

A COMPREHENSIVE ANALYSIS OF TRIGGERS AND RISK FACTORS FOR ASTHMA BASED ON MACHINE LEARNING AND LARGE HETEROGENEOUS DATA SOURCES¹

Wenli Zhang

Debbie and Jerry Ivy College of Business, Iowa State University, 2167 Union Drive,
Ames, IA 50011 U.S.A. {wlzhang@iastate.edu}

Sudha Ram

Eller College of Management, University of Arizona, 1130 E. Helen Street,
Tucson, AZ 85721 U.S.A. {ram@eller.arizona.edu}

Asthma is a common chronic health condition affecting millions of people in the United States. While asthma cannot be cured, it can be managed if we identify and understand triggers and risk factors that cause asthma exacerbations. However, this is challenging because these triggers and risk factors are complex and interconnected, and there are limitations to current mainstream approaches for identifying them. The recent availability of massive amounts of heterogeneous data has opened up new possibilities for asthma triggers and risk factors analyses. In this study, we introduce a data-driven framework, adapt and integrate multiple advanced machine learning techniques, and perform an empirical analysis to (1) derive characteristics of self-reported asthma patients from social media, (2) enable integration and repurposing of highly heterogeneous and commonly available datasets, and (3) uncover the sequential patterns of asthma triggers and risk factors, and their relative importance, both of which are difficult to achieve via retrospective cohort-based studies. Our methods and results can provide guidance for developing asthma management plans and interventions for specific subpopulations and, eventually, have the potential to reduce the societal burden of asthma.

Keywords: Chronic disease management, asthma triggers/risk factors, design science, machine learning, distant supervision, convolutional neural networks, sequential pattern mining, geometric inference, random forest

Introduction

Asthma is a common chronic disease that affects people of all ages. In the United States, more than 39.5 million people, including 10.5 million children, have been diagnosed with asthma in their lifetime. Of these people, 18.9 million adults and 7.1 million children are plagued by persistent asthma currently. Annually, it results in more than 2 million emergency department visits, half a million hospitalizations, and 3,500

deaths (CDC 2013). Due to causing repeated symptoms such as wheezing, chest tightness, shortness of breath, and nighttime or early morning coughing, asthma is considered as a leading cause of lost productivity with nearly 11 million missed school days and more than 14 million missed work days every year (Akinbami et al. 2011). The estimated cost of asthma to society from loss of productivity is about \$4 billion annually (Barnett and Nurmagambetov 2011).

Asthma has received a lot of attention from many medical and health researchers. However, the causes of asthma are still not completely understood and there is no cure for asthma (WHO 2017). Fortunately, we now have improved under-

¹The accepting senior editors for this paper were Indranil Bardhan, Hsinchun Chen, and Elena Karahanna.

standing of asthma as an inflammatory disease: it can be controlled with proper diagnosis, appropriate care, and personalized medication management plans to avoid *triggers* (i.e., momentary events that may immediately bring on symptoms, such as weather conditions) and *risk factors* (i.e., constant or chronic factors in one's life that may cause asthma exacerbation, such as race and smoking habits). In this study, we focus on identifying and understanding triggers and risk factors contributing to asthma exacerbations.

Identification of these triggers and risk factors is fraught with challenges. The first major challenge lies in the complexity of asthma. As with many chronic diseases, asthma triggers and risk factors are complex and varied (Figure 1), including multiple biological, demographic, behavioral, environmental, social, psychological, and infectious determinants. In addition to appropriate medication management, asthma patients must have logistical, financial, and cultural access to environments that foster avoidance of asthma triggers and risk factors and encourage good asthma management practices (Herman 2011). In view of this, a comprehensive analysis is needed to provide guidance for developing asthma management plans and population-level asthma interventions. The second challenge is that most existing asthma studies focus on one or a few triggers or risk factors (see Appendix A), and, therefore, fail to identify factors critical to asthma prevention and ignore connections among these factors. They also fail to identify the relative importance of risk factors, and thus are not sufficient to direct the development of asthma management strategies. The third challenge is the limitations of current mainstream approaches for asthma trigger and risk factor analyses. Most of these analyses (approximately 60%) are conducted using cohort studies relying on surveys (see Appendix A), which are limited by time, funding, or access to large populations.

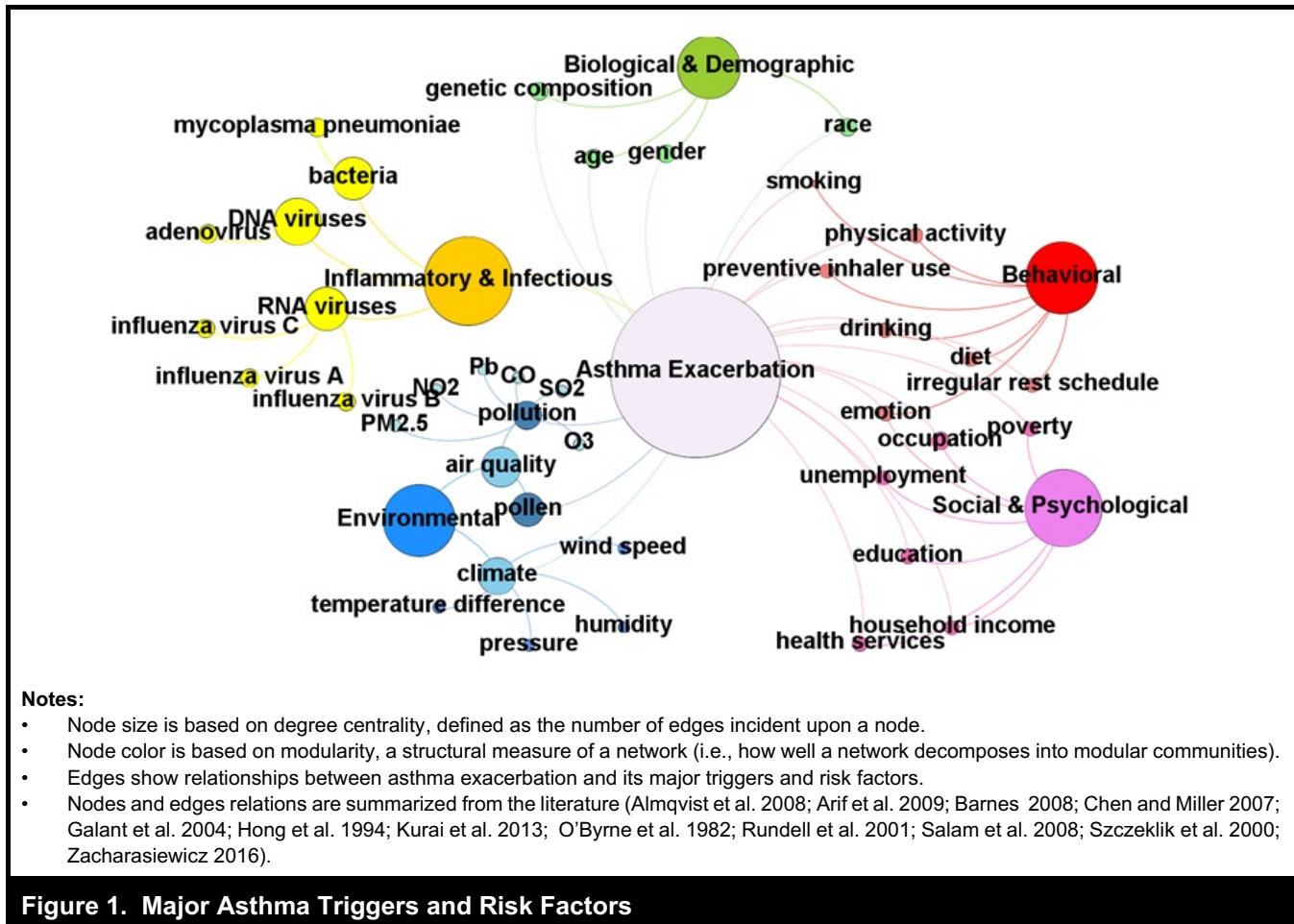
Due to the limitations of current studies, there is an imminent need to turn to additional sources of knowledge to economically identify asthma triggers and risk factors in a timely manner. Over the past two decades, the impact of big data (defined as data with large-volumes, arising from many independent sources, with distributed and decentralized control, used to explore complex and evolving relationships; Wu et al. 2014) for healthcare analytics has increased remarkably (Bates et al. 2014). Of patients with chronic conditions, 25% share their experience on social media sites (Andreu-Perez et al. 2015), which has opened up new possibilities to connect patients and health providers beyond the clinic. Wearable devices and environmental sensor data are readily available and relatively inexpensive to use. Recent studies demonstrated the use of remote sensing datasets to enhance chronic disease management (Eurowinter Group 1997; Kovats and Hajat, 2008; Ramachandran et al., 2013). All in all, a data-

driven solution holds great promise for helping analyze asthma risk factors from a more comprehensive perspective.

In this study, following the design science paradigm (Hevner et al. 2004), we propose a new data-driven approach to answer the following two research questions:

- (1) How can we identify asthma patients on social media and derive characteristics of asthma patients from social media, instead of using traditional data collection methods (e.g., questionnaires or surveys)? The implication is that the proposed methods will provide a new way to collect the background and demographic information from a large targeted population and may serve as a substitute or complement to traditional questionnaires or survey-based data collection methods.
- (2) How can we leverage existing knowledge about asthma and make full use of big data by repurposing and integrating multiple open data sources to identify asthma triggers and risk factors, their interconnecting relations, and their relative importance? The implication is that the proposed framework would provide a new way to understand the triggers and risk factors for chronic conditions and could be used as a complement to retrospective cohort studies on chronic conditions.

The contributions of this study are twofold. First, from the perspective of design science, we propose a new two-stage model to derive characteristics of self-reported asthma patients from social media. We propose the use of image recognition to enhance asthma patients' background information extraction. We show that our proposed model outperforms state-of-the-art methods. We incorporate the geometric inference algorithm into our proposed framework, which enables us to integrate and repurpose highly heterogeneous data from multiple open sources. We adapt sequential pattern mining to uncover the sequential patterns of asthma triggers and risk factors. We propose using the random forest algorithm to uncover the relative importance of asthma triggers and risk factors. Both of these are difficult to achieve using retrospective cohort studies, and hence our methods can be used as a complement to such studies. Second, from the perspective of chronic disease management, we propose a framework for comprehensive asthma triggers and risk factor analysis. We use a novel data source for deriving characteristics of self-reported asthma patients and provide a substitute for traditional survey-based data collection methods. The proposed framework is able to extract information from various open data sources with diverse spatial-temporal resolutions. A system based on the proposed framework has the potential to deliver guidance to stakeholders to manage and curb the societal burden of asthma.



In the remainder of this paper, we first discuss the challenges associated with analyzing asthma triggers and risk factors and the shortcomings of existing studies; we then provide a brief review of the design science methodology. Next, we develop and describe a new data-driven framework for conducting comprehensive asthma risk factors analysis. We implement and evaluate the proposed framework and demonstrate its feasibility and implications. Finally, we conclude our work with a summary and directions for future research.

Related Work

Asthma Triggers/Risk Factors and Limitations of Current Studies

Asthma and its associated triggers and risk factors have been a research focus for decades. An asthma trigger or risk factor is a momentary event or constant factor, including any attribute, characteristic, or exposure, that increases the likelihood

of exacerbating asthma for an individual (WHO 2017). Medical theories indicate that asthma triggers and risk factors include

- (1) *Biological and demographic factors.* Biological and demographic factors are essential to understanding asthma exacerbation. For instance, immune cells have been found to play a role in allergic asthma (Barnes 2008). Studies have shown gender differences to be related to the prevalence of asthma (Almqvist et al. 2008). The relationship between racial groups and asthma morbidity has been examined, reflecting potential genetic causes for asthma (Galant et al. 2004).
- (2) *Environmental factors.* Environmental factors have continuously been the focus of attention in asthma studies. Air pollution is suggested to be linked to asthma prevalence (Salam et al. 2008). Weather, especially cold air, could be a significant factor for asthma exacerbation (O'Byrne et al. 1982).

- (3) *Behavioral factors.* The interaction effects of behavioral factors on asthma exacerbation have been outlined by some researchers (Hong et al. 1994). In particular, smoking has consistently been shown to increase the risk of asthma in different studies (Zacharasiewicz 2016). Asthma patients who smoke appear to have an impaired response to the beneficial effects of anti-asthma drugs (Chaudhuri et al. 2003). Other behavioral factors include exercise (Rundell et al. 2001), irregular rest schedule (Kozyrskyj et al. 2009), and taking medications such as aspirin (Szczeklik et al. 2000).
- (4) *Social and psychological factors.* A number of studies highlight the importance of both social (e.g., social position or status; quality of social relationships; Wright et al. 1998) and psychological exposures (e.g., life stress, extreme emotions such as anger or fear; Chen and Miller 2007) in the exacerbation of asthma symptoms, which are important in understanding the rising asthma burden. In addition, certain occupational hazards and exposures are found to be related to asthma (Arif et al. 2009).
- (5) *Inflammatory and infectious factors.* Research indicates that infections by respiratory bacteria and/or viruses are important triggers of asthma exacerbations (CDC 2017b; Smith et al. 2017). For example, both rhinovirus, which causes the common cold, and influenza A, which causes the flu, can result in asthma exacerbations (Kurai et al. 2013). Nicholson et al. (1993) noted that about 25% of laboratory-confirmed viral acute upper respiratory infections are associated with almost half of the most severe asthma exacerbations.

Although important, these extant studies have substantial limitations. First, asthma triggers and risk factors in each of the above-mentioned categories are known to be related to each other. However, most studies have focused on one or two types of triggers or risk factors (84.7% of current studies focused on one type of asthma triggers or risk factors, 9.5% analyzed two types, and only 5.8% studied more than two types of asthma triggers or risk factors—see Appendix A) and, therefore, failed to discover the potential connections, especially the sequential patterns among them. Moreover, existing studies mostly depend on methods such as regression models with statistical significance (i.e., *p*-value) reported. They lack detailed assessments of risk factors (i.e., relative importance), and thus are not sufficient to direct the development of asthma management strategies. Most importantly, survey-based data collection methods and retrospective cohort studies (59.1%—see Appendix A) are still the most pervasive methods for asthma triggers and risk factors analysis. Although they have some important advantages, such as facilitating the control of factors, they have many disadvantages

including limited access to the population of concern, lack of time and funding necessary to conduct large-scale studies, not being amenable for tracking trends, and respondents not providing honest answers to lifestyle-, psychology-, and/or demography-related questions.

Asthma Trigger and Risk Factor Detection Using Novel Data Sources and Machine Learning

In recent years, asthma-related research has broadened to include novel data sources and various machine learning techniques to develop methods that integrate with our existing knowledge on asthma. Lee et al. (2011) proposed a pattern-based sequential mining framework by considering both biological and environmental factors. This was one of the first studies that applied the machine learning paradigm in this area. Sadat et al. (2015) proposed a spatial association rule-mining method to analyze the effect of air pollutants on asthma exacerbation. The novelty of this work is in adopting environmental sensing data and taking their spatial relations into account. Jalali et al. (2015) and Tang et al. (2015) pushed this research area forward by proposing an *events* (i.e., life events) model to examine the effect of environmental and demographic factors on asthma exacerbations. Their contributions are using social media data to obtain patients' life events data, adding the temporal relations of factors into the analysis, and computing the relative importance of asthma risk factors.

Although interesting and promising, the results of these studies are still preliminary. They also share some known issues and limitations. First, they rely heavily on traditional data collection methods (e.g., surveys or/and literature reviews) to obtain the patient background information needed to analyze asthma triggers and risk factors, which can be expensive and constrained by the original/principal purpose of the data being collected. Second, there is still no comprehensive study (one that focuses on multiple types of factors contributing to asthma exacerbations) and the sequential patterns of various risk factors are still not fully understood.

The deficiencies of existing asthma trigger and risk factor analyses, coupled with the challenges associated with building more effective methods, motivate our current work to develop a new data-driven framework (Lopez et al. 2006). The design science paradigm provides a good foundation for our current study. Design science is an outcome-based research methodology (Nunamaker et al. 1990). According to its definition, a *design* is both a *product* and a *process* (Hevner et al. 2004). The product is an artifact that can be broadly defined as construct, method, model, or instantiation (Simon 1996). The

process is a sequence of expert activities composed of the procedures taken to develop and evaluate the artifact (March and Smith 1995). In this study, the artifact we intend to deliver is a framework consisting of methods and instantiations that can be used to (1) target self-reported asthma patients and derive their characteristics from social media instead of using traditional survey-based data collection methods and (2) propose methods that can be used as a complement to traditional retrospective cohort studies and make full use of big data by integrating and repurposing data from diverse sources to help identify asthma triggers and risk factors, their interconnecting relations, and their relative importance.

Deriving User Characteristics from Social Media

One focus of this study, as mentioned earlier, is to extract characteristics (e.g., demographic, behavioral, social, and psychological characteristics) of self-reported asthma patients from social media datasets instead of using traditional survey-based data collection methods.

Social media is defined as web-based platforms (e.g., Twitter, Facebook, Instagram, etc.), where users (1) have unique user profiles; (2) have access to and/or provide digital content; and (3) can articulate a list of other users with whom they share a relational connection (Boyd and Ellison 2007; Kane et al. 2014). In this study, we are specifically interested in social media user profiles. Social media platforms typically provide a unique user profile that is constructed by the user, by the members' networks, and by the platform (Kane et al. 2014). Social media user profiles have the potential to capture and understand population characteristics that may otherwise be challenging to identify. The availability of massive social media datasets has given rise to a growing body of work that uses various techniques for the measurement of population characteristics on an unprecedented scale (Ruths and Pfeffer 2014). This is because, despite the possible privacy issues, a lot of users still tend to have *high self-presentation* (i.e., in social interactions, people have a desire to control the impressions other people form of them, and people wish to create an image that is consistent with their personal identity; Derlaga and Berg 1987) and *high self-disclosure* (i.e., a process of communication by which a person consciously or unconsciously reveals personal information that is consistent with the image they would like to give about him/herself to another; Dindia 2001) on social media sites (Kaplan and Haenlein 2010). This is driven by (1) a person's desire to present themselves in cyberspace (Schau and Gilly 2003); (2) the development of close relationships (Gibbs et al. 2006); and (3) the wish to influence other users to gain rewards

(Andrade et al. 2002). Research has shown that a large proportion of social media users tend to show authentic, unique, and stable demographic attributes that could explain their online behaviors (Kane et al. 2014; Xiang et al. 2017).

Hence, it is reasonable to assume that we can derive attributes of self-reported asthma patients from social media datasets. For example, many studies have explored ways to extract users' demographic attributes from social media, including geolocation, gender, age, race/ethnicity, and occupation (Burger et al. 2011; Chen et al. 2015; Huang et al. 2015; Mislove et al. 2011; Rao et al. 2010; Sloan et al. 2015). Increasingly, studies show that social media data can be used as a source to disclose a range of behavioral patterns, for example, work–rest schedule, eating and smoking habits, etc. (Abbar et al. 2015; Jamison-Powell et al. 2012; Myslín et al. 2013). Social media users' psychological status identification is another interesting area of research, including sentimental analysis, depression detection, and satisfaction with life (Fang et al. 2015; Ferguson et al. 2014; Mohammad et al. 2013; Nesi and Prinstein 2015). These studies demonstrate that social media has the potential to be a valuable source for uncovering underlying individual and population characteristics.

Identifying Asthma Risk Factors from Heterogeneous Data Sources

As stated earlier, another focus of this study is to identify risk factors for asthma using a variety of heterogeneous data sources. Because of the complexity of asthma triggers and risk factors, the proposed methods require features from various data sources. These data sources can represent multiple biological, demographic, behavioral, environmental, social, psychological, and infectious determinants. These data (a.k.a., big data) are at different resolutions with disparate spatial and temporal scales. Big data analytics is important for asthma trigger and risk factor analyses because (1) it can substantially improve decision-making to improve the development of the next generation of products and service (Manyika et al. 2011); (2) data-driven decisions can augment or occasionally overrule human judgment (Walker 2014); and (3) it provides opportunities for discovering new relationships, helping us gain an in-depth understanding of the hidden value (Chen et al. 2014).

Based on a review of extant literature, we note that asthma triggers and risk factors analysis remains an open research area with many challenging issues, such as providing a comprehensive view of asthma triggers and risk factors, uncovering the sequential patterns and the relative importance of different triggers and risk factors, and deriving characteristics of self-reported asthma patients from novel data

sources instead of using a traditional survey-based data collection method. Given the need for advanced analytics methods and the lack of systems that address these challenges, an obvious necessity arises. On the basis of the current understanding of asthma, we build on previous work to identify asthma triggers and risk factors using heterogeneous data sources and machine learning techniques. Specifically, we develop a new data-driven approach to examining triggers and risk factors contributing to asthma exacerbations.

Research Design and Framework for Asthma Triggers and Risk Factors Analysis

In this study, we propose a framework and methods to conduct comprehensive asthma triggers and risk factors analyses that will leverage existing knowledge regarding asthma and identify multiple types of asthma triggers and risk factors; have the ability to determine their relative importance; incorporate heterogeneous and autonomous data sources that contain useful information; and derive characteristics of self-reported asthma patients from novel data sources. We report on an empirical study and the results from analyzing 3 years of social media, environmental sensor, socioeconomic, and outpatient illness surveillance data from the United States. Figure 2 shows the proposed research framework and its comparison with retrospective cohort studies. Components of the framework are elaborated in the following subsections.

Heterogeneous and Open Data Sources

The data for this study are collected from four different sources: social media, environmental sensors, socioeconomic census, and outpatient illness surveillance. These datasets are collected for a period of 3 years from January 2013 to December 2016. In the following subsections, we describe the features extracted from these data sources and how they are used.

Social Media Data

Asthma-Related Twitter Stream Data: Twitter makes it convenient to collect data by providing APIs to access large volumes of user-filled text as well as automatically generated data (e.g., content creation time and geolocation). We collect a large dataset from Twitter using its streaming API to include tweets containing 1 or more of 18 asthma-related keywords (e.g., asthma, inhaler, wheezing, etc.) that are suggested by medical asthma specialists (Ram et al. 2015).

Self-Reported Asthma Patients' Twitter Data: Self-reported asthma patients (e.g., those who stated they had asthma attacks or asthma-related problems or report the use of asthma medication and/or inhalers) are identified from the asthma-related streaming datasets (described in a later section). For each individual, two types of data are collected: user profiles, from which we extract users' demographic information (Table 1 and Figure 3), and archived tweets (i.e., the entire set of tweets for each individual), from which we derive users' behavior information.

Environmental Sensor Data

Local air quality and weather conditions affect how people live and breathe. The effects of environmental factors on asthma exacerbations have consistently been identified in previous studies.

Local Weather Data: The daily average weather conditions, including temperature, dew point, humidity, winds, sky condition, weather type, atmospheric pressure, and more are collected from the U.S. National Centers for Environmental Information database.² This database provides a record of values for a specific weather station for each day (Menne et al. 2012). Each weather station has a unique identifier, WBAN (Weather Bureau Army Navy), with location (i.e., latitude and longitude).

Outdoor Air Quality Data: Air quality (or air pollution) data are collected from the U.S. Environmental Protection Agency databases.³ This database contains measures of six types of pollutants: particulate matter ($PM_{2.5}$ and PM_{10}), ground-level ozone (O_3), carbon monoxide (CO), sulfur oxides,⁴ nitrogen oxides,⁵ and lead (Pb). Indicators, such as air quality indexes (AQI) and concentration values, associated with these pollutants are collected. These data are available as hourly/daily average values. The higher the indicator value, the greater the level of air pollution and the greater the health concern (EPA 2014). Also, we are able to get the Air Quality System site ID. A site ID is associated with a specific location (with latitude and longitude).

²<https://www.ncdc.noaa.gov/data-access/quick-links>

³<https://www.epa.gov/outdoor-air-quality-data/download-daily-data>

⁴Sulfur oxides are a group of pollutants that contain both sulfur and oxygen molecules. Sulfur dioxide (SO_2) is the most common form in the lower atmosphere.

⁵Oxides of nitrogen are a mixture of gases that are composed of nitrogen and oxygen, including NO_3 , N_2O_3 , N_2O_4 , and N_2O_5 .

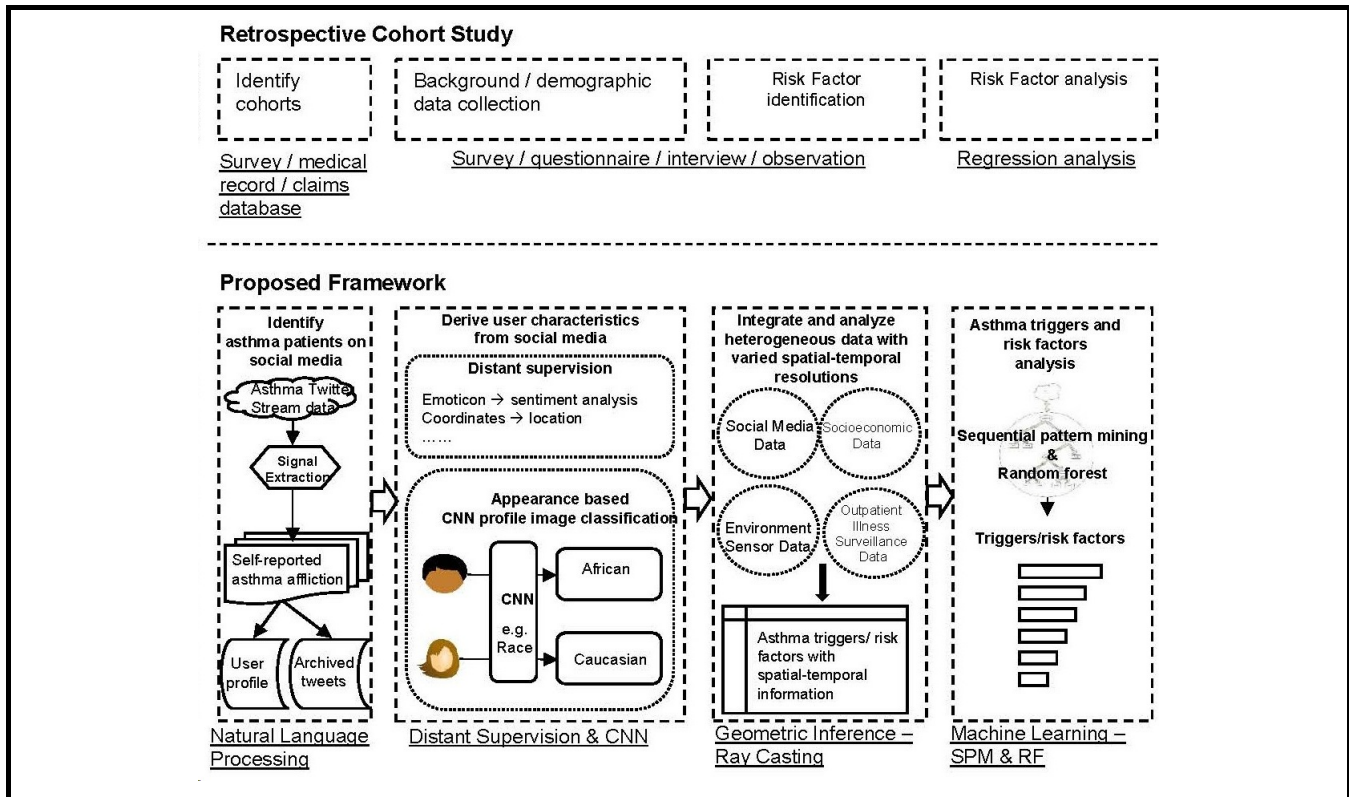


Figure 2. Asthma Trigger and Risk Factor Analyses Framework

Table 1. Twitter User Profile Fields* and Tweet Fields** Description

Profile Field	Type	Description
<i>id</i>	Integer	The unique identifier for a user
<i>name</i>	String	The name of the user, as they have defined it
<i>screen_name</i>	String	The screen name or handle with which users identify themselves
<i>url</i>	String	A URL provided by the user in association with their profile
<i>description</i>	String	The user-defined UTF-8 string describing their account
<i>location</i>	String	The user-defined location for this account's profile
<i>time_zone</i>	String	A string describing the time zone
<i>coordinates</i>	Float	The geographic location (longitude, latitude) reported by the user or client application
<i>profile_image_url</i>	String	A HTTP-based URL pointing to the user's profile image
<i>created_at</i>	String	UTC time when a Tweet was created
<i>text</i>	String	The UTF-8 text of the Tweet status update

*<https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object>

**<https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object>

Socioeconomic Census Data

A number of studies call our attention to the relationship between people's social position and asthma exacerbation. To

deal with asthma attacks, individuals should have logistical, financial, and cultural access to environments that encourage good asthma management practices. Hence, we also included a number of social economic data sources in our study.

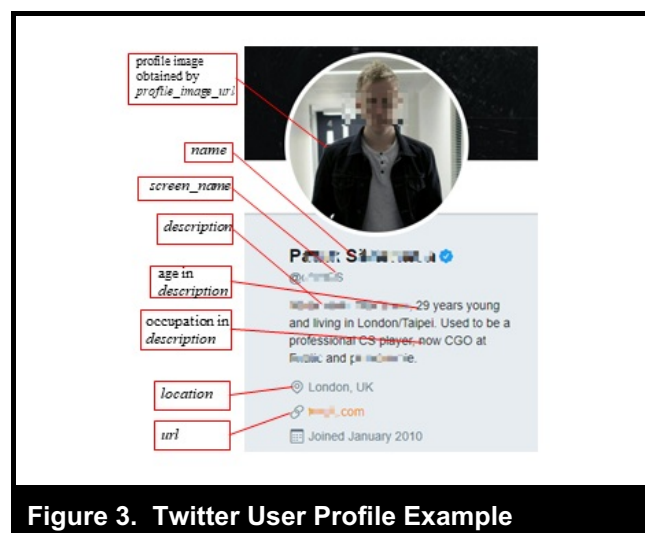


Figure 3. Twitter User Profile Example

Economic Research Service Data: U.S. Economic Research Service (ERS) is the principal agency that provides nationwide census information on various socioeconomic indicators. In this study, we use several vital social and economic factors, including unemployment rate, median household income, education levels, and poverty rate estimates. We collect these data from the ERS database,⁶ along with county-level rural/urban codes (with polygon coordinates).

County Health Rankings & Roadmap: County Health Rankings & Roadmap provides a summary of diverse health-related census data.⁷ We collect population-level data on health behaviors that may have a relationship with asthma exacerbation, including tobacco use, diet and exercise, alcohol and drug use, and others. We also identify critical healthcare indicators, which track accessibility to health services and the quality of healthcare, in a particular geographic region. Along with these data, the Federal Information Processing Standards (FIPS) codes, which uniquely identify counties in the United States (i.e., polygon coordinates), are also collected.

Outpatient Illness Surveillance Data

Rhinovirus and influenza virus infections can be more severe for asthmatic patients, even if their asthma is controlled by medication. Previous studies show that respiratory infections can trigger asthma exacerbations and worsen asthma symptoms.

⁶<https://www.ers.usda.gov/data-products/>

⁷<http://www.countyhealthrankings.org/>

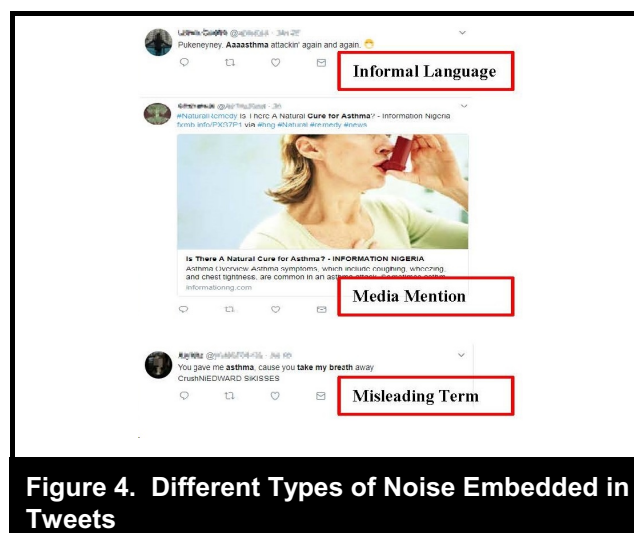


Figure 4. Different Types of Noise Embedded in Tweets

Influenza-Like Illness Surveillance Data: The U.S. Outpatient Influenza-Like Illness Surveillance Network⁸ collects and shares influenza-like illness (ILI, i.e., possible influenza or other illness causing a set of common symptoms) information on patient visits to healthcare providers. We are especially interested in the ILI activity level (i.e., the proportion of outpatient visits to healthcare providers) within a state. This data is provided on a weekly basis and state name is given for geolocation identification.

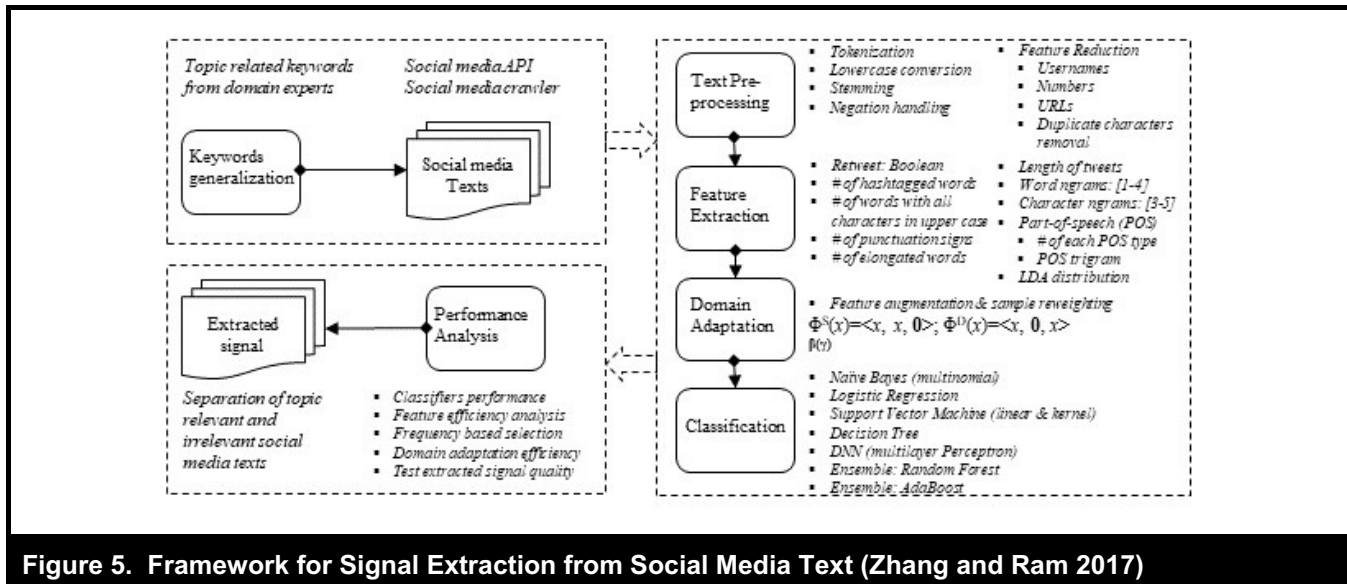
Weekly U.S. Influenza Surveillance Report: The CDC Weekly Influenza Surveillance Report⁹ provides an outline of the national and regional summary of select surveillance components. We collect the percentage of respiratory specimens of outpatients that are positive for flu in clinical laboratories. The collected data distinguish between influenza infections and other types of illness. This data is reported on a weekly basis and organized by 10 regions.

Identifying Self-Reported Asthma Patients on Social Media

As described in the previous section, we use a list of asthma-related keywords to generate a large, albeit noisy, asthma dataset from Twitter. There are three types of noise (see examples in Figure 4): (1) loosely structured informal languages, such as abbreviations, misspellings, punctuation errors, non-dictionary slang, and wordplay; (2) anomalous media spikes (people may include asthma-related terms in their posts,

⁸<https://gis.cdc.gov/grasp/fluview/main.html>

⁹<https://www.cdc.gov/flu/weekly/index.htm>



however, some of these may be from users who post asthma-related news; these do not reflect actual disease affliction); and (3) misleading terms and phrases (e.g., indicating awareness of asthma or using asthma as rhetoric, such as “Hope I won’t get asthma” or “I will call you asthma, because you take my breath away”—clearly about asthma but not about affliction). The first type of noise is challenging for machine interpretation. The latter two types of noise tend to overestimate population characteristics.

To identify potential asthma patients, a proper signal extraction process is essential for producing robust results because all of these types of noise need to be identified and removed. In previous research, Zhang and Ram (2015,2017) proposed a novel and efficient framework combining natural language processing, machine learning, and domain adaptation techniques to extract signal from social media text (Figure 5).

The proposed method was tested using several large real-world datasets from social media and outperforms other baseline methods by a large margin. After using this method, the correlation between aggregated asthma stream data and the actual U.S. adult asthma prevalence (CDC 2017a) was significantly improved from 30.3% to 69.2% (Table 2), which demonstrated the effectiveness of the framework. So it is reasonable to assume that the extracted signals do reflect *actual self-reported asthma affliction*. In this work, after applying this signal extraction process, we identify potential asthma patients from this cleaned dataset. The following criteria are developed to identify Twitter users as self-reported asthma patients:

- (1) Required condition: the tweets need to state that the individual has asthma or had an asthma attack recently. Supporting conditions: tweets may indicate severe difficulty in breathing as part of a discrete attack, shortness of breath, nighttime coughing with duration greater than one month; family history, or childhood history.
- (2) Required condition: the tweets indicate use of an inhaler. Supporting condition: tweets indicate no illicit drug use either explicitly stated or implied.

Deriving User Characteristics of Self-Reported Asthma Patients from Social Media

In this section, we describe methods to analyze social media data to automatically derive characteristics of potential asthma patients.

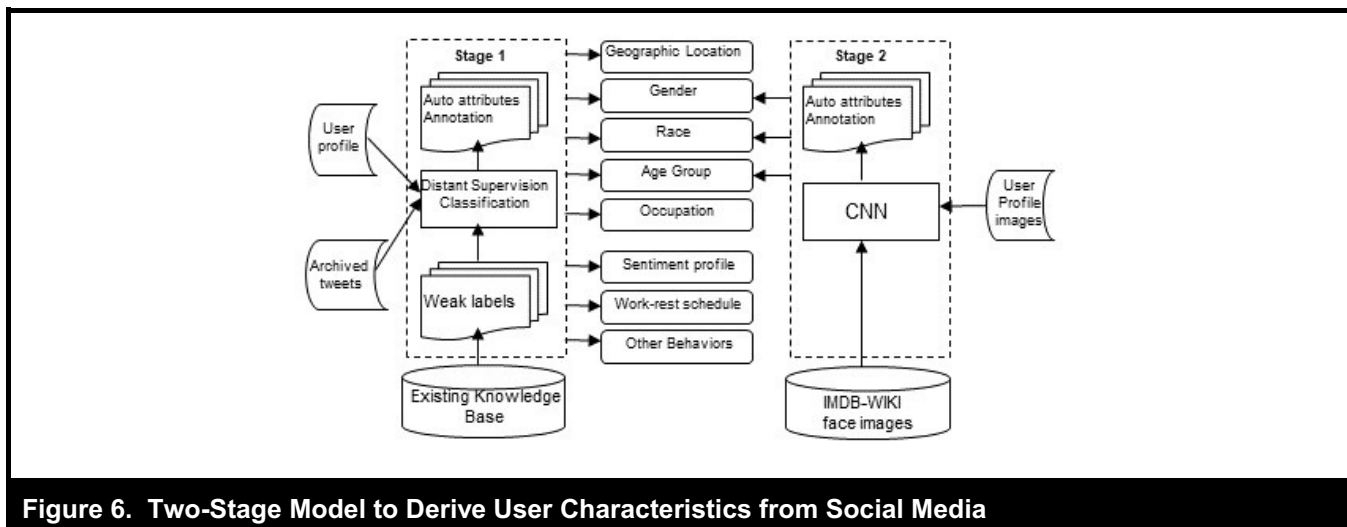
For many years, research studies have depended on traditional data collection methods (e.g., surveys) to obtain patients’ background to analyze asthma triggers and risk factors, making large-scale research prohibitively expensive and slow. Social media data, on the other hand, can reveal demographic characteristics and provide continuous signals of behavior. However, the effort associated with the collection, cleaning, and preparation of such data is not trivial. We propose an advanced two-stage classification model to automatically derive information that is vital for the asthma trigger and risk factor analyses (Figure 6). In stage one, we extract demographic attributes using weakly labeled data which makes our

Table 2. Correlation Between Asthma Prevalence and Twitter Asthma Dataset (Zhang and Ram 2017)

		<i>After Signal Extraction</i>	<i>Before Signal Extraction</i>
Asthma Prevalence 2013	Pearson Correlation	0.692**	0.303*
	Sig.	0	0.029
	N	52	52
Asthma Prevalence 2014	Pearson Correlation	0.701**	0.312*
	Sig.	0	0.028
	N	52	52

*Correlation is significant at the 0.05 level

**Correlation is significant at the 0.01 level



model scalable and efficient. In stage two, we use all available data and propose a convolutional neural network (CNN) based profile image classification method. This enhances our model's ability to identify gender, race and age group, by boosting the accuracy.

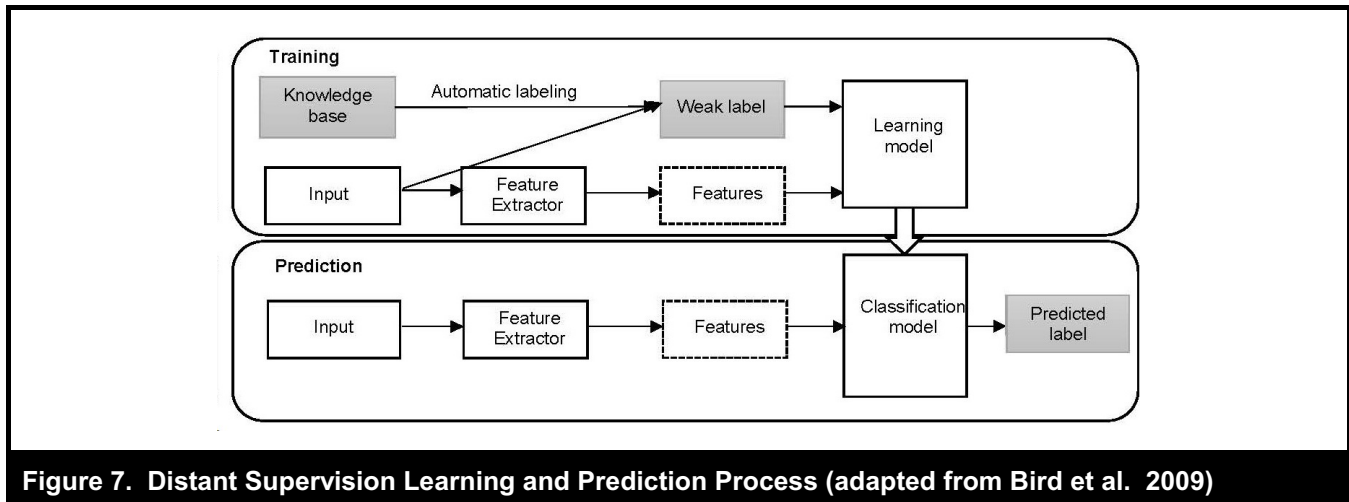
Stage 1: Distant Supervision

One of the major obstacles to extracting patients' characteristics from social media data is the lack of annotated datasets that provide ground truth for classification. The annotation process is extremely expensive, so the number of annotated samples tend to be very small compared to the number of data records in reality. In machine learning research, distant supervision is a subclass of semi-supervised learning where a classifier is trained on a weakly labeled (i.e., labels are incomplete or partially known) training dataset or existing knowledge base instead of manually labeled data (Craven and Kumlien 1999). It generates training data automatically by aligning existing knowledge and predicts labels for previously

unseen data. Although distant supervision models are trained in the presence of weak labels, they can make predictions with competitive accuracy compared to supervised learning. Most importantly, they have the ability to scale well for large datasets (Go et al. 2009).

In this study, we develop new distant supervision models for *social media user characteristics extraction*, where the models are learned using automatically labeled training sets (i.e., weak labels) based on biographical characteristics to avoid labor-intensive data annotation (Figure 7). We describe the knowledge base used, the weak label extracted and the learning target (i.e., label predicted for each characteristic of a social media user).

As stated previously, we collect two types of data for each identified individual asthma patient: *user profile* and *archived tweets*. We extract demographics from their social media user profile and behavioral information from their archived tweets.



The following demographic attributes (i.e., the learning target for each distant supervision task) are gleaned from user profile data:

- Geographic Location:** By using Nominatim¹⁰ as the knowledge base, the geographic location of each self-reported asthma patient can be identified via three Twitter user profile fields: *coordinates* (e.g., {-97.510, 35.465}), *location* (e.g., “San Francisco, CA”), and *time_zone*. For the latter two, there are considerable ambiguities (e.g., “I am somewhere on the earth”). Geolocation extraction was performed by using latitude and longitude mapping and distant supervision using state names or postal abbreviations for U.S. states (e.g., Arizona or AZ), county (e.g., Pima), and city names (e.g., Tucson) as weak labels. Users’ self-reported time zones were used to distinguish different cities with the same name.
- Race:** The genetic variants between racial groups may explain additional risk (e.g., the Hispanic population exhibited much higher asthma admission histories; Hunninghake et al. 2006). We employ users’ self-reported last name to infer racial groups (Wong et al. 2010). We collect weak labels from the 2000 U.S. census,¹² which summarizes the distribution of races for last names (Mislove et al. 2011). For example, the last name “Wood” was observed to correspond to Caucasian 90.06% of the time, African American 5.61%, Asian & Pacific Islander 0.51%, American Indian 0.82%, and Hispanic 1.60%. So the weak label of the race detection task for the last name “Wood” is assigned as “Caucasian.”
- Age group:** In this study, we classify users into two major demographic groups: below 20 or above 20. This binary classification is useful because studies show that asthma is more common among females than males after puberty (Almqvist et al. 2008). We depend on two Twitter user profile fields to discover their age groups: *description* and *url*. Many users include their age in the *description* field and linked their LinkedIn profiles or home pages (which may reveal age) in the *url* field (Rao et al. 2010).
- Occupation:** Certain occupational factors, such as chemicals used in farming and hairdressing, are asthma risk factors. We extract a user’s occupation by doing a fuzzy search (i.e., flexible matching) on the *description* field for occupations that are included in the U.S.

¹⁰<https://wiki.openstreetmap.org/wiki/Nominatim>

¹¹<https://www.ssa.gov/oact/babynames/>

¹²https://www.census.gov/topics/population/genealogy/data/2000_surnames.html

Occupational Employment Statistics list.¹³ The description field in some cases may contain more than one occupation. We only include the first one, assuming it most likely represents the primary appointment (Sloan et al. 2015).

Next, we derive behavioral attributes from the archived tweets of identified asthma patients.

- **Sentiment Profile:** Strong emotions and emotional reactions can be asthma risk factors. One's sentiment profile can be used to monitor stress level and emotional state (Liu et al. 2017). In this study, sentiment is defined as an individual's positive or negative feeling. We use emoticons as weak labels (e.g., positive: ☺, :D, etc.; negative: ☹, :-(, etc.). A sentiment profile is defined as one's collective sentiment each day: $s_i^d = \frac{1}{n} \sum_{archive} (t_p - t_n)$. For individual i , there are m tweets in day d , where t stands for a single tweet, p is positive and n is negative.
- **Work–Rest Schedule:** Irregular work–rest rhythms may exacerbate asthma. We examine users' work–rest schedules by recording their active periods on Twitter. The timestamp (i.e., *created_at* field) associated with each archived tweet is the natural weak label for this task. Activities during the time period 12:00 a.m. to 5:00 a.m. are recorded as an irregular work–rest status.
- **Other Behavioral Patterns:** Other behaviors, such as smoking, excessive drinking, taking medications, or physical exercise, are associated with asthma exacerbations. Related behavioral keywords are used as weak labels and information extraction techniques are used to detect such behaviors from the *text* field.

Stage 2: Appearance-Based CNN Profile Image Classification Model

In machine learning, convolutional neural network (CNN) is a class of deep neural networks (DNN) and is suitable for image analysis (Lecun et al. 1998). CNN uses special architectures (three basic ideas: local receptive fields, shared weights, and pooling) that are particularly well-adapted for image classification. Facial images (e.g., social media profile images) and CNN have been successfully combined for demographic attributes detection (Levi and Hassner 2015; Liu et al. 2015).

Our appearance-based CNN profile image classification model consists of three steps (Figure 8).

- **Profile image collection:** Over 90% of social media users display their own profile images publicly (Strano 2008). We collect users' Twitter profile images (400px × 400px) according to the Twitter user profile field *profile_image_url*.
- **Face alignment cascade classification** from OpenCV (Bradski 2000), an open source library with optimized algorithms for image analysis, is applied for facial recognition on profile images.
- **CNN image classification:** Adapting state-of-the-art techniques, we build a CNN-based profile image classification model. The network architecture is elaborated in Rothe et al. (2018). The model is trained on IMDB-WIKI face images dataset¹⁴ with age and gender labels (Rothe et al. 2018). Since the dataset is derived from celebrities' images on IMDb and Wikipedia, we further obtain race labels crawled from their profiles on Wikipedia when they are available.

A linear combination of Stage 1 and Stage 2 results is used for attribute classification, that is

$$\hat{y} = \arg \max_y \frac{\sum_{i=1}^2 \Pr_{stage(i)}}{2}$$

where P_r stands for probability and SVM classifier is employed for distant supervision classification. It does not provide probability estimates directly; probabilities are calculated using five-fold cross-validation. We use the following value sets for each attribute: gender = $y \in \{\text{male, female}\}$, age = $y \in \{\leq 20, > 20\}$, and race = $y \in \{\text{caucasian, african american, asian \& pacific islander, american indian, hispanic}\}$.

Repurposing, Integrating and Analyzing Heterogeneous Data with Varied Spatial–Temporal Resolutions

Due to the complexity of asthma triggers and risk factors, our methods require signals from multiple heterogeneous data sources and the capability to process data with a wide range of temporal (as is shown in Table 3; e.g., minute, hour, day, week, month, or year) and spatial resolutions (as is shown in Tables 3 and 4, including point and polygon coordinates). The following steps are taken to integrate and fuse these highly heterogeneous datasets.

- (1) Spatially, the point coordinates are interpolated to polygon coordinates at the *county level* by applying a

¹³https://www.bls.gov/oes/current/oes_stru.htm#00-0000

¹⁴<https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/>

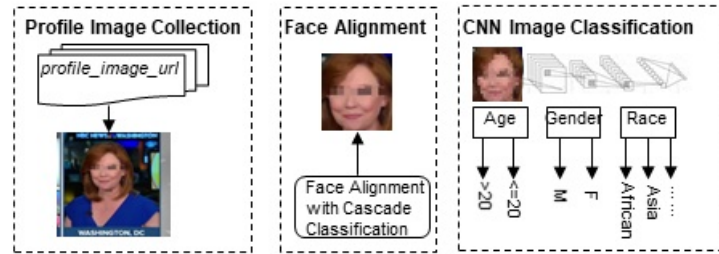


Figure 8. Appearance-Based Profile Image Classification

Table 3. Asthma Triggers/Risk Factors and Extracted Features

Factor	Data Type/Value	Features	Cat.	Data Source	Original Temporal Granularity	Original Spatial Granularity
Gender	M;F [†]	f_{gender}	1	Social Media User Profile		Point
Race	C;B;A;I;H [§]	f_{race}	1			Point
Age Group	≤ 20 ; > 20	f_{age}	1			Point
Occupation	{Occupations}	$f_{occupation}$	4			Point
Sentiment Profile	{0.0–1.0}	$f_{sentiment-x}$ where $x \in \{[high, middle, low]\}$	4	Social Media Archived Tweets	Minute	Point
Irregular Work-Rest	True; False	$f_{irregular}$	3		Minute	Point
Smoking	True; False	$f_{smoking}$	3		Minute	Point
Drinking	True; False	$f_{drinking}$	3		Minute	Point
Medication	True; False	$f_{medication}$	3		Minute	Point
Exercise	True; False	$f_{exercise}$	3		Minute	Point
Other Behaviors		$f_{beh-name}$	3		Minute	Point
Temperature	Floating	f_{term-x} where $x \in \{min, max, mean, 1, 3, 5, (+1), (+3), (+5) [high, middle, low]\}$	2	Local Weather Data	Hour	Point
Dew Point	Floating	$f_{dew-mean}$	2		Day	Point
Sky Condition	Enumerated	$f_{sky-condition-x}$ where $x \in \{see\ note^{\dagger}\}$	2		Day	Point
Humidity	Floating	$f_{humidity-x}$ where $x \in \{mean, [high, middle, low]\}$	2		Hour	Point
Snow Depth	Floating	f_{snow-x} where $x \in \{mean, [high, middle, low]\}$	2		Day	Point
Snow Fall	True; False	f_{snow}	2		Day	Point
Precipitation	Floating	$f_{precipitation}$ where $x \in \{mean, [high, middle, low]\}$	2		Day	Point
Precipitation Flag	True; False	f_{pre}	2		Day	Point
Temperature Departure	Floating	$f_{depart-x}$ where $x \in \{mean, [high, middle, low]\}$	2		Day	Point
Heat ($> 65^{\circ}\text{F}$)	Floating	f_{heat-x} where $x \in \{mean, [high, middle, low]\}$	2		Day	Point
Cool ($< 65^{\circ}\text{F}$)	Floating	f_{cool-x} where $x \in \{mean, [high, middle, low]\}$	2		Day	Point
Wind Speed	Floating	$f_{wind-speed-x}$ where $x \in \{mean, [high, middle, low]\}$	2		Day	Point
Direction of Wind	Enumerated	$f_{wind-direction}$	2		Day	Point
PM 2.5 AQI	Integer	$f_{aqi25-x}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Day	Point
PM 10 AQI	Integer	$f_{aqi10-x}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Day	Point

O ₃ AQI	Integer	$f_{aqi_{o_3-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2	Outdoor Air Quality Data	Day	Point
CO AQI	Integer	$f_{aqi_{co-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Day	Point
SO ₂ AQI	Integer	$f_{aqi_{so_2-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Day	Point
NO ₂ AQI	Integer	$f_{aqi_{no_2-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Day	Point
Pb AQI	Integer	$f_{aqi_{pb-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Day	Point
PM _{2.5} C	Floating	$f_{c_{25-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Hour	Point
PM ₁₀ C	Floating	$f_{c_{10-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Hour	Point
O ₃ C	Floating	$f_{c_{o_3-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Hour	Point
CO C	Floating	$f_{c_{co-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Hour	Point
SO ₂ C	Floating	$f_{c_{so_2-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Hour	Point
NO ₂ C	Floating	$f_{c_{no_2-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Hour	Point
Pb C	Floating	$f_{c_{pb-x}}$ where $x \in \{mean, (+1), (+3), (+5), [high, middle, low]\}$	2		Hour	Point
Unemployment Rate	Percentage	$f_{unemployment-x}$ where $x \in \{[high, middle, low]\}$	4	Economic Research Service Data	Year	Polygon
Median Household Income	Integer	f_{med_inc-x} where $x \in \{[high, middle, low]\}$	4		Year	Polygon
Education (> College)	Percentage	$f_{above-col-x}$ where $x \in \{[high, middle, low]\}$	4		Year	Polygon
Education (< High School)	Percentage	$f_{below-hig-x}$ where $x \in \{[high, middle, low]\}$	4		Year	Polygon
Poverty Rate	Percentage	$f_{poverty-x}$ where $x \in \{[high, middle, low]\}$	4		Year	Polygon
Tobacco Use	Percentage	$f_{tobacco-x}$ where $x \in \{[high, middle, low]\}$	3	County Health Rankings & Roadmap	Year	Polygon
Food Environment Index	Integer	$f_{food_environment_index-x}$ where $x \in \{[high, middle, low]\}$	3		Year	Polygon
Physical Inactivity	Percentage	$f_{physical_inactivity-x}$ where $x \in \{[high, middle, low]\}$	3		Year	Polygon
Alcohol Use	Percentage	$f_{alcohol-x}$ where $x \in \{[high, middle, low]\}$	3		Year	Polygon
Drug Use	Percentage	f_{drug-x} where $x \in \{[high, middle, low]\}$	3		Year	Polygon
Access to Health Services	Ratio	$f_{access_health_services-x}$ where $x \in \{[high, middle, low]\}$	4	Outpatient Illness Surveil- lance Data	Week	Polygon
Quality of Healthcare	Integer	$f_{quality_healthcare-x}$ where $x \in \{[high, middle, low]\}$	4		Week	Polygon
ILI Activity Level	Integer	$f_{ili_activity_level-x}$ where $x \in \{[1, 10]\}^*$	5		Week	Polygon
ILI Activity Level Label	Enumerated	$f_{ili_activity_label-x}$ where $x \in \{[insufficient data, minimal, low, moderate, high]\}^\#$	5	Outpatient Illness Surveil- lance Data	Week	Polygon
% Respiratory Specimens (Positive for Flu)	Percentage	$f_{flu_positive-x}$ where $x \in \{[high, middle, low]\}$	5		Week	Polygon

Factors: AQI = Air Quality Indexes; C = Concentration Values; ILI = Influenza-Like Illness.

Value: †: M = male; F = female.

[§]C = Caucasian; B = African American; A = Asian & Pacific Islander; I = American Indian; H = Hispanic.

[†]Sky Condition: FG = Fog, ice fog, or freezing fog (may include heavy fog); TS = Thunder; PL = Ice pellets, sleet, snow pellets or small hail; GR = Hail (may include small hail); GL = Glaze or rime; DU = Dust, volcanic ash, blowing dust, blowing sand or blowing obstruction; HZ = Smoke or haze; BLSN = Blowing or drifting snow; FC = Tornado, water spout or funnel cloud; WIND = High or damaging winds; BLPY = Blowing spray; BR = Mist; DZ = Drizzle; FZDZ = Freezing drizzle; RA = Rain; FZRA = Freezing rain; SN = Snow, snow pellets, snow grains or ice crystals; UP = Unknown precipitation; MIFG = Ground fog; FZFG = Ice fog or freezing fog.

[†]The 10 activity levels correspond to the number of standard deviations, *at or above* the mean for the current week compared with the mean of the non-influenza weeks.

[#]10 activity levels classified as minimal (levels 1–3), low (levels 4–5), moderate (levels 6–7), and high (levels 8–10).

Cat. = Category: 1 = Biological and demographic factors; 2 = Environmental factors; 3 = Behavioral factors; 4 = Social and psychological factors; 5 = Inflammatory and infectious factors.

geometric inference algorithm. Specifically, we employ a point-in-polygon method to ascertain if a discrete point lies within a particular polygon (Laurini and Thompson 1992). For instance, we may have a set of point coordinates representing the location of environmental sensors which need to be combined with socioeconomic census data that are available by areas (i.e., with polygon coordinates). The point-in-polygon method identifies which of the point coordinates fall into which polygon coordinates. To accomplish this, we adapt the ray casting algorithm (Shimrat 1962). In this research, we define a ray \vec{R} , let \vec{R} cast from the test point coordinates that serves as the origin of \vec{R} . Define $\vec{P}_i (i = 1, \dots, l)$ as the edges of the polygon (e.g., polygon coordinates of a county). Compute the intersections $n = \vec{R} \cap \vec{P}_i$, if $n = 2k + 1 (k \in \mathbb{N})$ then point-in-polygon is *true*. Figure 9 shows an example of the ray casting algorithm used in this study.

- (2) Temporally, all data are gathered from the same time period (2013–2016). All data records are *aggregated* (i.e., minutely and hourly data) or *decomposed* by day (all other data except minutely, hourly and daily data).
- (3) For real-valued factors, three features are extracted when feasible: f_{max} , f_{min} , and f_{mean} . We also compute the relative differences for weather factors, for example, f_{tem-3} indicates the absolute temperature difference between the average of the *current day* (a day with asthma exacerbation) and the average of *three days before the current day*. In addition, the cumulative mean of pollutants and temperature are also computed, for example, $f_{no_2-(3+)}$ represents the cumulative mean of NO_2 for three days. Besides the actual values, we also convert them from numerical to categorical values (high, middle and low) based on the first (Q1) and third quartiles (Q3), for example, $f_{tem-mean[high]}$ indicates that the average temperature experienced by an asthma patient on a particular day is above Q3 of the temperature (in the top 25th percentile of temperature range) for that patient's county.

In total, we extract 270 individual asthma triggers or risk factors, as shown in Table 3.

Asthma Trigger and Risk Factors Assessment with Relative Importance

In this section, we describe methods to discover the relationships among asthma risk factors and determine their relative importance.

First, we analyze features by adapting sequential pattern mining (SPM) to find statistically significant triggers and risk factors when they appear or occur in a chronological sequence. The identified sequential patterns are then input to a subsequent machine learning method to identify the relative importance of asthma triggers and risk factors (see Appendix B, Figure B1). SPM is one of the major machine learning approaches for detecting frequent subsequences in a set of data records, where all records are ordered events (Mabroukeh and Ezeife 2010). In this study, we adapt SPM for feature extraction. A feature is defined as a representative sequence of asthma triggers or risk factors that appear with a frequency no less than a specified threshold (Han et al. 2007). We use an efficient state-of-the-art method called PrefixSpan (Pei et al. 2001). One significant advantage of this method is that it works in a recursive and divide-and-conquer manner, making it amenable for parallel implementation.

This analysis is important because asthma patients may be exposed to multiple triggers or risk factors in a specific chronological order, thus making it necessary to investigate the sequential effects of risk factors on asthma exacerbation. Here we explain how we adapt SPM to our research setting. We identify an individual asthma patient as P_i . An asthma trigger or risk factor on day d is defined as f_t^d , where t is a feature type, for example $f_{temperature}$ indicates the feature: *temperature*. Let $F_{P_i} = \{f_t^{d_1}, f_t^{d_2}, \dots, f_t^{d_n}\}$ be the set of all risk factors associated with patient P_i , where $d_1 \leq d_2 \leq \dots \leq d_n$ indicates a chronological order of the feature values. A *sequence database* δ can be defined as $\delta = \{F_{P_1}, F_{P_2}, \dots, F_{P_m}\}$. A *factor-set* is a subset of F_{P_1} . A *sequence* α is an ordered list of a factor-set. Given a positive number ζ as the *support threshold*, a sequence α is defined as a *frequent sequential pattern* in the sequence database δ . Given a sequence database and a *min_support* threshold ζ_{min} , the problem of *sequential pattern mining* in our current research setting is to find the complete set of frequent sequential patterns $\{\alpha\}$ in the database δ .

We use random forest (RF) (Breiman 2001) to measure feature importance based on the decrease in average feature impurity (see Appendix B for more detailed explanation). RF, an ensemble-based machine learning method, is one of the most powerful algorithms for classification. RF fits a number of decision tree classifiers (Han et al. 2011) on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It can combine weak and strong learners, provide improved performance, and results in better scalability using fewer hyper-parameters (i.e., values are set before the learning process begins). It also provides a natural way to measure feature importance as the

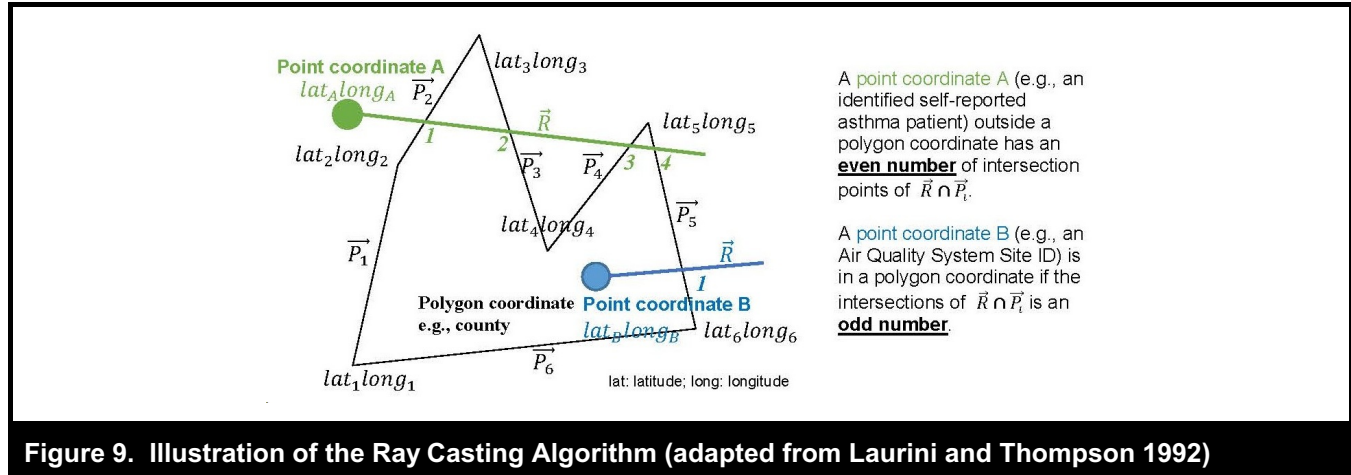


Figure 9. Illustration of the Ray Casting Algorithm (adapted from Laurini and Thompson 1992)

average feature impurity decreases (Strobl et al. 2007). There are two reasons why we use RF in this study:

- (1) As shown in Table 3, asthma risk factors have a large range of values of different data types. As a tree-based ensemble algorithm, RF is useful for integrating different data types and does not mandate any feature scaling.
- (2) Certain features may be closely related to each other, for example, the increase in precipitation (i.e., $f_{precipitation}$) generally increases relative humidity levels (i.e., $f_{humidity}$). Also, the feature space may be not linearly separable. RF makes no assumptions regarding linearity. Following the supervised learning paradigm, we define the problem as a probabilistic classification problem $P_r(Y|X)$ (P_r stands for probability). X denotes the input space (including individual features extracted from heterogeneous data sources (see Table 3) and sequential features obtained from SPM (see Table 6)), with the output space being $Y = \{-1: \text{no asthma affliction}, 1: \text{asthma affliction}\}$. For analysis, we collect an equal number of Twitter users who never reported an asthma affliction to balance the data set. We first define n tree classifiers, for the k^{th} tree, a bootstrap sample of X is generated as X_{Θ_k} , resulting in a classifier $P_r(Y|X_{\Theta_k})$. The classification result of RF is obtained by averaging the probabilistic prediction of all n trees as $\hat{y} = \arg \max_y \frac{1}{n} \sum_{k=1}^n P_r(Y|X_{\Theta_k})$, the predicted class is the one with the highest probability. The training process provides useful internal estimates of feature importance as the average feature impurity (i.e., tree node impurity) decreases. In this study, we adopt a widely used impurity measure, the Gini index: $I = 1 - \sum_{i=1}^{(-1,1)} p(i|t)^2$, where $p(i|t)$ is the proportion of samples that belong to one class for a particular node t in a tree. The Gini index is used for the calculation of splits during RF training. The Gini impurity for the descendent nodes in a tree is

always less than the parent node. A tree node with higher impurity indicates a node with higher importance.

Implementation and Evaluation

In the design science paradigm, the evaluation of an artifact provides feedback information and a better understanding of the problem in order to improve both the quality of the design product and the design process. Our evaluation plan and procedures are summarized in Figure 10.

In this study, two main parts need to be evaluated.

- (1) The two-stage model proposed for deriving user characteristics of self-reported asthma patients from social media. We collect annotated data from two researchers as the ground truth for each Twitter users' characteristics (i.e., gender, race, age group, occupation, sentiment, and behavior). We then use Cohen's Kappa coefficient (Cohen 1960) to measure the inter-rater agreement between human annotators and the difficulty of the classification tasks. Next, we compare our proposed two-stage model with state-of-the-art baseline methods by applying K-fold cross-validation.
- (2) Sequential pattern mining and random forest for extracting asthma triggers/risk factors, their interconnecting relations and their relative importance. First, we work with asthma specialists to evaluate the meaningfulness of extracted sequential patterns (i.e., results of SPM). Second, the classification results of random forest for comparing asthma versus non-asthma patients were provided to these experts to evaluate if the relative importance of asthma triggers and risk factors are meaningful.

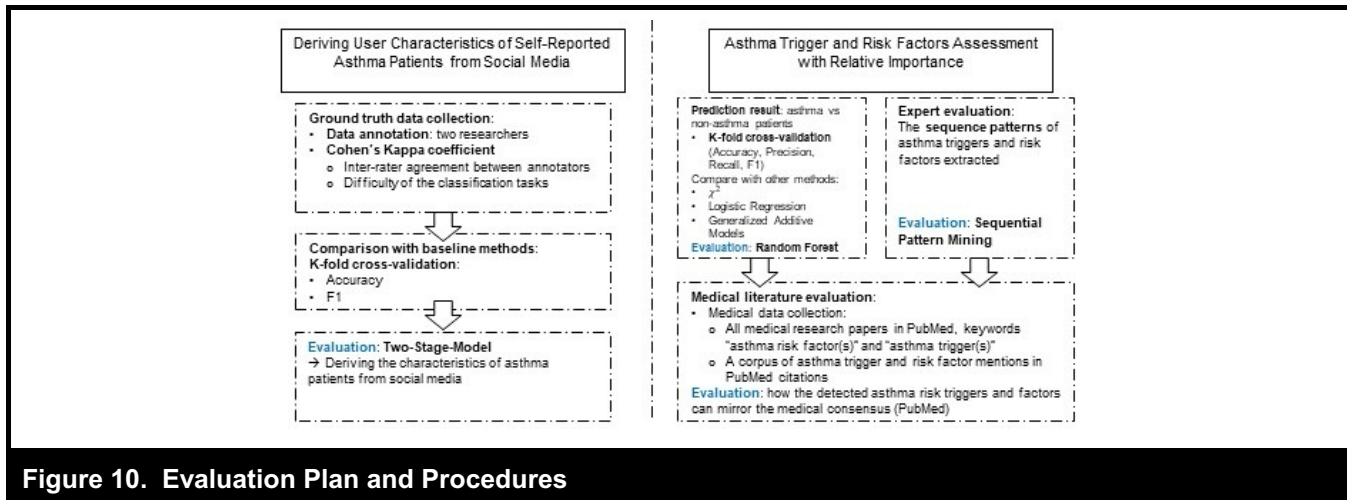


Figure 10. Evaluation Plan and Procedures

Third, medical articles (PubMed) are used to evaluate how the detected asthma risk triggers and factors can mirror the medical consensus.

We emphasize two techniques used in the evaluation process.

- (1) K-fold cross-validation is a well-known model evaluation technique. We first apply the k-fold cross-validation to evaluate the performance of proposed two-stage model for social media user attributes extraction. We also use k-fold cross-validation to evaluate the random forest classification results. In k-fold cross-validation, the original training dataset is randomly split into k equal sized subsamples without replacement. A single subsample is retained as the testing set, and the $k-1$ subsamples are used for model training. The process is then repeated k times (i.e., folds). The final estimation can be made based on averaged testing performance. The main reasons for adopting this validation technique are that it will not lose model significance and testing capability when the available annotated dataset is relatively small, and it is able to estimate how the performance of the model can be generalized to an independent dataset.
- (2) Medical literature-based evaluation. We first examine whether the extracted asthma risk factors agree with the general medical consensus on asthma trigger and risk factors in research articles from PubMed.¹⁵ This is accomplished by collecting a corpus of asthma trigger and risk factor mentions in PubMed citations. We also inves-

tigate if there are risk factors that our study identified, which have not received adequate attention in the biomedical literature, and evaluate if these identified triggers and risk factors and their relative importance extends medical knowledge.

Analysis, Results, and Discussion

In this section, we discuss the implementation process, the evaluation results of our methods, and the interpretation of the results.

Data Collection and Processing

We collected large datasets from social media, environmental sensors, socioeconomic census, and outpatient illness surveillance data for a three-year period from 2013 to 2016 (Table 4).

We applied the signal extraction process, described previously, to the social media streaming dataset to identify potential asthma patients. We were able to identify 9,096 potential asthma patients in the United States who self-reported asthma affliction. Moreover, after the data cleaning process, all datasets were integrated temporally and spatially (using methods we described previously in the "Research Design" section).

Characteristics of Potential Asthma Patients from Social Media

As described earlier, we relied on three types of signals to infer the characteristics of potential asthma patients from

¹⁵ PubMed is a service of the U.S. National Library of Medicine that provides free access to 27 million citations and abstracts for biomedical literature from MEDLINE, life science journals, and online books (www.ncbi.nlm.nih.gov/pubmed/).

Table 4. Datasets Description

<i>Dataset</i>	<i>Number of Records</i>		<i>Collection Period</i>	<i>Geo Area</i>	<i>Comments</i>	<i>Coordinate Type</i>
Asthma-related Twitter Stream Data	17,175,642		11/1/2013~12/30/2016	Worldwide	Twitter streaming API. Using 18 asthma-related keyword.	Point coordinates
Potential Asthma Patients' Twitter Archives	28,197,612			U.S.	9,096 self-reported asthma patients in the U.S. were identified.	Point coordinates
Local Weather Data	2,496,157		1/1/2013~12/30/2016	U.S.	Collected along with WBAN.	Point coordinates
Outdoor Air Quality Data	PM2.5	1,546,319	1/1/2013~12/30/2016	U.S.	Collected along with AQS-SITE-ID.	Point coordinates
	PM10	473,543				
	O ₃	1,450,491				
	CO	550,687				
	SO ₂	590,148				
	NO ₂	554,097				
	Pb	71,850				
Socio-Economic Census Data	Education	3,283	2013~2016	U.S.	Collected along with rural/urban codes code.	Polygon coordinates
	Poverty	3,194				
	Unemployment	3,274				
	Population	3,273				
County Health Rankings & Roadmap	12,548		2013~2016	U.S.	Collected along with FIPS code.	Polygon coordinates
Influenza-like illness surveillance data	14,026		2013~2016	U.S.	Collected along with state name.	Polygon coordinates
Weekly U.S. Influenza Surveillance Report	14,500		2013~2016	U.S.	Collected along with region number.*	Polygon coordinates

*Region classifications¹⁶

Region 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont

Region 2: New Jersey, New York, Puerto Rico, and the U.S. Virgin Islands

Region 3: Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia

Region 4: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee

Region 5: Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin

Region 6: Arkansas, Louisiana, New Mexico, Oklahoma, and Texas

Region 7: Iowa, Kansas, Missouri, and Nebraska

Region 8: Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming

Region 9: Arizona, California, Hawaii, and Nevada

Region 10: Alaska, Idaho, Oregon, and Washington

¹⁶<https://www.cdc.gov/flu/weekly/overview.htm>

social media data, namely social media user profiles, profile images, and archived tweets. To evaluate the performance of our two-stage model for social media users' characteristics extraction, two researchers were invited to annotate 1,200 Twitter users' attributes (e.g., age group, gender, race, etc.) based on their profiles and 1,500 tweets (i.e., tweets including users' sentiment and various behaviors). We used Cohen's Kappa coefficient¹⁷ to measure the inter-rater agreement between two human annotators (Kappa has a range from 0 to 1, with larger values indicating better reliability), and evaluate the difficulty of the classification tasks (as machine learning models are not likely to surpass human-level performance in this research setting and the level of disagreement between human annotators indicates the difficulty of the classification tasks).

Table 5 shows the annotation statistics. According to the interpretation of Landis and Koch (1977), Kappa results range from 0 to 1. The higher the value of Kappa, the stronger the agreement between annotators (< 0.20 slight agreement; 0.21–0.40 fair agreement; 0.41–0.60 moderate agreement; 0.61–0.80 substantial agreement; 0.81–1.00 perfect or almost perfect agreement). A Kappa coefficient above 0.2 is generally regarded as acceptable (Landis and Koch 1977). We observe that the Kappa coefficients of the multi-label classification tasks (i.e., race, Kappa = 0.37; occupation, Kappa = 0.22; and behavior, Kappa = 0.35) are consistently lower than the binary classification tasks (i.e., gender, Kappa = 0.73; age group, Kappa = 0.60; and sentiment, Kappa = 0.76), suggesting they may be challenging for both human and machine interpretations. In particular, occupation detection is a difficult task even for human annotators because of a lot of missing values.

We then tested our two-stage model on the annotated datasets. In stage 1, the widely used SVM classifier (Cortes and Vapnik 1995) was employed for distant supervision classification. For stage 2, the appearance based CNN profile image classification model was trained on the IMDB-WIKI dataset. We adopt accuracy and F1-score as our evaluation metrics. Accuracy is a description of model errors, defined as $accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n I(y, \hat{y})$, where $I(x)$ is the indicator function having a value of 1 if the predicted label \hat{y} matches with the true label y . The F1 score is defined as the harmonic mean of precision and recall $F1 = \frac{2PR}{P+R}$, where precision (P) is a measure of result relevance, while recall (R) is a measure of how many truly relevant results are returned. Ten-fold cross-validation is conducted and the results are presented in Figure 11.

¹⁷ $kappa = \frac{p_0 - p_e}{1 - p_e}$, where p_0 is the relative observed agreement among annotators (i.e., accuracy) and p_e is the hypothetical probability of chance agreement (Cohen 1960).

Two baselines, random choice and state-of-the-art performance, were adopted from the literature (Chen et al. 2015; Go et al. 2009; Sloan et al. 2015). The results show that our proposed model produces results that are comparable or outperform the state-of-the-art methods.

After analyzing the results in detail, we observed that social media users' attributes are often not explicitly provided by users through their profiles which leads to a less favorable result in terms of accuracy and F1 scores in stage 1 of the model, which relied on distant supervision classification with weak labels. We believe the reasons for this are threefold: (1) anonymity and pseudonymity: social media user profiles may disconnect from real-world identities despite what previous studies have indicated; (2) missing data: there are many missing values in some user profile fields (e.g., occupation); (3) weak labels may lack predictive power for certain tasks (e.g., last name as weak labels for race detection). Hence, incorporating profile images is a way to take advantage of all useful and available information. By combining the stage 1 (distant supervision) and stage 2 (CNN) of the two-stage model, the age group, gender, and race detection accuracy improved significantly, which suggests that the profile images are good complementary signals for detecting social media users' attributes. We show a subset of the demographic information extracted from social media data (Figure 12).

Asthma Risk Factors Assessment

Asthma triggers and risk factors assessment, as described in the previous sections, consists of two parts: SPM for uncovering the sequential patterns of triggers and risk factors contributing to asthma exacerbations as features, and RF for measuring the relative importance of asthma risk factors.

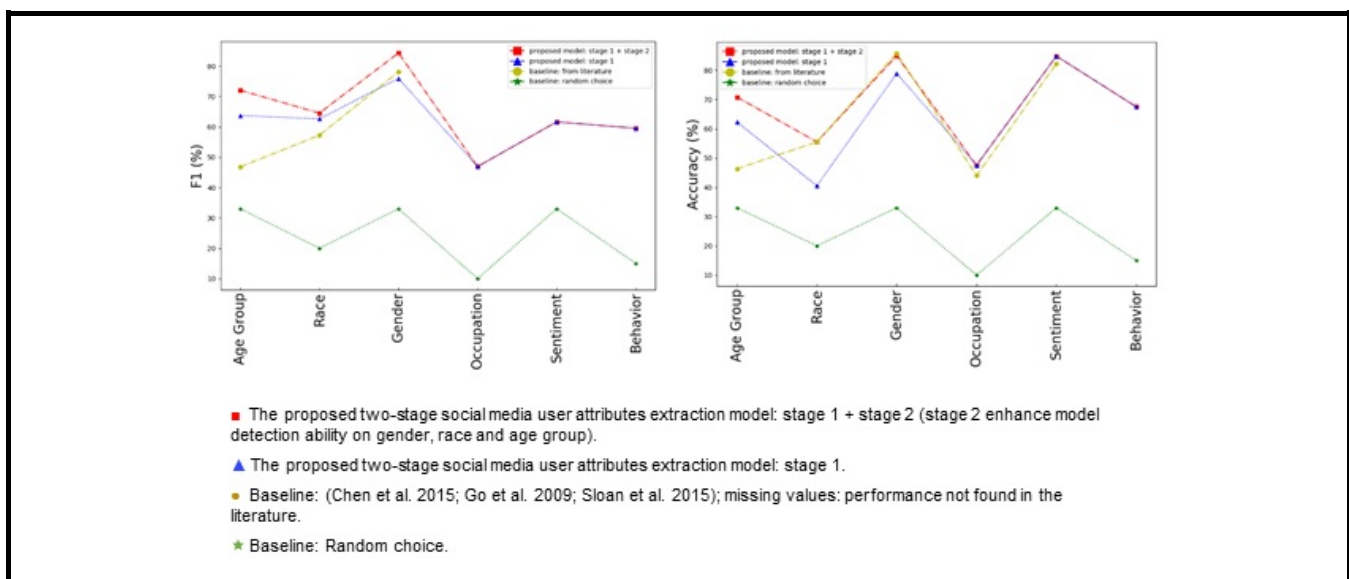
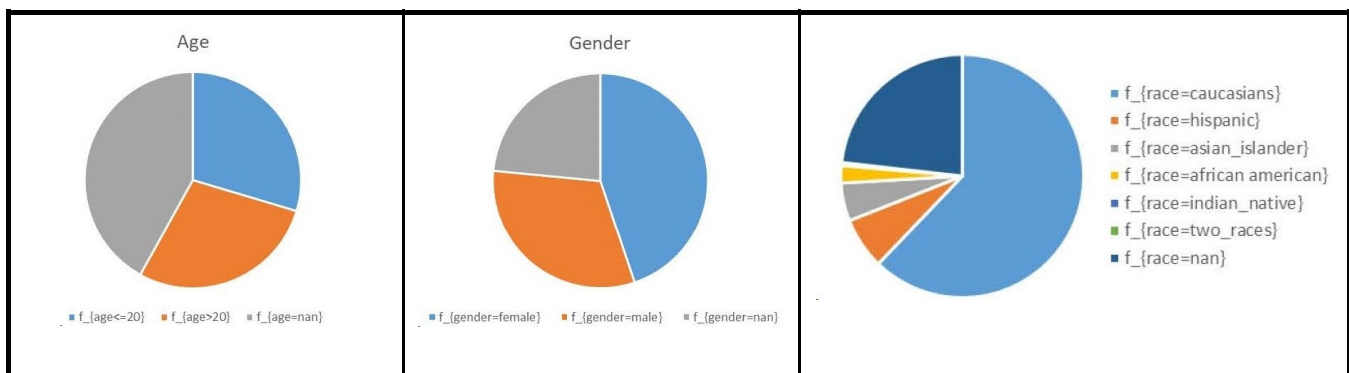
We first present the results of SPM regarding frequent sequential patterns of asthma risk factors. Two hyperparameters (i.e., values are set before the learning process begins) need to be specified: (1) The $\min_support$ threshold ξ_{min} for the PrefixSpan algorithm. We set $\xi_{min} = 0.05$ to ensure that we do not miss interesting patterns when the discovered features are frequent; (2) the observation interval λ , which represents the number of days before a self-reported asthma exacerbation episode. Based on suggestions from asthma specialists and also considering the length of time risk factors tend to linger, we set $\lambda = 5$ for environmental factors, $\lambda = 90$ for behavioral and psychological factors, $\lambda = 365$ for social factors, and $\lambda = \infty$ for biological factors. The results reveal a number of interesting patterns. All the results were evaluated and validated by asthma specialists who worked with us to interpret the results. In all, 88 sequential patterns were selected as features for further analysis (Table 6). Our asthma

Table 5. Annotation Statistics (Annotated Dataset for Two-Stage Model Evaluation)

Attributes	Gender	Race	Age Group	Occupation	Sentiment	Behavior
Annotations	M; F; nan	C;B;A;I;H; nan	≤20; >20; nan	Multi-labels (e.g., Chemists, Barbers, etc.)	Positive; Negative; nan	Multi-labels (e.g., drinking, smoking, etc.)
Kappa	0.73	0.37	0.60	0.22	0.76	0.35
Annotation Object	Twitter user profile & profile images			Tweets		
Number of Records	1200			1500		

Kappa value interpretation: < 0.20 slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; 0.81–1.00 almost perfect (Landis and Koch 1977).

Annotations: M = Male; F = Female; C = Caucasian; B = African America; A = Asian & Pacific Islander; I = American Indian; H = Hispanic; nan = Unknown or not available.

**Figure 11. Two-Stage Social Media Based User Attributes Extraction Model****Figure 12. Demographic Information Extracted from Social Media Data**

nan: Unknown or not available.

Table 6. Frequent Sequential Patterns Extracted

	Frequent Sequential Patterns Extracted	Cat.	Explanation
1	$f_{\{above_college=high\}}, f_{\{cat_alcohol_impaired=high\}}$	3 & 4	Percent of adults with a bachelor's degree or higher is high (> Q3); Alcohol-impaired deaths is high (> Q3)
2	$f_{\{age>20\}}, f_{\{gender=female\}}$	1	Age > 20; Gender is female
3	$f_{\{cat_alcohol_impaired=middle\}}, f_{\{exe=T\}}$	3	Alcohol-impaired deaths is middle (> Q1); Take exercise in the previous 90 days
4	$f_{\{cat_alcohol_impaired=middle\}}, f_{\{med=T\}}$	3	Alcohol-impaired deaths is middle (> Q1); Take medicine in the previous 90 days
5	$f_{\{cat_excessive_drinking=middle\}}, f_{\{cat_une_rate=middle\}}$	3 & 4	Percentage of adults reporting heavy drinking is middle (> Q1); Unemployment rate is middle (> Q1)
6	$f_{\{cat_excessive_drinking=middle\}}, f_{\{depart_mean=middle\}}$	2 & 3	Percentage of adults reporting heavy drinking is middle (> Q1); Average temperature's departure from normal temperature is middle (> Q1)
7	$f_{\{cat_excessive_drinking=middle\}}, f_{\{humidity_mean=high\}}$	2 & 3	Percentage of adults reporting heavy drinking is middle (> Q1); Humidity level is high (> Q3)
8	$f_{\{cat_excessive_drinking=middle\}}, f_{\{ili_act_label=minimal\}}$	3 & 5	Percentage of adults reporting heavy drinking is middle (> Q1); Influenza-like illness activity level above the mean by 1-3 deviations
9	$f_{\{cat_excessive_drinking=middle\}}, f_{\{sen=low\}}$	3 & 4	Percentage of adults reporting heavy drinking is middle (> Q1); Sentiment level is low (< Q1)
10	$f_{\{cat_food_environment_index=middle\}}, f_{\{cat_alcohol_impaired=middle\}}$	3	Food environment index is middle (< Q3); Alcohol-impaired deaths is middle (> Q1)
11	$f_{\{cat_physical_inactivity=low\}}, f_{\{cat_alcohol_impaired=middle\}}$	3	Percentage of adults reporting no leisure-time physical activity is low (< Q1); Alcohol-impaired deaths is middle (> Q1)
12	$f_{\{cat_physical_inactivity=low\}}, f_{\{cat_med_inc=high\}}$	3 & 4	Percentage of adults reporting leisure-time physical activity is low (< Q1); Median household income is high (> Q3)
13	$f_{\{cat_physical_inactivity=low\}}, f_{\{humidity_mean=high\}}$	2 & 3	Percentage of adults reporting leisure-time physical activity is low (< Q1); Humidity level is high (> Q3)
14	$f_{\{cat_physical_inactivity=low\}}, f_{\{humidity_mean=low\}}$	2 & 3	Percentage of adults reporting leisure-time physical activity is low (< Q1); Humidity level is low (< Q1)
15	$f_{\{cat_physical_inactivity=low\}}, f_{\{ili_act_label=minimal\}}$	2	Percentage of adults reporting leisure-time physical activity is low (< Q1); Influenza-like illness activity level above the mean by 1-3 deviations
16	$f_{\{depart_mean=middle\}}, f_{\{cool_mean=middle\}}$	2	Average temperature's departure from normal temperature is middle (> Q1); Cooling degree is middle (> Q1)
17	$f_{\{depart_mean=middle\}}, f_{\{heat_mean=middle\}}$	2	Average temperature's departure from normal temperature is middle (> Q1); Heating degree is middle (> Q1)
18	$f_{\{dew_mean=low\}}, f_{\{cool_mean=middle\}}, f_{\{daily_aqi_pm25_+1=high\}}$	2	Dew point temperature is low (< Q1); Cooling degree is middle (> Q1); Daily AQI pm25 previous 1 day is high (> Q3)
19	$f_{\{exe=T\}}, f_{\{med=T\}}$	3	Take exercise in the previous 90 days; Take medicine in the previous 90 days
20	$f_{\{exe=T\}}, f_{\{med=T\}}, f_{\{sen=low\}}$	3 & 4	Take exercise in the previous 90 days; Take medicine in the previous 90 days; Sentiment level is low (< Q1)
21	$f_{\{humidity_mean=high\}}, f_{\{concentration_pm25=high\}}$	2	Humidity level is high (> Q3); Daily concentration pm25 is high (> Q3)
22	$f_{\{humidity_mean=high\}}, f_{\{concentration_pm25_+5=high\}}$	2	Humidity level is high (> Q3); Daily concentration pm25 for previous 5 days are high (> Q3)
23	$f_{\{humidity_mean=high\}}, f_{\{cool_mean=middle\}}$	2	Humidity level is high (> Q3); Cooling degree is middle (> Q1)
24	$f_{\{humidity_mean=high\}}, f_{\{depart_mean=middle\}}$	2	Humidity level is high (> Q3); Average temperature's departure from normal temperature is middle (> Q1)

25	$f_{\{humidity_mean=high\}}, f_{\{heat_mean=middle\}}$	2	Humidity level is high (> Q3); Heating degree is middle (> Q1)
26	$f_{\{humidity_mean=high\}}, f_{\{tem_mean_+5=high\}}$	2	Humidity level is high (> Q3); Average temperature for previous 5 days are high (> Q3)
27	$f_{\{humidity_mean=low\}}, f_{\{concentration_pm25=high\}}$	2	Humidity level is low (< Q1); Daily concentration pm25 is high (> Q3)
28	$f_{\{humidity_mean=low\}}, f_{\{concentration_pm25_+5=high\}}$	2	Humidity level is low (< Q1); Daily concentration pm25 for previous 5 days are high (> Q3)
29	$f_{\{humidity_mean=low\}}, f_{\{cool_mean=middle\}}$	2	Humidity level is low (< Q1); Cooling degree is middle (> Q1)
30	$f_{\{humidity_mean=low\}}, f_{\{daily_aqi_pm25_+1=high\}}$	2	Humidity level is low (< Q1); Daily AQI pm25 previous 1 day is high (> Q3)
31	$f_{\{humidity_mean=low\}}, f_{\{depart_mean=middle\}}$	2	Humidity level is low (< Q1); Average temperature's departure from normal temperature is middle (> Q1)
32	$f_{\{humidity_mean=low\}}, f_{\{heat_mean=middle\}}$	2	Humidity level is low (< Q1); Heating degree is middle (> Q1)
33	$f_{\{humidity_mean=low\}}, f_{\{seafood=T\}}$	2 & 3	Humidity level is low (< Q1); Take seafood in the previous 90 days;
34	$f_{\{humidity_mean=low\}}, f_{\{tem_max_+5=low\}}$	2	Humidity level is low (< Q1); Maximum temperature for previous 5 days are low (< Q1)
35	$f_{\{humidity_mean=low\}}, f_{\{wind_speed=high\}}$	2	Humidity level is low (< Q1); Wind speed is high (> Q3)
36	$f_{\{ili_act_label=minimal\}}, f_{\{cat_pov_all=middle\}}, f_{\{cat_pov_017=middle\}}$	3 & 5	Influenza-like illness activity level above the mean by 1-3 deviations; Estimated percent of people of all ages in poverty is middle (> Q1); Estimated percent of people age 0-17 in poverty is middle (> Q1)
37	$f_{\{ili_act_label=minimal\}}, f_{\{cat_pov_all=middle\}}, f_{\{heat_mean=middle\}}$	2 & 3 & 5	Influenza-like illness activity level above the mean by 1-3 deviations; Estimated percent of people of all ages in poverty is middle (> Q1); Heating degree is middle (> Q1)
38	$f_{\{ili_act_label=minimal\}}, f_{\{cool_mean=middle\}}$	2 & 5	Influenza-like illness activity level above the mean by 1-3 deviations; Cooling degree is middle (> Q1)
39	$f_{\{ili_act_label=minimal\}}, f_{\{cool_mean=high\}}$	2 & 5	Influenza-like illness activity level above the mean by 1-3 deviations; Cooling degree is high (> Q3)
40	$f_{\{ili_act_label=minimal\}}, f_{\{exe=T\}}$	3 & 5	Influenza-like illness activity level above the mean by 1-3 deviations; Take exercise in the previous 90 days
41	$f_{\{ili_act_label=minimal\}}, f_{\{humidity_mean=high\}}$	2 & 5	Influenza-like illness activity level above the mean by 1-3 deviations; Humidity level is high (> Q3)
42	$f_{\{ili_act_label=minimal\}}, f_{\{med=T\}}, f_{\{sen=low\}}$	3 & 4 & 5	Influenza-like illness activity level above the mean by 1-3 deviations; Take medicine in the previous 90 days; Sentiment level is low (< Q1)
43	$f_{\{ili_act_label=minimal\}}, f_{\{positive_flu=middle\}}$	5	Influenza-like illness activity level above the mean by 1-3 deviations; Percentage of respiratory specimens positive for flu is middle (> Q1)
44	$f_{\{ili_act_label=minimal\}}, f_{\{tem_min_+5=middle\}}$	2 & 5	Influenza-like illness activity level above the mean by 1-3 deviations; Minimum temperature for previous 5 days are low (< Q1)
45	$f_{\{med=T\}}, f_{\{heat_mean=middle\}}$	2 & 3	Take medicine in the previous 90 days; Cooling degree is middle (> Q1)
46	$f_{\{med=T\}}, f_{\{humidity_mean=low\}}$	2 & 3	Take medicine in the previous 90 days; Humidity level is low (< Q1)
47	$f_{\{race=hispanics\}}, f_{\{cat_med_inc=low\}}$	1 & 4	Race is Hispanic; Median household income is low (< Q1)
48	$f_{\{sen=low\}}, f_{\{cool_mean=middle\}}$	2 & 4	Sentiment level is low (< Q1); Cooling degree is middle (> Q1)
49	$f_{\{sen=low\}}, f_{\{depart_mean=middle\}}$	2 & 4	Sentiment level is low (< Q1); Average temperature's departure from normal temperature is middle (> Q1)
50	$f_{\{sen=low\}}, f_{\{humidity_mean=low\}}$	2 & 4	Sentiment level is low (< Q1); Humidity level is low (< Q1)
51	$f_{\{sky_condition=BR\}}, f_{\{daily_aqi_pm25_+1=high\}}$	2	Sky condition is Mist; Daily AQI pm25 previous 1 day is high (> Q3)

52	$f_{\{sky_condition=RA\ BR\}},$ $f_{\{daily_aqi_pm25_+3=high\}}$	2	Sky condition is Rain & Mist; Daily AQI pm25 previous 3 days are high (> Q3)
53	$f_{\{sky_condition=RA\}}, f_{\{daily_aqi_pm25_+3=high\}}$	2	Sky condition is Rain; Daily AQI pm25 previous 3 days are high (> Q3)
54	$f_{\{tem_max=low\}}, f_{\{concentration_co_+5=high\}}$	2	Maximum temperature is low (< Q1); Daily concentration co for previous 5 days are high (> Q3)
55	$f_{\{tem_max_+1=low\}}, f_{\{daily_aqi_pm25_+3=high\}}$	2	Maximum temperature for previous 1 day is low (< Q1); Daily AQI pm25 previous 3 days are high (> Q3)
56	$f_{\{tem_max_+1=low\}}, f_{\{tem_max_+5=low\}},$ $f_{\{concentration_pm25_+5=high\}}$	2	Maximum temperature for previous 1 day is low (< Q1); Maximum temperature for previous 5 days are low (< Q1); Daily AQI pm25 previous 3 days are high (> Q3)
57	$f_{\{tem_max_+1=low\}}, f_{\{tem_mean_+1=low\}},$ $f_{\{concentration_pm25=high\}}$	2	Maximum temperature for previous 1 day is low (< Q1); Average temperature for previous 1 day is low (< Q1); Daily concentration pm25 is high (> Q3)
58	$f_{\{tem_max_+3=low\}}, f_{\{daily_aqi_no2_+1=high\}}$	2	Maximum temperature for previous 3 days are low (< Q1); Daily AQI no2 previous 1 day is high (> Q3)
59	$f_{\{tem_max_+3=low\}}, f_{\{tem_mean_+3=low\}},$ $f_{\{daily_aqi_no2=high\}}$	2	Maximum temperature for previous 3 days are low (< Q1); Average temperature for previous 3 days are low (< Q1); Daily AQI no2 previous 3 days are high (> Q3)
60	$f_{\{tem_max_+3=low\}}, f_{\{tem_min_+3=low\}},$ $f_{\{daily_aqi_pm25_+1=middle\}}$	2	Maximum temperature for previous 3 days are low (< Q1); Minimum temperature for previous 3 days are low (< Q1); Daily AQI pm25 previous 1 day is middle (> Q1)
61	$f_{\{tem_max_+5=low\}}, f_{\{daily_aqi_pm25_+3=high\}}$	2	Maximum temperature for previous 5 days are low (< Q1); Daily AQI pm25 previous 3 days are high (> Q3)
62	$f_{\{tem_max_+5=low\}}, f_{\{humidity_mean=low\}},$ $f_{\{daily_aqi_pm25=high\}}$	2	Maximum temperature for previous 5 days are low (< Q1); Humidity level is low (< Q1); Daily AQI pm25 previous 1 day is high (> Q3)
63	$f_{\{tem_max_+5=low\}}, f_{\{tem_mean_+5=low\}},$ $f_{\{concentration_co_+1=high\}}$	2	Maximum temperature for previous 5 days are low (< Q1); Average temperature for previous 5 days are low (< Q1); Daily AQI co previous 1 day is high (> Q3)
64	$f_{\{tem_max_+5=low\}}, f_{\{tem_min_+5=low\}},$ $f_{\{concentration_co_+3=high\}}$	2	Maximum temperature for previous 5 days are low (< Q1); Minimum temperature for previous 5 days are low (< Q1); Daily concentration co for previous 3 days are high (> Q3)
65	$f_{\{tem_mean_+1=low\}}, f_{\{concentration_co=high\}}$	2	Average temperature for previous 1 day is low (< Q1); Daily concentration co is high (> Q3)
66	$f_{\{tem_mean_+1=low\}}, f_{\{tem_max_+3=low\}},$ $f_{\{concentration_no2=high\}}$	2	Average temperature for previous 1 day is low (< Q1); Maximum temperature for previous 3 days are low (< Q1); Daily concentration no2 is high (> Q3)
67	$f_{\{tem_mean_+1=low\}}, f_{\{tem_mean_+3=low\}},$ $f_{\{concentration_pm25_+3=high\}}$	2	Average temperature for previous 1 day is low (< Q1); Average temperature for previous 3 days are low (< Q1); Daily concentration pm25 for previous 3 days are high (> Q3)
68	$f_{\{tem_mean_+1=low\}}, f_{\{tem_mean_+5=low\}},$ $f_{\{concentration_pm25_+5=high\}}$	2	Average temperature for previous 3 days are low (< Q1); Average temperature for previous 5 days are low (< Q1); Daily concentration pm25 for previous 5 days are high (> Q3)
69	$f_{\{tem_mean_+3=low\}}, f_{\{cool_mean=middle\}},$ $f_{\{daily_aqi_co_+3=high\}}$	2	Average temperature for previous 3 days are low (< Q1); Cooling degree is middle (> Q1); Daily AQI co previous 3 days are high (> Q3)
70	$f_{\{tem_mean_+3=low\}}, f_{\{daily_aqi_co_+1=high\}}$	2	Average temperature for previous 3 days are low (< Q1); Daily AQI co previous 1 day is high (> Q3)
71	$f_{\{tem_mean_+3=low\}}, f_{\{tem_mean_+5=low\}},$ $f_{\{daily_aqi_no2_+1=middle\}}$	2	Average temperature for previous 3 days are low (< Q1); Average temperature for previous 5 days are low (< Q1); Daily AQI no2 previous 1 day is middle (> Q1)

72	$f_{\{tem_mean_+5=low\}}, f_{\{cool_mean=middle\}}, f_{\{daily_aqi_no_2_+5=middle\}}$	2	Average temperature for previous 5 days are low (< Q1); Cooling degree is middle (> Q1); Daily AQI no2 previous 5 days are middle (> Q1)
73	$f_{\{tem_mean_+5=low\}}, f_{\{daily_aqi_no_2_+5=high\}}$	2	Average temperature for previous 5 days are low (< Q1); Daily AQI no2 previous 5 days are high (> Q3)
74	$f_{\{tem_mean_+5=low\}}, f_{\{humidity_mean=low\}}, f_{\{daily_aqi_no_2_+5=high\}}$	2	Average temperature for previous 5 days are low (< Q1); Humidity level is low (< Q1); Daily AQI no2 is high (> Q3)
75	$f_{\{tem_min_+1=low\}}, f_{\{cool_mean=middle\}}, f_{\{daily_aqi_pm25_+3=high\}}$	2	Minimum temperature for previous 1 day is low (< Q1); Cooling degree is middle (> Q1); Daily AQI pm25 previous 3 days are high (> Q3)
76	$f_{\{tem_min_+1=low\}}, f_{\{daily_aqi_pm25_+1=high\}}$	2	Minimum temperature for previous 1 day is low (< Q1); Daily AQI pm25 previous 1 day is high (> Q3)
77	$f_{\{tem_min_+1=low\}}, f_{\{humidity_mean=high\}}, f_{\{daily_aqi_pm25_+3=middle\}}$	2	Minimum temperature for previous 1 day is low (< Q1); Humidity level is high (> Q3); Daily AQI pm25 previous 3 days are middle (> Q1)
78	$f_{\{tem_min_+1=low\}}, f_{\{tem_mean_+1=low\}}, f_{\{daily_aqi_pm25_+1=middle\}}$	2	Minimum temperature for previous 1 day is low (< Q1); Average temperature for previous 1 day is low (< Q1); Daily AQI pm25 previous 1 day is high (> Q3)
79	$f_{\{tem_min_+3=low\}}, f_{\{cool_mean=middle\}}, f_{\{daily_aqi_pm25_+5=high\}}$	2	Minimum temperature for previous 3 days are low (< Q1); Cooling degree is middle (> Q1); Daily AQI pm25 previous 5 days are high (> Q3)
80	$f_{\{tem_min_+3=low\}}, f_{\{daily_aqi_pm25_+3=high\}}$	2	Minimum temperature for previous 3 days are low (< Q1); Daily AQI pm25 previous 3 days are high (> Q3)
81	$f_{\{tem_min_+3=low\}}, f_{\{tem_max_+5=low\}}, f_{\{daily_aqi_pm25=middle\}}$	2	Minimum temperature for previous 3 days are low (< Q1); Maximum temperature for previous 5 days are low (< Q1); Daily AQI pm25 is middle (> Q1)
82	$f_{\{tem_min_+3=low\}}, f_{\{tem_mean_+3=low\}}, f_{\{concentration_no2_+1=high\}}$	2	Minimum temperature for previous 3 days are low (< Q1); Average temperature for previous 3 days are low (< Q1); Daily concentration no2 for previous 1 day is high (> Q3)
83	$f_{\{tem_min_+3=low\}}, f_{\{tem_min_+5=low\}}, f_{\{concentration_no2_+5=high\}}$	2	Minimum temperature for previous 3 days are low (< Q1); Minimum temperature for previous 5 days are low (< Q1); Daily concentration no2 for previous 5 days are high (> Q3)
84	$f_{\{tem_min_+5=low\}}, f_{\{concentration_co_+1=high\}}$	2	Minimum temperature for previous 5 days are low (< Q1); Daily concentration co for previous 1 day is high (> Q3)
85	$f_{\{tem_min_+5=low\}}, f_{\{cool_mean=middle\}}, f_{\{concentration_no2=high\}}$	2	Minimum temperature for previous 5 days are low (< Q1); Cooling degree is middle (> Q1); Daily concentration no2 is high (> Q3)
86	$f_{\{tem_min_+5=low\}}, f_{\{humidity_mean=high\}}, f_{\{concentration_pm25_+3=high\}}$	2	Minimum temperature for previous 5 days are low (< Q1); Humidity level is high (> Q3); Daily concentration pm25 for previous 3 days are high (> Q3)
87	$f_{\{tem_min_+5=low\}}, f_{\{tem_mean_+5=low\}}, f_{\{concentration_pm25_+5=high\}}$	2	Minimum temperature for previous 5 days are low (< Q1); Average temperature for previous 5 days are low (< Q1); Daily concentration pm25 for previous 5 days are high (> Q3)
88	$f_{\{toy=T\}}, f_{\{sen=low\}}$	3 & 4	Mention toys in previous 90 days; Sentiment level is low (< Q1)

Cat. = Category: 1 = Biological and demographic factors; 2 = Environmental factors; 3 = Behavioral factors; 4 = Social and psychological factors; 5 = Inflammatory and infectious factors

Order: Sequential frequent patterns are sorted in lexicographically ascending order

Q1: The 25th percentile as the first quartile; Q3: The 75th percentile as the third quartile

specialists determined that these patterns are meaningful and useful; they also verified that some of these patterns extend their current understanding of asthma triggers and risk factors. The frequent patterns are sorted in lexicographic ascending order.

We also report the results from using RF to determine the relative importance of asthma triggers and risk factors. There are two major hyper-parameters (i.e., values are set before the learning process begins): the depth of the forest (d) and the number of trees (n), we set $d = 20$ and $n = 400$ by using 10-fold cross-validation. The tuned model was then used to evaluate the importance of features. All features shown in Table 3 combined with frequently occurring sequences of asthma triggers and risk factors, obtained from the sequential pattern mining process (Table 6), are the input space for the learner. A total of 358 features were analyzed and we report on the most significant asthma triggers and risk factors overall in the United States and also in 10 different U.S. regions (Figure 13). All the results were evaluated and validated by asthma experts who worked with us to interpret the results. As described in the previous section, feature importance is defined by the Gini impurity. The most significant features are reported based on the criteria suggested by asthma specialists. For the entire United States, we report features with Gini importance scores greater than 0.02; for the individual regions, we report features with Gini importance greater than 0.015. All reported asthma triggers and risk factors are sorted in descending order of Gini scores.

Evaluation of Methods and Results







In this subsection, we report on the evaluation results and determine how well they align with previously published medical literature.

As discussed in the previous section, the importance of asthma triggers and risk factors were determined using the RF classification process based on self-reported asthma patients identified from social media. Table 7 shows the RF classification results (based on data from the entire United States). A 10-fold cross-validation is applied. We first report the classification results by using all features from five categories of asthma triggers and risk factors. The classification results are found to be stable (see Table 7, Feature set = ALL, accuracy = 85.6%, F1(affliction) = 85.1%, F1(no-affliction) = 86.1%) which, in turn, show that the reported asthma triggers and risk factors are useful for distinguishing users with asthma affliction from those without.

We also report on classification results obtained from using each of the five categories of asthma triggers and risk factors

(see Table 7). When using either environmental (Cat. = 2) or behavioral (Cat. = 3) attributes, classification accuracy is moderate, ranging from 60.0% to 74.3%; when using biological and demographic factors (Cat. = 1), social and psychological factors (Cat. = 4) or inflammatory and infectious factors (Cat. = 5), classification accuracy is low (slightly over 50%). However, when all five categories are used together, classification performance improved significantly (85.6%) demonstrating that asthma triggers and risk factors are complex, and may have some interactions. The results also reinforce the deficiency of past studies which typically focused on one or two types of asthma triggers or risk factors.

In addition, we implemented a greedy feature selection method to gain insights into the contribution of different categories of asthma triggers and risk factors. We initialized the feature selection process with the most predictive feature group (i.e., the one with the best performance). Then we incrementally added another feature group that provided the highest accuracy improvement to the feature set. Note the results need to be interpreted with care. For example, if two feature groups are highly correlated, the algorithm would choose one at random and discard the other. Therefore, if a feature group is missing from the selection process it implies that the feature is either less useful, or strongly correlated with other features in the list. Table 8 summarizes the outcomes of this feature selection process. Category (Cat.) = x , $x \in \{1, 2, 3, 4, 5\}$ indicates each of the five asthma triggers and risk factors categories. Category (Cat.) = x^* or Category (Cat.) = $x \& y$, where $x \in \{1, 2, 3, 4, 5\}$, $y \in \{1, 2, 3, 4, 5\}$ specify the asthma triggers and risk factors patterns extracted from the frequent sequential mining process. For example, Cat. = 2^* shows frequent sequential patterns between two or more environmental factors (e.g., $f_{\{humidity_mean=high\}}$, $f_{\{concentration_pm25=high\}}$: Humidity level is high; Daily concentration pm2.5 is high); Cat. = $2 \& 3$ indicates frequent sequential patterns between environmental and behavioral factors (e.g., $f_{\{humidity_mean=low\}}$, $f_{\{seafood=T\}}$: Humidity level is low; consuming seafood in the previous 90 days is True). The results show that the selected features span all the five asthma triggers and risk factors categories. The vast majority are taken from the environmental features (Cat. = 2 and Cat. = 2^*). Behavioral factors (Cat. = 3), social and psychological factors (Cat. = 4) and their interconnections (Cat. = $3 \& 4$) also contribute to improved accuracy. Moreover, the frequent sequential patterns of asthma risk factors (Table 8, Cat. = $3 \& 4$, Cat. = $2 \& 4$, Cat. = $2 \& 3$, Cat. = $3 \& 5$, Cat. = $2 \& 5$, Cat. = 3^* , Cat. = 1^* , Cat. = $3 \& 4 \& 5$, and Cat. = 5^*) add a lot of value to the feature set, resulting in significant improvement in accuracy. The results show that asthma exacerbation may be related to multiple risk factors. This reinforces the need to investigate the sequential effects of triggers and risk factors on asthma exacerbation.

	<p>United States</p> <p> <i>f_{sky_condition=BR}</i> <i>f_{exe=T},f_{med=T},f_{sen=low}</i> <i>f_{ili_act_label=minimal},f_{med=T},f_{sen=low}</i> <i>f_{exe=T},f_{sen=low}</i> <i>f_{cat_alcohol_impaired=middle},f_{exe=T}</i> <i>f_{sen=low},f_{cool_mean=middle}</i> <i>f_{humidity_mean=low}</i> <i>f_{ili_act_label=minimal},f_{exe=T}</i> <i>f_{activity_level=low}</i> <i>f_{cat_excessive_drinking=middle},f_{sen=low}</i> <i>f_{depart_mean=middle},f_{cool_mean=middle}</i> </p>
	<p>Region 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont</p> <p> <i>f_{ili_act_label=minimal},f_{tem_min_+5=middle}</i> <i>f_{tem_min_+5=low},f_{concentration_co_+1=high}</i> <i>f_{exe=T},f_{med=T},f_{sen=low}</i> <i>f_{daily_aqi_so2=high}</i> <i>f_{age>20},f_{gender=female}</i> <i>f_{sen=low},f_{humidity_mean=low}</i> </p>
	<p>Region 2: New Jersey, New York, Puerto Rico, and the U.S. Virgin Islands</p> <p> <i>f_{humidity_mean=low},f_{daily_aqi_pm25_+1=high}</i> <i>f_{exe=T},f_{med=T},f_{sen=low}</i> <i>f_{tem_mean_+1=low},f_{concentration_co=high}</i> <i>f_{ili_act_label=minimal}</i> <i>f_{cat_physical_inactivity=low},f_{cat_alcohol_impaired=middle}</i> <i>f_{tem_mean_+3=low},f_{daily_aqi_co_+1=high}</i> </p>
	<p>Region 3: Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia</p> <p> <i>f_{sen=low},f_{depart_mean=middle}</i> <i>f_{concentration_pb_+5=high}</i> <i>f_{ili_act_label=minimal},f_{positive_flu=middle}</i> <i>f_{cat_excessive_drinking=middle},f_{ili_act_label=minimal}</i> <i>f_{cool_mean=middle}</i> <i>f_{exe=T},f_{sen=low}</i> </p>
	<p>Region 4: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee</p> <p> <i>f_{cat_physical_inactivity=low},f_{ili_act_label=minimal}</i> <i>f_{concentration_pm10_+3=high}</i> <i>f_{exe=T},f_{med=T},f_{sen=low}</i> <i>f_{cat_alcohol_impaired=middle},f_{exe=T}</i> <i>f_{humidity_mean=high},f_{concentration_pm25=high}</i> <i>f_{ili_act_label=minimal},f_{exe=T}</i> </p>
	<p>Region 5: Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin</p> <p> <i>f_{tem_max=low},f_{concentration_co_+5=high}</i> <i>f_{concentration_so2=high}</i> <i>f_{sky_condition=BR}</i> <i>f_{cat_med_inc=low}</i> <i>f_{activity_level=low}</i> <i>f_{cat_excessive_drinking=middle},f_{ili_act_label=minimal}</i> </p>






	<p>Region 6: Arkansas, Louisiana, New Mexico, Oklahoma, and Texas</p> <p> <i>f_{humidity_mean=low}</i> <i>f_{ili_act_label=minimal},f_{exe=T}</i> <i>f_{sky_condition=BR}</i> <i>f_{cat_excessive_drinking=middle},f_{sen=low}</i> <i>f_{depart_mean=middle},f_{cool_mean=middle}</i> <i>f_{risk_race=hispanic}</i> </p>
	<p>Region 7: Iowa, Kansas, Missouri, and Nebraska</p> <p> <i>f_{access_to_health_services=low}</i> <i>f_{humidity_mean=low},f_{concentration_pm25_+5=high}</i> <i>f_{med=T},f_{humidity_mean=low}</i> <i>f_{tem_mean_+1=low},f_{tem_mean_+5=low},f_{concentration_pm25_+5=high}</i> <i>f_{sen=low},f_{cool_mean=middle}</i> <i>f_{exe=T},f_{sen=low}</i> </p>
	<p>Region 8: Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming</p> <p> <i>f_{ili_act_label=minimal},f_{tem_min_+5=middle}</i> <i>f_{ili_act_label=minimal},f_{med=T},f_{sen=low}</i> <i>f_{tem_min_+1=low}</i> <i>f_{cat_physical_inactivity=low},f_{cat_alcohol_impaired=middle}</i> <i>f_{humidity_mean=low},f_{concentration_pm25_+5=high}</i> <i>f_{depart_mean=middle},f_{cool_mean=middle}</i> </p>
	<p>Region 9: Arizona, California, Hawaii, and Nevada</p> <p> <i>f_{daily_aqi_pm25_+5=high}</i> <i>f_{concentration_pb_+3=high}</i> <i>f_{tem_min_+1=high}</i> <i>f_{risk_race=hispanic}</i> <i>f_{sen=low},f_{cool_mean=middle}</i> <i>f_{sky_condition=BR},f_{daily_aqi_pm25_+1=high}</i> <i>f_{tem_max_+3=low},f_{tem_min_+3=low},f_{daily_aqi_pm25_+1=middle}</i> </p>
	<p>Region 10: Alaska, Idaho, Oregon, and Washington</p> <p> <i>f_{ili_act_label=minimal},f_{med=T},f_{sen=low}</i> <i>f_{cat_excessive_drinking=middle},f_{ili_act_label=minimal}</i> <i>f_{cat_excessive_drinking=middle},f_{sen=low}</i> <i>f_{concentration_pb_+3=high}</i> <i>f_{age>20},f_{gender=female}</i> <i>f_{cool_mean=high}</i> <i>f_{daily_aqi_pm10_+3=high}</i> </p>

Figure 13. Asthma Triggers and Risk Factors Based on Feature Importance

Feature importance: Gini Importance = Mean Decrease in Impurity.

Feature selection criteria: U.S.: Gini Importance > 0.02; Region 1–10: Gini Importance > 0.015.

Order: Features are sorted by the descending order of the feature importance.

Table 7. Random Forest Classification Results

Feature Set	Accuracy (%)	Affliction	No Affliction
		F1 (%)	F1 (%)
Cat. = 1	51.4	53.8	48.7
Cat. = 2	74.3	75.3	73.3
Cat. = 3	60.0	63.5	55.7
Cat. = 4	53.1	56.1	49.5
Cat. = 5	50.9	53.4	48.2
ALL	85.6	85.1	86.1

Feature Set: Cat. = x , $x \in \{1, 2, 3, 4, 5\} \rightarrow$ Asthma triggers and risk factors in category.

Cat. = Category: 1 = Biological and demographic factors; 2 = Environmental factors; 3 = Behavioral factors; 4 = Social and psychological factors; 5 = Inflammatory and infectious factors.

Accuracy: The description of model errors, defined as $accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n I(y, \hat{y})$, where $I(x)$ is the indicator function having the value 1 if the predicted label \hat{y} match with the true label y .

F1: The harmonic mean $\left(\frac{2PR}{P+R}\right)$ of precision (P) and recall (R).

Table 8. Summary of the Feature Selection Process

Iteration	Feature Set	Accuracy (%)
1	Cat. = 2	74.39
2	+ Cat. = 2*	78.94
3	+ Cat. = 3	82.19
4	+ Cat. = 4	82.91
5	+ Cat. = 3 & 4	83.51
6	+ Cat. = 2 & 4	83.93
7	+ Cat. = 2 & 3	84.23
8	+ Cat. = 3 & 5	84.47
9	+ Cat. = 1	84.71
10	+ Cat. = 2 & 5	84.89
11	+ Cat. = 3*	85.07
12	+ Cat. = 5	85.25
13	+ Cat. = 1*	85.31
14	+ Cat. = 3 & 4 & 5	85.37
15	+ Cat. = 5*	85.43
	ALL	85.61

Classifier: Random forest; 10-fold cross-validation

Cat. = Category: 1 = Biological and demographic factors; 2 = Environmental factors; 3 = Behavioral factors; 4 = Social and psychological factors; 5 = Inflammatory and infectious factors

Category (Cat.) = x , $x \in \{1, 2, 3, 4, 5\}$ indicate one of the five asthma triggers and risk factors categories. Category (Cat.) = x^* or Category (Cat.) = $x \& y$, where $x \in \{1, 2, 3, 4, 5\}$. $y \in \{1, 2, 3, 4, 5\}$ specify the frequent asthma triggers and risk factors patterns extracted from the frequent sequential mining process.

Additionally, since random forest is not the only method available for feature importance assessment, we provide comparison results with various other feature importance assessment algorithms, including χ^2 , logistic regression, and generalized additive model (see Table 9). χ^2 is a simple baseline approach for feature selection. It measures the worth of a feature by computing the value of the χ^2 statistic with respect to the class. Sparse estimators, such as logistic regression with the L1 penalty, are also useful for feature selection. Many of the estimated coefficients may be zero and the goal is to select the features with non-zero coefficients. Generalized additive models are state-of-the-art methods for high-dimensional non-parametric classification and feature selection. Precision at K (P@K) is introduced to compare the results of different feature selection methods. Precision at K shows the proportion (%) of the same feature in the top-k feature sets. The results (see Table 9) show that the feature importance derived from the three methods agree well with each other. In other words, 86.7% of the top 90 most important risk factors are the same between random forest and χ^2 ; 87.7% of the top 90 most important risk factors are the same between random forest and logistic regression (penalty = L1); and 82% of the top 50 most important risk factors are the same between random forest and generalized additive model. The results validate the choice of random forest as the feature importance assessment method.

Next, we evaluate how well the identified asthma risk factors mirror existing real world knowledge or medical consensus. To do this, we collected a comprehensive set of 137 medical research papers from PubMed (See Appendix A) which had the keywords “asthma trigger(s)” or “asthma risk factor(s)” in the title or abstract. We analyzed these research articles to extract the categories of asthma triggers and risk factors from them (see Figure 14). By comparing the categories of asthma risk factors identified from our study and the results from these medical articles, we found our methods are able to identify almost all of the important triggers and risk factors contributing to asthma exacerbations. This confirms the efficacy of our framework and methods. There is, however, a significant difference: most extracted triggers and factors in our study are environment-related factors, while current medical literature mainly focuses on biological and demographic factors. This may be because the effect of environmental factors cannot really be examined using survey-based or traditional data collection methods, hence researchers focus on fundamental biological and demographic factors; however, in reality, weather and air quality are very important risk factors; or because of some specific limitations of our data: although we made every attempt to include all relevant data sources, it is still extremely difficult to obtain personal biological data (due to privacy restrictions) and even harder to integrate them with other data sources. Hence, such factors

are underrepresented in our proposed framework. However, as an open framework, biological data can be imported and used when they are available (We are collaborating with asthma specialists and will add biological data in our future work).

Interpretations of the Results

Overall, our study has revealed several interesting insights related to asthma exacerbation.

- (1) The results confirm that asthma risk factors are complex and varied, including multiple biological, demographic, behavioral, environmental, social, psychological and infectious determinants (see Table 6 and Figure 13).
- (2) Asthma triggers and factors are often interconnected. Based on the results of SPM, for example, we find that viral infections are associated with environment, behavioral and social factors (Table 6, No. 36–44); while air quality factors interact with weather factors, resulting in asthma exacerbation, as shown in Table 6 (No. 51–87). Based on the results of RF, we find that a number of sequential patterns are critical in distinguishing asthma patients from non-asthma patients. For example, in the United States, there are very specific sequential patterns of asthma triggers and risk factors in different regions (Figure 13, United States, No. 2–8, 10–11).
- (3) According to our analysis, environmental factors, including weather and air quality, are the most important asthma risk factors. In particular, almost 55% of asthma exacerbations are caused by environment-related factors. In the United States, mist is one of the more important triggers for asthma exacerbation (Figure 13, United States, sky condition $BR = mist$, i.e., tiny water droplets suspended in the atmosphere). This has also been identified in medical studies (Kashiwabara et al. 2003). As a population-level intervention strategy, outdoor activities should be restricted in such conditions. Additionally, cold air is another significant trigger of asthma exacerbations as shown in Figure 13, United States, Region 2, 3, 5, 7, 8 and 9. Very high or very low humidity may also trigger asthma exacerbations (Figure 13, United States, Region 1, 2, 4, 6, 7, 8). The possible reasons are with low humidity, windy conditions and dry weather may spread pollutants quickly which may increase asthma exacerbations; with high humidity, moist air creates an ideal environment for dust mites to grow and multiply which may increase asthma exacerbations. Our results show that a key air pollutant, that is, small-particulate (i.e., PM2.5 and PM10, found in haze, smoke,

Table 9. Comparison Results with Different Feature Importance Assessment Algorithms

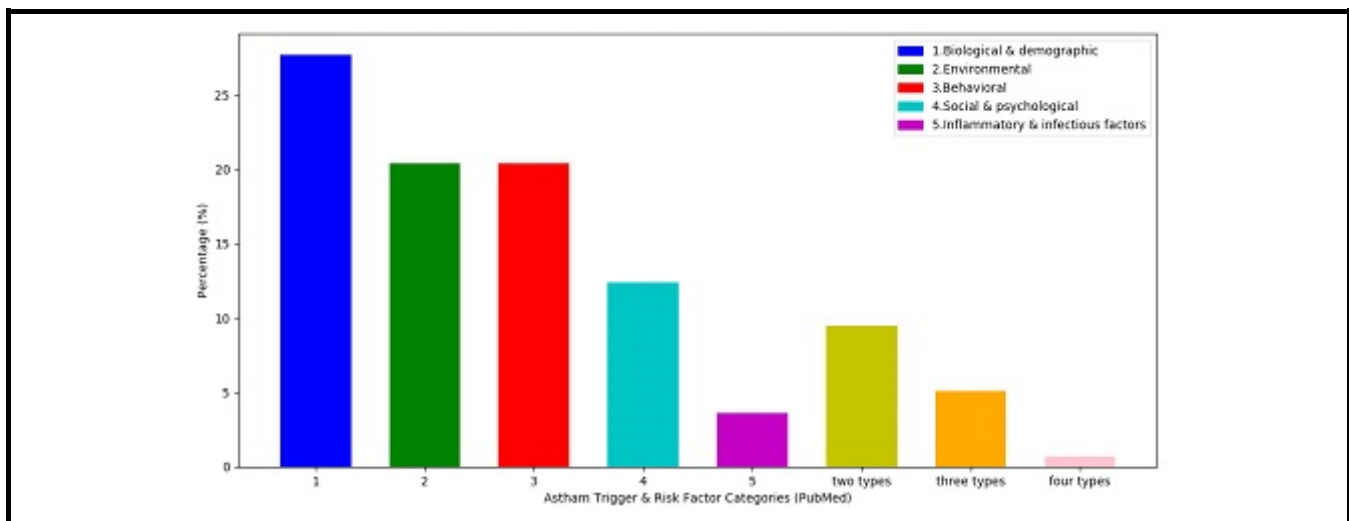
Feature Selection Compare with Random Forest P@K	χ^2 (%)	Logistic Regression (%)	Generalized Additive Model (%)
K = 20	45.0	55.0	55.0
K = 30	53.3	56.7	63.3
K = 40	58.3	60.0	75.0
K = 50	70.0	66.0	82.0
K = 60	66.6	73.3	
K = 70	78.6	74.2	
K = 80	81.3	80.0	
K = 90	86.7	87.7	

P@K: precision at k: the proportion of the same feature in the top-k feature sets.

χ^2 : measures the worth of a feature by computing the value of the statistic with respect to the class.

Logistic Regression: Penalty: L1 (L1: pushing feature coefficients to 0 and creating a method for feature selection). Feature relative importance assessment: standardized coefficients.

Generalized Additive Models: feature selection: p-value ($p < 0.05$). Feature relative importance assessment: standardized coefficients.

**Figure 14. Research Articles on Asthma Triggers and Risk Factors in PubMed**

including dust, and pollen), affects asthma exacerbations (Figure 13, Region 2, 4, 7, 8, 9, 10). As an intervention for asthmatic patients, such pollution information should be spatially and temporally analyzed, and communicated to community stakeholders.

- (3) The Hispanic population is disproportionately affected by asthma (Figure 13, Region 6 and 9) and researchers believe the reasons are yet-unidentified genetic variants (Hunninghake et al. 2006). Our extracted frequent risk factors sequence $f_{\{race = hispanics\}}$, $f_{\{cat_med_inc = low\}}$ (Table 6, No. 47) shows that this might be the

result of the collective effects of genetic factors and socioeconomic status (e.g., there may be low-income Hispanic families with limited healthcare benefits that have barriers to healthcare access). Population-level interventions should be provided to help such less privileged groups.

- (4) Exposure to indoor pollutants and allergens may also be important risk factors for asthma exacerbation. As is shown in Table 6 (No. 88), asthma patients may be exposed to mold through cotton toys.

- (5) Another interesting finding is that behavior modification interventions could be an important way to control asthma exacerbation. For example, behavioral factors such as drinking, smoking, exercise, and adhering to medications (Table 6, No. 3, 19, 20, etc.) are all important in managing asthma exacerbations. We believe, our study paves the way to understand these triggers and risk factors and to include them in devising effective asthma management plans.

Conclusions and Future Work

In this study, we focused on comprehensively identifying asthma triggers and risk factors, their relative importance, their interconnecting relationships, and the relative importance of asthma triggers and risk factors.

Our proposed framework is very promising for many reasons. Our major research contributions for chronic disease management are, first, we developed a framework for comprehensive asthma trigger and risk factor analysis by leveraging widely available open data sources about asthma and adapting multiple advanced machine learning techniques. Our framework is able to confirm existing understanding as well as discover new medical knowledge regarding asthma triggers and risk factors by determining the sequential patterns and relative importance of these risk factors. Second, we demonstrated the use of novel data sources, such as social media, for deriving characteristics, including demographics and behaviors, of asthma patients, as an alternative to traditional survey-based data collection methods. Third, we developed a framework for integrating and repurposing highly heterogeneous data from multiple sources with varied spatial-temporal resolutions, which can be used as a complement to retrospective cohort studies.

Our proposed framework can economically identify asthma risk factors in a timely manner. The results of this study can provide guidance for developing asthma management plans and population-level asthma interventions.

The technical contribution of this study is also significant. We propose a framework with four components along with systematic evaluations for each component. We first propose a new two-stage model to derive characteristics of self-reported asthma patients from social media with image recognition techniques to enhance asthma patients' background information extraction. The proposed two-stage model outperforms state-of-the-art methods. We adapt and integrate the ray casting algorithm to combine and repurpose highly

heterogeneous data from multiple open sources with disparate spatial-temporal resolutions. We use sequential pattern mining to determine the connections between asthma triggers and risk factors. We modify random forest to uncover the relative importance of asthma triggers and risk factors. These are typically challenging to understand using retrospective cohort studies, hence our framework and methods can be used as a complement to such studies.

While the results are encouraging, the proposed framework is not without limitations.

The first limitation stems from the use of social media data. Although we find social media is a valuable source for extracting characteristics of self-reported asthma patients, the datasets may be incomplete and have potential selection bias. To address the incompleteness problem and improve data quality, we recommend adding data from self-reported patients across multiple social media sites. To address the selection bias of social media, we propose using our framework as a complement to survey-based data collection methods to obtain more representative samples of individuals from the asthma population.

The second limitation is that some other important data sources are not included due to lack of availability. For example, since asthma is related to genetic factors, heterogeneous biological data sources (e.g., experimental results and genetic profiles) (Rharbi et al. 2012) may provide valuable information. Since the proposed framework is open, such data can be imported when available and integrated with other data sources. Additionally, we collected local weather and outdoor air quality data for our analyses. However, the indoor environment of asthma patients may be different from the outdoor environmental conditions. Wearable sensors that capture data about indoor environments, activity levels, as well as other behavior of individuals may enhance the risk factor analyses results using our proposed methods. Additionally, medication adherence related behavior, if available, may be incorporated into future studies.

In future work, our framework can be extended in the following ways: First, by integrating advanced machine learning paradigms, the framework can use real-time or near-real-time sensor and the Internet-based datasets to follow trends and changes in asthma risk factors as they evolve. Second, the framework may be generalized and extended to detect patterns, trends, and risk factors for other chronic conditions such as type 2 diabetes and obesity. Third, the framework can be used as a complement to prospective cohort studies because the conditions of self-reported asthma patients and other factors can be followed prospectively.

Acknowledgments

We thank medical research personnel from the College of Medicine at the University of Arizona who generously shared their insights and expertise on asthma research and management, greatly enhancing this work. We thank the *MISQ* special issue editors and the anonymous reviewers for constructive reviews and a developmental review process. We also thank participants at the *MIS Quarterly* Special Issue Workshop: The Role of Information Systems and Analytics in Chronic Disease Prevention & Management (2018) for valuable feedback.

This work was done when Wenli Zhang was a research assistant and doctoral student at the Department of MIS, Eller College of Management, University of Arizona.

References

- Abbar, S., Mejova, Y., and Weber, I. 2015. "You Tweet What You Eat: Studying Food Consumption Through Twitter," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, New York: ACM, pp. 3197-3206.
- Akinbami, L. J., Moorman, J. E., and Liu, X. 2011. "Asthma Prevalence, Health Care Use, and Mortality: United States, 2005–2009," *National Health Statistics Reports* (32), pp. 1-14.
- Almqvist, C., Worm, M., and Leynaert, B. 2008. "Impact of Gender on Asthma in Childhood and Adolescence: A GA²LEN Review," *Allergy* (63:1), pp. 47-57.
- Andrade, E. B., Kaltcheva, V., and Weitz, B. 2002. "Self-Disclosure on the Web: The Impact of Privacy Policy, Reward, and Company Reputation," *Advances in Consumer Research* (29:1), pp. 350-353.
- Andreu-Perez, J., Leff, D. R., Ip, H. M. D., and Yang, G.-Z. 2015. "From Wearable Sensors to Smart Implants—Toward Pervasive and Personalized Healthcare," *IEEE Transactions on Bio-Medical Engineering* (62:12), pp. 2750-2762.
- Arif, A. A., Delclos, G. L., and Serra, C. 2009. "Occupational Exposures and Asthma among Nursing Professionals," *Occupational and Environmental Medicine* (66:4), pp. 274-278.
- Barnes, P. J. 2008. "The Cytokine Network in Asthma and Chronic Obstructive Pulmonary Disease," *The Journal of Clinical Investigation* (118:11), pp. 3546-3556.
- Barnett, S. B. L., and Nurmagambetov, T. A. 2011. "Costs of Asthma in the United States: 2002–2007," *Journal of Allergy and Clinical Immunology* (127:1), pp. 145-152.
- Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., and Escobar, G. 2014. "Big Data In Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients," *Health Affairs* (33:7), pp. 1123-1131.
- Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, Sebastopol, CA: O'Reilly Media, Inc.
- Boyd, D. M., and Ellison, N. B. 2007. "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication* (13:1), pp. 210-230.
- Bradski, G. 2000. "The OpenCV Library," *Dr. Dobb's Journal: Software Tools for the Professional Programmer* (25:11), pp. 120-123.
- Breiman, L. 2001. "Random Forests," *Machine Learning* (45:1), pp. 5-32.
- Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. 2011. "Discriminating Gender on Twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA: Association for Computational Linguistics, pp. 1301-1309.
- CDC. 2013. "Asthma Facts—CDC's National Asthma Control Program Grantees," Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
- CDC. 2017a. "CDC—Asthma—Most Recent Asthma State Data," Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (https://www.cdc.gov/asthma/most_recent_data_states.htm; retrieved April 19, 2017).
- CDC. 2017b. "Flu and People with Asthma|Seasonal Influenza (Flu)," Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
- Chaudhuri, R., Livingston, E., McMahon, A. D., Thomson, L., Borland, W., and Thomson, N. C. 2003. "Cigarette Smoking Impairs the Therapeutic Response to Oral Corticosteroids in Chronic Asthma," *American Journal of Respiratory and Critical Care Medicine* (168:11), pp. 1308-1311.
- Chen, E., and Miller, G. E. 2007. "Stress and Inflammation in Exacerbations of sthma," *Brain, Behavior, and Immunity* (21:8), pp. 993-999.
- Chen, M., Mao, S., and Liu, Y. 2014. "Big Data: A Survey," *Mobile Networks and Applications* (19:2), pp. 171-209.
- Chen, X., Wang, Y., Agichtein, E., and Wang, F. 2015. "A Comparative Study of Demographic Attribute Inference in Twitter," in *Proceedings of the 9th International AAAI Conference on Web and Social Media*, Palo Alto, CA: AAAI, pp. 590-593.
- Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* (20:1), pp. 37-46.
- Cortes, C., and Vapnik, V. 1995. "Support-Vector Networks," *Machine Learning* (20:3), pp. 273-297.
- Craven, M., and Kumlien, J. 1999. "Constructing Biological Knowledge Bases by Extracting Information from Text Sources," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pp. 77-86.
- Derlaga, V. J., and Berg, J. H. 1987. *Self-Disclosure: Theory, Research, and Therapy, Perspectives in Social Psychology* (1st ed.), New York: Springer.
- Dindia, K. 2001. "Self-Disclosure Research: Knowledge through Meta-Analysis," in *Interpersonal Communication Research: Advances Through Meta-Analysis*, M. Allen, R. W. Preiss, B. M. Gayle, and N. Burrell (eds.), Oxford, UK: Taylor & Francis, pp. 169-185.
- EPA. 2014. "AQI Air Quality Index: A Guide to Air Quality and Your Health," Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Outreach and Information Division.
- Eurowinter Group. 1997. "Cold Exposure and Winter Mortality from Ischaemic Heart Disease, Cerebrovascular Disease,

- Respiratory Disease, and All Causes in Warm and Cold Regions of Europe," *The Lancet* (349), pp. 1341-1346.
- Fang, Q., Sang, J., Xu, C., and Hossain, M. S. 2015. "Relational User Attribute Inference in Social Media," *IEEE Transactions on Multimedia* (17:7), pp. 1031-1044.
- Ferguson, C. J., Muñoz, M. E., Garza, A., and Galindo, M. 2014. "Concurrent and Prospective Analyses of Peer, Television and Social Media Influences on Body Dissatisfaction, Eating Disorder Symptoms and Life Satisfaction in Adolescent Girls," *Journal of Youth and Adolescence* (43:1), pp. 1-14.
- Galant, S. P., Crawford, L. J. R., Morphew, T., Jones, C. A., and Bassin, S. 2004. "Predictive Value of a Cross-Cultural Asthma Case-Detection Tool in an Elementary School Population," *Pediatrics* (114:3), pp. e307-316.
- Gibbs, J. L., Ellison, N. B., and Heino, R. D. 2006. "Self-Presentation in Online Personals: The Role of Anticipated Future Interaction, Self-Disclosure, and Perceived Success in Internet Dating," *Communication Research* (33:2), pp. 152-177.
- Go, A., Bhayani, R., and Huang, L. 2009. "Twitter Sentiment Classification Using Distant Supervision," No. CS224N Project Report, Stanford, pp. 12-18.
- Han, J., Cheng, H., Xin, D., and Yan, X. 2007. "Frequent Pattern Mining: Current Status and Future Directions," *Data Mining and Knowledge Discovery* (15:1), pp. 55-86.
- Han, J., Pei, J., and Kamber, M. 2011. *Data Mining: Concepts and Techniques*, Amsterdam: Elsevier.
- Herman, E. J. 2011. "Conceptual Framework of the Controlling Asthma in American Cities Project," *Journal of Urban Health* (88:1), pp. 7-15.
- Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105.
- Hong, C. Y., Ng, T. P., Wong, M. L., Koh, K. T. C., Goh, L. G., and Ling, S. L. 1994. "Lifestyle and Behavioural Risk Factors Associated with Asthma Morbidity in Adults," *QJM: An International Journal of Medicine* (87:10), pp. 639-645.
- Huang, Y., Yu, L., Wang, X., and Cui, B. 2015. "A Multi-Source Integration Framework for User Occupation Inference in Social Media Systems," *World Wide Web* (18:5), pp. 1247-1267.
- Hunninghake, G. M., Weiss, S. T., and Celedón, J. C. 2006. "Asthma in Hispanics," *American Journal of Respiratory and Critical Care Medicine* (173:2), pp. 143-163.
- Jalali, L., Dao, M.-S., Jain, R., and Zettsu, K. 2015. "Complex Asthma Risk Factor Recognition from Heterogeneous Data Streams," in *Proceedings of the 2015 IEEE International Conference on Multimedia Expo Workshops*, Turin, Italy, June 29-July 3, pp. 214-219.
- Jamison-Powell, S., Linehan, C., Daley, L., Garbett, A., and Lawson, S. 2012. "'I Can't Get No Sleep': Discussing #Insomnia on Twitter," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York: ACM, pp. 1501-1510.
- Kane, G. C., Alavi, M., Labianca, G., and Borgatti, S. P. 2014. "What's Different About Social Media Networks? A Framework and Research Agenda," *MIS Quarterly* (38:1), pp. 275-304.
- Kaplan, A. M., and Haenlein, M. 2010. "Users of the World, Unite! The Challenges and Opportunities of Social Media," *Business Horizons* (53:1), pp. 59-68.
- Kashiwabara, K., Itonaga, K., and Moroi, T. 2003. "Airborne Water Droplets in Mist or Fog May Affect Nocturnal Attacks in Asthmatic Children," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (40:4), pp. 405-411.
- Kovats, R. S., and Hajat, S. 2008. "Heat Stress and Public Health: A Critical Review," *Annual Review of Public Health* (29), pp. 41-55.
- Kozyrskyj, A. L., Kendall, G. E., Zubrick, S. R., Newnham, J. P., and Sly, P. D. 2009. "Frequent Nocturnal Awakening in Early Life Is Associated with Nonatopic Asthma in Children," *The European Respiratory Journal* (34:6), pp. 1288-1295.
- Kurai, D., Saraya, T., Ishii, H., and Takizawa, H. 2013. "Virus-Induced Exacerbations in Asthma and COPD," *Frontiers in Microbiology* (4:293).
- Landis, J. R., and Koch, G. G. 1977. "An Application of Hierarchical Kappa-Type Statistics in the Assessment of Majority Agreement among Multiple Observers," *Biometrics* (33:2), pp. 363-374.
- Laurini, R., and Thompson, D. 1992. *Fundamentals of Spatial Information Systems*, Waltham, MA: Academic Press.
- Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE* (86:11), pp. 2278-2324.
- Lee, C.-H., Chen, J. C.-Y., and Tseng, V. S. 2011. "A Novel Data Mining Mechanism Considering Bio-Signal and Environmental Data with Applications on Asthma Monitoring," *Computer Methods and Programs in Biomedicine* (101:1), pp. 44-61.
- Levi, G., and Hassner, T. 2015. "Age and Gender Classification Using Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Boston, pp. 34-42.
- Liu, S., Yang, J., Huang, C., and Yang, M. H. 2015. "Multi-Objective Convolutional Learning for Face Labeling," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3451-3459.
- Liu, S., Zhu, M., Yu, D. J., Rasin, A., and Young, S. D. 2017. "Using Real-Time Social Media Technologies to Monitor Levels of Perceived Stress and Emotional State in College Students: A Web-Based Questionnaire Study," *JMIR Mental Health* (4:1).
- Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., and Murray, C. J. (eds.). 2006. *Global Burden of Disease and Risk Factors*, Washington, DC: World Bank.
- Mabroukeh, N. R., and Ezeife, C. I. 2010. "A Taxonomy of Sequential Pattern Mining Algorithms," *ACM Computing Surveys* (43:1), Article 3.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. 2011. "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, New York.
- March, S. T., and Smith, G. F. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), pp. 251-266.
- Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. 2012. "An Overview of the Global Historical Climatology Network-Daily Database," *Journal of Atmospheric and Oceanic Technology* (29:7), pp. 897-910.
- Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, N. 2011. "Understanding the Demographics of

- Twitter Users,” in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pp. 554-557.
- Mohammad, S. M., Kiritchenko, S., and Zhu, X. 2013. “NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets,” arXiv:1308.6242 (<http://arxiv.org/abs/1308.6242>).
- Myslin, M., Zhu, S.-H., Chapman, W., and Conway, M. 2013. “Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products,” *Journal of Medical Internet Research* (15:8).
- Nesi, J., and Prinstein, M. J. 2015. “Using Social Media for Social Comparison and Feedback-Seeking: Gender and Popularity Moderate Associations with Depressive Symptoms,” *Journal of Abnormal Child Psychology* (43:8), pp. 1427-1438.
- Nicholson, K. G., Kent, J., and Ireland, D. C. 1993. “Respiratory Viruses and Exacerbations of Asthma in Adults,” *BMJ* (307:6910), pp. 982-986.
- Nunamaker, J. F., Chen, M., and Purdin, T. D. M. 1990. “Systems Development in Information Systems Research,” *Journal of Management Information Systems* (7:3), pp. 89-106.
- O’Byrne, P. M., Ryan, G., Morris, M., McCormack, D., Jones, N. L., Morse, J. L. C., and Hargreave, F. E. 1982. “Asthma Induced by Cold Air and its Relation to Nonspecific Bronchial Responsiveness to Methacholine,” *American Review of Respiratory Disease* (125:3), pp. 281-285.
- Pei, J., Han, J., Mortazavi-asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M. 2001. “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth,” in *Proceedings of the 17th Annual International Conference on Data Engineering*, IEEE, pp. 215-224.
- Ram, S., Zhang, W., Williams, M., and Pengetnze, Y. 2015. “Predicting Asthma-Related Emergency Department Visits Using Big Data,” *IEEE Journal of Biomedical and Health Informatics* (19:4), pp. 1216-1223.
- Ramachandran, A., Snehalatha, C., Ram, J., Selvam, S., Simon, M., Nanditha, A., Shetty, A.S., Godslan, I. F., Chaturvedi, N., Majeed, A., Oliver, N. Toumazou, C., Alberti, K. G., and Johnston, D. G. 2013. “Effectiveness of Mobile Phone Messaging in Prevention of Type 2 Diabetes by Lifestyle Modification in Men in India: A Prospective, Parallel-Group, Randomized Controlled Trial,” *The Lancet Diabetes & Endocrinology* (1:3), pp. 191-198.
- Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. 2010. “Classifying Latent User Attributes in Twitter,” in *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, New York: ACM, pp. 37-44.
- Rharbi, A., Amine, K., Bakkoury, Z., Mikou, A., Kettani, A., and Betari, A. K. “Approaches to Access Biological Data Sources,” Chapter 6 in *Lipoproteins—Role in Health and Diseases*, S. Frank and G. Kostner (eds.) (<https://epdf.pub/lipoproteins-role-in-health-and-diseases.html>).
- Rothe, R., Timofte, R., and van Gool, L. 2018. “Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks,” *International Journal of Computer Vision* (126), pp. 144-157.
- Rundell, K. W., Im, J., Mayers, L. B., Wilber, R. L., Szmedra, L., and Schmitz, H. R. 2001. “Self-Reported Symptoms and Exercise-Induced Asthma in the Elite Athlete,” *Medicine and Science in Sports and Exercise* (33:2), pp. 208-213.
- Ruths, D., and Pfeffer, J. 2014. “Social Media for Large Studies of Behavior,” *Science* (346:6213), pp. 1063-1064.
- Sadat, Y. K., Nikaiein, T., and Karimipour, F. 2015. “Fuzzy Spatial Association Rule Mining to Analyze the Effect of Environmental Variables on the Risk of Allergic Asthma Prevalence,” *Geodesy and Cartography* (41:2), pp. 101-112.
- Salam, M. T., Islam, T., and Gilliland, F. D. 2008. “Recent Evidence for Adverse Effects of Residential Proximity to Traffic Sources on Asthma,” *Current Opinion in Pulmonary Medicine* (14:1), pp. 3-8.
- Schau, H., and Gilly, M. C. 2003. “We Are What We Post? Self-Presentation in Personal Web Space,” *Journal of Consumer Research* (30:3), pp. 385-404.
- Shimrat, M. 1962. “Algorithm 112: Position of Point Relative to Polygon,” *Communications of the ACM* (5:8), p. 434.
- Simon, H. A. 1996. *The Sciences of the Artificial*, Cambridge, MA: MIT Press.
- Sloan, L., Morgan, J., Burnap, P., and Williams, M. 2015. “Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data,” *PLoS ONE* (10:3), p. e0115545.
- Smith, D. K., Seales, S., and Budzik, C. 2017. “Respiratory Syncytial Virus Bronchiolitis in Children,” *American Family Physician* (95:2), pp. 94-99.
- Strano, M. M. 2008. “User Descriptions and Interpretations of Self-Presentation through Facebook Profile Images,” *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* (2:2), Article 5.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution,” *BMC Bioinformatics* (8:25).
- Szczeklik, A., Nizankowska, E., and Duplaga, M. 2000. “Natural History of Aspirin-Induced Asthma. AIANE Investigators. European Network on Aspirin-Induced Asthma,” *European Respiratory Journal* (16:3), pp. 432-436.
- Tang, M., Agrawal, P., and Jain, R. 2015. “Habits Vs. Environment: What Really Causes Asthma?,” Article 30 in *Proceedings of the ACM Web Science Conference*, New York: ACM.
- Walker, S. J. 2014. “Big Data: A Revolution That Will Transform How We Live, Work, and Think,” *International Journal of Advertising* (33:1), pp. 181-183.
- WHO. 2017. “WHO | Asthma,” World Health Organization, Geneva, Switzerland (<http://www.who.int/topics/asthma/en/>; retrieved July 5, 2017).
- Wong, E. C., Palaniappan, L. P., and Lauderdale, D. S. 2010. “Using Name Lists to Infer Asian Racial/Ethnic Subgroups in the Healthcare Setting,” *Medical Care* (48:6), pp. 540-546.
- Wright, R. J., Rodriguez, M., and Cohen, S. 1998. “Review of Psychosocial Stress and Asthma: An Integrated Biopsychosocial Approach,” *Thorax* (53:12), pp. 1066-1074.
- Wu, X., Zhu, X., Wu, G. Q., and Ding, W. 2014. “Data Mining with Big Data,” *IEEE Transactions on Knowledge and Data Engineering* (26:1), pp. 97-107.
- Xiang, L., Sang, J., and Xu, C. 2017. “Demographic Attribute Inference from Social Multimedia Behaviors: A Cross-OSN Approach,” in *Multimedia Modeling*, L. Amsaleg, G. Guo-

- mundsson, C. Gurrin, B. Jónsson, and S. Satoh, Cham, Switzerland: Springer Nature, pp. 515-526.
- Zacharasiewicz, A. 2016. "Maternal Smoking in Pregnancy and its Influence on Childhood Asthma," *ERJ Open Research* (2:3).
- Zhang, W., and Ram, S. 2015. "A Comprehensive Methodology for Extracting Signal from Social Media Text Using Natural Language Processing and Machine Learning," paper presented at the 25th Workshop on Information Technologies and Systems, Dallas, TX, December 12.
- Zhang, W., and Ram, S. 2017. "Domain Adaptation for Signal Extraction from Large Social Media Datasets," paper presented at the Conference on Information Systems and Technology, Houston, TX.

About the Authors

Wenli Zhang is an assistant professor at the Debbie and Jerry Ivy College of Business at Iowa State University. She received her Ph.D. from the University of Arizona in 2018. Her research interests revolve around the areas of data science and information system design, especially in developing techniques based on machine learning, natural language processing, network analysis, and distributed computing for solving real-world problems within the context of healthcare and other business concerns. Her work

appears in journals and conference proceedings including *IEEE Journal of Biomedical and Health Informatics* and *ACM Digital Health*. She received the Best Dissertation Proposal Award by the Workshop on Information Technologies and Systems and the Meritorious Mention by the Association for Information Systems SIG Health.

Sudha Ram is Anheuser-Busch Endowed Professor of MIS, and Entrepreneurship & Innovation in the Eller College of Management at the University of Arizona. She has a joint faculty appointment as a professor of Computer Science. She is the director of INSITE: Center for Business Intelligence and Analytics (www.insiteua.org) at the University of Arizona. Sudha received a Ph.D. from the University of Illinois at Urbana-Champaign in 1985. Her research is in the areas of network science, prediction modeling, machine learning, and big data analytics. Her work uses different methods such as machine learning, statistical approaches, ontologies, and conceptual modeling. She has published articles in such journals as *Communications of the ACM*, *IEEE Intelligent Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *Information Systems*, *Information Systems Research*, *Management Science*, and *MIS Quarterly*. She is an editor for *Journal of Business Analytics* and Fellow of the Association for Information Systems. She was a speaker for a TEDx Talk in December 2013 on "Creating a Smarter World with Big Data."

Appendix A

Existing Studies on Asthma Risk Factors and Triggers

<i>Methods</i>	<i>Cat.</i>	<i>References</i>	<i>Risk Factors/ Triggers</i>
Case-control study (13.87%)	1	Ardura-Garcia et al. 2015; Bener et al. 2007; Lødrup Carlsen et al. 2004; Malling et al. 2010; Ownby et al. 2015; Wang et al. 2001; Whu et al. 2007; Xu et al. 2016	Risk factors
		Lakhanpaul et al. 2017	Triggers
	2	Armentia et al. 2001; Boneberger et al. 2010	Risk factors
	3	Falliers 1973; Fredrickson et al. 2004; Mai et al. 2012	
	5	Sutherland et al. 2004	Triggers
		Webley and Aldridge 2015	
	1, 2, 3	Kamran et al. 2015	Risk factors
	1, 2, 3, 4	Mo et al. 2003	Triggers
	2, 3, 5	Sarafino et al. 2001	
Clinical trial study (1.46%)	2	Krieger et al. 2005	Triggers
	3	Valizadeh et al. 2014	
Cohort study (59.12%)	1	Agache and Ciobanu 2010; Arnedo et al. 2007; DeBaun et al. 2014; Del-Rio-Navarro et al. 2006; Dumanovsky and Matte 2007; Ergöz et al. 2014; Gonzales et al. 2007; Greenblatt et al. 2017; Klinnert et al. 2002; Kozyrskyj et al. 2003; Kühni and Sennhauser 1995; Lara et al. 2006; Loisel et al. 2011; Rolfsjord et al. 2015; Seo et al. 2015; Sims et al. 1999; Xu et al. 2016;	Risk factors
		Braman 2017; Luskin et al. 2014; Sarafino and Goldfedder 1995	Triggers
	2	Andrusaityte et al. 2016; Jackson et al. 2017; Quiralte et al. 2005; Sotir et al. 2003; Uthaisangsook 2010	Risk factors
		Banda et al. 2013; Cabana et al. 2004; Dong et al. 2018; Göksel et al. 2009; Martin et al. 2006; Martin et al. 2013; O'Leary et al. 2012; Shendell et al. 2007; Weiss et al. 2001	Triggers
	3	Acosta et al. 2008; Björkstén et al. 2011; Beuther and Sutherland 2007; Debley et al. 2012; Devereux and Seaton 2005 El-Zein et al. 2017; Gudelj et al. 2012; Hollams et al. 2017; Hung et al. 2010; Kelley et al. 2005; Kozyrskyj et al. 2007; Kozyrskyj et al. 2009; Li et al. 2005; Maher et al. 2004; Ross et al. 2009; Zedan et al. 2012	Risk factors
		Alhekail et al. 2017; Gruber et al. 2016; Peterson et al. 2012	Triggers
	4	Akpınar-Elci et al. 2002; Delclos et al. 2007; Horner 2008; Hovland et al. 2015; Kozyrskyj et al. 2008; Wendt et al. 2012	Risk factors
		Kakumanu et al. 2017; Ritz et al. 2014; Ritz, Kullowatz, Bobb et al. 2008; Ritz et al. 2016; Takaro et al. 2004; Turyk et al. 2013; Vazquez et al. 2017	Triggers
	5	Lukkarinen et al. 2017; Pereira et al. 2007	Risk factors
		Sarafino and Dillon 1998	Triggers
	1, 2	Peroni et al. 2009	Risk factors
	1, 2, 3	Agrawal et al. 2013	

	2, 3	Janssens and Harver 2015; Price et al. 2014; Ritz et al. 2006; Washington et al. 2012	Triggers
	2, 3, 4	Reddy et al. 2017; Ritz, Kullowatz, Kannies, et al. 2008	
	2, 4	Harris et al. 2017	
	2, 4, 5	Chipps et al. 2018	
	2, 5	Janssens et al. 2017	
	3, 4	Warman et al. 2006	Risk factors
Cross-sectional study (5.11%)	1	Higgins et al. 2005; Stridsman et al. 2017	Risk factors
			Triggers
	2	García-Marcos et al. 2005; Rojas Molina et al. 2001	Risk factors
	4	Mohammad et al. 2017	Triggers
	1, 2	Bener et al. 1996	Risk factors
Experiment (2.19%)	3, 5	Rank et al. 2010	Triggers
	1	Polley et al. 2017; Worgall 2017	Risk factors
	2	Redlich 2010	
Literature review (14.60%)		Dautel et al. 1999	Triggers
	1	Durham et al. 2011; Holt and Sly 2011; Howell 2008; Subbarao et al. 2009	Risk factors
	2	Dietert 2011; Horner et al. 2002	Risk factors
		Crocker et al. 2011; Gautier and Charpin 2017	Triggers
	3	Farber et al. 1998; Forno et al. 2014; Peroni et al. 2002; Reisman et al. 2006	Risk factors
		Covar et al. 2005	Triggers
	4	Cullinan 2005; Gergen and Togias 2015	Risk factors
		Janssens and Ritz 2013	Triggers
	1, 2	Guibas et al. 2015; Toskala and Kennedy 2015	Risk factors
	2, 3, 5	Vernon et al. 2012	Triggers
Modeling analysis (2.19%)	3, 5	McCarty and Ferguson 2014	Triggers
	2	Liu et al. 2016	Risk factors
		Brown et al. 2014; Myatt et al. 2008	Triggers

Cat. = Category: 1 = Biological and demographic factors; 2 = Environmental factors; 3 = Behavioral factors; 4 = Social and psychological factors; 5 = Inflammatory and infectious factors.

84.7% of current studies focused on one type of asthma triggers or risk factors, 9.5% analyzed two types and only 5.8% of them studied more than two types of asthma triggers or risk factors.

PubMed Keywords: "asthma trigger(s)" or "asthma risk factor(s)" (accessed April 2018).

References for Appendix A

- Acosta, L. M., Acevedo-García, D., Perzanowski, M. S., Mellins, R., Rosenfeld, L., Cortés, D., Gelman, A., Fagan, J. K., Bracero, L. A., Correa, J. C., Reardon, A. M., and Chew, G. L. 2008. "The New York City Puerto Rican Asthma Project: Study Design, Methods, and Baseline Results," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (45:1), pp. 51-57.
- Agache, I., and Ciobanu, C. 2010. "Risk Factors and Asthma Phenotypes in Children and Adults with Seasonal Allergic Rhinitis," *The Physician and Sportsmedicine* (38:4), pp. 81-86.
- Agrawal, S., Pearce, N., and Ebrahim, S. 2013. "Prevalence and Risk Factors for Self-Reported Asthma in an Adult Indian Population: A Cross-Sectional Survey," *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease* (17:2), pp. 275-282.
- Akpınar-Elci, M., Cimrin, A. H., and Elci, O. C. 2002. "Prevalence and Risk Factors of Occupational Asthma among Hairdressers in Turkey," *Journal of Occupational and Environmental Medicine* (44:6), pp. 585-590.
- Alhekail, G. A., Althubaiti, A., and AlQueflie, S. 2017. "The Association between Body Mass Index and Frequency of Emergency Department Visits and Hospitalization for Asthma Exacerbation in a Pediatric Population," *Annals of Saudi Medicine* (37:6), pp. 415-419.

- Andrusaityte, S., Grazuleviciene, R., Kudzyte, J., Bernotiene, A., Dedele, A., and Nieuwenhuijsen, M. J. 2016. "Associations between Neighbourhood Greenness and Asthma in Preschool Children in Kaunas, Lithuania: A Case-Control Study," *BMJ Open* (6:4), p. e010341.
- Ardura-Garcia, C., Vaca, M., Oviedo, G., Sandoval, C., Workman, L., Schuyler, A. J., Perzanowski, M. S., Platts-Mills, T. A. E., and Cooper, P. J. 2015. "Risk Factors for Acute Asthma in Tropical America: A Case-Control Study in the City of Esmeraldas, Ecuador," *Pediatric Allergy and Immunology: Official Publication of the European Society of Pediatric Allergy and Immunology* (26:5), pp. 423-430.
- Armentia, A., Bañuelos, C., Arranz, M. L., Del Villar, V., Martín-Santos, J. M., Gil, F. J., Vega, J. M., Callejo, A., and Paredes, C. 2001. "Early Introduction of Cereals into Children's Diets as a Risk-Factor for Grass Pollen Asthma," *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology* (31:8), pp. 1250-1255.
- Arnedo, A., Bellido, J. B., Pac, M. R., Artero, A., Campos, J.-B., Museros, L., Puig-Barberà, J., Tosca, R., and Tornador, E. 2007. "Incidence of Asthma and Risk Factors in a Cohort of Schoolchildren Aged from 6-7 Years Old to 14-15 Years Old in Castellón (Spain) Following the International Study of Asthma and Allergies in Childhood (ISAAC)," *Medicina Clinica* (129:5), pp. 165-170.
- Banda, E., Persky, V., Chisum, G., Damitz, M., Williams, R., and Turyk, M. 2013. "Exposure to Home and School Environmental Triggers and Asthma Morbidity in Chicago Inner-City Children," *Pediatric Allergy and Immunology: Official Publication of the European Society of Pediatric Allergy and Immunology* (24:8), pp. 734-741.
- Bener, A., Abdulrazzaq, Y. M., Al-Mutawwa, J., and Debuse, P. 1996. "Genetic and Environmental Factors Associated with Asthma," *Human Biology* (68:3), pp. 405-414.
- Bener, A., Ehlayel, M., and Sabbah, A. 2007. "The Pattern and Genetics of Pediatric Extrinsic Asthma Risk Factors in Polluted Environment," *European Annals of Allergy and Clinical Immunology* (39:2), pp. 58-63.
- Beuther, D. A., and Sutherland, E. R. 2007. "Overweight, Obesity, and Incident Asthma: A Meta-Analysis of Prospective Epidemiologic Studies," *American Journal of Respiratory and Critical Care Medicine* (175:7), pp. 661-666.
- Björkstén, B., Ait-Khaled, N., Innes Asher, M., Clayton, T. O., Robertson, C., and ISAAC Phase Three Study Group. 2011. "Global Analysis of Breast Feeding and Risk of Symptoms of Asthma, Rhinoconjunctivitis and Eczema in 6-7 Year Old Children: ISAAC Phase Three," *Allergologia Et Immunopathologia* (39:6), pp. 318-325.
- Boneberger, A., Radon, K., Baer, J., Kausel, L., Kabesch, M., Haider, D., Schierl, R., von Kries, R., and Calvo, M. 2010. "Asthma in Changing Environments--Chances and Challenges of International Research Collaborations between South America and Europe—Study Protocol and Description of the Data Acquisition of a Case-Control-Study," *BMC Pulmonary Medicine* (10), p. 43.
- Braman, S. S. 2017. "Asthma in the Elderly," *Clinics in Geriatric Medicine* (33:4), pp. 523-537.
- Brown, K. W., Minegishi, T., Allen, J. G., McCarthy, J. F., Spengler, J. D., and MacIntosh, D. L. 2014. "Reducing Patients' Exposures to Asthma and Allergy Triggers in Their Homes: An Evaluation of Effectiveness of Grades of Forced Air Ventilation Filters," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (51:6), pp. 585-594.
- Cabana, M. D., Slish, K. K., Lewis, T. C., Brown, R. W., Nan, B., Lin, X., and Clark, N. M. 2004. "Parental Management of Asthma Triggers within a Child's Environment," *The Journal of Allergy and Clinical Immunology* (114:2), pp. 352-357.
- Chipps, B. E., Haselkorn, T., Rosén, K., Mink, D. R., Trzaskoma, B. L., and Luskin, A. T. 2018. "Asthma Exacerbations and Triggers in Children in TENOR: Impact on Quality of Life," *The Journal of Allergy and Clinical Immunology: In Practice* (6:1), pp. 169-176.e2.
- Collaborating Group ISRDCE. 1997. "Frequency of Risk Factors in Bronchial Asthma in Various Regions of Italy," *Epidemiologia E Prevenzione* (21:4), pp. 243-251.
- Covar, R. A., Macomber, B. A., and Szeffler, S. J. 2005. "Medications as Asthma Triggers," *Immunology and Allergy Clinics of North America* (25:1), pp. 169-190.
- Crocker, D. D., Kinyota, S., Dumitru, G. G., Ligon, C. B., Herman, E. J., Ferdinands, J. M., Hopkins, D. P., Lawrence, B. M., Sipe, T. A., and Task Force on Community Preventive Services. 2011. "Effectiveness of Home-Based, Multi-Trigger, Multicomponent Interventions with an Environmental Focus for Reducing Asthma Morbidity: A Community Guide Systematic Review," *American Journal of Preventive Medicine* (41:2 Suppl. 1), pp. S5-32.
- Cullinan, P. 2005. "Occupational Asthma: Risk Factors, Diagnosis and Preventive Measures," *Expert Review of Clinical Immunology* (1:1), pp. 123-132.
- Dautel, P. J., Whitehead, L., Tortolero, S., Abramson, S., and Sockrider, M. M. 1999. "Asthma Triggers in the Elementary School Environment: A Pilot Study," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (36:8), pp. 691-702.
- DeBaun, M. R., Rodeghier, M., Cohen, R., Kirkham, F. J., Rosen, C. L., Roberts, I., Cooper, B., Stocks, J., Wilkey, O., Inusa, B., Warner, J. O., and Strunk, R. C. 2014. "Factors Predicting Future ACS Episodes in Children with Sickle Cell Anemia," *American Journal of Hematology* (89:11), pp. E212-217.
- Debley, J., Stanojevic, S., Filbrun, A. G., and Subbarao, P. 2012. "Bronchodilator Responsiveness in Wheezy Infants and Toddlers Is Not Associated with Asthma Risk Factors," *Pediatric Pulmonology* (47:5), pp. 421-428.
- Delclos, G. L., Gimeno, D., Arif, A. A., Burau, K. D., Carson, A., Lusk, C., Stock, T., Symanski, E., Whitehead, L. W., Zock, J.-P., Benavides, F. G., and Antó, J. M. 2007. "Occupational Risk Factors and Asthma among Health Care Professionals," *American Journal of Respiratory and Critical Care Medicine* (175:7), pp. 667-675.
- Del-Rio-Navarro, B., Berber, A., Blandón-Vijil, V., Ramírez-Aguilar, M., Romieu, I., Ramírez-Chanona, N., Heras-Acevedo, S., Serrano-Sierra, A., Barraza-Villareal, A., Baeza-Bacab, M., and Sienra-Monge, J. J. L. 2006. "Identification of Asthma Risk Factors in Mexico City in an International Study of Asthma and Allergy in Childhood Survey," *Allergy and Asthma Proceedings* (27:4), pp. 325-333.

- Devereux, G., and Seaton, A. 2005. "Diet as a Risk Factor for Atopy and Asthma," *The Journal of Allergy and Clinical Immunology* (115:6), pp. 1109-1118.
- Dieterl, R. R. 2011. "Maternal and Childhood Asthma: Risk Factors, Interactions, and Ramifications," *Reproductive Toxicology* (32:2), pp. 198-204.
- Dong, Z., Nath, A., Guo, J., Bhaumik, U., Chin, M. Y., Dong, S., Marshall, E., Murphy, J. S., Sandel, M. T., Sommer, S. J., Ursprung, W. W. S., Woods, E. R., Reid, M., and Adamkiewicz, G. 2018. "Evaluation of the Environmental Scoring System in Multiple Child Asthma Intervention Programs in Boston, Massachusetts," *American Journal of Public Health* (108:1), pp. 103-111.
- Dumanovsky, T., and Matte, T. D. 2007. "Variation in Adult Asthma Prevalence in Hispanic Subpopulations in New York City," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (44:4), pp. 297-303.
- Durham, A. L., Wiegman, C., and Adcock, I. M. 2011. "Epigenetics of Asthma," *Biochimica Et Biophysica Acta* (1810:11), pp. 1103-1109.
- El-Zein, M., Conus, F., Benedetti, A., Menzies, D., Parent, M.-E., and Rousseau, M.-C. 2017. "Association Between Bacillus Calmette-Guérin Vaccination and Childhood Asthma in the Quebec Birth Cohort on Immunity and Health," *American Journal of Epidemiology* (186:3), pp. 344-355.
- Ergöz, N., Seymen, F., Gencay, K., Tamay, Z., Deeley, K., Vinski, S., and Vieira, A. R. 2014. "Genetic Variation in Ameloblastin Is Associated with Caries in Asthmatic Children," *European Archives of Paediatric Dentistry: Official Journal of the European Academy of Paediatric Dentistry* (15:3), pp. 211-216.
- Falliers, C. J. 1973. "Aspirin and Subtypes of Asthma: Risk Factor Analysis," *The Journal of Allergy and Clinical Immunology* (52:3), pp. 141-147.
- Farber, H. J., Johnson, C., and Beckerman, R. C. 1998. "Young Inner-City Children Visiting the Emergency Room (ER) for Asthma: Risk Factors and Chronic Care Behaviors," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (35:7), pp. 547-552.
- Forno, E., Young, O. M., Kumar, R., Simhan, H., and Celedón, J. C. 2014. "Maternal Obesity in Pregnancy, Gestational Weight Gain, and Risk of Childhood Asthma," *Pediatrics* (134:2), pp. e535-546.
- Fredrickson, D. D., Molgaard, C. A., Dismuke, S. E., Schukman, J. S., and Walling, A. 2004. "Understanding Frequent Emergency Room Use by Medicaid-Insured Children with Asthma: A Combined Quantitative and Qualitative Study," *The Journal of the American Board of Family Practice* (17:2), pp. 96-100.
- Garcia-Marcos, L., Castro-Rodríguez, J. A., Suarez-Varela, M. M., Garrido, J. B., Hernandez, G. G., Gimeno, A. M., González, A. L., Ruiz, T. R., and Torres, A. M. 2005. "A Different Pattern of Risk Factors for Atopic and Non-Atopic Wheezing in 9-12-Year-Old Children," *Pediatric Allergy and Immunology: Official Publication of the European Society of Pediatric Allergy and Immunology* (16:6), pp. 471-477.
- Gautier, C., and Charpin, D. 2017. "Environmental Triggers and Avoidance in the Management of Asthma," *Journal of Asthma and Allergy* (10), pp. 47-56.
- Gergen, P. J., and Togias, A. 2015. "Inner City Asthma," *Immunology and Allergy Clinics of North America* (35:1), pp. 101-114.
- Göksel, O., Celik, G. E., Erkekol, F. O., Güllü, E., Mungan, D., and Misirligil, Z. 2009. "Triggers in Adult Asthma: Are Patients Aware of Triggers and Doing Right?," *Allergologia Et Immunopathologia* (37:3), pp. 122-128.
- Gonzales, M., Malcoe, L. H., Myers, O. B., and Espinoza, J. 2007. "Risk Factors for Asthma and Cough among Hispanic Children in the Southwestern United States of America, 2003-2004," *Revista Panamericana De Salud Publica = Pan American Journal of Public Health* (21:5), pp. 274-281.
- Greenblatt, R., Mansour, O., Zhao, E., Ross, M., and Himes, B. E. 2017. "Gender-Specific Determinants of Asthma among U.S. Adults," *Asthma Research and Practice* (3), p. 2.
- Gruber, K. J., McKee-Huger, B., Richard, A., Byerly, B., Raczkowski, J. L., and Wall, T. C. 2016. "Removing Asthma Triggers and Improving Children's Health: The Asthma Partnership Demonstration Project," *Annals of Allergy, Asthma & Immunology: Official Publication of the American College of Allergy, Asthma, & Immunology* (116:5), pp. 408-414.
- Gudelj, I., Mrkić Kobal, I., Munivrana Škvorc, H., Miše, K., Vrbica, Z., Plavec, D., and Tudorić, N. 2012. "Intraregional Differences in Asthma Prevalence and Risk Factors for Asthma among Adolescents in Split-Dalmatia County, Croatia," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* (18:4), pp. PH43-50.
- Guibas, G. V., Megremis, S., West, P., and Papadopoulos, N. G. 2015. "Contributing Factors to the Development of Childhood Asthma: Working toward Risk Minimization," *Expert Review of Clinical Immunology* (11:6), pp. 721-735.
- Harris, D. A., Mainardi, A., Iyamu, O., Rosenthal, M. S., Bruce, R. D., Pisani, M. A., and Redlich, C. A. 2017. "Improving the Asthma Disparity Gap with Legal Advocacy? A Qualitative Study of Patient-Identified Challenges to Improve Social and Environmental Factors That Contribute to Poorly Controlled Asthma," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma*, pp. 1-9.
- Higgins, P. S., Wakefield, D., and Cloutier, M. M. 2005. "Risk Factors for Asthma and Asthma Severity in Nonurban Children in Connecticut," *Chest* (128:6), pp. 3846-3853.
- Hollams, E. M., Teo, S. M., Kusel, M., Holt, B. J., Holt, K. E., Inouye, M., De Klerk, N. H., Zhang, G., Sly, P. D., Hart, P. H., and Holt, P. G. 2017. "Vitamin D over the First Decade and Susceptibility to Childhood Allergy and Asthma," *The Journal of Allergy and Clinical Immunology* (139:2), pp. 472-481.e9.

- Holt, P. G., and Sly, P. D. 2011. "Interaction between Adaptive and Innate Immune Pathways in the Pathogenesis of Atopic Asthma: Operation of a Lung/Bone Marrow Axis," *Chest* (139:5), pp. 1165-1171.
- Horner, S. D. 2008. "Childhood Asthma in a Rural Environment: Implications for Clinical Nurse Specialist Practice," *Clinical Nurse Specialist* (22:4), pp. 192-198; quiz 199-200.
- Horner, S. D., Surratt, D., and Smith, S. B. 2002. "The Impact of Asthma Risk Factors on Home Management of Childhood Asthma," *Journal of Pediatric Nursing* (17:3), pp. 211-221.
- Hovland, V., Riiser, A., Mowinckel, P., Carlsen, K.-H., and Lødrup Carlsen, K. C. 2015. "Early Risk Factors for Pubertal Asthma," *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology* (45:1), pp. 164-176.
- Howell, G. 2008. "Nonadherence to Medical Therapy in Asthma: Risk Factors, Barriers, and Strategies for Improving," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (45:9), pp. 723-729.
- Hung, Y.-L., Hsieh, W.-S., Chou, H.-C., Yang, Y.-H., Chen, C.-Y., and Tsao, P.-N. 2010. "Antenatal Steroid Treatment Reduces Childhood Asthma Risk in Very Low Birth Weight Infants without Bronchopulmonary Dysplasia," *Journal of Perinatal Medicine* (38:1), pp. 95-102.
- Jackson, D. J., Gern, J. E., and Lemanske, R. F. 2017. "Lessons Learned from Birth Cohort Studies Conducted in Diverse Environments," *The Journal of Allergy and Clinical Immunology* (139:2), pp. 379-386.
- Janssens, T., Caris, E., Van Diest, I., and Van den Bergh, O. 2017. "Learning to Detect Triggers of Airway Symptoms: The Role of Illness Beliefs, Conceptual Categories and Actual Experience with Allergic Symptoms," *Frontiers in Psychology* (8), p. 926.
- Janssens, T., and Harver, A. 2015. "Effects of Symptom Perception Interventions on Trigger Identification and Quality of Life in Children with Asthma," *Pulmonary Medicine*.
- Janssens, T., and Ritz, T. 2013. "Perceived Triggers of Asthma: Key to Symptom Perception and Management," *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology* (43:9), pp. 1000-1008.
- Kakumanu, S., Antos, N., Szefer, S. J., and Lemanske, R. F. 2017. "Building School Health Partnerships to Improve Pediatric Asthma Care: The School-Based Asthma Management Program," *Current Opinion in Allergy and Clinical Immunology* (17:2), pp. 160-166.
- Kamran, A., Hanif, S., and Murtaza, G. 2015. "Risk Factors of Childhood Asthma in Children Attending Lyari General Hospital," *JPM: The Journal of the Pakistan Medical Association* (65:6), pp. 647-650.
- Kelley, C. F., Mannino, D. M., Homa, D. M., Savage-Brown, A., and Holguin, F. 2005. "Asthma Phenotypes, Risk Factors, and Measures of Severity in a National Sample of US Children," *Pediatrics* (115:3), pp. 726-731.
- Klennert, M. D., Price, M. R., Liu, A. H., and Robinson, J. L. 2002. "Unraveling the Ecology of Risks for Early Childhood Asthma among Ethnically Diverse Families in the Southwest," *American Journal of Public Health* (92:5), pp. 792-798.
- Kozyrskyj, A. L., Ernst, P., and Becker, A. B. 2007. "Increased Risk of Childhood Asthma from Antibiotic Use in Early Life," *Chest* (131:6), pp. 1753-1759.
- Kozyrskyj, A. L., Kendall, G. E., Zubrick, S. R., Newnham, J. P., and Sly, P. D. 2009. "Frequent Nocturnal Awakening in Early Life Is Associated with Nonatopic Asthma in Children," *The European Respiratory Journal* (34:6), pp. 1288-1295.
- Kozyrskyj, A. L., Mai, X.-M., McGrath, P., Hayglass, K. T., Becker, A. B., and Macneil, B. 2008. "Continued Exposure to Maternal Distress in Early Life Is Associated with an Increased Risk of Childhood Asthma," *American Journal of Respiratory and Critical Care Medicine* (177:2), pp. 142-147.
- Kozyrskyj, A. L., Mustard, C. A., and Becker, A. B. 2003. "Childhood Wheezing Syndromes and Healthcare Data," *Pediatric Pulmonology* (36:2), pp. 131-136.
- Krieger, J. W., Takaro, T. K., Song, L., and Weaver, M. 2005. "The Seattle-King County Healthy Homes Project: A Randomized, Controlled Trial of a Community Health Worker Intervention to Decrease Exposure to Indoor Asthma Triggers," *American Journal of Public Health* (95:4), pp. 652-659.
- Kühni, C. E., and Sennhauser, F. H. 1995. "The Yentl Syndrome in Childhood Asthma: Risk Factors for Undertreatment in Swiss Children," *Pediatric Pulmonology* (19:3), pp. 156-160.
- Lakhanpaul, M., Culley, L., Robertson, N., Bird, D., Hudson, N., Johal, N., McFeeters, M., Angell, E., Hamlyn-Williams, C., Abbas, N., Manikam, L., and Johnson, M. 2017. "A Qualitative Study to Identify Parents' Perceptions of and Barriers to Asthma Management in Children from South Asian and White British Families," *BMC Pulmonary Medicine* (17:1), p. 126.
- Lara, M., Akinbami, L., Flores, G., and Morgenstern, H. 2006. "Heterogeneity of Childhood Asthma among Hispanic Children: Puerto Rican Children Bear a Disproportionate Burden," *Pediatrics* (117:1), pp. 43-53.
- Li, Y.-F., Langholz, B., Salam, M. T., and Gilliland, F. D. 2005. "Maternal and Grandmaternal Smoking Patterns Are Associated with Early Childhood Asthma," *Chest* (127:4), pp. 1232-1241.
- Liu, A. H., Babineau, D. C., Krouse, R. Z., Zoratti, E. M., Pongracic, J. A., O'Connor, G. T., Wood, R. A., Khurana Hershey, G. K., Kerckmar, C. M., Gruchalla, R. S., Kattan, M., Teach, S. J., Makhija, M., Pillai, D., Lamm, C. I., Gern, J. E., Sigelman, S. M., Gergen, P. J., Togias, A., Visness, C. M., and Busse, W. W. 2016. "Pathways through Which Asthma Risk Factors Contribute to Asthma Severity in Inner-City Children," *The Journal of Allergy and Clinical Immunology* (138:4), pp. 1042-1050.
- Lødrup Carlsen, K. C., Pettersen, M., and Carlsen, K.-H. 2004. "Is Bronchodilator Response in 2-Yr-Old Children Associated with Asthma Risk Factors?," *Pediatric Allergy and Immunology: Official Publication of the European Society of Pediatric Allergy and Immunology* (15:4), pp. 323-330.

- Loisel, D. A., Tan, Z., Tisler, C. J., Evans, M. D., Gangnon, R. E., Jackson, D. J., Gern, J. E., Lemanske, R. F., and Ober, C. 2011. "IFNG Genotype and Sex Interact to Influence the Risk of Childhood Asthma," *The Journal of Allergy and Clinical Immunology* (128:3), pp. 524-531.
- Lukkarinen, M., Koistinen, A., Turunen, R., Lehtinen, P., Vuorinen, T., and Jartti, T. 2017. "Rhinovirus-Induced First Wheezing Episode Predicts Atopic but Not Nonatopic Asthma at School Age," *The Journal of Allergy and Clinical Immunology* (140:4), pp. 988-995.
- Luskin, A. T., Chipps, B. E., Rasouliyan, L., Miller, D. P., Haselkorn, T., and Dorenbaum, A. 2014. "Impact of Asthma Exacerbations and Asthma Triggers on Asthma-Related Quality of Life in Patients with Severe or Difficult-to-Treat Asthma," *The Journal of Allergy and Clinical Immunology in Practice* (2:5), pp. 544-552.e1-2.
- Maher, J. E., Mullooly, J. P., Drew, L., and DeStefano, F. 2004. "Infant Vaccinations and Childhood Asthma among Full-Term Infants," *Pharmacoepidemiology and Drug Safety* (13:1), pp. 1-9.
- Mai, X.-M., Langhammer, A., Camargo, C. A., and Chen, Y. 2012. "Serum 25-Hydroxyvitamin D Levels and Incident Asthma in Adults: The HUNT Study," *American Journal of Epidemiology* (176:12), pp. 1169-1176.
- Malling, T. H., Sigsgaard, T., Andersen, H. R., Deguchi, Y., Brandslund, I., Skadhauge, L., Thomsen, G., Baelum, J., Sherson, D., and Omland, O. 2010. "Differences in Associations between Markers of Antioxidative Defense and Asthma Are Sex Specific," *Gender Medicine* (7:2), pp. 115-124.
- Martin, M. A., Hernández, O., Naureckas, E., and Lantos, J. 2006. "Reducing Home Triggers for Asthma: The Latino Community Health Worker Approach," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (43:5), pp. 369-374.
- Martin, M. A., Thomas, A. M., Mosnaim, G., Greve, M., Swider, S. M., and Rothschild, S. K. 2013. "Home Asthma Triggers: Barriers to Asthma Control in Chicago Puerto Rican Children," *Journal of Health Care for the Poor and Underserved* (24:2), pp. 813-827.
- McCarty, J. C., and Ferguson, B. J. 2014. "Identifying Asthma Triggers," *Otolaryngologic Clinics of North America* (47:1), pp. 109-118.
- Mo, F., Robinson, C., Choi, B. C., and Li, F. C. 2003. "Analysis of Prevalence, Triggers, Risk Factors and the Related Socio-Economic Effects of Childhood Asthma in the Student Lung Health Survey (SLHS) Database, Canada 1996," *International Journal of Adolescent Medicine and Health* (15:4), pp. 349-358.
- Mohammad, Y., Rafea, S., Latifeh, Y., Khaddam, A., Sawaf, B., Zakaria, M. I., Al Masalmeh, M. S., Fawaz, Y., Allaham, A., Almani, I., El-Tarcheh, H., Ghazal, A., Zaher, A., Rifai, H., Joumah, H., Glockler-Lauf, S. D., and To, T. 2017. "Uncontrolled and Under-Diagnosed Asthma in a Damascus Shelter during the Syrian Crisis," *Journal of Thoracic Disease* (9:9), pp. 3415-3424.
- Myatt, T. A., Minegishi, T., Allen, J. G., and Macintosh, D. L. 2008. "Control of Asthma Triggers in Indoor Air with Air Cleaners: A Modeling Analysis," *Environmental Health: A Global Access Science Source* (7), p. 43.
- O'Leary, R., Wallace, J., and BREATH Study Research Group. 2012. "Asthma Triggers on the Cheyenne River Indian Reservation in Western South Dakota: The Breathing Relief Education and Tribal Health Empowerment (BREATHE) Study," *South Dakota Medicine: The Journal of the South Dakota State Medical Association* (65:2), pp. 57, 59, 61 passim.
- Ownby, D. R., Tingen, M. S., Havstad, S., Waller, J. L., Johnson, C. C., and Joseph, C. L. M. 2015. "Comparison of Asthma Prevalence among African American Teenage Youth Attending Public High Schools in Rural Georgia and Urban Detroit," *The Journal of Allergy and Clinical Immunology* (136:3), pp. 595-600.e3.
- Pereira, M. U., Sly, P. D., Pitrez, P. M., Jones, M. H., Escuto, D., Dias, A. C. O., Weiland, S. K., and Stein, R. T. 2007. "Nonatopic Asthma Is Associated with Helminth Infections and Bronchiolitis in Poor Children," *The European Respiratory Journal* (29:6), pp. 1154-1160.
- Peroni, D. G., Chatzimichail, A., and Boner, A. L. 2002. "Food Allergy: What Can Be Done to Prevent Progression to Asthma?," *Annals of Allergy, Asthma & Immunology: Official Publication of the American College of Allergy, Asthma, & Immunology* (89:6 Suppl 1), pp. 44-51.
- Peroni, D. G., Piacentini, G. L., Bodini, A., and Boner, A. L. 2009. "Preschool Asthma in Italy: Prevalence, Risk Factors and Health Resource Utilization," *Respiratory Medicine* (103:1), pp. 104-108.
- Peterson, M. G. E., Gaeta, T. J., Birkhahn, R. H., Fernández, J. L., and Mancuso, C. A. 2012. "History of Symptom Triggers in Patients Presenting to the Emergency Department for Asthma," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (49:6), pp. 629-636.
- Polley, D. J., Mihara, K., Ramachandran, R., Vliagoftis, H., Renaux, B., Saifeddine, M., Daines, M. O., Boitano, S., and Hollenberg, M. D. 2017. "Cockroach Allergen Serine Proteinases: Isolation, Sequencing and Signalling via Proteinase-Activated Receptor-2," *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology* (47:7), pp. 946-960.
- Price, D., Dale, P., Elder, E., and Chapman, K. R. 2014. "Types, Frequency and Impact of Asthma Triggers on Patients' Lives: A Quantitative Study in Five European Countries," *The Journal of Asthma* (51:2), pp. 127-135.
- Quiralte, J., Llanes, E., Barral, P., Arias de Saavedra, J. M., Sáenz de San Pedro, B., Villalba, M., Florido, J. F., Rodríguez, R., Lahoz, C., and Cárda, B. 2005. "Ole e 2 and Ole e 10: New Clinical Aspects and Genetic Restrictions in Olive Pollen Allergy," *Allergy* (60:3), pp. 360-365.
- Rank, M. A., Wollan, P., Li, J. T., and Yawn, B. P. 2010. "Trigger Recognition and Management in Poorly Controlled Asthmatics," *Allergy and Asthma Proceedings* (31:6), pp. 99-105.
- Reddy, A. L., Gomez, M., and Dixon, S. L. 2017. "An Evaluation of a State-Funded Healthy Homes Intervention on Asthma Outcomes in Adults and Children," *Journal of Public Health Management and Practice* (23:2), pp. 219-228.
- Redlich, C. A. 2010. "Skin Exposure and Asthma: Is There a Connection?," *Proceedings of the American Thoracic Society* (7:2), pp. 134-137.

- Reisman, J., Schachter, H. M., Dales, R. E., Tran, K., Kourad, K., Barnes, D., Sampson, M., Morrison, A., Gaboury, I., and Blackman, J. 2006. "Treating Asthma with Omega-3 Fatty Acids: Where Is the Evidence? A Systematic Review," *BMC Complementary and Alternative Medicine* (6), p. 26.
- Ritz, T., Bobb, C., and Griffiths, C. 2014. "Predicting Asthma Control: The Role of Psychological Triggers," *Allergy and Asthma Proceedings* (35:5), pp. 390-397.
- Ritz, T., Kullowatz, A., Bobb, C., Dahme, B., Magnussen, H., Kanniss, F., and Steptoe, A. 2008. "Psychological Triggers and Hyper-ventilation Symptoms in Asthma," *Annals of Allergy, Asthma & Immunology: Official Publication of the American College of Allergy, Asthma, & Immunology* (100:5), pp. 426-432.
- Ritz, T., Kullowatz, A., Kanniss, F., Dahme, B., and Magnussen, H. 2008. "Perceived Triggers of Asthma: Evaluation of a German Version of the Asthma Trigger Inventory," *Respiratory Medicine* (102:3), pp. 390-398.
- Ritz, T., Steptoe, A., Bobb, C., Harris, A. H. S., and Edwards, M. 2006. "The Asthma Trigger Inventory: Validation of a Questionnaire for Perceived Triggers of Asthma," *Psychosomatic Medicine* (68:6), pp. 956-965.
- Ritz, T., Wittchen, H.-U., Klotzsch, J., Mühlig, S., Riedel, O., and sap-NEEDs study group. 2016. "Asthma Trigger Reports Are Associated with Low Quality of Life, Exacerbations, and Emergency Treatments," *Annals of the American Thoracic Society* (13:2), pp. 204-211.
- Rojas Molina, N., Legorreta Soberanis, J., and Olvera Guerra, F. 2001. "Prevalence and Asthma Risk Factors in Municipalities of the State of Guerrero, Mexico," *Revista Alergia Mexico* (48:4), pp. 115-118.
- Rolfjord, L. B., Skjerven, H. O., Bakkeheim, E., Carlsen, K.-H., Hunderi, J. O. G., Kvenshagen, B. K., Mowinkel, P., and Lødrup Carlsen, K. C. 2015. "Children Hospitalised with Bronchiolitis in the First Year of Life Have a Lower Quality of Life Nine Months Later," *Acta Paediatrica* (104:1), pp. 53-58.
- Ross, K. R., Hart, M. A., Storf-Isner, A., Kibler, A. M. V., Johnson, N. L., Rosen, C. L., Kerckmar, C. M., and Redline, S. 2009. "Obesity and Obesity Related Co-Morbidities in a Referral Population of Children with Asthma," *Pediatric Pulmonology* (44:9), pp. 877-884.
- Sarafino, E. P., and Dillon, J. M. 1998. "Relationships among Respiratory Infections, Triggers of Attacks, and Asthma Severity in Children," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (35:6), pp. 497-504.
- Sarafino, E. P., Gates, M., and DePaulo, D. 2001. "The Role of Age at Asthma Diagnosis in the Development of Triggers of Asthma Episodes," *Journal of Psychosomatic Research* (51:5), pp. 623-628.
- Sarafino, E. P., and Goldfeder, J. 1995. "Genetic Factors in the Presence, Severity, and Triggers of Asthma," *Archives of Disease in Childhood* (73:2), pp. 112-116.
- Seo, Y., Nonaka, M., Tagaya, E., Tamaoki, J., and Yoshihara, T. 2015. "Eosinophilic Otitis Media Is Associated with Asthma Severity and Smoking History," *ORL; Journal for Oto-Rhino-Laryngology and Its Related Specialties* (77:1), pp. 1-9.
- Shendell, D. G., Rawling, M.-M., Foster, C., Bohlke, A., Edwards, B., Rico, S. A., Felix, J., Eaton, S., Moen, S., Roberts, E. M., and Love, M. B. 2007. "The Outdoor Air Quality Flag Program in Central California: A School-Based Educational Intervention to Potentially Help Reduce Children's Exposure to Environmental Asthma Triggers," *Journal of Environmental Health* (70:3), pp. 28-31.
- Sims, J. R., Tibbles, P. M., and Jackman, R. P. 1999. "A Descriptive Analysis of Asthma in the U.S. Navy Submarine Force," *Aviation, Space, and Environmental Medicine* (70:12), pp. 1214-1218.
- Sotir, M., Yeatts, K., and Shy, C. 2003. "Presence of Asthma Risk Factors and Environmental Exposures Related to Upper Respiratory Infection-Triggered Wheezing in Middle School-Age Children," *Environmental Health Perspectives* (111:4), pp. 657-662.
- Stridsman, C., Dahlberg, E., Zandrén, K., and Hedman, L. 2017. "Asthma in Adolescence Affects Daily Life and School Attendance—Two Cross-Sectional Population-Based Studies 10 Years Apart," *Nursing Open* (4:3), pp. 143-148.
- Subbarao, P., Becker, A., Brook, J. R., Daley, D., Mandhane, P. J., Miller, G. E., Turvey, S. E., and Sears, M. R. 2009. "Epidemiology of Asthma: Risk Factors for Development," *Expert Review of Clinical Immunology* (5:1), pp. 77-95.
- Sutherland, E. R., Brandorff, J. M., and Martin, R. J. 2004. "Atypical Bacterial Pneumonia and Asthma Risk," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (41:8), pp. 863-868.
- Takaro, T. K., Krieger, J. W., and Song, L. 2004. "Effect of Environmental Interventions to Reduce Exposure to Asthma Triggers in Homes of Low-Income Children in Seattle," *Journal of Exposure Analysis and Environmental Epidemiology* (14 Suppl 1), pp. S133-143.
- Toskala, E., and Kennedy, D. W. 2015. "Asthma Risk Factors," *International Forum of Allergy & Rhinology* (5 Suppl 1), pp. S11-16.
- Turyk, M., Banda, E., Chisum, G., Weems, D., Liu, Y., Damitz, M., Williams, R., and Persky, V. 2013. "A Multifaceted Community-Based Asthma Intervention in Chicago: Effects of Trigger Reduction and Self-Management Education on Asthma Morbidity," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (50:7), pp. 729-736.
- Uthaisangsook, S. 2010. "Risk Factors for Development of Asthma in Thai Adults in Phitsanulok: A University-Based Study," *Asian Pacific Journal of Allergy and Immunology* (28:1), pp. 23-28.
- Valizadeh, L., Zarei, S., Zamanazadeh, V., Bilan, N., Nasiri, K., and Howard, F. 2014. "The Effects of Triggers' Modifying on Adolescent Self-Efficacy with Asthma: A Randomized Controlled Clinical Trial," *Journal of Caring Sciences* (3:2), pp. 121-129.
- Vazquez, K., Sandler, J., Interian, A., and Feldman, J. M. 2017. "Emotionally Triggered Asthma and Its Relationship to Panic Disorder, Ataques de Nervios, and Asthma-Related Death of a Loved One in Latino Adults," *Journal of Psychosomatic Research* (93), pp. 76-82.
- Vernon, M. K., Wiklund, I., Bell, J. A., Dale, P., and Chapman, K. R. 2012. "What Do We Know about Asthma Triggers? A Review of the Literature," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (49:10), pp. 991-998.

- Wang, T. N., Chao, Y. Y., Wang, T. H., Chen, C. J., and Ko, Y. C. 2001. "Familial Risk of Asthma among Adolescents and Their Relatives in Taiwan," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (38:6), pp. 485-494.
- Warman, K., Silver, E. J., and Wood, P. R. 2006. "Asthma Risk Factor Assessment: What Are the Needs of Inner-City Families?," *Annals of Allergy, Asthma & Immunology: Official Publication of the American College of Allergy, Asthma, & Immunology* (97:1 Suppl 1), pp. S11-15.
- Washington, D., Yeatts, K., Sleath, B., Ayala, G. X., Gillette, C., Williams, D., Davis, S., and Tudor, G. 2012. "Communication and Education about Triggers and Environmental Control Strategies during Pediatric Asthma Visits," *Patient Education and Counseling* (86:1), pp. 63-69.
- Webley, W. C., and Aldridge, K. L. 2015. "Infectious Asthma Triggers: Time to Revise the Hygiene Hypothesis?," *Trends in Microbiology* (23:7), pp. 389-391.
- Weiss, S. T., Horner, A., Shapiro, G., Sternberg, A. L., and Childhood Asthma Management Program (CAMP) Research Group. 2001. "The Prevalence of Environmental Exposure to Perceived Asthma Triggers in Children with Mild-to-Moderate Asthma: Data from the Childhood Asthma Management Program (CAMP)," *The Journal of Allergy and Clinical Immunology* (107:4), pp. 634-640.
- Wendt, J. K., Symanski, E., and Du, X. L. 2012. "Estimation of Asthma Incidence among Low-Income Children in Texas: A Novel Approach Using Medicaid Claims Data," *American Journal of Epidemiology* (176:8), pp. 744-750.
- Whu, R., Cirilo, G., Wong, J., Finkel, M. L., Mendez, H. A., and Leggiadro, R. J. 2007. "Risk Factors for Pediatric Asthma in the South Bronx," *The Journal of Asthma: Official Journal of the Association for the Care of Asthma* (44:10), pp. 855-859.
- Worgall, T. S. 2017. "Sphingolipids, ORMDL3 and Asthma: What Is the Evidence?," *Current Opinion in Clinical Nutrition and Metabolic Care* (20:2), pp. 99-103.
- Xu, H., Radabaugh, T., Lu, Z., Galligan, M., Billheimer, D., Vercelli, D., Wright, A. L., Monks, T. J., Halonen, M., and Lau, S. S. 2016. "Exploration of Early-Life Candidate Biomarkers for Childhood Asthma Using Antibody Arrays," *Pediatric Allergy and Immunology: Official Publication of the European Society of Pediatric Allergy and Immunology* (27:7), pp. 696-701.
- Zedan, Magdy, Nasef, N., El-Bayoumy, M., El-Assmy, M., Attia, G., Zedan, Mohamed, AlWakeel, A., Kandil, S., Laimon, W., and Fouda, A. 2012. "Does Decline of Lung Function in Wheezy Infants Justify the Early Start of Controller Medications?," *Indian Journal of Pediatrics* (79:9), pp. 1176-1180.

Appendix B

Random Forest for Asthma Triggers/Risk Factors Assessment with Relative Importance

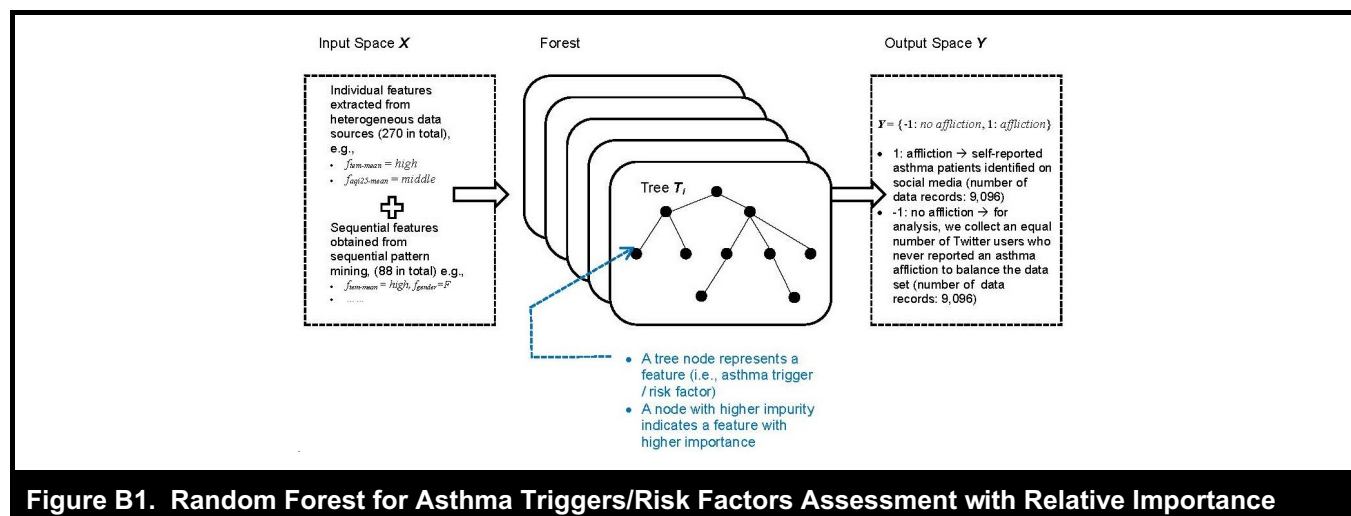
In this study, we use random forest to measure feature importance based on the decrease in average feature impurity.

Using random forest for feature importance measuring has been receiving increased attention in many domains (Strobl et al. 2007). The advantages of using random forest rather than multiple linear regression methods are that it allows nonlinearities and interactions among the data without explicitly identifying them (Grömping 2009) and it requires very little feature engineering and parameter tuning (Breiman 2001).

The Gini importance for classification is a well-known variable importance metric in random forests (Breiman 2001). It is operationalized as follows:

A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \theta_k), k = 1, 2, 3, \dots\}$ where $\{\theta_k\}$ are independent and identically distributed random vectors and each tree $h(x)$ casts a unit vote for the most popular class for input x (Breiman 2001). Every node in $h(x)$ is a condition on a single feature. In a two-way classification problem, every node in the tree $h(x)$ splits the input data x into two so that data records with similar response values are in the same set. Impurity is the measure used to choose the local optimal condition. Gini impurity (originally used by the CART (Chebrolu et al. 2005) algorithm) measures how often a randomly chosen data record from the input x is incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. In a two-way classification problem, the Gini impurity can be calculated as $I_G(p) = 1 - \sum_{i=1}^{I-1} p_i^2$, where p_i is the fraction of data records annotated with class i in the data set. Gini impurity (I_G) reaches zero (the minimum) when all the data records in a tree node fall into a single class. Thus, when training a tree $h(x)$, we can calculate how much each feature decreases the weighted impurity in a tree. For a forest $\{h(x, \theta_k), k = 1, 2, 3, \dots\}$, the impurity decrease from each feature can be averaged and the features are ranked according to Gini impurity.

In this study, we have already identified self-reported asthma patients on social media (see “Identifying Self-Reported Asthma Patients on Social Media” in the paper). So the random forest classification target is already known (i.e., patients with asthma affliction or non-affliction). The focus of this study is to measure feature (i.e., asthma triggers and risk factors) importance based on Gini impurity. Figure B1 illustrates the use of random forest in this study.



References for Appendix B

- Breiman, L. 2001. "Random Forests," *Machine Learning* (45:1), pp. 5-32.
- Chebrolu, S., Abraham, A., and Thomas, J. P. 2005. "Feature Deduction and Ensemble Design of Intrusion Detection Systems," *Computers & Security* (24:4), pp. 295-307.
- Grömping, U. 2009. "Variable Importance Assessment in Regression: Linear Regression versus Random Forest," *The American Statistician* (63:4), pp. 308-319.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics* (8:1), p. 25.

Copyright of MIS Quarterly is the property of MIS Quarterly and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.