PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Socioeconomic bias in influenza surveillance

Samuel V. Scarpino[1,2,3,4,5], **James G. Scott**[6], **Rosalind M. Eggo**[7], **Bruce Clements**[8], **Nedialko B. Dimitrov**[9], **Lauren Ancel Meyers**[10,11]*

1 Network Science Institute, Northeastern University, Boston, Massachusetts, United States of America, **2** Marine & Environmental Sciences, Northeastern University, Boston, Massachusetts, United States of America, **3** Physics, Northeastern University, Boston, Massachusetts, United States of America, **4** Health Sciences, Northeastern University, Boston, Massachusetts, United States of America, **5** ISI Foundation, Turin, Italy, **6** Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, Texas, United States of America, **7** Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom, **8** Pediatric Healthcare Connection, Austin, Texas, United States of America, **9** Department of Operations Research, The University of Texas at Austin, Austin, Texas, United States of America, **10** Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, United States of America, **11** Santa Fe Institute, Santa Fe, New Mexico, United States of America

* laurenmeyers@austin.utexas.edu

OPEN ACCESS

Citation: Scarpino SV, Scott JG, Eggo RM, Clements B, Dimitrov NB, Meyers LA (2020) Socioeconomic bias in influenza surveillance. PLoS Comput Biol 16(7): e1007941. https://doi.org/10.1371/journal.pcbi.1007941

Editor: Matthew (Matt) Ferrari, The Pennsylvania State University, UNITED STATES

Received: October 12, 2018

Accepted: May 11, 2020

Published: July 9, 2020

Copyright: © 2020 Scarpino et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: I. Hospitalization data must be requested from the Texas Department of State Health Services Texas Health Care Information Collection (https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm). II. ILINet data must be request from the Texas Department of State Health Services Influenza Surveillance Coordinator: Hailey Rucas at 512-776-6358 (https://www.dshs.texas.gov/idcu/disease/influenza/surveillance/). III. All other data associated with this study are provided in a GitHub repository: https://github.com/Emergent-Epidemics/influenza_surveillance_bias_texas.

Abstract

Individuals in low socioeconomic brackets are considered at-risk for developing influenza-related complications and often exhibit higher than average influenza-related hospitalization rates. This disparity has been attributed to various factors, including restricted access to preventative and therapeutic health care, limited sick leave, and household structure. Adequate influenza surveillance in these at-risk populations is a critical precursor to accurate risk assessments and effective intervention. However, the United States of America's primary national influenza surveillance system (ILINet) monitors outpatient healthcare providers, which may be largely inaccessible to lower socioeconomic populations. Recent initiatives to incorporate Internet-source and hospital electronic medical records data into surveillance systems seek to improve the timeliness, coverage, and accuracy of outbreak detection and situational awareness. Here, we use a flexible statistical framework for integrating multiple surveillance data sources to evaluate the adequacy of traditional (ILINet) and next generation (BioSense 2.0 and Google Flu Trends) data for situational awareness of influenza across poverty levels. We find that ZIP Codes in the highest poverty quartile are a critical vulnerability for ILINet that the integration of next generation data fails to ameliorate.

Author summary

Public health agencies maintain increasingly sophisticated surveillance systems, which integrate diverse data streams within limited budgets. Here we develop a method to design robust and efficient forecasting systems for influenza hospitalizations. With these forecasting models, we find support for a key data gap namely that the USA's public health surveillance data sets are much more representative of higher socioeconomic subpopulations and perform poorly for the most at-risk communities. Thus, our study

Funding: The authors received funding from the National Institutes of Health (MIDAS U01 GM087719 to LAM; https://www.nigms.nih.gov/Research/specificareas/MIDAS/), Postdoctoral Fellowship Omidyar Group (to SVS; https://www.omidyargroup.com/), Postdoctoral Fellowship Santa Fe Institute (https://santafe.edu/), Startup Funds University of Vermont (https://www.uvm.edu/), and Startup Funds Northeastern University (https://www.northeastern.edu/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

highlights another related socioeconomic inequity—a reduced capability to monitor outbreaks in at-risk populations—which impedes effective public health interventions.

Introduction

As part of a broader national security strategy, US President Obama created the first *National Strategy for Biosurveillance*, outlining the nation's key strategic goals in disease surveillance [1]. As a core component of this strategy, President Obama listed taking "full advantage of the advanced technologies. . . that can keep our citizens safe." The surveillance systems outlined by the president are targeted at both recurring diseases, such as influenza, and newly emerging infections. Biosurveillance using advanced technologies may be most important in lower socioeconomic areas, where influenza burden tends to be highest [2–4].

This article assesses the capacity for traditional and novel data sources to provide real-time influenza risk assessments in under-served populations. Using a combination of public health, health care, and Internet-source data available between 2007 and 2012 to make short-term predictions of influenza-related hospitalizations, we compare forecasting accuracy across socioeconomic groups in the Dallas-Fort Worth metro area of Texas, USA. Traditional influenza surveillance is based on primary healthcare provider reports, which may be biased towards serving populations with higher socioeconomic status because of the costs and accessibility of healthcare [5, 6]. Next generation data sources provide promise for improving the timeliness and statistical power of surveillance systems. However, a systematic evaluation of the current surveillance system is needed to evaluate where it falls short, and whether new data can fill gaps.

New technologies have fueled a rapid expansion of data sources that can be acquired quickly and inexpensively for public health surveillance. For example, Google Flu Trends used Internet search queries of influenza-related terms for surveillance [7]. Following the introduction of Google Flu Trends, digital disease surveillance has exploded [8–10] with efforts focused on data from search engines [11, 12], crowd-sourced participatory surveillance (e.g., Flu Near You, InfluenzaNet) [13–15], Twitter (e.g., MappyHealth) [16, 17], Facebook [18, 19], Wikipedia access logs [20, 21], and a variety of other sources (as reviewed in [22, 23]). There is evidence that essentially all of these next-generation surveillance data streams correlate to some degree with epidemiological time-series during typical seasonal outbreaks.

However, there are at least two recent findings worth considering with respect to the these high-tech surveillance systems: 1.) the performance of Google Flu Trends has been unreliable during anomalous influenza outbreaks [24–26] and 2:) it is unclear who is responsible for maintaining these systems [22], especially considering that Google Flu Trends was recently taken offline.

Newly upgraded hospital information systems are another promising source of surveillance data. For example, the United States Centers for Disease Control and Prevention (CDC) launched the BioSense 2.0 program, a set of cooperative agreements between the Department of Veterans Affairs, the Department of Defense, and civilian hospitals from around the country. Through the cooperative agreements, the BioSense 2.0 program creates a "collaborative data exchange system that allows users to track health issues as they evolve" [27]. Whereas BioSense 2.0 provides real-time data on severe cases, the CDC's primary influenza surveillance system, the influenza-like-illness network (ILINet), provides weekly estimates of number of patients presenting with influenza-like-illness symptoms at primary care clinics. Integrating potentially complementary information from new and traditional systems like BioSense 2.0

and ILINet, along with publicly available Internet-source data, like Google Flu Trends, may provide a more timely, comprehensive, and robust picture of disease activity. To this end, the Defense Threat Reduction Agency has begun a national effort to build the Biosurveillance Ecosystem, an integrated disease surveillance system providing access to diverse data sources and powerful analytics [28].

Here, we build and evaluate a multi-source influenza surveillance system that leverages traditional surveillance, electronic health records, and Internet-source data. It is designed to provide short-term forecasts of influenza-related inpatient hospitalizations once an epidemic is underway rather that provide early warning of emerging influenza threats. At the state and multi-county regional levels, these data sources provide effective situational awareness (as compared to early detection of outbreaks). However, we find that they are much more representative of higher socioeconomic sub-populations and perform poorly for the most at-risk communities. Thus, the integration of Internet and electronic medical records data into surveillance systems may improve timeliness and accuracy, but fail to remedy a critical surveillance bias.

Materials and methods

Ethics statement

The Texas Department of State Health Services Institutional Review Board #1 approved this project. The associated reference number is IRB# 12-051. An informed consent waiver was approved by the IRB.

Data sources

We used the following sources, which contained data primarily from Dallas, Tarrant, Denton, Ellis, Johnson, and Parker counties in Texas, between 2007 and 2012:

1. Weekly BioSense 2.0 data were extracted from an online repository [29]. Data are the percent of emergency department (ED) visits for upper respiratory infection, based on classification of free-text chief complaint entries. Although ZIP Code level data are available, we used county-level aggregates in our analysis. Because these data are hosted on a publicly accessible site, we make them available in a CSV file hosted here: https://github.com/Emergent-Epidemics/US_influenza_data_1998_09-2019.

2. ILINet gathers data from thousands of healthcare providers across the USA. Throughout influenza season, participating providers are asked to report weekly the number of cases of influenza-like illness treated and total number of patients seen, by age group. The case definition requires fever in excess of 100˚F with a cough and/or a sore throat without another known cause. The Texas Department of State Health Services (DSHS) provided weekly ILINet records from 2007–2012. In the main text, we use county-level aggregates and provide results with ZIP Code level aggregates in S5 Text.

3. Google Flu Trends (GFT) estimated the number of ILI patients per 100, 000 people based on the daily number of Google search terms associated with signs, symptoms, and treatment for acute respiratory infections. Although GFT is no longer active, past data are available for download from Google.org and have been shown to reliably estimate seasonal influenza activity [7, 30], but be unreliable for the 2009 H1N1 pandemic [31] and during more recent influenza seasons [25]. We considered six different GFT time series, corresponding to the state of Texas and five cities in the Dallas-Fort Worth area: Fort Worth (Tarrant county), Irving (Dallas county), Plano (Collin and Denton counties), Addison

(Dallas county) and Dallas (Dallas county). Google searches are geo-located using the IP address of the device [7]. We used one state-level and six city-level GFT data in all models.

The surveillance models predict hospitalizations that have been aggregated by income quartile. We obtained hospital discharge records from Texas Health Care Information Collection (THCIC), filtered for influenza-related principal diagnostic codes of ICD-9 487.*, which includes 487.0 (with pneumonia), 487.1 (with other respiratory manifestations) and 487.8 (with other manifestations). The data are aggregated into weeks and by patient ZIP Code. Patient ZIP Codes were then combined into income quartiles based on US Census estimates.

Fig 1 presents aggregate counts from the BioSense 2.0, ILINet, GFT, and hospitalization data used in the study for the Dallas-Fort Worth region. We grouped ZIP Codes into quartiles, based on the percentage of the population living in poverty reported in the 2011 American Community Survey [32]. We estimated age distributions within ZIP Codes from the 2011 American Community Survey and the 2010 Census.

Short term predictions

We used generalized additive models to make short-term predictions of influenza-related hospitalizations in the study populations. First, we partitioned ZIP Codes into four poverty quartiles. To predict hospitalizations for group i, we use the Poisson generalized linear model given by

$$y_t^{(i)} \sim \text{Poisson}(\lambda_t^{(i)}) \,, \qquad \log \lambda_t^{(i)} = \alpha^{(i)} + \sum_{k=1}^{D} h_k^{(i)}(x_{k,t}) \,, \tag{1}$$

where $y_t^{(i)}$ is the total number of hospitalizations in group i at time t, $x_{k,t}$ is the kth predictor for hospitalizations at time t, α_i is a background hospitalization rate for group i, and $h_k^{(i)}(\cdot)$ is some potentially nonlinear function (specific to group i) that maps predictors to expected hospitalization counts. Intuitively, the $x_{k,t}$ scalars capture all the information used by the surveillance model to predict hospitalizations. Here t indexes the time of the prediction and k the particular data source—for example, GFT data from two weeks prior. We fit the $h_k^{(i)}(\cdot)$ by expanding each predictor in a third-order B-spline basis with six degrees of freedom. The result of this expansion is that each predictor is now also represented by a number of new predictors, which functionally allow for non-linear associations between the original predictors and influenza hospitalizations. To avoid overfitting, we regularized the spline coefficients using a lasso penalty, with the regularization parameter chosen by cross-validation.

Let $y_i = (y_{i1}, \ldots, y_{iN})^T$ be a vector of counts for income quartile i across all weeks. Let X be an $N \times D$ matrix of surveillance variables used as predictors, where rows are weeks and columns are variables. We considered one-week-ahead forecasts, thus entry t in y_i corresponds to this week's hospitalization count, while row t of the X matrix (used to forecast y_{it}) corresponds to information based on surveillance variables up through week $t - 1$ only. Two-week-ahead forecasts were similar, but with the X matrix containing data only through week $t - 2$.

We considered six different model variations, each using a distinct combination of data from BioSense 2.0, ILINet, and GFT. Importantly, these three data sets included multiple time series. For example, BioSense 2.0 provided hospitalization counts for all six counties in the study area. Additionally, for each time series we added three columns in the X matrix: the level (actual value of the time series in the trailing week), the slope of that variable (first difference over the trailing two weeks at the time of prediction), and the acceleration (second difference over the trailing three weeks at the time of prediction). The columns of X corresponded to the predictors in the model, and we considered six sets of predictors: (i) ILINet alone (15

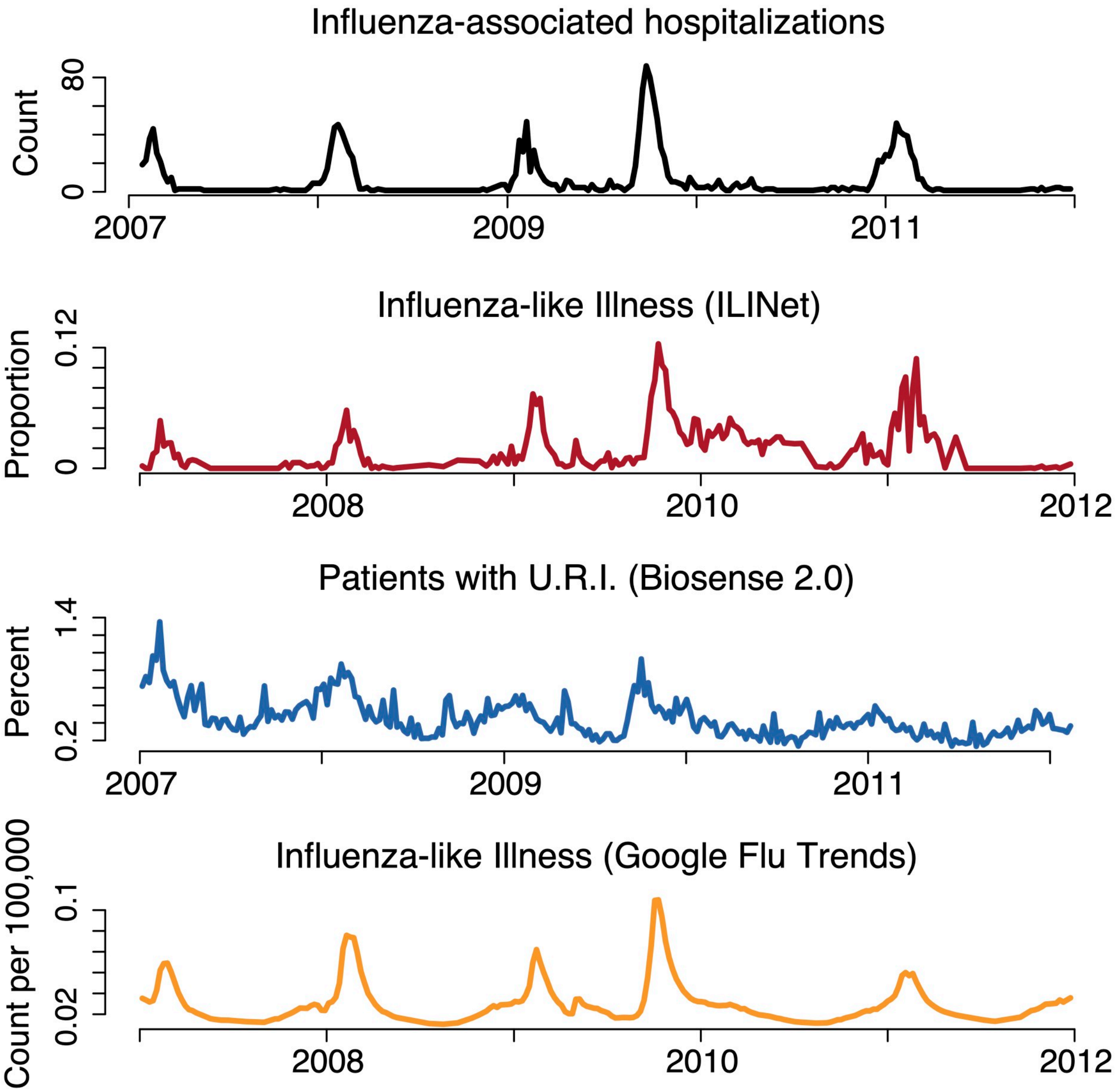

Fig 1. Datasets used in the analysis. The top panel shows influenza-associated inpatient hospitalizations in black, as defined by ICD9 codes 486 and 487, the next panel shows ILINet in red where the proportion of doctor visits are for influenza-like illness, the next panel shows BioSense 2.0 in blue, which is the proportion of ED visits per week that are for an upper respiratory infection. The final panel shows the GFT estimate, in orange, of the number of influenza-like-illness cases per 100,000 people.

https://doi.org/10.1371/journal.pcbi.1007941.g001

predictors), (ii) BioSense 2.0 alone (18 predictors), (iii) GFT alone (18 predictors), (iii) ILINet+ BioSense 2.0 (33 predictors), (iv) ILI + GFT (33 predictors), (v) BioSense + GFT (36 predictors) and (vi) GFT + ILINet+ BioSense 2.0 (51 predictors). In addition, the B-spline expansions provided another 6 variables for each set of predictors, for example, the fully expanded version of model (vi) would have 51 + 306 predictors for a total of 357.

Across the 6 models, we fitted separate models to each group *i*; these group-level models shared the same predictors, but result in different regression coefficients from B-spline expansions of each partial response function. Overall, we fitted 16 models, one for every combination of ZIP Code group (*i*) and candidate predictor set described. Given that we had 188 weeks of data and between 15 and 357 predictors per model, we regularized the coefficient estimates in order to avoid over-fitting. Specifically, we applied a lasso penalty on the coefficient vector β, by minimizing the objective function

$$f(\beta) = l(\beta) + \lambda\ p(\beta)\,,$$

where $l(\beta)$ is the negative log likelihood arising from the Poisson model, $p(\beta)$ is the lasso penalty function, and λ is a scalar that governs the strength of regularization. We select λ for each regression separately using cross validation. See [33] for further details of the model-fitting algorithm. A similar procedure to avoid over-fitting associated with influenza forecasting was utilized by [34].

Predictive performance

To evaluate the predictive performance of the models, we calculated out-of-sample RMSE (ORMSE). Let $\hat{y}_{it}$ be the predicted hospitalization count for group i on week t, generated from fitting the model to every data point except week t. The quantity

$$e_{it} = y_{it} - \hat{y}_{it}$$

is the out-of-sample prediction error. We refitted the model 188 times, one for each week that is removed; this is repeated for every group and every combination of surveillance variables. The ORMSE for a group of ZIP Codes i is given by

$$\mathrm{ORMSE}^{(i)} = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N} e_{it}^2}}{\mathrm{Pop}^{(i)}}\,,$$

where N is the number of weeks, and $\mathrm{Pop}^{(i)}$ is the total population of the group. This can be interpreted as the average predictive error of the model. The units are hospitalization counts per person. Although the groups have approximately the same population size, normalizing by the population of the group is essential. Without normalization, predictions for a large population may appear worse than predictions for a small population, simply because more hospitalizations occur in the larger group. We corroborated our ORMSE results using a log-likelihood analysis (see S2 Text).

To determine whether performance differences between poverty groups were statistically significant, we ran a permutation test with 10,000 repeats, by randomly assigning ZIP Codes into four equally sized groups, and re-fitting the model to each randomized group, following the original procedure, including cross-validation regularization. We then calculated ORMSE for each group, and also the difference between the best ORMSE and the worst ORMSE among the four groups.

For each of the four model variants, we (1) used this procedure to generate null distributions of test statistics for each of our four model variants, (2) calculated the difference between the ORMSE measured for the highest poverty quartile and that measure in the lowest poverty quartiles (according to the original grouping), and (3) determined the proportion of the null distribution less than this difference. This proportion was the P-value used to determine statistical significance.

Results

We evaluated the performance of BioSense 2.0, GFT and ILINet data sources, with respect to short-term predictions of influenza-related hospitalizations in the six-county region surrounding the Dallas-Fort Worth metropolitan area (Fig 2). This region included 305 ZIP Codes and

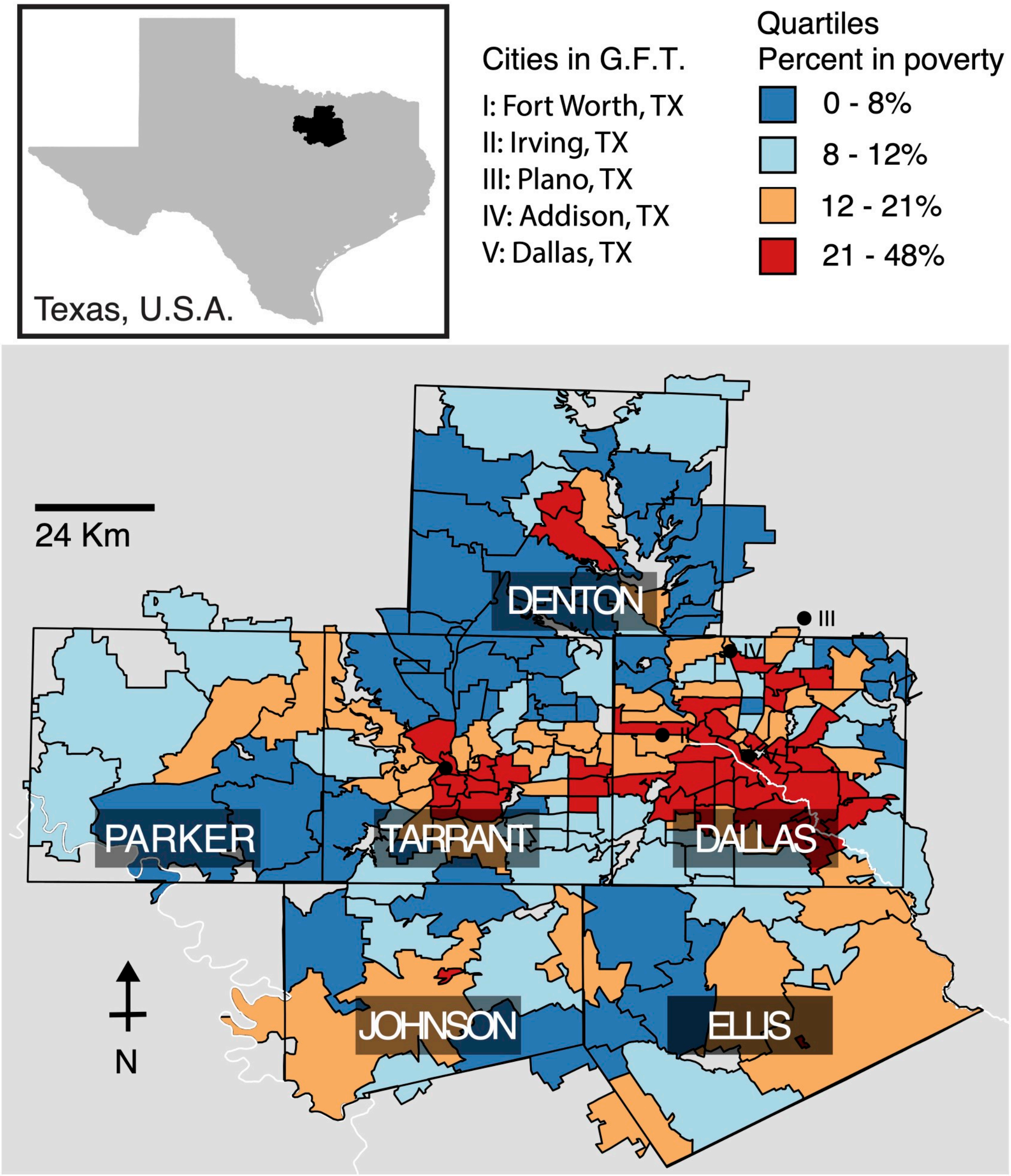

Fig 2. The six counties in northeast Texas included in this study (Dallas, Tarrant, Parker, Denton, Johnson, and Ellis). Zip codes are colored by their poverty quartile, [0-8) (dark blue), [8-12) (light blue), [12-21) (orange), >21 (red) percent of residents below the poverty line. In addition to the state-level Google Flu Trends (GFT) time series, we used the five city-level time series most closely associated with our study area: Fort Worth (I), Irving (II), Plano (III), Addision (IV), and Dallas (V).

https://doi.org/10.1371/journal.pcbi.1007941.g002

all of the emergency departments reporting to the Texas BioSense 2.0 system during the five-year study period (2007-2012).

Influenza burden by poverty level and age

We estimated the influenza hospitalization rate per 1,000 people in each ZIP Code. Throughout the region, we find that influenza hospitalization rates exhibit a significant positive correlation with both poverty level and the proportion of the 2010 census population over age 65

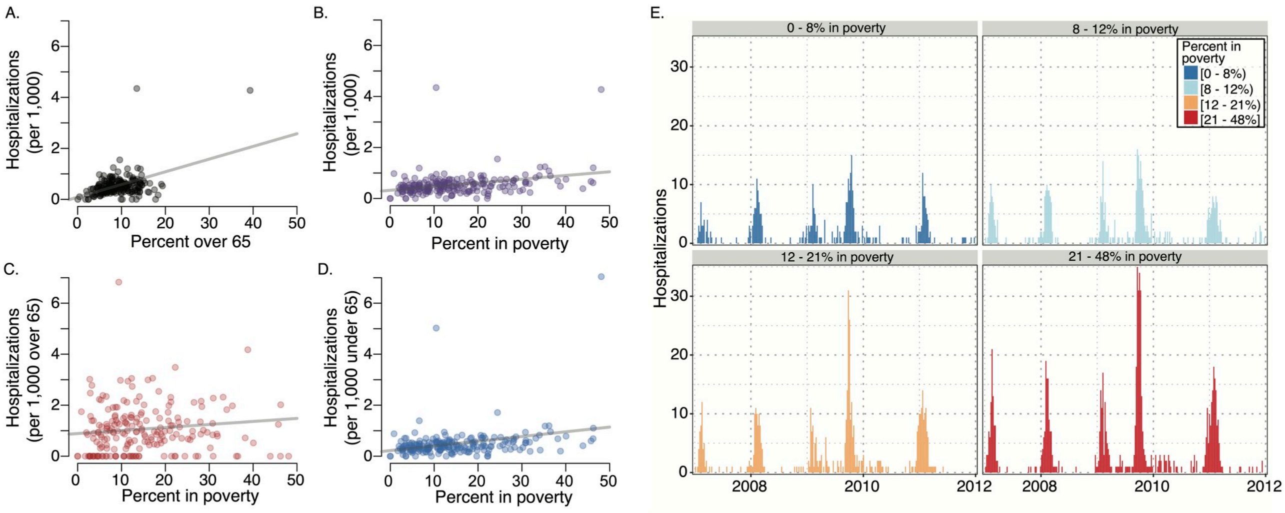

Fig 3. Relationship between age, poverty level, and influenza hospitalizations across 305 ZIP Codes from 2007 to 2012. Demographic data are based on 2010 Census. (A) Influenza hospitalizations increase with the size of the over 65 population (p <.001). (B) Influenza hospitalizations increase with the percent of the population under the federal poverty level (p <.001). (C) Influenza hospitalizations in over 65 year olds does not significantly increase with poverty (p = .11). (D) Influenza hospitalizations in under 65 year olds does significantly increase with poverty (p <.001). (E) The weekly number of hospitalizations across each of the four poverty quartiles. Because the quartiles were selected to include comparable population sizes, we find 2—3 times higher rates of inpatient hospitalizations in the most impoverished quartile (red) as compared to the least (dark blue).

https://doi.org/10.1371/journal.pcbi.1007941.g003

(Fig 3), consistent with recent literature [2, 3, 35]. After controlling for age, we find that poverty and influenza burden are significantly correlated in the under age 65 population but not the over age 65 population (2011 American Community Survey estimates) (p<.001). We established this result with a multivariate regression of hospitalization rate at the zip-code level with the proportion of the ZIP Code living below the poverty line and the proportion of the ZIP Code over 65 (Table 1).

Forecast quality by poverty quartile

We classified ZIP Codes into quartiles based on the proportion of the population living below the federally defined poverty line and fitted separate generalized additive forecasting models to the data in each of the quartiles. In comparisons between model predictions and hospitalization data, we find that the data become less informative as the poverty level increases (Fig 4 and Tables 2 and 3). The models make the best predictions in the most affluent 25% of ZIP Codes—with poverty levels between 0% and 7.5%— and the worst predictions in the most impoverished 25% of ZIP Codes—those with poverty levels between 21.2% and 48.1%, regardless of the data sources included as predictors. Additionally, in an attempt to reduce the forecasting bias, we included the out-of-sample predictions for the three lower poverty quartiles as candidate predictors for highest poverty quartile. This model did not

Table 1. A multivariate regression of hospitalization rate at the zip-code level with the proportion of the ZIP Code living below the poverty line and the proportion of the ZIP Code over 65.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3656	0.0978	3.74	0.0002
Proportion over 65	-0.0001	0.0094	-0.01	0.9911
Proportion in poverty	-0.0096	0.0043	-2.25	0.0258
Interaction	0.0022	0.0003	6.26	0.0000

https://doi.org/10.1371/journal.pcbi.1007941.t001

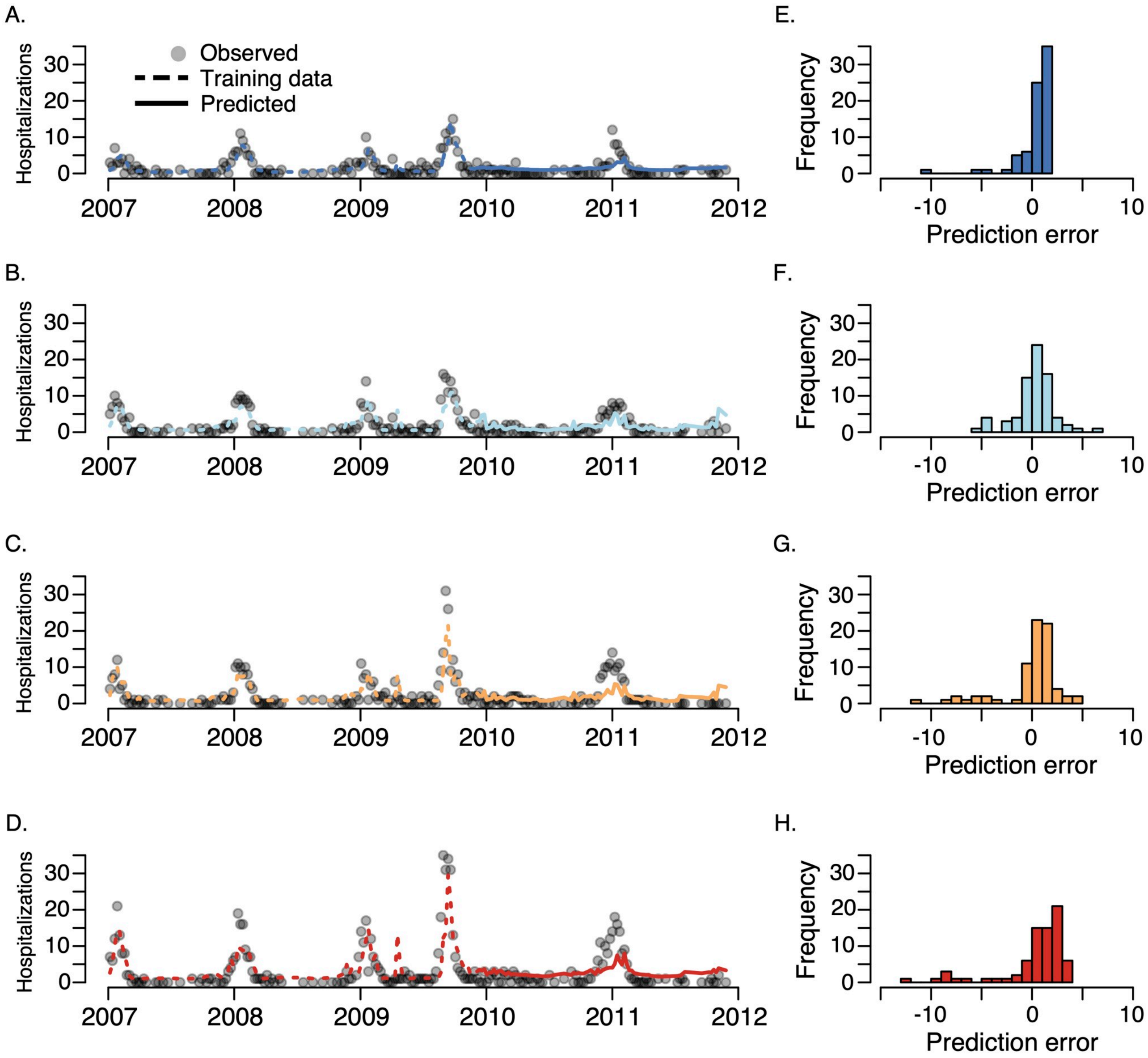

Fig 4. Comparison between one-week ahead model predictions and the total number of weekly observed influenza hospitalizations for each of the four poverty quartiles (A) upper quartile (i.e. least impoverished), (B) upper-middle quartile, (C) lower-middle quartile, (D) lowest quartile (most impoverished) and the distribution of out-of-sample prediction errors (observed—predicted) for the (E) upper quartile, (F) upper-middle quartile, (G) lower-middle quartile, and (H) lowest quartile. The model was trained on the first 60% of the data (dashed lines) and evaluated on the remaining 40% of the data (solid lines). Qualitatively similar results were obtained with n-fold (leave-one-out) cross-validation, see Tables 2 and 3 and S2 Text. Across all four quartiles, the model was unbiased according to a re-sampling test on the residuals, see S3 Text.

https://doi.org/10.1371/journal.pcbi.1007941.g004

improve the forecast accuracy in the highest poverty quartile (see S1 Text). The differences in prediction errors between the upper and lower poverty quartiles are statistically significant ($P < 0.0001$, bootstrap analysis and S1 Fig). This trend is confirmed by generalized linear Poisson and negative binomial models–along with generalized additive Poisson and Gaussian models–fit using one-week-ahead forecasts and evaluated using leave-one-out root-mean-square-error and log-likelihood (all evaluated on out-of-sample data, see S2 Text and S4 Fig). Finally, to address potential biases arising from aggregating of ILINet data from ZIP code level to county level, we re-ran the analyses using ZIP Code level time ILINet time

Table 2. Out-of-sample (60/40 training/testing) root mean-squared error (ORMSE) using a Poisson generalized additive model. Values are normalized by the population size of each ZIP Code quartile and then multiplied by 10^6 to obtain ORMSE per one million residents. The rightmost column gives aggregate ORMSE across all ZIP Codes included in our study area. The quartiles contained: [0-8) (1st quartile), [8-12) (2nd quartile), [12-21) (3rd quartile), and >21 (4th quartile) percent of residents below the poverty line.

Surveillance Data Sources	1st quartile	2nd quartile	3rd quartile	4th quartile	Combined
ILI	1.69	2.41	2.29	5.12	2.22
BioSense	1.55	1.95	2.51	2.60	2.01
GFT	1.38	1.34	2.16	2.68	1.74
ILI + BioSense	1.46	1.68	2.30	3.81	1.94
ILI + GFT	1.42	1.35	2.17	2.75	1.74
BioSense + GFT	1.44	1.58	2.11	2.64	1.79
ILI + BioSense + GFT	1.44	1.53	2.12	2.64	1.72

https://doi.org/10.1371/journal.pcbi.1007941.t002

series. We again found that out-of-sample forecast accuracy was lowest in the most impoverished 25% of ZIP Codes, (see S5 Text).

Synchrony within poverty quartiles

One possible explanation for the observed bias in forecast accuracy by income quartile is that the most impoverished ZIP Codes are either out-of-sync with each other or are more widely distributed across the study area. We tested the hypothesis that the most disadvantaged quartile exhibits greater asynchrony in influenza hospitalization rates and/or are located further away from each other and found the opposite: ZIP Codes in the most impoverished quartile are more synchronous and are located no less closely together as compared to more affluent ZIP Codes.

We define asynchrony as the average pair-wise correlation between ZIP Codes. Based on data visualization and pairwise correlation analyses among ZIP Codes, we failed to find evidence in support of this hypothesis (Fig 5). In fact, influenza hospitalization patterns exhibited significantly more similarity within the lowest poverty quartile than within the less impoverished quartiles. To test for significance, we randomly assigned ZIP Codes to income quartiles 5,000 times and repeated the analysis. The observed mean correlation among the most impoverished quartiles was higher than all of the 5,000 randomizations (i.e., $p < 0.0002$) and the observed median was higher than all but 2 of the simulations (i.e., $p = 0.0004$).

For each poverty quartile, we also performed a principal-component analysis of ZIP Code level hospitalization counts. That is, we calculated the principal components of the matrix $Y^{(i)}$ of hospitalization counts whose rows are weeks, and whose columns are the ZIP Codes within

Table 3. Out-of-sample (leave-one-out) root mean-squared error (ORMSE) for each Poisson generalized additive model. Values are normalized by the population size of each ZIP Code quartile and then multiplied by 10^6 to obtain ORMSE per one million residents. The rightmost column gives aggregate ORMSE across all ZIP Codes included in our study area. The quartiles contained: [0-8) (1st quartile), [8-12) (2nd quartile), [12-21) (3rd quartile), and >21 (4th quartile) percent of residents below the poverty line.

Surveillance Data Sources	1st quartile	2nd quartile	3rd quartile	4th quartile	Combined
ILINet	1.45	1.81	2.63	4.04	2.20
BioSense 2.0	1.69	2.05	2.66	4.23	2.52
GFT	1.33	1.64	2.61	3.63	2.00
ILINet+ BioSense 2.0	1.55	1.81	2.49	3.88	2.18
ILINet+ GFT	1.33	1.59	2.65	3.29	2.00
BioSense 2.0 + GFT	1.39	1.91	2.59	4.03	2.35
ILINet+ BioSense 2.0 + GFT	1.39	1.91	2.52	4.03	2.35

https://doi.org/10.1371/journal.pcbi.1007941.t003

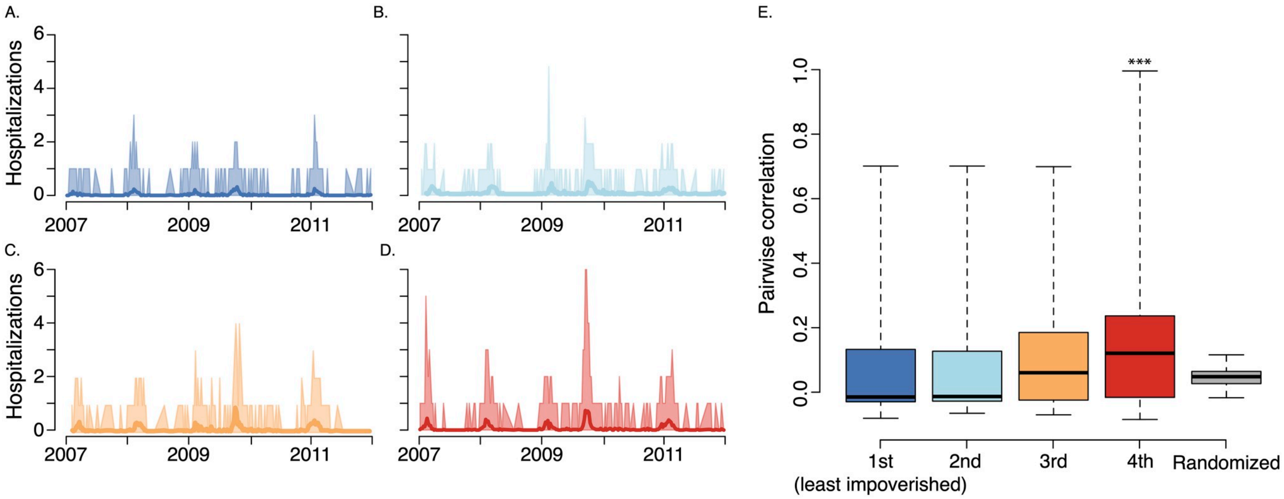

Fig 5. Influenza synchrony among ZIP Codes within each poverty quantile. The quartiles contained: [0-8) (A., 1st quartile), [8-12) (B., 2nd quartile), [12-21) (C., 3rd quartile), and >21 (D., 4th quartile) percent of residents below the poverty line. (A-D) Range of influenza activity (shades) around mean (solid line) at the ZIP Code level, in least impoverished to most impoverished quartiles, respectively. (E) Boxplots of correlation coefficient among pairs of ZIP Codes within each quartile. The most impoverished quartile exhibited the greatest synchrony. To test for significance, we randomly assigned ZIP Codes to income quartiles 5,000 times and repeated the analysis. The observed mean correlation among the most impoverished quartiles was higher than all of the 5,000 randomizations (i.e., $p < 0.0002$) and the observed median was higher than all but 2 of the simulations (i.e., $p = 0.0004$). The median value for each quartile is indicated with a solid, black line, the boxes enclosed the inter-quartile range, and the whiskers cover the entier distribution. We include the distribution of randomized median correlations on the far right (gray box plot). As discussed in the Results, we confirmed these results using a principle component analysis.

https://doi.org/10.1371/journal.pcbi.1007941.g005

poverty quartile *i*. The highest poverty quartile has the highest percent variation explained by these leading components (33.2%), as compared to 13.2% in the least impoverished, 19.3% in the upper-middle, and 26.0% in the lower-middle quartiles, indicating greater synchrony in influenza trends within more impoverished populations. Thus, we do not believe that the reduced performance in lower socioeconomic groups stems from greater variation in temporal flu trends. We utilized the same permutation-based approach described above to test for a significant difference in the principle component analysis results ($p < 0.001$).

Next, we considered the hypothesis that geographic clustering might explain the discrepancies in forecast accuracy. One might intuitively expect nearby populations to exhibit similar influenza patterns; consequently, spatially aggregated populations should be more amenable to forecasting than more dispersed populations. We found that the lowest poverty ZIP Codes have similar patterns of spatial aggregation as the other quartiles (using both Moran's I [36] and inter-centroid distances, see S5 Fig and S4 Text). However, the majority of these ZIP Codes are clustered in Dallas and Tarrant Counties, which is well represented in both our predictor and hospitalization data. To confirm that the uneven distribution of the poverty quartiles across counties (see S5 Fig) did not bias our results, we fit separate prediction models to Dallas and Tarrant counties. In both cases, we confirm our results that forecast accuracy decreases as poverty level increases (see S4 Text).

Finally, to further evaluate alternative explanations for the observed bias, we conducted a simulation experiment to address the role of other factors, such as reduced rates of ILI primary care in lower socioeconomic groups [37, 38], lower correlation between ILI-related Internet searches and actual ILI in lower socioeconomic groups [39, 40], socioeconomic differences in vaccination levels [41, 42], and/or socioeconomic differences in underlying health conditions [43]. The results of this simulation experiment demonstrate that, when all else is equal, a higher hospitalization rate should increase statistical power and provide greater prediction precision (S6 Text, S2 and S3 Figs).

Discussion

Populations with lower socioeconomic status have higher hospitalization rates across a range of diseases [4, 44], caused in part by reduced access to healthcare [37]. Our analysis suggests a similar disparity in the accuracy of public health outbreak surveillance.

Specifically, a combination of clinical symptom reports, Internet searches, and electronic emergency room data can predict week-ahead inpatient influenza hospitalizations more reliably in higher socioeconomic than in lower socioeconomic populations. Given this performance discrepancy, we were surprised to find that high poverty ZIP Codes exhibit much more synchronous influenza hospitalization patterns than low poverty ZIP Codes and are geographically clustered. Thus, the failure likely stems from data bias or under-sampling of at-risk populations. We speculate that GFT (which tallies the number of influenza related Google searches) and ILINet (which collects data from volunteer outpatient clinics) provide low coverage of at-risk populations [5, 6, 45], while BioSense 2.0 may be biased by an excess in non-emergency visits to emergency rooms among uninsured and Medicaid recipients [46].

Over 100 years of epidemiological study demonstrates a consistent, positive association between health and economic prosperity [47, 48]. In many settings, lower socioeconomic status has been linked to both reduced access to healthcare and increased burden of both infectious and chronic diseases [37, 49–51]. For example, the REACH 2010 surveillance program in the U.S.A. found that, “More minorities reported being in fair or poor health, but they did not see a doctor because of the cost.” [49] and a recent study on neonatal intensive care in the US found that, “Black, Hispanic, and Asian infants were segregated across NICUs [neonatal intensive care units], reflecting the racial segregation of minority populations in the United States,” which translated into lower-quality care for infants in the most at-risk populations [52]. In this vein, we found a positive correlation between poverty and influenza hospitalization rates in study populations under age 65, which is consistent with a three-fold excess in pediatric influenza-related hospitalizations estimated for a Connecticut at-risk community [53]. However, it is unknown which of many possible factors—including differences in access to care, vaccine coverage, or prevalence of underlying conditions—are driving this disparity.

Our study identifies another related socioeconomic inequity—a reduced capability to detect and monitor outbreaks in at-risk populations—which impedes effective public health interventions. An analogous surveillance gap has been identified for cancer [54]. Ironically, surveillance systems seem to neglect communities most in need of intervention. New methods for designing and optimizing disease data collection have focused on state-level coverage [55–59] or assumed that risk was evenly spread across well-mixed populations [60], but could be adapted to identify data sources that remedy critical gaps or biases.

We recognize several important limitations of our study. First, our goal was to forecast inpatient hospitalizations for influenza. It is likely that different forms or amounts of bias might manifest themselves had we focused on a different objective. Second, our analysis was restricted to the Dallas-Fort Worth region from which we obtained BioSense 2.0 data, and may not generalize to the rest of the USA nor globally. Third, since we could not access BioSense 2.0 with influenza diagnoses, we used upper respiratory infection data as a proxy. We expect that influenza-specific BioSense 2.0 records would generally improve one-week-ahead predictions, but may or may not close the surveillance poverty gap. Fourth, we did not consider many other data sets, some of which might provide more representative coverage of at-risk populations, including public health laboratory data [61], pharmacy sales [62], school absenteeism records [63, 64], or other Internet-sourced or social media data [22, 65]. Fifth, because we used a lasso penalty to regularize the regression coefficients–implying that the number of degrees of freedom does not necessarily increase with the number of predictors–we could not apply standard model selection methods, such as Akaike Information Criteria, to compare the performance across models (rows of Table 2). Although BioSense 2.0 yields slightly higher performance scores across all poverty quartiles, we leave a definitive comparison among different combinations of surveillance data sources for future study. Sixth, we did not have individual-level patient socioeconomic and/or ZIP Code information from ILINet, BioSense 2.0, and GFT, and thus we were unable to assess directly whether lower socioeconomic groups are underrepresented. However, prior studies suggest that lower socioeconomic groups use the Internet less frequently than higher socioeconomic groups, and that disease-related signals derived from Internet-search data poorly reflect incidence in lower socioeconomic communities [39, 40, 45]. Interestingly, our results suggest that predictions based solely on GFT performed no worse in the lowest income quartile than did other candidate predictors. Researchers with access to individual-level BioSense 2.0 and GFT data–or other systems such as FilmArray Trend [66]–could test our hypothesis, and perhaps develop methods for subsampling the data to improve predictive performance in low income areas. Finally, the Texas inpatient hospitalization data did not indicate whether patients were admitted through an emergency department. Therefore, we were unable to determine whether visitation rate to emergency departments for influenza varied by socioeconomic status. We note that the majority of inpatient hospitalizations in the US are not preceded by an emergency department visit [67].

A growing community of researchers and practitioners across public health, medicine, science, military, and non-governmental organizations are developing and deploying technology-enabled surveillance systems [22] to support adaptive management of infectious diseases [68] and deliver actionable forecasts [69–76]. Many of these efforts focused on improving the timeliness and accuracy of bioevent detection, situational awareness, and forecasting [34, 77]. However, our results suggest a different, and arguably more important priority: improving coverage in at-risk populations. Gaps in both traditional and early next generation surveillance systems compound health disparities in populations with reduced access to healthcare or higher rates of severe disease. Thus, as surveillance systems are upgraded and expanded to incorporate novel data sources, and crowd-sourced/participatory systems are deployed, particular attention should be paid to improving equity, in addition to other performance goals [22, 25]. We further argue that our results highlight the critical need for more research into drivers of disease dynamics and studies to measure the burden of disease–across severity levels–in at-risk communities.

Conclusions

We introduce a robust and flexible method for improving and bench marking situational awareness. Our method offers a general statistical model for short-term prediction, that can systematically integrate diverse data sources, including traditional surveillance data, electronic medical records and Internet-source digital data. We used the method to construct a surveillance system that made one-week-ahead predictions of influenza hospitalizations from real-time BioSense 2.0, Google Flu Trends and ILINet data. While overall performance was reasonable, we discovered a critical data vulnerability in Dallas-Fort Worth's most at-risk populations. This surveillance design framework can be readily applied to evaluate and integrate new data sources that address this challenge.

Supporting information

S1 Text. This document contains supplemental information regarding the use of hospitalization forecasts as predictors.
(PDF)

S2 Text. This document contains supplemental information detailing additional statistical model fits and evaluations.
(PDF)

S3 Text. This document contains supplemental information regarding subsampling residuals to test for model bias.
(PDF)

S4 Text. This document contains supplemental information regarding models fit only to Dallas and Tarrant County hospitalizations and a spatial autocorrelation analysis using Moran's I.
(PDF)

S5 Text. This document contains supplemental information regarding models fit using ZIP Code level ILINet data.
(PDF)

S6 Text. This document contains supplemental information regarding simulations to evaluate the sensitivity of model predictions to hospitalization rate and surveillance detection rate.
(PDF)

S1 Fig. Result of the permutation test for the ILI + BioSense + GFT model across 10,000 Monte Carlo samples. The vertical red line is at 3.3, the observed value based on the poverty grouping. The results indicate that it is unlikely for the observed value to arise by chance. The Monte Carlo p-value is 0.0001, with only of our randomized permutations yielding an ORMSE gap at least as large as 3.3.
(TIF)

S2 Fig. Simulating disparate case hospitalization rates. The curves illustrate a typical simulation. The left-hand panel depicts the Influenza-Like-Illness (ILI) time series (blue) for populations A and B, and surveillance time series for A (red) and B (green) derived by stochastically sampling the ILI time series, assuming that 10% of cases are detected by the system (for example, via internet use or physician visits). The right-hand panel depicts the hospitalization time series and predicted hospitalizations for populations A and B, which had hospitalization rates of 0.1 and 0.9, respectively. The hospitalization curves were generated by stochastically sampling the ILI curve in the left-hand panel and the predictions were created using the same regression model as in the main analysis. The average R^2 over 10,000 simulations for these predictions are 0.9986 and 0.9993, for A and B, respectively.
(TIF)

S3 Fig. As the hospitalization rate or the surveillance detection rate drops, the predictions become less precise. For each combination of surveillance detection rate and hospitalization rate, we run 100 simulations to estimate the expected R^2. These simulations are conducted assuming $\beta = 0.076$ and $\gamma = 0.07$ and the results are qualitatively the same for other values of these parameters.
(TIF)

S4 Fig. Comparison between one-week ahead model predictions and the total number of weekly observed influenza hospitalizations for each of the four poverty quartiles (A) upper quartile, (B) upper-middle quartile, (C) lower-middle quartile, (D) lowest quartile and the distribution of out-of-sample prediction errors (observed—predicted) for the (E) upper quartile, (F) upper-middle quartile, (G) lower-middle quartile, and (H) lowest quartile. Across all four quartiles, the model was unbiased according to a resampling test on the residuals.
(TIF)

S5 Fig. Geographic distribution of ZIP Codes by poverty quartile. A. Boxplots of pairwise distances between ZIP Codes in the four poverty quartiles. ZIP Codes in the highest poverty quartile (red) are significantly closer than ZIP Codes in the other three quartiles (ANOVA and Tukey Honest Test $p < 0.001$). B. Distribution of ZIP Codes in each poverty quartile by county. The most impoverished quartile (21-48%) is over-represented in Dallas County.
(TIF)

Author Contributions

Conceptualization: Samuel V. Scarpino, Rosalind M. Eggo, Bruce Clements, Nedialko B. Dimitrov, Lauren Ancel Meyers.

Formal analysis: Samuel V. Scarpino, James G. Scott, Rosalind M. Eggo.

Funding acquisition: Bruce Clements, Lauren Ancel Meyers.

Investigation: Samuel V. Scarpino, James G. Scott, Rosalind M. Eggo, Nedialko B. Dimitrov.

Methodology: Samuel V. Scarpino, James G. Scott, Rosalind M. Eggo, Nedialko B. Dimitrov.

Supervision: Bruce Clements, Nedialko B. Dimitrov, Lauren Ancel Meyers.

Visualization: Samuel V. Scarpino, Rosalind M. Eggo.

Writing – original draft: Samuel V. Scarpino, James G. Scott, Rosalind M. Eggo, Nedialko B. Dimitrov, Lauren Ancel Meyers.

Writing – review & editing: Samuel V. Scarpino, James G. Scott, Rosalind M. Eggo, Bruce Clements, Nedialko B. Dimitrov, Lauren Ancel Meyers.

References

1. US G. National Strategy for Biosurveillance; 2012. https://obamawhitehouse.archives.gov/sites/default/files/National_Strategy_for_Biosurveillance_July_2012.pdf.
2. Sloan C, Chandrasekhar R, Mitchel E, Schaffner W, Lindegren ML. Socioeconomic disparities and influenza hospitalizations, Tennessee, USA. Emerging infectious diseases. 2015; 21(9):1602. https://doi.org/10.3201/eid2109.141861 PMID: 26292106
3. Tam K, Yousey-Hindes K, Hadler JL. Influenza-related hospitalization of adults associated with low census tract socioeconomic status and female sex in New Haven County, Connecticut, 2007-2011. Influenza and other respiratory viruses. 2014; 8(3):274–281. https://doi.org/10.1111/irv.12231 PMID: 24382111
4. Placzek H, Madoff L. Effect of race/ethnicity and socioeconomic status on pandemic H1N1-related outcomes in Massachusetts. American journal of public health. 2014; 104(1):e31–e38. https://doi.org/10.2105/AJPH.2013.301626 PMID: 24228651
5. Nsoesie EO, Brownstein JS. Computational Approaches to Influenza Surveillance: Beyond Timeliness. Cell host & microbe. 2015; 17(3):275–278. https://doi.org/10.1016/j.chom.2015.02.004
6. Stoto MA. The effectiveness of US public health surveillance systems for situational awareness during the 2009 H1N1 pandemic: a retrospective analysis. PloS one. 2012; 7(8):e40984. https://doi.org/10.1371/journal.pone.0040984 PMID: 22927904

7. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232):1012–1014. https://doi.org/10.1038/nature07634 PMID: 19020500

8. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—harnessing the Web for public health surveillance. New England Journal of Medicine. 2009; 360(21):2153–2157. https://doi.org/10.1056/NEJMp0900702 PMID: 19423867

9. Salathe M, Bengtsson L, Bodnar T, Brewer D, Brownstein J, Buckee C, et al. Digital epidemiology. PLoS computational biology. 2012; 8:e1002616. https://doi.org/10.1371/journal.pcbi.1002616 PMID: 22844241

10. Milinovich GJ, Williams GM, Clements AC, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. The Lancet infectious diseases. 2014; 14(2):160–168. https://doi.org/10.1016/S1473-3099(13)70244-5 PMID: 24290841

11. Polgreen PM, Chen Y, Pennock D, Nelson F, Weinstein R. Using internet searches for influenza surveillance. Clinical infectious diseases. 2008; 47(11):1443–1448. https://doi.org/10.1086/593098 PMID: 18954267

12. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in china with search query from baidu. PloS one. 2013; 8(5):e64323. https://doi.org/10.1371/journal.pone.0064323 PMID: 23750192

13. Chunara R, Aman S, Smolinski M, Brownstein JS. Flu Near You: An Online Self-reported Influenza Surveillance System in the USA. Online Journal of Public Health Informatics. 2012; 5(1). https://doi.org/10.5210/ojphi.v5i1.4456

14. Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, Gomes G, et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clinical Microbiology and Infection. 2014; 20(1):17–21. https://doi.org/10.1111/1469-0691.12477 PMID: 24350723

15. Crawley A, Paolotti D, Dalton C, Brownstein J, Smolinski M. Global flu view: a platform to connect crowdsourced disease surveillance around the world. International Journal of Infectious Diseases. 2019; 79:7. https://doi.org/10.1016/j.ijid.2018.11.036

16. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PloS one. 2013; 8(12):e83672. https://doi.org/10.1371/journal.pone.0083672 PMID: 24349542

17. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS currents. 2014; 6.

18. Gittelman S, Lange V, Crawford CAG, Okoro CA, Lieb E, Dhingra SS, et al. A new source of data for public health surveillance: Facebook likes. Journal of medical Internet research. 2015; 17(4). https://doi.org/10.2196/jmir.3970 PMID: 25895907

19. Jha A, Lin L, Savoia E. The use of social media by state health departments in the US: analyzing health communication through Facebook. Journal of community health. 2016; 41(1):174–179. https://doi.org/10.1007/s10900-015-0083-4 PMID: 26318742

20. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS computational biology. 2014; 10(4):e1003581. https://doi.org/10.1371/journal.pcbi.1003581 PMID: 24743682

21. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS computational biology. 2014; 10(11):e1003892. https://doi.org/10.1371/journal.pcbi.1003892 PMID: 25392913

22. Althouse BM, Scarpino SV, Meyers LA, Ayers JW, Bargsten M, Baumbach J, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Science. 2015; 4(1):1.

23. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM, et al. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PLOS ONE. 2015; 10(10):1–20. https://doi.org/10.1371/journal.pone.0139701

24. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol. 2013; 9(10):e1003256. https://doi.org/10.1371/journal.pcbi.1003256 PMID: 24146603

25. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. Science. 2014; 343(6176):1203–1205. https://doi.org/10.1126/science.1248506 PMID: 24626916

26. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google Flu Trends? American journal of preventive medicine. 2014; 47(3):341–347. https://doi.org/10.1016/j.amepre.2014.05.020 PMID: 24997572

27. US C. BioSense 2.0; 2013. https://stacks.cdc.gov/view/cdc/13171.

28. US D. The Biosurveillance Ecosystem (BSVE); 2014. http://www.dtra.mil/Portals/61/Documents/bsve-fact-sheet_draft_05-01-2014_pa-cleared-distro-statement.pdf.

29. NA. Biosense Google Public Data Explorer; 2012. http://www.google.com/publicdata/overview?ds=z46e2n1b69u8mu_.

30. Valdivia A, Lopez-Alcalde J, Vicente M, Pichiule M, Ruiz M, Ordobas M. Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks-results for 2009-10. Eurosurveillance. 2010; 15(29):2–7. https://doi.org/10.2807/ese.15.29.19621-en

31. Baker M, Wilson N, Huang Q, Paine S, Lopez L, Bandaranayake D, et al. Pandemic influenza A (H1N1) v in New Zealand: the experience from April to August 2009. Eurosurveillance. 2008; 14(34):127–136.

32. 2007–2011 American Community Survey;. http://ftp2.census.gov/.

33. Mazumder R, Friedman JH, Hastie T. SparseNet: Coordinate descent with nonconvex penalties. Journal of the American Statistical Association. 2012;.

34. Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, et al. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. JMIR public health and surveillance. 2018; 4(1). https://doi.org/10.2196/publichealth.8950 PMID: 29317382

35. Thompson WW, Shay DK, Weintraub E, Brammer L, Bridges CB, Cox NJ, et al. Influenza-associated hospitalizations in the United States. Jama. 2004; 292(11):1333–1340. https://doi.org/10.1001/jama.292.11.1333 PMID: 15367555

36. Bivand RS, Pebesma E, Gomez-Rubio V. Applied spatial data analysis with R, Second edition. Springer, NY; 2013. Available from: http://www.asdar-book.org/.

37. Shi L, Starfield B, Kennedy B, Kawachi I. Income inequality, primary care, and health indicators.(Original Research). Journal of Family Practice. 1999; 48(4):275–285. PMID: 10229252

38. Wilkinson RG, Pickett KE. Income inequality and population health: a review and explanation of the evidence. Social science & medicine. 2006; 62(7):1768–1784. https://doi.org/10.1016/j.socscimed.2005.08.036

39. Richiardi L, Pizzi C, Paolotti D. Internet-based Epidemiology. In: Handbook of Epidemiology. Springer; 2014. p. 439–469.

40. Silver MP. Socio-economic status over the lifecourse and internet use in older adulthood. Ageing and Society. 2014; 34(06):1019–1034. https://doi.org/10.1017/S0144686X12001420

41. Fiscella K, Franks P, Gold MR, Clancy CM. Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care. Jama. 2000; 283(19):2579–2584. https://doi.org/10.1001/jama.283.19.2579 PMID: 10815125

42. Abbas KM, Kang GJ, Chen D, Werre SR, Marathe A. Demographics, perceptions, and socioeconomic factors affecting influenza vaccination among adults in the United States. PeerJ. 2018; 6:e5171. https://doi.org/10.7717/peerj.5171 PMID: 30013841

43. Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. Health affairs. 2002; 21(2):60–76. https://doi.org/10.1377/hlthaff.21.2.60 PMID: 11900187

44. Billings J, Zeitel L, Lukomnik J, Carey TS, Blank AE, Newman L. Impact of socioeconomic status on hospital use in New York City. Health affairs. 1993; 12(1):162–173. https://doi.org/10.1377/hlthaff.12.1.162 PMID: 8509018

45. Nsoesie E, Hawkins F, Maharana J, Skotnes A, Marinho T, JS B. Social Media as a Sentinel for Disease Surveillance: What Does Sociodemographic Status Have to Do with It? PLOS Currents Outbreaks. 2016;. https://doi.org/10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6

46. Gandhi SO, Grant LP, Sabik LM. Trends in nonemergent use of emergency departments by health insurance status. Medical Care Research and Review. 2014; p. 1077558714541481. https://doi.org/10.1177/1077558714541481 PMID: 25006044

47. Farmer P. Infections and inequalities: The modern plagues. Univ of California Press; 2001.

48. Marmot M. Social determinants of health inequalities. The Lancet. 2005; 365(9464):1099–1104. https://doi.org/10.1016/S0140-6736(05)71146-6

49. Liao Y, Tucker P, Okoro CA, Giles WH, Mokdad AH, Harris VB. REACH 2010 Surveillance for Health Status in Minority Communities—United States, 2001–2002. Morbidity and mortality weekly report Surveillance summaries (Washington, DC: 2002). 2004; 53(6):1–36.

50. Mensah GA, Mokdad AH, Ford ES, Greenlund KJ, Croft JB. State of disparities in cardiovascular health in the United States. Circulation. 2005; 111(10):1233–1241. https://doi.org/10.1161/01.CIR.0000158136.76824.04 PMID: 15769763

51. Kandula NR, Lauderdale DS, Baker DW. Differences in self-reported health among Asians, Latinos, and non-Hispanic whites: The role of language and nativity. Annals of epidemiology. 2007; 17(3):191–198. https://doi.org/10.1016/j.annepidem.2006.10.005 PMID: 17320786

52. Horbar JD, Edwards EM, Greenberg LT, Profit J, Draper D, Helkey D, et al. Racial segregation and inequality in the neonatal intensive care unit for very low-birth-weight and very preterm infants. JAMA pediatrics. 2019; 173(5):455–461. https://doi.org/10.1001/jamapediatrics.2019.0241 PMID: 30907924

53. Yousey-Hindes KM, Hadler JL. Neighborhood socioeconomic status and influenza hospitalizations among children: New Haven County, Connecticut, 2003–2010. American journal of public health. 2011; 101(9):1785–1789. https://doi.org/10.2105/AJPH.2011.300224 PMID: 21778498

54. Glaser SL, Clarke CA, Gomez SL, O'Malley CD, Purdie DM, West DW. Cancer surveillance research: a vital subdiscipline of cancer epidemiology. Cancer Causes & Control. 2005; 16(9):1009–1019. https://doi.org/10.1007/s10552-005-4501-2

55. Polgreen PM, Chen Z, Segre AM, Harris ML, Pentella MA, Rushton G. Optimizing influenza sentinel surveillance at the state level. American journal of epidemiology. 2009; p. kwp270. https://doi.org/10.1093/aje/kwp270 PMID: 19822570

56. Scarpino SV, Dimitrov NB, Meyers LA. Optimizing provider recruitment for influenza surveillance networks. PLoS Comput Biol. 2012; 8(4):e1002472. https://doi.org/10.1371/journal.pcbi.1002472 PMID: 22511860

57. Fairchild G, Polgreen PM, Foster E, Rushton G, Segre AM. How many suffice? A computational framework for sizing sentinel surveillance networks. International journal of health geographics. 2013; 12 (1):56. https://doi.org/10.1186/1476-072X-12-56 PMID: 24321203

58. Scarpino SV, Meyers LA, Johansson MA. Design strategies for efficient arbovirus surveillance. Emerging infectious diseases. 2017; 23(4):642. https://doi.org/10.3201/eid2304.160944 PMID: 28322711

59. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS computational biology. 2015; 11(10):e1004513. https://doi.org/10.1371/journal.pcbi.1004513 PMID: 26513245

60. Pelat C, Ferguson NM, White PJ, Reed C, Finelli L, Cauchemez S, et al. Optimizing the precision of case fatality ratio estimates under the surveillance pyramid approach. American journal of epidemiology. 2014; p. kwu213. https://doi.org/10.1093/aje/kwu213 PMID: 25255809

61. Navarro-Marí J, Pérez-Ruiz M, Cantudo-Muñoz P, Petit-Gancedo C, Jiménez-Valera M, Rosa-Fraile M, et al. Influenza-like illness criteria were poorly related to laboratory-confirmed influenza in a sentinel surveillance study. Journal of clinical epidemiology. 2005; 58(3):275–279. https://doi.org/10.1016/j.jclinepi.2004.08.014 PMID: 15768487

62. Das D, Metzger K, Heffernan R, Balter S, Weiss D, Mostashari F. Monitoring over-the-counter medication sales for early detection of disease outbreaks—New York City. MMWR Morb Mortal Wkly Rep. 2005; 54(Suppl):41–46.

63. Schmidt W, Pebody R, Mangtani P. School absence data for influenza surveillance: a pilot study in the United Kingdom. Euro surveillance. 2010; 15(3). PMID: 20122378

64. Eggo RM, Scott JG, Galvani AP, Meyers LA. Respiratory virus transmission dynamics determine timing of asthma exacerbation peaks: Evidence from a population-level model. Proceedings of the National Academy of Sciences. 2016; 113(8):2194–2199. https://doi.org/10.1073/pnas.1518677113

65. Cowling B, Wong I, Ho L, Riley S, Leung G. Methods for monitoring influenza surveillance data. International journal of epidemiology. 2006; 35(5):1314–1321. https://doi.org/10.1093/ije/dyl162 PMID: 16926216

66. Meyers L, Ginocchio CC, Faucett AN, Nolte FS, Gesteland PH, Leber A, et al. Automated Real-Time Collection of Pathogen-Specific Diagnostic Data: Syndromic Infectious Disease Epidemiology. JMIR public health and surveillance. 2018; 4(3). https://doi.org/10.2196/publichealth.9876 PMID: 29980501

67. Schuur JD, Venkatesh AK. The growing role of emergency departments in hospital admissions. New England Journal of Medicine. 2012; 367(5):391–393. https://doi.org/10.1056/NEJMp1204431 PMID: 22784039

68. Shea K, Tildesley MJ, Runge MC, Fonnesbeck CJ, Ferrari MJ. Adaptive management and the value of information: learning via intervention in epidemiology. PLoS biology. 2014; 12(10):e1001970. https://doi.org/10.1371/journal.pbio.1001970 PMID: 25333371

69. Biggerstaff M, Alper D, Dredze M, Fox S, Fung ICH, Hickmann KS, et al. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. BMC infectious diseases. 2016; 16(1):357. https://doi.org/10.1186/s12879-016-1669-x PMID: 27449080

70. Brownstein JS, Chu S, Marathe A, Marathe MV, Nguyen AT, Paolotti D, et al. Combining participatory influenza surveillance with modeling and forecasting: three alternative approaches. JMIR public health and surveillance. 2017; 3(4). https://doi.org/10.2196/publichealth.7344

71. Ertem Z, Raymond D, Meyers LA. Optimal multi-source forecasting of seasonal influenza. PLoS computational biology. 2018; 14(9):e1006236. https://doi.org/10.1371/journal.pcbi.1006236 PMID: 30180212

72. Viboud C, Vespignani A. The future of influenza forecasts. Proceedings of the National Academy of Sciences. 2019; p. 201822167. https://doi.org/10.1073/pnas.1822167116

73. Reich NG, McGowan CJ, Yamana TK, Tushar A, Ray EL, Osthus D, et al. A Collaborative Multi-Model Ensemble for Real-Time Influenza Season Forecasting in the US. PNAS. 2019; p. 201812594. https://doi.org/10.1073/pnas.1812594116

74. Zhang Q, Perra N, Perrotta D, Tizzoni M, Paolotti D, Vespignani A. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In Proceedings of the 26th international conference on world wide web. 2017; p. 311–319.

75. Yang C, Chen R, Chou W, Lee Y, Lo Y. An Integrated Influenza Surveillance Framework Based on National Influenza-Like Illness Incidence and Multiple Hospital Electronic Medical Records for Early Prediction of Influenza Epidemics: Design and Evaluation. Journal of medical Internet research. 2019; 21:e12341. https://doi.org/10.2196/12341 PMID: 30707099

76. Scarpino S, Petri G. On the predictability of infectious disease outbreaks. Nature communications. 2019; 10:898–892. https://doi.org/10.1038/s41467-019-08616-0 PMID: 30796206

77. Baltrusaitis K, Brownstein JS, Scarpino SV, Bakota E, Crawley AW, Conidi G, et al. Comparison of crowd-sourced, electronic health records based, and traditional health-care based influenza-tracking systems at multiple spatial resolutions in the United States of America. BMC infectious diseases. 2018; 18(1):403. https://doi.org/10.1186/s12879-018-3322-3 PMID: 30111305