

# SCIENTIFIC REPORTS



OPEN

## Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance

M. Santillana<sup>1,2,3</sup>, A. T. Nguyen<sup>3</sup>, T. Louie<sup>4</sup>, A. Zink<sup>5</sup>, J. Gray<sup>5</sup>, I. Sung<sup>5</sup> & J. S. Brownstein<sup>1,2</sup>

**Accurate real-time monitoring systems of influenza outbreaks help public health officials make informed decisions that may help save lives. We show that information extracted from cloud-based electronic health records databases, in combination with machine learning techniques and historical epidemiological information, have the potential to accurately and reliably provide near real-time regional estimates of flu outbreaks in the United States.**

Received: 31 December 2015

Accepted: 20 April 2016

Published: 11 May 2016

Influenza is a leading cause of death in the United States (US), where up to 50,000 are killed each year by influenza-like illnesses (ILI)<sup>1</sup>. Therefore, monitoring, early detection, and prediction of influenza outbreaks are crucial to public health. Disease detection and surveillance systems provide epidemiologic intelligence that allows health officials to deploy preventive measures and help clinic and hospital administrators make optimal staffing and stocking decisions<sup>2</sup>.

The US Centers for Disease Control and Prevention (CDC) monitors ILI in the US by gathering information from physicians' reports about patients with ILI seeking medical attention<sup>3</sup>. CDC's ILI data provides useful estimates of influenza activity; however, its availability has a known time lag of one to two weeks. This time lag is far from optimal since public health decisions need to be made based on information that is two weeks old. Systems capable of providing real-time estimates of influenza activity are, thus, critical.

Many attempts have been made to design methods capable of providing real-time estimates of ILI activity in the US by leveraging Internet-based data sources that could potentially measure ILI in an indirect manner<sup>4–11</sup>. Google Flu Trends (GFT), a digital disease detection system that used Internet searches to predict ILI in the US, became the most widely used of these non-traditional methods in the past few years<sup>12</sup>. In August of 2015, GFT was shut down, opening opportunities for novel and reliable methods to fill the gap. Many lessons have been learned in the field of digital disease detection from multiple updates to GFT, proposed not only by Google, but also by other researchers<sup>13–18</sup>. The performance of some of these updated models has substantially improved, for example, by including historical flu activity information as input, and by dynamically recalibrating the models, in order to not only incorporate the most up-to-date clinical information but also to adapt to new behavior in the population (for example, how Internet users search for health-related terms)<sup>16–18</sup>. Finally, very accurate real-time ILI estimates can now be produced, at the national-level, by combining disparate data streams in the US as shown in<sup>19</sup>.

National-level flu estimates are hard to translate into actionable information that enable local health officials to make better decisions during a surge in clinical visits, for example<sup>20</sup>. As a consequence, accurate flu estimates at finer spatial resolutions are desirable. Cities within the same geographical region tend to have positively correlated estimates of epidemiologic model parameters such as the basic reproduction number,  $R_0$ , and such estimates tend to differ from region to region in the US<sup>21</sup>. Additionally, knowledge of the influenza level in one region, when used in conjunction with a network model drawn from social network analysis, has been shown to improve the accuracy of GFT's ILI forecasts nationally as well as in multiple regions<sup>22</sup>. Unfortunately, the accuracies of existing GFT-like systems degrade substantially at the regional and local resolutions<sup>23</sup>.

Diverse studies have shown *retrospectively* the high correlation between aggregated data obtained from electronic health records (EHR) and flu syndromic surveillance systems<sup>24–26</sup>, thus, suggesting the feasibility of using EHR data for disease tracking for both local and regional spatial resolutions. In the past, EHR data have not been used for real-time surveillance, due to reporting lag times of 1 to 2 weeks. Near real-time access to EHR records data would address this issue.

<sup>1</sup>Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. <sup>2</sup>Harvard Medical School, Boston, MA, USA. <sup>3</sup>Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA. <sup>4</sup>Harvard School of Public Health, Boston, MA, USA. <sup>5</sup>athenaResearch at athenahealth, Watertown, MA, USA. Correspondence and requests for materials should be addressed to M.S. (email: msantillana@gmail.com)

Here, we demonstrate EHR data collected and distributed in near real-time by an electronic health records and cloud services company, athenahealth, combined with historical patterns of flu activity using a suitable machine learning algorithm, can accurately track real-time influenza activity (as reported by the CDC), at the regional scale in the United States. Additionally, we show that the signal to noise ratio in this data source is high. EHR data provides us with an “early count” of ILI activity in much the same way as exit polls enable us to forecast election results. We build a machine learning model that optimally exploits the data by building a system as timely as GFT used to be, yet as stable and reliable as CDC validated data sources. Although our model is capable of forecasting influenza levels weeks into the future, we decided to present here the real-time monitoring of ILI. We name our model ARES, which stands for AutoRegressive Electronic health record Support vector machine.

## Methods

**Data.** Athenahealth is a provider of cloud-based services and mobile applications for medical groups and health systems (<http://www.athenahealth.com>). Its electronic health records, medical billing, and care coordination services are organized around a single cloud network, allowing for the collection of unique insights related to patient-provider encounter data for more than 72,000 healthcare providers in medical practices and health systems nationwide. This database includes claims data for over 64 million lives and electronic health records for over 23 million lives.

In collaboration with athenahealth, we obtained weekly total visit counts, flu vaccine visit counts, flu visit counts, ILI visit counts, and unspecified viral or ILI visit counts. While these variables are captured and aggregated in near-real time by the athenahealth research team, we obtained the weekly aggregated data of a given Sunday-Saturday week on the immediately subsequent Monday during the 2014–2015 flu-season. This means that athenahealth data was available at least one week ahead of the publication of CDC’s ILI reports.

The athenahealth ILI rates are based on visits to primary care providers on the athenahealth network, for the period between 6/28/09 and 10/15/15. These providers see patients mostly in office-based settings (73.4% of the visits), though they also see patients in the following settings: inpatient hospital (11.3%), outpatient hospital (6.6%), nursing facility (1.5%), emergency room (1.4%), and other (5.8%). The age statistics for the athenahealth network are as follows: younger than 15 years old (15%), 15 to 24 years (6%), 25 to 44 years (14%), 45 to 64 years (29%), and 65 years or older (36%). The place of service and age distribution statistics are close to the statistics reported by the CDC’s National Ambulatory Medical Care Survey (NAMCS). This suggests that medical visits to athenahealth’s medical provider network provide a good representation of ambulatory care, including flu activity, in the US.

There are two main differences to note between the athenahealth and NAMCS statistics. First, providers in the athenahealth network see patients that are slightly older than average, though re-weighting the data change the athenahealth ILI rates very little. Second, emergency departments are underrepresented in the athenahealth network, causing the raw athenahealth %ILI estimates to be consistently lower than the values reported by the CDC. This indicates that the athenahealth data could be adapted to build specialized models that mimic ILI rates.

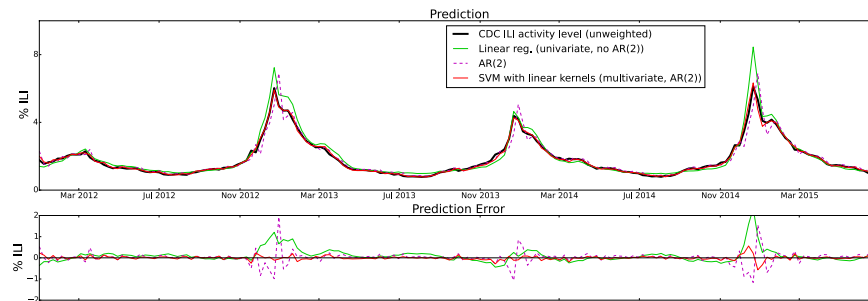
To develop these models, we used the following list of athenahealth variables as input: *Flu Vaccine Visit Count*, which accounts for the number of visits where a flu vaccine was administered; *Flu Visit Count*, the number of visits where the patient had a flu diagnosis; *ILI Visit Count*, the number of visits where the patient had either a flu diagnosis or a fever diagnosis with an accompanying sore throat or cough diagnosis; and *Unspecified Viral or ILI Visit Count*, the number of visits where the patient had either an unspecified viral diagnosis, a flu diagnosis, or a fever diagnosis with an accompanying sore throat or cough diagnosis. It is important to note that *Unspecified Viral or ILI Visit Count* contains *ILI Visit Count*, which itself contains *Flu Visit Count*. In<sup>19</sup> only *ILI Visit Count* was used as an independent variable.

We also obtained the national and regional ILI weekly values from the CDC ([gis.cdc.gov/grasp/fluview/fluportaldashboard.html](http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html)), for the same time period, to use as both a comparator as well as to provide historical independent variables for our models. We note that using weekly information from reports published by the CDC as a gold standard for US national and regional influenza activity may have limitations. An evaluation of the CDC’s data collection approach has been published and improvements suggested in<sup>27</sup>.

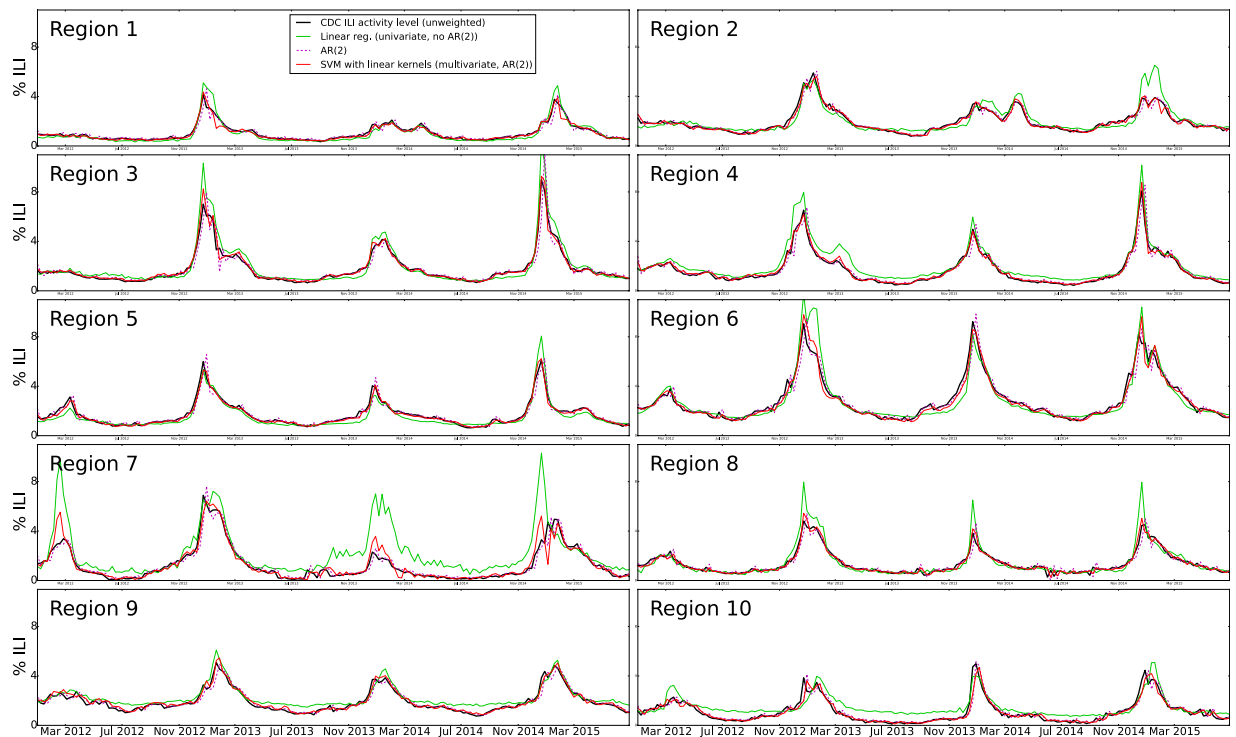
We obtained historical GFT data for the same time period to use as a comparison from the Google Flu Trends website (<http://www.google.org/flutrends>). All of the data used was downloaded on Oct 15, 2015 and all of the experiments were conducted in Python 2.7.

**Analysis.** We built a collection of regional models that dynamically recalibrate every week in order to optimally include all of the available data up to the week of flu estimation. This dynamic recalibration is inspired by data assimilation techniques used in industrial and financial time series forecasting<sup>28</sup> and weather forecasting<sup>29</sup>. The independent variables used by ARES to produce real-time estimates of ILI activity at time  $t$  include: *Unspecified Viral or ILI Visit Count*, *ILI Visit Count*, and *Flu Visit Count* from athenahealth, for weeks  $(t-2)$ ,  $(t-1)$  and  $(t)$ . We also incorporated (autoregressive) historical information from the unweighted percentage ILI estimates obtained from the CDC for weeks  $(t-2)$  and  $(t-1)$ . The variable capturing “flu vaccine visit counts” was not used in the models since it showed no significant correlation with the historical CDC-reported %ILI.

Our ARES models map the variables described above into a %ILI real-time estimate, using a support vector machine (SVM) model<sup>30</sup>. Support vector machine models are similar to multivariate regression models with the important difference that non-linear functions can be learned via the kernel trick, which implicitly maps the independent variables to a higher dimensional feature space. The independent variables can even be mapped to an infinite dimensional feature space with the use of a radial basis function kernel. SVM models are fitted by minimizing an epsilon insensitive cost function where errors of magnitude less than epsilon are ignored by the cost function, leading to better generalization of the learned model.



**Figure 1.** The CDC's ILI estimates, baseline linear regression and AR(2) autoregressive model estimates, and ARES estimates are displayed as a function of time for the national level on the top panel. The errors associated with the linear regression and autoregressive model baselines, and ARES are shown on the bottom panel.



**Figure 2.** The CDC's ILI estimates, baseline linear regression and AR(2) autoregressive model estimates, and ARES estimates are displayed as a function of time for each of the 10 US regions defined by the HHS.

For comparison purposes, we produced real-time estimates using two baseline methodologies: (a) a dynamically-trained autoregressive model that only used historical CDC information, called AR(2) throughout the paper; and (b) a dynamically-trained linear model that used athenahealth's %ILI information onto CDC's ILI, as introduced in<sup>19</sup>. The independent variables used to produce ILI estimates for the week at time  $t$  are: (a) for the AR(2) model: CDC's ILI for weeks  $(t-2)$  and  $(t-1)$ , and (b) for the dynamically trained linear model: the value of athenahealth's ILI value at time  $t$ .

## Results

The training period for our first flu estimate consisted of data from 6/28/2009 through 1/1/2012 for all the implemented models. This choice was made so that our training period was at least two years in length as in<sup>17</sup>. The hyperparameters of our SVM models were chosen via cross-validation on the training data. While the values of these hyperparameters could have been updated dynamically to optimize performance every week, we decided to keep their values constant throughout the out-of-sample estimation time period for reproducibility purposes. We used linear kernels, margin width values of 0.1, and regularization parameter values of 1.0.

Our first out-of-sample real-time estimate of ILI was produced for the week of 1/8/2012. Time series of real-time (out-of-sample) estimates using ARES, the AR(2) model, and the linear model, were generated up to and including the week of June 28, 2015. Figure 1 shows the national level real-time estimates produced by ARES

Algorithm	RMSE			Rel. RMSE (%)			Correlation		
	2012–2013	2013–2014	2014–2015	2012–2013	2013–2014	2014–2015	2012–2013	2013–2014	2014–2015
National									
GFT	2.16	0.39	0.36	62.73%	13.21%	12.31%	0.932	0.968	0.986
Linear (univariate)	0.45	0.20	0.54	12.20%	9.00%	12.22%	0.996	0.985	0.986
AR(2)	0.46	0.30	0.45	11.77%	9.53%	11.40%	0.940	0.940	0.937
SVM (linear) + AR(2)	<b>0.10</b>	<b>0.09</b>	<b>0.17</b>	<b>4.15%</b>	<b>5.12%</b>	<b>4.62%</b>	<b>0.997</b>	<b>0.995</b>	<b>0.991</b>
Region 1									
GFT	2.87	0.27	0.41	107.01%	21.20%	26.01%	0.789	0.881	0.951
Linear (univariate)	0.51	0.22	0.36	25.54%	21.11%	20.80%	<b>0.974</b>	0.965	<b>0.973</b>
AR(2)	0.40	0.23	0.32	18.35%	16.88%	17.31%	0.897	0.856	0.926
SVM (linear) + AR(2)	<b>0.26</b>	<b>0.13</b>	<b>0.24</b>	<b>13.80%</b>	<b>10.44%</b>	<b>10.30%</b>	0.964	<b>0.974</b>	0.960
Region 2									
GFT	2.22	0.64	1.18	46.54%	27.37%	38.64%	0.960	0.833	0.938
Linear (univariate)	0.38	0.49	0.97	14.24%	20.16%	28.98%	<b>0.983</b>	0.877	0.940
AR(2)	0.42	0.27	0.27	11.32%	10.12%	9.73%	0.949	0.922	0.937
SVM (linear) + AR(2)	<b>0.30</b>	<b>0.19</b>	<b>0.23</b>	<b>9.22%</b>	<b>7.02%</b>	<b>8.61%</b>	0.975	<b>0.970</b>	<b>0.956</b>
Region 3									
GFT	1.97	0.33	0.63	78.06%	24.18%	21.19%	0.914	0.984	0.983
Linear (univariate)	0.90	0.36	0.81	22.97%	16.86%	20.51%	<b>0.984</b>	0.986	<b>0.993</b>
AR(2)	0.71	0.24	0.88	19.97%	9.96%	16.59%	0.908	0.965	0.900
SVM (linear) + AR(2)	<b>0.43</b>	<b>0.15</b>	<b>0.26</b>	<b>12.66%</b>	<b>6.57%</b>	<b>7.07%</b>	0.976	<b>0.988</b>	0.992
Region 4									
GFT	1.84	0.36	0.45	58.99%	27.05%	16.71%	0.891	0.958	0.974
Linear (univariate)	1.02	0.38	0.48	47.33%	34.95%	13.33%	0.979	0.973	0.986
AR(2)	0.57	0.37	0.74	15.07%	15.21%	18.29%	0.924	0.941	0.903
SVM (linear) + AR(2)	<b>0.21</b>	<b>0.16</b>	<b>0.27</b>	<b>8.70%</b>	<b>11.41%</b>	<b>9.18%</b>	<b>0.991</b>	<b>0.989</b>	<b>0.989</b>
Region 5									
GFT	2.18	0.36	0.46	63.18%	20.44%	21.70%	0.887	0.962	0.970
Linear (univariate)	0.28	0.43	0.53	13.58%	23.34%	15.22%	<b>0.989</b>	0.951	0.983
AR(2)	0.46	0.33	0.42	12.86%	11.59%	11.95%	0.927	0.886	0.940
SVM (linear) + AR(2)	<b>0.23</b>	<b>0.20</b>	<b>0.18</b>	<b>7.51%</b>	<b>8.22%</b>	<b>6.68%</b>	0.985	<b>0.964</b>	<b>0.991</b>
Region 6									
GFT	3.74	0.75	1.49	66.62%	14.31%	25.84%	0.921	0.968	0.923
Linear (univariate)	1.29	0.73	0.66	21.17%	18.53%	13.38%	0.965	0.964	<b>0.963</b>
AR(2)	0.69	0.58	0.68	14.39%	10.38%	13.31%	0.937	0.949	0.935
SVM (linear) + AR(2)	<b>0.49</b>	<b>0.34</b>	<b>0.57</b>	<b>10.26%</b>	<b>8.90%</b>	<b>10.31%</b>	<b>0.982</b>	<b>0.987</b>	0.957
Region 7									
GFT	0.77	0.92	1.67	22.54%	85.14%	72.62%	0.942	<b>0.968</b>	0.695
Linear (univariate)	0.63	2.43	1.95	29.42%	315.08%	124.32%	0.975	0.958	0.569
AR(2)	0.59	<b>0.24</b>	<b>0.55</b>	16.87%	<b>35.87%</b>	<b>22.99%</b>	0.948	0.910	<b>0.919</b>
SVM (linear) + AR(2)	<b>0.37</b>	0.45	0.81	<b>13.03%</b>	54.24%	29.14%	<b>0.981</b>	0.935	0.827
Region 8									
GFT	0.84	0.39	0.43	27.29%	28.97%	24.09%	0.920	0.951	0.953
Linear (univariate)	0.77	0.54	0.78	23.73%	20.16%	25.38%	0.973	0.942	0.894
AR(2)	0.36	0.41	0.38	15.01%	19.64%	15.78%	0.961	0.843	0.930
SVM (linear) + AR(2)	<b>0.24</b>	<b>0.21</b>	<b>0.26</b>	<b>10.09%</b>	<b>12.99%</b>	<b>11.47%</b>	<b>0.986</b>	<b>0.972</b>	<b>0.970</b>
Region 9									
GFT	2.84	0.80	0.42	72.09%	24.93%	15.59%	0.922	0.946	0.934
Linear (univariate)	0.43	0.38	0.29	19.28%	20.79%	13.37%	<b>0.975</b>	0.927	0.965
AR(2)	0.37	0.29	0.35	11.92%	13.42%	10.82%	0.934	0.935	0.942
SVM (linear) + AR(2)	<b>0.28</b>	<b>0.21</b>	<b>0.21</b>	<b>10.81%</b>	<b>10.57%</b>	<b>7.79%</b>	0.970	<b>0.971</b>	<b>0.981</b>
Region 10									
GFT	2.85	0.73	0.50	181.88%	91.96%	30.46%	0.866	0.953	0.955
Linear (univariate)	0.75	0.49	0.48	125.35%	82.45%	31.04%	0.737	<b>0.976</b>	0.908
AR(2)	0.49	0.57	0.41	37.20%	30.14%	27.99%	0.867	0.881	0.920
SVM (linear) + AR(2)	<b>0.42</b>	<b>0.46</b>	<b>0.34</b>	<b>26.92%</b>	<b>21.07%</b>	<b>21.54%</b>	<b>0.904</b>	0.922	<b>0.947</b>

**Table 1.** Accuracy metrics between ARES and CDC's ILI for all geographic regions, for the three flu seasons spanning 2012–2015. For comparison purposes, we have included GFT's historical predictions, and the two baseline models: dynamic linear regression (mapping athenahealth's ILI onto CDC's ILI), and a two term autoregressive model, AR(2). Values with best performance appear in bold face.

and the target CDC ILI signal. Estimates produced by the two baseline models described in the previous section are included for comparison purposes as well as their respective errors when compared to the CDC ILI. Figures 2 and 3 show the same results but at a regional resolution for each of the 10 regions defined by the Health and Human Services (HHS).

Table 1 shows the accuracy metrics, as defined in<sup>19</sup>, between the models' estimates, GFT, and the target signal, CDC's ILI. For transparency, we displayed the values of these accuracy metrics for each flu season. ARES provides accurate real-time estimates of ILI activity in all ten HHS regions as well as at the national level. The average Pearson correlation across all ten regions and across all flu seasons is 0.972, and the national Pearson correlation is 0.996. The average root mean square error (RMSE) across all ten regions in the whole time period is 0.261, and the national RMSE is 0.10. The average relative RMSE across all ten regions is 18.28%, and the national relative RMSE is 4.63%. The average regional error is heavily biased due to Region 7 inaccurate estimates.

In order to understand the predictive power of each of the independent variables to estimate CDC's ILI, we plotted the values of the coefficient associated with each variable in the (dynamically-trained) linear models as a function of time. These values are presented in the multiple heatmaps of Fig. 1 in the Supplementary materials. These heatmaps show that the variables with highest predictive power are CDC's ILI value during the previous week of flu estimation ( $t-1$ ), athenahealth's viral visit counts during the week of the estimates ( $t$ ), and athenahealth's ILI during the week of the estimates ( $t$ ).

## Discussion

In this study we have shown that EHR data in combination with historical patterns of flu activity and a robust dynamical machine learning algorithm, are capable of accurately predicting real-time influenza activity at the national and regional scales in the US.

While the ability of athenahealth's ILI data to predict CDC's ILI *nationally* was established using a dynamically-trained linear model in<sup>19</sup>, here we show that incorporating CDC's ILI historical information and more of the available EHR information, using a suitable machine learning methodology, can improve flu estimates. Specifically, as shown in the heatmaps of Fig. 1 in the Supporting Material, adding CDC's ILI for the previous week of flu estimation and the variable associated to athenahealth's *viral visits* during the week of the estimates typically improve results.

Flu estimates using ARES lead to 2–3 fold error reductions in the national and regional point estimates, when compared to the dynamic linear model introduced in<sup>19</sup> for mapping athenahealth's ILI data onto CDC's ILI. Pearson correlations improved nationally and in the majority of regions and flu seasons when comparing ARES to the dynamic linear model in<sup>19</sup>. In cases when correlations were higher for the dynamic linear model introduced in<sup>19</sup> over ARES, correlations were comparable in magnitude, showing at least a comparable performance between the two methods. The exceptions to these cases were regions 7 and 10 during the 2013–2014 flu seasons.

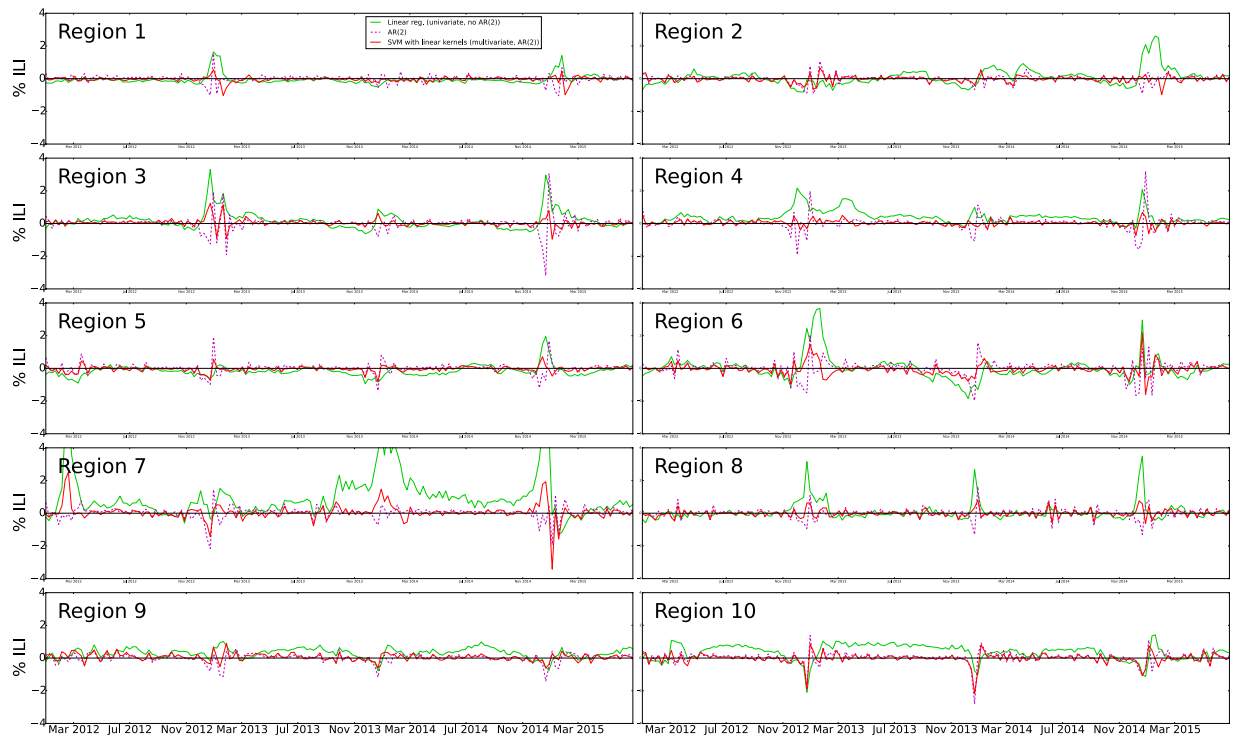
Models that use only historical information to predict future ILI typically show high Pearson correlation and not very large (RMSE) error values when compared with CDC's ILI; however, they consistently show lags of 1 to 2 weeks with respect to the observed CDC's ILI values, making them systematically inaccurate. This happens to our baseline AR(2) model implementation across regions, as shown in Figures 1, 2 and 3. Interestingly, when using ARES to combine CDC's ILI historical information with EHR data, flu estimates improved. The improvement, achieved when historical information is added, has been previously observed in methodologies that use Google searches<sup>17</sup> and Twitter<sup>31</sup> to estimate flu activity. Intuitively speaking, historical information maintains the estimates within a reasonable range, while EHR data improves the timely responsiveness of the model to changes. The improved responsiveness and stability of ARES can be identified when comparing the estimates of ARES with the two baseline methodologies, as shown in Figures 1, 2 and 3.

Specifically, using ARES improved Pearson correlation (from 0.958 to 0.996) and reduced the error (RMSE) three-fold for the national level when compared to the autoregressive, AR(2), model. Pearson correlation across regions was generally improved, with the largest improvement in region 3 (from 0.932 to 0.987) and the mildest taking place in region 2 (from 0.96 to 0.978). The average error (RMSE) generally improved, with the greatest performance in region 4, where more than a two-fold reduction was achieved, and the mildest reduction in region 9, where a 20% reduction in error was achieved.

The only region where the combination of historical CDC data and EHR data did not lead to improvements when compared to the AR(2) model was region 7, where correlations went from 0.958 to 0.938, and the average error (RMSE) went from 0.4 to 0.51 (%ILI). This may be explained by the fact that athenahealth manages very few facilities in the Midwestern United States.

ARES correctly estimated the timing and magnitude of the national peak week of the 3 flu seasons. For 6 out of the 10 regions, ARES correctly estimated the timing of the peak week at least 2 (out of the 3) flu seasons; for regions 2 and 9 ARES only estimated the timing of 1 peak correctly, and for regions 7 and 10, ARES did not capture this feature. When ARES did not get the timing correctly, the estimated peak happened 1.4 weeks off, on average. The peak values were estimated with 0.148% error on the national level, and with 0.445% error on average for the 10 regions.

While the historical predictions of the now discontinued Google Flu Trends system have been shown to contain multiple inconsistencies<sup>13–18</sup>, we decided to include them as an additional source for comparison with ARES in Table 1, as some public health officials still equate digital disease detection systems (such as ARES) with GFT. We found that ARES is capable of estimating national ILI activity with an almost ten-fold reduction in the average error (GFT RMSE = 0.96 and ARES RMSE = 0.1) with respect to the historically reported estimates by the now discontinued GFT system, during our study period. Table 1 shows that ARES is capable of outperforming GFT with significant (5–10 fold) reductions in the average error (RMSE) during the three flu seasons for all the regional flu estimates. Substantial improvements in Pearson correlation are observed from historical GFT near real-time flu estimates and ARES. Nationally, for example, GFT's Pearson correlation is 0.91 while ARES's is 0.996. Overall, ARES consistently outperforms GFT's historical estimates in all statistics.



**Figure 3.** The errors associated with the linear regression and AR(2) autoregressive model baselines, and ARES are displayed as a function of time for each of the 10 US regions defined by the HHS.

We would like to note that in this study we used the *revised* version of historical CDC ILI reports to dynamically train all of our models. While this *revised* information was not technically available at the time the flu estimates were produced (it was probably available a week or two later), our main point in this work is to show the methodological advantages of combining historical data with EHR data to improve flu estimates. This goal is achieved since all models (including the baseline ones) were trained with the same *revised* data. However, our comparison with the historical GFT values may not be strictly fair. It is important to highlight that this decision was not taken lightly. Our experience training near real-time flu prediction models<sup>17,19</sup> has shown us that the results of our predictions training with only the historically available (*revised* and *unrevised*) CDC's ILI values, at the time of prediction, versus training with only *revised* CDC's ILI values change minimally. Specifically, as we have reported in<sup>17</sup>, the performance of our models is almost unchanged (correlation slightly improving from 0.985 to 0.986, for example) for national predictions when using only *revised* historical CDC reports to train them, as opposed to using only the available ones (*revised* and *unrevised*) at the time. See the differences between Table 1 and Table S1 in<sup>17</sup>, for example.

The discrepancies between the flu estimates using ARES and the observed CDC values, as captured by Pearson correlation and RMSE values, are comparable to those shown in<sup>19</sup> where an ensemble approach was used to combine multiple models based on disparate data sources to obtain optimal predictions. The resulting ensemble real-time flu estimates outperformed any estimates produced with any single-source model<sup>19</sup>. We suspect that using ARES as input in a new ensemble approach will lead to improved predictions. Real-time flu estimates produced using ARES can be found in the website: <http://www.healthmap.org/flu-trends/>. In this live implementation, the hyperparameters of the SVM linear model are chosen dynamically with 10 fold cross validation. The predictions with dynamic hyperparameters have historically led to similar performance to the flu estimates presented in this study, perhaps with slight improvements.

Finally, ARES avoids many of the limitations faced by most non-traditional digital disease surveillance systems, such as a lack of specificity, while providing the main advantage of non-traditional systems, timeliness. The most limiting aspect of ARES is that it relies on electronic health records, data that are not generally publically available. At the time of the writing of this report, athenahealth data is provided to several groups of flu researchers around the country but is not publically available. Future work includes integrating ARES into a network model and testing the accuracy of ARES at the state and city levels, in other countries, as well as on other communicable diseases.

## Conclusion

We have shown that EHR data in combination with historical patterns of flu activity and a robust dynamical machine learning algorithm provide a novel and promising way of monitoring infectious diseases at the national and local level. Our methodology provides timely flu estimates with the accuracy and specificity of sentinel systems like the CDC's ILI surveillance network. This demonstrates the value of cloud-based electronic health records databases for public health surveillance at the local level.

## References

- Centers for Disease Control and Prevention. Estimating Seasonal Influenza-Associated Deaths in the United States: CDC Study Confirms Variability of Flu. March 18, 2015. Accessed June 21, (2015) [http://www.cdc.gov/flu/about/disease/us\\_flu-related\\_deaths.htm](http://www.cdc.gov/flu/about/disease/us_flu-related_deaths.htm).
- Brownstein, J. S. & Mandl, K. D. Reengineering real time outbreak detection systems for influenza epidemic monitoring. *Am Med Inform Assoc, Annual Symposium Proceedings* vol. 2006, p. 866 (2006).
- Centers for Disease Control and Prevention. Overview of Influenza Surveillance in the United States. January 27, 2015. Accessed June 21, <http://www.cdc.gov/flu/weekly/overview.htm>, (2015).
- Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D. & Weinstein, R. A. Using internet searches for influenza surveillance. *Clin Infect Dis.* **47**(11), 1443–1448, doi: 10.1086/593098 PMID: 18954267 (2008).
- Broniatowski, D. A., Paul, M. J. & Dredze, M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012–2013 Influenza Epidemic. *PLoS ONE* **8**(12), e83672, doi: 10.1371/journal.pone.0083672 (2013).
- Lamb, A., Paul, M. J. & Dredze, M. Separating Fact from Fear: Tracking Flu Infections on Twitter. *Proc of HLT-NAACL* **13**(1), 789–795 (2013).
- Santillana, M., Nsoesie, E. O., Mekaru, S. R., Scales, D. & Brownstein, J. S. Using Clinicians' Search Query Data to Monitor Influenza Epidemics. *Clin Infect Dis.* **59**(10), 1446–1450, doi: 10.1093/cid/ciu647 PMID: 25115873 (2014).
- McIver, D. J. & Brownstein, J. S. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLoS Comput. Biol.* **10**, e1003581, doi: 10.1371/journal.pcbi.1003581 PMID: 24743682 (2014).
- Smolinski, M. S. *et al.* Flu Near You: Crowd-sourced Symptom Reporting Spanning Two Influenza Seasons. *Am J Public Health* **105**(10), e1–e7 (2015).
- Yuan, Q. *et al.* Monitoring influenza epidemics in China with search query from Baidu. *PLoS One* **8**, e64323, doi: 10.1371/journal.pone.0064323 PMID: 23750192 (2013).
- Nagar, R. *et al.* A Case Study of the New York City 2012–2013 Influenza Season With Daily Geocoded Twitter Data From Temporal and Spatiotemporal Perspectives. *J Med Internet Res.* **16**(10), e236, doi: 10.2196/jmir.3416 (2014).
- Ginsberg, J. *et al.* Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014, doi: 10.1038/nature07634 PMID: 19020500 (2009).
- Cook, S., Conrad, C., Fowlkes, A. L. & Mohebbi, M. H. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE* **6**, e23610, doi: 10.1371/journal.pone.0023610 PMID: 21886802 (2011).
- Copeland, P. *et al.* Google disease trends: an update. *Int Soc Negl Trop Dis.* **3**, (2013).
- Lazer, D. M., Kennedy, R., King, L. & Vespignani, A. The parable of Google flu: traps in big data analysis. *Science* **343**, 1203–1205, doi: 10.1126/science.1248506 PMID: 24626916 (2014).
- Santillana, M., Zhang, D. W., Althouse, B. M. & Ayers, J. W. What can digital disease detection learn from (an external revision to) Google flu trends? *Am. J. Prev. Med.* **47**, 341–347, doi: 10.1016/j.amepre.2014.05.020 PMID: 24997572 (2014).
- Yang, S., Santillana, M. & Kou, S. C. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci.* **112**, no. 47, 14473–14478, doi: 10.1073/pnas.1515373112 (2015).
- Lampos, V., Miller, A. C., Crossan, S. & Stefansen, C. Advances in nowcasting influenza-like illness rates using search query logs. *Sci. Rep.* **5**, doi: 10.1038/srep12760 (2015).
- Santillana, M. *et al.* Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol* **11**, no. 10, e1004513, doi: 10.1371/journal.pcbi.1004513 (2015).
- Lipsitch, M., Finelli, L., Heffernan, R. T., Leung, G. M. & Redd, S. C. Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. *Biosecurity and bioterrorism: biodefense strategy, practice, and science* **9**, no. 2, 89–115, doi: 10.1089/bsp.2011.0007 (2011).
- Yang, W., Lipsitch, M. & Shaman, J. Inference of seasonal and pandemic influenza transmission dynamics. *Proc Natl Acad Sci.* **112**, no. 9, 2723–2728, doi: 10.1073/pnas.1415012112 (2015).
- 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**, doi: 10.1038/srep08154 (2015).
- Olson, D. R. *et al.* Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput Biol* **9**, no. 10, e1003256, doi: 10.1371/journal.pcbi.1003256 (2013).
- Lazarus, R., Kleinman, K. P., Dashevsky, I., DeMaria, A. & Platt, R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC public health* **1**, no. 1, 9, doi: 10.1186/1471-2458-1-9 (2001).
- Hripcsak, G. *et al.* Syndromic surveillance using ambulatory electronic health records. *J Am Med Inform Assoc.* **16**, no. 3, 354–361, doi: 10.1197/jamia.M2922 (2009).
- Viboud, C. *et al.* Demonstrating the use of high-volume electronic medical claims data to monitor local and regional influenza activity in the US. *PLoS One*, e102429, doi: 10.1371/journal.pone.0102429 (2014).
- Scarpino, S. V., Dimitrov, N. B. & Meyers, L. A. Optimizing provider recruitment for influenza surveillance networks. *PLoS Comput Biol* **8**, no. 4, e1002472, doi: 10.1371/journal.pcbi.1002472 (2012).
- Burkom, H. & Murphy, S. P. Automated time series forecasting for biosurveillance. *Statistics in Medicine, Stat Medic* **26**(22), 06–035 (2007).
- Ghil, M. & Malanotte-Rizzoli, P. Data assimilation in meteorology and oceanography. *Adv. Geophys* **33**, 141–266, doi: 10.1016/S0065-2687(08)60442-2 (1991).
- Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat Comput* **14**, no. 3, 199–222 (2004).
- Paul, M. J., Dredze, M. & Broniatowski, D. Twitter Improves Influenza Forecasting. *PLoS currents* **6**, doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117 (2014).

## Author Contributions

M.S., J.S.B., I.S. and J.G. conceived the research. I.S., A.Z. and J.G. collected and prepared the aggregated electronic health records data. M.S., A.T.N. and T.L. performed the analysis and developed the predictive algorithms. M.S. and A.T.N. wrote the main manuscript text. T.L. prepared the figures and tables. All authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** I.S., A.Z. and J.G. were employed by athenahealth at the time this research was performed. M.S., A.T.N., T.L. and J.S.B. declare no potential conflict of interest.

**How to cite this article:** Santillana, M. *et al.* Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. *Sci. Rep.* **6**, 25732; doi: 10.1038/srep25732 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>