Abstract
Background: Traditionally, public health agencies track seasonal influenza activity by collecting information from clinics, hospitals, and laboratories. The inherent slowness of the processes used to collect influenza activity data limits the ability of public health agencies to adapt to unexpected changes in influenza activity in near real-time. In recent years, new influenza surveillance methods that use nontraditional data sources, such as Google searches, have been proposed to successfully estimate influenza activity in near real-time. However, most of these methods have been designed for and implemented in high-income countries even though influenza disease burden remains high in low- to middle-income countries. Here, we seek to predict influenza activity in near real-time in Africa using machine learning models that combine Google searches with traditional epidemiological data.

Methods: We extend the AutoRegression with Google search data (ARGO) model to track influenza activity in near-real-time in Africa. The ARGO model, which was originally designed to predict influenza activity in the United States, combines influenza-related Google searches with historical laboratory-confirmed influenza trends. We evaluate the predictive performance of the ARGO model and compare it with several benchmark models in Algeria, Ghana, Morocco, and South Africa. We also explore the advantages and limitations of using Google search data to monitor influenza activity.

Results: In South Africa, Algeria, and Morocco, the ARGO model outperforms all benchmark models, suggesting that incorporating influenza-related Google search information in predictive models in these countries leads to improved predictions. In Ghana, however, the ARGO model and the autoregressive model of historical influenza activity have comparable performances.

Conclusions: These results demonstrate that the quality of the ARGO predictions is higher in regions where influenza activity is seasonal, historical influenza activity is recorded consistently, and the volume of influenza-related Google search queries is enough to appear as non-zero in the Google Trends tool.
Keywords
Real-time disease surveillance, Digital epidemiology, Google Flu Trends, Influenza monitoring, Seasonal influenza

Corresponding author: Mauricio Santillana (msantill@g.harvard.edu)

Author roles: Mejia K: Data Curation, Formal Analysis, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Viboud C: Funding Acquisition, Writing – Review & Editing; Santillana M: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The study was funded in part by the Bill and Melinda Gates Foundation (OPP 1195154).

Copyright: © 2019 Mejia K et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

How to cite this article: Mejia K, Viboud C and Santillana M. Leveraging Google search data to track influenza outbreaks in Africa [version 1; peer review: awaiting peer review] Gates Open Research 2019, 3:1653 (https://doi.org/10.12688/gatesopenres.13072.1)

Introduction

According to estimates from the U.S.A. Centers of Disease Control and Prevention (CDC), up to 650,000 people around the world die each year from complications related to seasonal influenza. The primary methods of influenza prevention include vaccination and good health habits, such as hand washing and covering one’s mouth when sneezing. Furthermore, populations must be revaccinated annually to keep up with the rapid evolution of influenza virus strains. The influenza vaccine is used routinely in high-income countries but less so in low- to middle-income countries. The constant threat of an influenza pandemic highlights the need for accurate influenza surveillance methods to guide local and global policies for preventing and mitigating the spread of influenza.

Traditionally, public health agencies, such as the U.S.A. CDC, track influenza activity by collecting respiratory specimen information and syndromic data from hospitals, clinics, and laboratories and analyzing them for the presence of the influenza virus. Unfortunately, these important and necessary healthcare-based approaches lag behind influenza activity in real-time by at least one or two weeks because of the inherent processing times of lab testing, data collection, data aggregation, and quality control. Indeed, first, an infected patient visits a clinic or hospital where health providers obtain a specimen to send to laboratories for testing. Then, the laboratory tests the specimen for influenza and the results of the specimen are reported to public health agencies which aggregate the results to the regional and national levels. This time delay limits the ability of public health agencies to respond to sudden and unexpected changes in influenza activity in real-time. To address this problem, researchers in the past decade have introduced influenza surveillance methods that combine information from nontraditional Internet-based data sources to estimate influenza activity in near-real-time.

In 2008, Google launched Google Flu Trends (GFT), an Internet-based tool which leveraged Google search queries of influenza-related terms to estimate influenza activity in near real-time in the United States. In subsequent years, GFT expanded to provide influenza activity estimates for more than 25 countries. Despite early interest around GFT, the failure of GFT to capture the influenza pandemic of 2009 as well as the tendency of GFT to overestimate peak influenza activity in the 2013-2014 influenza season led to public speculation of the value of using nontraditional influenza surveillance methods. As described by Santillana, Google discontinued GFT in 2015 after multiple revisions to their original approach.

In response to the shortcomings of GFT, researchers introduced promising methodological improvements to Google’s original approach. However, most of the proposed methods have been implemented in countries in the Americas and Europe even though sub-Saharan Africa, the western Pacific, and Southeast Asia have high influenza-related mortality rates. Surveillance methods that can help estimate influenza activity in near real-time in countries where traditional viral surveillance is sparse or delayed could be useful to local public health officials. In addition, although the use of the Google Search tool and social media preferences vary geographically, we expect the reliability of such methods to increase over time as Internet coverage continues to increase rapidly in these regions.

Our contribution

We extend the AutoRegression with Google search data (ARGO) model introduced by Yang et al. to track influenza viral activity in near real-time in four African countries: Algeria, Ghana, Morocco, and South Africa. The predictions of the ARGO model rely on a combination of influenza-related Google searches and historical influenza trends. In order to understand the value of Google search information in monitoring influenza activity, we evaluate the out-of-sample prediction performance of the ARGO model during the study period and compare it with benchmark predictive models that only use either historical influenza activity or Google search information. Our findings suggest that incorporating influenza-related Google search information in predictive models leads to improved predictions in three out of the four African countries we studied.

Methods

Influenza surveillance data

We measured influenza activity using data from FluNet, a publicly available influenza surveillance database maintained by the World Health Organization (WHO) that tracks the weekly number of processed respiratory specimens (NPS) and confirmed influenza cases (NCC) in countries around the world. The NPS is the number of specimens from patients with influenza-like symptoms processed at national laboratories by type and subtype while the NCC is the subset of the NPS that tested positive for influenza. We obtained all available weekly FluNet reports from January 8, 2012 to October 7, 2018 for Algeria, Ghana, Morocco, and South Africa. The remaining African countries were excluded because they either had a large proportion of missing data (see Extended data, fluenet heatmap and Supplementary Figure 1) during the study period or lacked influenza-related Google search data.

We used the complete case ratio (CCR) to measure influenza activity in each country. The CCR is defined as the NCC normalized by the NPS from the entire influenza season. Specifically, the CCR in week j of influenza season i is equal to the NCC in week j of influenza season i divided by the NPS in influenza season i as shown below:

\[
\text{CCR}_{ij} = \frac{\text{NCC}_{ij}}{\text{NPS}_{ij}}
\]

Hospitals in resource-limited regions frequently have limited numbers of influenza testing kits during each influenza season. Consequently, changes in the NPS and NCC over time may track the available number of testing kits and not actual changes in influenza activity. By using the CCR, we monitor the changes in the total confirmed influenza cases relative to the number of suspected cases or available testing kits.

Google search data

We used Google Correlate (GC) or Google Trends (GT) to identify influenza-related Google search terms in Algeria, Ghana, Morocco, and South Africa. GC and GT are publicly available.
tools developed by Google for analyzing Google Search data. GC allows users to obtain the most correlated search terms to any user-provided time series while GT allows users to input terms and obtain the most popular search queries made by users who also searched for the inputted terms. Since Morocco is the only African country supported by GC, we used the time series of the NPS reported in Morocco from January 6th, 2008 to January 6th, 2013 to obtain the most correlated search terms in Morocco. For the remaining countries, we used GT to obtain the most popular search queries made by users who also searched for influenza, influenza symptoms, and influenza-like illnesses in each country’s top official languages from January 6th, 2008 to January 6th, 2013. We manually excluded spurious terms such as “yuppie flu” and “rsvp” as well as Arabic terms since we are not literate in Arabic. See Extended data (Search terms and Supplementary Tables 1–4) for a list of the Google search terms identified for Algeria, Ghana, Morocco, and South Africa.

We accessed the Google Health Trends (GHT) API on November 12, 2018 to obtain the search frequencies from January 8, 2012 to December 31, 2017 of the search terms identified for Algeria, Ghana, Morocco, and South Africa. GHT is a private application programming interface (API) developed by Google that allows academic researchers to obtain the relative search frequencies of any search term. The search frequencies from GHT are relative because they represent the proportions of searches from a sample of all Google searches. We used the search frequencies as a proxy for public concern for influenza under the assumptions that search activities reflect public concern and that public concern for influenza is positively associated with influenza activity.

The search frequencies from GHT are left-censored because the API sets the search frequencies to zero when the number of searches does not exceed the privacy threshold set by Google to prevent researchers from dissecting the searches of individual users. As a result, models based on search terms with search frequencies that include zeros may underestimate the importance of those search terms because zero values do not always represent zero searches. In addition, since the results from the API only represent a sample of all Google searches, then results for search terms with low search volumes will have higher variability and thus introduce more error than results for search terms with high search volumes. To minimize error due to sampling variability, we calculated the proportion of zeros for each search term over the initial training period from January 8, 2012 to January 5, 2014 and manually excluded search terms with proportions of zeros exceeding 0.25.

Models
We implemented the ARGO model and several benchmark models to obtain retrospective out-of-sample estimates of the weekly CCR in Algeria, Ghana, Morocco, and South Africa under the assumption that we only had access to data available at the time of estimation. More specifically, we assumed that in order to produce predictions for time \( t \), we had access to epidemiological data for the prior week of estimation (i.e., week \( t-1 \)) and influenza-related Google search activity up to the week of estimation (i.e., week \( t \)). For each country, the training and test periods were based on the availability of the historical influenza activity data reported by FluNet. We used moving training windows of 52 weeks for Morocco and moving training windows of 104 weeks for Algeria, Ghana, and South Africa. All analyses were implemented in Python 3.6.2 using scikit-learn version 0.20.0\(^27\). Underlying data shows the data used in this analysis\(^28\).

ARGO model
The ARGO model introduced by Yang et al. (2015) is a dynamic penalized multivariate linear regression that combines historical influenza activity data with influenza-related Google search frequencies to predict influenza activity in the United States\(^27\). We adapted the ARGO framework to predict influenza activity in Algeria, Ghana, Morocco, and South Africa by combining historical influenza activity data reported by FluNet with influenza-related Google search frequencies. The ARGO model is motivated by two assumptions: (a) unobserved influenza activity at time \( t \) depends on previously observed influenza activity, and (b) changes in influenza activity at time \( t \) lead to changes in influenza-related Google search activity at time \( t \). In other words, we assume that if more people are affected by influenza, then more people will search for influenza-related information on the Internet. These assumptions can be formalized as a hidden Markov model

\[
\begin{align*}
y_i, \mu_i, \beta_i, N &\rightarrow y_i, 2N + 1 \rightarrow \ldots \rightarrow y_i, (T - N + 1)T \\
\downarrow & \downarrow \\
X_i, N &\rightarrow X_i, N + 1 \rightarrow \ldots \rightarrow X_i, T
\end{align*}
\]

which results in the following mathematical representation of the ARGO model

\[
y_i = \mu_i + \sum_{j = 1}^{K} \alpha_j y_{i-j} + \sum_{k = 1}^{K} \beta_k X_{i-k} + \epsilon_i, \epsilon_i \sim N(0, \sigma) \tag{2}
\]

where \( y_i \) represents the CCR, \( X_{i-k} \) represents the standardized Google search frequencies of term \( k \) at week \( t \), and \( \epsilon_i \) represents the noise which captures the complexity of the observed data that cannot be explained by the model. \( X_{i-k} \) is standardized to have mean zero and variance one. \( \epsilon_i \) is assumed to be a Gaussian white noise process with mean zero and constant variance.

Since the number of predictors may exceed the number of observations, using least squares to estimate the model parameters will result in overfitting due to the curse of dimensionality. For that reason, the ARGO model imposes an \( L_1 \) penalty to estimate the parameters \( \beta_i = (\beta_1, \ldots, \beta_K) \) that minimize the objective function

\[
\sum_{i = 1}^{T} (y_i - \mu_i - \sum_{j = 1}^{K} \alpha_j y_{i-j} - \sum_{k = 1}^{K} \beta_k X_{i-k})^2 + \lambda_\alpha |\alpha| + \lambda_\beta |\beta| \tag{3}
\]

with a moving training window and hyperparameters \( \lambda_\alpha \) and \( \lambda_\beta \). For simplicity, we restricted the hyperparameters to \( \lambda_\alpha = \lambda_\beta \) and...
used 10-fold cross validation to choose the hyperparameters that produce the minimum mean squared error on the training set.

Performance metrics
We assessed model performance with the root mean squared error (RMSE), mean absolute error (MAE), and Pearson correlation (CORR).

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}
\]

(4)

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|
\]

(5)

\[
\text{CORR} = \text{Corr}(\hat{y}_i, y_i) = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

(6)

The RMSE and MAE measure how well a model describes the historical influenza activity data while the CORR measures how well a model tracks the trends in the influenza activity data. We used the RMSE as the primary metric for assessing model performance because the objective function of each model minimizes the RMSE. A model that describes historical influenza activity well will have RMSE and MAE values closer to zero, while a model that tracks historical influenza activity trends well will have a CORR value closer to one. For each model, we calculated the performance metrics between predicted and official influenza activity over the test period from January 6th, 2014 to January 5th, 2018.

Benchmark models
In addition to computing performance metrics, we compared the performance of ARGO to three benchmark models. We implemented an autoregressive (AR) model based on historical influenza surveillance data, a penalized multivariate regression model that only relies on GT data, and the historical predictions from GFT which are only available for South Africa.

AR model. We constructed the AR model using historical influenza activity data from FluNet to nowcast weekly influenza activity. The AR model assumes that unobserved influenza activity in the current week depends on previously observed influenza activity by representing unobserved influenza activity at week t as a linear combination of previously observed influenza activity plus some noise. The AR model of order p can be represented mathematically as

\[
y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + w_t, w_t \sim N(0, \sigma^2)
\]

(7)

where p ∈ N and \(\alpha_1, \ldots, \alpha_p \in \mathbb{R}\) such that \(\alpha_i \neq 0\) are the autoregressive coefficients. As in the ARGO model, \(y_t\) in the AR model represents the historical influenza activity and the noise \(w_t\) is a Gaussian white noise process with mean zero and constant variance.

We used partial autocorrelograms (given in Extended data, Supplementary Figure 2) to choose the number of autoregressive terms \(p\) to include in the AR model for each country. Partial autocorrelograms plot the empirical partial autocorrelation \(\hat{p}(h)\) for different autoregressive terms \(h\) where the partial autocorrelation refers to the correlation between a variable and its \(h\)-th autoregressive term after removing the linear dependence of the shorter autoregressive terms.

\[
\hat{p}(h) = \text{Corr}(y_{t}, y_{t-h}|y_{t-1}, \ldots, y_{t-h+1})
\]

(8)

In other words, the partial autocorrelation between a variable and its \(h\)-th autoregressive term quantifies the degree to which they are linearly related after removing the impact of the shorter autoregressive terms. We modeled influenza activity Algeria, Ghana, Morocco, and South Africa with an AR(1) model because only the partial autocorrelation for the first autoregressive term was significantly different from zero for each country. We trained the AR model using linear regression with a moving time window to compare the AR model to the ARGO model. Since the ARGO model is an autoregressive model that leverages exogenous GT data, comparing the ARGO model to the AR model will demonstrate the additional predictive value of including the GT data.

GT model. We implemented a GT model that uses influenza-related Google search frequencies to predict influenza activity without taking historical influenza activity data into account. The GT model assumes that the observed frequencies of influenza-related Google search terms at week \(t\) can be used as a proxy for influenza activity at week \(t\) by representing unobserved influenza activity at week \(t\) as a linear combination of the search frequencies of the Google search terms at week \(t\) plus some noise. The GT model can be represented mathematically as

\[
y_t = \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_p X_{p,t} + v_t, v_t \sim N(0, \sigma^2)
\]

(9)

where \(y_t\) represents historical influenza activity at time \(t\), \(X_i\) represents the Google search frequency of term \(i\) at week \(t\), and the noise \(v_t\) is a Gaussian white noise process with mean zero and constant variance. We trained the GT model using LASSO regression with a moving time window and dynamic variable transformations. Comparing the GT model to the ARGO model will demonstrate the predictive value of the GT data in the absence of autoregressive information.

GFT. GFT for South Africa, which launched in 2010 as a result of a collaboration between Google and the National Institute for Communicable Diseases (NICD), combined Google search data with influenza surveillance data from the NICD to track influenza activity in South Africa at the national level as well as the provincial level for the provinces of Gauteng, KwaZulu-Natal, and the Western Cape. Although GFT for South Africa was discontinued in 2015, the historical GFT estimates for South Africa from January 1st, 2006 to August 9th, 2015 remain publicly available. Since the estimates reported by GFT are not on the same scale as those reported by ARGO, we dynamically rescaled the GFT estimates
to fit the historical influenza activity data from FluNet using the same moving time windows applied to the other models. Comparing historical GFT estimates to ARGO estimates will demonstrate whether ARGO avoids the pitfalls of GFT.

Results

Algeria

As shown in Table 1, the ARGO model outperforms the AR and GT models across every metric (RMSE, MAE, and CORR) from January 5, 2014 to January 3, 2016 in Algeria. The ARGO model yields a 7% lower RMSE than the AR model and a 35% lower RMSE than the GT model. As shown in Figure 1, the prediction error of each model increases during the periods leading up to and following the peak observed in the official CCR around May 2016 because the peak estimated by the GT model anticipates the peak observed in the official CCR while the peak estimated by the AR and ARGO models lags behind the peak observed in the official CCR. The fact that the Google search data anticipate the peak in the official CCR reported by FluNet, unlike the AR and ARGO models, suggests that influenza-related Google searches could potentially provide valuable information but that the observed volume of influenza-related Google searches in Algeria may be too low to be informative. The heatmap of the ARGO coefficients for each predictor over time in Figure 1 shows that although the ARGO estimates consistently rely on the search frequency of ‘grippe’ and the first autoregressive term, the weight of the first autoregressive term is always approximately twice that of the search frequency of ‘grippe’. Furthermore, some influenza-related search terms such as ‘vaccin’ and ‘la grippe’ have negative coefficient values before December 2012. These results further confirm that the observed influenza-related Google searches do not provide valuable information for estimating the CCR in Algeria even though the ARGO model yields a slightly lower RMSE than the AR model.

Ghana

In Ghana, the ARGO and AR models outperform the GT model across all performance metrics from January 3, 2014 to December 31, 2017. The ARGO and AR models perform equally well according to the RMSE and MAE, but the AR model outperforms the ARGO model according to the CORR. The AR and ARGO models yield a 30% lower RMSE than the GT model which suggests that the Google search data do not provide valuable information for estimating the CCR in Ghana.

Table 1. Comparison of performance metrics of the AutoRegression with Google search data (ARGO), Google Trends (GT), and autoregressive (AR) models in South Africa, Algeria, Morocco, and Ghana. For each metric, the bold values denote the best performance among all models. Google Flu Trends (GFT) estimates were only available for South Africa.

<table>
<thead>
<tr>
<th></th>
<th>Full Period</th>
<th>GFT Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>South Africa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GFT</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AR</td>
<td>0.0017</td>
<td>0.001</td>
</tr>
<tr>
<td>GT</td>
<td>0.0022</td>
<td>0.0016</td>
</tr>
<tr>
<td>ARGO</td>
<td>0.0016</td>
<td>0.0011</td>
</tr>
<tr>
<td>Algeria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>0.0061</td>
<td>0.0033</td>
</tr>
<tr>
<td>GT</td>
<td>0.0088</td>
<td>0.006</td>
</tr>
<tr>
<td>ARGO</td>
<td>0.0057</td>
<td>0.0032</td>
</tr>
<tr>
<td>Ghana</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>0.0016</td>
<td>0.0012</td>
</tr>
<tr>
<td>GT</td>
<td>0.0023</td>
<td>0.0018</td>
</tr>
<tr>
<td>ARGO</td>
<td>0.0016</td>
<td>0.0012</td>
</tr>
<tr>
<td>Morocco</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>0.0032</td>
<td>0.002</td>
</tr>
<tr>
<td>GT</td>
<td>0.004</td>
<td>0.0031</td>
</tr>
<tr>
<td>ARGO</td>
<td>0.0029</td>
<td>0.0021</td>
</tr>
</tbody>
</table>
Figure 1. The estimated complete case ratio in Algeria from the AutoRegression with Google search data (ARGO), Google Trends (GT), and autoregressive (AR) models. The complete case ratio from FluNet, the prediction errors associated with each model, and the heatmap of the ARGO coefficients over time are shown for reference.
As shown in Figure 2, the estimated CCR from the GT model over time does not correspond to the observed CCR and the ARGO estimates rely almost entirely on the autoregressive data which confirms the relative unimportance of the Google search data compared to the autoregressive data. These results are not surprising considering that most of the Google search terms for Ghana are related to non-influenza infections generating influenza-like symptoms, such as yellow fever and meningitis.

**Morocco**

In Morocco, the ARGO model outperforms every benchmark model according to the RMSE and CORR while the AR...
model outperforms the ARGO and GT models according to the MAE from September 4, 2016 to December 31, 2017. The ARGO model yields a 9% lower RMSE than the AR model and a 28% lower RMSE than the GT model. Even though the heatmap of the ARGO coefficients over time in Figure 3 shows that 'vaccin' and 'grippe' have a non-zero weight over time, the weight of the first autoregressive term is consistently approximately twice that of 'vaccin' and 'grippe'. In addition, the GT and ARGO models tend to overestimate the CCR during period of low volumes which may be a symptom of having

![Figure 3](image_url)

**Figure 3.** The estimated complete case ratio in Morocco from the AutoRegression with Google search data (ARGO), Google Trends (GT), and autoregressive (AR) models. The complete case ratio from FluNet, the prediction errors associated with each model, and the heatmap of the ARGO coefficients over time are shown for reference.
small sample sizes in both the observed CCR values as well as the influenza-related Google searches in Morocco.

South Africa
In South Africa, the ARGO model outperforms every benchmark model according to the RMSE and CORR while the AR model outperforms the AR and GT models according to the MAE from January 5, 2014 to December 3, 2017. In particular, the ARGO model outperforms GFT in South Africa during the period from January 5, 2014 to August 9, 2015 when GFT published influenza activity estimates for South Africa. The ARGO model yields a 6% lower RMSE than the AR model and a 27% lower RMSE than the GT model during the full period. During the GFT period, the ARGO model yields a 5% lower RMSE than the AR model, a 27% lower RMSE than the GT model, and a 59% lower RMSE than GFT. As shown in Figure 4, the GT model tends to underestimate the official CCR during periods of peak influenza activity and occasionally under- or overestimate the

---

**Figure 4.** The estimated complete case ratio in South Africa from the AutoRegression with Google search data (ARGO), Google Trends (GT), and autoregressive (AR) models. The complete case ratio from FluNet, the estimates from Google Flu Trends (GFT), the prediction errors associated with each model, and the heatmap of the ARGO coefficients over time are shown for reference.
offical CCR during periods of low influenza activity. The heatmap of the ARGO coefficients over time in Figure 4 shows that the relationship between the official CCR reported by FluNet and the Google search data changes over time. For example, the search frequency of 'swine flu' was as important as the autoregressive terms for predicting the CCR from August 2015 to May 2017 but not as important afterwards. The heatmap of the ARGO coefficients also visualizes how the ARGO model reduces dimensionality through $L_1$ regularization.

**Discussion**

Influenza surveillance methods that provide accurate influenza activity estimates in near real-time can help public health agencies adapt quickly to unexpected changes in influenza activity to minimize the ever-present risk of an influenza pandemic. We demonstrated that ARGO, which was originally conceived by Yang et al.\textsuperscript{17} to estimate influenza activity in near real-time in the United States, can be extended to monitor influenza activity in African countries. Specifically, our results show that ARGO outperforms every benchmark model according to the RMSE in South Africa, Algeria, and Morocco which are countries with clear seasonal influenza activity. However, in countries where reported influenza activity seemed erratic (i.e., with sharp changes in influenza case counts from week to week), ARGO predictions did not properly track the observed case counts. Another important observation to highlight in our modeling efforts is that the GT model, which only relies on synchronous Google search activity, can track the salient features of influenza activity in Algeria, Morocco and South Africa. This finding suggests that, in the absence of timely healthcare-based influenza estimates, public health officials could use near real-time estimates derived from Google search activity to get a sense of the unfolding trends in an ongoing outbreak.

The performance of the ARGO model is limited by the low sample size in both the observed historical influenza data and the influenza-related Google searches. As shown in Table 2, the median yearly NCC of South Africa, Algeria, Morocco, and Ghana ranges from 632 in Algeria to 7,388 in South Africa. Also, the number of influenza-related search terms identified in these countries ranges from 8 search terms in Morocco to 102 search terms in South Africa. To put that into perspective, the number of influenza-related search terms identified in the United States was 100\textsuperscript{17} and the median yearly NCC in the United States from 2014 to 2018 is at least twice that of South Africa. It is also worth noting that the study period of the historical influenza surveillance data and Google search data varied substantially between countries, from one year in Morocco to five years in South Africa, which affects the robustness of model predictions.

ARGO performed best in countries with seasonal influenza activity, increased Internet access, and higher sampling of both historical influenza and Google search data as shown in Table 2. However, it appears that the effect of Internet access on ARGO performance is modified by the literacy rate because ARGO performance is not as impressive in Morocco compared to South Africa even though both Morocco and South Africa have comparable Internet penetration rates and strong seasonality in influenza activity. Interestingly, although the ARGO model

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>South Africa</th>
<th>Algeria</th>
<th>Morocco</th>
<th>Ghana</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARGO correlation</td>
<td>0.91</td>
<td>0.87</td>
<td>0.84</td>
<td>0.70</td>
</tr>
<tr>
<td>Median yearly case count</td>
<td>7,388</td>
<td>632</td>
<td>1,337</td>
<td>3,207</td>
</tr>
<tr>
<td>Seasonality</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Internet penetration</td>
<td>54%</td>
<td>43%</td>
<td>58%</td>
<td>35%</td>
</tr>
<tr>
<td>Google market share</td>
<td>95%</td>
<td>97%</td>
<td>97%</td>
<td>89%</td>
</tr>
<tr>
<td>Literacy rate</td>
<td>99%</td>
<td>80%</td>
<td>69%</td>
<td>77%</td>
</tr>
<tr>
<td>Average population (millions)</td>
<td>57.5</td>
<td>42.2</td>
<td>35.7</td>
<td>28.3</td>
</tr>
<tr>
<td>Country size ($10^3$ mi$^2$)</td>
<td>471</td>
<td>919.6</td>
<td>274.5</td>
<td>92.5</td>
</tr>
<tr>
<td>Population (millions / mi$^2$)</td>
<td>47</td>
<td>15.9</td>
<td>50</td>
<td>262.9</td>
</tr>
<tr>
<td>GDP per capita</td>
<td>$13,840</td>
<td>$4,669</td>
<td>$8,959</td>
<td>$2,081</td>
</tr>
<tr>
<td>Poverty headcount ratio</td>
<td>56%</td>
<td>6%</td>
<td>15%</td>
<td>24%</td>
</tr>
<tr>
<td>Conflict zone</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
in Algeria did not perform particularly well, it was still able to capture two closely spaced peaks of influenza activity in the 2016-2017 influenza season in Ghana which suggests that this approach could work in countries with less defined seasonality. In Algeria, the historical influenza trends were unusual with several consecutive recent years without significant influenza activity which could be a symptom of sampling issues. We hypothesize that ARGO performance will improve when national disease surveillance systems adopt accurate and consistent reporting standards. Moreover, the ongoing increase of Internet penetration and literacy rates may also lead to better predictions in the near future.

Future work should explore expanding the Google Search data to include other languages commonly spoken in Africa such as Arabic in Northern Africa or indigenous languages of Sub-Saharan Africa. It would also be interesting to experiment with assigning different $L_i$ penalties to the historical influenza data and the influenza-related Google search data given that the data originate from different sources. Specifically, the performance of the ARGO model may be improved by penalizing the exogenous Google search data while not necessarily penalizing the autoregressive terms that were identified as having a significant effect on the estimation of future influenza activity as done by Yang et al. Alternatively, future work should consider restricting the $L_i$ penalties so that the variables that are consistently identified as having a significant effect on the estimation of influenza activity are constrained to have a fraction of the penalty applied to the rest of the terms as done by Lu et al.

Furthermore, future work should assess the robustness of the ARGO model to variability in the Google search data by implementing the ARGO model using multiple samples of the Google search data. As shown in Yang et al., different samples of the Google search data can be obtained by downloading the Google search data from GHT at different times of the week since the search activity data from GHT are based on a sample of all Google searches. Also, since influenza activity data aggregated at the national level are not necessarily representative of actual influenza activity at the local level, future work could investigate the feasibility of implementing ARGO at smaller geographic scales as shown by Lu et al. for state-level predictions in the USA and the Boston metropolis.

In conclusion, we find that the ARGO algorithm provided modest but significant improvement in monitoring influenza activity in four African countries compared to models that did not include Google search queries. It will be interesting to see how the performances of digital or social media-based surveillance approaches evolve over time as the quality of influenza surveillance data improves and Internet coverage expands. Further research could test whether these algorithms may be useful in monitoring emerging infections that may be perceived as more urgent health threats than influenza in Africa and other low-income settings. We hypothesize that ARGO performance will improve when national disease surveillance systems adopt accurate and consistent reporting standards. Moreover, the ongoing increase of Internet penetration and literacy rates may also lead to better predictions in the near future.

Data availability

Underlying data


This project contains the following underlying data:
- Algeria_data (data for Algeria; please see the notes field of the record for a description of each heading)
- Ghana_data (data for Ghana)
- Morocco_data (data for Morocco)
- South_africa_data (data for South Africa)

Extended data


This project contains the following extended data:
- Algeria_pacf (partial autocorrelogram for Algeria)
- Flunet_heatmap (number of processed specimens reported from January 8, 2012 to October 7, 2018 for each country)
- Ghana_pacf (partial autocorrelogram for Ghana)
- Morocco_pacf (partial autocorrelogram for Morocco)
- Search_terms (Google search terms for Algeria, Ghana, Morocco and South Africa)
- South_africa_pacf (partial autocorrelogram for South Africa)
- Supplementary_Data (containing Supplementary Figures 1 and 2, and Supplementary Tables 1–4)
- Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

Acknowledgements

We are grateful to the Google Health Trends team for granting us access to the API.

Disclaimer

This study does not necessarily represent the views of the NIH or the US government.
References


