KDD 2014 Workshop KDDBHI

Big Data Analytic Technology For Bioinformatics and Health Informatics (KDDBHI)

Proceedings of Workshop

Big Data Analytic Technology for Bioinformatics and Health Informatics

For better quality of life and healthier world

Editors: Xin Deng and Donghui Wu

August 2014

New York, New York

Organization Committee

Chair: Dr. Xin Deng, Research Scientist, LexisNexis | Risk Solutions | Healthcare, Orlando, FL Xin.Deng@lexisnexis.com

Dr. Donghui Wu, Senior Director, Statistical Modeling, LexisNexis | Risk Solutions | Healthcare, Orlando, FL Donghui.Wu@lexisnexis.com,

Dr. Jianlin Cheng, Associate Professor, Computer Science Department, Informatics Institute, C. Bond Life Science Center, University of Missouri- Columbia

Dr. Sumit Kumar Jha, Assistant Professor, Computer Science Department, University of Central Florida, Orlando, FL

Program Committee

Chair: Dr. Xin Deng, Research Scientist, LexisNexis | Risk Solutions | Healthcare, Orlando, FL

Dr. Mohammed J. Zaki, Professor, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY

Dr. Donghui Wu, Senior Director, Statistical Modeling, LexisNexis | Risk Solutions | Healthcare

Dr. Ran Duan, Research Scientist, Eli Lilly and Company, Indianapolis, Indiana

Dr. Nan Zhao, Assistant Research Professor, Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University

Jing Han, Doctoral Student, University of Missouri-Columbia, Columbia, MO

Faraz Hussain, Graduate Student, EECS Department, University of Central Florida, Orlando, FL

Dr. Sumit Kumar Jha, Assistant Professor, Computer Science Department, University of Central Florida, Orlando, FL

Sponsor

LexisNexis Risk Solutions and HPCC Systems

lexisnexis.com/risk hpccsystems.com

Workshop Schedule

2:00 – 2:10 **Opening remark, Dr. Xin Deng (Chair of KDDBHI Workshop)**

2:10 – 2:40 **Keynote: Dr. Aidong Zhang, SUNY Distinguished Professor and Chair**

Evolutionary Analysis of Functional Modules in Dynamic Protein Interaction Networks and Its Applications in Health

2:40 – 4:00 **Paper session**

2:40 – 3:00 **Bayesian Model Averaging methods and R package for gene network construction**
Ka Yee Yeung, Chris Fraley, William Chad Young, Roger Bumgarner and Adrian E.Raftery.

3:00 – 3:20 **Multiuse Data Mining Framework and Infrastructure for a Medical Registry**
Hubert Kordylewski, Richard Dutton and Benjamin Westlake.

3:20 – 3:40 **Discovery of Disease Co-occurrence Patterns from Electronic Healthcare Reimbursement Claims Data**
Arvind Ramanathan, Laura L. Pullum, Tanner C. Hobson, Shannon P. Quinn, Chakra S. Chennubhotla and Silvia Valkova

3:40 – 4:00 **Towards Automatic Annotation of Clinical Interview Transcripts**
Alexander Kotov, April Carcone and Ming Dong.

4:00 – 4:30 **Break (HPCC Demo)**

4:30 – 5:00 **Keynote: Dr. Flavio Villanstre, VP Technology, LexisNexis Risk Solutions**

Uncovering HealthCare fraud and crime through Big Data analytics

5:00 – 6:00 **Panel Discussion**

Moderator: Dr. Donghui Wu

Panelist: Dr. Aidong Zhang

Dr. Flavio Villanustre

Dr. Yindalon Aphinyanaphongs

Dr. Paul Bradley

6:00 **Workshop conclusion**

Workshop Keynote Presentations

Title: Evolutionary Analysis of Functional Modules in Dynamic Protein Interaction Networks and Its Applications in Health

Dr. Aidong Zhang

SUNY Distinguished Professor and Chair

Department of Computer Science and Engineering

State University of New York at Buffalo

Abstract: In recent years, with the high-throughput gene and protein screening and detection techniques, the volume of the gene and protein data has been increased dramatically. There are many gene and protein databases publicly available. Such massive data provide us with the opportunity to systematically analyze the structure of a large living system and also allows us to use it to understand essential principles like genetic interactions, functions, functional modules, gene products, and cellular pathways. In this talk, I will present our ideas and approaches to tracking the evolutionary process of protein functional modules over time. Analyzing the evolutionary pattern of protein functional modules detected over time can help us discover underlying evolutionary trends or behaviors of functional modules in response to different diseases. By learning the evolution pattern of protein function modules for various diseases, we can solve some classical health problems with new perspectives. By detecting the most stable/unstable functional modules and by examining our module evolution nets we can provide new findings to classical health problems. For example, we can find out some key points which could distinguish two ambiguous diseases or understand the underlying similarity between two diseases and what make them different/similar. We can perform early detection of a certain kind of disease. We can focus on the outstanding differences of the patient's PPI networks at the critical beginning timestamp of a disease to find the change pattern of functionally related modules during the early disease stage. Given the patient's early stage PPI network of an unclear disease, we could tell that what the disease is. In addition, we could predict the mutation risk of a certain kind of disease.

Short Bio: Dr. Aidong Zhang is a SUNY Distinguished Professor and Chair of Computer Science and Engineering at State University of New York (SUNY) at Buffalo (the highest academic rank available for any faculty member in the State University of New York System). Her research interests include bioinformatics, health Informatics, data mining,

multimedia and database systems, and content-based image retrieval. She is an author of over 250 research publications in these areas. She has chaired or served on over 100 program committees of international conferences and workshops, and currently serves several journal editorial boards. She has published two books "Protein Interaction Networks: Computational Analysis" (Cambridge University Press, 2009) and "Advanced Analysis of Gene Expression Microarray Data" (World Scientific Publishing Co., Inc. 2006). Dr. Zhang is a recipient of the National Science Foundation CAREER award and State University of New York (SUNY) Chancellor's Research Recognition award. Dr. Zhang is an IEEE Fellow.

Title: Uncovering HealthCare fraud and crime through Big Data analytics

Dr. Flavio Villanustre

VP Technology Architecture & Product

LexisNexis and HPCC Systems

Abstract: Drug trafficking, fraud, waste and abuse are significant challenges to the Public Health System across the world, costing hefty sums of money and lives every year. Unfortunately, in many cases, traditional indicators are ineffective to identify and prevent these cases, or even mitigate their impact. Industry experts agree that these types of activities usually don't happen in isolation, and that more than one individual are often accomplices in these crimes. In addition to this, social influence can be traced to the way these groups of individuals connect and organize.

Recent developments in Big Data analytics and large-scale graph processing have introduced novel approaches to solving this set of problems, helping identify non-obvious relationships in the data, to uncover organized crime groups behind these types of criminal activities. During this presentation, the audience will be introduced to leading-edge research in this field, real-world use cases and some of the challenges and opportunities when using Big Data to tackle this significant burden to society.

Short Bio: Dr. Flavio Villanustre, VP Technology Architecture & Product, for LexisNexis and HPCC Systems. In this position, Flavio is responsible for Information and Physical Security, overall infrastructure strategy and new product development. Prior to 2001, Dr. Villanustre served in different companies in a variety of roles in infrastructure, information security and information technology. In addition, Dr. Villanustre has been involved with the open source community for over 15 years through multiple initiatives. Some of these include founding the first Linux User Group in Buenos Aires (BALUG) in 1994, releasing several pieces of software under different open source licenses, and evangelizing open source to different audiences through conferences, training and education. Prior to his technology career, Dr. Villanustre was a neurosurgeon.

Panelists:

Dr. Yindalon Aphinyanaphongs

Dr. Yindalon Aphinyanaphongs is a Research Assistant Professor in the Department of Medicine at NYU. He is co-Director of the Evidence-Based Medicine Information Retrieval and Scientometrics Laboratory (EBMIRSL) in the Center for Health Informatics and Bioinformatics (CHIBI). Before coming to NYU, he earned BS and MS degrees in engineering, and later MD, MS, and PhD degrees in Biomedical Informatics at Vanderbilt University.

Dr. Aphinyanaphongs current research interests include medical predictive modeling in support of clinical decisions, information retrieval in support of evidence based medicine, and substance abuse followup and monitoring through text messaging and social media.

Dr. Paul Bradley

Dr. Paul Bradley is a co-founder, and Chief Data Scientist at MethodCare where he oversees the research and development functions of MethodCare, including the development of new processes, technologies and products. Paul also keeps MethodCare at the forefront of the most recent predictive analytics, data mining advances and industry trends.

Dr. Bradley is Industry & Government Track Invited Talks Chairs of KDD 2014, also served as Area Editor for the International Journal on Data Mining and Knowledge Discovery and as Associate Editor of SIGKDD Explorations, and was KDD-2003 Industrial Track Co-chair, KDD-2001 Exhibits Chair and KDD-2000 Publicity Chair.

Discovery of Disease Co-occurrence Patterns from Electronic Healthcare Reimbursement Claims Data

Arvind Ramanathan*, Laura L. Pullum*, Tanner C. Hobson
Computational Science & Engineering Division, Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
{ramanathana,pullumll,hobsontc}@ornl.gov

Shannon P. Quinn, Chakra S. Chennubhotla
Department of Computational & Systems Biology, University of Pittsburgh,
Pittsburgh, Pennsylvania, USA
{spq,chakracs}@pitt.edu

Silvia Valkova
IMS Government Solutions
Plymouth Meeting, Pennsylvania, USA
SValkova@theimsinstitute.org

ABSTRACT

Effective public health surveillance is important for national security. With novel emerging infectious diseases being reported across different parts of the world, there is a need to build effective bio-surveillance systems that can track, monitor and report such events in a timely manner. Additionally, there is a need to identify susceptible geographic regions/populations where these diseases may have a significant impact and design preemptive strategies to tackle them. With the digitization of health related information through electronic health records (EHR) and electronic healthcare claim reimbursements (eHCR), there is a tremendous opportunity to exploit these datasets for public health surveillance. In this paper, we present our analysis on the use of eHCR data for bio-surveillance by studying the 2009-2010 H1N1 pandemic flu season. We present a novel approach to extract spatial and temporal patterns of flu incidence across the United States (US) from eHCRs and find that a small, but distinct set of break-out patterns govern the flu and asthma incidence rates across the entire country. Further, we observe a distinct temporal lag in the onset of flu when compared to asthma across geographic regions in the US. The patterns extracted from the data collectively indicate how these break-out patterns are coupled, even though the flu represents an infectious disease whereas asthma represents a typical chronic condition. Taken together, our approach demonstrates how mining eHCRs can provide novel insights in tackling public health concerns.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Clustering

Keywords

Public health surveillance, electronic healthcare claims reimbursement, spatial and temporal patterns

KDD '14 New York, USA

INTRODUCTION

Infectious diseases pose a serious challenge for public health officials and governments [16]; in particular, the emergence (and re-emergence) of novel strains of various pathogens including flu and West Nile viruses, and multi-drug resistant strains of bacteria (causing tuberculosis and other diseases) pose serious threats for national security [24, 14, 11]. Additionally, there has been a significant increase in the number of allergies (such as asthma and food related allergies [13]) and other chronic conditions (e.g., diabetes, cancer) within the United States (US) in the last decade [6]. With an estimated 50-60 million patients diagnosed every year and over $100 billion spent on medical expenses yearly, the combined effect of these diseases creates an extraordinary socio-economic and financial burden [6, 13]. Therefore, there is an urgent need to develop effective bio-surveillance systems that can identify, monitor and track such diseases [12].

Digital public health surveillance is emerging as an important tool for tracking, monitoring and driving decisions regarding emerging infectious disease spread within geographically distributed populations [7]. Many bio-surveillance systems rely on the use of event-based, unstructured digital data (such as news feed aggregators, internet search patterns of users, and social media) [7]. However, with the digitization of health-related information through electronic health records (EHR) and electronic healthcare claim reimbursements (eHCR), there is a tremendous opportunity to seek, collect, monitor and analyze these large-scale datasets for public health surveillance. EHRs refer to an individual patient's detailed medical history, collected and aggregated at medical facilities, whereas eHCRs refer to electronic records of claim transactions processed by retail pharmacies to dispense prescription drugs to patients. eHCR transactions can also include diagnostic information (e.g., when a patient visits his/her doctor) and therefore captures rich and timely information regarding prevailing medical conditions within any given geographic location.

In spite of these advantages, the use of EHR and eHCR datasets for bio-surveillance is still in its early stages [8]. Privacy and security concerns within EHR and eHCR systems have made it tremendously challenging to engage local and public health departments in effectively collecting, sharing and disseminating bio-surveillance related data [15]. eHCR transaction datasets have been routinely used in the context of tracking and analyzing pharmacy prescriptions and understanding drug efficacy (e.g., [21, 10, 5, 3, 4]); however, very little research has been carried out in terms of using them as potential data sources for digital public health surveillance. A

recent study showed that retail pharmacy sales data can be used as a reliable measure for syndromic surveillance; specifically, the aggregate counts of prescription sales of four antiviral drugs for influenza correlated well with Google Flu Trends [9, 20]. However, given the concerns with Google Flu [17], there is a need to develop alternate strategies to evaluate eHCRs in tracking flu (and other diseases).

In this paper, we present an analysis of eHCRs from the 2009-2010 H1N1 pandemic flu season. Using eHCR datasets from IMS Health that capture electronic diagnostic reimbursement claims, we show that the influenza like illnesses (ILI) indicators determined from eHCRs correlate well with the publicly available Centers for Disease Control (CDC) ILINet surveillance data [1]. We also study co-occurring patterns of ILI (infectious disease) and asthma (chronic condition) using the IMS Health eHCRs. We also describe an approach to automatically identify spatial and temporal patterns from the eHCR datasets for the 2009-2010 influenza and study its interrelationship with asthma incidence in the same time period. Apart from discovering a very small number of distinct spatial and temporal patterns, our analysis shows a distinct lag in the temporal patterns of asthma and flu; i.e., we find that a peak in the number of diagnosed flu cases followed a peak in the number of diagnosed asthma cases. Further, we also show that a small number of specific regions within the US have vulnerability to the co-occurrence of flu and asthma, indicating a prior susceptibility for respiratory conditions to co-occur. Taken together, our results show how eHCR datasets can be analyzed for public health surveillance.

DATA

IMS Health is a leading consolidator of eHCRs within the US. With over 55-60 million eHRCs collected every week, this proprietary dataset constitutes a unique resource for public health surveillance. The diagnosis eHRCs (referred to as Dx data) process data from over 500,000 medical practitioners/year and received from all parts of the US, including rural areas. The Dx dataset consists of over 1 billion eHRCs collected annually and represents more than 165 million unique patients. Additionally, IMS Health uses proprietary technology to protect patient privacy concerns and all of the data is HIPAA-compliant.

In this study, we analyzed the IMS Health Dx data from the 2009 - 2010 pandemic (H1N1) flu season. The specific dates covered as part of the study include Apr 1, 2009 - Mar 31, 2010 with a total of nearly one billion records. We processed the Dx data and parsed out for influenza (ICD9 codes 486XX and 488XX) and asthma (ICD9 codes 493XX) related records. Although ICD9 codes for flu can potentially include other diagnostic codes, we specifically chose only those ICD9 codes that corresponded to hospital diagnosed cases of the flu. For flu, we obtained a total of over 6 million individual records (throughout the US). For asthma, we obtained a total of over 10 million individual records. In order to organize the data based on location specific information, we used the zip code corresponding to the patient's service provider (i.e., a medical practitioner/physician), since the provider's five digit zip code is more specific than the patient's three digit zip code directly accessible from the Dx data. It is reasonable to assume that the location of the patient's service provider/pharmacy is most likely to be co-located, unless the patient remotely consults with his/her service provider (only 0.0001% of the total records had different 3 digit zip codes available for the patient and service provider).

We organized the resulting flu and asthma datasets into corresponding matrices, $\mathbf{A}_{flu}$ and $\mathbf{A}_{asthma}$, where the rows represented the number of days and the columns represented the total number of zip codes. In order to characterize the co-occurrence of asthma and flu, we obtained a list of common zip codes between $\mathbf{A}_{flu}$ and $\mathbf{A}_{asthma}$ and considered only those zip codes that had more than 10 reported cases of either diagnostic code set. Of the 29,761 general zip codes in the US, the IMS diagnostic dataset covered 14,098 zip codes with statistically significant data for both flu and asthma, covering about 47% of the US.

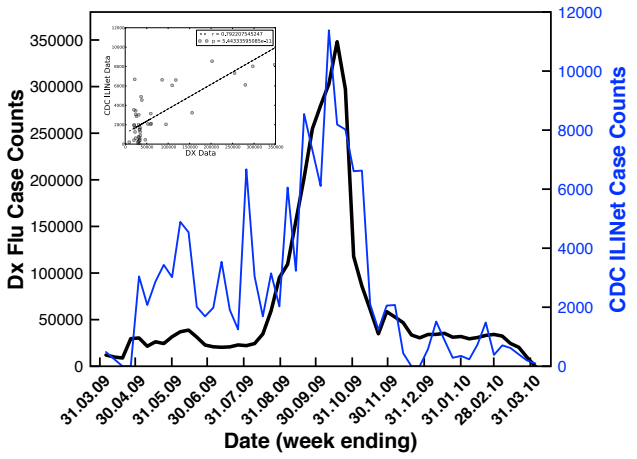


Figure 1: A summary of flu incidence counts from IMS Health Dx data (black line) compared against CDC ILINet data (blue line) showing similar temporal trends during the 2009-2010 pandemic flu season.

RESULTS

Flu and asthma case-counts during the 2009-2010 H1N1 pandemic season

Comparison of IMS Health Dx data with CDC influenza surveillance network

We began by examining if we could rely on the IMS Health Dx data to provide insights into case counts of the 2009-2010 flu season. This is particularly important to ensure that the data is indeed comparable to traditional methods of data collection. We measure the total number of cases (per week) from the Dx data, based on the case definition described in the Data section and compare it against the CDC ILINet data for the same period. ILINet has been a standard way of monitoring the entire nation for the flu and this comparison serves our purpose of checking if the data obtained from Dx and CDC ILINet are similar. Instead of comparing the actual numbers, we observe whether the overall trends in the rise and fall of the flu cases are similar. As shown in Fig. 1, the overall trend within the CDC data matches closely with respect to the Dx data, for the H1N1 flu within the US. The flu season of 2009-2010 was particularly severe as a consequence of the novel H1N1 viral strain of that season, which was a unique combination of flu viruses never before identified in animals or humans [2]. This viral strain spread throughout the world. Within the US, the estimated number of cases of the H1N1 flu was between 42 and 86 million, with about 192,000 to 398,000 hospitalizations. As shown above, although the magnitudes of the IMS Health Dx data and CDC ILINet data are different, the trend observed in the rise/fall of the flu pandemic within the US is similar (inset in Fig. 1, correlation value of 0.8; p-value=5.43E-11). It is important to note that the Dx data is more comprehensive (covers about 47% of the entire country on

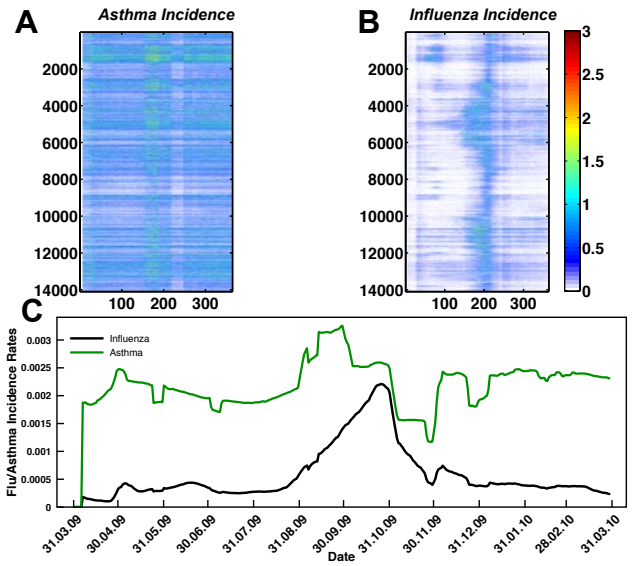


Figure 2: A summary of the temporal trend of flu and asthma incidence from the Dx dataset at the zip code level. (A) Summary of case counts of asthma incidences whereas (B) shows the number of influenza cases reported per zip code (both panels are in log scale). (C) Summary of temporal trends observed from the flu (black) and asthma (green) case counts; note that we have summarized the data using a moving average window of 7 days (to overcome gaps in the IMS Dx data based on reports received throughout the week) and normalized the results based on fraction of total case counts.

an average for the time-period examined) than the CDC ILINet data (we only use publicly available information from the CDC website) and hence one may need to interpret the similarity with caution. A similar analysis for asthma, however, is difficult to perform with the current data because the CDC reports, and IMS Health Dx data reflect different metrics. Hence, for this study, we have not performed a comparative analysis of the asthma incidence.

Summary of flu and asthma incidence rates from diagnostic data

We summarize the flu and asthma from the diagnostics (Dx) data as a function of daily incidence and as a matrix in Fig. 2. Clearly, the number of cases for both asthma (Fig. 2A) and flu (Fig. 2B) at a local zip-code level is sparse. While influenza rates rise sharply during the Aug-Sep 2009 time-frame, the number of asthma cases observed from the data shows more or less a uniform distribution throughout the year, except for a slight increase and decrease around the time of the pandemic flu season. Further, we find that there is a temporal trend where we observe that the peak number of asthma incidences (within the Dx data) lags behind by about 3 weeks when compared to the peak number of influenza incidences (Fig. 2C). Note that for Fig. 2C, we present the data that was temporally averaged by 7 days (to account for lag times within diagnostic data reporting within the IMS Health datasets). We note that even without the temporal averaging, these trends are observed (both at state and national levels).

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

Temporal patterns in flu and asthma incidence

Identifying optimal subspace and cross validation

The dimensionality of the data for each of the matrices ($\mathbf{A}_{flu,asthma}$) is $N_z \times N_t$ where N_z represents the total number of zip codes (14,098) and N_t represents the time points (365 days starting from Apr 1 2009 to Mar 31 2010). We hypothesized that the flu incidence patterns would be composed of discrete spatial and temporal patterns, especially given the geographic size and spread of the US as well as the fact that influenza occurrence is a highly complex process. Further, given prior knowledge that there were distinct 'peaks' associated with the 2009-2010 pandemic, it was reasonable for us to use techniques that could elucidate discrete, yet sparse spatial and temporal patterns from this high dimensional data. Additionally, the entries within each of these matrices are non-negative (i.e., it is not possible to obtain a negative count for the number of patients with the flu or asthma). For this purpose, we used non-negative matrix factorization (NMF), a technique that can extract low-rank approximations from the data.

Given a data matrix $\mathbf{A}$ with non-negative entries ($N_z \times N_t$ dimensions), NMF finds low-rank approximations of the form $\mathbf{A} \approx \mathbf{WH}$, where $\mathbf{W}$ ($N_z \times s$) captures spatial patterns and $\mathbf{H}$ ($s \times N_t$) represents temporal patterns within the data. We used the alternate least squares algorithm proposed by Paatero [19, 18], available as part of the standard Matlab (Mathworks, Inc.). Although the size of the $A_{asthma,flu}$ are quite large, we did not find the speed of convergence as a significant problem. We used a stopping value of 1000 as the maximum number of iterations. To identify the appropriate subspace (s) dimensions for the original data, we iterated over $s = 1 \ldots 15$ for both matrices, dividing the data into random yet equal-sized training and testing data. We tracked the residual errors using Frobenius norm for both training and testing data. For each choice of s, we repeated this process 100 times. Once the optimal s was selected, we report the most stable version of the basis matrices by computing the KL divergence between every pair of the 100 instances of $\mathbf{W}$ from the training dataset and picking the $\mathbf{W}$ with the lowest KL divergence value.

Distinct breakout patterns govern flu and asthma incidence

NMF offers a convenient framework to interpret the incidence of flu and asthma throughout the US during the 2009-2010 time period. In particular, it provides a small number of basis vectors that describe temporal ($\mathbf{H}_{flu,asthma}$) and spatial ($\mathbf{W}_{flu,asthma}$) breakout patterns. In this context, break-out patterns refer to only the total case counts that have been obtained from the diagnostic data. Note that while flu is an infectious disease, asthma is typically a chronic condition; hence, the breakout patterns described here do not capture the traditional epidemiological definitions/measures of disease spread, but indicate the global trends in occurrences of either flu or asthma.

For both asthma and flu, based on the procedure outlined above we decided that the optimal subspace, $s = 5$ sufficiently captured the underlying spatial and temporal patterns in the data. A summary of the five temporal patterns is depicted in Fig. 3 for both (A) asthma and (B) flu. One of the notable observations from Fig. 3 is that the temporal signatures are very distinct in terms of describing the overall occurrence of flu and asthma in the 2009-2010 season. The asthma incidence patterns suggest a strong peak around days 150-180 (Aug-Sep 2009) in both $\mathbf{H}^3_{asthma}$ and $\mathbf{H}^4_{asthma}$ basis vectors. Additionally in basis vector $\mathbf{H}^5_{asthma}$, we observe a high incidence of asthma around days 10-45 (Apr-May 2009) time-frame.

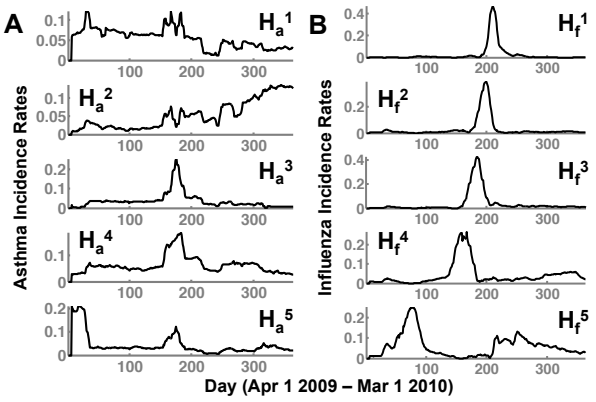


Figure 3: NMF summary of temporal patterns (H) in both asthma (left hand panel) and flu (right hand panel).

The flu occurrence across the nation indicates at least three distinct peaks, ranging from days 180-210 (Sep-Oct 2009), 150-180 (Aug-Sep 2009) and 90-100 (Jun-Jul 2009). The temporal patterns from the flu data indicate that there is a distinct early onset of the epidemic ($\mathbf{H}^5_{flu}$), followed by several waves at later time-periods ($\mathbf{H}^{1-4}_{flu}$), which all have their own distinct temporal signatures. Thus, each of the basis vectors (in the flu dataset) captures a unique temporal break-out pattern that captures ağ different phase of the 2009-2010 flu epidemic, similar to previously reported studies in the spread of influenza [23].

Comparing the flu and asthma break-out patterns suggests that there is a distinct overlap in the incidence of flu and asthma around the Aug-Sep 2009 time-frame. Further, comparing $\mathbf{H}^5_{asthma}$ and $\mathbf{H}^5_{flu}$ also indicates that even during the early onset of the flu (days 90-100; Jun-Jul 2009), there is a marked increase in the asthma incidence rates around days 10-45 (Apr-May 2009). Although from Fig. 2C we see that the overall trend indicates that the peak of asthma incidence precedes the peak of flu season, the analysis presented here further suggests that this precedence may be a distinct factor influencing the susceptibility of flu occurrence within some regions.

Geographic patterns of flu and asthma incidence

In the previous section, we described the temporal patterns of flu and asthma incidence within the US. In this section, we examine the spatial patterns observed from the data. In particular, we present insights into the spatial patterns that we observe from NMF. We also discuss specific regions within the US that showed elevated flu levels following an asthma outbreak. Together, these observations provide for developing a comprehensive picture of how these two diseases may be correlated.

Distinct spatial patterns of flu occurrence

The spatial patterns summarized by NMF depict a distinct separation between the asthma and flu incidence. As shown in Fig. 4, each $\mathbf{W}$, can be mapped onto the specific zip code and provides a geographic interpretation of the results presented above. Each dot represents a specific zip code examined and the intensity of the color indicates a higher occurrence of the flu/asthma (blue indicates lower and red indicates higher incidence). Note that both the asthma and flu incidence maps are drawn to the same color scale (as indicated by the color bar in Fig. 4).

We note that densely populated areas (such as New York, Florida and California) constitute common grounds for the temporal patterns observed in Fig. 3. In particular, throughout the northeast, southeast, west and central US, asthma patterns are widespread; however, the spatial patterns for influenza across the entire US are quite discrete. Notably, several north-east states do not exhibit any patterns observed in $\mathbf{W}^{2-4}_{flu}$. Additionally, the occurrence of $\mathbf{W}^4_{flu}$ is almost exclusively in the southern regions, with cases detected in both southeast and southwest (California). Interestingly, the temporal patterns from the south east constitute the time-frame of Aug-Sep 2009, which signified the beginning of school season within the same region, leading to the unique spatial patterns observed here. The other interesting aspect observed from our analysis is the early onset of the flu in some northeastern states (notably New York and New Jersey) as well as southwest (California), is captured by $\mathbf{W}^5_{flu}$, indicating that this early onset also meant a sustained flu in the later part of the season (around Feb-Mar 2010) in these regions (Fig. 3B, $\mathbf{H}^5_{flu}$).

As part of the analysis, we have also highlighted zip codes where flu and asthma patterns occur concurrently. These regions include parts of the northeast (specifically, southeastern New York, New Jersey, Delaware, southern parts of New Hampshire, Connecticut and Pennsylvania), southeast (Tennessee, Georgia, North and South Carolinas, Florida, and south central parts of Virginia) and the west-coast area (California, Oregon and Washington states). These large, geographic regions constitute a majority of the places where the co-occurrence of the flu and asthma follow a clear temporal trend where by a peak in asthma diagnoses is subsequently followed by a peak in the flu diagnoses. At this time, because we have not integrated socio-economic/census data into our analysis, is difficult for us to speculate whether particular demographic factors (e.g., age-group, socio-economic background or other factors), population density or other environmental and climatic factors within these regions lead to the observed patterns.

We also note the sparse coverage with respect to the diagnostic data across the rest of the country - and these regions also constitute large parts of the US where the population density is also quite low. A more systematic analysis of the variation in population of these regions, followed by a statistical comparison with the flu and asthma diagnostic data would be necessary to draw additional conclusions regarding these spatial and temporal patterns to understand further details on the epidemiological significance of these spatial and temporal patterns.

SUMMARY

In this study we examined whether eHCRs can be used for public health surveillance. For this purpose, we examined the eHCR transactions provided by IMS Health and showed that the diagnostics (Dx) data, which summarizes primary care visits by patients are comparable to standard public health surveillance data such as the CDC ILINet. We have shown that consolidated eHCRs at local (zip code level information) to regional (county, metropolitan, city, state, etc.) to national levels can be used to assess how infectious diseases like the flu may spread. Unlike aggregating web-based search patterns by users, or the use of social media, our approach relies on using traditional methods of surveillance and in particular, uses patients' visits to their medical practitioners for surveillance of diseases such as influenza. Further, the approach here shows that we can track these conditions (and others) without compromising individual privacy (or medical history).

We have also shown that eHCRs could be used to study co-occurrence patterns of asthma and influenza. While both affect the respiratory systems of patients, asthma is a chronic condition whereas the flu is an infectious disease. From the analysis of the

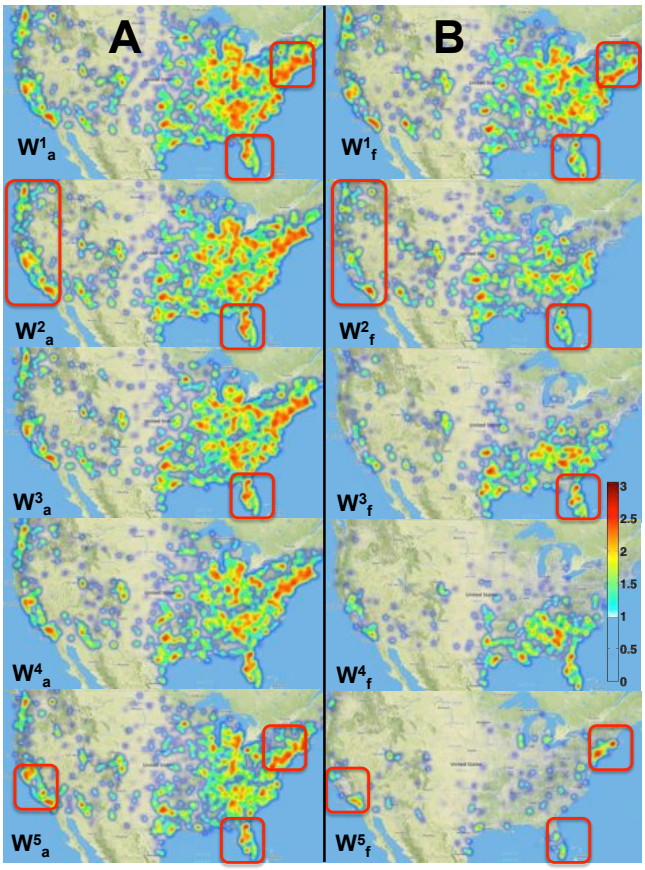


Figure 4: A visual summary of spatial patterns (W) in both (A) asthma and (B) flu diagnoses across the entire nation.

Dx data, we showed that it is possible to summarize the spatial and temporal patterns from these two conditions into a small number of categorical dimensions, each showing a distinct (temporal and spatial) signature with respect to the occurrence of these diseases. In particular, for the flu season of 2009, NMF was able to identify distinct temporal signatures that corresponded well with the early and late part of the flu. Additionally, we also showed a distinct lag in the peak times of asthma occurrence that preceded the peak time of influenza season. It is also interesting to note that this observation holds both in the early part of the year (May-Jun 2009 followed by a rise in the flu in Jun-July 2009) and the later parts of the year (Aug-Sep peak of asthma followed by a dramatic rise in the number of influenza infections in Sep-Oct 2009), across several parts of the US.

The analysis of the spatial patterns for flu and asthma revealed that there are distinct geographic locations (albeit a very small number of them) that perhaps show more than one temporal pattern in their occurrence data. Further analysis of these regions will be necessary to understand the origins of such "mixing". In particular, as part of our analysis, we did not examine patient age or history to understand how a specific group of patients (or a demographic) may be more susceptible to asthma or the flu in particular. Patients with one or more pre-existing respiratory conditions can be more susceptible to either flu or asthma and hence these factors would have to be taken into account to further understand the co-occurrence patterns observed during the 2009-2010 flu season. The spatial patterns that we discovered also highlight specific vulnerable regions within the US where the incidence of asthma and flu are inter-related. Although in this paper, we do not describe the many confounding factors (e.g., environmental factors/ climate factors that have a strong influence on the occurrence of asthma) that may play a role in the co-occurrence of asthma and flu, the ability to discover such complex associations from eHCRs provides an added capability for public health surveillance systems to monitor and quickly identify vulnerable geographic areas/population for preemptive intervention.

We must note here that a more detailed analysis of the spatio-temporal patterns is required. In particular, for this paper we have not quantitatively examined how these temporal patterns match up against other known temporal mining algorithms and even other unsupervised machine learning techniques such as principal component analysis. Additionally, within the scope of this paper, we have not examined whether these patterns correspond to other well known algorithms such as Google Flu. Finally, we must also note that the predictive aspects of our algorithm have also not been fully explored for two reasons: (1) the data available to us is only from the 2009-2010 flu season and (2) it is difficult to obtain a baseline behavior based on a year that showed highly anomalous behavior in terms of the overall flu incidence across the entire country. We will explore these questions in greater detail in a following publication.

While diagnostic information (from Dx data) can be helpful for public health surveillance, additional analyses of the prescription datasets (Rx) from IMS Health is necessary to obtain more accurate information regarding epidemic spread. The prescription transactions record the dosage and medicines provided to a patient and hence can provide tighter bounds on the number of estimated people infected and measure the intensity of spread. Such a collective integration of Dx and Rx datasets can provide novel insights not only in the context of understanding the flu, but can have a wide impact in general for more complex disease etiologies and chronic disease conditions.

The analytic techniques outlined here are part of the data analytic platform for public health surveillance that we have been

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

ACKNOWLEDGEMENTS

The preparation of this paper was funded in part by the Defense Threat Reduction Agency (DTRA) under the interagency agreement with Department of Energy (DOE) for DOE proposal number 2276-V643-13 as authorized by DOE contract number DE-AC05-00OR22725 at Oak Ridge National Laboratory (ORNL). The contents of this publication are the responsibility of the authors and do not necessarily represent the official views of DTRA.

1. REFERENCES

[1] *Flu Activity and Surveillance: Reports and Surveillance Methods in the United States*. Centers for Disease Control, 2009.

[2] *The 2009 H1N1 Pandemic: Summary Highlights, April 2009-April 2010*. Centers for Disease Control (CDC), June 2010.

[3] M. Aitken, E. R. Berndt, and D. M. Cutler. Prescription drug spending trends in the united states: Looking beyond the turning point. *Health Affairs*, 28(1):w151–w160, 2009.

[4] G. Alexander, N. Sehgal, R. Moloney, and R. Stafford. National trends in treatment of type 2 diabetes mellitus, 1994-2007. *Archives of Internal Medicine*, 168(19):2088–2094, 2008.

[5] C. Atkins, A. Patel, J. T. A. Taylor, M. Biggerstaff, T. L. Merlin, S. Dulin, B. Erickson, R. Borse, R. Hunkler, and M. Meltzer. Estimating effect of antiviral drug use during pandemic (h1n1) 2009 outbreak, united states. *Emerging Infectious Diseases*, 17(9):1591, 2011.

[6] T. Bodenheimer, E. Chen, and H. D. Bennett. Confronting the growing burden of chronic disease: Can the u.s. health care workforce do the job? *Health Affairs*, 28(1):64–74, 2009.

[7] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. Digital disease detection — harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157, 2009. PMID: 19423867.

[8] A. Chiolero, V. Santschi, and F. Paccaud. Public health surveillance with electronic medical records: at risk of surveillance bias and overdiagnosis. *The European Journal of Public Health*, 23(3):350–351, 2013.

[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 02 2009.

[10] A. Hersh, M. Stefanick, and R. Stafford. National use of postmenopausal hormone therapy: Annual trends and response to recent evidence. *J Amer Med Assoc*, 291(1):47–53, 2004.

[11] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, 02 2008.

[12] L. M. Lee and S. B. Thacker. Public health surveillance and knowing about health in the context of growing sources of health data. *American Journal of Preventive Medicine*, 41(6):636 – 640, 2011.

[13] A. H. Liu, R. Jaramillo, S. H. Sicherer, R. A. Wood, S. A. Bock, A. W. Burks, M. Massing, R. D. Cohn, and D. C. Zeldin. National prevalence and risk factors for food allergy and relationship to asthma: Results from the national health and nutrition examination survey 2005-2006. *Journal of Allergy and Clinical Immunology*, 126(4):798 – 806.e14, 2010.

[14] D. M. Morens, G. K. Folkers, and A. S. Fauci. The challenge of emerging and re-emerging infectious diseases. *Nature*, 430(6996):242–249, 07 2004.

[15] J. Myers, T. R. Frieden, K. M. Bherwani, and K. J. Henning. Ethics in public health research. *American Journal of Public Health*, 98(5):793–801, 2014/06/23 2008.

[16] I. of Medicine (US). *Global Infectious Disease Surveillance and Detection: Assessing the Challenges—Finding Solutions, Workshop Summary*. National Academies Press (US), 2007.

[17] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Comput Biol*, 9(10):e1003256 EP –, 10 2013.

[18] P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics Int. Lab. Sys.*, 37:23–35, 1997.

[19] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[20] A. Patwardhan and R. Bilkovski. Comparison: Flu prescription sales data from a retail pharmacy in the us with google flu trends and us ilinet (cdc) data as flu activity indicator. *PLoS ONE*, 7(8):e43611, 08 2012.

[21] D. Radley, S. Finkelstein, and R. Stafford. Off-label prescribing among office-based physicians. *Archives of Internal Medicine*, 166(9):1021–1026, 2006.

[22] A. Ramanathan, L. Pullum, C. Steed, S. Quinn, and C. Chennubhotla. Oak ridge bio-surveillance toolkit:. In *IEEE VAST Workshop on Public Health's Wicked Problems: Can InfoVis Save Lives?*, 2013.

[23] C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451, 2006.

[24] R. A. Weiss and A. J. McMichael. Social and environmental risk factors in the emergence of infectious diseases. *Nat Med*, 10:S70–S76, 2004.

KDDBHI Workshop

Big Data Analytic Technology For Bioinformatics and Health Informatics (KDDBHI)

For better quality of life and healthier world

Welcome to join us in KDDBHI, the premier international forum for professionals, researchers, clinicians and data scientists in the field of Bioinformatics and Health Informatics to exchange ideas and share research results as well as experiences.

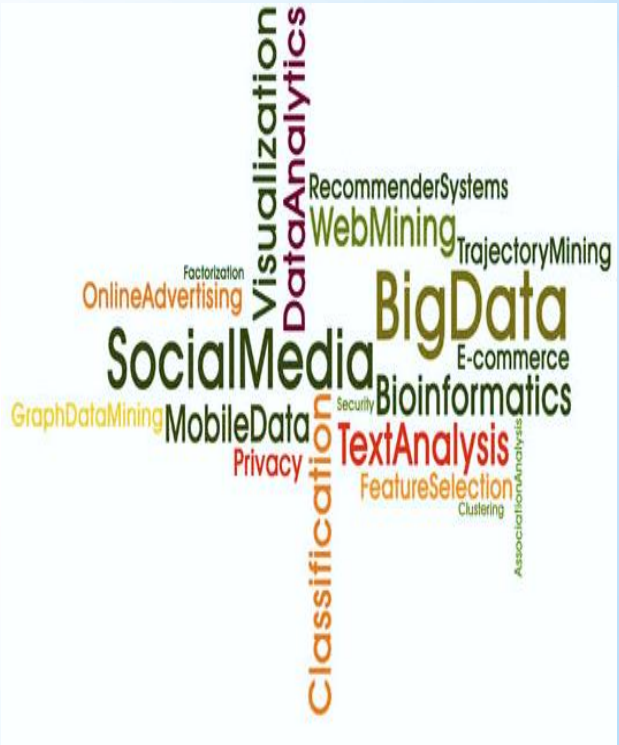
Organization Chairs:

Dr. Xin Deng, Research Scientist, LexisNexis | Risk Solutions | Healthcare, Orlando, FL

Dr. Donghui Wu, Senior Director, Statistical Modeling, LexisNexis | Risk Solutions | Healthcare, Orlando, FL

Featured Topics:

- Healthcare and healthcare delivery
- Healthcare policy research
- Healthcare outcomes research, monitoring and evaluation
- Health Analytics and Informatics
- Hospital Information System
- Electronic Medical Record and Electronic Health Record
- Population Health and Public Health Management
- Mobile Health and Sensor Applications
- Other areas related to healthcare
- Protein structure prediction
- Protein function analysis
- Drug design
- RNAseq and microarray gene expression data analysis
- Gene regulatory network construction
- Next-generation sequencing(NGS) analysis
- Functional Genomics
- Population, Evolution, and Comparative Genomics
- Transnational Bioinformatics
- Other areas related to proteomics and genomics
- Other related areas with applications in big data technology



Workshop Website: http://www.kddbhi.com/