INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Zimmer C, Leuba SI, Yaesoubi R, Cohen T. 2018 Use of daily Internet search query data improves real-time projections of influenza epidemics. *J. R. Soc. Interface* **15**: 20180220.
http://dx.doi.org/10.1098/rsif.2018.0220

Received: 28 March 2018
Accepted: 11 September 2018

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
computational biology, biomathematics

Keywords:
influenza, transmission dynamics, forecasting, data resolution, Wikipedia

Author for correspondence:
Christoph Zimmer
e-mail: christoph.zimmer@yale.edu; christoph.zimmer@de.bosch.com

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4238648.

THE ROYAL SOCIETY PUBLISHING

Use of daily Internet search query data improves real-time projections of influenza epidemics

Christoph Zimmer[1,3], Sequoia I. Leuba[1], Reza Yaesoubi[2] and Ted Cohen[1]

[1]Epidemiology of Microbial Diseases, and [2]Health Policy and Management, Yale School of Public Health, New Haven, CT, USA
[3]Bosch Center for Artificial Intelligence, Robert Bosch GmbH, Renningen, Germany

CZ, 0000-0001-5362-6160

Seasonal influenza causes millions of illnesses and tens of thousands of deaths per year in the USA alone. While the morbidity and mortality associated with influenza is substantial each year, the timing and magnitude of epidemics are highly variable which complicates efforts to anticipate demands on the healthcare system. Better methods to forecast influenza activity would help policymakers anticipate such stressors. The US Centers for Disease Control and Prevention (CDC) has recognized the importance of improving influenza forecasting and hosts an annual challenge for predicting influenza-like illness (ILI) activity in the USA. The CDC data serve as the reference for ILI in the USA, but this information is aggregated by epidemiological week and reported after a one-week delay (and may be subject to correction even after this reporting lag). Therefore, there has been substantial interest in whether real-time Internet search data, such as Google, Twitter or Wikipedia could be used to improve influenza forecasting. In this study, we combine a previously developed calibration and prediction framework with an established humidity-based transmission dynamic model to forecast influenza. We then compare predictions based on only CDC ILI data with predictions that leverage the earlier availability and finer temporal resolution of Wikipedia search data. We find that both the earlier availability and the finer temporal resolution are important for increasing forecasting performance. Using daily Wikipedia search data leads to a marked improvement in prediction performance compared to weekly data especially for a three- to four-week forecasting horizon.

1. Introduction

Seasonal influenza remains an important infectious cause of morbidity and mortality [1,2]. In the USA alone, estimates of annual incidence range from 9.2 million to 35.6 million cases, resulting in 140 000 to 710 000 hospitalizations and 12 000 to 56 000 deaths [3].

Efforts to improve *nowcasts* and short-term predictions of influenza activity are motivated by the need to anticipate intensive care unit crowding and surges in vaccine demand. The US Centers for Disease Control and Prevention (CDC) recognize the importance of improving methods for assessing short-term predictions of influenza activity by hosting an annual influenza forecasting challenge [4]. Several data sources have been used to monitor the influenza activity in the USA and as calibration data for forecasting models. Data include both direct measures of influenza such as the official CDC influenza-like illness (ILI) data [5] and indirect measures such as Internet search queries. The validity of Internet search queries, which have included Google [6–12], Twitter [13,14] and Wikipedia [15,16], as proxy measures of influenza activity have been the subject of some debate [12]. However, the rapid availability, low cost of acquisition and potential to access Internet search query data at relatively high spatial

resolution make these data sources attractive options for many teams participating in the annual CDC's influenza prediction challenge [17].

The approaches utilized for influenza prediction in the CDC challenge vary widely, with methods that range from compartmental dynamic transmission models [18–22] to non-mechanistic approaches [23,24]. While previously described models utilize weekly aggregated data on influenza activity [16,18–20,23,24], we note that Internet search query data are often available at much higher temporal resolution, and we sought to understand whether leveraging daily search data can improve short-term predictions of influenza activity. Here we evaluate the comparative performance of models utilizing CDC ILI data to models which utilize Wikipedia data that despite being a more indirect measure of influenza are available more rapidly than ILI data and at finer temporal resolution.

2. Material and methods

Here we describe the availability of Wikipedia search data and demonstrate its utility as a proxy measure of influenza activity, introduce a mechanistic model of influenza transmission and define our framework for model calibration and prediction.

2.1. Wikipedia data

2.1.1. Weekly aggregated search data

The CDC defines ILI as a 'fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat without a KNOWN cause other than influenza' [25] and per cent ILI is the percentage of the total patient visits related to an ILI.

The correspondence of Wikipedia search data [26] and CDC ILI data has been previously demonstrated by others [15]. In more recent work, Wikipedia data have been used as the basis for influenza forecasting in the USA. Hickmann *et al.* [16] developed a linear regression to map searches of selected Wikipedia articles to CDC ILI data. This model calculates a CDC ILI estimate based on the previous week's CDC ILI, ILI_{-1} and current week's numbers of Wikipedia searches for relevant Wikipedia articles. The most predictive article names included 'Human Flu', 'Influenza', 'Influenza A virus', 'Influenza B virus' and 'Oseltamivir'. Their regression reads as

$$\widehat{ILI}_0 = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6ILI_{-1}, \quad (2.1)$$

where $\widehat{ILI}_0$ is the current ILI estimate based on Wikipedia data, x_i is the ratio of the number of visits of these articles to the total number of visited pages, and the regression coefficients are $b_0 = 0.0063$, $b_1 = 17517.3$, $b_2 = 3206.1$, $b_3 = 41258.9$, $b_4 = -71428.7$, $b_5 = -17410.9$ and $b_6 = 0.955$ [16]. In equation (2.1), the unit of ILI is '% influenza among total visits', and the unit of $x_1, \ldots, x_5$ is '# page searches/# total searches', the unit of $b_1, \ldots, b_5$ consequently is '% influenza among total visits/(# page searches/# total searches)', the unit of b_0 '% influenza among total visits' and b_6 is dimensionless.

Even though this model is purely phenomenological, Hickmann *et al.* show that this model produces a good fit to CDC ILI data [16]. However, Hickmann *et al.* [16] also note that the regression coefficients should not be used to infer an importance of the predictors as page visits and ILI are not on the same scale. In figure 1, we visualize this fit for the 2008/2009 season (other years produce similarly good fits as shown in electronic supplementary material, figure A2).

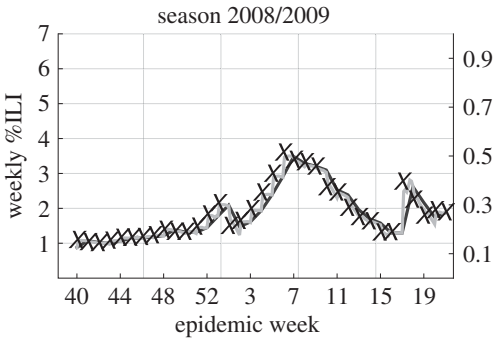

Figure 1. Wikipedia search data provide good fits to CDC ILI data. The CDC ILI data from the season 2008/2009 are in black crosses, estimates obtained from Wikipedia searches with weekly aggregation are in dark grey and estimates obtained from Wikipedia searches in daily resolution are in light grey.

2.1.2. Daily search data

While previous work has used weekly aggregated Wikipedia search data, we note that the data are available at finer temporal resolution. We extend the formula that links weekly Wikipedia data to CDC ILI to create a daily ILI estimate (figure 1) using the coefficients determined in the previous regression:

$$\left.\begin{aligned} \overline{ILI}_0^d &= \frac{b_0}{7} + b_1x_1^d + b_2x_2^d + b_3x_3^d + b_4x_4^d + b_5x_5^d + \frac{b_6}{7}\widetilde{ILI}_- \\ \text{and} \qquad \widehat{ILI}_0^d &= \overline{ILI}_0^d \frac{\widehat{ILI}_0}{\sum_{d\in W} \overline{ILI}_0^d} \end{aligned}\right\} \quad (2.2)$$

where d stands for the day number of the current week, $x_1^d, \ldots, x_5^d$ are the daily page visit ratios and $\overline{ILI}_0^d$ is the unnormalized daily ILI estimate based on Wikipedia, with $\widetilde{ILI}_- = (1-\alpha)ILI_{-2} + \alpha ILI_{-1}$. The set W consists of the days corresponding to the week of data collection. In equation (2.2), the unit of the daily ILI estimate, $\overline{ILI}_d^0$, is '% influenza-related visits among total visits per day', the unit of $\widetilde{ILI}_-$ is as in equation (2.1) '% influenza among total visits (per week)', the unit of $x_1^d, \ldots, x_5^d$ is '# page searches/# total searches (per day)' and consequently the unit of $b_1, \ldots, b_5$ is '% influenza among total visits (per day)/(# page searches/# total searches (per day))', the unit of b_0 is '% influenza-related visits among total visits per day' and b_6 is dimensionless.

The normalization (second line of equation (2.2)) allows the daily Wikipedia data to sum up to the values of the weekly Wikipedia data. ILI_{-1} denotes the previous week's CDC's ILI and ILI_{-2}, the CDC's ILI from two weeks ago. Therefore, $\widetilde{ILI}_-$ is a α weighted combination. The smoothing factor α can be set to

(a) $\alpha = 1$ (no smoothing)
(b) $\alpha = d/7$ (smoothing) or
(c) $\alpha = \begin{cases} 1, & ILI_{-1} \geq ILI_{-2} \\ d/7, & ILI_{-1} < ILI_{-2} \end{cases}$ (partial smoothing)

Partial smoothing is thus equal to no smoothing as long as the data trajectory is rising and equal to smoothing if the data trajectory is falling. The reason to introduce these different kinds of smoothing is the following: the daily estimate of the last day of a week depends only on the CDC's ILI of the beginning of the week, while 1 day later, the daily estimate of a first day of a week depends only on the CDC's ILI estimate of the next week. This sort of discontinuity can lead to less smooth behaviour as depicted in figure 1 and can be addressed by introducing the smoothing factor α. We will use the partial smoothing in the main text as it seems to have the best performance

(electronic supplementary material, figure A6). If a data point, x_i^d, is missing (this happens for 8 days in epidemic weeks 16 and 17 of 2009 and for 4 days in December 2011), the last previous valid data point is used.

2.2. Computational model

A humidity-based susceptible–infected–recovered–susceptible (SIRS) influenza model has been developed by Shaman and colleagues [18,19]. The model includes the following transitions:

$$S \xrightarrow{\beta(t)*S*I/N} I \tag{2.3}$$

$$I \xrightarrow{I/\gamma} R \tag{2.4}$$

and

$$R \xrightarrow{R/\alpha} S \tag{2.5}$$

with an average duration of immunity α, a mean infectious period γ, and a transmission rate $\beta(t)$ which is defined as $\beta(t) = R_0(t)/\gamma$ with an $R_0(t) = \exp(-180q(t) + \log(R_{0\max} - R_{0\min})) + R_{0\min}$ with a maximal and minimal daily reproductive number $R_{0\max}$ and $R_{0\min}$, and a function q describing the absolute humidity (see electronic supplementary material, Humidity data). The effective reproductive number is $R_{\text{eff}}(t) = S(t)/N \cdot R_0(t)$.

Assuming constant population size $N_{\text{pop}} = S + I + R$, it holds that $R = N_{\text{pop}} - S - I$, and we can reduce the model by eliminating the third equation and replacing R by $N_{\text{pop}} - S - I$. We summarize the state of the epidemic consisting of a continuous relaxation of the number of susceptibles, infected and recovered, as $\nu = (\nu^{(S)}, \nu^{(I)}, \nu^{(R)})$. As the initial states are unknown, we treat them as additional model parameters and we summarize the parameters in one parameter vector θ, $\theta = (\alpha, \gamma, R_{0\max}, R_{0\min}, \nu_0^{(S)}, \nu_0^{(I)})$.

The CDC ILI data (and the Wikipedia proxy) reflect the daily or weekly number of incident infections. To apply the MSS method, we need to specify how to map observations y_i, e.g. daily or weekly number of incident infections, to states ν_i (the number of susceptibles $\nu_i^{(S)}$, infected $\nu_i^{(I)}$ and recovered $\nu_i^{(R)}$). As in [27], we describe the new cases y_i in a time interval $[t_{i-1}, t_i]$ as the difference between the number of infected at the end, $\nu_i^{(I)}$, and number of infected at the beginning, $\nu_{i-1}^{(I)}$, plus the number of recoveries, denoted by R_i^*. This number of new recoveries is not simply the difference in numbers of people in the R compartment between the beginning and end of the interval $[t_{i-1}, t_i]$ as people might lose immunity and, hence, move from R to S. Therefore, we track the number of recoveries in each time interval $[t_{i-1}, t_i]$ by introducing an artificial compartment R^* of newly recovered people that is initialized with 0 at the start of each interval.

Our model does not differentiate between different influenza strains or non-influenza causes for ILI. We note that while this is a strong simplifying modelling assumption, it appears to work sufficiently well as demonstrated by [16] and also in our work.

The model is considered as a continuous time stochastic model. Forward simulation will be created by using Gillespie's stochastic simulation algorithm [28]. The calibration method will be introduced in the next section.

2.3. Calibration and prediction

Observations $y_1, \ldots, y_n$ are recorded at time points $t_1, \ldots, t_n$. The observations consist in this study of weekly or daily new ILI cases. We use an iterative procedure to update our knowledge at the time of each new observation. We use a prior distribution $\pi_0(\theta)$ to represent our existing knowledge on the epidemic parameters before the first observation. We use flat uniform priors for the parameters $\theta = (\alpha, \gamma, R_{0\max}, R_{0\min}) \sim U([365, 3650] \times [1.5, 7] \times [1.3, 4] \times [0.8, 1.2])$ and the initial states S_0, $I_0 \sim U([50\,000,\ 70\,000] \times [10, 100])$.

As each new observation y_i accumulates, we update our knowledge on the parameter θ by multiplying our prior with the probability of observing y_i:

$$\left.\begin{aligned} \pi_i(\theta | y_i, y_{i-1}, \ldots, y_1) \ = \\ \pi_{i-1}(\theta | y_{i-1}, \ldots, y_1) \mathcal{P}(y_i | y_1, \ldots, y_{i-1}; \theta) \end{aligned}\right\} \tag{2.6}$$

Note that the posterior at time t_{i-1}, π_{i-1}, also serves as the prior at time t_i. For $i = 1$ we set $\pi_0(\theta \mid y_0) = \pi_0(\theta)$ as we do not have any observations at time t_0. The two following subsections explain how we set up a suitable approximation for $\mathcal{P}$ and how we propagate the distribution π_i through time.

2.3.1. Likelihood approximation

We use the multiple shooting for stochastic systems (MSS) method to approximate $\mathcal{P}$. The MSS method is fast enough to be computationally feasible and accurate enough to allow for reliable calibration and prediction. MSS was initially developed in a systems biology context [29–31] and has been successfully applied to calibration and prediction of epidemics models [27,32]. This subsection will briefly summarize MSS; for full details, we refer the reader to the original publications [28–31].

Given accumulated observations $y_1, \ldots, y_i$, the probability distribution for the epidemic states is called the 'belief state' and we denote it with $\Pi(\cdot | y_1, y_2, \ldots, y_i)$ at time t_i. The belief state Π is a probability distribution assigning each state ν_i its probability conditioned on previous observations, $\Pi(\nu_i \mid y_1, y_2, \ldots, y_i)$. By conditioning on the epidemic state at time t_{i-1}, denoted by ν_{i-1}, and the epidemic state ν_i at time t_i, the probability function $\mathcal{P}(y_i \mid y_1, y_2, \ldots, y_{i-1}; \theta)$ in equation (2.5) can be calculated as:

$$\mathcal{P}(y_i \mid y_1, y_2, \ldots, y_{i-1}; \theta) = \sum_{\nu_i \in \Omega_i} \sum_{\nu_{i-1} \in \Omega_{i-1}} P(y_i \mid \nu_i, \nu_{i-1}; \theta) p(\nu_i \mid \nu_{i-1}; \theta) \Pi(\nu_{i-1} \mid y_1, y_2, \ldots, y_{i-1}; \theta). \tag{2.7}$$

where Ω_i is the support of the belief state at time t_i, and p is the transition probability to move from state ν_{i-1} at time t_{i-1} to state ν_i at time t_i. In case of the above mentioned SIRS model, an epidemic state ν_i is a vector corresponding to the number of people in the compartments $(S(t_i), I(t_i), R(t_i))$. P is the observation probability mapping the state ν_i to the observation y_i incorporating any additional uncertainty in the data collection such as reporting errors. The observation probability P for the new cases y_i is assumed to be normally distributed with a mean $\nu_i^{(I)} - \nu_{i-1}^{(I)} + R_i^*$ and variance 10 (as previously assumed in MSSa version in Zimmer *et al.* [27]).

As in previous work [30–32], we employ a linear noise approximation (LNA) method to approximate the transition probability p (of equation (2.6)). The LNA assumes that the probability distribution of $\nu_i \mid \nu_{i-1}$ can be properly approximated by a normal distribution $\mathcal{N}(x_i, \Sigma_i)$ where x_i is the solution of the ordinary differential equation (ODE) representation of the system on the interval $[t_{i-1}, t_i]$

$$\left.\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} x(t, \nu_{i-1}; \theta) = \Gamma \Lambda(x(t, \nu_{i-1}; \theta), \theta), \\ \text{and} \qquad x(0, \nu_{i-1}; \theta) = \hat{\nu}_{i-1}. \end{aligned}\right\} \tag{2.8}$$

where Γ is a matrix describing the instantaneous change of each transition on each compartment and the vector Λ the rate of the instantaneous change of each transition. The initialization of the equation with $\hat{\nu}_{i-1}$ will be discussed below in equation (2.10).

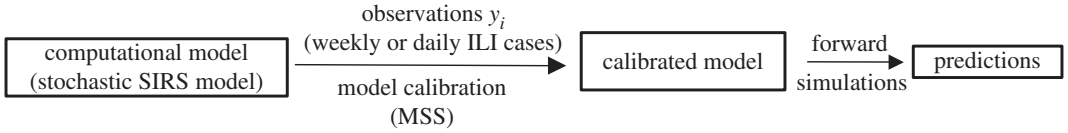

Figure 2. Workflow steps.

The rate vector Λ in equation (2.7) and the instantaneous change matrix Γ are defined as

$$\Lambda(x(t,\nu;\theta),\theta) = \begin{bmatrix} \beta(t)\nu^{(S)}(t)\frac{\nu^{(I)}(t)}{N(t)} \\ \frac{1}{\gamma}\,\nu^{(I)}(t) \\ \frac{1}{\alpha}\,\nu^{(R)}(t) \end{bmatrix} \text{ and } \Gamma = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \quad (2.9)$$

where the columns in Γ correspond to the transitions (namely becoming infected, recovering and loosing immunity) and the rows in Γ correspond to the compartments (namely S, I and R). The state ν_i consists of the number of susceptibles $\nu_i^{(S)}$, infected $\nu_i^{(I)}$ and recovered $\nu_i^{(R)}$. The entry -1 in the first row and first column, hence, describes the effect of the first transition on the first compartment. Here, the first transition means becoming infected and this reduces the number of people in the first compartment by one.

According to [33,34], the covariance matrix Σ can be calculated by solving the following ODE system

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t,\nu_{i-1};\theta) = J(x,\theta)\Sigma(t,\nu_{i-1};\theta) + \Sigma(t,\nu_{i-1};\theta)J(x,\theta)^{\mathrm{T}} + D(x;\theta)$$
$$\Sigma(0,\nu_{i-1};\theta) = 0. \quad (2.10)$$

Here, $J(x,\theta) = \Gamma\,\mathrm{d}/\mathrm{d}x\;\Lambda(x,\theta)$. D is matrix with the (i,j) entry equal to $\sum_{k=1}^{K}\Gamma_{jk}\Gamma_{jk}\Lambda(x,\theta)$.

We calculate the initial values for equation (2.7) recursively. At time t_i, we use the previous state estimate $\hat{\nu}_{i-1}$ to calculate the probability to observe the current observation y_i as $P(y_i \mid \nu_i, \hat{\nu}_{i-1})p(\nu_i \mid \hat{\nu}_{i-1};\theta)$. At time t_i, we then choose the state ν_i as our state estimate $\hat{\nu}_i$ which maximizes this probability, namely

$$\hat{\nu}_i = \underset{\nu_i \in \Omega_i}{\arg\max} P(y_i|\nu_i,\hat{\nu}_{i-1})p(\nu_i|\hat{\nu}_{i-1};\theta). \quad (2.11)$$

We can use this state estimate to strongly reduce the computational complexity of the second summation in equation (2.6) by using a belief state $\Pi(\nu_i \mid y_1,\ldots,y_i)$ that yields 1 at $\hat{\nu}_i$ and 0 elsewhere (point distribution):

$$\Pi(\nu_i \mid y_1, y_2, \ldots, y_i;\theta) = \begin{cases} 1 & \nu_i = \hat{\nu}_i \\ 0 & \text{else.} \end{cases}$$

As we use parameter samples to update the prior distribution recursively (equation (2.5)), using 1000 samples of parameter vectors to forward the recursion (equation (2.5)). We additionally use a mechanism against filter degeneracy (electronic supplementary material, Mechanism against filter degeneracy).

2.3.2. Prediction

The calibration is carried out iteratively according to equation (2.5) as previously described [27]. Next, we make predictions for targets denoted by Z based on the posterior distribution of the parameters π and state estimates $\hat{\nu}_i$ using the simulation model.

$$P(Z \mid y_1, \ldots, y_i) = \int_{\theta\in\Theta}\int_{\nu_i\in\Omega_i} P_{\mathrm{Sim}}(Z \mid \nu_i;\theta)\;\Pi(\nu_i \mid y_1,\ldots,y_i;\theta)\;\pi_i(\theta \mid y_1,\ldots,y_i)\,\mathrm{d}\nu_i\,\mathrm{d}\theta \quad (2.12)$$

where Θ is the parameter vector space and Ω_i the state vector space.

Specifically, we sample $M' = 100$ (as in [27]) parameter vectors $\theta^{(1)},\ldots,\theta^{(M')}$ from the parameter posterior π_i and state vectors $\hat{\nu}_i^{(1)},\ldots,\hat{\nu}_i^{(M')}$ from the belief state. For each of these M' epidemic scenarios, we carry out simulations with a stochastic model P_{Sim} resulting in M' target values $Z^{(1)},\ldots,Z^{(M')}$. These target values are used for the posterior distributions P of our prediction targets Z.

The CDC ILI data are published after a one-week lag [15] while the Wikipedia data are available immediately. Therefore, current Wikipedia data can be used to *nowcast* the current CDC ILI data. *Nowcast* means to estimate a CDC ILI value (based on another data source such as Wikipedia data) before the CDC publishes the official value. We described an approach to estimate CDC ILI based on Wikipedia data in the 'Wikipedia Data' section and use $\widehat{ILI}_0$ or $\widehat{ILI}_0^d$ from equation (2.1) or equation (2.2) as *nowcasts*. The uncertainty around this estimate can be characterized using knowledge of its performance from past seasons (electronic supplementary material, Nowcasting).

Figure 2 summarizes the main steps of our workflow.

3. Results

3.1. Evaluation scheme

We assume that CDC ILI data serves as a gold standard for influenza activity and we evaluate all predictions relative to the CDC ILI data. For all of our comparisons, we assume a one-week lag in publication of CDC ILI data. While this one-week lag in reporting of CDC data accurately reflects the reporting delay, we note that the CDC ILI data are also sometimes revised at a later date [14]. Thus, while the corrected CDC ILI data may be delayed more than a week, we conservatively assume that the final ILI data are available after 7 days.

We perform influenza forecasting retrospectively for the seasons between 2008/2009 and 2015/2016, excluding the pandemic season 2009/2010. We define an influenza season to begin at epidemic week 40 and last for 33 weeks. We assess the performance of the forecasts by calculating a log-score measure [35] which has been used for the judging of the annual CDC Influenza Prediction Challenge [4]. The log-score measure categorizes the per cent ILI in bins of size 0.1 (e.g. [1.0% ILI, 1.1% ILI], [1.1% ILI, 1.2% ILI]). The score is obtained by summing over the forecasting distribution that falls within the bin containing the true value plus the five preceding and five subsequent bins (electronic supplementary materials, Scoring System and figure A1 for details). While the log-score is a valuable instrument for comparing predictions, it is not an intuitive measure, so we also report the reduction in inter-quantile distance which we calculate as the difference between the 95%-quantile and the 5%-quantile of the posterior distribution.

We consider the predictive value of Wikipedia data in three sequential scenarios:

(a) weekly aggregated Wikipedia (ignoring the earlier availability of this data compared with CDC ILI data),
(b) weekly aggregated Wikipedia, now including its immediate availability,

(c) daily Wikipedia data, including its immediate availability.

This approach allows us to determine (a) whether our prediction suffers if we use Wikipedia as a proxy of influenza activity (and fail to take advantage of the earlier availability of Wikipedia data compared with CDC ILI data); (b) how much we gain by leveraging the earlier availability of the Wikipedia data; and (c) how much additional gain we achieve by including finer temporal resolution of Wikipedia data.

Similar to the CDC Influenza Prediction Challenge [4], we use one- to four-week predictions as our targets for our comparison. As we only use data available at the time of forecast, the forecasts early in the season are based on very few data points and the forecasts later in the season are based on the data of up to one influenza season.

Electronic supplementary material, figure A3 allows us to visualize two-week predictions of ILI and associated prediction intervals when using daily Wikipedia data.

3.2. Quantifying improvements in forecasting performance

First, we examine our predictions which use weekly aggregated Wikipedia data without its *nowcasting* feature (scenario a); this is shown in green in figure 3 and electronic supplementary material, figure A4. The quality of these predictions is similar to the CDC, the ILI baseline, which demonstrates that the weekly aggregated Wikipedia data appears sufficient to replicate predictions based on CDC ILI data, but failing to leverage its early availability also eliminates any advantage of this data source.

Second, there is a large improvement when we take advantage of the more rapid availability of weekly Wikipedia data compared to the CDC ILI data (scenario b); this is shown in blue in figure 3 and electronic supplementary material, figure A4. The log-score is substantially improved compared to predictions that use only CDC ILI in all seasons.

Finally, we find that leveraging the daily Wikipedia data (scenario c) leads to additional predictive ability in most seasons, with an average gain over weekly aggregated Wikipedia data (scenario b) of 21% across the seasons we modelled; this is shown in red in figure 3 and electronic supplementary material, figure A4.

We also note that the improvement associated with use of either the weekly aggregated or daily versions of Wikipedia data is most apparent for one- and two-week predictions and is more modest for the four-week predictions. The earlier data availability transforms one-week predictions to *nowcasts* and also trims a week off of two-, three- and four-week predictions. The benefit decreases over time because the relative shortening of the prediction period decreases with increasing absolute length. Therefore, the gain associated with the earlier availability of Wikipedia data decreases as well. However, we note that even for the four-week predictions, the gain in performance associated with earlier availability of the Wikipedia data may still be important.

A closer analysis of the incremental benefit of using daily data reveals that the performance of weekly Wikipedia data (scenario b) and daily Wikipedia data (scenario c) is identical for one-week predictions. This occurs because one-week 'predictions' are no longer real predictions due the earlier data availability; these are *nowcasts* and the same *nowcasting* scheme is used for both (scenario b and scenario c).

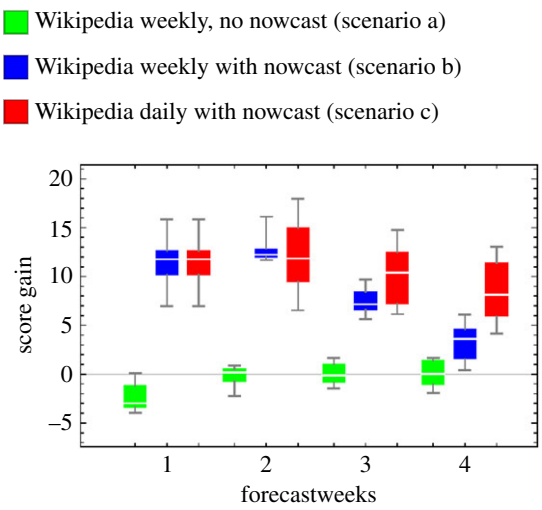

Figure 3. Using Wikipedia data leads to a gain in prediction of influenza. This figure summarizes influenza predictions over different years depicted in electronic supplementary material, figure A4. The following relations are significant (p-value < 0.05) based on the Wilcoxon signed-rank test: Weekly Wikipedia without *nowcasting* (green) is worse than the CDC ILI baseline for one-week forecasts. Weekly (blue) and daily (red) Wikipedia with *nowcasts* is always better than the CDC ILI baseline and the weekly Wikipedia without *nowcasting* (green). Daily Wikipedia (red) is better than weekly Wikipedia (blue) for three- and four-week forecasts. (Online version in colour.)

Comparing using weekly Wikipedia data with using daily Wikipedia data for predictions, we note that the use of daily data is advantageous for all years and the incremental benefit of the daily data is strongest for four-week predictions with an average 409% improvement compared to weekly data. The gain in performance for three-week predictions is on average 38% with daily data performing better than weekly data in five out of seven seasons. For two-week predictions, using daily data was better than using weekly in only three out of seven seasons and performed on average 6% worse.

While the use of the log-score is a well-established approach to compare the performance of prediction systems, it is a non-intuitive measure. Therefore, we also visualize the reduction in the prediction uncertainty when using daily Wikipedia data by reporting the inter-quantile distance (between 5%- and 95%-quantile) of the posterior distribution to provide an illustration of the distribution width (electronic supplementary material, Technical details). By using daily Wikipedia data, we are able to achieve an absolute reduction in inter-quantile distance by an average 0.24% ILI for the two-week predictions, a 0.76% ILI reduction for the three-week predictions and a 1.64% ILI reduction for the four-week predictions (figure 4). This narrower posterior translates to a reduction in prediction uncertainty. We note that this improvement does not lead to more frequent prediction failure (i.e. true value outside of confidence interval) since for both weekly and daily data, the coverage for all targets remains above 90% (electronic supplementary material, table A1).

As noted previously, the use of daily data (compared with weekly aggregated data) produces the greatest gains for the three- and four-prediction horizon (figures 3 and electronic supplementary material, figure A4, the comparison between the red and blue bars). To further investigate the source of this benefit, we analysed the posterior distributions

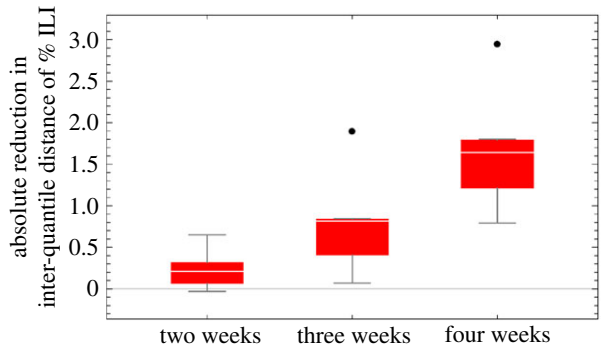

Figure 4. Using daily Wikipedia data versus weekly Wikipedia data reduces the inter-quantile distance of % ILI forecasts. For all seven seasons, using daily versus weekly Wikipedia data reduces the inter-quantile distance for two-, three- and four-week forecasts. The reduction using daily Wikipedia data was significant (p-value <0.05) for two-, three- and four-week forecasting targets based on the Wilcoxon signed-rank test. (Online version in colour.)

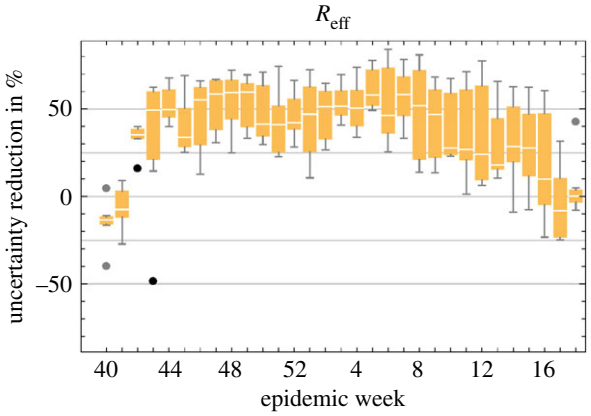

Figure 5. Using daily instead of weekly Wikipedia data reduces parameter uncertainty. Relative reduction in inter-quantile distance (5%- to 95%-quantile) for estimates of the effective reproductive number, R_{eff}, over the seven seasons indicates using daily versus weekly Wikipedia data reduces parameter uncertainty.

of the parameters. Since the true values of parameters are not known, we cannot use a log-score measure, and instead we use inter-quantile distance to evaluate the parameter distributions. There is a substantial reduction in parameter uncertainty for R_{eff} (figure 5) which can be largely attributed to an uncertainty reduction in the time-varying $R_0(t)$ (electronic supplementary material, figure A5). This reduction in parameter uncertainty in R_{eff} likely explains the reduction in prediction uncertainty when using daily Wikipedia data.

All of our model-based predictions can be found in the electronic supplementary material, Prediction results in influenza prediction challenge format; we use the same format as the CDC Influenza Prediction Challenge to facilitate replication and comparison by other modellers [4].

4. Discussion

In this study, we demonstrate and quantify the improvement in model-based influenza prediction that may be achieved by using an immediately available data source rather than the CDC ILI surveillance system which reports weekly aggregated data after a one-week delay (figure 3 and electronic supplementary material, figure A4). Further, we find that the temporal resolution of observations matters for prediction quality, and an additional gain in precision can be achieved by using daily-resolved data (versus weekly aggregated data) (figure 4). As all our forecasts are available in the electronic supplementary material, we anticipate that this work will be compared to future approaches for model-based influenza prediction (electronic supplementary material, Prediction results in influenza prediction challenge format).

Much recent research activity has focused on short- and medium-term predictions of influenza activity [2,36] and the importance of *nowcasting* [6–15,37]. While several previous studies have used similar indirect sources of influenza activity data for forecasting [16,19,24], none of these studies have used daily data or quantified the improvements associated with using these more readily available data sources. Other work [21,22,38,39] has used data with a daily resolution as a basis for forecasting, but these studies have not evaluated the performance over several seasons and do not report gains in prediction performance associated with the higher resolution observation frequency.

Our study clearly shows that influenza prediction can be improved by using a data source that is updated daily and available in near real-time. On average, over the seven seasons we studied, the improvements of daily versus weekly Wikipedia data were up to 409% for four-week predictions (measured in log-score) and 38% for three-week predictions. We note that Wikipedia data are readily available at a daily resolution, so this gain in prediction performance can be achieved without any further cost in data collection.

As our study uses Wikipedia data, it suffers from similar limitations to other modelling studies that use Internet search data to predict influenza activity. For example, it is possible that varying intensity of search activity over a season may compromise the utility of this data source [40]. A Wikipedia-specific limitation is that global article searches for English language articles are aggregated, and searches originating specifically in the USA are not available. This aggregation could erode the value of this data source for predicting specifically US epidemics.

We also note that the improvement associated with using daily-resolved data has been achieved using our calibration and prediction framework as in Zimmer *et al.* [27] and in the previously described methods. We did not investigate whether similar improvements would be seen if other calibration and prediction approaches were used.

We found that the strongest gain in performance associated with use of daily data compared to weekly data is for the three- and four-week forecasting horizons (figures 3 and 4). This suggests that daily-resolved data may help with resource planning given that this horizon seems feasible for public health planning. Whether such improvements in prediction allow for the deployment of more efficient or effective interventions is not directly addressed by our current investigation.

In summary, we find that the use of near real-time daily Internet search data improves the precision of short- and medium-term forecasts of influenza activity. Given the free and ubiquitous nature of this type of information, we expect that future predictions which leverage data at finer temporal resolution and with limited reporting delay will produce epidemic predictions with less uncertainty to better inform reactive public health policy.

Data accessibility. All data is accessible as part of the electronic supplementary material.

Authors' contributions. C.Z., R.Y. and T.C. designed the study. S.I.L. collected the humidity data. C.Z. performed the calculations. C.Z. and

T.C. performed the analysis. All authors wrote and approved the final manuscript.

Competing interests. We declare we have no competing interests.

Funding. Funding was provided by the German research foundation (DFG) for C.Z. and National Institutes of Health (NIH) U54GM088558 from the National Institute of General Medical Sciences (C.Z., R.Y. and T.C.) and 1K01AI119603 from the National Institute of Allergy and Infectious Disease (R.Y.).

Acknowledgements. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

1. Simonsen L, Clarke MJ, Williamson GD, Stroup DF, Arden NH, Schonberger LB. 1997 The impact of influenza epidemics on mortality: introducing a severity index. *Am. J. Public Health* **87**, 1944–1950. (doi:10.2105/ajph.87.12.1944)
2. Chretien JP, George D, Shaman J, Chitale RA, McKenzie FE. 2014 Influenza forecasting in human populations: a scoping review. *PLoS ONE* **9**, e94130. (doi:10.1371/journal.pone.0094130)
3. Centers for Disease Control and Prevention. 2017 *Disease burden of influenza*. See https://www.cdc.gov/flu/about/disease/burden.htm (accessed 30 May 2017).
4. Epidemic Prediction Initiative BETA. 2016 See https://predict.phiresearchlab.org/legacy/flu/evaluation.html (accessed 2 February 2017).
5. Centers for Disease Control and Prevention. 2017 *Flu activity and surveillance*. See https://www.cdc.gov/flu/weekly/fluactivitysurv.htm (accessed 31 May 2017).
6. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009 Detecting influenza epidemics using search engine query data. *Nat. Lett.* **457**, 1012–1014. (doi:10.1038/nature07634)
7. Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH. 2011 Monitoring influenza activity in the united states: a comparison of traditional surveillance systems with Google Flu Trends. *PLoS ONE* **6**, e18687. (doi:10.1371/journal.pone.0018687)
8. Preis T, Moat HS. 2014 Adaptive nowcasting of influenza outbreaks using Google searches. *R. Soc. open sci.* **1**, 140095. (doi:10.1098/rsos.140095)
9. Cho S, Sohn CH, Jo MW, Shin S-Y, Lee JH, Ryoo SM, Kim WY, Seo D-W. 2013 Correlation between national influenza surveillance data and Google Trends in South Korea. *PLoS ONE* **8**, e81422. (doi:10.1371/journal.pone.0081422)
10. Kang M, Zhong H, He J, Rutherford S, Yang F. 2013 Using Google Trends for influenza surveillance in South China. *PLoS ONE* **8**, e55205. (doi:10.1371/journal.pone.0055205)
11. Martin LJ, Lee BE, Yasui Y. 2016 Google Flu Trends in canada: a comparison of digital disease surveillance data with physician consultations and respiratory virus surveillance data, 2010–2014. *Epidemiol. Infect.* **144**, 325–332. (doi:10.1017/S0950268815001478)
12. Lazer D, Kennedy R, King G, Vespignani A. 2014 The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205. (doi:10.1126/science.1248506)
13. Broniatowski DA, Paul MJ, Dredze M. 2013 National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. *PLoS ONE* **8**, e83672. (doi:10.1371/journal.pone.0083672)
14. Paul MJ, Dredze M, Broniatowski D. 2014 Twitter improves influenza forecasting. *PLoS Curr. Outbreaks* **1**. (doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117)
15. McIver DJ, Brownstein JS. 2014 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)
16. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, Del Valle SY. 2015 Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Comput. Biol.* **11**, e1004239. (doi:10.1371/journal.pcbi.1004239)
17. Biggerstaff M *et al.* 2016 Results from the Centers for Disease Control and Prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infect. Dis.* **16**, 357. (doi:10.1186/s12879-016-1669-x)
18. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. 2013 Real-time influenza forecasts during the 2012–2013 season. *Nat. Commun.* **4**, 2837. (doi:10.1038/ncomms3837)
19. Yang W, Karspeck A, Shaman J. 2014 Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput. Biol.* **10**, e1003583. (doi:10.1371/journal.pcbi.1003583)
20. Yang W, Cowling BJ, Lau EH, Shaman J. 2015 Forecasting influenza epidemics in Hong Kong. *PLoS Comput. Biol.* **11**, e1004383. (doi:10.1371/journal.pcbi.1004383)
21. Ong JB, Chen MI, Cook AR, Lee HC, Lee VJ, Lin RT, Tambyah PA, Goh LG. 2010 Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. *PLoS ONE* **5**, e10036. (doi:10.1371/journal.pone.0010036)
22. Dukic V, Lopes HF, Polson NG. 2012 Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *J. Am. Stat. Assoc.* **107**, 1410–1426. (doi:10.1080/01621459.2012.713876)
23. Viboud C, Boëlle P, Carrat F, Valleron A, Flahault A. 2003 Prediction of the spread of influenza epidemics by the method of analogues. *Am. J. Epidemiol.* **158**, 996–1006. (doi:10.1093/aje/kwg239)
24. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. 2015 Flexible modeling of epidemics with an empirical bayes framework. *PLoS Comput. Biol.* **11**, e1004382. (doi:10.1371/journal.pcbi.1004382)
25. Centers for Disease Control and Prevention. Overview of influenza surveillance in the USA. See https://www.cdc.gov/flu/weekly/overview.htm.
26. Page view statistics for wikimedia projects. 2017 See https://dumps.wikimedia.org/other/pagecounts-raw/ (accessed January–March 2017).
27. Zimmer C, Leuba SI, Yaesoubi R, Cohen T. In press. Accurate quantification of uncertainty in epidemic parameter estimates and predictions using stochastic compartmental models. *Stat. Method. Med. Res.*
28. Gillespie DT. 1976 A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434. (doi:10.1016/0021-9991(76)90041-3)
29. Zimmer C, Sahle S. 2012 Parameter estimation for stochastic models of biochemical reactions. *J. Comp. Sci. Syst. Biol.* **6**, 011–021. (doi:10.4172/jcsb.1000095)
30. Zimmer C, Sahle S. 2015 Deterministic inference for stochastic systems using multiple shooting and a linear noise approximation for the transition probabilities. *IET Syst. Biol.* **9**, 181–192. (doi:10.1049/iet-syb.2014.0020)
31. Zimmer C. 2015 Reconstructing the hidden states in time course data of stochastic models. *Math. Biosci.* **269**, 117–129. (doi:10.1016/j.mbs.2015.08.015)
32. Zimmer C, Yaesoubi R, Cohen T. 2017 A likelihood approach for real-time calibration of stochastic compartmental epidemic models. *PLoS Comput. Biol.* **13**, e1005257. (doi:10.1371/journal.pcbi.1005257)
33. Thomas P, Matuschek H, Grima R. 2012 Intrinsic Noise Analyzer: a software package for the exploration of stochastic biochemical kinetics using the system size expansion. *PLoS ONE* **7**, e38518. (doi:10.1371/journal.pone.0038518)
34. Van Kampen NG. 1992 *Stochastic processes in physics and chemistry*, vol. 1. Amsterdam, The Netherlands: Elsevier.
35. Gneiting T, Raftery AE. 2007 Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378. (doi:10.1198/016214506000001437)
36. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. 2013 A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza Other Resp. Viruses* **8**, 309–316. (doi:10.1111/irv.12226)
37. Nunes B, Natàriob I, Carvalho ML. 2013 Nowcasting influenza epidemics using non-homogeneous hidden Markov models. *Stat. Med.* **32**, 2643–2660. (doi:10.1002/sim.5670)
38. Jiang X, Wallstrom G, Cooper GF, Wagner MM. 2009 Bayesian prediction of an epidemic curve. *J. Biomed. Inform.* **42**, 90–99. (doi:10.1016/j.jbi.2008.05.013)
39. Rhodes CJ, Hollingsworth TD. 2009 Variational data assimilation with epidemic models. *J. Theor. Biol.* **258**, 591–602. (doi:10.1016/j.jtbi.2009.02.017)
40. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. 2013 Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput. Biol.* **9**, e1003256. (doi:10.1371/journal.pcbi.1003256)