Improved state-level influenza activity nowcasting in the United States leveraging Internet-based data sources and network approaches via ARGONet

Fred S. Lu[1], Mohammad W. Hattab[2], Leonardo Clemente[3], Mauricio Santillana[1,4*]

[1] Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA
[2] Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Boston, MA, USA
[3] Tecnológico de Monterrey, Monterrey, Mexico
[4] Department of Pediatrics, Harvard Medical School, Boston, MA, USA
* Corresponding author: Mauricio Santillana, msantill@fas.harvard.edu

Abstract. In the presence of population-level health threats, precision public health approaches seek to provide the right intervention to the right population at the right time. Accurate real-time surveillance methodologies that can estimate infectious disease activity ahead of official healthcare-based reports, in relevant spatial resolutions, are critical to eventually achieve this goal. We introduce a novel methodological framework for this task which dynamically combines two distinct flu tracking techniques, using ensemble machine learning approaches, to achieve improved flu activity estimates at the state level in the US. The two predictive techniques behind the proposed ensemble methodology, named ARGONet, utilize (1) a dynamic and self-correcting statistical approach to combine flu-related Google search frequencies, information from electronic health records, and historical trends within a given state, as well as (2) a data-driven network-based approach that leverages spatial and temporal synchronicities observed in historical flu activity across states to improve state-level flu activity estimates. The proposed ensemble approach considerably outperforms each individual method and any previously proposed state-specific method for flu tracking, with higher correlations and lower prediction errors.

Introduction

The Internet has enabled near-real time access to multiple sources of medically relevant information, from cloud-based electronic health records to environmental conditions, social media activity, and human mobility patterns. These data streams, combined with an increase in computational power and our ability to process and analyze them, promise to revolutionize the way we treat individual patients and communities in the presence of health threats. As the field of precision medicine [1] continues to yield important medical insights as a consequence of recent improvements in the quality and cost of genetic sequencing as well as advances in bioinformatics methodologies, *precision public health* efforts aim to eventually provide the right intervention to the right population at the right time [2]. In this context, real-time disease surveillance systems capable of delivering early signals of disease activity at the local level may give local decision-makers, such as governments, school districts, and hospitals, valuable and timely information to better mitigate the effects of disease outbreaks. Our work focuses on a methodology aimed at achieving this for influenza activity surveillance.

Influenza has a large seasonal burden across the United States, infecting up to 35 million people and causing between 12000 and 56000 deaths per year [3]. Limiting the spread of outbreaks and reducing morbidity in those already infected are crucial steps for mitigating the impact of influenza. To guide this effort, public health officials, as well as the general public, should have access to localized, real-time indicators of influenza activity. Established influenza reporting systems currently exist over large geographic scales in the United States, coordinated by the Centers for Disease Control and Prevention (CDC). These systems provide weekly reports of influenza statistics, aggregated over the national, regional (10 groups as defined by the Health and Human Services), and starting in fall 2017, state level. Of particular interest, U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) records the percentage of patients reporting to outpatient clinics with symptoms of influenza-like illness (ILI), which is defined by fever over 100 °F in addition to sore

throat or cough, over the total number of patient visits [4]. While these measurements are an established indicator of historical ILI activity, they require around a week to collect from individual health-care providers across the country, analyze, and report and are frequently revised. This delay and potential subsequent revisions can reduce the utility of the system for real-time situational awareness and data analysis.

To address this delay, research teams have devised methods to estimate ILI a week ahead of healthcare-based reports and in near-real time, termed "nowcasting", at the national and regional levels. These methods incorporate a variety of techniques from statistical modeling and machine learning [5–8], to mechanistic and epidemiological models [8–10]. Many utilize innovative web-based data sources such as Internet search frequencies and electronic health records [5]. Some have also taken into account historically-observed spatial and temporal synchronicities in flu activity [11, 12] to improve the accuracy of existing flu surveillance tools [13, 14]. Because influenza transmission occurs locally and is spread from person to person, the timing of outbreaks and resulting infection rate curves can significantly differ from state to state. Thus, despite the aforementioned successes in national and regional surveillance, these spatial resolutions are likely not enough to aid decision-making at smaller geographic scales, since important information about localized conditions is lost in regional or national aggregates.

The first influenza nowcasting system at the state level across the United States was Google Flu Trends (GFT). GFT reported a number each week representing influenza activity for each state, various cities, in addition to the national and regional levels, using Google search activity as a predictor. While innovative at the time, studies have pointed out its large prediction errors when tested in real time and proposed alternative methodologies that can incorporate Google searches more effectively at the national level [6, 15–17]. A feasibility proposal replacing GFT for flu detection, at the state level, was published last year by Kandula, Hsu, and Shaman, who presented retrospective out-of-sample flu estimates, over the 2005-2011 flu seasons, using a random forest methodology based on Google searches and historical flu activity as predictors [18]. While this study showed promise, the authors did not report significant improvements to GFT and provided only aggregate distributional metrics to evaluate the performance of their models over conglomerates of states (as opposed to state-level metrics), making it challenging to replicate or improve their results for any given state.

Approach. In this study, we provide a solution for localized flu nowcasting by first extending to each state a proven methodology for inferring flu activity, named ARGO, which combines information from flu-related Google search frequencies, electronic health records, and historical flu trends. Next, we develop a spatial network approach, named Net, which refines ARGO's flu estimates by incorporating structural spatio-temporal synchronicities observed historically in flu activity. Finally, we introduce ARGONet, a novel ensemble approach that combines estimates from ARGO and Net using a dynamic, out-of-sample learning method. We produce retrospective estimates using ARGO from September 2012 to May 2017 and show that ARGO alone demonstrates strong improvement over GFT and an autoregressive benchmark. Then we generate retrospective flu estimates using ARGONet from September 2014 to May 2017 and show further improvements in accuracy over ARGO in over 75% of the states studied. We present detailed metrics and figures over each state to enable analysis as well as future refinement of our methods.

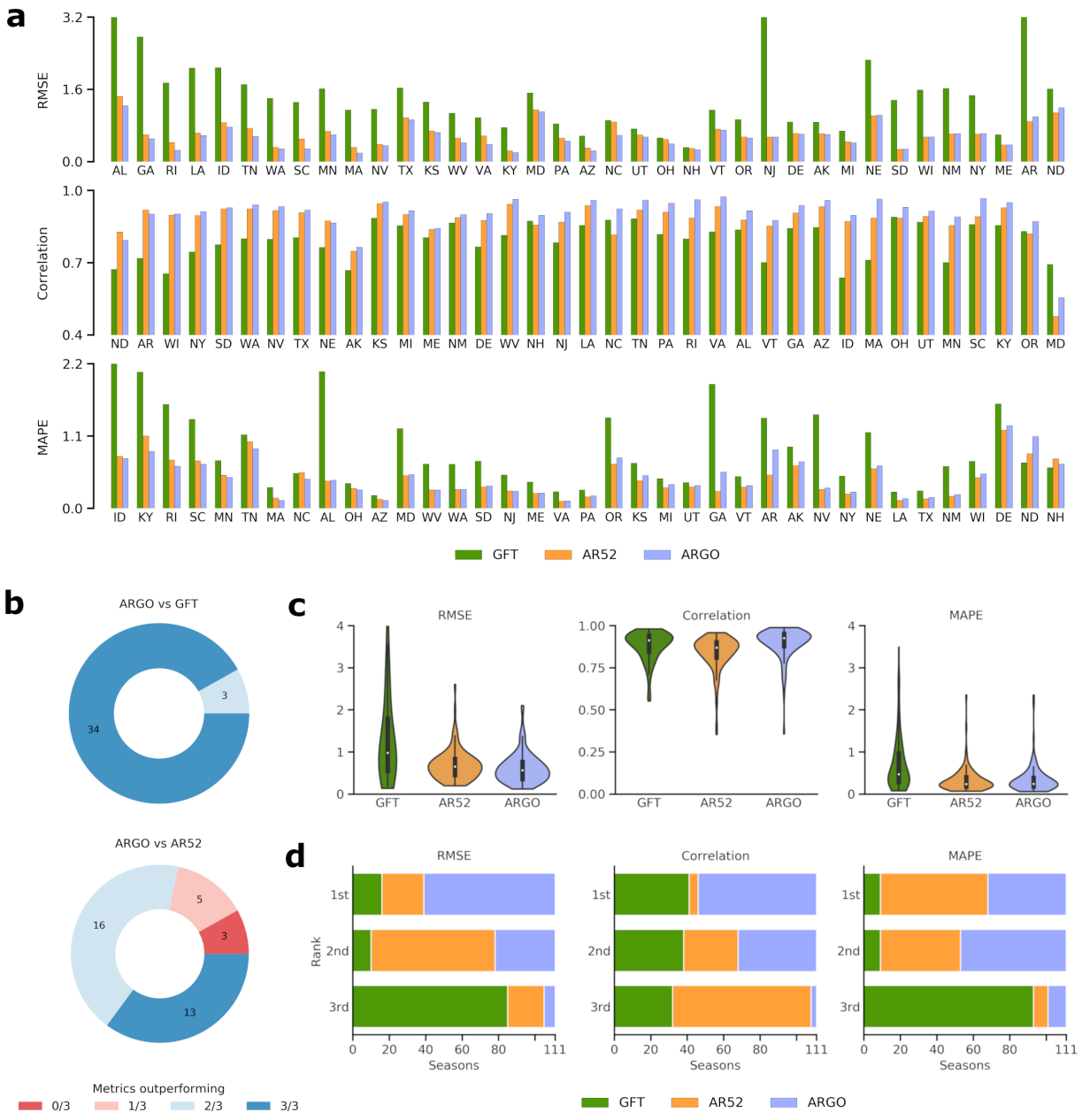
Results

State-level ARGO models outperform existing benchmarks

We first adapted the ARGO (AutoRegression with General Online information) methodology to state-level flu detection. ARGO has previously demonstrated the ability to infer flu activity with high precision over a variety of geographical areas and scales [5, 19, 20]. The adapted model dynamically fits a regularized multivariable regression on state-level Google search engine frequencies, electronic health record reports from athenahealth, and historical CDC %ILI estimates (see Methods section). We trained a separate ARGO model for each state and used them to generate retrospective out-of-sample estimates from September 30, 2012 to May 14, 2017 for each state in the study.

To assess the predictive performance of ARGO, we also produced retrospective estimates using two benchmarks: a) GFT, a scaled version of the Google Flu Trends time series for each state, fitted to match the

Fig 1: a) State-level performance of ARGO benchmarked with GFT and AR52, as measured by RMSE (top), Pearson correlation (middle), and MAPE (bottom), over the period from September 30, 2012 to August 15, 2015. Extreme GFT error values are displayed up to a cutoff point. States are ordered by ARGO performance relative to the benchmarks to facilitate comparison. b) The proportion of states where ARGO outperforms GFT (top) or AR52 (bottom) in 0, 1, 2, or all 3 metrics. c) The distribution of values for each metric for each model, over the 111 state-seasons during the same period. Numerical values are reported in Table A1. d) The distribution of ranks attained by each model over the 111 state-seasons for each metric.



scale of each state's CDC %ILI; and b) AR52, an autoregressive model which uses the CDC %ILI from the previous 52 weeks in a regularized multivariable regression to predict %ILI of the current week (see Methods section for details on both). Since autoregression is an important component of ARGO itself, improvement over AR52 indicates the effective contribution of real-time Google search and electronic health record data. Fig. 1a shows a graphical comparison of ARGO, AR52, and GFT for each state over the time window when GFT estimates were available (September 30, 2012 to August 15, 2015). The three panels display the root mean square error (RMSE), Pearson correlation, and mean absolute percent error (MAPE) over the period (defined under Comparative Analyses in the Methods section). ARGO models outperform GFT in RMSE in every state, in correlation in all but one state, and in MAPE in all but two states. Furthermore, ARGO reduces the RMSE of GFT by $> 50\%$ in 23 states and increases correlation by $> 10\%$ in 25 states. ARGO also performs comparably or better in RMSE and correlation than AR52, although it does not generally outperform AR52 in MAPE. In all but 8 states, ARGO beats AR52 in a majority (2 or all 3) of metrics (Fig. 1b).

Attention to flu activity is typically heightened during flu seasons (between week 40 of one year and week 20 of the next), as the majority of seasonal flu cases occur within this time frame. We assessed ARGO performance over each flu season within the study period, namely the 2012-13 to 2014-15 seasons inclusive. With three seasons where comparison with GFT is available and 37 states, this yields 111 state-seasons. Of these, ARGO outperforms GFT in 94 state-seasons in RMSE, 69 in correlation, and 97 in MAPE. ARGO also surpasses AR52 in 83 state-seasons in RMSE, 104 in correlation, and 47 in MAPE (aggregated from Table A3). Correspondingly, ARGO outperforms the benchmarks in terms of median and interquartile range over the seasons, with the exception of MAPE against AR52 (Fig. 1c), and ranks first over the majority of state-seasons in RMSE and correlation (Fig. 1d). Interestingly, despite lower quartile values, GFT has a better tail spread than ARGO in correlation, though not in RMSE or MAPE.

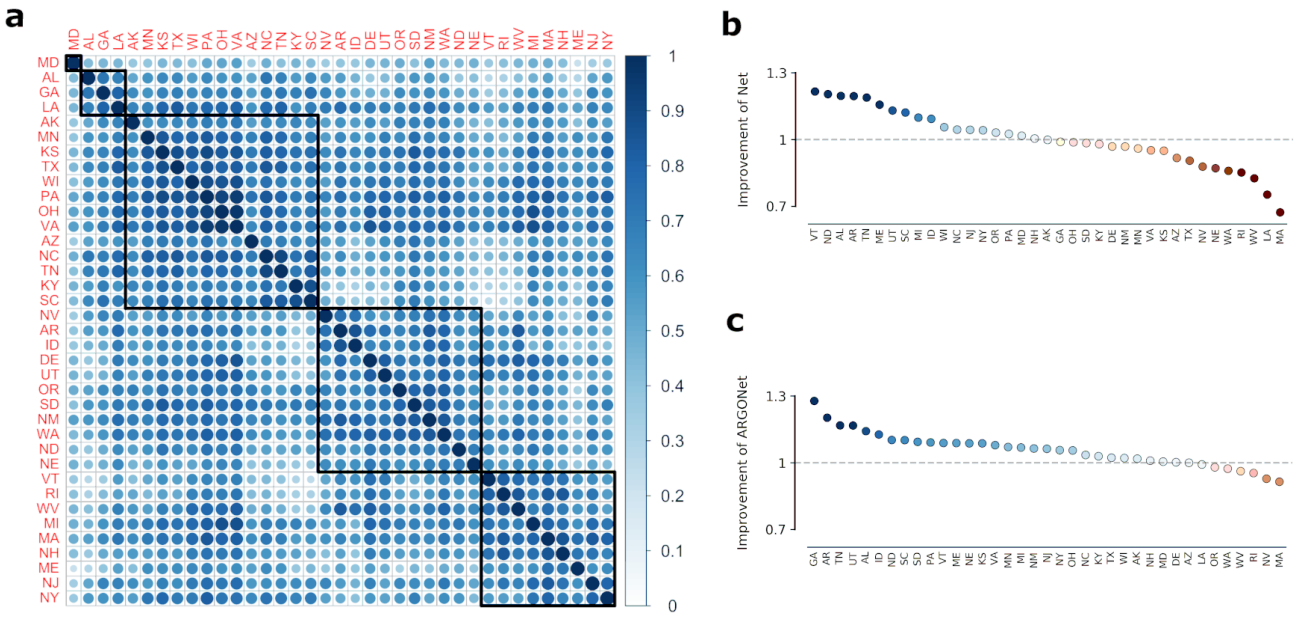


Fig 2: a) Heatmap of pairwise %ILI correlations between all states in the study over the period September 30, 2012 to May 14, 2017. Five clusters of inter-correlated states are denoted by black boxes. b) RMSE improvement of Net over ARGO over the period September 28, 2014 to May 14, 2017. The improvement of Net is here defined as the inverse RMSE ratio of Net and ARGO, so values above 1 indicate improvement. c) RMSE improvement of ARGONet over ARGO over the same period.

Incorporating spatio-temporal structure in state-level flu activity

Because ARGO models the flu activity within a given state using only data specific to that state, a natural question is whether information from other states across time can be used to improve the accuracy of flu predictions. As shown in Fig. 2a, historical CDC %ILI observations show synchronous correlations between states. The clustering of intercorrelated states from the same regions suggests that geographical spatio-temporal structure can be exploited as a correctional effect.

Inspired by this finding, we developed a network-based model on each state, which incorporates multiple weeks of historical %ILI activity from all other states in a regularized multivariable regression. Out-of-sample estimates from this model, denoted Net, improve on the RMSE of ARGO on half of the states over the period of Sept. 28, 2014 to May 14, 2017, but show a comparable increase in error on the other half of the states (Fig. 2b). Because ARGO and Net dramatically outperform each other over distinct states, we investigated whether an ensemble combining the relative strengths of each model could lead to significant improvement.

ARGONet ensemble improves on state-level ARGO models

The resulting ensemble, denoted ARGONet, dynamically selects either ARGO's or Net's prediction in each week and state based on the past performance of each model over a suitable training space (see Methods section for details). Over the period where ARGONet estimates were generated (Sept. 28, 2014 to May 14, 2017 after a 2-year training window), we found that this approach resulted in out-of-sample improvement in RMSE over ARGO in all but 8 states (Fig. 2c). Furthermore, in these 8 states, the error increase of ARGONet is relatively controlled compared to the error increase of Net.

In addition to RMSE, ARGONet also displays general improvement in correlation and MAPE over both ARGO and the AR52 benchmark (Fig. 3a). We previously noted that ARGO did not outperform AR52 in MAPE despite being superior in terms of RMSE, which suggests that ARGO is more accurate than AR52 during periods of high flu incidence and less accurate during low flu incidence. On the other hand, by incorporating spatio-temporal structure, ARGONet is able to achieve lower MAPE than AR52 over both the entire time period of Sept. 2014 - May 2017 and the 108 state-seasons within this period (three states are missing %ILI data for the 2016-17 season, resulting in fewer state-seasons compared to the previous analysis) (Fig. 3a-c). Note that while ARGO and Net outperforms the AR52 benchmark by a majority of metrics in 32 and 30 states, respectively, ARGONet does the same in 36 out of 37 states (Fig. 3b).

Interestingly, the performance increase of ARGONet does not appear to stem from being simultaneously more accurate than ARGO and Net over a majority of state-seasons. Note in Fig. 3d that while ARGONet tends to rank first in a smaller proportion of state-seasons than ARGO or Net, ARGONet ranks either first or second in a far larger proportion of state-seasons than the other two models, indicating that the ensemble's overall success comes from increased consistency. Finally, Fig. 3e subdivides the states by the fraction of seasons (out of 3) where each model outperforms AR52. We see that ARGONet performs favorably (wins 2 or 3 out of 3 seasons) in the vast majority of states, with considerably better distribution in terms of MAPE than ARGO or Net. Refer to Table A3 for numerical metrics over each state and season.

Detailed time series comparisons of ARGO and ARGONet relative to the official CDC-reported %ILI values are shown in Fig. 4. Note that our models consistently track the CDC %ILI curve during both high and low periods of ILI activity, whereas GFT often significantly overpredicts during season peaks. Time series plots specifically comparing ARGONet and ARGO over Sept. 2014 - May 2017 are presented in Fig. A1 and better enable the reader to visually inspect ARGONet's improvement over ARGO. In concordance with previous results, ARGONet tracks the CDC %ILI curve more accurately than ARGO over some periods of time, while over other periods the curves are identical. This is an expected result of our winner-takes-all ensemble methodology. Some states that especially highlight these patterns are Arkansas (AR), Georgia (GA), New York (NY), and Vermont (VT).

Fig 3: a) State-level performance of ARGONet benchmarked with ARGO and AR52, as measured by RMSE (top), Pearson correlation (middle), and MAPE (bottom), over the period from September 28, 2014 to May 14, 2017. States are ordered by ARGONet performance relative to the benchmarks to facilitate comparison. b) The proportion of states where ARGO (top), Net (middle), or ARGONet (bottom) outperforms AR52 in 0, 1, 2, or all 3 metrics. c) The distribution of values for each metric for each model, over the 108 state-seasons during the same period. Numerical values are reported in Table A1. d) The distribution of ranks attained by each model over the 108 state-seasons for each metric. e) The proportion of states where each model outperforms AR52 in 0, 1, 2, or all 3 flu seasons within the same period. Only 34 states have data for all three flu seasons available in this period.

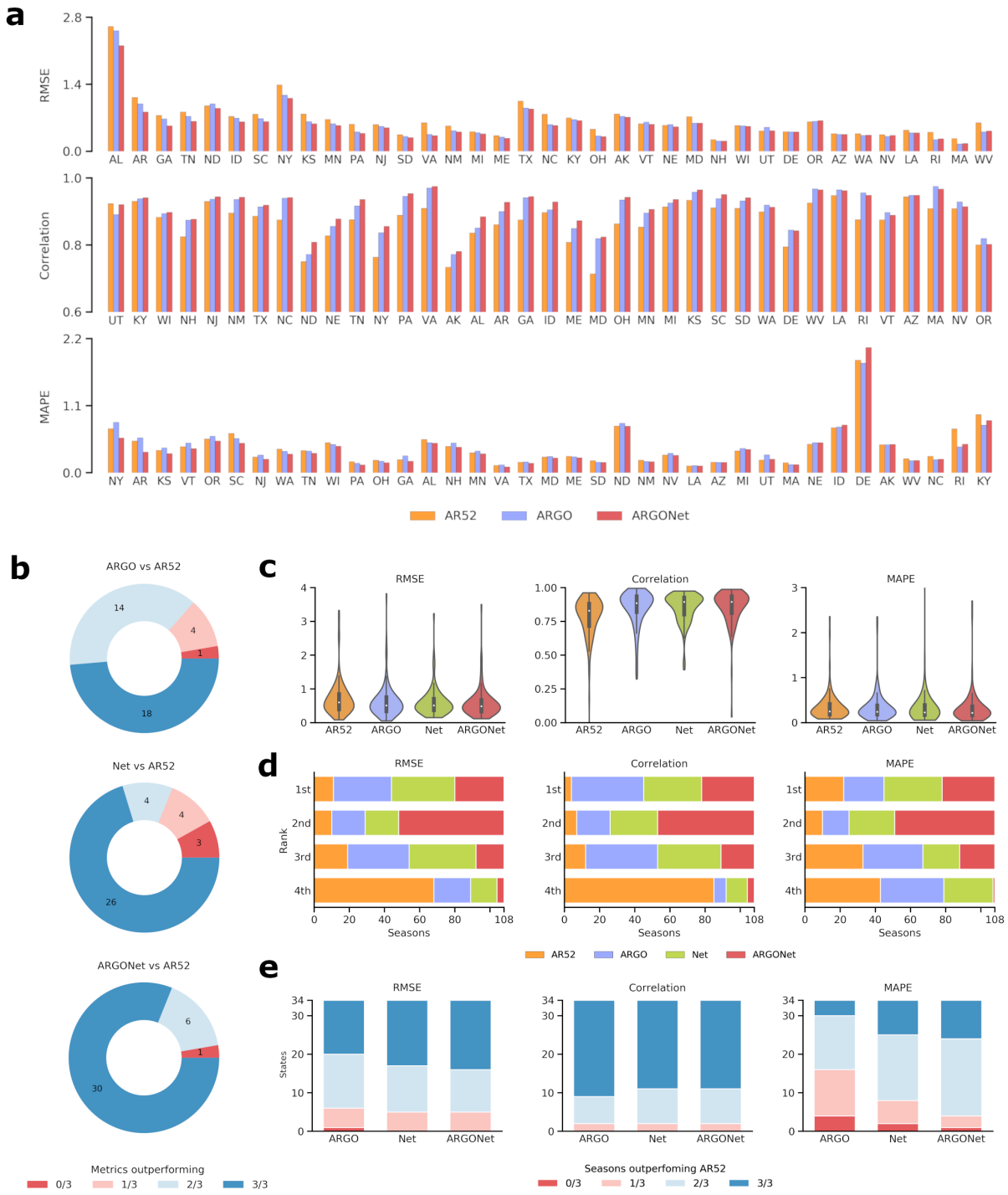
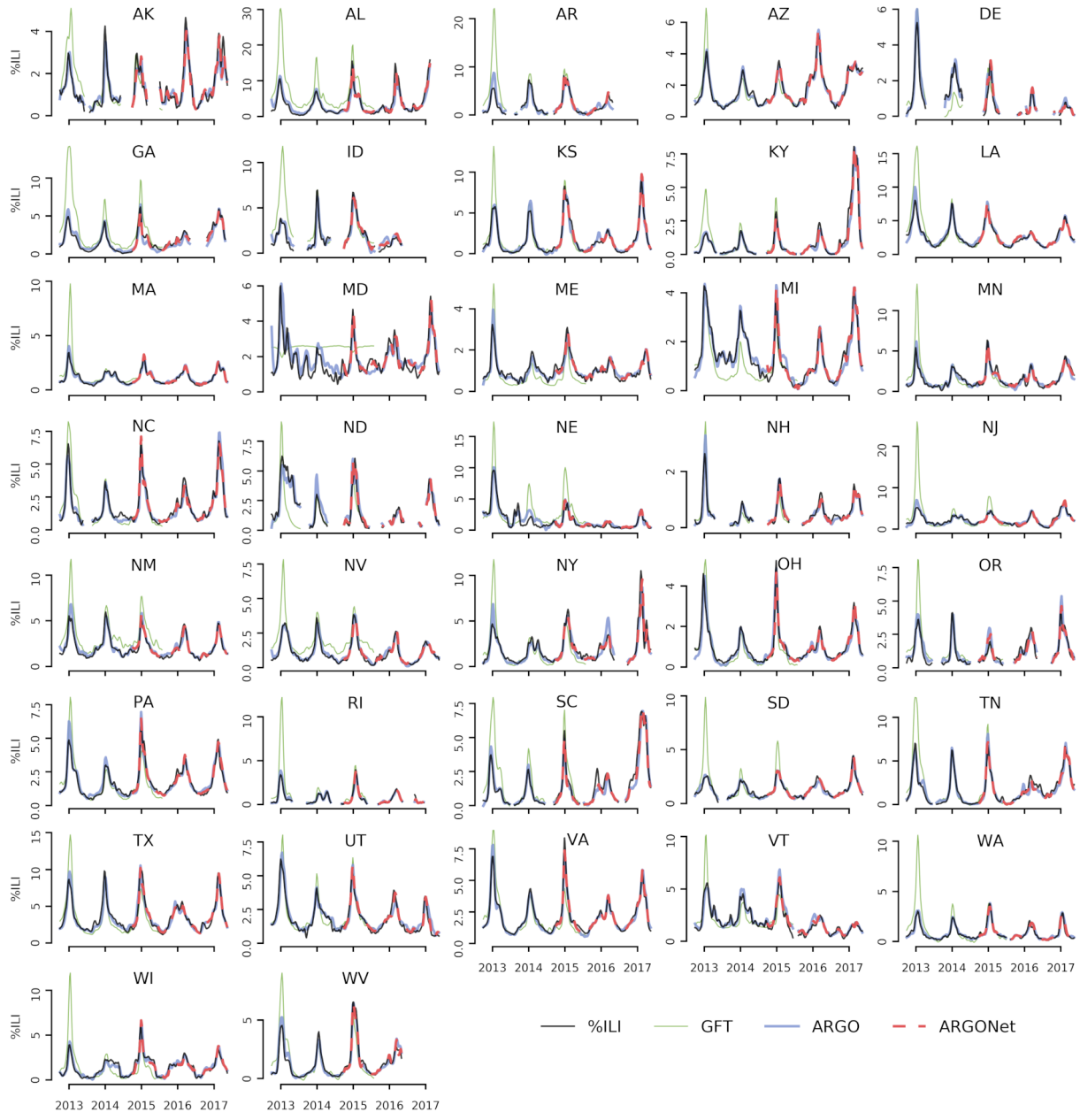


Fig 4: Time series plots displaying the performance of ARGO and ARGONet relative to the official CDC %ILI time series over the entire study period (September 30, 2012 to May 14, 2017). The GFT benchmark is also shown. Refer to Fig. A1 for more detailed figures from Sept. 2014 - May 2017.



Discussion

Our ensemble, ARGONet, successfully combines Google search frequencies and electronic health record data with spatial and temporal trends in flu activity to produce accurate forecasts for current ILI activity at the state level. We believe that the accuracy of our method involves a balance between responsiveness and robustness. Real-time data sources such as Google searches and electronic health records provide information about the present, allowing the model to immediately respond to current flu trends. On the other hand, using the values of past CDC flu reports in an autoregression adds robustness by preventing our models from creating outsize errors in prediction. Similarly, incorporating spatial synchronicities adds stability by maintaining state-level inter-correlations evident in historical flu activity. Our results suggest that dynamic learning ensembles incorporating real-time Internet-based data sources can surpass any individual methods in inferring flu activity.

Previous work has shown the versatility of ARGO, one of the component models in our ensemble, over a variety of disease estimation scenarios [5, 20–22]. At the state level, where it had not been applied before, it clearly outperformed existing benchmarks over the study period, namely Google Flu Trends and an autoregression. While ARGO alone performs better than study benchmarks, we also found that spatio-temporal synchrony could be used to improve model accuracy (Net). Combining web-based data sources with this structural network-based approach (ARGONet) further improves prediction accuracy and suggests the future study of synchronous network effects at varying geographical scales. Future work should explore adding similar approaches to flu nowcasting systems at finer spatial resolutions, such as the city level.

Accurate flu monitoring at the state level faces challenges due to higher variance in data quality across states. The official %ILI reporting system within each state varies in reporting coverage and consistency, and thus the magnitudes of flu activity may not reflect actual differences of flu activity between states. In addition, the quality of Google Trends frequencies and the prevalence of clinics reporting to athenahealth (the provider of our electronic health record data) vary considerably from state to state, affecting the ability of our models to extract useful information from these data sources. In past work, we hypothesized that better model performance could be explained as a consequence of higher Internet-based data quality. As a result, we believe that geographical improvement of ARGONet over the benchmark AR52 (as defined by percent reduction of RMSE) may also be associated with factors that serve as proxies of Google Trends or athenahealth data quality, namely detectable flu-related search terms from Google Trends and estimated athenahealth population coverage (Fig. A2a-d). Regression analysis indicates a moderate association of ARGONet improvement with athenahealth coverage (Fig. A2e) and a weak association with detectable flu-related Google search activity (Fig. A2f). Interestingly, flu-related search activity has a very strong correlation with state population (Fig. A2g), which suggests that larger pools of Internet users result in better signal-to-noise ratio in search activity. Future analysis can examine the interplay of these factors with CDC %ILI report quality and structural spatial correlations. For example, we hypothesize that the Net model contributes strongly in states with lesser-quality data which are adjacent to states with high-quality data. Clustering states by pairwise %ILI correlation (Fig. 2a) shows that geographic proximity is a relevant synchronous factor. Within the boxed clusters, we see that Southeastern states, Western states, and New England/mid-Atlantic states often group together.

Localized, accurate surveillance of flu activity can set a foundation for precision public health in infectious diseases. Important developments in this field can involve emerging methodologies for tracking disease at fine-grained spatial resolutions, rapid analysis and response to changing dynamics, and targeted, granular interventions in disparate populations, each of which has the potential to complement traditional public health methods to increase effectiveness of outcomes [23]. We believe that the use of our system can produce valuable real-time subregional information and is a step toward this direction. At the same time, the performance of ARGONet depends directly on the availability and quality of Internet-based input data and also relies on a consistently reporting (even if lagged) healthcare-based surveillance system. We anticipate that data sources will improve over time, for example, if athenahealth continues to gain a larger market share over the states or more Google Trends information becomes available. If these conditions hold, or as more web-based data sources including other electronic health record systems become available in real time, the accuracy of our methods may continue to increase.

Methods

Data Acquisition

Three data sources were used in our models: influenza-like illness rates from ILINet, Internet search frequencies from Google Trends, and electronic health records from athenahealth. Weekly information from each data source was collected for the time period of October 4, 2009 to May 14, 2017.

Influenza-like illness rates

Weekly influenza rates reported by outpatient clinics and health providers for each available state were used as the epidemiological target variable of this study. The weekly rate, denoted %ILI, is computed as the number of visits for influenza-like illness divided by the total number of visits. Data from October 4, 2009 to May 14, 2017 for 37 states were obtained from the CDC. For inclusion in the study, a state must have data from October 2009 to May 2016, with no influenza seasons (week 40 of one year to week 20 of the next) missing. Some states were missing data, usually due to not reporting in the off-season (between week 20 and week 40 of each year). Missing or unreported weeks, as well as weeks where 0 cases were reported, were excluded from analysis on a state-by-state basis. While in real-time ILI values for a given week may be revised in subsequent weeks, we only had access to the revised version of these historical values.

Internet search frequencies

Search volumes for specific queries in each state were downloaded through Google Trends, which returns values in the form of frequencies scaled by an unknown constant. While our pipeline used the Google Trends API for efficiency, search volumes can be publicly obtained from `www.trends.google.com` for reproducibility. Relevant search terms were identified by downloading a complete set of 287 flu-related search queries for each state, and keeping the terms that were not completely sparse. Because Google Trends left-censors data below an unknown threshold, replacing values with 0, a query with high sparsity indicates low frequency of searches for that query within the state.

In an ideal situation, relevant search queries at the state-level resolution would be obtained by passing the historical %ILI time series for each state into Google Correlate, which returns the most highly correlated search frequencies to an input time series. However, such functionality is only supported at the national level, at least in the publicly accessible tool. Given this limitation, we used two strategies to select search terms:

1. A initial set of 128 search terms was taken from previous studies tracking influenza at the US national level [6].

2. To search for additional terms, we submitted multiple state %ILI time series into the Google Correlate and extracted flu-related terms, under the assumption that some of the state-level terms would show up at the national level.

To minimize overfitting on recent information, the %ILI time series inputted into Google Correlate were restricted from 2009-2013. State-level search frequencies for the union of these terms and the 128 previous terms were then downloaded from the Google Trends API, resulting in 282 terms in total (Table A2).

Electronic health records

Influenza rates for patients within the athenahealth provider network are provided weekly from athenahealth on each Monday. Three types of syndromic reports were used as variables: 'influenza visit counts', 'ILI visit counts', and 'unspecified viral or ILI visit counts', which were converted into percentages by dividing by the total patient visit counts for each week. The athenahealth network and influenza rate variables are detailed in Santillana et al. [19]

Google Flu Trends

In addition to the above data sources, we downloaded GFT estimates as a benchmark for our models. GFT provided a public flu prediction system for each state until its discontinuation in August 2015 [24]. GFT values were downloaded and scaled using the same initial training period of 104 weeks used in all of our models (October 4, 2009 to September 23, 2012).

Models

ARGO

The time series prediction framework ARGO (AutoRegression with General Online information) issues flu predictions by fitting a multivariable linear regression each week on the most recent available Internet predictors and the previous 52 %ILI values. Because of many potentially redundant variables, L1 regularization (Lasso) was applied to produce a parsimonious model by setting the coefficients for weak predictors to 0. The model was re-trained each week on a shifting 104-week training window in order to adapt to the most recent two years of data, and the regularization hyperparameter was selected using 10-fold cross validation on each training set. Details about the ARGO model and its applicability in monitoring infectious diseases such as influenza, dengue, and zika are presented in previous work [5, 21, 22]. Refer to the Appendix for a detailed mathematical formulation of ARGO.

To fine-tune predictive performance, adjustments to ARGO were introduced on a state-by-state basis:

- Filtering features by correlation: For each week, non-autoregressive features ranked outside the top 10 by correlation were removed to reduce noise from poor predictors. Based on previous research, this complementary feature selection process benefits the performance of lasso, which can be unstable in variable selection [20, 21].

- Regularization hyperparameters: Features with high correlation to the target variable over the training set received a lower regularization weight, which makes them less subject to the L1 penalty (see the Appendix for derivation).

- Weighting recent observations: Although ARGO dynamically trains on the last 104 weeks of observations, more recent observations likely contain more relevant information. Thus the most recent 4 weeks of data received a higher weight (set to be twice the weight of the other variables) in the training set.

Network-based approach

Historical CDC ILI observations show synchronous relationships between states, as shown in Fig. 2. To identify these relationships with the goal of improving our %ILI predictions, for each state, we dynamically constructed a regularized multi-linear model for each week that has the following predictors: %ILI terms for the previous 52 weeks for the target state, and the synchronous (same week's values) and the past three week's of observed CDC's %ILI terms for the other states. Notice that to produce predictions of %ILI for a given state in a given week, the model requires synchronous %ILI from the other states, which would not be available in real-time. Instead, we used ARGO predictions for the current week as surrogates for these unobserved values. This model was re-trained weekly using all previously observed data with 10-fold cross-validation to determine the L1 regularization term (formulation in Appendix). This model is denoted Net.

Ensemble approach

In order to optimally combine the predictive power of ARGO and Net, we trained an ensemble approach based on a winner-takes-all voting system, which we named ARGONet. ARGONet's prediction for a given week is assigned to be Net's prediction if Net produces lower error relative to the observed CDC %ILI (specifically root mean square error, as defined in Comparative analyses) in the previous K predictions as compared to ARGO. Otherwise, ARGONet's prediction is assigned to be ARGO's prediction. To determine

the hyper-parameter K for each state, we trained ARGONet using the first 104 out-of-sample predictions of ARGO. Here K can take the value of either 1, 2 or 3. The value of K that yielded the lowest RMSE between ARGONet and CDC's %ILI over the training period was chosen to produce out-of-sample predictions in the unseen time period (September 28, 2014 onward).

Comparative analyses

To assess the predictive performance of the models, we produced state-level retrospective estimates using two benchmarks: a) "AR52", an autoregressive model, which uses the %ILI from the previous 52 weeks in a LASSO regression to predict %ILI of the current week, and b) "GFT", made by scaling each state's Google Flu Trends time series to its official revised %ILI using the same initial training period of 104 weeks used in our models as discussed in the Methods section.

The performances of all models and benchmarks compared to the official (revised) %ILI were scored using three metrics: root mean squared error (RMSE), Pearson correlation coefficient, and mean absolute percent error (MAPE). These were computed over the entire study period (September 30, 2012 to May 14, 2017) and over each influenza season (defined as week 40 of one year to week 20 of the next) within the study period.

The models and benchmarks were further scored over two specific sub-periods: 1) the window when GFT was available (September 30, 2012 to August 9, 2015), and 2) the window starting with the first available ARGONet prediction (September 28, 2014 to May 14, 2017).

ARGO model estimates were generated using scikit-learn in Python 2.7 [25], while Net and ensemble models were generated in R 3.4.1. Analysis was conducted in Python except for Fig. 2a, which used the R package corrplot [26].

Acknowledgments

This work was partially funded by the Centers for Disease Control and Prevention's Cooperative Agreement PPHF 11797-998G-15. The authors thank Josh Gray, Anna Zink, and Dorrie Raymond for the collection and processing of the data from athenahealth.

The authors would also like to thank Matthew Biggerstaff from the National Center for Immunization and Respiratory Disease, Centers for Disease Control and Prevention, for great insights and guidance on the direction and objectives of this study.

References

1. Collins FS, Varmus H. A new initiative on precision medicine. New England Journal of Medicine. 2015;372(9):793–795.

2. Khoury MJ, Iademarco MF, Riley WT. Precision public health for the era of precision medicine. American journal of preventive medicine. 2016;50(3):398–401.

3. Disease Burden of Influenza — Seasonal Influenza (Flu) — CDC; 2018. `https://www.cdc.gov/flu/about/disease/burden.htm`.

4. Overview of Influenza Surveillance in the United States — Seasonal Influenza (Flu) — CDC; 2017. `https://www.cdc.gov/flu/weekly/overview.htm`.

5. Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC. Using electronic health records and Internet search information for accurate influenza forecasting. BMC Infect Dis. 2017;17(1):332.

6. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc Natl Acad Sci U S A. 2015;112(47):14473–14478.

7. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLoS Comput Biol. 2015;11(8):e1004382.

8. Yang W, Karspeck A, Shaman J. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. PLoS Comput Biol. 2014;10(4):e1003583.

9. Gog JR, Ballesteros S, Viboud C, Simonsen L, Bjornstad ON, Shaman J, et al. Spatial Transmission of 2009 Pandemic Influenza in the US. PLoS Comput Biol. 2014;10(6):e1003635.

10. Yang W, Lipsitch M, Shaman J. Inference of seasonal and pandemic influenza transmission dynamics. Proc Natl Acad Sci U S A. 2015;112(9):2723–2728.

11. Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. Synchrony, waves, and spatial hierarchies in the spread of influenza. Science. 2006;312(5772):447–451.

12. Charu V, Zeger S, Gog J, Bjørnstad ON, Kissler S, Simonsen L, et al. Human mobility and the spatial transmission of influenza in the United States. PLoS Comput Biol. 2017;13(2):e1005382.

13. Davidson MW, Haim DA, Radin JM. Using networks to combine "big data" and traditional surveillance to improve influenza predictions. Sci Rep. 2015;5:8154.

14. Zou B, Lampos V, Cox I. Multi-Task Learning Improves Disease Models from Web Search. In: Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee; 2018. p. 87–96.

15. Lampos V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using search query logs. Sci Rep. 2015;5:12760.

16. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol. 2013;9(10):e1003256.

17. Santillana M. Editorial Commentary: Perspectives on the Future of Internet Search Engines and Biosurveillance Systems. Clin Infect Dis. 2017;64(1):42–43.

18. Kandula S, Hsu D, Shaman J. Subregional Nowcasts of Seasonal Influenza Using Search Trends. J Med Internet Res. 2017;19(11):e370.

19. Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. Sci Rep. 2016;6:25732.

20. Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, et al. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. JMIR Public Health Surveill. 2018;4(1):e4.

21. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. PLoS Negl Trop Dis. 2017;11(1):e0005295.

22. Yang S, Kou SC, Lu F, Brownstein JS, Brooke N, Santillana M. Advances in using Internet searches to track dengue. PLoS Comput Biol. 2017;13(7):e1005607.

23. Dolley S. Big Data's Role in Precision Public Health. Frontiers in public health. 2018;6:68.

24. Google Flu Trends;. `https://www.google.org/flutrends/about/`.

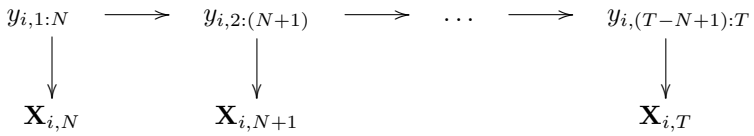
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825–2830.

26. Wei T, Simko V. R package "corrplot": Visualization of a Correlation Matrix; 2017. Available from: `https://github.com/taiyun/corrplot`.

Appendix

ARGO formulation

Let $y_{i,t}$ be the CDC %ILI in state i at time t, and let $\mathbf{X}_{i,t} = \{X_{i,t,k}\}_{k\in\{1,\dots,M\}}$ be the vector of Internet-based data in corresponding state and time. ARGO assumes a hidden Markov model based on an autoregressive structure with N lags, as shown below:



Here, the vectors $\{y_{i,(t-N+1):t}\}_{t\geq N}$ follow the Markov property, and at time T, the Tth such vector has not yet been fully observed. Meanwhile, the observed variables $\mathbf{X}_{i,T}$ depend only the corresponding hidden $y_{i,T}$.

ARGO model

The above formulation results in the model

$$y_{i,t} = \mu_i + \sum_{j=1}^{N} \alpha_j y_{i,t-j} + \sum_{k=1}^{M} \beta_k \mathbf{X}_{i,t,k} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

We take $N = 52$ to incorporate short-term and seasonal autoregressive trends within the past year of data, and $M = 285$ at maximum, corresponding to the number of Google Trends and athenahealth variables. The high number of input variables gives us a $p > n$ situation, so we impose L_1 regularization. Therefore, we can solve for parameters μ_i, $\alpha = (\alpha_1, \dots, \alpha_N)$, and $\beta = (\beta_1, \dots, \beta_M)$ which minimize the objective function

$$\sum_t \left(y_{i,t} - \mu_i - \sum_{j=1}^{N} \alpha_j y_{i,t-j} + \sum_{k=1}^{M} \beta_k \mathbf{X}_{i,t,k} \right)^2 + \lambda_\alpha ||\alpha_j||_1 + \lambda_\beta ||\beta_k||_1 \tag{2}$$

using a rolling training window consisting of the 104 weeks prior to time t, with hyperparameters λ_α and λ_β.

Hyperparameters

The parameters in (2) are governed by hyperparameters λ_α and λ_β; however, we introduced a modification to the model. Rather than adhering to the groups α and β, we replaced them with more flexible groups in the following manner:

Let $\gamma = \{\alpha_1, \dots, \alpha_N, \beta_1, \dots, \beta_M\}$ be the set of regularized parameters. Take $p \subseteq \gamma$ to be a set of "priority" parameters, and $q = \gamma - p$ to be the remaining parameters. Letting γ_p and γ_q represent the parameter vectors corresponding to these sets, the objective function simply becomes

$$\sum_t \left(y_{i,t} - \mu_i - \sum_{j=1}^{N} \alpha_j y_{i,t-j} + \sum_{k=1}^{M} \beta_k \mathbf{X}_{i,t,k} \right)^2 + \lambda_p ||\gamma_p||_1 + \lambda_q ||\gamma_q||_1 \tag{3}$$

To limit the model space, we allowed p to take one of 5 configurations and selected the one with best out-of-sample performance within each state. The configurations are the parameters corresponding to:

1. the 10 highest correlated input variables to the %ILI vector over the training set
2. athenahealth variables
3. athenahealth variables and the most correlated autoregressive terms, namely $\{y_{t-i}\}_{i\in\{1,2,3,4,12,26,52\}}$

4. the 3 most correlated Google Trends variables over the training set and the most correlated autoregressive terms

5. athenahealth variables, the 3 most correlated Google Trends variables, and the most correlated autoregressive terms.

To further constrain the search space, we maintain a hyperparameter ratio $\lambda_p/\lambda_q = 1/10$. Since λ_p is smaller, this allows the priority parameters to take larger values, effectively giving them more weight in the regression. The single hyperparameter was then determined using 10-fold cross-validation over the training set. In practice, prediction accuracy was robust to the specific value of the ratio.

Net formulation

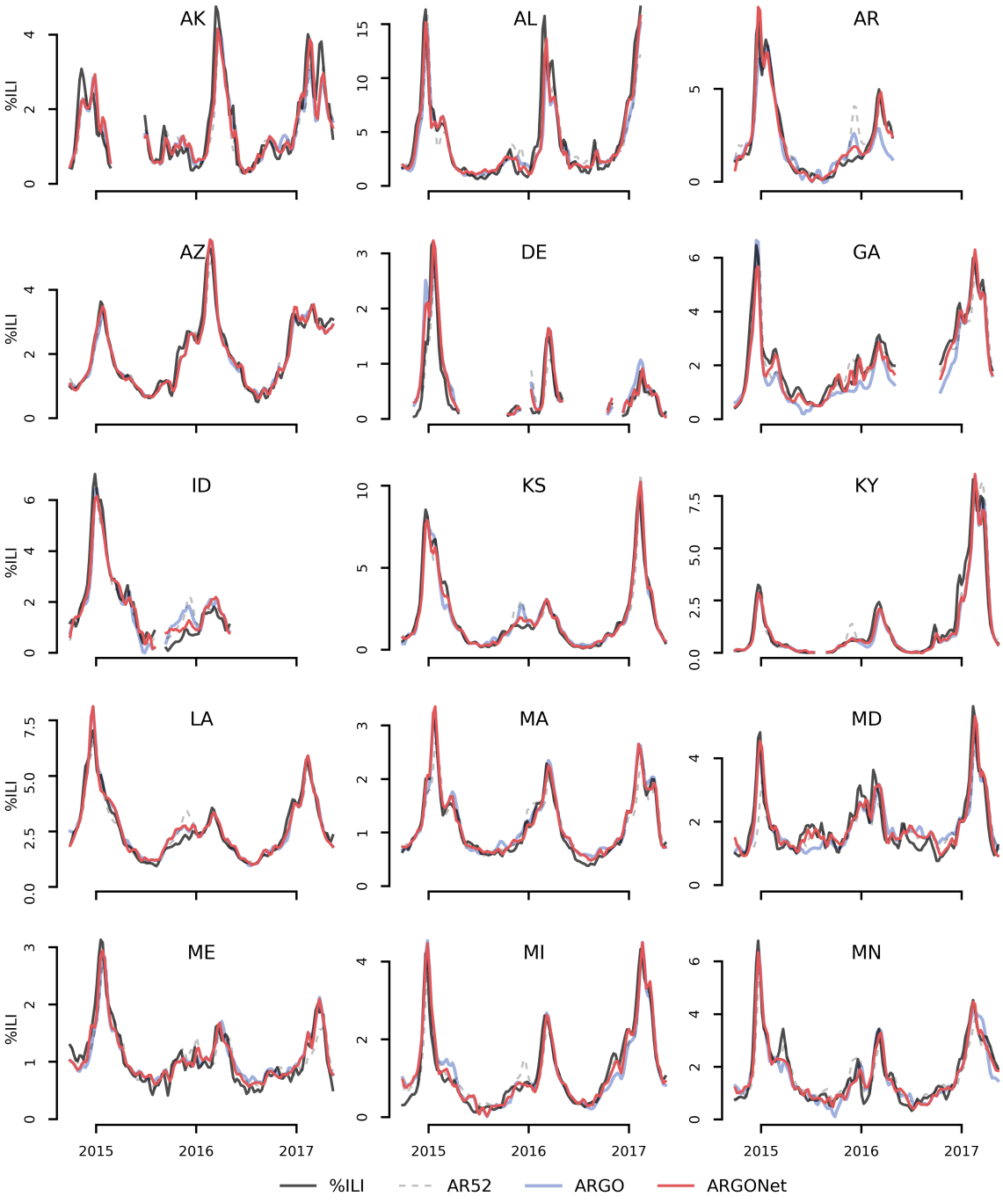
While the ARGO model for any given state i is constrained to the data within state i, we hypothesized that $y_{i,t}$ shows spatio-temporal structure with other states $s \neq i$ in the short term (i.e. over the past 4 weeks). Adopting the previous notation, the Net model is then

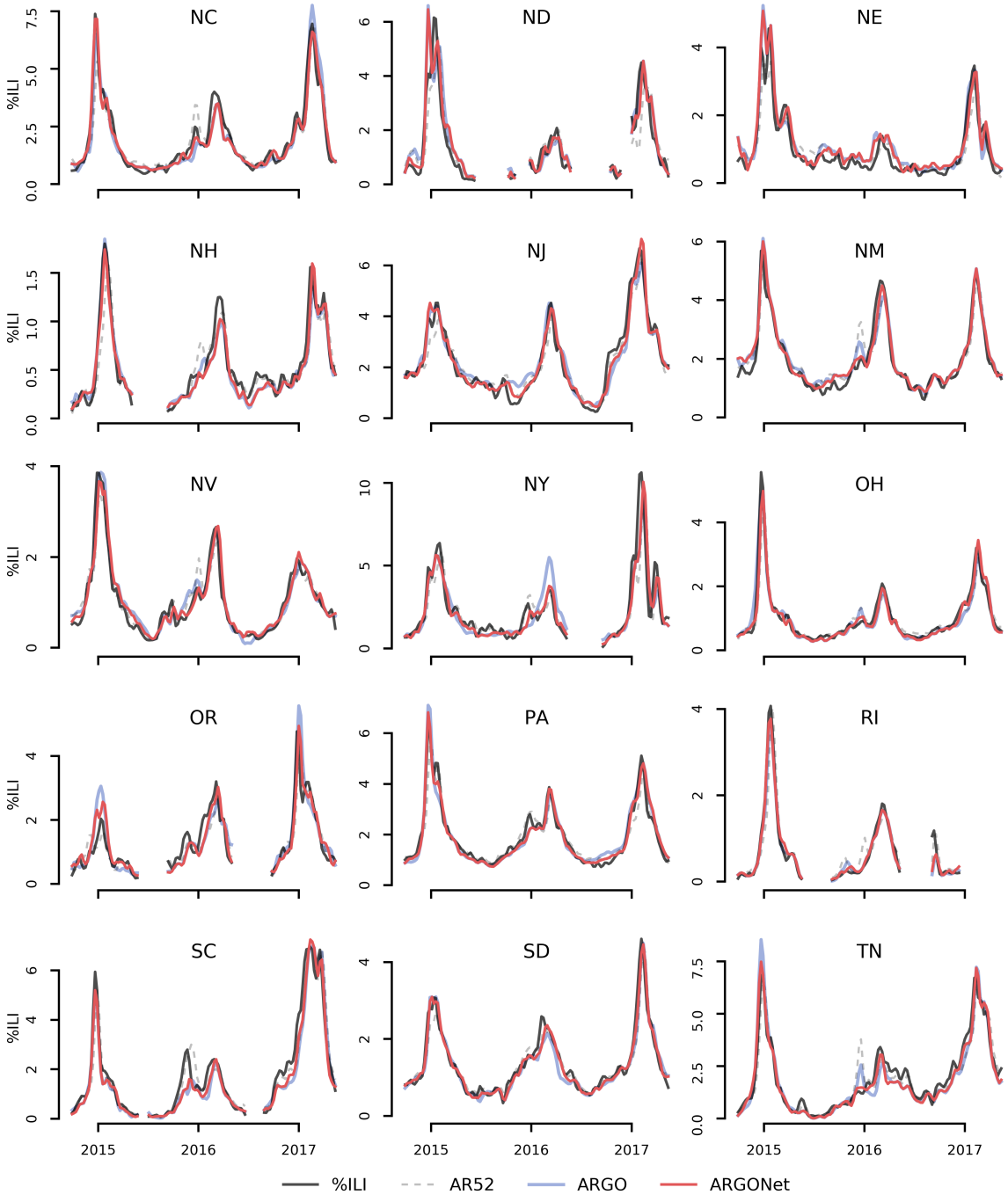
$$y_{i,t} = \mu_i + \sum_{j=1}^{52} \alpha_j y_{i,t-j} + \sum_{s \neq i} \sum_{k=0}^{3} \beta_{s,k} y_{s,t-k} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

This model was fit using an expanding training window consisting of all previously observed data with 10-fold cross-validation for the regularization hyperparameter. However, for prediction at a given time t, the concurrent %ILI values $y_{s,t}$ are not yet observed, so we replace them with the corresponding real-time ARGO estimates $\hat{y}_{s,t}$. This substitution inherently assumes that ARGO is unbiased, i.e. $y_{i,t} = \hat{y}_{i,t} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

Appendix figures and tables

Fig A1: Time series plots displaying the performance of ARGO and ARGONet relative to the official CDC %ILI time series over the period September 28, 2014 to May 14, 2017. The AR52 benchmark is also displayed.





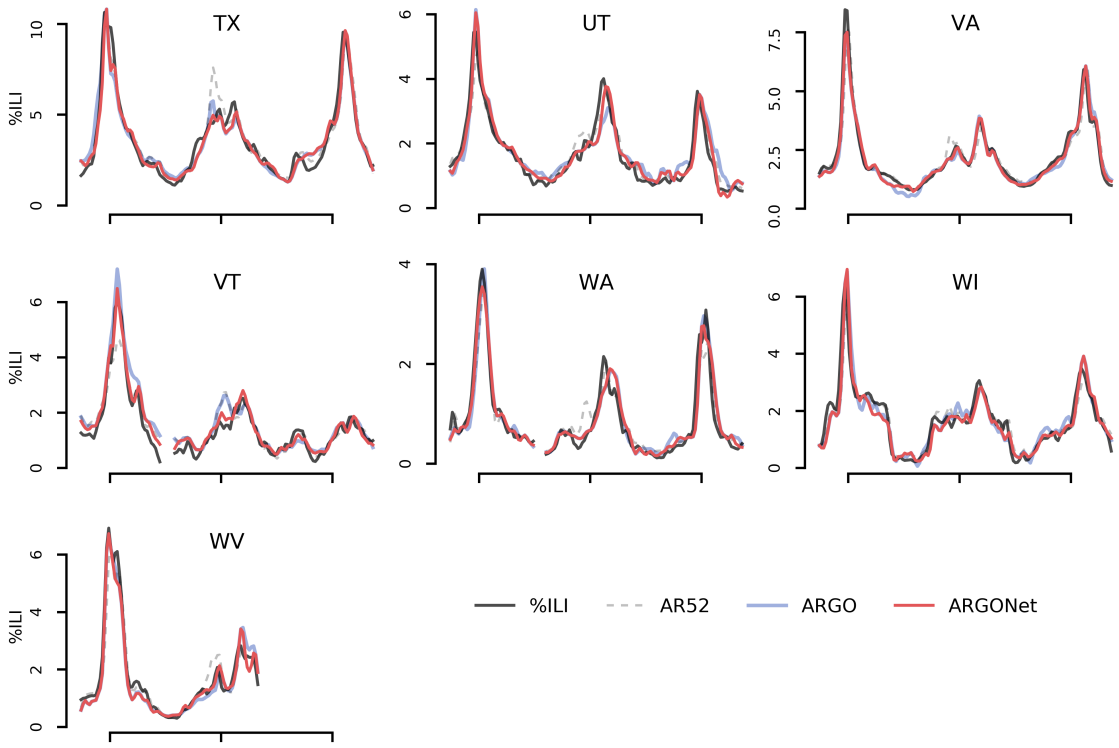


Fig A2: a) Geographical heatmap of the improvement (%RMSE reduction) of ARGONet over AR52. Possible explanatory factors for improvement over an autoregressive benchmark are b) population, which can affect data quality or spatial structure, c) the number of detectable flu-related Google Trends terms per state, which can be taken as a proxy of search data quality, and d) athenahealth coverage per state, calculated as average visits per thousand. e-g) Regression analysis indicates the presence of associations between e) ARGONet improvement and athenahealth coverage, with one outlier (Wisconsin, colored red) removed, f) ARGONet improvement and Google Trends quality, and g) state population and Google Trends quality.

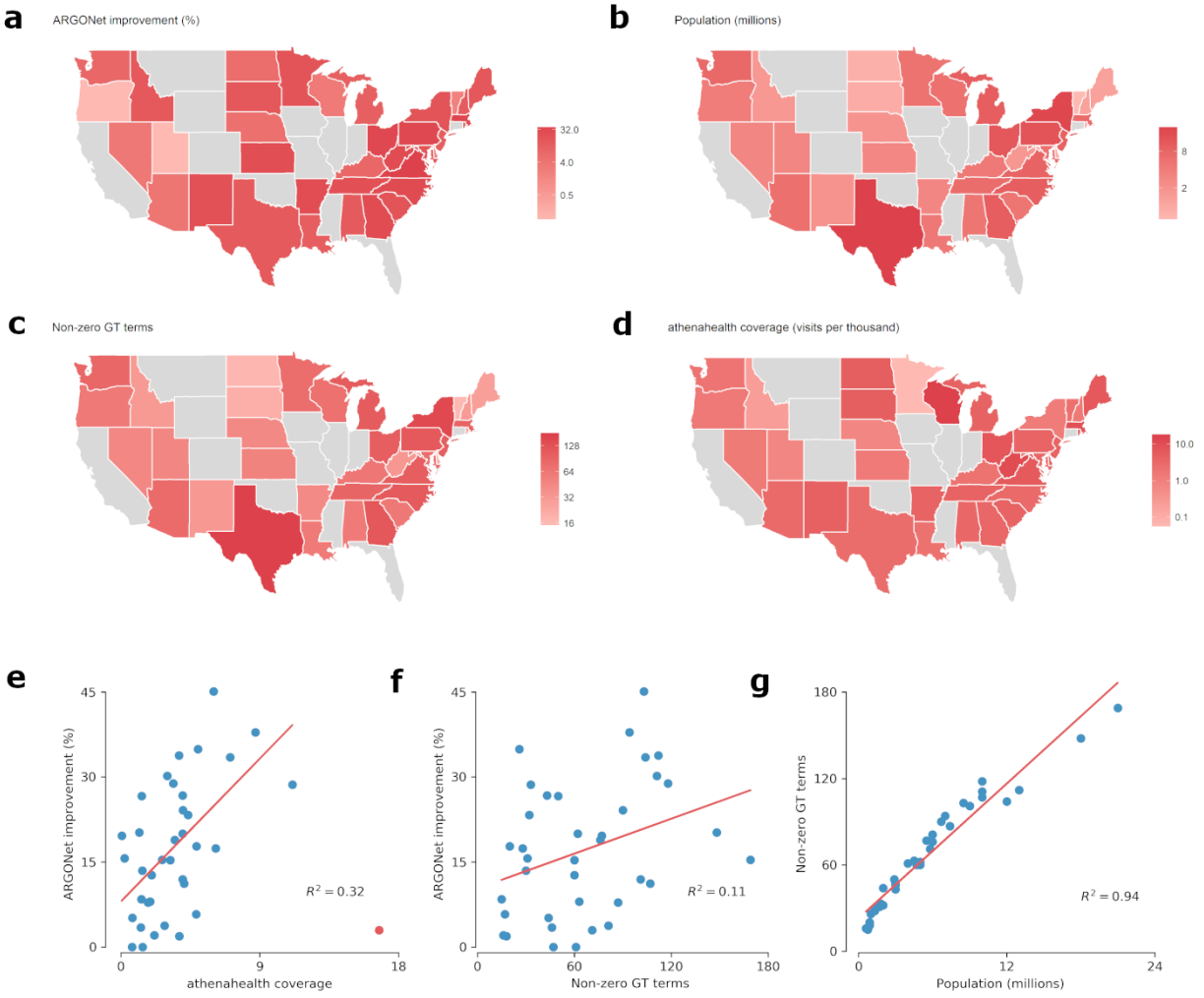


Table A1: Aggregate metrics over each flu season for each state corresponding to the violin plots in Figures 1c and 3c. The reported numbers are the median and interquartile ranges of the aggregates. The final two rows are the best reported aggregates for each metric from Kandula et al. for comparison. These were not discussed in the main paper because of the different time period.

Model	RMSE	Correlation	MAPE
2014-15 to 2016-17 seasons			
ARGONet	0.48 (0.33-0.67)	0.90 (0.81-0.94)	0.22 (0.12-0.36)
Net	0.51 (0.37-0.71)	0.90 (0.80-0.93)	0.23 (0.17-0.39)
ARGO	0.51 (0.34-0.75)	0.89 (0.82-0.94)	0.24 (0.18-0.38)
AR52	0.60 (0.39-0.85)	0.83 (0.72-0.88)	0.25 (0.18-0.42)
2012-13 to 2014-15 seasons			
ARGO	0.56 (0.36-0.77)	0.93 (0.88-0.95)	0.25 (0.18-0.38)
AR52	0.65 (0.46-0.84)	0.87 (0.81-0.90)	0.24 (0.18-0.40)
GFT	0.98 (0.55-1.80)	0.91 (0.84-0.94)	0.47 (0.29-0.96)
2005-06 to 2010-11 seasons			
GFT	0.93 (0.66-1.33)	0.89 (0.80-0.94)	0.71 (0.44-1.51)
Kandula et al.	0.84 (0.54-1.25)	0.86 (0.75-0.91)	0.54 (0.33-0.90)

Table A2: Complete search queries downloaded from Google Trends

thermoscan, how long is the flu contagious, how contagious is the flu, baby has rsv, anas barbariae hepatis, tamiflu suspension, flu food, what to do if you have the flu, coughing remedies, treatment for rsv, having the flu, influenza b symptoms, z pack and alcohol, how long does the flu last in adults, sinus infection cure, robitussin cough, strep, fever and cough, signs of flu, flu fever, cough headache, flu recovery time, flu how long, how to reduce fever, symptoms of rsv in adults, braun thermoscan, sinus infections, how long flu, treatment for the flu, upper respiratory, strep throat, generic tamiflu, intestinal flu, ear thermometer, cold versus flu, how long does flu last, cold or flu, the flu symptoms, expectorant, fight the flu, how long contagious, how to get rid of the flu, tamiflu pediatric dosing, influenza a contagious, how to get over flu, treat flu, can dogs get the flu from humans, rsv symptoms in adults, influenza a treatment, tylenol sinus, how long am i contagious, influenza contagious, flu in children, reduce fever, flu gestation, reduce a fever, is influenza contagious, how long does it take to get over the flu, over the counter flu, how to treat flu, medicine for the flu, contagious flu, flu care, flu or cold, can adults get rsv, is the flu contagious?, rsv in adults, influenza a symptoms, flu relief, fever cough, signs of rsv, oscillococcinum, type a influenza, flu symptoms, painful cough, flu incubation, how to get over the flu, cold and flu, acute bronchitis, fever flu, tamiflu in children, what is flu a, what to eat when you have the flu, flu remedy, type a flu, baby rsv, tamiflu side effects, flu shot symptoms, flu and fever, symptoms of flu, cold, how long does the flu last?, taking temperature, type b flu, flu and bronchitis, flu versus cold, get rid of the flu, medicine for flu, flu children, treatment for flu, cough and fever, positive flu test, influenza type b, flu shots, influenza type a, high fever, symptoms of rsv, flu cold, flu remedies, flu recovery, dangerous fever, do i have the flu, how to cure the flu, rsv contagious, get over the flu fast, influenza treatment, flu treatment, body temperature, remedies for the flu, biaxin side effects, duration of the flu, flu incubation period, flu test, symptoms of bronchitis, tamiflu dose, robitussin, rsv treatment, influenza symptoms, tamiflu drug interactions, tamiflu dosage, walking pneumonia, oscillo, how to treat the flu, getting over the flu, break a fever, influenza a and b, incubation period for flu, clarithromycin, over the counter flu medicine, the flu virus, cold with fever, tessalon, when is the flu contagious, symptoms of pneumonia, stomach flu, tussin, how long does the flu last, treating the flu, type b flu symptoms, flu virus, how long are you contagious, is tamiflu an antibiotic, flu lasts, rsv baby, influenza b treatment, flu shot, chest cold, remedies for flu, rsv infection, flu treatments, flu like symptoms, tamiflu while pregnant, flu contagious, symptoms of influenza, flu while pregnant, flu or pneumonia, flu report, flu reports, symptoms of influenza b, type a and b flu, i have the flu, is the flu contagious, type a flu symptoms, tamiflu children, influenza treatment guidelines, breaking a fever, flu type b symptoms, flu complications, flu contagious period, incubation period for the flu, flu cough, what is influenza, rsv, flu type a symptoms, tussionex, tamiflu during pregnancy, pneumonia, how long am i contagious with the flu, fever and cold, is rsv contagious, signs of the flu, influenza a incubation, influenza a incubation period, tamiflu generic, recovering from flu, rapid flu, influenza incubation period, how to treat flu symptoms, cure the flu, incubation for flu, how long are you contagious with the flu, exposed to flu, robitussin cf, flu a symptoms, get rid of flu, fever reducer, type b influenza, exposure to flu, bronchitis, tamiflu and alcohol, flu length, cough fever, tamiflu drug, flu and cold, flu cures, sinus, how to break a fever, tamiflu and breastfeeding, what is influenza b, tamiflu contagious, the flu, what is influenza a, get rid of the flu fast, pregnant and have the flu, baby with rsv, incubation period flu, flu a, influenza incubation, when you have the flu, normal body temperature, how long does influenza last, gripe, flu duration, treat flu symptoms, flu or strep, when is the flu no longer contagious, rsv symptoms, flu germs, how long is the flu, flu symptoms in toddlers, oseltamivir, get over the flu, how long is flu contagious, treating flu, flu prophylaxis, home remedies for flu, flu a and b, flu vaccine, cold vs flu, respiratory flu, a influenza, human temperature, tamiflu and pregnancy, influenza a, influenza b, is influenza a contagious, flu vs cold, cure flu, how to get rid of flu, normal body, flu medicine, cold symptoms, flu type, what is type a flu, b flu, flu and strep, rsv infant, tamiflu wiki, stomach flu symptoms, how long does rsv last, how long is rsv contagious, treat the flu, symptoms of the flu, early flu symptoms, flu headache, flu type b, flu type a, what is rsv, how to break a fever in adults, incubation period for influenza, tamiflu in pregnancy, is flu contagious

Table A3: See separate supplementary excel file.