

Original Paper

# Differences in Regional Patterns of Influenza Activity Across Surveillance Systems in the United States: Comparative Evaluation

Kristin Baltrusaitis<sup>1</sup>, PhD; Alessandro Vespignani<sup>2</sup>, PhD; Roni Rosenfeld<sup>3</sup>, PhD; Josh Gray<sup>4</sup>, MBA; Dorrie Raymond<sup>4</sup>, MA; Mauricio Santillana<sup>5,6</sup>, PhD

<sup>1</sup>Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States

<sup>2</sup>Institute for Network Science, Northeastern University, Boston, MA, United States

<sup>3</sup>School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States

<sup>4</sup>athenaResearch at athenahealth, Watertown, MA, United States

<sup>5</sup>Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

<sup>6</sup>Department of Pediatrics, Harvard Medical School, Boston, MA, United States

**Corresponding Author:**

Mauricio Santillana, PhD

Computational Health Informatics Program

Boston Children's Hospital

1 Autumn St

Boston, MA, 02215

United States

Phone: 1 617 599 5460

Email: [msantill@fas.harvard.edu](mailto:msantill@fas.harvard.edu)

## Abstract

**Background:** The Centers for Disease Control and Prevention (CDC) tracks influenza-like illness (ILI) using information on patient visits to health care providers through the Outpatient Influenza-like Illness Surveillance Network (ILINet). As participation in this system is voluntary, the composition, coverage, and consistency of health care reports vary from state to state, leading to different measures of ILI activity between regions. The degree to which these measures reflect actual differences in influenza activity or systematic differences in the methods used to collect and aggregate the data is unclear.

**Objective:** The objective of our study was to qualitatively and quantitatively compare national and region-specific ILI activity in the United States across 4 surveillance data sources—CDC ILINet, Flu Near You (FNY), athenahealth, and HealthTweets.org—to determine whether these data sources, commonly used as input in influenza modeling efforts, show geographical patterns that are similar to those observed in CDC ILINet's data. We also compared the yearly percentage of FNY participants who sought health care for ILI symptoms across geographical areas.

**Methods:** We compared the national and regional 2018-2019 ILI activity baselines, calculated using noninfluenza weeks from previous years, for each surveillance data source. We also compared measures of ILI activity across geographical areas during 3 influenza seasons, 2015-2016, 2016-2017, and 2017-2018. Geographical differences in weekly ILI activity within each data source were also assessed using relative mean differences and time series heatmaps. National and regional age-adjusted health care-seeking percentages were calculated for each influenza season by dividing the number of FNY participants who sought medical care for ILI symptoms by the total number of ILI reports within an influenza season. Pearson correlations were used to assess the association between the health care-seeking percentages and baselines for each surveillance data source.

**Results:** We observed consistent differences in ILI activity across geographical areas for CDC ILINet and athenahealth data. ILI activity for FNY displayed little variation across geographical areas, whereas differences in ILI activity for HealthTweets.org were associated with the total number of tweets within a geographical area. The percentage of FNY participants who sought health care for ILI symptoms differed slightly across geographical areas, and these percentages were positively correlated with CDC ILINet and athenahealth baselines.

**Conclusions:** Our findings suggest that differences in ILI activity across geographical areas as reported by a given surveillance system may not accurately reflect true differences in the prevalence of ILI. Instead, these differences may reflect systematic collection and aggregation biases that are particular to each system and consistent across influenza seasons. These findings are potentially relevant in the real-time analysis of the influenza season and in the definition of unbiased forecast models.

**KEYWORDS**

digital disease surveillance; influenza; surveillance; participatory syndromic surveillance; disease modeling

## Introduction

### Background

Influenza epidemics are responsible for a significant public health burden that includes an estimated 12,000 to 56,000 deaths each year in the United States [1]. Consequently, timely and reliable surveillance of influenza activity is essential for local, state, and national public health officials to monitor and respond to outbreaks. In the United States, the Centers for Disease Control and Prevention (CDC) collects and analyzes information on influenza activity throughout the year. As part of this national surveillance system, patients seeking medical attention for influenza-like illness (ILI) are tracked through the Outpatient Influenza-like Illness Surveillance Network (ILINet). This system contains thousands of volunteer health care specialists, including individual providers, group practices, and hospital-based clinics located throughout all 50 states, Puerto Rico, the District of Columbia, and the US Virgin Islands. As participation in ILINet is voluntary and each state is responsible for their own recruitment of health care providers, the composition of provider types, coverage of geographical regions, and consistency of provider reporting vary from state to state. This convenience sample-driven model of surveillance results in certain parts of the population being over- or underrepresented in the reported influenza activity [2-4].

At both national and Health and Human Services-defined regional levels (conglomerates of 2-8 states), the CDC routinely reports the weekly percentage of patients presenting with ILI to health care providers. In addition, the CDC calculates and reports region-specific baselines, using influenza activity data from previous seasons, to identify the beginning and end of the influenza season and contextualize the severity of a given region-specific outbreak. These baselines vary widely across regions, and the degree to which the differences in baselines, as well as the percentage of ILI visits during an influenza season, reflect actual differences in influenza activity or systematic differences in the methods used to collect the data is unclear. Recent models suggest that the spatial patterns in US sentinel ILI surveillance may be the result of socioenvironmental factors, state-specific health policies, and sampling [3]. Identifying and characterizing the presence of potential methodological measurement biases in ILINet is important, as it is frequently used as an indicator of influenza activity for decision-making purposes and as the ground truth in mechanistic and statistical predictive modeling efforts aimed at understanding disease transmission dynamics and monitoring and forecasting influenza activity [5-15]. Furthermore, because these models typically leverage data from outside of the public health systems, such as Google internet searches [15,16], participatory syndromic surveillance systems [17,18], Twitter [19], and electronic health record (EHR) [14,20], it is important to understand if input sources show similar structural aggregation patterns.

### Objectives

In this study, we qualitatively and quantitatively compared national and region-specific baselines and ILI activity during 3 influenza seasons across 4 surveillance data sources—CDC ILINet; Flu Near You (FNY), a crowd-sourced participatory syndromic surveillance system; athenahealth, a provider of cloud-based EHR services; and HealthTweets.org, a research platform that shares health trends data from Twitter—to determine whether these surveillance data sources, commonly used as input in influenza modeling efforts, show regional structural patterns that are similar to those observed in CDC ILINet's data. We also compared yearly self-reported health care-seeking rates of FNY participants to determine if this factor can better characterize the differences in ILI activity across geographic areas.

## Methods

### Data

#### *Centers for Disease Control and Prevention Outpatient Influenza-Like Illness Surveillance Network*

The CDC reports the weighted percentage of patient visits to health care providers presenting ILI symptoms on a weekly basis at the national and regional levels. These values are weighted on the basis of state population and represent the percentage of patient visits to health care providers that present as ILI, defined as fever (temperature of 100°F [37.8°C] or greater) plus a cough and/or a sore throat without a known cause other than influenza.

#### *Flu Near You*

FNY is a participatory syndromic surveillance system that allows volunteers in the United States to report health information of the user and their family using brief weekly surveys [21]. Through these surveys, FNY users report any symptoms that they or any registered household members experienced during the previous week (Monday through Sunday). For all reported symptoms, FNY users are asked to provide the date of symptom(s) onset and whether or not they received medical care for the symptom(s). The national and regional percentage of ILI symptoms reported is calculated by dividing the number of participants reporting ILI, as defined by reporting fever plus cough and/or sore throat, in a given week by the total number of FNY participant reports in that same week. FNY participants are assigned to a region based on the zip code provided at registration. Unweighted FNY percentage of ILI symptoms is used to maintain consistency across previous studies and the FNY website.

athenahealth is a provider of cloud-based services and mobile apps for medical groups and health systems. National and regional percentage of visits for ILI is calculated by dividing the unspecified viral or ILI visit count, which is equal to the number of visits where the patient had an unspecified viral

diagnosis, an influenza diagnosis, or a fever diagnosis with an accompanying sore throat or cough diagnosis, by the total number of visits for each week.

### HealthTweets.org

This dataset is generated by a Web-based research platform (HealthTweets.org) that shares the output of Twitter data mining algorithms with researchers and public officials [19]. We use weekly aggregated trends data from each state to calculate the influenza prevalence measure for each region. Weekly national and regional influenza prevalence measures are calculated by normalizing the number of influenza infection tweets in the health stream by the total number of tweets in the general stream during the same week [22].

### Statistics of Datasets

#### Baseline Comparison

Baselines are used as a single quantitative measure that compares ILI activity during noninfluenza weeks across geographical areas within each surveillance data source. The CDC ILINet national and regional baselines for the 2018-2019 influenza season are available on the CDC website [23].

**Figure 1.** Equation for the mean relative difference.

$$\text{Mean Relative Difference} = \frac{1}{K} \sum_{\text{week}_k} \frac{ILI_{ik} - ILI_{jk}}{\max(ILI_{ik}, ILI_{jk})} \text{ for } i, j \in \{\text{Region 1: 10, National}\}$$

Mean relative differences within each surveillance data source are summarized using heatmaps, where the geographical areas along the rows are represented by  $i$  in the equation and the geographical areas along the columns are represented by  $j$ . Geographical areas that have consistently higher weekly ILI activity compared with other geographical areas have positive mean relative differences, indicated by red shades across the row in the heatmap, whereas geographical areas that have consistently lower weekly ILI activity have negative mean relative differences, indicated by blue shades across the row. Time series heatmaps are also presented to qualitatively compare weekly ILI activity across geographical areas for each surveillance data source.

#### Health Care-Seeking Behavior

National and regional health care-seeking percentages are calculated for each influenza season by dividing the number of FNY participants who sought medical care for ILI symptoms, as defined above, by the total number of ILI reports within an influenza season. As health care-seeking behavior varies by age [24], health care-seeking percentages are adjusted by age group (<18 years, 18-49 years, 50-64 years, and ≥65 years) using population data from the 2010 US census [25]. We use

National and regional baselines for FNY, athenahealth, and HealthTweets.org are estimated following the CDC's baseline definition. A baseline is defined as the mean percentage of ILI activity during *noninfluenza weeks*, for the previous 3 seasons, plus 2 SDs. *Noninfluenza weeks* during these seasons are the same for all 3 systems and are delineated, by the CDC, as periods of 2 or more consecutive weeks in which each week accounted for less than 2% of the season's total number of specimens that tested positive for influenza in public health laboratories. We used region-specific noninfluenza weeks. Descriptive statistics of baselines are presented as median (interquartile range, IQR).

#### Influenza-Like Illness Activity Comparison

Differences in ILI activity during noninfluenza as well as influenza weeks across geographical areas within each surveillance data source are assessed using data from the start of the 2015-2016 influenza season (week of October 5, 2015) to the end of the 2017-2018 influenza season (week of October 1, 2018). Weekly ILI activity across geographical areas within each data source is quantitatively compared by dividing the difference in ILI activity between 2 areas by the maximum within each week, defined by Figure 1.

Pearson correlation to assess the association between the adjusted health care-seeking percentages and baselines for each surveillance data source. All analyses are performed using R version 3.3.2. [26].

## Results

### Centers for Disease Control and Prevention Outpatient Influenza-Like Illness Surveillance Network

Table 1 and Figure 2 provide the ILI activity baselines for each surveillance data source across geographical areas. The national baseline for CDC ILINet during the 2018-2019 influenza season is 2.2, and the median CDC ILINet regional baseline is 2.1 (IQR 1.8-2.3). Region 10 has the smallest baseline, 1.1, whereas region 6 has the largest baseline, 4.0. As shown in Figure 3 regions 2 and 6 have consistently higher weekly percentage of ILI visits compared with other regions, indicated by the red shades across the row, whereas regions 1, 8, and 10 have consistently lower weekly percentage of ILI visits, indicated by the blue shades across the row. This pattern is also shown qualitatively in both Figure 4 and Multimedia Appendix 1, where darker shades of red, as seen for regions 2, 6, and 9, correspond to higher percentage of ILI visits.

**Table 1.** Regional and national influenza-like illness activity baselines for the 2018-2019 influenza season for Centers for Disease Control and Prevention Outpatient Influenza-like Illness Surveillance Network, Flu Near You, athenahealth, and HealthTweets.org.

Geographical area	Centers for Disease Control and Prevention Outpatient Influenza-like Illness Surveillance Network	Flu Near You	athenahealth	HealthTweets.org
Region 1 <sup>a</sup>	1.8	2.1	1.3	0.8
Region 2 <sup>b</sup>	3.1	2.4	1.7	0.4
Region 3 <sup>c</sup>	2.0	2.4	1.5	0.5
Region 4 <sup>d</sup>	2.2	2.7	1.4	0.6
Region 5 <sup>e</sup>	1.8	2.6	1.1	0.5
Region 6 <sup>f</sup>	4.0	2.6	1.9	0.7
Region 7 <sup>g</sup>	1.6	2.5	1.0	0.7
Region 8 <sup>h</sup>	2.2	2.9	1.0	0.8
Region 9 <sup>i</sup>	2.3	2.5	1.7	0.6
Region 10 <sup>j</sup>	1.1	2.5	0.6	0.7
National	2.2	2.3	1.4	0.5

<sup>a</sup>Region 1 includes Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont.

<sup>b</sup>Region 2 includes New Jersey, New York, Puerto Rico, and US Virgin Islands.

<sup>c</sup>Region 3 includes Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia.

<sup>d</sup>Region 4 includes Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee.

<sup>e</sup>Region 5 includes Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin.

<sup>f</sup>Region 6 includes Arkansas, Louisiana, New Mexico, Oklahoma, and Texas.

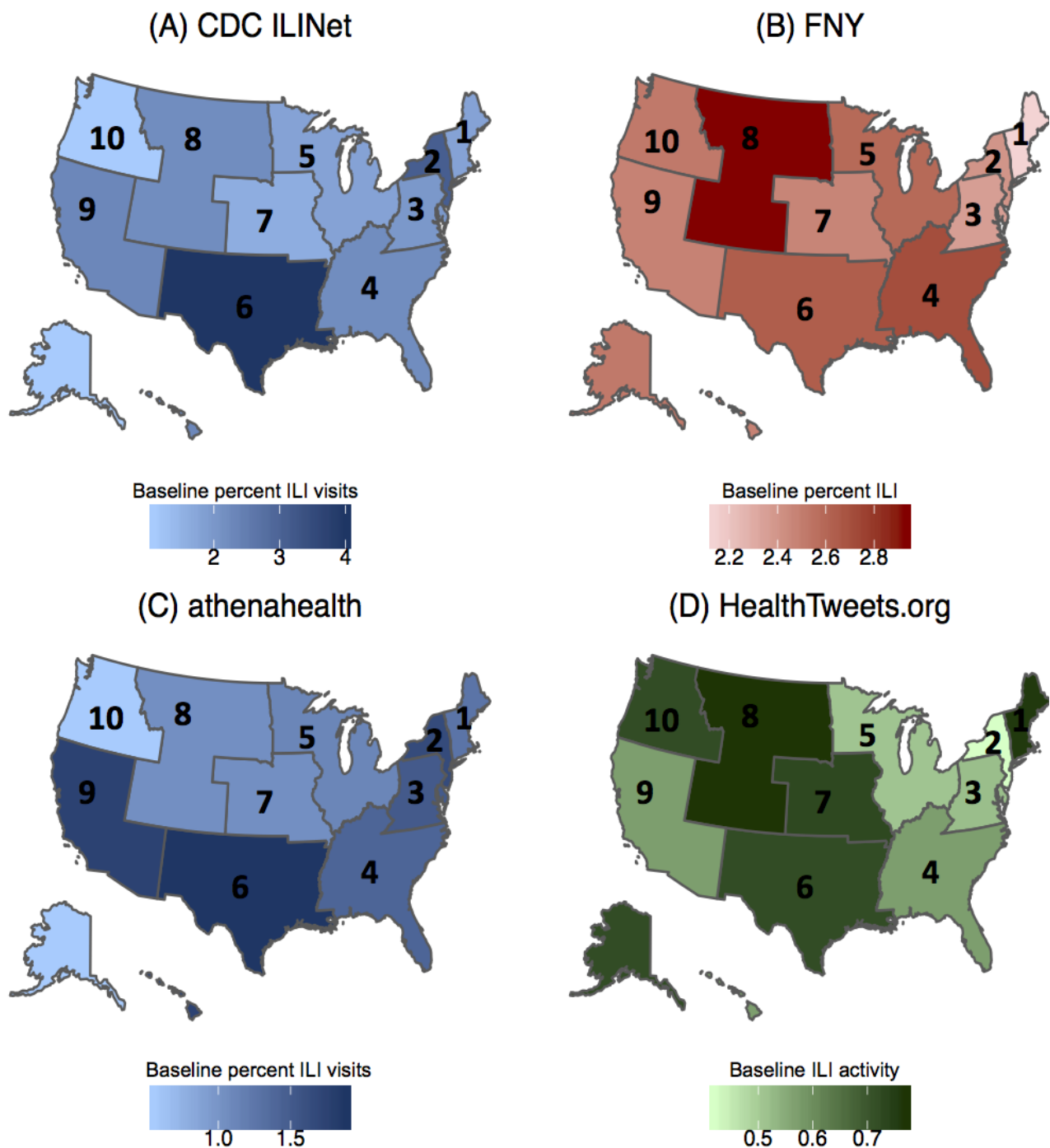
<sup>g</sup>Region 7 includes Iowa, Kansas, Missouri, and Nebraska.

<sup>h</sup>Region 8 includes Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming.

<sup>i</sup>Region 9 includes Arizona, California, Guam, Hawaii, and Nevada.

<sup>j</sup>Region 10 includes Alaska, Idaho, Oregon, and Washington.

**Figure 2.** Spatial heatmaps of US regional baseline influenza-like illness activity for the 2018-2019 influenza season for (A) Centers for Disease Control and Prevention Influenza-like Illness Surveillance Network, (B) Flu Near You, (C) athenahealth, and (D) HealthTweets.org. CDC: Centers for Disease Control and Prevention; FNY: Flu Near You; ILINet: Influenza-like Illness Surveillance Network.



**Flu Near You**

For FNY, the national baseline is 2.3, and the median regional baseline is 2.5 (IQR 2.4-2.6). The minimum baseline is 2.1, region 1, and the maximum baseline is 2.9, region 8. Compared with other data sources, the mean relative differences for FNY in Figure 3 show less heterogeneity and no consistent patterns in the percentage of ILI across geographical areas. Although the timing of peaks in the percentage of ILI varies between regions, the relative percentage of ILI is consistent across

geographical areas and seasons (Figure 4 and Multimedia Appendix 2).

The national baseline for athenahealth is 1.4, and the median regional baseline is 1.3 (IQR 1.0-1.6). Region 10 has the minimum baseline of 0.6, and region 6 has the maximum baseline of 1.9. Similar to CDC ILINet, regions 2, 6, and 9 have consistently higher weekly percentage of ILI visits compared with other regions, and regions 7, 8, and 10 have consistently lower weekly percentage of ILI visits. This pattern is reflected in Figure 4 and Multimedia Appendix 3, as regions 2, 6, and 9

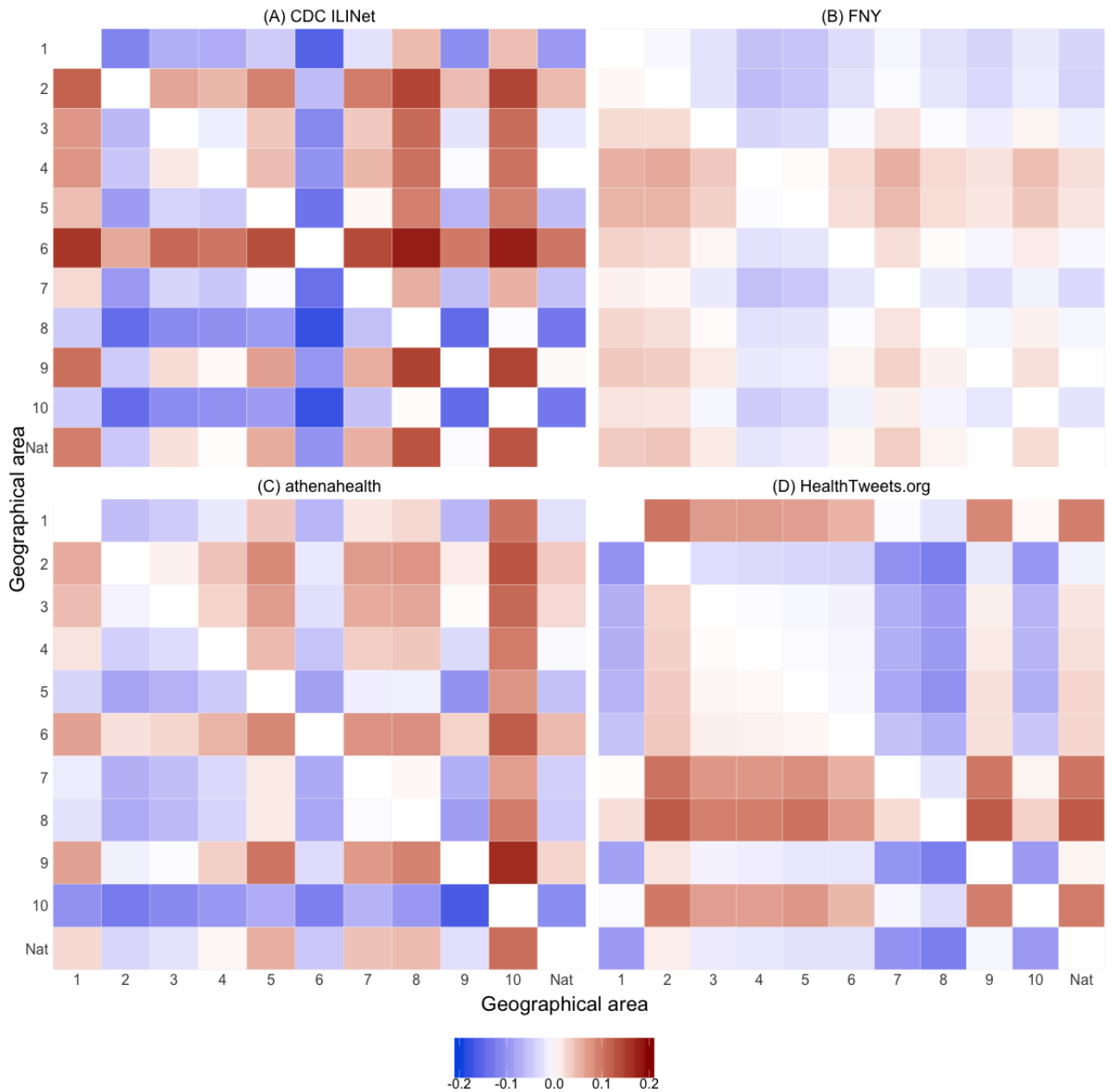
have consistently higher percentage of ILI visits across all seasons.

**HealthTweets.org**

The national baseline is 0.5, the median baseline is 0.6 (IQR 0.5-0.7), the minimum baseline is 0.4 (region 2), and the maximum baseline is 0.8 (region 8). Unlike CDC ILINet and

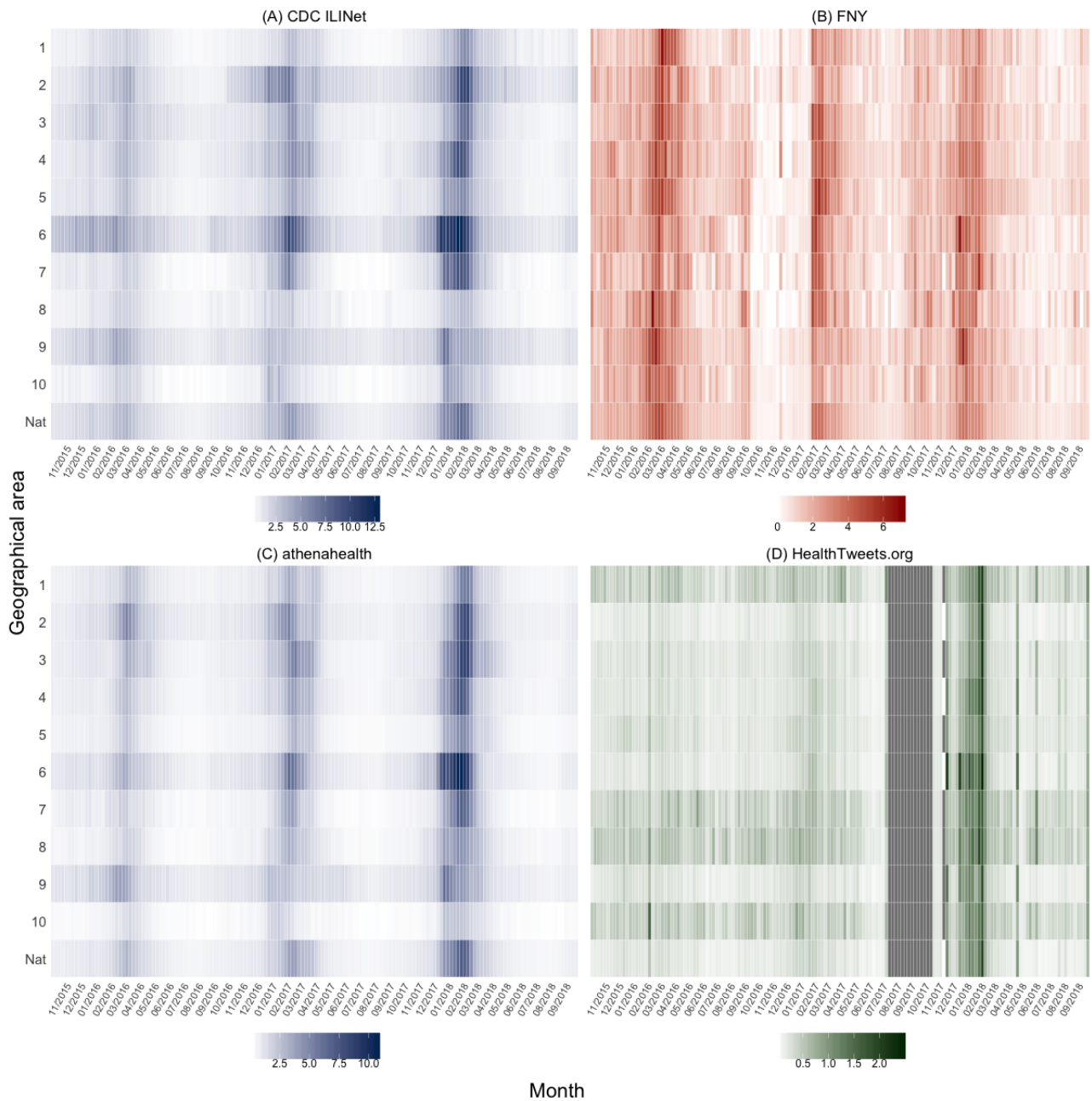
athenahealth, HealthTweets.org shows higher ILI activity in regions 1, 7, 8, and 10 (Figure 3). These regions have mean normalizing constants that are less than half the mean normalizing constants of other regions (Table 2). As shown in Figure 4 and Multimedia Appendix 4, this pattern is consistent across seasons.

**Figure 3.** Heatmaps of the mean relative difference of influenza-like illness activity across geographical areas for (A) Centers for Disease Control and Prevention Influenza-like Illness Surveillance Network, (B) Flu Near You, (C) athenahealth, and (D) HealthTweets.org. CDC: Centers for Disease Control and Prevention; FNY: Flu Near You; ILINet: Influenza-like Illness Surveillance Network.





**Figure 4.** Time series heatmaps of influenza-like illness activity across geographical areas for (A) Centers for Disease Control and Prevention Influenza-like Illness Surveillance Network, (B) Flu Near You, (C) athenahealth, and (D) HealthTweets.org. CDC: Centers for Disease Control and Prevention; FNY: Flu Near You; ILI: influenza-like illness ILINet: Influenza-like Illness Surveillance Network.



**Table 2.** Descriptive statistics of the HealthTweets.org normalizing constant at the national and regional levels.

Geographical area	Normalizing constant, mean (SD)
Region 1 <sup>a</sup>	210.82 (114.917)
Region 2 <sup>b</sup>	627.69 (330.270)
Region 3 <sup>c</sup>	599.53 (293.320)
Region 4 <sup>d</sup>	1103.78 (553.374)
Region 5 <sup>e</sup>	798.25 (387.266)
Region 6 <sup>f</sup>	845.30 (414.785)
Region 7 <sup>g</sup>	171.05 (82.077)
Region 8 <sup>h</sup>	121.96 (63.936)
Region 9 <sup>i</sup>	5848.54 (3775.923)
Region 10 <sup>j</sup>	181.33 (97.756)
National	6352.25 (3351.390)

<sup>a</sup>Region 1 includes Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont.

<sup>b</sup>Region 2 includes New Jersey, New York, Puerto Rico, and US Virgin Islands.

<sup>c</sup>Region 3 includes Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia.

<sup>d</sup>Region 4 includes Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee.

<sup>e</sup>Region 5 includes Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin.

<sup>f</sup>Region 6 includes Arkansas, Louisiana, New Mexico, Oklahoma, and Texas.

<sup>g</sup>Region 7 includes Iowa, Kansas, Missouri, and Nebraska.

<sup>h</sup>Region 8 includes Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming.

<sup>i</sup>Region 9 includes Arizona, California, Guam, Hawaii, and Nevada.

<sup>j</sup>Region 10 includes Alaska, Idaho, Oregon, and Washington.

### Health Care–Seeking Behavior

The age-adjusted estimates of the percentage of FNY participants who sought health care for ILI symptoms are shown by season and across all seasons in [Table 3](#) and [Figure 5](#). At the national level, a higher age-adjusted percentage of participants sought health care for ILI symptoms during the 2016-2017 season, 35.1%, compared with the 2015-2016 and 2017-2018 seasons, 21.7% and 29.2%, respectively. Within each season, regions 2, 4, and 6 have the highest age-adjusted percentages of participants who sought health care, whereas regions 1, 5, 9,

and 10 have the smallest age-adjusted percentages of participants who sought health care.

As shown in [Figure 6](#), the age-adjusted estimates of the percentage of individuals who sought health care for ILI symptoms is significantly correlated with the baselines for CDC ILINet ( $P=.03$ ) and is borderline significant for athenahealth ( $P=.08$ ). There is no evidence of an association between the age-adjusted estimates of the percentage of individuals who sought health care and the baselines for FNY ( $P=.68$ ) and HealthTweets.org ( $P=.76$ ).



**Table 3.** Age-adjusted regional and national estimates of the percentage of Flu Near You participants who sought health care for influenza-like illness symptoms.

Geographical area	All seasons	2015-2016	2016-2017	2017-2018
Region 1 <sup>a</sup>	25.98	20.82	33.29	27.77
Region 2 <sup>b</sup>	29.97	26.05	36.03	31.79
Region 3 <sup>c</sup>	28.66	22.07	37.03	31.73
Region 4 <sup>d</sup>	32.61	25.47	43.23	34.77
Region 5 <sup>e</sup>	26.43	21.53	34.59	26.73
Region 6 <sup>f</sup>	35.17	28.58	44.83	37.47
Region 7 <sup>g</sup>	30.93	23.79	41.95	32.09
Region 8 <sup>h</sup>	25.50	22.74	30.86	26.16
Region 9 <sup>i</sup>	22.49	19.06	27.77	24.69
Region 10 <sup>j</sup>	20.03	17.03	23.39	22.33
National	27.12	21.73	35.06	29.23

<sup>a</sup>Region 1 includes Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont.

<sup>b</sup>Region 2 includes New Jersey, New York, Puerto Rico, and US Virgin Islands.

<sup>c</sup>Region 3 includes Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia.

<sup>d</sup>Region 4 includes Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee.

<sup>e</sup>Region 5 includes Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin.

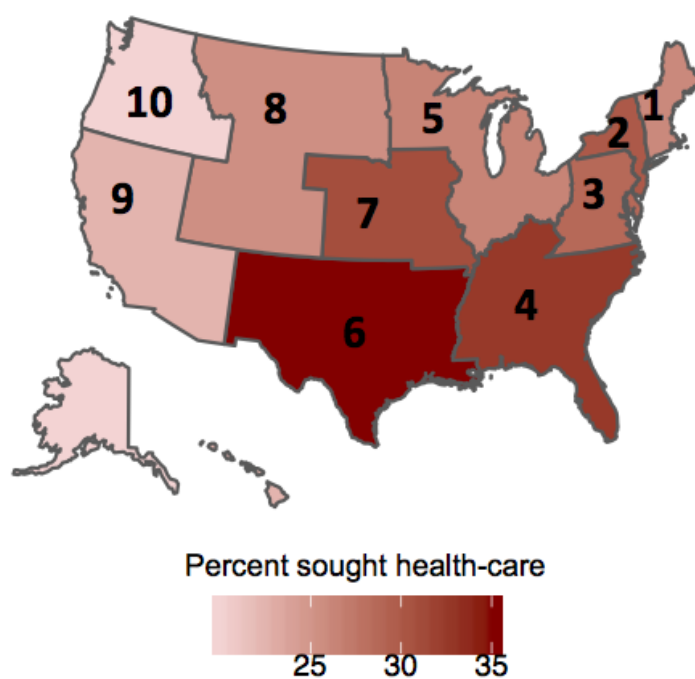
<sup>f</sup>Region 6 includes Arkansas, Louisiana, New Mexico, Oklahoma, and Texas.

<sup>g</sup>Region 7 includes Iowa, Kansas, Missouri, and Nebraska.

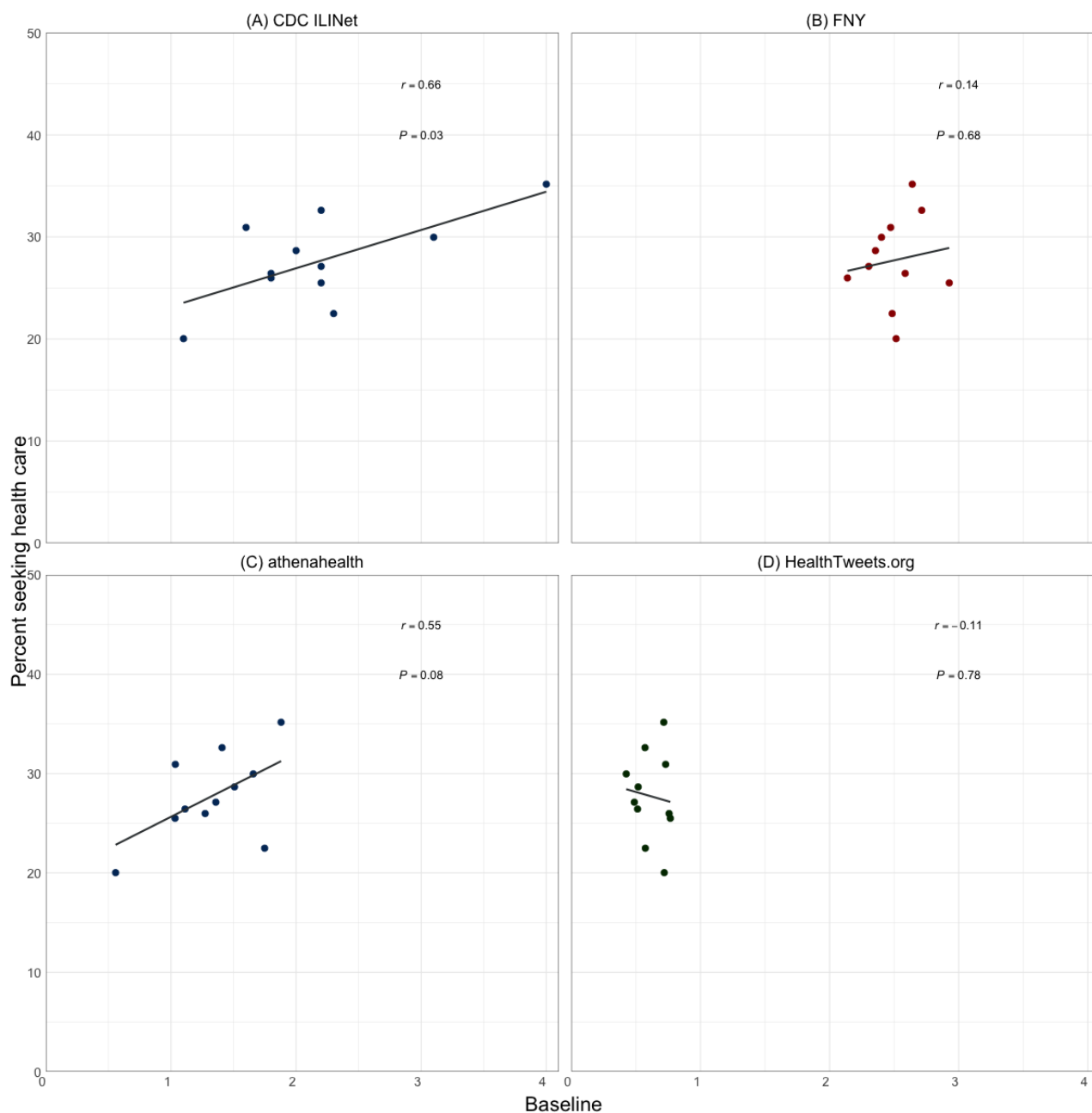
<sup>h</sup>Region 8 includes Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming.

<sup>i</sup>Region 9 includes Arizona, California, Guam, Hawaii, and Nevada.

<sup>j</sup>Region 10 includes Alaska, Idaho, Oregon, and Washington.

**Figure 5.** Spatial heatmap of age-adjusted regional percentage of Flu Near You participants who sought health care for ILI symptoms across all seasons.

**Figure 6.** Scatterplot of the age-adjusted percentage of individuals who sought health care for influenza-like illness symptoms versus baseline influenza-like illness activity for (A) Centers for Disease Control and Prevention Influenza-like Illness Surveillance Network, (B) Flu Near You, (C) athenahealth, and (D) HealthTweets.org. CDC: Centers for Disease Control and Prevention; FNY: Flu Near You; ILINet: Influenza-like Illness Surveillance Network.



## Discussion

### Principal Findings

Our findings show that differences in ILI activity across regions, as reported by a given surveillance system, are not consistent across surveillance platforms. In other words, regions that show larger baselines (and thus higher overall historical ILI activity) in one surveillance system appear to be different from their counterparts in other surveillance systems. The heterogeneity of recruitment practices of health care providers for each system, the composition of provider types, and the variability and consistency of coverage of geographical regions have the potential to contribute substantially to these systematic differences in baselines [3]. As such, our findings suggest that

these structural differences reflect methodological collection practices rather than actual differences in influenza activity across regions. The observed structural patterns within each surveillance system are consistent across individual influenza seasons (Multimedia Appendix 5), which implies that the differences are not reflecting a specific time-period heterogeneity.

Specifically, baselines from CDC ILINet vary across different geographical areas, and the geographical areas with the largest baselines also have a consistently larger percentage of ILI visits during the influenza season. Conversely, FNY's baselines and the percentage of ILI were similar across geographical areas. This similarity is captured by the homogeneity in the mean relative differences. One potential contributing factor to the

observed differences in patterns between these surveillance systems is the activity being measured. CDC ILINet measures the number of visits with ILI out of the total number of patient visits, whereas FNY measures the number of ILI reports out of enrolled persons who submitted a report. Furthermore, the population under surveillance also differs as FNY includes individuals who may not seek medical attention and FNY has a different demographic profile compared with CDC ILINet. For example, females and middle-aged participants are overrepresented in FNY [27].

Although not identical, athenahealth shows similar patterns in both baseline measures as well as the percentage of ILI visits to CDC ILINet across geographical areas. Both CDC ILINet and athenahealth use data from individuals seeking medical care. However, athenahealth has only a partially overlapping coverage of health care providers, and the proportion of visit settings differs slightly between the 2 systems. Most of athenahealth's providers see patients in office-based settings. Other settings, such as emergency room and nursing facilities, are underrepresented compared with CDC ILINet [28].

Unlike FNY, patterns across geographical areas within HealthTweets.org ILI activity appear to be the opposite of the patterns shown by CDC ILINet and athenahealth, as areas with consistently lower HealthTweets.org ILI activity have a consistently higher percentage of ILI visits for CDC ILINet and athenahealth, and vice versa. One potential reason for the differences in patterns in ILI activity across data sources is the difference in the activity being measured. As mentioned above, both CDC ILINet and athenahealth measure the number of ILI visits out of total visits, whereas HealthTweets.org normalizes the number of influenza infection tweets by the total number of tweets in the general stream. In addition, the groups most susceptible to influenza illness, young children and the elderly, may be underrepresented on Twitter. Furthermore, we found that smaller normalizing constants correspond to higher values of ILI activity.

### Comparison With Previous Work

Despite the differences in patterns of ILI activity within systems, current research shows that these alternative data sources track CDC ILINet at both the national and regional levels. At the national level, the correlation between CDC ILINet and athenahealth is 0.97, and regional correlations range from 0.90 to 0.97 [29]. The correlation between CDC ILINet and FNY at the national level is 0.81, and regional correlations range from 0.64 to 0.81 [29]. Twitter-based influenza prevalence measures show a correlation of 0.93 with CDC ILINet at the national level and a correlation of 0.88 with New York City's weekly emergency department visits for ILI [22].

Compared with other recent publications, the percentage of FNY participants who sought medical care for ILI is less than reported estimates. A recent meta-analysis that used estimates from multiple countries across different influenza seasons estimated an overall pooled health care-seeking rate of 0.52 (95% CI 0.46-0.59) [30]. In the United States, national reported health care-seeking percentages for children were 56% and 57% during the 2009-2010 and 2010-2011 influenza seasons, respectively. Among adults, 40% reported that they sought

health care during the 2009-2010 influenza season and 45% reported that they sought health care during the 2009-2010 influenza season [24,31]. Interestingly, the percentage of FNY participants who sought health care for ILI symptoms differs slightly across geographical areas. These differences may contribute to the differences in CDC ILINet and athenahealth baseline activity, as health care-seeking percentages are positively correlated with both CDC ILINet and athenahealth baselines.

From a predictive modeling perspective, our findings may explain why certain approaches designed to predict CDC ILINet values for the *Predict the Influenza season challenge*, weeks ahead of the publication of official CDC reports, may work better than others. As discussed in the 2 existing reports that document the performance of different methodologies to predict influenza activity, models that rely on local statistical approaches that exploit region-specific autoregressive information and historically observed ILI activity from previous seasons, as well as external predictors (such as humidity data, Google searches, and Wikipedia) [9,11], outperform mechanistic agent-based stochastic susceptible-infected-recovered (SIR) models that aim at modeling individual humans' behavior to infer epidemic activity across spatial resolutions [7,8,10]. The former modeling approaches are *trained* to track ILI levels in a region-specific fashion (frequently ignoring inconsistency across spatial resolutions), whereas the latter agent-based stochastic SIR models aim to predict the whole national epidemic outbreak across geographic areas. In other words, if the ILI activity report varies depending on how data are aggregated, then even a very accurate agent-based model may not be able to capture influenza activity correctly for every geographic area.

### Limitations

Our study has several limitations. During the beginning of the 2015-2016 season, there were errors in FNY data collection, resulting in an underestimation in the weekly percentage of ILI reports. We did not adjust the estimates of these weeks. There was also an issue in data collection during the week of August 28, 2017. We adjusted the estimates for this week by taking the average percentage of ILI reports of the previous and subsequent weeks. In addition, there were a few weeks during the summer of 2017 during which there were no reports of ILI activity for HealthTweets.org. We did not input or estimate these missing weeks. As the overall patterns of ILI activity were similar across seasons (Multimedia Appendix 5), we do not suspect that these data issues affected our overall conclusions.

In addition, FNY relies on self-reported data that are subject to recall and social desirability bias. FNY participant reporting is also not consistent throughout the influenza season. Although previous studies have used various methods, including restricting analyses to cohorts of users that report regularly [32-34], dropping the first report of all users, and applying a spike detector [21], we did not adjust for these potential reporting biases because reporting habits are consistent across regions [27]. Finally, because each system has a different measure of ILI activity, we cannot directly compare measures across systems.

## Conclusions

Although ILI activity differs across geographical areas and data sources, the general region-specific seasonal trends are similar

and provide valuable information about changes in influenza activity. Together, these platforms offer a more comprehensive view of influenza surveillance that helps public health offices monitor and respond to seasonal influenza epidemics.

## Acknowledgments

This work was supported by National Institute of Health, National Institute of General Medical Sciences (grant number T32GM074905 to KB). AV is partly funded by Models of Infectious Disease Agent Study, National Institute of General Medical Sciences (grant number U54GM111274). MS and AV were partially funded by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM130668. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge all the participants who contributed their time and information to the FNY system.

## Authors' Contributions

KB, MS, AV, and RR conceived the research. KB conducted the statistical analysis. KB, MS, AV, RR, and JG drafted the manuscript. JG and DR made substantial contributions to the acquisition and collection of data. All authors critically revised the intellectual content of the manuscript and approved the final version.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Time series plots of weekly percentage of influenza-like illness visits from Centers for Disease Control and Prevention Influenza-like Illness Surveillance Network across 3 influenza seasons (2015-2016, 2016-2017, and 2017-2018) with baselines. Geographical areas on the columns are presented in black, and geographical areas on the rows are presented in blue.

[\[PNG File 426 KB-Multimedia Appendix 1\]](#)

## Multimedia Appendix 2

Time series plots of weekly percentage of influenza-like illness from Flu Near You across 3 influenza seasons (2015-2016, 2016-2017, and 2017-2018) with baselines. Geographical areas on the columns are presented in black, and geographical areas on the rows are presented in red.

[\[PNG File 462 KB-Multimedia Appendix 2\]](#)

## Multimedia Appendix 3

Time series plots of weekly percentage of influenza-like illness visits from athenahealth across 3 influenza seasons (2015-2016, 2016-2017, and 2017-2018) with baselines. Geographical areas on the columns are presented in black, and geographical areas on the rows are presented in blue.

[\[PNG File 396 KB-Multimedia Appendix 3\]](#)

## Multimedia Appendix 4

Time series plots of weekly influenza-like illness activity from HealthTweets.org across 3 influenza seasons (2015-2016, 2016-2017, and 2017-2018) with baselines. Geographical areas on the columns are presented in black, and geographical areas on the rows are presented in green.

[\[PNG File 386 KB-Multimedia Appendix 4\]](#)

## Multimedia Appendix 5

Heatmaps of the mean relative difference of ILI activity across geographical areas for (A) Centers for Disease Control and Prevention Influenza-like Illness Surveillance Network, (B) Flu Near You, (C) athenahealth, and (D) HealthTweets.org for each influenza season.

[\[PNG File 113 KB-Multimedia Appendix 5\]](#)

## References

1. Molinari NM, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, et al. The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine* 2007 Jun 28;25(27):5086-5096. [doi: [10.1016/j.vaccine.2007.03.046](https://doi.org/10.1016/j.vaccine.2007.03.046)] [Medline: [17544181](https://pubmed.ncbi.nlm.nih.gov/17544181/)]

2. Polgreen PM, Chen Z, Segre AM, Harris ML, Pentella MA, Rushton G. Optimizing influenza sentinel surveillance at the state level. *Am J Epidemiol* 2009 Nov 15;170(10):1300-1306 [FREE Full text] [doi: [10.1093/aje/kwp270](https://doi.org/10.1093/aje/kwp270)] [Medline: [19822570](https://pubmed.ncbi.nlm.nih.gov/19822570/)]
3. Lee EC, Arab A, Goldlust SM, Viboud C, Grenfell BT, Bansal S. Deploying digital health data to optimize influenza surveillance at national and local scales. *PLoS Comput Biol* 2018 Mar;14(3):e1006020 [FREE Full text] [doi: [10.1371/journal.pcbi.1006020](https://doi.org/10.1371/journal.pcbi.1006020)] [Medline: [29513661](https://pubmed.ncbi.nlm.nih.gov/29513661/)]
4. Scarpino SV, Dimitrov NB, Meyers LA. Optimizing provider recruitment for influenza surveillance networks. *PLoS Comput Biol* 2012;8(4):e1002472 [FREE Full text] [doi: [10.1371/journal.pcbi.1002472](https://doi.org/10.1371/journal.pcbi.1002472)] [Medline: [22511860](https://pubmed.ncbi.nlm.nih.gov/22511860/)]
5. Zhang Q, Gioannini C, Paolotti D, Perra N, Perrotta D, Quaggiotto M, et al. Social Data Mining and Seasonal Influenza Forecasts: The FluOutlook Platform. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2015 Presented at: ECML PKDD'15; September 7-11, 2015; Porto, Portugal p. 237-240 URL: [https://doi.org/10.1007/978-3-319-23461-8\\_21](https://doi.org/10.1007/978-3-319-23461-8_21) [doi: [10.1007/978-3-319-23461-8\\_21](https://doi.org/10.1007/978-3-319-23461-8_21)]
6. 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 Presented at: WWW'17; April 3-7, 2017; Perth, Australia p. 311-319 URL: <https://doi.org/10.1145/3038912.3052678> [doi: [10.1145/3038912.3052678](https://doi.org/10.1145/3038912.3052678)]
7. Biggerstaff M, Alper D, Dredze M, Fox S, Fung IC, Hickmann KS, Influenza Forecasting Contest Working Group. Results from the centers for disease control and prevention's predict the 2013-2014 influenza season challenge. *BMC Infect Dis* 2016 Jul 22;16:357 [FREE Full text] [doi: [10.1186/s12879-016-1669-x](https://doi.org/10.1186/s12879-016-1669-x)] [Medline: [27449080](https://pubmed.ncbi.nlm.nih.gov/27449080/)]
8. Biggerstaff M, Johansson M, Alper D, Brooks LC, Chakraborty P, Farrow DC, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics* 2018 Sep;24:26-33 [FREE Full text] [doi: [10.1016/j.epidem.2018.02.003](https://doi.org/10.1016/j.epidem.2018.02.003)] [Medline: [29506911](https://pubmed.ncbi.nlm.nih.gov/29506911/)]
9. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Comput Biol* 2015 Aug;11(8):e1004382 [FREE Full text] [doi: [10.1371/journal.pcbi.1004382](https://doi.org/10.1371/journal.pcbi.1004382)] [Medline: [26317693](https://pubmed.ncbi.nlm.nih.gov/26317693/)]
10. Tizzoni M, Bajardi P, Poletto C, Ramasco JJ, Balcan D, Gonçalves B, et al. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC Med* 2012 Dec 13;10:165 [FREE Full text] [doi: [10.1186/1741-7015-10-165](https://doi.org/10.1186/1741-7015-10-165)] [Medline: [23237460](https://pubmed.ncbi.nlm.nih.gov/23237460/)]
11. Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biol* 2010 Feb 23;8(2):e1000316 [FREE Full text] [doi: [10.1371/journal.pbio.1000316](https://doi.org/10.1371/journal.pbio.1000316)] [Medline: [20186267](https://pubmed.ncbi.nlm.nih.gov/20186267/)]
12. Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC. Using electronic health records and internet search information for accurate influenza forecasting. *BMC Infect Dis* 2017 May 8;17(1):332 [FREE Full text] [doi: [10.1186/s12879-017-2424-7](https://doi.org/10.1186/s12879-017-2424-7)] [Medline: [28482810](https://pubmed.ncbi.nlm.nih.gov/28482810/)]
13. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci U S A* 2015 Nov 24;112(47):14473-14478 [FREE Full text] [doi: [10.1073/pnas.1515373112](https://doi.org/10.1073/pnas.1515373112)] [Medline: [26553980](https://pubmed.ncbi.nlm.nih.gov/26553980/)]
14. 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 Comput Biol* 2015 Oct;11(10):e1004513 [FREE Full text] [doi: [10.1371/journal.pcbi.1004513](https://doi.org/10.1371/journal.pcbi.1004513)] [Medline: [26513245](https://pubmed.ncbi.nlm.nih.gov/26513245/)]
15. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google flu trends? *Am J Prev Med* 2014 Sep;47(3):341-347. [doi: [10.1016/j.amepre.2014.05.020](https://doi.org/10.1016/j.amepre.2014.05.020)] [Medline: [24997572](https://pubmed.ncbi.nlm.nih.gov/24997572/)]
16. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* 2009 Feb 19;457(7232):1012-1014. [doi: [10.1038/nature07634](https://doi.org/10.1038/nature07634)] [Medline: [19020500](https://pubmed.ncbi.nlm.nih.gov/19020500/)]
17. Wójcik OP, Brownstein JS, Chunara R, Johansson MA. Public health for the people: participatory infectious disease surveillance in the digital age. *Emerg Themes Epidemiol* 2014;11:7 [FREE Full text] [doi: [10.1186/1742-7622-11-7](https://doi.org/10.1186/1742-7622-11-7)] [Medline: [24991229](https://pubmed.ncbi.nlm.nih.gov/24991229/)]
18. Leal-Neto OB, Dimech GS, Libel M, Oliveira W, Ferreira JP. Digital disease detection and participatory surveillance: overview and perspectives for Brazil. *Rev Saude Publica* 2016;50:17 [FREE Full text] [doi: [10.1590/S1518-8787.2016050006201](https://doi.org/10.1590/S1518-8787.2016050006201)] [Medline: [27191153](https://pubmed.ncbi.nlm.nih.gov/27191153/)]
19. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. *PLoS Curr* 2014 Oct 28;6:ecurrents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117 [FREE Full text] [doi: [10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117](https://doi.org/10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117)] [Medline: [25642377](https://pubmed.ncbi.nlm.nih.gov/25642377/)]
20. 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 Surveill* 2018 Jan 9;4(1):e4 [FREE Full text] [doi: [10.2196/publichealth.8950](https://doi.org/10.2196/publichealth.8950)] [Medline: [29317382](https://pubmed.ncbi.nlm.nih.gov/29317382/)]
21. Smolinski MS, Crawley AW, Baltrusaitis K, Chunara R, Olsen JM, Wójcik O, et al. Flu near you: crowdsourced symptom reporting spanning 2 influenza seasons. *Am J Public Health* 2015 Oct;105(10):2124-2130. [doi: [10.2105/AJPH.2015.302696](https://doi.org/10.2105/AJPH.2015.302696)] [Medline: [26270299](https://pubmed.ncbi.nlm.nih.gov/26270299/)]



22. 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 [FREE Full text] [doi: [10.1371/journal.pone.0083672](https://doi.org/10.1371/journal.pone.0083672)] [Medline: [24349542](https://pubmed.ncbi.nlm.nih.gov/24349542/)]
23. Centers for Disease Control and Prevention. Overview of Influenza Surveillance in the United States URL: <https://www.cdc.gov/flu/weekly/overview.htm> [accessed 2018-11-01] [WebCite Cache ID 75R8li3IP]
24. Biggerstaff M, Nhung MA, Reed C, Fry AM, Balluz L, Finelli L. Influenza-like illness, the time to seek healthcare, and influenza antiviral receipt during the 2010-2011 influenza season-United States. *J Infect Dis* 2014 Aug 15;210(4):535-544 [FREE Full text] [doi: [10.1093/infdis/jiu224](https://doi.org/10.1093/infdis/jiu224)] [Medline: [24731959](https://pubmed.ncbi.nlm.nih.gov/24731959/)]
25. American FactFinder - Census Bureau. Annual Estimates of the Resident Population for Selected Age Groups by Sex for the United States, States, Counties, and Puerto Rico Commonwealth and Municipios: April 1, 2010 to July 1, 2018 URL: <http://factfinder.census.gov/> [accessed 2018-11-01] [WebCite Cache ID 75R9KGNMr]
26. R: The R Project for Statistical Computing. 2019. URL: <https://www.R-project.org/> [accessed 2018-11-01] [WebCite Cache ID 75R9dPErN]
27. Baltrusaitis K, Santillana M, Crawley AW, Chunara R, Smolinski M, Brownstein JS. Determinants of participants' follow-up and characterization of representativeness in flu near you, a participatory disease surveillance system. *JMIR Public Health Surveill* 2017 Apr 7;3(2):e18 [FREE Full text] [doi: [10.2196/publichealth.7304](https://doi.org/10.2196/publichealth.7304)] [Medline: [28389417](https://pubmed.ncbi.nlm.nih.gov/28389417/)]
28. Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, et al. Cloud-based electronic health records for real-time, region-specific influenza surveillance. *Sci Rep* 2016 May 11;6:25732 [FREE Full text] [doi: [10.1038/srep25732](https://doi.org/10.1038/srep25732)] [Medline: [27165494](https://pubmed.ncbi.nlm.nih.gov/27165494/)]
29. 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 Infect Dis* 2018 Aug 15;18(1):403. [doi: [10.1186/s12879-018-3322-3](https://doi.org/10.1186/s12879-018-3322-3)] [Medline: [30111305](https://pubmed.ncbi.nlm.nih.gov/30111305/)]
30. Ma W, Huo X, Zhou M. The healthcare seeking rate of individuals with influenza like illness: a meta-analysis. *Infect Dis (Lond)* 2018 Oct;50(10):728-735. [doi: [10.1080/23744235.2018.1472805](https://doi.org/10.1080/23744235.2018.1472805)] [Medline: [30009680](https://pubmed.ncbi.nlm.nih.gov/30009680/)]
31. Biggerstaff M, Nhung M, Kamimoto L, Balluz L, Finelli L. Self-reported influenza-like illness and receipt of influenza antiviral drugs during the 2009 pandemic, United States, 2009-2010. *Am J Public Health* 2012 Oct;102(10):e21-e26. [doi: [10.2105/AJPH.2012.300651](https://doi.org/10.2105/AJPH.2012.300651)] [Medline: [22897525](https://pubmed.ncbi.nlm.nih.gov/22897525/)]
32. Chunara R, Goldstein E, Patterson-Lomba O, Brownstein JS. Estimating influenza attack rates in the United States using a participatory cohort. *Sci Rep* 2015 Apr 2;5:9540 [FREE Full text] [doi: [10.1038/srep09540](https://doi.org/10.1038/srep09540)] [Medline: [25835538](https://pubmed.ncbi.nlm.nih.gov/25835538/)]
33. Cantarelli P, Debin M, Turbelin C, Poletto C, Blanchon T, Falchi A, et al. The representativeness of a European multi-center network for influenza-like-illness participatory surveillance. *BMC Public Health* 2014 Sep 20;14:984 [FREE Full text] [doi: [10.1186/1471-2458-14-984](https://doi.org/10.1186/1471-2458-14-984)] [Medline: [25240865](https://pubmed.ncbi.nlm.nih.gov/25240865/)]
34. Patterson-Lomba O, van Noort S, Cowling BJ, Wallinga J, Gomes MG, Lipsitch M, et al. Utilizing syndromic surveillance data for estimating levels of influenza circulation. *Am J Epidemiol* 2014 Jun 1;179(11):1394-1401 [FREE Full text] [doi: [10.1093/aje/kwu061](https://doi.org/10.1093/aje/kwu061)] [Medline: [24748609](https://pubmed.ncbi.nlm.nih.gov/24748609/)]

## Abbreviations

- CDC:** Centers for Disease Control and Prevention
- EHR:** electronic health record
- FNY:** Flu Near You
- ILI:** influenza-like illness
- ILINet:** Influenza-like Illness Surveillance Network
- IQR:** interquartile range
- SIR:** susceptible-infected-recovered

*Edited by G Eysenbach; submitted 16.01.19; peer-reviewed by M Biggerstaff, O Leal Neto; comments to author 28.04.19; revised version received 02.07.19; accepted 19.07.19; published 14.09.19*

### *Please cite as:*

Baltrusaitis K, Vespignani A, Rosenfeld R, Gray J, Raymond D, Santillana M  
*Differences in Regional Patterns of Influenza Activity Across Surveillance Systems in the United States: Comparative Evaluation*  
*JMIR Public Health Surveill* 2019;5(4):e13403  
URL: <https://publichealth.jmir.org/2019/4/e13403>  
doi: [10.2196/13403](https://doi.org/10.2196/13403)  
PMID: [31579019](https://pubmed.ncbi.nlm.nih.gov/31579019/)



©Kristin Baltrusaitis, Alessandro Vespignani, Roni Rosenfeld, Josh Gray, Dorrie Raymond, Mauricio Santillana. Originally published in JMIR Public Health and Surveillance (<http://publichealth.jmir.org>), 14.09.2019. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Public Health and Surveillance, is properly cited. The complete bibliographic information, a link to the original publication on <http://publichealth.jmir.org>, as well as this copyright and license information must be included.