

# Discovering multi-scale co-occurrence patterns of asthma and influenza with oak ridge bio-surveillance toolkit

Arvind Ramanathan<sup>1,2\*</sup>, Laura L. Pullum<sup>1,2\*</sup>, Tanner C. Hobson<sup>1</sup>, Christopher G. Stahl<sup>1</sup>, Chad A. Steed<sup>1</sup>, Shannon P. Quinn<sup>3</sup>, Chakra S. Chennubhotla<sup>3</sup> and Silvia Valkova<sup>4</sup>

<sup>1</sup> Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA, <sup>2</sup> Health Data Sciences Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA, <sup>3</sup> Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA, <sup>4</sup> IMS Health Government Solutions, Plymouth Meeting, PA, USA

## OPEN ACCESS

### Edited by:

Alina Deshpande,  
Los Alamos National Laboratory, USA

### Reviewed by:

Li Xu,  
Cornell University, USA  
Sara Y. Del Valle,  
Los Alamos National Laboratory, USA

### \*Correspondence:

Arvind Ramanathan and  
Laura L. Pullum,  
Computational Science and  
Engineering Division, Oak Ridge  
National Laboratory, Health Data  
Sciences Institute, P.O. Box 2008,  
MS-6085, Oak Ridge,  
TN 37831-6085, USA  
ramanathana@ornl.gov;  
pulluml@ornl.gov

### Specialty section:

This article was submitted to  
Infectious Diseases, a section of the  
journal *Frontiers in Public Health*

**Received:** 18 January 2015

**Accepted:** 10 July 2015

**Published:** 03 August 2015

### Citation:

Ramanathan A, Pullum LL,  
Hobson TC, Stahl CG, Steed CA,  
Quinn SP, Chennubhotla CS and  
Valkova S (2015) Discovering  
multi-scale co-occurrence patterns of  
asthma and influenza with oak ridge  
bio-surveillance toolkit.  
*Front. Public Health* 3:182.  
doi: 10.3389/fpubh.2015.00182

We describe a data-driven unsupervised machine learning approach to extract geo-temporal co-occurrence patterns of asthma and the flu from large-scale electronic healthcare reimbursement claims (eHRC) datasets. Specifically, we examine the eHRC data from 2009 to 2010 pandemic H1N1 influenza season and analyze whether different geographic regions within the United States (US) showed an increase in co-occurrence patterns of the flu and asthma. Our analyses reveal that the temporal patterns extracted from the eHRC data show a distinct lag time between the peak incidence of the asthma and the flu. While the increased occurrence of asthma contributed to increased flu incidence during the pandemic, this co-occurrence is predominant for female patients. The geo-temporal patterns reveal that the co-occurrence of the flu and asthma are typically concentrated within the south-east US. Further, in agreement with previous studies, large urban areas (such as New York, Miami, and Los Angeles) exhibit co-occurrence patterns that suggest a peak incidence of asthma and flu significantly early in the spring and winter seasons. Together, our data-analytic approach, integrated within the Oak Ridge Bio-surveillance Toolkit platform, demonstrates how eHRC data can provide novel insights into co-occurring disease patterns.

**Keywords:** disease co-occurrence, non-negative matrix factorization, public health surveillance, asthma, flu, electronic healthcare reimbursement claims

## Introduction

The digitization of health records has spurred the systematic collection and archival of massive datasets, both within hospital and government computer systems (1–6). Therefore, digital public health surveillance is emerging as an important tool for tracking, monitoring, and driving decisions regarding emerging infectious disease spread within geographically distributed populations (7). Many bio-surveillance systems rely on the use of event-based, unstructured digital data, such as news feed aggregators, internet search patterns of users, and social media (7). However, with the availability of electronic health records (EHR) and electronic healthcare reimbursement claims (eHRC), there is a tremendous opportunity to seek, collect, monitor, and analyze these large-scale datasets for public health surveillance. While EHRs capture a patient's full medical history, eHRCs capture only the healthcare reimbursements processed by insurance companies. In particular,

eHRCs serve as a data warehouse that include claim transactions processed: (a) when patients visit their providers' (e.g., doctor's/nurse practitioner's) office and/or (b) when retail pharmacies dispense prescription drugs to patients.

In this paper, we present a novel data-driven approach to extract co-occurring patterns of influenza-like illnesses (ILI) and asthma using eHRC datasets. While both conditions represent significant respiratory ailments, ILI occurs seasonally and asthma is a chronic condition that can result in wheezing, breathlessness, and cough. With an increasing number of patients being diagnosed with asthma since 2000 (8–10), we wanted to explore the relationship between the occurrence of asthma with the flu during the 2009–2010 pandemic flu season. In particular, the risks associated with influenza in young children susceptible to asthma have been well studied (11–14). However, the co-occurrence of flu and asthma in adults is less understood (15, 16). Therefore, we hypothesized that during the 2009–2010 H1N1 pandemic flu season, people who were more susceptible to asthma were likely to also be affected with the flu. Further, we hypothesized that the continued incidence of asthma within specific geographic regions in the US would predict which regions could be significantly affected by the flu.

To evaluate these hypotheses, we describe a novel unsupervised, machine learning approach to automatically identify spatial and temporal patterns from large-scale eHRC datasets for the 2009–2010 influenza and study its inter-relationship with asthma incidence during the same time period. Apart from discovering a small number of distinct geo-temporal patterns, our analysis shows a distinct lag in the temporal patterns of asthma and flu, i.e., we find that a peak in the number of diagnosed flu cases followed a peak in the number of diagnosed asthma cases. Our results further include an exploratory analysis into the demographic features of why such a peak may have been observed. In particular, we observe a behavior of the epidemic within large urban areas where environmental factors may have a significant impact (in addition to other factors) in influencing the total number of patients with flu and asthma.

## Materials and Methods

### Data

Prior to our study, we obtained internal Institutional Review Board approval for analyzing the IMS Health datasets. In this study, we analyzed the IMS Health ambulatory care reimbursement claims data from the 2009–2010 pandemic (H1N1) flu season; the details of the datasets are provided in our previous paper (17). Note that the data from IMS Health are already processed by a third party to remove any form of personally identifiable information before it receives the claims data from its suppliers. The study included eHRC from April 1, 2009 to March 31, 2010 with a total of nearly one billion records. We processed the ambulatory care reimbursement claims data and parsed out influenza (ICD9 codes 486XX and 488XX) and asthma (ICD9 codes 493XX) related records. We specifically chose those ICD9 codes that corresponded to hospital diagnosed cases of the flu. For flu, we obtained a total of over six million individual records (throughout

the US). For asthma, we obtained a total of over 10 million individual records. We used the zip code corresponding to the patient's service provider (i.e., a medical practitioner/physician), since the provider's five-digit zip code is more specific than the patient's three digit zip code directly accessible from the data. Only 0.0001% of the total records had different three-digit zip codes available for the patient and service provider.

The resulting flu and asthma datasets were stored as matrices,  $\mathbf{A}_f$  and  $\mathbf{A}_a$ , respectively, where the rows represent the number of days and the columns represent the total number of zip codes. Note that the datasets are proprietary to IMS Health and therefore we cannot freely share the derived datasets used in this analysis. In order to characterize the co-occurrence of asthma and flu, we obtained a list of common zip codes between  $\mathbf{A}_f$  and  $\mathbf{A}_a$  and considered only those zip codes that had more than 10 reported cases of either diagnostic code set. The IMS diagnostic dataset covered 14,098 zip codes with statistically significant data for both flu and asthma, covering about 47% of the US.

### Identifying Geo-Temporal Patterns Using Non-Negative Matrix Factorization

Our primary hypothesis from the flu and asthma incidence patterns was to observe if the people susceptible to asthma were more likely to be infected with the flu, during the 2009–2010 H1N1 pandemic flu. We also wanted to understand if the continued presence of asthma within specific geographic regions would be predictive of the flu incidence in that area. To answer these questions, we used non-negative matrix factorization (NMF) to extract a small set of spatial and temporal patterns from the flu and asthma eHRCs. As we have shown in our previous paper (17), we chose NMF as an unsupervised machine learning technique to analyze the data primarily based on several empirical observations about the data. First, the data matrix  $\mathbf{A}$  consists of only non-negative entries – because the total number of patients at any given zip code will be  $\geq 0$ . Further, we observed that the individual zip codes exhibit a small number of distinct patterns in the occurrence of the flu (17), suggesting that in spite of a high dimensional setting of the  $\mathbf{A}_{\text{asthma}}$  and  $\mathbf{A}_{\text{flu}}$  matrices (with more than 14,000 individual zip codes and 365 days), there might be only a small number of geo-temporal patterns that could best capture the co-occurrence of these two conditions. Second, while several other types of analyses are possible to examine the data, our choice of analysis was motivated by the need to discover the underlying geo-temporal patterns in an unsupervised manner. Techniques, such as principal component analysis, which represents one of the most widely used unsupervised analysis technique, pursue variance blindly (18, 19) may fail to capture the intrinsic orientations in the high dimensional data space.

Given a data matrix  $\mathbf{A}$ , with non-negative entries,  $N_z \times N_t$  dimensions where  $N_z$  represents the number of zip codes and  $N_t$  represents time (in days), NMF finds low-rank approximations in  $s$  dimensions of the form  $\mathbf{A} \approx \mathbf{WH}$ , where  $\mathbf{W}$  with  $N_z \times s$  dimensions represents spatial patterns and  $\mathbf{H}$  with  $s \times N_t$  captures temporal patterns within the data matrix. We used the alternate least squares algorithm proposed by Paatero (20, 21), available

as part of the Matlab package. We ran NMF for a total of 1,000 iterations. To find the appropriate low-rank approximation ( $s$ ), we varied  $s = 1, \dots, 15$ , dividing the original data into training and testing data (50% training and 50% testing). Tracking the residual errors using the Frobenius norm for both training and testing data, we performed a total of 250 iterations. In our analysis, as shown in previous work (17, 22), we identified  $s$  to be 5. Once we chose  $s$ , the most stable version of the basis matrices ( $\mathbf{W}$ ,  $\mathbf{H}$ ) by computing the Kullback–Leibler (KL) divergence between every pair of the 250 instances of  $\mathbf{W}$  (or  $\mathbf{H}$ ) from the training set and picking  $\mathbf{W}$  (or  $\mathbf{H}$ ) with the lowest KL divergence value.

## Results

### Flu and Asthma Case-Counts During the 2009–2010 H1N1 Pandemic Season

We summarize the flu and asthma from the ambulatory care reimbursement claims data as a function of daily incidence. While influenza rates rise sharply during the August–September 2009 time-frame, the number of asthma cases observed from the data shows more or less a uniform distribution throughout the year, except for a slight increase and decrease around the time of the pandemic flu season. Further, we find that the peak number of asthma incidences lags behind by about 3 weeks when compared to the peak number of influenza incidences (Figure 1). Except for the beginning of the winter season in Figure 1, highlighted as  $A_3$  and  $F_3$ , where the asthma and flu incidence rates coincide, in the other two cases, highlighted by  $A_{1,2}$  and  $F_{1,2}$ , the peak incidence of asthma occurs earlier than the peak incidence of the flu. Note that for Figure 1, we present the data that were temporally averaged by 7 days (to account for lag times within diagnostic data reporting within the IMS Health datasets). We note that even without the

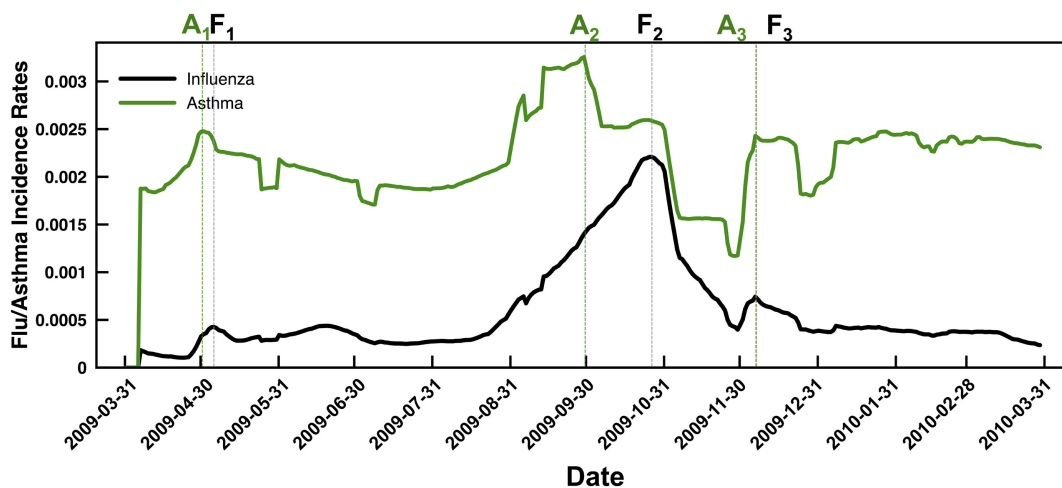
temporal averaging, these trends are observed (both at state and national levels).

Once we examined the temporal trends in the flu and asthma datasets, we then extended the analysis to examine the demographic data for the total number of cases observed (both for H1N1 flu and asthma). Note that the IMS Health data include only information regarding the age of the patient and their sex, but do not include any other demographic information. As summarized in Table 1, we find that a larger proportion of children show increased co-occurrence of flu and asthma symptoms. Although the number of adults diagnosed with asthma is higher, only a small proportion of patients are co-diagnosed with both flu and asthma during the 2009–2010 flu season. Interestingly, in our analysis of the data, girls tend to exhibit a higher risk (over 90% of girls are susceptible to both flu and asthma in Table 1). Similarly, within adults co-diagnosed with the flu and asthma, female patients tended to be higher in ratio compared to men. Table 1 also indicates that the total number of diagnosed cases with both flu and asthma conditions is very infrequent and indicate that children seem to be at a greater risk than adults.

An interesting question that arises from the above analysis is whether there are specific geographic regions within the US (or time windows), which exhibit a concurrent occurrence of the flu and asthma. We present an approach to discover such co-occurring patterns in the next section.

### Temporal Patterns in Flu and Asthma Incidence Identifying Optimal Subspace and Cross Validation

The dimensionality of the data for each of the matrices ( $A_{f,a}$ ) is  $N_z \times N_t$  where  $N_z$  represents the total number of zip codes (14,098) and  $N_t$  represents the time points (365 days starting



**FIGURE 1 | Summary of temporal trends observed from the flu (black line) and asthma (green line) case counts indicate a distinct delay in the peak incidence of flu compared to asthma.** Note that we have reported the data using a moving average window of 7 days (to overcome gaps in the IMS ambulatory care reimbursement claims data based on reports received throughout the week) and normalized the results based on the fraction of total case counts received at every zip code.

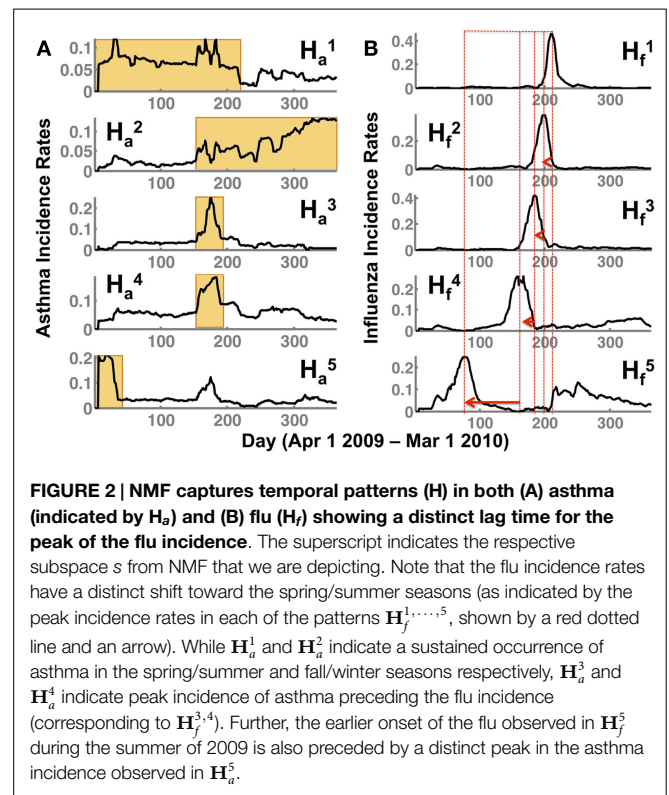
Dotted lines are used to the respective peak incidence rates of asthma (green) and flu (black). Note that in both the spring season (April–May 2009) and the fall season (September–October 2009), the asthma incidence (indicated by  $A_1$  and  $A_2$ , respectively) peaks before the peak in flu incidence (indicated by  $F_1$  and  $F_2$ , respectively). Only for the winter season, the peaks in asthma and flu incidence rates coincide (highlighted by  $A_3$  and  $F_3$  respectively).

**TABLE 1 | Demographic summary of H1N1 and asthma case count summary observed from eHRC data.**

Child attributes	Flu	Asthma	Flu and asthma
Mean age	7	7	
<1 year	109586	171117	6972
1–2 years	279806	663782	27206
3–5 years	466094	1109834	51155
>5 years	1241575	2870801	118951
Total	2097061	4815534	204284
No. (%) of girls	48	47	92
Adult attributes	Flu case counts	Asthma case counts	Flu and asthma
Mean age	42	51	
18–24 years	226075	616275	12543
25–30 years	181821	536832	8962
31–35 years	138167	530525	7565
36–40 years	144418	648801	8543
41–45 years	133488	741015	8762
46–50 years	131245	856720	9093
>50 years	393298	4080939	25450
Total	1348512	8011107	80918
No. (%) of females	60	69	70

from April 1 2009 to March 31 2010). The subscripts  $f$  and  $a$  for the **A** matrix represent the two conditions examined, namely the flu and asthma, respectively. We hypothesized that the flu incidence patterns would be composed of discrete spatial and temporal patterns, given the geographic size and spread of the US. Further, given prior knowledge that there were distinct “peaks” associated with the 2009–2010 pandemic, it is reasonable to use techniques that could elucidate discrete, yet sparse spatial and temporal patterns from this high dimensional data. Additionally, the entries within each of these matrices are non-negative (i.e., it is not possible to obtain a negative count for the number of patients with the flu or asthma). For this purpose, we used non-negative matrix factorization (NMF), a technique that can extract low-rank approximations from the data.

Given a data matrix **A** with non-negative entries ( $N_z \times N_t$  dimensions), NMF finds low-rank approximations of the form  $A \approx WH$ , where **W** ( $N_z \times s$ ) captures spatial patterns and **H** ( $s \times N_t$ ) represents temporal patterns within the data. We used the alternate least squares algorithm proposed by Paatero (20, 21), available as part of standard Matlab (Mathworks, Inc.). Although the size of  $A_{a,f}$  are quite large, we did not find the speed of convergence a significant problem. We used a stopping value of 1,000 as the maximum number of iterations. To identify the appropriate subspace ( $s$ ) dimensions for the original data, we iterated over  $s = 1 \dots 15$  for both matrices, dividing the data into random yet equal-sized training and testing data. We tracked the residual errors using Frobenius norm for both training and testing data. For each choice of  $s$ , we repeated this process 100 times. Using this procedure, we chose the optimal  $s = 5$ , based on the most stable version of the basis matrices by computing the Kullback–Leibler (KL) divergence between every pair of the 100 instances of **W** from the training dataset and picking the **W** with the lowest KL divergence value.



**FIGURE 2 | NMF captures temporal patterns (H) in both (A) asthma (indicated by  $H_a$ ) and (B) flu ( $H_f$ ) showing a distinct lag time for the peak of the flu incidence.** The superscript indicates the respective subspace  $s$  from NMF that we are depicting. Note that the flu incidence rates have a distinct shift toward the spring/summer seasons (as indicated by the peak incidence rates in each of the patterns  $H_f^1 \dots H_f^5$ , shown by a red dotted line and an arrow). While  $H_a^1$  and  $H_a^2$  indicate a sustained occurrence of asthma in the spring/summer and fall/winter seasons respectively,  $H_a^3$  and  $H_a^4$  indicate peak incidence of asthma preceding the flu incidence (corresponding to  $H_f^3, H_f^4$ ). Further, the earlier onset of the flu observed in  $H_f^5$  during the summer of 2009 is also preceded by a distinct peak in the asthma incidence observed in  $H_a^5$ .

### Distinct Break-out Patterns Govern Flu and Asthma Incidence

NMF offers a convenient framework to interpret the incidence of flu and asthma throughout the US during the 2009–2010 time period. In particular, it provides a small number of basis vectors that describe temporal ( $H_{f,a}$ ) and spatial ( $W_{f,a}$ ) break-out patterns. Note that the subscripts used,  $f$  and  $a$ , correspond to the two conditions tracked, the flu and asthma, respectively. Based on the procedure outlined above, we selected the optimal subspace to be  $s = 5$  as it sufficiently captured the underlying spatial and temporal patterns in the data, while providing an intuitive description how flu and asthma co-occurred at any given time period (or spatial location). As shown in **Figure 2**, one of the notable observations is that the temporal signatures are distinct in capturing the occurrence of flu and asthma in the 2009–2010 season.

The temporal patterns for asthma incidence,  $H_a^1$  and  $H_a^2$ , show a seasonal rise in summer and winter season, respectively. In particular, the orange rectangles highlight the rise and sustained occurrence of asthma cases for the respective seasons. Interestingly, there is a very short overlap (of about 30 days from September to October) between the summer and winter where the asthma occurrence from one season overlaps with the other. This intersecting time period is captured as increased incidence rates in both  $H_a^3$  and  $H_a^4$ . Additionally, in basis vector,  $H_a^5$ , we observe a high incidence of asthma around days 10–45 (April–May 2009) time-frame.

The break-out patterns for the flu across the US indicates at least three distinct peaks, ranging from days 180 to 210 (September–October 2009), 150 to 180 (August–September 2009), and



90–100 (June–July 2009). The temporal patterns from the flu data indicate that there is a distinct early onset of the epidemic ( $H_f^5$ ), followed by several waves at later time-periods ( $H_f^{1-4}$ ), which all have their own distinct temporal signatures. Thus, each of the basis vectors (in the flu dataset) captures a unique temporal break-out pattern that captures a different phase of the 2009–2010 flu epidemic, similar to previously reported studies in the spread of influenza (23).

Comparing the flu and asthma break-out patterns suggests that there is an overlap in the incidence of flu and asthma around August–September, described by  $H_a^{3,5}$  and  $H_f^{3,4}$ , respectively. Further, comparing  $H_a^5$  and  $H_f^5$  also indicates that even during the early onset of the flu (days 90–100; June–July 2009), there is a marked increase in the asthma incidence rates around days 10–45 (April–May 2009). Although from **Figure 1**, we see that the overall trend indicates that the peak of asthma incidence precedes the peak of flu season, the analysis presented here further suggests that this precedence may be a distinct factor influencing the susceptibility of flu occurrence within some regions.

### Geographic Patterns of Flu and Asthma Incidence

The spatial patterns summarized by NMF depict a distinct separation between the asthma and flu incidence. As shown in **Figure 3**, each  $W$  can be mapped onto the specific zip code and provides a geographic interpretation of the results presented above. Each dot represents a specific zip code examined, and the intensity of the color indicates a higher occurrence of the flu/asthma (blue indicates lower and red indicates higher incidence). Note that both the asthma and flu incidence maps are drawn to the same color scale (as indicated by the color bar in **Figure 3**).

We note that densely populated areas (such as New York, Florida, and California) constitute common grounds for the co-occurrence of the flu and asthma. In particular, throughout the north-east, southeast, west, and central US, asthma patterns are widespread. The spatial patterns for influenza across the entire US are, however, discrete. Several of the north-east states do not exhibit any patterns observed in  $W_f^{2-4}$ , meaning that during the latter half of the year, there were no many cases of the flu reported (except in urban areas). Within the urban areas of the north-east, however, the occurrence of asthma is quite widespread and occurs throughout the year.

The occurrence of  $W_f^4$  is almost exclusive in the southern regions, with cases detected in both southeast and southwest (California). The temporal patterns from the south-east constitute the time-frame of August–September 2009, which signified the beginning of school season within the same region, leading to the unique spatial patterns observed here. Within the north-east (specifically in New York), the H1N1 pandemic was detected early (in April 2009) and a suitable warning was also issued (24), perhaps leading to a low number of observed flu cases during the peak time (September–October 2009) of the flu in these regions. The other interesting aspect observed from our analysis is the early onset of the flu in some north-eastern states (notably New York and New Jersey) as well as southwest (California), is captured by  $W_f^5$ , indicating that this early onset also meant a sustained flu

in the later part of the season (around February–March 2010) in these regions (**Figure 2B**,  $H_f^5$ ).

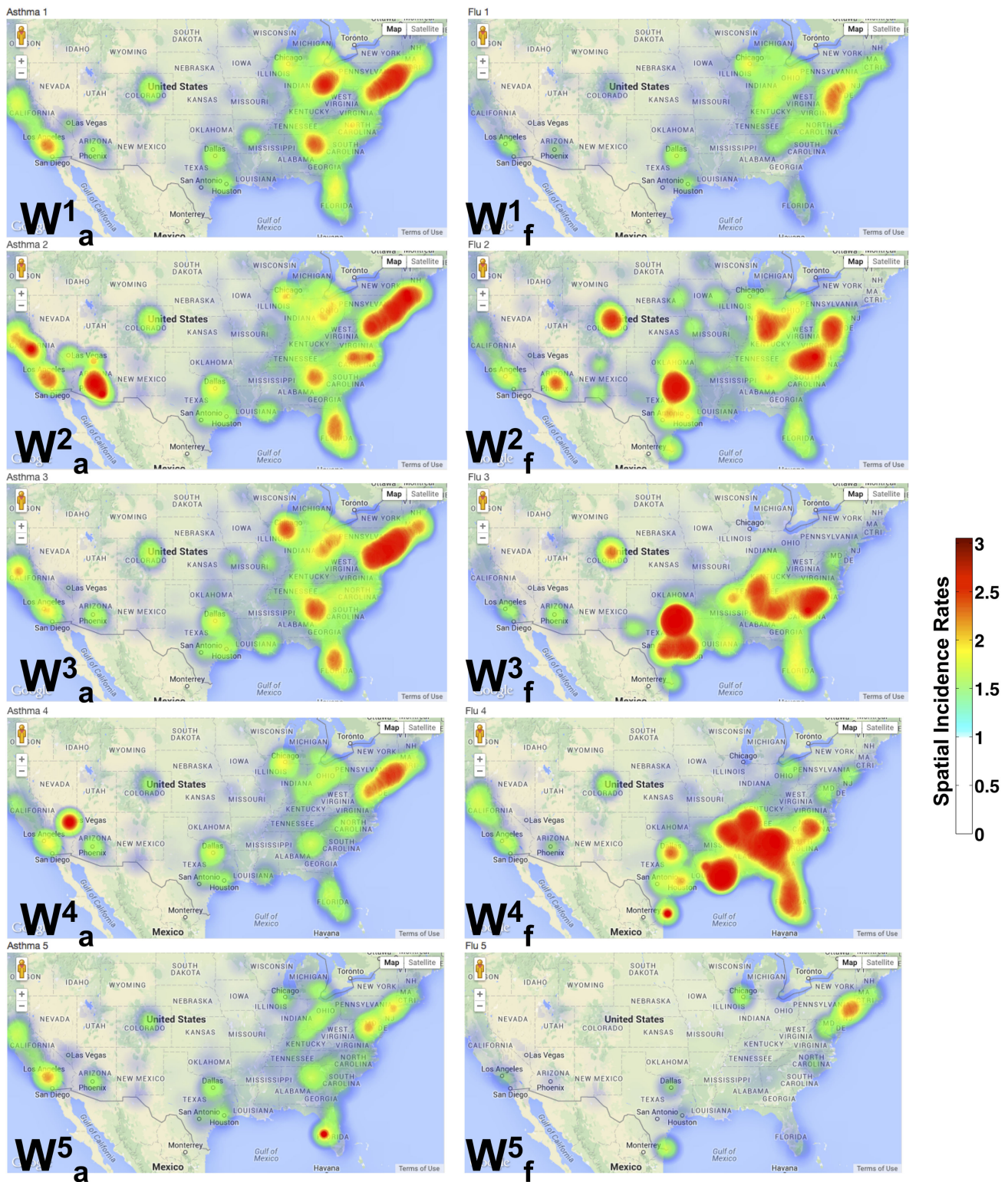
We also examined the zip codes where flu and asthma patterns occur concurrently. These regions include urban areas within the north-east (specifically, southeastern New York, New Jersey, Delaware, southern parts of New Hampshire, Connecticut, and Pennsylvania), southeast (Tennessee, Georgia, North and South Carolinas, Florida, and south central parts of Virginia), and the west-coast area (California, Oregon, and Washington states). These urban areas constitute a majority of the places where the co-occurrence of the flu and asthma exhibit a clear trend, i.e., a peak in asthma diagnoses is subsequently followed by a peak in the flu diagnoses. This is in agreement with previous studies that have showed that air quality, local weather, and pollen fluctuations, as well as presence of environmental pollutants within urban areas can significantly impact patients with asthma (25).

## Discussion and Conclusion

### Comparison with Previous Work

The use of EHR and eHRC datasets for bio-surveillance is relatively new (26). Privacy and security concerns within EHR and eHRC systems have made it tremendously challenging to engage local and public health departments in effectively collecting, sharing, and disseminating bio-surveillance-related data (27). Although eHRC transaction datasets have been routinely used in the context of tracking and analyzing pharmacy prescriptions and understanding drug efficacy [e.g., Ref. (28–32)], very little research has been carried out in terms of using them as potential data sources for digital public health surveillance. eHRCs have a distinct lag time associated with claims processing and being made available for analysis. Therefore, the timeliness of their availability from the claims processor may have a significant impact on assessing the data real-time, i.e., as epidemics are spreading through the population. A recent study showed that retail pharmacy sales data can be used as a reliable measure for syndromic surveillance; specifically, the aggregate counts of prescription sales of four antiviral drugs for influenza correlated well with Google Flu Trends (33, 34). However, given the concerns with Google Flu (35), there is a need to develop alternate strategies to evaluate eHRCs in tracking flu (and other diseases). It should be noted here that these papers make use of standard time-series algorithms and/or other signal processing techniques to model the temporal trends and report correlations with existing and available CDC ILINet datasets.

Of the many approaches used to analyze bio-surveillance-related datasets, supervised and unsupervised machine learning techniques have been made use of in classifying text messages from various social media sources (such as Twitter) (36, 37). In addition, these techniques are used to rank search results of various bio-surveillance terms (from either a pool of Twitter documents or other internet-based surveillance sources) (38) to aid analysts in identifying the most relevant documents for decision making. However, these techniques have not been used in the context to identify co-occurring disease patterns for bio-surveillance.



**FIGURE 3 | Spatial patterns from NMF indicate distinct pockets of urban areas showing co-occurrence of flu and asthma.** A geographic incidence map of the flu ( $W_f^1, \dots, W_f^5$ ) and asthma ( $W_a^1, \dots, W_a^5$ ) shows the common areas of co-occurrence as described in our analysis of the diagnostic data. The spatial incidence is summarized in increasing color intensity shown on the color map. It

is interesting to observe that the flu incidence gradually progresses to the south-east from  $W_f^1$  through  $W_f^4$ . Further,  $W_f^5$  almost exclusively describes the occurrence of the flu in only large urban areas of the country. Similar patterns are also observed in the asthma incidence with local incidences being concentrated around urban areas.



In our previous work, we showed how diagnostic eHRC transactions are comparable to standard public health surveillance data, such as the CDC ILINet (17). Further, we also showed that the consolidated eHRCs at local (zip code level information), regional (county, metropolitan, city, state, etc.), and national levels can be used to assess how infectious diseases like the flu may spread. Unlike aggregating web-based search patterns by users (33, 35), or the use of social media (39–42), where the use of such data can significantly overestimate the flu incidence patterns (43, 44), using eHRCs provides a more accurate indication and potential predictors. To our knowledge, this study is perhaps the first to use eHRCs for identifying co-occurrence patterns of flu and asthma at the national scale.

Another body of literature examines how asthma and the flu represent syndemic conditions, meaning that both afflictions are linked and interact synergistically contributing to an excess burden of disease (14, 25, 45, 46). In this context, the mechanisms by which influenza can exacerbate asthma in patients have been well documented (47, 48). Studies have examined clinical strategies to vaccinate patients (both children and adults) against influenza that have asthma so that adverse reactions can be prevented (11, 15, 49–52) and also evaluated the general safety of the influenza vaccines and other treatments (such as anti-viral drugs) for patients with asthma (53, 54). Although our study did not examine whether the flu and asthma are syndemic within particular patient populations, we showed that the girls and, in general, female patients were more susceptible to be co-diagnosed with the flu and asthma during the 2009 H1N1 pandemic season. We believe that further analysis would be necessary, including the use of prescription eHRC datasets to glean whether treatments, such as anti-viral medicines or vaccinations for these susceptible patient sub-populations, were effective in controlling the pandemic spread.

From the analysis of the diagnostic data, we showed that it is possible to summarize the spatial and temporal patterns from these two conditions into a small number of categorical dimensions, each showing a distinct (temporal and spatial) signature with respect to the occurrence of asthma and flu. By examining the demographics of flu and asthma occurrence in both children and adult populations, we observed that a major proportion of girls and women were more susceptible to their co-occurrence. While it is widely acknowledged that older women are more susceptible to asthma in later ages (25), the co-occurrence of asthma and flu within younger females correlates well on the statistics in recent years, showing a higher percentage of girls affected with asthma attacks (10). Further analysis into the nature of incidence and reports would be needed, and we propose to examine this as part of future publications.

## Perspective and Potential Limitations

The analysis of the spatial patterns for flu and asthma revealed that there are distinct geographic locations (albeit a very small number of them, about 4,000/14,000 zip codes) that show more than one temporal signatures in the flu/asthma incidence patterns. Further analysis of these regions will be necessary to understand the origins of such “mixing.” In particular, as part of our analysis, we did not examine patient age or history to understand how

a specific group of patients (or a demographic) may be more susceptible to asthma or the flu. Patients with one or more pre-existing respiratory conditions can be more susceptible to either flu or asthma and hence these factors would have to be taken into account to further understand the co-occurrence patterns observed during the 2009–2010 flu season. At the time of writing this paper, this information was not available.

We note here that a more detailed analysis of the spatio-temporal patterns is required. In particular, for this paper, we have not quantitatively examined how these temporal patterns match up against other known temporal mining algorithms and even other unsupervised machine learning techniques, such as principal component analysis. We also note that the predictive aspects of our algorithm have also not been fully explored for two reasons: (1) the data available to us are only from the 2009–2010 flu season and (2) it is difficult to obtain a baseline behavior based on a year that showed highly anomalous behavior in terms of the overall flu incidence across the entire country. We will explore these questions in greater detail in a following publication. Another potential limitation of our study is that we chose to aggregate our data based on individual zip codes. While the use of other aggregation techniques (e.g., HHS regions, or state-level) are more appropriate for epidemiological purposes, our goal within this study was to demonstrate how we can extract constituent patterns of asthma/flu incidence and observe correlated behaviors at this spatial resolution. We propose to examine standard approaches of epidemiological data aggregation in further studies.

The analytic techniques outlined here are part of the data analytic platform for public health surveillance that we have been developing (22). The platform was designed specifically to bring together heterogeneous datasets, such as social media and eHRCs, and analyze these datasets to gather insights into emerging public health concerns. In this study, we used asthma and influenza as specific examples to understand co-occurrence patterns across the US. However, the techniques are quite general and can be integrated with visual analytic tools to summarize, navigate, and interpret large volumes of complex healthcare datasets. We believe that the availability of unique datasets and data analyses techniques outlined above can lead to better public health surveillance systems and have a positive impact on the nation's health.

## Source Code and Availability

The Oak Ridge Bio-surveillance Toolkit (ORBiT) (17, 22) and the tools implemented as part of this paper will be made available as an open source Python-based package from our website (<http://cda.ornl.gov>). Data (used as part of this paper) can be requested through IMS Health Institute.

## Author Contributions

AR and LP conceived and designed the study. SV provided the data. TH and LP processed and stored the eHRC datasets for analysis. AR, LP, SQ, and CC developed the analysis techniques. TH, CAS, and CGS developed the user interface components for visualizing the results. AR, CC, SQ, and TH analyzed the data. AR, LP, TH, CGS, CAS, SQ, SV, and CC wrote the paper.

## Acknowledgments

**Funding:** Preparation of this paper was funded by ORNL internal SEED project number 7280, “Demonstrating a Novel Bio-Defense Capability using Public Health Data Informatics.” This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher,

by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

## References

- Jha AK, Ferris TG, Donelan K, DesRoches C, Shields A, Rosenbaum S, et al. How common are electronic health records in the United States? A summary of the evidence. *Health Aff* (2006) **25**:w496–507. doi:10.1377/hlthaff.25.w496
- Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, et al. Use of electronic health records in U.S. hospitals. *N Engl J Med* (2009) **360**:1628–38. doi:10.1056/NEJMsa0900592
- Ambinder EP. Electronic health records. *J Oncol Pract* (2005) **1**:57–63. doi:10.1200/JOP.1.2.57
- DesRoches CM, Campbell EG, Rao SR, Donelan K, Ferris TG, Jha A, et al. Electronic health records in ambulatory care – a national survey of physicians. *N Engl J Med* (2008) **359**:50–60. doi:10.1056/NEJMsa0802005
- Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* (2008) **77**:291–304. doi:10.1016/j.ijmedinf.2007.09.001
- Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. *Med Care Res Rev* (2009) **66**:611–38. doi:10.1177/1077558709332440
- Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection – harnessing the web for public health surveillance. *N Engl J Med* (2009) **360**:2153–7. doi:10.1056/NEJMp0900702
- Moorman J, Rudd R, Johnson C, King M, Minor P, Bailey C, et al. National surveillance for asthma – United States, 1980–2004. *MMWR Surveill Summ* (2007) **56**:1–54.
- Moorman J, Zahran H, Truman B, Molla M. Current asthma prevalence – United States, 2006–2008. *MMWR Surveill Summ* (2011) **60**:84–6.
- Moorman J, Person CJ, Zahran H. Asthma attacks among persons with current asthma – United States, 2001–2010. *MMWR Surveill Summ* (2013) **62**:93–8.
- Kramarz P, DeStefano F, Gargiullo PM, Davis RL, Chen RT, Mullooly JP, et al. Influenza vaccination in children with asthma in health maintenance organizations. *Vaccine* (2000) **18**:2288–94. doi:10.1016/S0264-410X(99)00551-4
- Chung EK, Casey R, Pinto-Martin JA, Pawlowski NA, Bell LM. Routine and influenza vaccination rates in children with asthma. *Ann Allergy Asthma Immunol* (1998) **80**:318–22. doi:10.1016/S1081-1206(10)62976-6
- Bhat N, Wright JG, Broder KR, Murray EL, Greenberg ME, Glover MJ, et al. Influenza-associated deaths among children in the United States, 2003–2004. *N Engl J Med* (2005) **353**:2559–67. doi:10.1056/NEJMoa051721
- Tuffaha A, Gern JE, Lemanske RF Jr. The role of respiratory viruses in acute and chronic asthma. *Clin Chest Med* (2000) **21**:289–300. doi:10.1016/S0272-5231(05)70267-7
- Lu PJ, Euler GL, Callahan DB. Influenza vaccination among adults with asthma. *Am J Prev Med* (2009) **37**:109–15. doi:10.1016/j.amepre.2009.03.021
- Moore MML, Parker MAL. Influenza vaccine compliance among pediatric asthma patients: what is the better method of notification? *Pediatr Asthma Allergy Immunol* (2006) **19**:200–4. doi:10.1089/pai.2006.19.200
- Ramanathan A, Pullum L, Hobson T, Steed C, Quinn S, Chennubhotla C. Orbit: oak ridge bio-surveillance toolkit for public health dynamics. *BMC Bioinformatics* (2015).
- Ramanathan A, Savol AJ, Langmead CJ, Agarwal PK, Chennubhotla CS. Discovering conformational sub-states relevant to protein function. *PLoS One* (2011) **6**:e15827. doi:10.1371/journal.pone.0015827
- Castro JB, Ramanathan A, Chennubhotla CS. Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS One* (2013) **8**:e73289. doi:10.1371/journal.pone.0073289
- Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* (1994) **5**:111–26. doi:10.1002/env.3170050203
- Paatero P. Least squares formulation of robust non-negative factor analysis. *Chemometr Intell Lab Syst* (1997) **37**:23–35. doi:10.1016/S0169-7439(96)00044-5
- Ramanathan A, Pullum L, Steed C, Quinn S, Chennubhotla C. Oak ridge bio-surveillance toolkit. *IEEE VAST Workshop on Public Health's Wicked Problems: Can InfoVis Save Lives?* (2013). Available from: [www.cc.gatech.edu/gvu/iil/PublicHealthVis/Papers/OrBIT\\_PHDecisionMaking\\_F.pdf](http://www.cc.gatech.edu/gvu/iil/PublicHealthVis/Papers/OrBIT_PHDecisionMaking_F.pdf)
- Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* (2006) **312**:447–51. doi:10.1126/science.1125237
- Centers for Disease Control. Swine-origin influenza A (H1N1) virus infections in a school – New York city, April 2009. *MMWR Morb Mortal Wkly Rep* (2009) **58**(17):470–2.
- Yeatts K, Sly P, Shore S, Weiss S, Martinez F, Geller A, et al. A brief targeted review of susceptibility factors, environmental exposures, asthma incidence, and recommendations for future asthma incidence research. *Environ Health Perspect* (2006) **114**:634–40. doi:10.1289/ehp.8381
- Chiolero A, Santschi V, Paccaud F. Public health surveillance with electronic medical records: at risk of surveillance bias and overdiagnosis. *Eur J Public Health* (2013) **23**:350–1. doi:10.1093/eurpub/ckt044
- Myers J, Frieden TR, Bherwani KM, Henning KJ. Ethics in public health research. *Am J Public Health* (2008) **98**:793–801. doi:10.2105/AJPH.2006.107706
- Radley D, Finkelstein S, Stafford R. Off-label prescribing among office-based physicians. *Arch Intern Med* (2006) **166**:1021–6. doi:10.1001/archinte.166.9.1021
- Hersh A, Stefanick M, Stafford R. National use of postmenopausal hormone therapy: annual trends and response to recent evidence. *JAMA* (2004) **291**:47–53. doi:10.1001/jama.291.1.47
- Atkins C, Patel A, Taylor JTA, Biggerstaff M, Merlin TL, Dulin S, et al. Estimating effect of antiviral drug use during pandemic (H1N1) 2009 outbreak, United States. *Emerg Infect Dis* (2011) **17**:1591. doi:10.3201/eid1709.110295
- Aitken M, Berndt ER, Cutler DM. Prescription drug spending trends in the United States: looking beyond the turning point. *Health Aff* (2009) **28**:w151–60. doi:10.1377/hlthaff.28.1.w151
- Alexander G, Sehgal N, Moloney R, Stafford R. National trends in treatment of type 2 diabetes mellitus, 1994–2007. *Arch Intern Med* (2008) **168**:2088–94. doi:10.1001/archinte.168.19.2088
- Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature* (2009) **457**:1012–4. doi:10.1038/nature07634
- Patwardhan A, Bilkovski R. Comparison: flu prescription sales data from a retail pharmacy in the US with Google flu trends and us ilinet (cdc) data as flu activity indicator. *PLoS One* (2012) **7**:e43611. doi:10.1371/journal.pone.0043611
- 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**:e1003256. doi:10.1371/journal.pcbi.1003256
- Waters J, Howard K, Baker H, Brown C. Developing a social media system for biosurveillance. *Online J Public Health Inform* (2014) **6**:e129. doi:10.5210/ojphi.v6i1.5072
- Hartley DM, Nelson NP, Arthur RR, Barboza P, Collier N, Lightfoot N, et al. An overview of internet biosurveillance. *Clin Microbiol Infect* (2013) **19**:1006–13. doi:10.1111/1469-0691.12273



38. Torii M, Yin L, Nguyen T, Mazumdar CT, Liu H, Hartley DM, et al. An exploratory study of a text classification framework for internet-based surveillance of emerging epidemics. *Int J Med Inform* (2011) **80**:56–66. doi:10.1016/j.ijmedinf.2010.10.015
39. Kamel Boulos M, Resch B, Crowley D, Breslin J, Sohn G, Burtner R, et al. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, ogc standards and application examples. *Int J Health Geogr* (2011) **10**:67. doi:10.1186/1476-072x-10-67
40. Signorini A, Segre AM, Polgreen PM. The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza a h1n1 pandemic. *PLoS One* (2011) **6**:e19467. doi:10.1371/journal.pone.0019467
41. Cassa C, Chunara R. Twitter as a sentinel in emergency situations: lessons from the Boston marathon explosions. *PLoS Curr* (2013) **1**:5. doi:10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b
42. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *Am J Trop Med Hyg* (2012) **86**:39–45. doi:10.4269/ajtmh.2012.11-0597
43. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus a (h1n1) pandemic. *PLoS One* (2011) **6**:e23610. doi:10.1371/journal.pone.0023610
44. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. *Science* (2014) **343**:1203–5. doi:10.1126/science.1248506
45. Singer M, Clair S. Syndemics and public health: reconceptualizing disease in bio-social context. *Med Anthropol Q* (2003) **17**:423–41. doi:10.1525/maq.2003.17.4.423
46. Singer M. *Introduction to Syndemics: A Critical Systems Approach to Public and Community Health*. San Francisco: John Wiley & Sons Ltd (2009).
47. Pattemore P, Johnston S, Bardin P. Viruses as precipitants of asthma symptoms. I. epidemiology. *Clin Exp Allergy* (1992) **22**:325–36. doi:10.1111/j.1365-2222.1992.tb03094.x
48. Gern JE. Viral respiratory infection and the link to asthma. *Pediatr Infect Dis J* (2004) **23**:S78–86. doi:10.1097/01.inf.0000108196.46134.a6
49. Bueving HJ, Bernsen RMD, de Jongste JC, van Suijlekom-Smit LWA, Rimmelzwaan GF, Osterhaus AD, et al. Influenza vaccination in children with asthma. *Am J Respir Crit Care Med* (2004) **169**:488–93. doi:10.1164/rccm.200309-1251OC
50. Glezen WP. Asthma, influenza, and vaccination. *J Allergy Clin Immunol* (2006) **118**:1199–206. doi:10.1016/j.jaci.2006.08.032
51. Hanania NA, Atmar RL, Castro M. Influenza vaccine in patients with asthma. *Expert Rev Vaccines* (2006) **5**:111–8. doi:10.1586/14760584.5.1.111
52. Rank MA, Li JT. Clinical pearls for preventing, diagnosing, and treating seasonal and 2009 H1N1 influenza infection in patients with asthma. *J Allergy Clin Immunol* (2009) **124**:1123–6. doi:10.1016/j.jaci.2009.09.028
53. Castro M, Dozor A, Fish J, Irvin C, Scharf S, Scheipeter M, et al. The safety of inactivated influenza vaccine in adults and children with asthma. *N Engl J Med* (2001) **345**:1529–36. doi:10.1056/NEJMoa011961
54. Johnston SL, Ferrero F, Garcia ML, Dutkowski R. Oral oseltamivir improves pulmonary function and reduces exacerbation frequency for influenza-infected children with asthma. *Pediatr Infect Dis J* (2005) **24**:225–32. doi:10.1097/01.inf.0000154322.38267.ce

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Ramanathan, Pullum, Hobson, Stahl, Steed, Quinn, Chennubhotla and Valkova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.