; Lu et al.[21] is evaluated from Oct 11, 2014 to May 14, 2017 due to its availability.The figure was generated by the programming language `R`. The US maps were drawn based on the publicly available `R` package `urbnmapr`, which uses map shapefiles from the US Census Bureau (https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html).

GFT, time-series-based vector autoregression (VAR), and another recent Internet-search-based method developed in Lu et al.[21]. ARGOX achieves its high estimation accuracy through a few features: (i) it automatically selects the most relevant search queries to address the problem of lower-quality Google search information at state or regional level; (ii) it incorporates time-series momentum of flu activity; (iii) it pools the multi-resolution information by combining the national-, regional-, and state-level data; (iv) it explicitly models the spatial correlation structure of state-level flu activities; (v) it adapts to the evolution in people's search pattern, Google's search engine algorithms, epidemic trends, and other time-varying factors[24] with a dynamic two-year rolling window for training; and (vi) it achieves selective pooling of most immediately relevant information for a handful of stand-alone states (details in Methods).

Results

We conducted retrospective estimation of the weekly %ILI at the US state level—50 states excluding Florida whose ILI data is not available from CDC, plus Washington DC and New York City—for the period of Oct 11, 2014 to March 21, 2020. For each week during this period, we only used the data that would have been available—the historical CDC's ILI reports up to the previous week and Google search data up to the current week—to estimate state-level %ILI of the current week. To evaluate the accuracy of our estimation, we compared the estimates with the actual %ILI released by CDC weeks later in multiple metrics, including the mean squared error (MSE), the mean absolute error (MAE), and the correlation with the actual %ILI (detailed in Methods). We also compared the performance of ARGOX with several benchmark methods, including (a) GFT (last estimate available: the week ending on August 15, 2015), (b) estimates by the lag-1 vector autoregressive model (VAR model), (c) the naive estimates, which for each state without any modeling effort simply use CDC's reported %ILI of the previous week as the estimate for the current week, and (d) a recent Internet-search-based state-level estimation model developed in Lu et al.[21]. As ARGOX uses a two-year training window, for fair comparison we keep the

	Whole period	'14–'15	'15–'16	'16–'17	'17–'18	'18–'19	'19–'20
MSE							
ARGOX	**0.340**	**0.488**	**0.217**	**0.421**	**0.445**	**0.301**	**0.835**
VAR	1.556	1.606	0.819	1.629	2.615	1.277	3.747
GFT	–	2.186	–	–	–	–	–
naive	0.473	0.665	0.257	0.551	0.779	0.434	1.150
MAE							
ARGOX	**0.340**	**0.380**	**0.311**	**0.407**	**0.423**	**0.359**	**0.580**
VAR	0.597	0.633	0.516	0.693	0.825	0.668	1.058
GFT	–	0.944	–	–	–	–	–
naive	0.393	0.435	0.340	0.464	0.547	0.443	0.696
Correlation							
ARGOX	**0.949**	**0.914**	**0.832**	**0.875**	**0.937**	**0.921**	**0.902**
VAR	0.857	0.806	0.693	0.752	0.854	0.813	0.772
GFT	–	0.904	–	–	–	–	–
naive	0.931	0.885	0.803	0.842	0.902	0.890	0.874

Table 1. Comparison of different methods for state-level %ILI estimation. The evaluation is based on the average of 51 US states/district/city in multiple periods and multiple metrics. The MSE, MAE, and correlation are reported. The method with the best performance is highlighted in boldface for each metric in each period. Methods considered here include ARGOX, VAR, GFT, and the naive method. All comparisons are conducted on the original scale of CDC's %ILI. The whole period is Oct 11, 2014 to March 21, 2020. Columns 3 to 8 correspond to the regular flu seasons (week 40 to week 20 next year, defined by CDC's Morbidity and Mortality Weekly Report; 19'-20' season is up to March 21, 2020).

	Overall ('14–'17)	'14–'15	'15–'16	'16–'17
MSE				
ARGOX	**0.269**	**0.406**	**0.163**	**0.339**
Lu et al.[21]	0.418	0.467	0.528	0.544
Correlation				
ARGOX	**0.919**	**0.914**	**0.836**	**0.890**
Lu et al.[21]	0.912	0.912	0.808	0.858

Table 2. Comparison of ARGOX to the method of Lu et al.[21] for state-level %ILI estimation. The numbers of Lu et al. are directly obtained from Ref.[21], which reported its estimation results of 37 states over three flu seasons: 2014–2017. For fair comparison, the result of ARGOX is restricted to the same 37 states and the same time period to match Ref.[21]. The method with best performance for each metric in each period is highlighted in boldface.

same two-year training window for VAR as well. Also for fair comparison, the numerical results of the method of Lu et al.[21] were directly quoted from the article (which reported results through May 14, 2017).

Table 1 summarizes the overall results of ARGOX, VAR, GFT, and the naive method, averaging over the 51 states/district/city for the whole period of 2014 to 2020 (up to March 21, 2020). Table 2 summarizes the comparison between ARGOX and the method of Lu et al.[21], averaging over 37 states for the period of 2014 to 2017. We need to compare ARGOX with Lu et al.[21] in a separate Table 2 because the results of Lu et al.[21] are only available for 37 states and only for the period of 2014 to 2017.

Table 1 shows that ARGOX gives the leading performance uniformly through all flu seasons in all metrics. Particularly, ARGOX achieves up to 28% error reduction in MSE and about 15 % error reduction in MAE compared to the best alternative in the whole period. ARGOX also keeps consistent season-by-season performance, with at least 15% error reduction in MSE compared to the best alternative method in every season from 2014 to 2019. For the 2019–2020 flu season with the (onset of) COVID-19 pandemic, ARGOX's accuracy still maintains. Compared with other benchmarks, ARGOX's advantages in state-level flu tracking are substantial. VAR and GFT fail to outperform the naive method in any of the evaluated flu seasons; both methods have MSE two or three times larger than the naive method. Table 2 shows that ARGOX also uniformly outperforms Lu et al.[21] in all three seasons when the benchmark is available. More detailed results comparing ARGOX with the benchmarks can be found in the Supplementary Information (Table S4). The advantage of ARGOX over the method of Lu et al.[21] could be attributed to (i) incorporating multi-resolution information in the modeling that pools national, regional and state-level information together, (ii) capturing the spatio-temporal information using one joint statistically structured variance-covariance matrix as opposed to ad hoc regression of each individual state's %ILI on other states', and (iii) using a statistically principled and interpretable method to dichotomously

select between either joint modeling for statistically "connected" states or stand-alone modeling for statistically/geographically "disconnected" states.

Among all the methods that we numerically compared, ARGOX is the only one that uniformly outperforms the naive method in all 51 states/district/city in terms of MSE for the whole period of evaluation. Figure 1 plots the state-by-state estimation results, showing the ratio of the MSE of a given method to the MSE of the naive method. The results of four methods are plotted: ARGOX, VAR, GFT, and Lu et al.[21] For each state, a blue color means that the MSE of a method is smaller (better) than the MSE of the naive method for that state, and a red color means the MSE of the method is larger (worse) than the MSE of the naive method. Darker blue means more advantage over the naive method, while darker red means more disadvantage than the naive method. It is noteworthy that ARGOX with all blue colors is the only method that gives uniformly better performance than the native method across all states. All other methods in comparison fail to do so for a large portion of the states investigated. Note that the naive method provides a model-free baseline benchmark that solely relies on information from CDC's flu reports. Therefore, ARGOX is the only method that effectively utilizes the Internet data to uniformly improve flu tracking from the traditional surveillance system, indicating ARGOX's reliability and adaptability. With its universally enhanced accuracy over the alternative methods for real-time state-level flu situation estimate, it appears that ARGOX could aid timely, proper public health decision making for the local monitoring and control of the disease.

Detailed numerical results for each state and for each flu season are reported in Tables S5–S55 and the figures in Supporting Information (SI), where ARGOX holds lead over other methods in the vast majority of the cases, further revealing its robustness over geographical and seasonal variability in flu epidemics.

In addition to the point estimate, ARGOX also provides 95% confidence intervals for each week's estimates. For the entire period from 2014 to 2020, over all 51 states/district/city, the intervals provided by ARGOX successfully cover the actual %ILI in 92.5% of the cases (Table S1), which is close to the nominal 95%, demonstrating ARGOX's accurate uncertainty quantification.

Discussion

ARGOX effectively combines state-, regional-, and national-level publicly available data from Google searches and CDC's traditional flu surveillance system. It incorporates geographical and temporal correlation of flu activities to provide accurate, reliable real-time flu tracking at the state level. Across all the available states, ARGOX outperforms time-series-based benchmark models, GFT, and the method of Lu et al.[21] ARGOX's weekly %ILI estimations are accompanied by reliable interval estimates as a measure for uncertainty. The state-level real-time tracking of flu epidemics by ARGOX could help public health officials and the general public to make more informed decisions to control and prevent the flu epidemics at the state or local levels. In particular, with the real-time estimates of flu activities by ARGOX in their home states and neighboring states, local public health officials could make more proper and timely decisions on the allocation of relevant resources, such as vaccines, hospitalization, medical equipment, personnel, etc. Also, informed with the current local flu situation provided by ARGOX, the general public could take necessary measures accordingly, such as taking the flu shot, social distancing, and mask wearing to reduce the risk of contracting flu; knowing the real-time flu severity at other states could help the general public make travel decisions and plan/arrange care for relatives and friends. More discussion on the usefulness of influenza forecasts to public health decision making can be found in Ref.[25] and Ref.[23].

ARGOX's adaptive pooling of the most-relevant information among the 51 US states/district/city plays an important role in its performance. To avoid the possibility of overfitting, a structured covariance matrix on the %ILI increments is utilized. Such structured dynamic modeling of the cross-state covariance serves to capture the ever-changing geographic spread pattern of the flu. It aggregates state-to-state, time-varying connectivity factors such as commuting traffic, airline frequency, geographic proximity, and climatic patterns. The utilization of cross-state correlation also helps pool information from different states, regions and the entire nation in addition to the information at a given state. The pooling from national and regional level estimates incorporates the shared seasonality component in flu trends across all the states, which further helps reduce the risk of overfitting.

ARGOX operates in two steps: the first step extracts Internet search information at the state level, and the second step enhances the estimates using cross-state and cross-resolution information (detailed in Methods). Such two-step design of ARGOX has broad applicability. With the general availability of ubiquitous Internet search data, ARGOX's two-step framework could be flexibly adapted to track flu activities at even higher resolutions, such as county or city levels, when such weekly %ILI data become available. In addition, the first step could be substituted by other models or include other data sources, while the second step remains adaptable for multi-resolution spatial-temporal boosting. A wide spectrum of flu estimation models, including susceptible-infectious-recovered-susceptible model[7], empirical Bayes method[16], Wisdom-of-crowds forecast[17], or ensemble of them[26] can be fitted into the cross-state boosting step (the second step) of ARGOX.

Like all big-data-based models, our result has certain limitations. ARGOX's accuracy depends on the reliability of its inputs—Google Trends data and historical %ILI data from CDC. Google Trends data have increasing amount of missing data and zero counts as the resolution goes from national to regional and state levels (Table S3). Such degeneracy in data quality is a challenge for high-resolution inference. Google search information could also be sensitive to media coverage[27–29]. Furthermore, Google search data may only be representative of the search interests among Google users rather than the entire population. In states with less Internet penetration, such Google search data may be less predictive of the overall %ILI. The L_1 penalty and the dynamic training of ARGOX aims to correct for the sparsity, over-shooting, and representative issues of Google data, where only the most relevant search terms to %ILI estimation are selected at each state's level. Models to further alleviate or eliminate the bias in Internet search data (e.g. by incorporating data on media coverage intensity) could be an interesting future direction. In addition, we should be aware that our estimation target, the CDC's %ILI, is

only a proxy for the true flu incidence in the population, as it's calculated from a sample of outpatient visits with influenza-like symptoms. The reported %ILI at the state level could have (i) high noise due to its limited sample size, (ii) subsequent revision when healthcare providers update their information, and (iii) bias towards those with easy healthcare access. Nevertheless, accurate estimation of CDC's %ILI at the state level is valuable for optimizing resource allocations. More detailed discussion about the importance of alternative indicators for flu incidence in the population can be found in Ref.[30–32].

ARGOX is accurate, reliable, flexible and generalizable, making it adaptable to other spatial and temporal resolutions for tracking or forecasting other diseases and social/economic events that leave traces on people's Internet activity records. The ARGOX framework can be potentially adapted for COVID-19 tracking by incorporating additional coronavirus-related query terms at city, state, regional, and national level[33]. With the current development of COVID-19 pandemic, it is likely that the coronavirus would come back in the future winters. In light of this, accurate localized tracking of epidemic activity has become more important than ever before.

Methods

CDC's ILINet data. Every Friday, CDC releases a report of %ILI for the previous week, which gives the percent of outpatient visits with influenza-like illness for the whole nation, each HHS region, each state (except Florida), Washington DC, and New York City (separated from New York State) (http://www.cdc.gov/flu/weekly/overview.htm). CDC also revises the initial report numbers in the subsequent weeks when more information becomes available (gis.cdc.gov/grasp/fluview/fluportaldashboard.html). Consequently, CDC's %ILI data lag behind real-time for up to 2 weeks and are less accurate for more recent weeks. CDC's %ILI data for this study were downloaded on Mar 27, 2020.

Google data. The Internet search volume data from Google are publicly available through Google Trends (trends.google.com). A user can specify the desired query term, geographical location, and time frame on Google Trends; the website then will return a (weekly) time series in integer values from 0 to 100, which corresponds to the normalized search volume of the query term within the specified time frame, where 100 represents the historical maximum, and 0 represents missing data due to inadequate search intensity. This integer-valued time series from Google Trends is based on sampling Google's raw search logs.

The search query terms that we use are based on previous work for national and regional flu estimation[9,22]. We also included several additional queries and topics in this study, which were obtained from "Related queries" and "Related topics" on the Google Trends website when searching for flu related information. Table S2 in the Supplementary Information lists these search terms.

As one benchmark, we downloaded the discontinued Google Flu Trends (GFT) data (https://www.google.org/flutrends/about/data/flu/us/data.txt). GFT has national, regional, and state-level prediction for the weekly %ILI from Jan 1, 2004 to August 9, 2015.

Google search data may only be representative of the search interests among Google users rather than the entire population. ARGOX attempts to correct for such potential bias in the modeling.

Regional-Enrichment of state-level Google search data. Google Trends provides (normalized) search volume data at both national and state levels. However, for the state-level data, there is a high level of sparsity (i.e., zero observations) among the returned integer-valued time series (see Table S3). These zeros, which correspond to missing data due to inadequate search intensity, significantly lower the data quality at the state level (compared to the national level), which in turn severely reduces the prediction accuracy at the state level. To enhance the predictive power of state-level Google data, we use a simple approach to borrow information from the regional level. First, we reconstruct regional-level search frequency for each region in the US by weighting the state-level search frequencies within a given region, where the weights are proportional to the state's population. Second, instead of using the state-level Google Trends time-series, for each search term, we use a weighted average of the state-level search frequency (2/3 weight) and the regional-level search frequency (1/3 weight) as the input for state-level %ILI estimation. We carry out this regional-enrichment process for all states/district/city, except seven states—Hawaii (HI), Alaska (AK), Vermont (VT), Montana (MT), North Dakota (ND), Maine (ME), and South Dakota (SD)—because these seven states are modeled with a separate stand-alone model (as detailed in the following sections). For these seven states, the raw Google Trends state-level times series, not the regional-enriched time series, are used as input.

Evaluation metrics. We use three metrics to evaluate the accuracy of an estimate against the actual %ILI released by CDC: the mean squared error (MSE), the mean absolute error (MAE), and the Pearson correlation (Correlation). MSE between an estimate $\hat{p}_t$ and the true value p_t over period $t = 1, \ldots, T$ is $\frac{1}{T}\sum_{t=1}^{T}\left(\hat{p}_t - p_t\right)^2$. MAE between an estimate $\hat{p}_t$ and the true value p_t over period $t = 1, \ldots, T$ is $\frac{1}{T}\sum_{t=1}^{T}\left|\hat{p}_t - p_t\right|$. Correlation is the Pearson correlation coefficient between $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_T)$ and $\boldsymbol{p} = (p_1, \ldots, p_T)$.

Prediction model of ARGOX. ARGOX operates in two steps: the first step extracts Internet search information at the state level, and the second step enhances the estimates using cross-state and cross-resolution information.

At the second step, we take a dichotomous approach for the 51 US states/district/city (50 states except Florida, which does not have %ILI data, plus Washington DC and New York City). We set apart seven states: HI, AK, VT, MT, ND, ME, and SD. The first two (HI and AK) are geographically separated from the contiguous US. The last five (VT, MT, ND, ME, and SD) are the states that have the lowest multiple correlations (a.k.a. the R) in %ILI to the %ILI of the entire nation, the %ILI of the other states, and the %ILI of the other regions (detailed calculation

method is given in Supplementary Information). A low multiple correlation of a state implies that the state's flu activity is not well correlated with other states' or other regions'. For these seven states, due to either the geological discontinuity or the low multiple correlation, it is not clear if using information cross the other states or other regions can help the state-level %ILI estimation. Therefore, we adopt the dichotomous approach: For the 44 states/district/city (the vast majority), we apply a joint estimation approach at the second step to enhance the state-level %ILI estimation by using all information, including information from other states and other regions; for the above-mentioned seven states, we use a stand-alone estimation approach at the second step to enhance the %ILI estimation (not using information from other states and regions). The two steps of ARGOX are detailed below.

First step: extracting Internet search information at the state level. This step concerns extracting Google search information at each state. In particular, for a given state/district/city $m, m = 1, \ldots, 51$, let $X_{i,t,m}$ be the logarithm of 1 plus the state-level Google Trends data of search term i at week t (note: 1 is added to each state-level Google Trends data point to avoid taking logarithm of zero); let $y_{t,m}$ be the logit-transformation of CDC's %ILI at time t for state m. To estimate $y_{T,m}$, an L_1 regularized linear estimator is used in the first step based on the vector $\boldsymbol{X}_{T,m} = (x_{i,T,m})$:

$$\hat{y}_{T,m} = \hat{\beta}_{0,m} + \boldsymbol{X}_{T,m}^{\intercal}\hat{\boldsymbol{\beta}}_m,$$

where the coefficients $(\hat{\beta}_{0,m}, \hat{\boldsymbol{\beta}}_m)$ are obtained via

$$\underset{\beta_{0,m},\boldsymbol{\beta}_m}{\operatorname{argmin}} \sum_{t=T-N}^{T-1} \left(y_{t,m} - \beta_{0,m} - \boldsymbol{X}_{t,m}^{\intercal}\boldsymbol{\beta}_m\right)^2 + \lambda\|\boldsymbol{\beta}_m\|_1. \tag{1}$$

We set $N = 104$, i.e., a two-year window, as recommended in previous studies[9,22,24]. We set λ through cross-validation.

In addition, we obtain an accurate estimate $\hat{p}_T^{nat}$ for the national %ILI by using the ARGO method[9], which uses national level Google search data. We also obtain an estimate $(\hat{p}_{T,1}^{reg}, \ldots, \hat{p}_{T,10}^{reg})$ for the ten HHS regional %ILI by the first step of ARGO2 method[22], which uses aggregated regional level Google search data.

Second step: joint model for the 44 states/district/city other than HI, AK, ND, VT, MT, ME, and SD. For the 44 states, let $\boldsymbol{p}_t = (p_{t,1}, \ldots, p_{t,44})^{\intercal}$ denote CDC's %ILI at the state level; they are related to $y_{t,m}$ through $p_{t,m} = \exp(y_{t,m})/(1 + \exp(y_{t,m}))$. Our raw estimate for $\boldsymbol{p}_t$ from the first step is $\hat{\boldsymbol{p}}_t^{GT} = (\hat{p}_{t,1}, \ldots, \hat{p}_{t,44})^{\intercal}$, where $\hat{p}_{t,m} = \exp(\hat{y}_{t,m})/(1 + \exp(\hat{y}_{t,m}))$. Our estimate of the national %ILI from the first step is $\hat{p}_t^{nat}$. Let the boldface $\hat{\boldsymbol{p}}_t^{nat}$ denote the length-44 vector $\hat{\boldsymbol{p}}_t^{nat} = (\hat{p}_t^{nat}, \ldots, \hat{p}_t^{nat})^{\intercal}$. We also have the regional %ILI estimate $(\hat{p}_{t,1}^{reg}, \ldots, \hat{p}_{t,10}^{reg})$ from the first step. Let $\hat{\boldsymbol{p}}_t^{reg}$ denote the length-44 vector $\hat{\boldsymbol{p}}_t^{reg} = (\hat{p}_{t,r_1}^{reg}, \ldots, \hat{p}_{t,r_{44}}^{reg})^{\intercal}$, where r_m is the region number for state m.

Estimating $\boldsymbol{p}_t$ is equivalent to estimating the time series increment $\Delta\boldsymbol{p}_t = \boldsymbol{p}_t - \boldsymbol{p}_{t-1}$. We denote $\boldsymbol{Z}_t = \Delta\boldsymbol{p}_t$ for notational simplicity. For the estimation of $\boldsymbol{Z}_t$, we want to incorporate the cross-state, cross-source correlations. We have four predictors for $\boldsymbol{Z}_t$ after the first step: (i) $\boldsymbol{Z}_{t-1} = \Delta\boldsymbol{p}_{t-1}$, (ii) $\hat{\boldsymbol{p}}_t^{GT} - \boldsymbol{p}_{t-1}$, (iii) $\hat{\boldsymbol{p}}_t^{reg} - \boldsymbol{p}_{t-1}$, and (iv) $\hat{\boldsymbol{p}}_t^{nat} - \boldsymbol{p}_{t-1}$; they represent time series information, information from the state level Google search, information from the regional level estimation, and information from the national level estimation, respectively. Let $\boldsymbol{W}_t$ denote the collection of these four vectors $\boldsymbol{W}_t = (\boldsymbol{Z}_{t-1}^{\intercal}, (\hat{\boldsymbol{p}}_t^{GT} - \boldsymbol{p}_{t-1})^{\intercal}, (\hat{\boldsymbol{p}}_t^{reg} - \boldsymbol{p}_{t-1})^{\intercal}, (\hat{\boldsymbol{p}}_t^{nat} - \boldsymbol{p}_{t-1})^{\intercal})^{\intercal}$.

To combine the four predictors, we use the best linear predictor formed by them:

$$\hat{\boldsymbol{Z}}_t = \boldsymbol{\mu}_Z + \Sigma_{ZW}\Sigma_{WW}^{-1}(\boldsymbol{W}_t - \boldsymbol{\mu}_W), \tag{2}$$

where μ_Z and μ_W are the mean vectors of $\boldsymbol{Z}$ and $\boldsymbol{W}$ respectively, and Σ_{ZZ}, Σ_{ZW}, and Σ_{WW} are the covariance matrices of and between $\boldsymbol{Z}$ and $\boldsymbol{W}$. The best linear predictor gives the optimal way to linearly combine the four predictors to form a new one. The variance of $\hat{\boldsymbol{Z}}_t$ is

$$\mathrm{Var}(\hat{\boldsymbol{Z}}_t|\boldsymbol{W}_t) = \Sigma_{ZZ} - \Sigma_{ZW}\Sigma_{WW}^{-1}\Sigma_{WZ}. \tag{3}$$

Consistent with the first step, we adopt a sliding two-year training window to estimate μ_Z, μ_W, Σ_{ZZ}, Σ_{ZW}, and Σ_{WW} in Eq. (2) and (3). For μ_Z and μ_W, we use the empirical mean of the corresponding variables as the estimates. However, for the covariance matrices, due to their large sizes and the small number of observations, we need to structure the covariance matrices for reliable estimation.

We assume the following structure:

1. The covariances between the time series increments satisfy $\mathrm{Var}(\boldsymbol{Z}_t) = \mathrm{Var}(\boldsymbol{Z}_{t-1}) = \Sigma_{ZZ}$ and $\mathrm{Cov}(\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}) = \rho\Sigma_{ZZ}$, where $0 < \rho < 1$. This essentially assumes that the time series increments are stationary and have a stable autocorrelation across time and states.
2. Independence among the different sources of information: time series increment, the estimation error of the first-step state-level estimate, the estimation error of the regional estimate, and the estimation error of the national estimate, i.e., $\boldsymbol{Z}_t, \hat{\boldsymbol{p}}_t^{GT} - \boldsymbol{p}_t, \hat{\boldsymbol{p}}_t^{reg} - \boldsymbol{p}_t, \hat{\boldsymbol{p}}_t^{nat} - \boldsymbol{p}_t$ are all mutually independent.

The covariance matrices are thereby simplified as:

$$\Sigma_{ZW} = \begin{pmatrix}\rho\Sigma_{ZZ} & \Sigma_{ZZ} & \Sigma_{ZZ} & \Sigma_{ZZ}\end{pmatrix} \tag{4}$$

$$\Sigma_{WW} = \begin{pmatrix} \Sigma_{ZZ} & \rho\Sigma_{ZZ} & \rho\Sigma_{ZZ} & \rho\Sigma_{ZZ} \\ \rho\Sigma_{ZZ} & \Sigma_{ZZ} + \Sigma^{GT} & \Sigma_{ZZ} & \Sigma_{ZZ} \\ \rho\Sigma_{ZZ} & \Sigma_{ZZ} & \Sigma_{ZZ} + \Sigma^{reg} & \Sigma_{ZZ} \\ \rho\Sigma_{ZZ} & \Sigma_{ZZ} & \Sigma_{ZZ} & \Sigma_{ZZ} + \Sigma^{nat} \end{pmatrix} \tag{5}$$

where $\Sigma^{reg} = \mathrm{Var}(\hat{\boldsymbol{p}}_t^{reg} - \boldsymbol{p}_t)$, $\Sigma^{nat} = \mathrm{Var}(\hat{\boldsymbol{p}}_t^{nat} - \boldsymbol{p}_t)$, and $\Sigma^{GT} = \mathrm{Var}(\hat{\boldsymbol{p}}_t^{GT} - \boldsymbol{p}_t)$. To further control the estimation stability, we incorporate a ridge-regression-inspired shrinkage[34] to the linear predictor (2), replacing the joint covariance matrix of $(\boldsymbol{Z}_t^\intercal, \boldsymbol{W}_t^\intercal)^\intercal$ by the average of the structured covariance matrix and its empirical diagonal. Effectively, in Eq. (2), Σ_{ZW} is replaced by $\frac{1}{2}\Sigma_{ZW}$, and Σ_{WW} is replaced by $(\frac{1}{2}\Sigma_{WW} + \frac{1}{2}D_{WW})$, where D_{WW} is the diagonal of the empirical covariance of $\boldsymbol{W}_t$:

$$\hat{\boldsymbol{Z}}_t = \boldsymbol{\mu}_Z + \frac{1}{2}\Sigma_{ZW}(\frac{1}{2}\Sigma_{WW} + \frac{1}{2}D_{WW})^{-1}(\boldsymbol{W}_t - \boldsymbol{\mu}_W). \tag{6}$$

$\Sigma_{ZZ}, \Sigma^{nat}, \Sigma^{reg}, \Sigma^{GT}$ and D_{WW} are estimated by the corresponding sample covariance from the data in the most recent 2-year training window; ρ is estimated by minimizing the Frobenius norm (L_2 distance) between the empirical correlation and structured correlation. Based on Eq. (3), the variance estimate is similarly updated by

$$\mathrm{Var}(\hat{\boldsymbol{Z}}_t|\boldsymbol{W}_t) = \Sigma_{ZZ} - \frac{1}{2}\Sigma_{ZW}(\frac{1}{2}\Sigma_{WW} + \frac{1}{2}D_{WW})^{-1}\frac{1}{2}\Sigma_{WZ}.$$

Our final state-level %ILI estimate for week T after the second step is:

$$\hat{\boldsymbol{p}}_T = \boldsymbol{p}_{T-1} + \hat{\boldsymbol{\mu}}_Z + \hat{\Sigma}_{ZW}(\hat{\Sigma}_{WW} + \hat{D}_{WW})^{-1}(\boldsymbol{W}_T - \hat{\boldsymbol{\mu}}_W), \tag{7}$$

with corresponding 95% interval estimate

$$\left[\hat{\boldsymbol{p}}_T \pm 1.96 \cdot \sqrt{\mathrm{diagonal}\left(\hat{\Sigma}_{ZZ} - \frac{1}{2}\hat{\Sigma}_{ZW}(\hat{\Sigma}_{WW} + \hat{D}_{WW})^{-1}\hat{\Sigma}_{WZ}\right)}\right].$$

Second step: stand-alone model for HI, AK, ND, VT, MT, ME and SD. For $m \in$ {HI, AK, ND, VT, MT, ME, SD}, we take a stand-alone modeling approach. For each of these states, which is either non-contiguous or has the lowest multiple correlation with out-of-state %ILI (detailed in Supplementary Information), we focus on estimating the individual state's %ILI by integrating the within-state and national information in the second step. Thereby, our target is a scalar $Z_t^{(m)} = p_{t,m} - p_{t-1,m}$, the state's %ILI increment at the current week. The predictor vector in the second step for state m is $\boldsymbol{W}_t^{(m)} = (Z_{t-1}^{(m)}, (\hat{p}_{t,m}^{GT} - p_{t-1,m}), (\hat{p}_t^{nat} - p_{t-1,m}))$, where the regional terms are dropped. The best linear predictor with ridge-regression inspired shrinkage is then used to get the final estimate

$$\hat{Z}_t^{(m)} = \mu_Z^{(m)} + \frac{1}{2}\Sigma_{ZW}^{(m)}(\frac{1}{2}\Sigma_{WW}^{(m)} + \frac{1}{2}D_{WW}^{(m)})^{-1}(\boldsymbol{W}_t^{(m)} - \boldsymbol{\mu}_W^{(m)}). \tag{8}$$

The corresponding covariance matrices between the components $\Sigma_{ZW}^{(m)} = \mathrm{Cov}(Z^{(m)}, \boldsymbol{W}^{(m)})$, $\Sigma_{WW}^{(m)} = \mathrm{Var}(\boldsymbol{W}^{(m)})$, and $D_{WW}^{(m)} = \mathrm{diagonal}(\Sigma_{WW}^{(m)})$ are estimated by the corresponding sample covariance from the data in the most recent 2-year training window.

The final state-level %ILI estimate for week T after the second step for $m \in$ {HI, AK, ND, VT, MT, ME, SD} is:

$$\hat{p}_{T,m} = p_{T-1,m} + \hat{\mu}_Z^{(m)} + \hat{\Sigma}_{ZW}^{(m)}(\hat{\Sigma}_{WW}^{(m)} + \hat{D}_{WW}^{(m)})^{-1}(\boldsymbol{W}_T^{(m)} - \hat{\boldsymbol{\mu}}_W^{(m)}), \tag{9}$$

with corresponding 95% interval estimate

$$\left[\hat{p}_{T,m} \pm 1.96 \cdot \sqrt{\hat{\Sigma}_{ZZ}^{(m)} - \frac{1}{2}\hat{\Sigma}_{ZW}^{(m)}(\hat{\Sigma}_{WW}^{(m)} + \hat{D}_{WW}^{(m)})^{-1}\hat{\Sigma}_{WZ}^{(m)}}\right],$$

where $\Sigma_{ZZ}^{(m)} = \mathrm{Var}(Z^{(m)})$ is the scalar variance of the univariate time series $Z_t^{(m)}$.

Received: 5 November 2020; Accepted: 28 January 2021
Published online: 17 February 2021

References

1. US Centers for Disease Control and Prevention (CDC). Past seasons estimated influenza disease burden. https://www.cdc.gov/flu/about/burden/past-seasons.html (2020). Accessed: 2020-05-07.
2. Ginsberg, J. *et al.* Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
3. Yang, S. *et al.* Advances in using internet searches to track dengue. *PLoS Comput. Biol.* **13**, e1005607 (2017).
4. Scott, S. L. & Varian, H. R. Predicting the present with Bayesian structural time series. *Int. J. Math. Modell. Numer. Optim.* **5**, 4–23 (2014).
5. Scott, S. L. & Varian, H. R. Bayesian variable selection for nowcasting economic time series. In *Economic Analysis of the Digital Economy* (eds Goldfarb, A. *et al.*) 119–135 (University of Chicago Press, Chicago, 2015).
6. Wu, L. & Brynjolfsson, E. The future of prediction: how Google searches foreshadow housing prices and sales. In *Economic Analysis of the Digital Economy* (eds Avi Goldfarb, S. G. & Tucker, C.) 89–118 (University of Chicago Press, Chicago, 2015).

7. Shaman, J. & Karspeck, A. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences* **109**, 20425–20430 (2012). http://www.pnas.org/content/109/50/20425.full.pdf+html.
8. McNeil, D. G. Can smart thermometers track the spread of the coronavirus? https://www.nytimes.com/2020/03/18/health/coronavirus-fever-thermometers.html (2020). Accessed: 2020-04-12.
9. Yang, S., Santillana, M. & Kou, S. C. Accurate estimation of influenza epidemics using google search data via argo. *Proc. Natl. Acad. Sci.* **112**, 14473–14478 (2015).
10. Yang, S. *et al.* Using electronic health records and internet search information for accurate influenza forecasting. *BMC Infect. Dis.* **17**, 332. https://doi.org/10.1186/s12879-017-2424-7 (2017).
11. Yang, W., Lipsitch, M. & Shaman, J. Inference of seasonal and pandemic influenza transmission dynamics. *Proc. Natl. Acad. Sci.* **112**, 2723–2728 (2015).
12. Shaman, J., Karspeck, A., Yang, W., Tamerius, J. & Lipsitch, M. Real-time influenza forecasts during the 2012–2013 season. *Nat. Commun.* **4**, 2837. https://doi.org/10.1038/ncomms3837 (2013).
13. Yang, W., Karspeck, A. & Shaman, J. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput. Biol.* **10**, e1003583 (2014).
14. Shaman, J. & Kandula, S. Improved discrimination of influenza forecast accuracy using consecutive predictions. *PLoS Curr. Outbreaks* https://doi.org/10.1371/currents.outbreaks.8a6a3df285af7ca973fab4b22e10911e (2015).
15. Flusight: Flu forecasting | CDC. https://www.cdc.gov/flu/weekly/flusight/index.html (2020). Accessed: 2020-04-12.
16. Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J. & Rosenfeld, R. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Comput. Biol.* **11**, e1004382 (2015).
17. Farrow, D. C. *et al.* A human judgment approach to epidemiological forecasting. *PLoS Comput. Biol.* **13**, e1005248 (2017).
18. Yang, W., Olson, D. R. & Shaman, J. Forecasting influenza outbreaks in boroughs and neighborhoods of New York City. *PLoS Comput. Biol.* **12**, e1005201 (2016).
19. Davidson, M. W., Haim, D. A. & Radin, J. M. Using networks to combine "big data" and traditional surveillance to improve influenza predictions. *Sci. Rep.* **5**, 8154 (2015).
20. Zou, B., Lampos, V. & Cox, I. Multi-task learning improves disease models from web search. In *Proceedings of the 2018 World Wide Web Conference*, 87–96 (2018).
21. Lu, F. S., Hattab, M. W., Clemente, C. L., Biggerstaff, M. & Santillana, M. Improved state-level influenza nowcasting in the united states leveraging internet-based data and network approaches. *Nat. Commun.* **10**, 1–10 (2019).
22. Ning, S., Yang, S. & Kou, S. Accurate regional influenza epidemics tracking using internet search data. *Sci. Rep.* **9**, 5238 (2019).
23. Reich, N. G. *et al.* Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the us. *PLoS Comput. Biol.* **15**, e1007486 (2019).
24. Burkom, H. S., Murphy, S. P. & Shmueli, G. Automated time series forecasting for biosurveillance. *Stat. Med.* **26**, 4202–4218 (2007).
25. Biggerstaff, M. *et al.* Results from the Centers for Disease Control and Prevention's predict the 2013–2014 influenza season challenge. *BMC Infect. Dis.* **16**, 1–10 (2016).
26. Santillana, M. *et al.* Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput. Biol.* **11**, e1004513 (2015).
27. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
28. Butler, D. When Google got flu wrong. *Nature* **494**, 155–156 (2013).
29. Lampos, V. *et al.* Tracking covid-19 using online search. arXiv preprint arXiv:2003.08086 (2020).
30. Lipsitch, M. *et al.* Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. *Biosecur. Bioterrorism Biodefense Strategy Pract. Sci.* **9**, 89–115 (2011).
31. Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N. & Marathe, M. V. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza Other Resp. Viruses* **8**, 309–316 (2014).
32. Chretien, J.-P., George, D., Shaman, J., Chitale, R. A. & McKenzie, F. E. Influenza forecasting in human populations: a scoping review. *PLoS ONE* **9**, e94130 (2014).
33. Stephens-davidowitz, S. Google searches can help us find emerging covid-19 outbreaks. https://www.nytimes.com/2020/04/05/opinion/coronavirus-google-searches.html (2020). Accessed: 2020-05-07.
34. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
35. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2016).
36. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

Acknowledgements

SCK's research was supported in part by National Science Foundation grant DMS-1810914. The authors thank Professor Herman Chernoff for helpful comments. All analyses were performed with the R statistical software[35]. The R package that implements the ARGOX method is available on CRAN at https://cran.r-project.org/web/packages/argo/, which uses the `glmnet` package[36]. All datasets analyzed in the current study are available in the Harvard Dataverse repository, https://doi.org/10.7910/DVN/2IVDGK.

Author contributions

S.Y. and S.N. contributed equally to this work. S.Y., S.N., and S.C.K. designed the research; S.Y., S.N., and S.C.K. performed the research; S.Y. and S.N. analyzed data; and S.Y., S.N., and S.C.K. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-83084-5.

Correspondence and requests for materials should be addressed to S.C.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.