Predicting Asthma-Related Emergency Department Visits Using Big Data

Sudha Ram, *Member, IEEE*, Wenli Zhang, Max Williams, and Yolande Pengetnze

***Abstract*—Asthma is one of the most prevalent and costly chronic conditions in the United States, which cannot be cured. However, accurate and timely surveillance data could allow for timely and targeted interventions at the community or individual level. Current national asthma disease surveillance systems can have data availability lags of up to two weeks. Rapid progress has been made in gathering nontraditional, digital information to perform disease surveillance. We introduce a novel method of using multiple data sources for predicting the number of asthma-related emergency department (ED) visits in a specific area. Twitter data, Google search interests, and environmental sensor data were collected for this purpose. Our preliminary findings show that our model can predict the number of asthma ED visits based on near-real-time environmental and social media data with approximately 70% precision. The results can be helpful for public health surveillance, ED preparedness, and targeted patient interventions.**

***Index Terms*—Asthma, big data, emergency department (ED) visits, environmental sensors, predictive modeling, social media.**

I. Introduction

ASTHMA is one of the most prevalent and costly chronic conditions in the United States, with 25 million people affected [1]. Asthma accounts for about two million emergency department (ED) visits, half a million hospitalizations, and 3 500 deaths [1], and incurs more than 50 billion dollars in direct medical costs annually [2]. Moreover, asthma is a leading cause of loss productivity with nearly 11 million missed school days and more than 14 million missed work days every year due to asthma [1]. Although asthma cannot be cured, many of its adverse events can be prevented by appropriate medication use and avoidance of environmental triggers [3]. The prediction of population- and individual-level risk for asthma adverse events using accurate and timely surveillance data could guide timely and targeted interventions, to reduce the societal burden of asthma.

At the population level, current national asthma disease surveillance programs rely on weekly reports to the Centers for Disease Control and Prevention (CDC) of data collected from various local resources by state health departments [4]. Notoriously, such data have a lag time of weeks, therefore, providing retrospective information that is not amenable to proactive and timely preventive interventions. At the individual level, known predictors of asthma ED visits and hospitalizations include past acute care utilization, medication use, and sociodemographic characteristics [5]–[7]. Common data sources for these variables include electronic medical records (EMR), medical insurance claims data, and population surveys, all of which, also, are subject to significant time lag. In an ongoing quality improvement project for asthma care, Parkland Center for Clinical Innovation (PCCI), Dallas, TX, USA, researchers have built an asthma predictive model relying on a combination of EMR and claim data to predict the risk for asthma-related ED visits within three months of data collection [Unpublished reports from PCCI]. Although the model performance (C-statistic 72%) and prediction timeframe (three months) are satisfying, a narrower prediction timeframe potentially could provide additional risk-stratification for more efficiency and timeliness in resource deployment. For instance, resources might be prioritized to first serve patients at high risk for an asthma ED visit within two weeks of data collection, while being safely deferred for patients with a later predicted high-risk period.

Novel sources of timely data on population- and individual-level asthma activities are needed to provide additional temporal and geographical granularity to asthma risk stratification. Short of collecting information directly from individual patients (a time- and resource-intensive endeavor), readily available public data will have to be repurposed intelligently to provide the required information.

There has been increasing interest in gathering nontraditional, digital information to perform disease surveillance. These include diverse datasets such as those stemming from social media, internet search, and environmental data. Twitter is an online social media platform that enables users to post and read 140-character messages called "tweets." It is a popular data source for disease surveillance using social media since it can provide nearly instant access to real-time social opinions. More importantly, tweets are often tagged by geographic location and time stamps potentially providing information for disease surveillance [8], [9]. Another notable nontraditional disease surveillance system has been a data-aggregating tool called Google Flu Trends, which uses aggregated search data to estimate flu activity [10], [11]. Google Trends was quite successful in its estimation of influenza-like illness. It is based on Google's search

Manuscript received September 29, 2014; revised December 4, 2014; accepted February 6, 2015. Date of publication February 19, 2015; date of current version July 23, 2015. This work was supported in parts by the National Center for Advancing Translational Sciences under Grant 1UL1TR001105-01, National Institutes of Health, and by Grant 10-D-028227 from the W.W. Caruth, Jr. Foundation at Communities Foundation of Texas, to PCCI.

S. Ram and W. Zhang are with the Department of Management Information Systems, University of Arizona, Tucson, AZ 85721 USA (e-mail: ram@eller.arizona.edu; wenlizhang@email.arizona.edu).

M. Williams is with the Parkland Center for Clinical Innovation, Dallas, TX 75247 USA (e-mail: max.williams@phhs.org).

Y. Pengetnze is with the Parkland Center for Clinical Innovation, Dallas, TX 75247 USA and also with the Children's Medical Center of Dallas, Dallas, TX 75235 USA (e-mail: yolande.pengetnzes@phhs.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2404829

engine, which tracks how often a particular search-term is entered relative to the total search-volume across a particular area. This enables access to the latest data from web search interest trends on a variety of topics, including diseases like asthma. Air pollutants are known triggers for asthma symptoms and exacerbations [12]. The United States Environmental Protection Agency (EPA) provides access to monitored air quality data collected at outdoor sensors across the country, which could be used as a data source for asthma prediction. Meanwhile, as health reform progresses, the quantity and variety of health records being made available electronically are increasing dramatically [13]. In contrast to the traditional disease surveillance systems, these new data sources have the potential to enable health organizations to respond to chronic conditions, like asthma, in real time. This in turn implies that health organizations can appropriately plan for staffing and equipment availability in a flexible manner. They can also provide early warning signals to the people at risk for asthma adverse events, and enable timely, proactive, and targeted preventive and therapeutic interventions.

Our research objective is to leverage social media, internet search, and environmental air quality data to estimate ED visits for asthma in a relatively discrete geographic area (a metropolitan area) within a relatively short time period (days). To this end, we have gathered asthma-related ED visits data, social media data from Twitter, internet users' search interests from Google and pollution sensor data from the EPA, all from the same geographic area and time period, to create a model for predicting asthma-related ED visits. This study is different from extant studies that typically predict the spread of contagious diseases using social media such as Twitter. Unlike influenza or other viral diseases, asthma is a noncommunicable health condition and we demonstrate the utility and value of linking big data from diverse sources in developing predictive models for noncommunicable diseases with a specific focus on asthma.

II. Background

A number of research studies have explored the use of novel data sources to propose rapid, cost-effective health status surveillance methodologies. Some of the early studies rely on document classification suggesting that Twitter data can be highly relevant for early detection of public health threats [14]. Others employ more complex linguistic analysis, such as the Ailment Topic Aspect Model [15], which is useful for syndrome surveillance. This type of analysis is useful for demonstrating the significance of social media as a promising new data source for health surveillance.

Other recent studies have linked social media data with real-world disease incidence to generate actionable knowledge useful for making health care decisions. These include [16], which analyzed Twitter messages related to influenza and correlated them with reported CDC statistics. Similarly, a study by Chew [17] during the 2009 H1N1 flu pandemic, validated Twitter as a real-time content, sentiment, and public attention trend-tracking tool. Collier [18] employed supervised classifiers (SVM and Naive Bayes) to classify tweets into four self-reported protective behavior categories. This study adds to evidence supporting a high degree of correlation between prediagnostic social media signals and diagnostic influenza case data.

While, these disease surveillance systems, including Google Flu trends [10], based on novel data sources have shown significant promise, other studies have challenged the accuracy of these systems for two reasons [19], [25]. 1) *Anomalous media spikes:* People searching flu terms may have had symptoms—but many users might have been simply looking for news stories about an anomalous season. Media attention might increase tweets about specific diseases but may not necessarily reflect actual disease affliction. 2) *Misleading information:* Tweets indicating awareness of disease, e.g., "Hope I don't get asthma," or using disease as rhetoric, e.g., "He is so cute I think I got asthma," are clearly about a specific disease but are not about actual disease affliction. These kinds of signals can be misleading and can mask signs of actual disease affliction. These issues and challenges are being addressed by several studies including [19] and [20]. For instance, Google Flu Trends engineers announced a redesign of their algorithm by dampening media attention and using Elastic Net, rather than regression, for estimation [19]. Commonly used Twitter techniques such as keyword matching or linear regression can widely overestimate the prevalence of disease. Broniatowski and Paul [8] have specifically addressed this issue and built models to determine if tweets were relevant to health, influenza, or an actual infection. In addition, they used geographic information associated with tweets for influenza surveillance.

Building on these techniques, our work uses a combination of data from multiple sources to predict the number of asthma-related ED visits in near real time. In doing so, we exploit geographic information associated with each dataset. We describe the techniques to process multiple types of datasets, to extract signals from each, integrate, and feed into a prediction model using machine learning algorithms, and demonstrate the feasibility of such a prediction.

III. Methods

We have examined the feasibility of using multiple data sources for predicting the number of asthma-related ED visits. A preliminary prediction model was built for this purpose. Semisupervised classification models were applied on data streams stemming from Twitter to distinguish tweets indicating asthma affliction from tweets just including keywords related to asthma. We also processed air quality data obtained from sensors, historical electronic health records indicating asthma-related visits to an emergency room, and Google search trends, all from the same specific geographic area, during the same time period.

A. Data Collection and Processing

1) ED Data: Deidentified aggregate data on ED visits for asthma as a primary diagnosis (International Classification of Disease Ninth [ICD9] code 493.00 to 493.99) to the Children's Medical Center (CMC) of Dallas were collected between October 2013 and December 2013 for a Quality Improvement initiative (see Fig. 1). Additional data were collected between November and December 2013 on ED visits for constipation

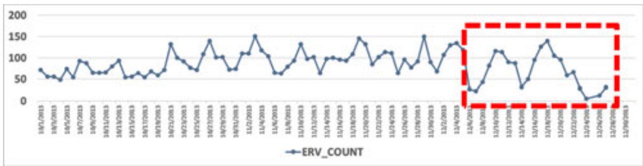


Fig. 1. Hospital administrative data on ED visits for asthma, Dallas, TX, USA, October 1, 2013 ∼ December 24, 2013.

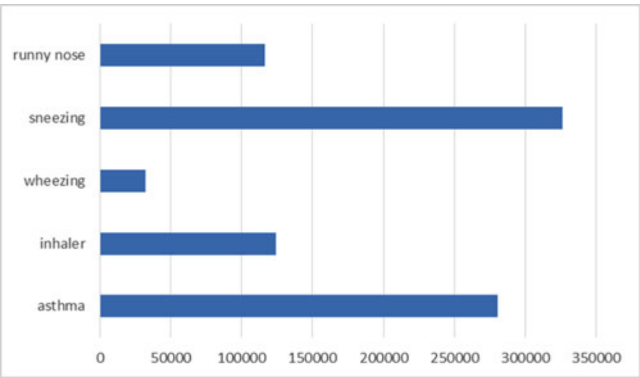


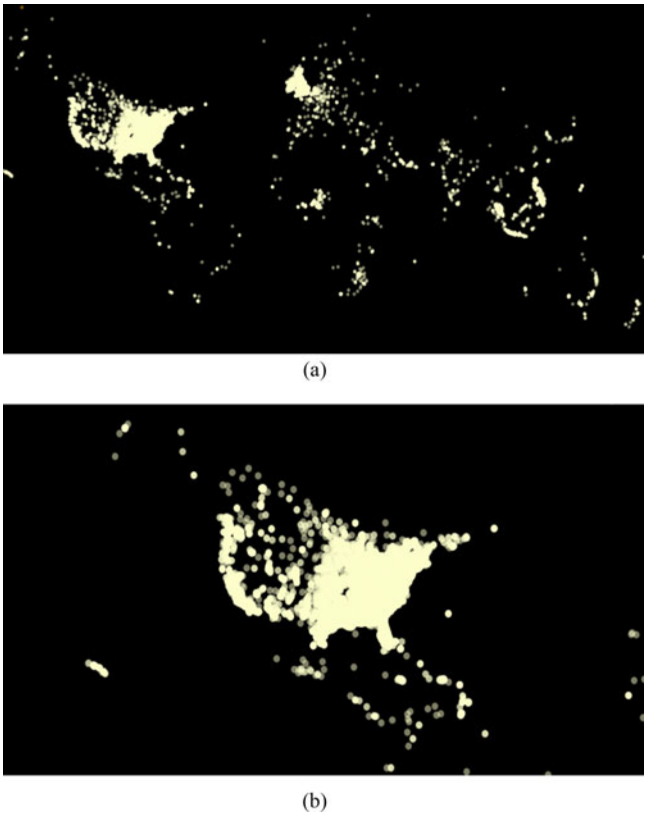
Fig. 2. Examples of keywords and frequency of tweets (Asthma stream, October 11, 2013 – December 31, 2013).

(ICD9 564.00 to 564.09) or abdominal pain (ICD9 789.00 to 789.09) to serve as controls for background variations in the number ED visits unrelated to asthma activity. The two control conditions were chosen because they affect a different organ system than asthma and are unlikely to be related to asthma. The study was approved by the CMC Institutional Review Board.

It should be noted that the average and standard deviation of the ED visits data from December (ave = 81, std = 40) are significantly different from the first two months' data (ave = 94, std = 25). So we excluded December's ED visits data from one of our correlation analyses.

2) Twitter Data: Twitter offers streaming APIs to give developers and researchers low latency access to its global stream of data. Public streams, which can provide access to the public data flowing through Twitter, were used in this study. Studies have estimated that using Twitter's Streaming API, researchers can expect to receive 1% of the tweets in near real time. Twitter4j, an unofficial Java library for the Twitter API, was used to access tweet information from the Twitter Streaming API. Two different Twitter datasets were collected in this study. 1) *The general twitter stream:* A simple collection of JSON grabbed from the general Twitter stream. The general tweet counts were used to estimate the Twitter population and further normalize asthma tweet counts. 2) *The asthma-related stream:* to collect only tweets containing any of 19 related keywords that were suggested by our clinical collaborators from PCCI. The asthma stream is limited to 1% of full tweets as well.

Fig. 2 shows the number of tweets in our asthma stream for some of the keywords used in data collection. Our Twitter dataset for this study was collected from October 11, 2013, through December 31, 2013, and contains 464 845 785 general tweets and 1 315 390 asthma-related tweets. On average, 15 000 asthma-related tweets were generated from all over the world per day. This demonstrates that Twitter is a promising data source for asthma surveillance and should be explored further.


(a)

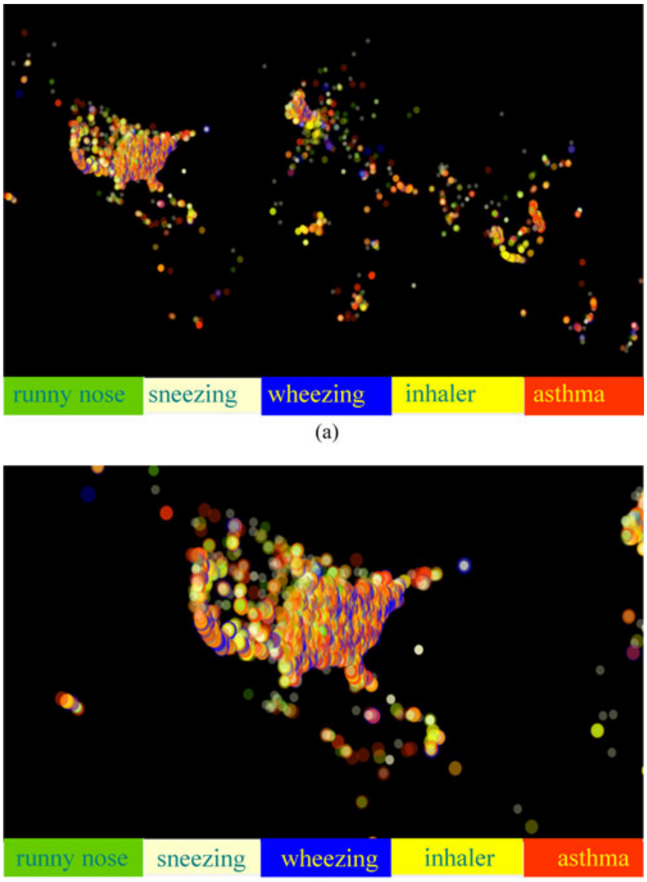
(b)

Fig. 3. (a) Global Asthma related tweets (Asthma stream, October 11, 2013–December 31, 2013). (b) Asthma related tweets, United States, (Asthma stream, October 11, 2013–December 31, 2013).

The geographic location of each tweet is identified via two fields: coordinates and location. Coordinates indicate the longitude and latitude of the tweet's location, e.g., {"coordinates": [−97.51087576, 35.46500176]}. Unfortunately, only a small percentage of tweets expose their coordinates. Among all the asthma-related tweets we collected, only 2% (35152/1315390) of the tweets revealed their coordinates as shown in Fig. 3(a). Most of these tweets are from English-speaking countries and approximately 60% of them are from the United States [see Fig. 3(b)]. We further analyzed these tweets based on a subset of our keywords [see Fig. 4(a) and (b)].

Locations indicate the cities and states where the users, who posted tweets, reside, e.g., {"location": "San Francisco, CA, USA"}. This information was collected from Twitter users' public biographic profiles. To estimate the prevalence of asthma-related tweets in a geographic region, we only included tweets from that particular region. We confined our analysis to the Twitter streaming data collected from the Dallas-Fort Worth (DFW) Metropolitan area in Texas, to closely match the geographical origins of patients in our clinical data sample. The boundary of the region and geographical "coordinates" are shown in Fig. 5.

As previously mentioned, not many people divulge their location in their tweets; consequently, in our dataset, only



(a)

(b)

Fig. 4. (a) Global Asthma-related tweets based on keywords (Asthma stream, October 11, 2013–December 31, 2013). (b) Asthma-related tweets based on keywords, United States (Asthma stream, October 11, 2013–December 31, 2013).

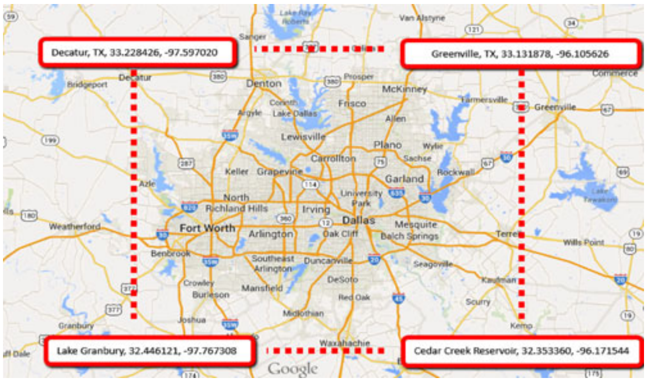

Decatur, TX, 33.228426, -97.597020
Greenville, TX, 33.131878, -96.105626
Lake Granbury, 32.446121, -97.767308
Cedar Creek Reservoir, 32.353360, -96.171544
Denton
McKinney
Fort Worth
Dallas
Google

Fig. 5. Geo boundaries defined for “coordinates” of Twitter asthma stream. Northwest: Decatur, TX, USA, 33.228426, −97.597020; Northeast: Greenville, TX, 33.131878, −96.105626; Southwest: Lake Granbury, 32.446121, −97.767308; Southeast: Cedar Creek Reservoir, 32.353360, −96.171544.

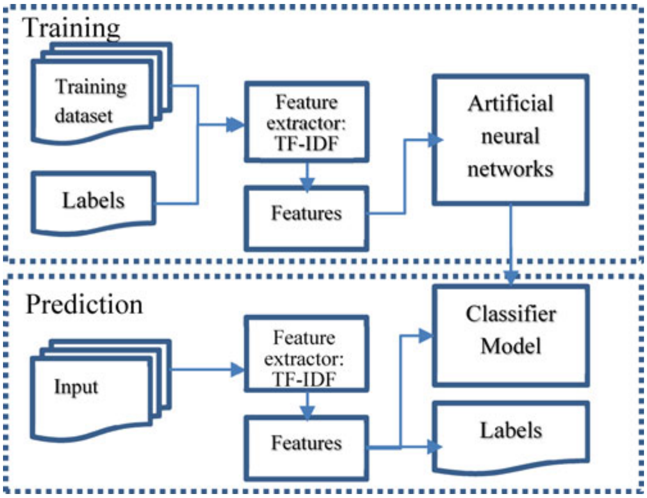


Fig. 6. Extracting signal from noisy twitter dataset.

892 asthma-related tweets were actually identified to fall within the geographic boundaries of interest to our study. Hence, we examined the dataset in more detail, and collected profiles of users tweeting about asthma. By examining these user profiles, we were able to extract users’ locations from their profile information and identified additional tweets stemming from our location of interest. We were, thus, able to identify 3 768 additional tweets from the asthma stream in the area of interest, and a total of 1 953 402 tweets from the general stream in the same area.

One of the challenges we needed to address was to extract signal from the noisy Twitter dataset, i.e., to distinguish tweets that are relevant to asthma from tweets that mentioned asthma in an irrelevant context. Fig. 6 shows the process used for cleaning the Twitter dataset. First, non-English tweets and retweets were excluded. The exclusion of non-English tweets is not expected to have a major impact on the analysis as 95% of tweets originating in the USA, including our geographical region of interest, are in English [21]. We transformed tweets to lower case, e.g., removing all the special characters, targets (@), hashtags (#), URLs, and emoticons. Each tweet was then tokenized by splitting based on nonletter characters. The tokens were used to generate a vector numerically representing each tweet. TF-IDF (term frequency–inverse document frequency) [22] was used for this purpose. All the words were stemmed by applying Porter’s algorithm [23] and English stop-words [24] were filtered out.

We then employed a machine learning classification technique called artificial neural networks (ANN) to accurately identify relevant tweets, using the process shown in Fig. 6. ANN is a supervised classification technique requiring a training dataset. First, we extracted a dataset containing 4 500 tweets from our asthma stream described earlier, i.e., each tweet in the training dataset contained at least one asthma-related keyword. This dataset was divided into three parts and each tweet was manually labeled by three researchers as “asthma relevant” or “asthma irrelevant.” The annotation criteria for “asthma relevant” tweets included: a statement that the individual has had asthma; and supporting criteria included: 1) severe difficulty in breathing as part of a discrete attack, (2) shortness of breath-triggered by exercise, stress, smoke, or irritants, 3) night time coughing duration greater than 1 month, 4) family history or childhood history of asthma, or 5) use of an inhaler.

The classifier was trained on this dataset. A tenfold cross validation was executed to evaluate the performance. The results showed a high overall accuracy of 85.78%. The

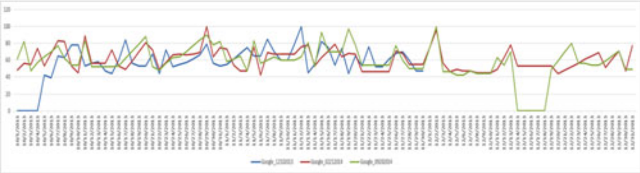


Fig. 7. Google search trends for Asthma-related keywords.

precision for "asthma irrelevant" class was 86.71% and the precision for "asthma relevant" class was 66.67%. The recall for "true asthma irrelevant" class was as high as 98.15%, whereas the recall for "true asthma relevant" class was 19.72%, indicating that there is a lot of noise in the data. Despite the latter low recall, the large Twitter dataset provided a sufficient number of: "true" asthma-related tweets for the analyses.

This complex cleaning process resulted in a dataset from which we were able to extract a sufficient number of asthma-related tweets in the geographic area of interest along with their specific timestamps.

3) Google Data: Google Trends analyzes its search engine traffic to determine the usage frequency of specific search-terms by individual users as compared to the total number of Google searches performed during a specific time period. To make it easier to compare data on different keywords, results in Google Trends are normalized using their total search traffic. Using the keywords from our Twitter asthma stream collection process, we extracted data from Google Trends. To align with the ED visits data, Google search interests were accessed for the same time period and in the same location as the Twitter data (see Fig. 7).

We retrieved Google search interest data by accessing the Google Trends website (www.google.com/trends) on three specific dates, December 10, 2013, February 21, 2014, and September 20, 2014 with the same query. For reasons unknown to us, the results are different on each of the three days. We used each of the three datasets for our analysis as described later.

4) Sensor Data: Air pollution data were collected from the EPA databases (www.epa.gov). The dataset contains measures of six types of pollutants, i.e., particulate matter, ground-level ozone, carbon monoxide, sulfur oxides, nitrogen oxides, and lead. The air quality indexes (AQI) associated with these pollutants were used in our model. The higher the AQI value, the greater the level of air pollution and the greater the health concern. Along with the AQI, we were able to get the AQS-SITE-ID (Air Quality System, site identification) from the EPA database. A Site ID is associated with a specific physical location and address. The site latitude and longitude also are provided. Using this information, we collected AQI data from 27 sites in the DFW area. The sites closest to the zip codes of origin of asthma patients were retained for analysis. Using this data, we calculated daily average AQI for our prediction model.

B. Prediction Model

We first analyzed the association between the asthma-related ED visits and data from Twitter, Google trends, and Air Quality sensors, using the Pearson correlation coefficient. We also examined the association between asthma-related tweet counts and ED visit counts for abdominal pain/constipation patients, to control for nonasthma-specific variations in ED visit counts. Then, we designed and implemented a prediction model to estimate the incidence of asthma ED visits at CMC using a combination of independent variables from the aforementioned data sources.

Since each dataset is from a different source and has different levels of granularity with respect to time and location, we first performed some transformations on each dataset to make them compatible. An important transformation was to normalize each dataset using a standard normalization technique, i.e., z-score

$$z = \frac{x - u}{\sigma} \tag{1}$$

where u is the mean and σ is the standard deviation.

Additionally, Twitter data were normalized by calculating the ratio of asthma-related tweets to the total number of tweets in the general twitter stream collected from the same geo-region in a given time.

For the prediction model, we employed four different classification methods: Decision tree, Naive Bayes, SVM, and ANN, and compared their classification accuracy. We also used techniques called adaptive boosting and stacking, to reduce classification errors. The ED visit counts were converted from numerical to categorical values based on the z-values, where the observations were labeled as "High," "Medium," or "Low." Our model was used to classify the predicted variable, i.e., number of daily ED visits, into one of three complementary and mutually exclusive classes—High, Medium, or Low. The Naive Bayes technique requires nominal data, hence, another transformation was used to convert all numerical data values into categorical values based on the z values similar to the transformation used for ED visit counts.

IV. Analysis and Results

A. Relationship Between ED Visits and Individual Types of Data

We first report on the relationship between asthma ED visits and each individual type of dataset, i.e., Twitter, Google trends, and Air Quality sensors.

Of note, Twitter data were only available beginning on October 11, 2013, and ED visits data were not available after December 24, 2013. We, therefore, performed the correlation analysis based on 74 days' worth of data. Our results indicate that absolute asthma tweets count is correlated with the asthma ED visit counts ($r = 0.328$, $p < 0.01$) (see Table I). After normalization of the number asthma tweets using the daily number of general tweets, the correlation coefficient improved ($r = 0.378$, $p < 0.01$) (see Table I).

Given that the average and standard deviation of the asthma ED visits data for December 2013 (ave = 81, std = 40) are significantly different from the first two months' data (ave = 94, std = 25), we did a sensitivity analysis excluding data from December, which left us 50 observations. The 50 observations showed further improvement of correlation between the number of asthma tweets and asthma ED visit counts (see Table II).

TABLE I
CORRELATION RESULTS BETWEEN TWITTER DATA (74 OBSERVATIONS) AND ASTHMA ED VISITS

		# of asthma affiliation tweets	# of normalized tweets
# of ED visits	Pearson Correlation	0.328^{**}	0.378^{**}
	Sig.	0.004	0.001
	N	74	74

** Correlation is significant at the 0.01 level.

TABLE II
CORRELATION RESULTS BETWEEN TWITTER DATA (50 OBSERVATIONS) AND ASTHMA ED VISITS

		# of Asthma Tweets	# of Normalized Asthma Tweets
# of ED Visits	Pearson Correlation	0.409^{**}	0.363^{**}
	Sig.	0.003	0.009
	N	50	50

** Correlation is significant at the 0.01 level.

TABLE III
CORRELATION RESULTS BETWEEN TWITTER DATA AND ABDOMINAL PAIN/CONSTIPATION ED VISITS

		# of Asthma Tweets	# of Normalized Asthma Tweets
# of Abdominal Pain/Constipation Patients ED visits	Pearson Correlation	0.084	0.075
	Sig.	0.697	0.729
	N	24	24

Note: We were only able to get the abdominal pain/constipation ED visits data from December 01–December 24, 2013.

TABLE IV
CORRELATION RESULTS BETWEEN AIR POLLUTION DATA AND ASTHMA ED VISITS

		CO	NO_2	PM2.5
# of ED Visits	Pearson Correlation	0.332^{**}	0.316^{**}	0.239^{*}
	Sig.	0.002	0.003	0.027
	N	85	85	85

** Correlation is significant at the 0.01 level.
* Correlation is significant at the 0.05 level.

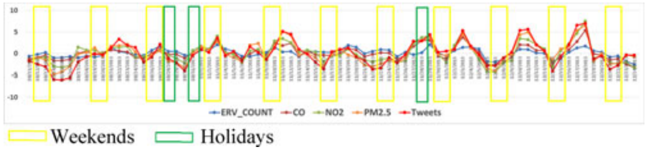


Fig. 8. Air pollution data, Twitter data and asthma ED visits.

TABLE V
CORRELATION RESULTS BETWEEN GOOGLE DATA AND ASTHMA ED VISITS

		Google 12/10/2013	Google 02/21/2014	Google 09/20/2014
# of ED visits	Pearson Correlation	0.298^{**}	0.049	0.049
	Sig.	0.002	0.654	0.658
	N	61	85	85

** Correlation is significant at the 0.01 level.

TABLE VI
BACKWARD FEATURE SELECTION TO DEMONSTRATE USEFULNESS OF ATTRIBUTES

Classification Methods	Attributes		
	CO+NO2+PM2.5	Tweets	ALL
Decision Tree	**65.18%**	63.93%	**65.18%**
ANN	61.25%	62.68%	**66.25%**
ANN (Adaptive Boosting)	62.50%	62.14%	**66.43%**
Stacking (ANN + Decision Tree)	61.07%	64.86%	**66.25%**

Evaluation metric: Accuracy; Design: tenfold cross validation.

Meanwhile, as a control, we also examined the relationship between asthma tweets count and abdominal pain/constipation ED visit counts. The absence of correlation speaks to the specificity of the association between asthma-related tweets and asthma ED visits (see Table III).

We next report on the correlation between the pollutant indexes and asthma ED visits (pollutant data after December 24, 2013 were removed since there was no ED visits data available after that date). Three pollutant indexes, i.e., CO, NO_2, and PM2.5 show significant correlation with asthma ED visits (see Table IV and Fig. 8).

As stated earlier, we had three different datasets collected from Google trends. We examined the relationship between each dataset and asthma ED visits (Table V). The asthma ED visits were only correlated with one of the Google datasets, hence, Google data were not included in the final prediction model.

B. Prediction Results

Based on the results from the correlation analysis, asthma tweets, CO, NO_2, and PM2.5 were selected as inputs into our prediction model. We are only reporting results for the Decision Tree and ANN techniques, as the Naive Bayes and SVM techniques did not yield good prediction results.

First, backward feature selection algorithm was used to examine if the addition of Twitter data would improve the prediction. As shown in Table VI, combining air quality data with tweets resulted in higher prediction accuracy.

Additionally, we evaluated prediction precision. Given that our prediction task is for a three-way classification, each technique resulted in different prediction and/or precision in different classes (see Table VII). Decision Tree performed well in predicting the "High" class, while ANN, after Adaptive Boosting, worked well for the "Low" class. Stacking the two techniques performed well for the "Medium" class.

The results of our analysis are promising because they perform with a fairly high level of accuracy overall. As noted in the introduction, traditional asthma ED visit models are useful for predicting events in a three-month window and have an accuracy of approximately 70%. It is to be noted that "traditional models" estimate a risk score for asthma ED visit

TABLE VII
PREDICTION RESULTS

Classification Methods	Class	Class Precision
Decision Tree	High	72.73%
ANN	Low	71.43%
ANN (Adaptive Boosting)	Low	72.73%
Stacking (ANN + Decision Tree)	Medium	75.00%

Evaluation metric: Precision; Design: tenfold cross validation.

for each individual patient, whereas our "Twitter/Environmental data model" predicts the risk for a daily number of ED visits being High, Low, or Medium. The former is an individual-level risk model, while the latter is a population-level risk model. Our population-level asthma risk prediction model has the potential for complementing current individual-level models, and may lead to a shorter time window and better accuracy of prediction. This in turn has implications for better planning and proactive treatment in specific geo-locations at specific time periods.

V. Discussion, Implications, and Limitations

Although preliminary, the findings of this study are very promising for many reasons. As asthma prevalence continues to rise, novel and coordinated strategies are required at the public health and clinical levels to curb the societal burden of asthma adverse outcomes. Readily available, real-time or near-real-time, environmental and internet-based data offer a unique opportunity for early identification of clusters of patients or communities at high risk for severe asthma events at a given time. Interventions would be prioritized in time and place to reduce the risk for asthma ED visits. For instance, public health resources could be used to reach out to patients from high-risk clusters or communities at any given time, and direct them toward less costly and more efficient care sites such as their primary care provider offices. Moreover, predicted risks could be spatially and temporally visualized, and made available to community stakeholders through various media sources. Clinical resources could be prioritized to offer earlier clinic appointments to patients with impending risk for failure, and later, slots to patients with deferred risk. Additionally, hospitals and EDs could use such risk stratification for optimal resource planning, such as ED staffing or opening observation units.

The limitations of our study include: First, this study is limited to English tweets, which might not accurately represent the Twitter activity of non-English speakers, although, the latter are likely represented in our ED sample. However, we do not expect this to have a significant impact on the final results given that non-English tweets represent less than 5% of tweets in the USA, as discussed in the "Methods" section. Second, this study was limited to ED visits data from one hospital only which did not allow us to examine variations around different clinical care sites. A larger study is underway to validate these preliminary findings in a larger clinical sample spanning a wider geographical area over a longer timeframe. Third, the current model is designed for a noncommunicable disease (asthma) with a significant prevalence in the community. Although similar models could be developed for other noncommunicable diseases such as diabetes and chronic obstructive pulmonary disease (COPD), this model might not be suitable for communicable diseases or for diseases with low prevalence and high social media activity reflecting public awareness rather than actual disease activity.

VI. Conclusion and Future Research

In this study, we have provided preliminary evidence that social media and environmental data can be leveraged to accurately predict asthma ED visits at a population level.

We are in the process of confirming these preliminary findings by collecting larger clinical datasets across different seasons and multiple hospitals. Our continued work is focused on extending this research to propose a temporal prediction model that analyzes the trends in tweets and AQI changes, and estimates the time lag between these changes and the number of asthma ED visits. We also are collecting AQI data over a longer time period to examine the effects of seasonal variations. In addition, we would like to explore the effect of relevant data from other types of social media interactions, e.g., blogs and discussion forums, on our asthma visit prediction model. Additional studies are needed to examine how combining real-time or near-real-time social media and environmental data with more traditional data might affect the performance and timing of the current individual-level prediction models for asthma, and eventually, for other chronic conditions. In future projects, we intend to extend our work to diseases with geographical and temporal variability, e.g., COPD and diabetes.

Acknowledgment

The authors would like to thank Dr. P. Jamieson and Ms. A. Decrescendo from PCCI for helping with the annotation and cleaning of asthma tweets.

References

[1] L. J. Akinbami, J. E. Moorman, and X. Liu, "Asthma prevalence, health care use, and mortality: United States, 2005–2009," National Center for Health Statistics, Hyattsville, MD, USA, National health statistics reports no. 32, 2011.

[2] "Vital signs: Asthma prevalence, disease characteristics, and self-management education: United States, 2001–2009," Centers for Disease Control and Prevention Atlanta, GA, USA, Morbidity and mortality weekly report, vol. 60, no. 17, pp. 547, 2011.

[3] "Guidelines for the Diagnosis and Management of Asthma," National Institutes of Health, Bethesda, MD, USA, Expert Panel Report 3, vol. 2, 1997.

[4] Centers for Disease Control and Prevention. (2010). About the Morbidity and Mortality Weekly Report (MMWR) Series. [Online]. Available: http://www.cdc.gov/mmwr/about.html

[5] G. R. Pesola, F. Xu, H. Ahsan, P. Sternfels, I. H. Meyer, and J. G. Ford, "Predicting asthma morbidity in Harlem emergency department patients," *Acad. Emergency Med.*, vol. 11, no. 9, pp. 944–950, 2004.

[6] M. Schatz, R. S. Zeiger, W. M. Vollmer, D. Mosen, G. Mendoza, A. J. Apter, T. B. Stibolt, A. Leong, M. S. Johnson, and E. F. Cook, "The controller-to-total asthma medication ratio is associated with patient-centered as well as utilization outcomes," *Chest J.*, vol. 130, no. 1, pp. 43–50, 2006.

[7] C. Tolomeo, C. Savrin, M. Heinzer, and A. Bazzy-Asaad, "Predictors of asthma-related pediatric emergency department visits and hospitalizations," *J. Asthma*, vol. 46, no. 8, pp. 829–834, 2009.

[8] D. A. Broniatowski, M. J. Paul, and M. Dredze, "National and local influenza surveillance through twitter: An analysis of the 2012–2013 influenza epidemic," *PloS One*, vol. 8, no. 12, p. e83672, 2013.

[9] E.-K. Kim, J. H. Seok, J. S. Oh, H. W. Lee, and K. H. Kim, “Use of hangeul twitter to track and predict human influenza infection,” *PloS one*, vol. 8, no. 7, p. e69305, 2013.

[10] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2008.

[11] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi, “Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic,” *PloS One*, vol. 6, no. 8, p. e23610, 2011.

[12] L. Trasande and G. D. Thurston, “The role of air pollution in asthma and other pediatric morbidities,” *J. Allergy Clin. Immunol.*, vol. 115, no. 4, pp. 689–699, 2005.

[13] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: Using analytics to identify and manage high-risk and high-cost patients,” *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[14] M. Krieck, J. Dreesman, L. Otrusina, and K. Denecke, “A new age of public health: Identifying disease outbreaks by analyzing tweets,” presented at the Health Web-Science Workshop, ACM Web Science Conf., Koblenz, Germany, 2011.

[15] M. J. Paul and M. Dredze, “You are what you Tweet: Analyzing twitter for public health,” in *Proc. Int. Conf. Weblogs Social Media* , 2011, pp. 265–272.

[16] C. Aron, “Towards detecting influenza epidemics by analyzing Twitter messages,” in *Proc. ACM 1st Workshop Social Media Anal.*, 2010, pp. 115–122.

[17] C. Chew and G. Eysenbach, “Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak,” *PloS one*, vol. 5, no. 11, e14118, 2010.

[18] C. Nigel, N. T. Son, and N. M. Nguyen, “OMG U got flu? Analysis of shared health messages for bio-surveillance,” *J. Biomed. Semantics*, vol. 2, no. S-5, S9, 2011.

[19] D. M. Lazer, R. Kennedy, G. King, and A. Vespignani. (2014). Google Flu trends still appears sick: An evaluation of the 2013–2014 flu season. [online]. Available: http://ssrn.com/abstract = 2408560

[20] M. Mark, S.-H. Zhu, W. Chapman, and M. Conway, “Using Twitter to examine smoking behavior and perceptions of emerging tobacco products,” *J. Med. Internet Res.*, vol. 15, no. 8, e174, 2013.

[21] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, “The twitter of Babel: Mapping world languages through microblogging platforms,” *PLoS ONE*, vol. 8, no. 4, 2013.

[22] G. Salton and M. J. MacGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[23] M. Porter. (2006). The Porter stemming algorithm. [online]. Available: http://tartarus.org/martin/PorterStemmer/

[24] C. Fox, “A stop list for general text,” *ACM SIGIR Forum*, vol. 24, no. 1–2, pp. 19–21, 1989.

[25] D. Scanfeld, V. Scanfeld, and E. L. Larson, “Dissemination of health information through social networks: Twitter and antibiotics,” *Amer. J. Infection Control*, vol. 38, no. 3, pp. 182–188, 2010.

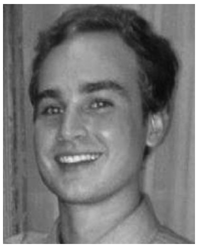
Sudha Ram (M’85) received the Ph.D. degree from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1985.

She is the Anheuser-Busch Endowed Professor of management information systems, and entrepreneurship & innovation in the Eller College of Management, University of Arizona, Tucson, AZ, USA. She has joint faculty appointment as a Professor of computer science. She is the Director of the Advanced Database Research Group and a Codirector of INSITE: Center for Business Intelligence and Analytics (www.insiteua.org), University of Arizona. Her research interests include the areas of enterprise data management, business intelligence, large scale networks, and data analytics. Her work uses different methods such as machine learning, statistical approaches, ontologies, and conceptual modeling. She has published articles in journals such as the *Communications of the ACM*, IEEE INTELLIGENT SYSTEMS, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Information Systems*, *Information Systems Research*, *Management Science*, and *MIS Quarterly*. Her research has been highlighted in several media outlets including NPR news,

Dr. Ram was a Speaker for a TED talk in December 2013 on “Creating a Smarter World with Big Data.”

Wenli Zhang is currently working toward the Ph.D. degree in management information systems at the University of Arizona, Tucson, AZ, USA.

Her main research interests include big data in healthcare, analyzing and mining social networks.



Max Williams received the B.Sc. degree in biomedical sciences from Texas A&M University, College Station, TX, USA, in 2013.

He is currently working as a Research Fellow at the Parkland Center for Clinical Innovation, Dallas, TX. He plans to pursue a career in public health.

Yolande Pengetnze received the M.D. degree in 1998 from the University of Yaounde, Yaounde, Cameroon. She completed Pediatric Residency training in 2008 from Maimonides Medical Center, New York, NY, USA, then General Pediatric/Health Services Research Fellowship training combined with a Masters of Sciences in Clinical Sciences in 2013 from the University of Texas Southwestern Medical Center (UTSW), Dallas, TX, USA.

She joined Parkland Center for Clinical Innovation (PCCI), Dallas, in December 2013, as a Physician Scientist and holds a Clinical Faculty position at UTSW. Her research interests include the use of advanced predictive analytics integrating traditional data sources and novel ..Big data.. sources to improve health outcomes at the individual and population level. She is currently leading multiple projects at PCCI, including population health quality improvement projects in pediatric asthma using both traditional and nontraditional data sources.