# Forecasting Seasonal Influenza Fusing Digital Indicators and a Mechanistic Disease Model

Qian Zhang
Northeastern University, USA
qi.zhang@neu.edu

Nicola Perra
Northeastern University, USA
University of Greenwich, UK
n.perra@gre.ac.uk

Daniela Perrotta
ISI Foundation, Turin, Italy
daniela.perrotta@isi.it

Michele Tizzoni
ISI Foundation, Turin, Italy
michele.tizzoni@isi.it

Daniela Paolotti
ISI Foundation, Turin, Italy
daniela.paolotti@isi.it

Alessandro Vespignani
Northeastern University, USA
a.vespignani@neu.edu

## ABSTRACT

The availability of novel digital data streams that can be used as proxy for monitoring infectious disease incidence is ushering in a new era for real-time forecast approaches to disease spreading. Here, we propose the first seasonal influenza forecast framework based on a stochastic, spatially structured mechanistic model (individual level microsimulation) initialized with geo-localized microblogging data. The framework provides for more than 600 census areas in the United States, Italy and Spain, the initial conditions for a stochastic epidemic computational model that generates an ensemble of forecasts for the main indicators of the epidemic season: peak time and intensity. We evaluate the forecasts accuracy and reliability by comparing the results with the data from the official influenza surveillance systems in the US, Italy and Spain in the seasons 2014/15 and 2015/16. In all countries studied, the proposed framework provides reliable results with leads of up to 6 weeks that became more stable and accurate with progression of the season. The results for the United States have been generated in real-time in the context of the Centers for Disease Control and Prevention "Forecasting the Influenza Season Challenge". A characteristic feature of the mechanistic modeling approach is in the explicit estimate of key epidemiological parameters relevant for public health decision-making that cannot be achieved with statistical models that do not consider the disease dynamic. Furthermore, the presented framework allows the fusion of multiple data streams in the initialization stage and can be enriched with census, weather and socioeconomic data.

## Keywords

Computational epidemiology; Influenza modeling; Real-time forecasting; Social media


*WWW 2017,* April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052678

## 1. INTRODUCTION

Seasonal influenza annually results in up to 5 million severe illness, half million deaths, and increases visits to emergency departments [1, 2, 3, 4]. Real-time forecast of major influenza indicators, such as peak time, and peak intensity can provide key information for public health interventions, such as resources allocation for influenza prevention, control and the public communication of health risk. Although substantial work has been carried out in the field of infectious disease analysis and modeling in the past, real-time forecasting is now fueled by the availability of novel digital data streams generated by human activities that can provide real-time surrogates for the clinically-based reporting of influenza-like-ilness (ILI) [5, 6]. Typical examples are provided by Google Flu Trends (GFT) and similar approaches based on Twitter, Wikipedia and other datasets originated by human activities in the digital world. Combined with traditional surveillance these novel data streams are being used to provide local and timely information about disease and health indicators in populations around the world [7, 8]. Although novel digital data streams may suffer from a number of limitations including signal drifts that might affect the reliability of their forecasts [9, 10], they have increasingly large data volumes, are highly contextual, geo-localized, and allows an unprecedented real-time access to information that can improve forecasting methodologies.

Novel data streams have been generally used as input for statistical models that do not take into account a detailed individual level description of the disease dynamics. Here, we introduce a general framework that, fusing digital indicators with a stochastic, spatially structured, individual level, disease dynamic model, produces short and long term predictions of seasonal influenza at different granularities. In particular, we use geo-localized microblogging data from Twitter to quantify relative geographical incidence of influenza in a given country. These estimates serve as inputs for a mechanistic ILI dynamic model: the global epidemic and mobility model (GLEAM) [11, 12, 13]. We explore the disease dynamic by a latin hypercube sampling of the model's parameter and initial conditions space. For each sampled point we generate 500 identically initialized Monte Carlo simulations of the epidemic spreading that can be aggregated at different geographical resolutions. Finally, we perform a likelihood analysis assimilating official ILI surveillance data up to the time of forecast, thus selecting the ensemble of numer-

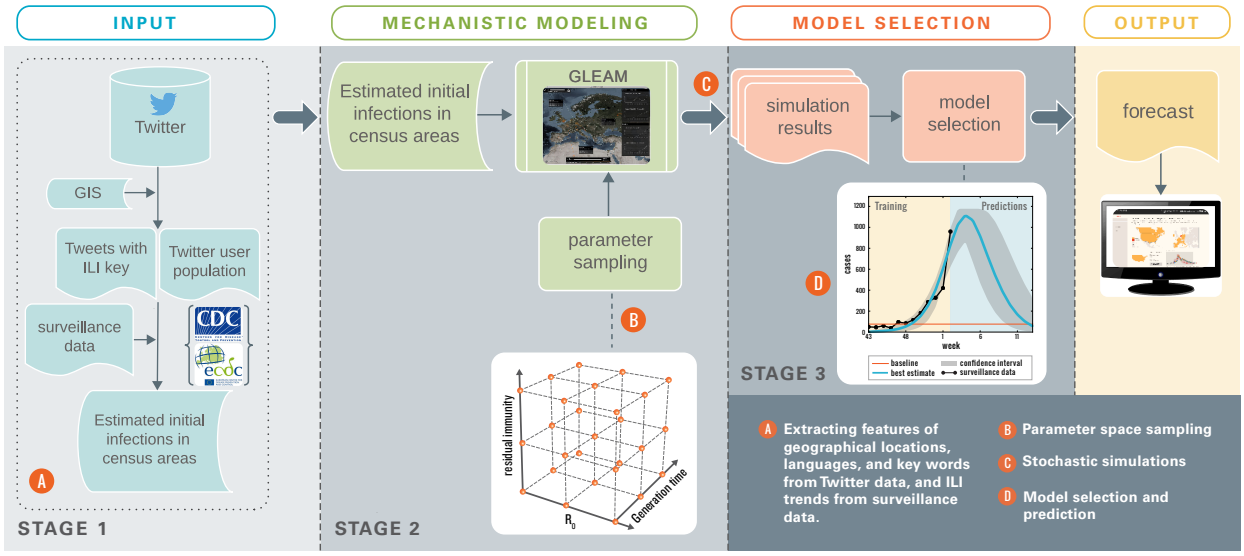


**Figure 1: Illustration of the forecast framework.**

ical simulations that better match the evolving dynamics of the epidemic. Weekly forecasts are then generated by using such ensemble. This approach uses microblogging data only in defining the initial conditions of the disease model thus reducing the dependability of the approach from the data streams, possible biases, spurious correlations and signal drifts due to changing user behavior during anomalous events [9, 10, 14]. Most importantly, the realistic data-driven epidemiological model provides estimates the key epidemiological parameter such as the reproductive number of the disease, serial interval, and residual immunity. These parameters cannot be obtained with statistical models as they do not consider the disease dynamics explicitly.

Here, we present the results obtained with the proposed forecasting framework for the United States, Italy and Spain at national level during the influenza seasons 2014/15 and 2015/16. It is worth remarking that the results concerning the United States have been obtained in real-time in the context of the Centers for Disease Control and Prevention (CDC) "forecasting influenza season challenge" [15]. The presented framework allows the forecast and estimation of a number of indicators of the seasonal epidemic curve. Here, we report the results for the peak timing, and the intensity of the epidemic. Real-time forecasts predict with good accuracy ($\pm$ 1 week) the onset and peak timing up to 6 weeks in advance, and as expected the statistical accuracy of the ensemble forecasts increase as the season progresses. The framework clearly identifies epidemiological differences across seasons and countries, such as different reproductive numbers and peak intensity.

The findings indicate that reliable predictions for the flu season can be generated by using large-scale mechanistic models, thus leaving room to the generalization to other countries, geographical resolutions and epidemic indicators accessible through the dynamic disease model. Finally, we shall stress that as digital indicators are used only in the definition of the initial conditions of the model, the framework can be easily adapted to assimilate other data stream or indicators from digital surveillance systems.

## 2. RELATED WORK

Traditionally, the main approach adopted to model seasonal influenza stems from mathematical descriptions of the disease dynamic often complemented with statistical inference methods [16, 17]. More recently, advances in computing and the access to unprecedented amount of data describing human mobility, census, and interactions allowed the development of realistic data-driven computational models able to capture the unfolding of ILI at different geographical scales [18, 19, 20, 11]. However, seasonal influenza planning and analysis have mostly relied on the situational awareness provided by surveillance data, with mathematical and computational model approaches hindered by the lack of adequate data for real-time forecast. The recent access to novel data streams, such as search queries, microblogging and pages views in Wikipedia is however opening the door to novel approaches to real-time infectious disease forecast. In these approaches traditional and internet-derived data are combined with statistical modeling strategies to provide short and long term forecasts of the evolution of infectious diseases. Several efforts used statistical models based on the digital surrogates of ILI, such as search queries, tweets, or page views in Wikipedia related to the flu [7, 21, 22, 23, 24, 25, 26]. These approaches are generally agnostic to the details of the disease transmission and tailored to provide short terms predictions of the ongoing season. Furthermore, they might be affected by non trivial biases in the data used that could significantly lower their accuracy [9, 10, 27]. In order to limit the effects of such biases, approaches based on the Bayesian fusion of different models have been proposed [28], showing noticeable improvement in the stability and reliability of forecasts. Other approaches fuse the output of regressions tools as GFT with generative epidemiological models [29, 30, 31]. These strategies

improve significantly the accuracy and the horizon of the predictions. Interestingly forecasting approaches based on novel data streams have been readily extended to a range of epidemics and emerging infectious diseases threats [32, 33]. Nevertheless, the landscape of predictive tools is still at an early stage of development. Further work aimed at refining forecasting methodology and identifying the limits and best practices of forecasting approaches in public health is needed. For this reason initiative such as the CDC influenza season challenge [15] are providing nationally coordinated efforts aimed at exploring and assessing forecasting framework and improve their usefulness to public health decision making. The methodology we are presenting is the first to provide seasonal influenza forecasts by fusing a stochastic, spatially structured mechanistic model (individual level microsimulation) with geo-localized microblogging data.

# 3. PRELIMINARIES

## 3.1 Influenza-like Illness datasets

In the United States, CDC measures the intensity of ILI with the percentage of patients visiting for ILI reported by the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). It defines ILI as a symptomatic diagnosis of fever above 37.8 °C with cough or sore throat. Note this definition may include patients of other respiratory diseases who present similar symptoms. CDC usually updates the surveillance data of ILI activity weekly, and the data are free to access via ILINet. The surveillance data we use for Italy and Spain are from European Center for Disease Control and Prevention (ECDC). The ECDC monitors and reports the influenza activity by collecting data from the member States national surveillance systems. These are based on networks of general practitioners reporting the weekly number of patients visited with influenza-like-illness (ILI) or acute respiratory infection (ARI), depending on the country. Here we use the weekly ILI consultation rate per 1,000 and 100,000 individuals for Italy and Spain respectively [34]. The ECDC defines a ILI case as the sudden onset of symptoms with one or more systemic symptoms (fever or feverishness, malaise, headache, myalgia) plus one or more respiratory symptoms (cough, sore throat, shortness of breath) [35].

## 3.2 Microblogging dataset

The dataset of tweets used in this study was extracted from the raw Twitter Gardenhose feed [36]. The Gardenhose is an unbiased sample of about 10% of the entire tweet database, thus providing a statistically significant real-time view of all Twitter account activity [37, 38]. On average 2% of tweets contain GPS information. The accuracy of the modern GPS technology can be just a few meters with 95% confidence [39]. In order to select relevant tweets we consider a list of ILI-related keywords compiled from the literature [40, 41, 42, 43, 44, 45, 46, 47]. The full list of keywords can be accessed in [48]. To better identify the tweets containing ILI keywords, we first detect the language in each tweet and keep only the tweets of the major speaking language in a given country (e.g., English in the United States). The language detection was performed with the same method in [38]. In the United States, for instance, in 2014 there were on average 850,000 tweets in English containing GPS locations and about 330,000 unique Twitter users whose tweets contain GPS information per day. For each tweet with GPS location posted in a time window at the beginning of the flu season (see below for details), we filter it with a list of ILI-related keywords. If a tweet contains any of these keywords, we consider it as a piece of sensory information indicating a potential initial infection in the corresponding area. Note we do not perform further data cleaning to exclude retweets and successive posts from the same users suggested by [24]. In our filtered tweets, the average percentage of retweets over the total tweets with both GPS-location and ILI-related keywords in the beginning of the flu seasons is 1.54% with standard deviation 0.25%, which is much lower than it was reported in [24] (12%). Moreover, the average percent of tweets that were posted by the same user in one week is 5% with standard deviation 1%. We assume such the small fraction of retweets and multiple encounters will not impact our final results.

## 3.3 Mechanistic disease model

The global epidemic and mobility model (GLEAM) is a data-driven spatial, stochastic and individual based epidemic model, in which the world is divided into geographical regions defining a subpopulation network, where connections among subpopulations represent real population traffic flows due to transportation and mobility infrastructures. The model's technical details and the algorithms underpinning the computational implementation are extensively reported in the literature [11, 12, 13, 49]. By using real demographics, the model divides the world population into geographic census areas (basins) that are defined around transportation hubs and connected by mobility fluxes, resulting in an infectious disease metapopulation network model [50, 51, 52]. The model is fully stochastic and from any nominally identical initialization (initial conditions and disease model) generates an ensemble of possible epidemic evolution for epidemic observables, such as newly generated cases.

The disease model within each subpopulation assumes a compartmental classification of the disease under study. The epidemic evolution is modeled using an individual dynamic where transitions are mathematically defined by chain binomial and multinomial processes [53] to preserve the discrete and stochastic nature of the processes. Each subpopulation's disease dynamic is coupled with the other subpopulations through the mechanistically simulated travel and commuting patterns of disease carriers. GLEAM is able to provide high resolution predictions of ILI spreading, which makes possible generating predictions at any region or state level if surveillance data are available. Specifically, GLEAM produces simulation results at the level of subpopulation, that can be aggregated at the different geographical levels.

In the application to the seasonal influenza, the disease dynamics is modeled with a Susceptible-Latent-Infectious-Recovered (SLIR) compartmental scheme, typical of ILI. In each subpopulation, each individual can be in one of these four discrete disease states at each discrete time step [12]. The model assumes homogeneous mixing of individuals inside each subpopulation $j$. The disease transmission rate of symptomatic infected individuals is $\beta$, and for asymptomatic infected individuals, the transmission rate is rescaled with $r_\beta = 0.5$ [54, 55, 12]. We consider that a fraction $r$ of the population is not susceptible to the disease. This is because of residual immunity from previous seasons or vaccination. Given the force of infection $\lambda_j$, each individual in the susceptible compartment ($S_j$) contacts the infection with prob-

ability $\lambda_j \Delta t$ and enters the latent compartment ($L_j$), where $\Delta t = 1$ day is the time interval considered and the index $j$ indicates the specific subpopulation $j$. The latency period is the viral incubation period $1/\varepsilon$ and we consider $\varepsilon = 1.5$ here. Latent individuals exit the compartment with probability $\varepsilon \Delta t$, and transit to asymptomatic infectious compartment ($I_j^a$) with probability $p_a = 1/3$ [56], or with the complementary probability $1 - p_a$, become symptomatic infectious. To reflect the changes of human traveling behavior after the onset of symptoms, infected individuals with symptoms are further divided into two categories: those who can travel ($I_j^t$) with probability $p_t = 0.5$ [54, 12], and those who are travel-restricted ($I_j^{nt}$) with probability $1 - p_t$. All the infected individuals permanently recover after the average infectious period $1/\mu$, entering the recovered compartment ($R_j$). The basic reproduction number $R_0$, defined as the average number of secondary cases generated by a infected individual in a fully susceptible population [16], for seasonal influenza might change from season to season. For example, for the United States it varies from 0.9 to 2.1 [57]. Thus we define the effective reproduction number $R^{\text{eff}} = (1 - r)R_0$, where $R_0 = (r_\beta p_a + (1 - p_a))\beta/\mu$. The number of subpopulations considered in the model is 582, 39 and 38 in the United States, Italy and Spain, respectively.

# 4. METHODS

In order to provide real-time forecasts of seasonal influenza we combine digital indicators, surveillance reports and the mechanistic modeling approach in a three stages framework as presented in Figure 1. In the following we detail the methodology implemented at each stage of the framework.

## 4.1 Stage 1: initialization of the model

Once the initial conditions for the epidemic are known, mechanistic models can numerically generate the epidemic progression by explicitly simulating the transmission dynamics of the disease in the population. In the case of the seasonal flu, the initial conditions are not localized and one needs to map the flu incidence across all the geographical regions included in the model. This is a particularly difficult task as we generally lack surveillance data providing this information at the required spatial granularity. In order to estimate the relative incidence of the flu across regions at any given point in time we use microblogging data from the Twitter platform. For each time window considered for the initialization of the mechanistic model we mine the Twitter dataset (see Section 3), and define $\omega_{l,w}^C$ as the number of GPS localized tweets in the country matching ILI-related keyword $l$, in week $w$ in a given country $C$. By analyzing the time series $\omega_{l,w}^C$ of each keyword with the official surveillance data $d_w$ (for instance in the US, the percentage of patient visits for ILI reported from ILINet to CDC) in week $w$, we can evaluate the coefficient of determination

$$R_l^2 = 1 - \frac{\sum_w (\omega_{l,w}^C - \bar{\omega}_{l,w}^C)^2}{\sum_w (\omega_{l,w}^C - d_w)^2}. \quad (1)$$

This quantity tells us how correlated the two time series are. A keyword with the value closer to 1 indicates the temporal trend of the volume of tweets containing this keyword is more consistent with the temporal trend of the surveillance data, i.e., the keyword is a better proxy for the seasonal flu. From the full dataset, we extract only tweets with GPS location and create for each keyword $l$, week $w$, subpopulation

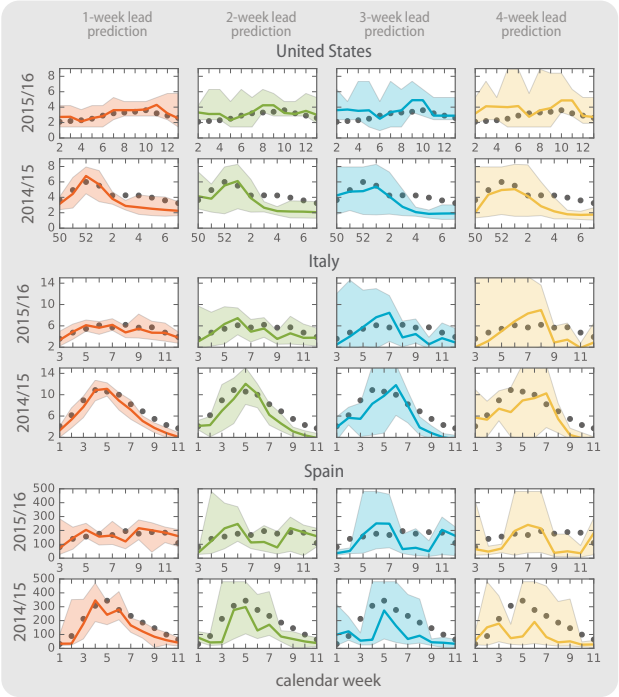


**Figure 2: Forecast results for the entire season The column (I), (II), (III) and (IV) correspond to the epidemic intensity predicted 1, 2, 3, and 4 week ahead respectively. Dots represent observed surveillance data at each week. Dash lines and shadow areas show the average predictions and confidence intervals given by selected ensemble of models respectively.**

$k$ a time series, $\omega_{l,w}^k$, and define the estimator of the number of infected individuals in a given week $w$, basin $k$, as:

$$I_{k,w}^C = \left( \sum_l \omega_{l,w}^C R_l^2 \right) \alpha_k Y. \quad (2)$$

The r.h.s. of the equation considers all the flu keywords, by weighting each one of those according to the corresponding coefficient of determination. The coefficient $\alpha_k$ is the ratio of census population to the total number of Twitter users that are estimated to live in the subpopulation $k$. This rescaling is important to consider the heterogeneous penetration of Twitter[58, 38]. $Y$ is a free parameter necessary to fine tune the correlation between the rescaled number of ILI related tweets and the actual number of infected individuals. The value of $w$ in Eq. 2 describes the starting week in the simulations that we call *seeding week*. In principle its value could be fixed considering the first available data in the surveillance. However, at the early stages of the season reports are quite noisy and result of a balanced mix of different ILI. For this reason, the value of $w$ is shifted forward in time until the onset of the flu season is reached. After that point $w$ is set to be four weeks before the onset week. The onset week of the flu season corresponds to the first week for which the ILI percentage reported by the official surveillance system crosses the seasonal threshold. Each country, depending on

the methodology and data collection of the surveillance system has a specific threshold value.

## 4.2 Stage 2: parameters phase exploration

In the second stage we explore a range mechanistic models by sampling a 4-dimensional phase space defined by the vector $\vec{\theta} = r \times \mu \times R_0 \times Y$ The variable $r$ describes the fraction of the population not susceptible to the disease. This is a combination of the residual immunity and the fraction vaccinated population. The estimates of this last quantity vary from 25% to 45% [59]. We consider $r \in [0.0, 0.45]$. The inverse of the recovery rate $\mu$ defines the infectious period that typically vary from 2 to 5 days. We consider $\mu \in [0.2, 0.5]$. The parameter $\beta$ defines the disease transmission rate, and together with $r$, $\mu$, is determined by the effective reproduction number, $R^{\text{eff}}$. The value of $R^{\text{eff}}$ typically varies from 0.9 to 2.1 [57]; here we consider $R^{\text{eff}} \in [0.8, 3.0]$. The parameter $Y$ is the rescaling factor that provides the number of infected individuals in each basin given the incidence of ILI related tweets. We consider $Y \in [10^{-6}, 10]$. Each parameter is sampled according to different resolution and the total the phase space is formed by $58,000$ sampling points. For each sampled point we generate a statistical ensemble of 500 identically initialized Monte Carlo simulations. Each simulations provides for each subpopulation $i$ the number of new flu cases in time $G_i(t : r, \mu, R_0, Y)$, among other indicators. The signal can be aggregated to provide the epidemic profiles as a function of time at variable geographical resolutions. Here, for sake of simplicity, we will consider only the country level aggregation $G(t) = \sum_i G_i(t)$.

## 4.3 Stage 3: parameters selection and predictions

In the third and final stage we use a statistical inference approach based on the data available to select the ensemble of simulations used to generate predictions. In order to assimilate the surveillance ILI incidence observations into our model, it is necessary to rescale the epidemic profile generated by mechanistic model, expressed as the total number of new flu cases per week, in a profile using the same scale adopted by the surveillance system. Formally, the incidence profile output of the mechanistic model $G(t)$ is rescaled to $G^{'}$: $G^{'} = \mathcal{F}(G)$, where $\mathcal{F}(x) = \alpha x + b$. To evaluate the rescaling factor and the offset term, we consider the average peak intensities and off-season average of the ILI surveillance signal from season 2003/04 to 2012/13, excluding the season 2009/10 because of the H1N1 pandemics. Since the rescaling can fluctuate year by year for reasons inherent to the surveillance systems, we also consider the standard deviation of the peak values in past seasons and consider variations in the rescaling factor of the order of the standard deviation. Given the set of ILI surveillance data available we can evaluate the goodness of fit of the rescaled results by using a multi-model information approach, based on the Akaike information Criterion (AIC)[60]. This approach selects models with the minimum loss of information corresponding to the maximum likelihood with respect to the real data. The available ILI data $x_0, x_1, ..., x_{T-1}$, defines the training window, shown in yellow in Figure 1 (D). For each training window we select models within the 1/33 evidence ratio. This selected ensemble of models is then used to forecast the epidemic in the following weeks $T, T+1, ...,$ , and estimate peak time and peak intensity of the influenza season. As the season progress, the model selection process is repeated at each week as new data from the ILI surveillance system is available. The model selection process is therefore more and more stringent as more data are available, thus reducing uncertainties and stabilizing the forecast.

# 5. RESULTS

Here, we report the forecasts for the 2014/15 and 2015/16 influenza seasons in the context of the CDC's "forecasting the influenza season challenge" for the United States and for the 2014/15 and 2015/16 seasons for Spain and Italy. We use the weekly data of the weighted percentage of patients visiting for ILI from ILINet for the United States, and the the weekly ILI consultation rate per 1,000 and 100,000 individuals for Italy and Spain, respectively. In 2014/15 season, the forecasts started with ILI data of week 45, released on October 31 2014, for the United States, and with data of week 47, released on November 25 2014, for Italy and Spain. In 2015/16 season, the forecasts started with ILI data of week 45, released on October 30 2015, for the United States, and with data of week 52, released on December 30 2015, for Italy and Spain. The proposed forecast framework provides the full epidemic profile of the influenza season and thereby both short and long term forecasts of the epidemic intensity. As shown in Figure 2 the 95% confidence interval inferred from the selected ensemble of models is able to forecast the range of the entire epidemic profile. In the figure we show the values of the forecast along the entire season by considering $x$-week lead predictions ($x - wlp$ for short) $x \in [1, 4]$. For all the countries and seasons considered the empirical observations lay within the confidence intervals for most of the weeks. In order to provide a quantitative comparison between our forecasts and the real epidemic curves we report the values of the Pearson correlation coefficient and the mean absolute percentage error (MAPE) in Table 1. As expected, the correlation generally decreases and the MAPE increases as the epidemic profiles is made by forecasts considering larger lead. Furthermore, a close inspection to the values reveals how the performance of our methodology is tied to the severity of the season. In fact, in case of a severe season, as in 2014-15, our predictions show high values of correlations and small MAPE even in $4 - wlp$. In case of a very mild season instead, as in 2015-16 for Italy and Spain, our method provides more reliable predictions with one or two weeks lead. This is especially so in the case of a mild flu season when the dynamic pattern of the epidemic and the surveillance data signals are affected by large relative fluctuations. While the overall behavior of an ILI season is obviously dominated by the flu season, it is worth remarking that the ILI rate reported by official surveillance systems compounds together with the flu several other pathogens like rhinovirus and respiratory syncytial virus. For this reason a single flu model like the one used here should be generalized to multiple pathogens in order to improve the accuracy also at the very beginning and end of the season.

## 5.1 Predictions of influenza season indicators

In order to deepen the analysis of the accuracy of the model we report weekly forecasts for two main indicators of the season: peak week and peak intensity. In Figure 3 we show the weekly forecasts for the peak week as a function of the lead time. The grey strip in each plot describes the actual empirical values recorded at the end of the season,

**Table 1: Person correlations and mean absolute percentage errors (MAPE) obtained by comparing the forecast results and the official ILI surveillance data along the entire season in each country.**

| | | Pearson correlation | | | | MAPE | | | |
|---|---|---|---|---|---|---|---|---|---|
| country | season | 1-wlp | 2-wlp | 3-wlp | 4-wlp | 1-wlp | 2-wlp | 3-wlp | 4-wlp |
| USA | 2015/16 | 0.83 | 0.49 | 0.34 | 0.25 | 0.12 | 0.21 | 0.30 | 0.37 |
| USA | 2014/15 | 0.95 | 0.81 | 0.75 | 0.90 | 0.22 | 0.27 | 0.27 | 0.32 |
| Italy | 2015/16 | 0.76 | 0.68 | 0.55 | 0.72 | 0.09 | 0.16 | 0.29 | 0.36 |
| Italy | 2014/15 | 0.98 | 0.94 | 0.88 | 0.90 | 0.19 | 0.25 | 0.28 | 0.32 |
| Spain | 2015/16 | 0.58 | 0.46 | 0.23 | 0.11 | 0.17 | 0.33 | 0.47 | 0.50 |
| Spain | 2014/15 | 0.93 | 0.82 | 0.53 | 0.36 | 0.28 | 0.53 | 0.68 | 0.62 |

**Table 2: Accuracy of weekly prediction of peak week for the United States, Italy and Spain at the national level.**

| | | | Percent predictions accurate within 1 week (sorted by weeks prior to observed peak week) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | -6 | -5 | -4 | -3 | -2 | -1 |
| season | country | observed peak week | | | | | | |
| 15/16 | USA | 10 | 57 | 21 | 75 | 97 | 94 | 96 |
| 14/15 | USA | 52 | 50 | 50 | 61 | 86 | 99 | 90 |
| 15/16 | Italy | 8 | 28 | 41 | 63 | 61 | 58 | 47 |
| 14/15 | Italy | 4 | 57 | 21 | 61 | 72 | 69 | 100 |
| 15/16 | Spain | 8 | 8 | 16 | 32 | 48 | 27 | 54 |
| 14/15 | Spain | 5 | 10 | 13 | 90 | 98 | 100 | 100 |

**Table 3: Accuracy of weekly prediction of the peak intensity for the United States, Italy and Spain at the national level.**

| | | | Percent predictions accurate within 20% (sorted by weeks prior to observed peak week) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | -6 | -5 | -4 | -3 | -2 | -1 |
| season | country | observed peak intensity | | | | | | |
| 15/16 | USA | 3.6 | 17 | 25 | 12 | 20 | 88 | 89 |
| 14/15 | USA | 6.0 | 44 | 44 | 33 | 58 | 75 | 84 |
| 15/16 | Italy | 6.2 | 17 | 17 | 21 | 24 | 100 | 100 |
| 14/15 | Italy | 10.8 | 46 | 48 | 43 | 40 | 50 | 66 |
| 15/16 | Spain | 195 | 30 | 17 | 25 | 40 | 100 | 100 |
| 14/15 | Spain | 343 | 10 | 13 | 90 | 98 | 100 | 100 |

and the box and whisker plots describe the distribution of forecasts in the selected ensemble of simulations. With the exception of few points, for all countries and seasons considered, the real value lay within the 95% confidence range of the ensemble forecast. Forecasts generated on week $w$ are based on the surveillance data collected up to that week. Access to more surveillance data it is expected to result in small confidence intervals of predicted outcomes. However, surveillance data for a specific week could be revised by authorities inducing fluctuations in prediction's accuracy respect to the final value of each indicator. Despite these complications, in all the cases considered predictions stabilize towards the correct value as the season progresses.

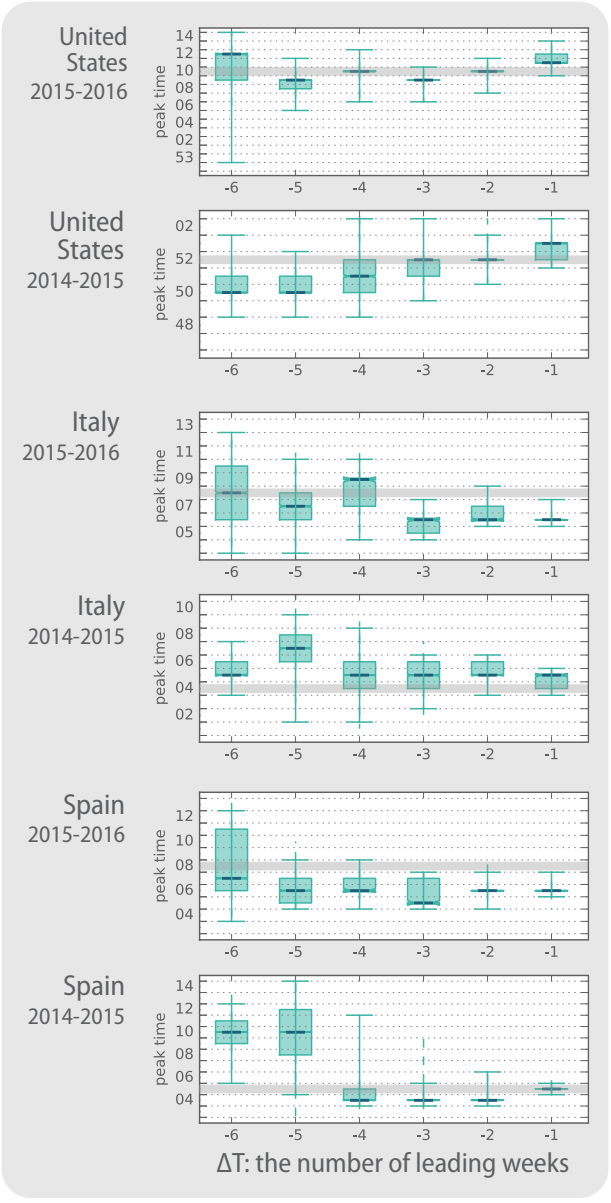


**Figure 3: The boxplots of weekly prediction results for peak week as function of weeks prior to observed peak week. Light grey horizontal bars refer to the final observation for each quantity.**

In Tables 2 and 3, we quantify more rigorously the accuracy of our approach. We define the accuracy of forecasts as the percentage of the selected ensemble of simulations providing predictions within one week for peak time and within 20% error for the peak intensity. As shown in Table 2 the accuracy of our predictions for the peak week is consistently above 60% up to four weeks of lead in the case of USA (both seasons), Italy and Spain in 2014-15. Remarkably, in the case of USA in 2014-15 the accuracy is 50% or higher up to six weeks of lead. In 2015-16 the accuracy is overall lower with the exception of USA.

As shown in Table 3, the accuracy for the peak intensity is on average smaller respect to the other indicators and more dependent on the season and country considered. However, for both seasons and all countries the accuracy of our predictions is at least 50% with two weeks of lead. Furthermore, in the case of Spain during the 2014-15 season the accuracy is 90% or above up to four weeks of lead.

From these analyses we can conclude that our methodology provides reliable predictions often with significant lead times. It is important to stress how the accuracy depends on the severity of the season under consideration. Indeed, the large fluctuations in the surveillance data and the confounding factors introduced by other ILI pathogens that are predominant in mild influenza seasons reverberate in the ensemble selection process inducing unstable predictions.

## 5.2 Estimation of epidemiological parameters

**Table 4: The estimated effective reproductive number $R^{\text{eff}}$, residual immunity and serial interval from the model ensemble at the week of season onset (Median [95% CI])**

| | season | USA | Italy | Spain |
|---|---|---|---|---|
| $R^{\text{eff}}$ | 14/15 | 1.80 [1.50, 2.20] | 1.50 [1.40, 1.50] | 2.00 [1.80, 2.20] |
| | 15/16 | 1.30 [1.20, 1.40] | 1.20 [1.10, 1.30] | 1.30 [1.20, 1.30] |
| residual immunity | 14/15 | 0.15 [0.05, 0.35] | 0.20 [0.05, 0.40] | 0.15 [0.00, 0.30] |
| | 15/16 | 0.30 [0.10, 0.40] | 0.25 [0.00, 0.40] | 0.10 [0.05, 0.35] |
| average infectious time | 14/15 | 4.00 [2.50, 5.00] | 3.60 [2.80, 5.00] | 3.30 [2.50, 4.00] |
| | 15/16 | 5.00 [3.60, 5.00] | 3.30 [2.00, 5.00] | 3.30 [2.50, 4.00] |

Besides the ability of producing reliable predictions with several weeks lead, our approach has another major advantage with respect to statistical models: it allows estimating the key epidemiological parameters of an influenza season reported in Table 4. As expected the non susceptible population tends to be larger for weak flu seasons and varying between 10 to 40% of the total population. These values are compatible with vaccination rate and efficacy and with previous studies [31]. We report also the estimate for the average infectious period ($\mu^{-1}$) that represent the duration of the host infectiousness. Also in this case the results are very close to what is generally reported for influenza in the literature, especially if we consider that the model assumes a 1.5 days of latency, thus yielding a serial interval generally close to 4 or 5 days such as in [31]. A crucial quantity that determines the intensity and severity of the influenza season is the effective reproductive number $R^{\text{eff}}$. This quantity is defined as

$$R^{\text{eff}} = (1 - r)R_0, \tag{3}$$

The estimation of $R^{\text{eff}}$ is particularly interesting to the public health community because it assesses the transmission rate of an infectious disease and the possibility of outbreak. In Table 4, we also report the median and 95% confidence interval for $R^{\text{eff}}$ estimated from the selected ensemble of models at the beginning of the season in each of the considered countries. For the United States at the national level, the $R^{\text{eff}}$ in 2014/15 season is estimated to be considerably larger than in the season 2015/16. Interestingly, this reflects the fact that the 2015/16 season was indeed milder in all the countries considered. In fact, in Italy and Spain the 2015/16 season has been extremely mild with peak activity close to factor two smaller than in the 2014/15 season. Remarkably, such observations are totally consistent with our estimates. Furthermore, the estimated $R^{\text{eff}}$ for the United States are within the range of values reported by retrospective epidemiological studies [57]. The season-length collection of social media data can be helpful in the estimate of epidemiological parameters like serial intervals and attack rate [61]. Our framework combining social media data and mechanistic models shows that it is possible to achieve predictive power for both flu season indicators and key epidemiological parameters.

# 6. DISCUSSION

In this study, we present a framework to forecast the unfolding of seasonal influenza adopting digital ILI surrogates, real surveillance data, and a mechanistic epidemiological model. The predictions focus on weekly season intensity and two indicators: peak week and intensity of the season. We test the framework considering its performance during the 2014/15 and 2015/16 season in the United States, Italy and Spain. Remarkably, our predictions are in good agreement with real observations. In particular, the large majority of real data fall within the 95% confidence range of our forecasts up to four week lead. The framework provides forecasts that stabilizes and are more accurate as the season progresses. The forecasts quality is related to the intensity of the epidemic, and for the 2015/16 season that has been particularly weak we see that the fluctuations in the signal generates a larger uncertainty in the predictions. The importance of accurate forecasts is certainly higher during severe influenza season and it is encouraging to observe that the forecast framework performs better exactly when it is most needed. Our approach presents a number of advances with respect to previous work. We adopt an individual based data driven epidemiological model in which the census areas, metropolitan areas, are coupled by real mobility data. Thus, our forecasts capture the geospatial spreading patterns in each country, and can be applied to any other country where digital surrogates are available [62]. We uses digital surrogates just to initialize the epidemiological model. Thus, predictions are just indirectly influenced by such surrogates, and the dependence from their availability is limited to just few weeks in the early stages of the season. Furthermore, we are not limited to specific source of digital surrogates. For instance, the approach has been already extended to consider data from digital surveillance [62]. Finally the presented framework can be easily refined with multiple data sources, such weather data information, specific contact matrices and school calendars.

## 7. ACKNOWLEDGMENTS

We acknowledge funding from the NIH U54GM111274 and the EU Cimplex Grant agreement n. 641191 under the H2020 Framework program. We are grateful to the past and current members of the Influenza Forecasting Contest Working Group for many stimulating discussions and valuable technical suggestions. We are grateful for N. Samay for her help in visualizations.

## 8. REFERENCES

[1] `http://www.who.int/mediacentre/factsheets/fs211/en/`, March 2014.

[2] C.A. Glaser and et al. Medical care capacity for influenza outbreaks, Los Angeles. *Emerging infectious diseases*, 8(6):569–574, 2002.

[3] M.J. Schull, M.M. Mamdani, and J. Fang. Community influenza outbreaks and emergency department ambulance diversion. *Annals of emergency medicine*, 44(1):61–67, 2004.

[4] W.M. McDonnell, D.S. Nelson, and J.E. Schunk. Should we fear "flu fear" itself? Effects of H1N1 influenza fear on ED use. *The American journal of emergency medicine*, 30(2):275–282, 2012.

[5] J-P. Chretien, D. George, J. Shaman, R.A. Chitale, and F.E. McKenzie. Influenza forecasting in human populations: a scoping review. *PloS ONE*, 9(4):e94130, 2014.

[6] E. Nsoesie, J. Brownstein, N. Ramakrishnan, and M. Marathe. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses*, 8(3):309–316, 2014.

[7] M. Salathe and et al. Digital epidemiology. *PLoS Comput Biol*, 8(7):e1002616, 2012.

[8] B. M. Althouse and et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science*, 4(1):17, 2015.

[9] T. Bodnar and M. Salathé. Validating models for disease detection using twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 699–702. ACM, 2013.

[10] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of Google Flu: traps in big data analysis. *Science*, 343(14 March), 2014.

[11] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J.J. Ramasco, and A. Vespignan. Multiscale mobility networks and the spatial spreading of infectious diseases. *PNAS*, 106(51):21484–21489, 2009.

[12] D. Balcan and et al. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine*, 7(1):45, 2009.

[13] D. Balcan, B. Gonçalves, H. Hu, J.J. Ramasco, V. Colizza, and A. Vespignani. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of Computational Science*, 1(3):132–145, aug 2010.

[14] M. Santillana, D.W. Zhang, B.M. Althouse, and J.W. Ayers. What can digital disease detection learn from (an external revision to) Google Flu Trends? *American journal of preventive medicine*, 47(3):341–347, 2014.

[15] M. Biggerstaff and et al. Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge. *BMC Infectious Diseases*, 16(1):357, 2016.

[16] R.M. Anderson and R.M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[17] E.L. Ionides, C. Bretó, and A.A. King. Inference for nonlinear dynamical systems. *PNAS*, 103(49):18438–18443, 2006.

[18] C.L. Barrett and et al. Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, page 37. IEEE Press, 2008.

[19] D.L. Chao, M.E. Halloran, V.J Obenchain, and I.M. Longini. Flute, a publicly available stochastic influenza epidemic simulation model. *PLoS Comput Biol*, 6(1):e1000656, 2010.

[20] S. Merler and M. Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proc. R. Soc. B*, 277(1681):557–565, 2010.

[21] J. Ginsberg and et al. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009.

[22] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of 1st Workshop on Social Media Analytics (SOMA '10)*, 2010.

[23] D.A. Broniatowski, M.J. Paul, and M. Dredze. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PLOS ONE*, 12(8):e83672, 2013.

[24] H. Achrekar, A. Gandhe, R. Lazarus, Y. Ssu-Hsin, and L. Benyuan. Predicting Flu Trends using Twitter data. In *2011 IEEE Conference on Computer Communications Workshops*, pages 702–707, 2011.

[25] K. Hickmann and et al. Forecasting the 2013-2014 influenza season using wikipedia. *PLoS Comput Biol*, 11(5):e1004239, 05 2015.

[26] D.J. McIver and J.S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS Comput Biol*, 10(4):e1003581, 2014.

[27] T. Bodnar and et al. On the ground validation of online diagnosis with twitter and medical records. In *WWW2014 Companion*, pages 651–656, 2014.

[28] P. Chakraborty and et al. Forecasting a moving target: Ensemble models for ili case count predictions. *Proceedings of the 2014 SIAM International Conference on Data Mining.*, pages 262–270, 2014.

[29] J. Shaman and A. Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, Nov 2012.

[30] J. Shaman and et al. Real-time influenza forecasts during the 2012−2013 season. *Nat. Comms*, 4, Dec 2013.

[31] W. Yang, M. Lipsitch, and J. Shaman. Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*, 112(9):2723–2728, 2015.

[32] R. Gluskin and et al. Evaluation of Internet-based dengue query data: Google Dengue Trends. *PLoS Negl Trop Dis*, 8(2):e2713, 2014.

[33] M.S. Majumder and et al. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak. *JMIR public health and surveillance*, 2(1):e30, 2016.

[34] `http://ecdc.europa.eu/en/healthtopics/influenza/EISN/Pages/index.aspx`, Oct. 2016.

[35] `http://ecdc.europa.eu/en/healthtopics/influenza/surveillance/Pages/influenza_case_definitions.aspx`, Oct. 2016.

[36] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonąlves, S. Patil, A. Flammini, and F. Menczer. Truthy: Mapping the spread of astroturf in microblog streams. In *WWW2011*, 2011.

[37] Q. Samantha. Guide to the Twitter API Part 3 of 3: An Overview of Twitters Streaming API. http://blog.gnip.com/tag/gardenhose/, January 2014.

[38] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE*, 8(4):e61981, 04 2013.

[39] GPS Accuracy. `http://www.gps.gov/systems/gps/performance/accuracy/`, January 2014.

[40] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. In *2010 2nd International Workshop on Cognitive Information Processing*, pages 411–416. IEEE, 2010.

[41] M.J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272, 2011.

[42] E. de Quincey and P. Kostkova. Early warning and outbreak detection using social networking websites: The potential of twitter. In *International Conference on Electronic Healthcare*, pages 21–24. Springer, 2009.

[43] C. Corley, A.R. Mikler, K.P. Singh, and D.J. Cook. Monitoring influenza trends through mining social media. In *BIOCOMP*, pages 340–346, 2009.

[44] D. Ediger and et al. Massive social network analysis: Mining twitter for social good. In *2010 39th International Conference on Parallel Processing*, pages 583–593. IEEE, 2010.

[45] C.D. Corley, D.J. Cook, A.R. Mikler, and K.P. Singh. Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*, 7(2):596–615, 2010.

[46] A. Hulth, G. Rydevik, and A. Linde. Web queries as a source for syndromic surveillance. *PloS ONE*, 4(2):e4378, 2009.

[47] A. Signorini, A.M. Segre, and P.M. Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS ONE*, 6(5):e19467, 2011.

[48] Q. Zhang. *Contagion and ranking processes in complex networks: the role of geography and interaction strength*. PhD thesis, Northeastern University, 2014.

[49] M. Tizzoni, P. Bajardi, C. Poletto, J. J. Ramasco, D. Balcan, B. Gonçalves, N. Perra, V. Colizza, and A. Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC Medicine*, 10(1):165, 2012.

[50] L.A. Rvachev and I.M. Longini. A mathematical model for the global spread of influenza. *Mathematical Biosciences*, 75(1):3–22, jul 1985.

[51] A. Flahault and A.J. Valleron. A method for assessing the global spread of HIV-1 infection based on air travel. *Mathematical Population Studies*, 3(3):161–171, feb 1992.

[52] V. Colizza and et al. Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *Plos Med*, 4(1):e13, 2007.

[53] M.E. Halloran, I.M. Longini, and C.J. Struchiner. Binomial and Stochastic Transmission Models. In *Design and Analysis of Vaccine Studies*, pages 63–84. Springer Science Business Media, sep 2009.

[54] I.M. Longini and et al. Containing pandemic influenza at the source. *Science*, 309:1083–1087, 2005.

[55] C. Fraser and et al. Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*, 324(5934):1557–1561, 2009.

[56] C. Fabrice and et al. Time Lines of Infection and Disease in Human Influenza: A Review of Volunteer Challenge Studies. *American Journal of Epidemiology*, 167(7):775–785, 2008.

[57] G. Chowell, M.A. Miller, and C. Viboud. Seasonal influenza in the United States, France, and Australia: transmission and prospects for control. *Epidemiology & Infection*, 136:852–864, 6 2008.

[58] A. Mislove, S. Lehmann, Y.Y. Ahn, J.P. Onnela, and J.N. Rosenquist. Understanding the Demographics of Twitter Users. In *ICWSM'11*, Barcelona, July 2011.

[59] CDC. National Early Season Flu Vaccination Coverage, United States, November 2013.

[60] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[61] E. Yom-Tov, I. Johansson-Cox, V. Lampos, and A. C. Hayward. Estimating the secondary attack rate and serial interval of influenza-like illnesses using social media. *Influenza and other respiratory viruses*, 9(4):191–199, 2015.

[62] Q. Zhang, C. Gioannini, D. Paolotti, N. Perra, D. Perrotta, M. Quaggiotto, M. Tizzoni, and A. Vespignani. Social data mining and seasonal influenza forecasts: the FluOutlook platform. In *ECML-PKDD*, pages 237–240. Springer, 2015.