Spatio-temporal Analysis for New York State SPARCS Data

Xin Chen, MS, Yu Wang, BS, Elinor Schoenfeld, PhD, Mary Saltz, MD, Joel Saltz, MD, PhD, Fusheng Wang, PhD
Stony Brook University, Stony Brook, NY

Abstract

Increased accessibility of health data provides unique opportunities to discover spatio-temporal patterns of diseases. For example, New York State SPARCS (Statewide Planning and Research Cooperative System) data collects patient level detail on patient demographics, diagnoses, services, and charges for each hospital inpatient stay and outpatient visit. Such data also provides home addresses for each patient. This paper presents our preliminary work on spatial, temporal, and spatial-temporal analysis of disease patterns for New York State using SPARCS data. We analyzed spatial distribution patterns of typical diseases at ZIP code level. We performed temporal analysis of common diseases based on 12 years' historical data. We then compared the spatial variations for diseases with different levels of clustering tendency, and studied the evolution history of such spatial patterns. Case studies based on asthma demonstrated that the discovered spatial clusters are consistent with prior studies. We visualized our spatial-temporal patterns as animations through videos.

Introduction

Open data initiatives supported by the governments are providing unprecedented information about our health. New York State SPARCS (Statewide Planning and Research Cooperative System[1]) data, for example, collects patient level detail on patient characteristics, diagnoses and treatments, services, and charges for each hospital inpatient stay and outpatient (emergency department, ambulatory surgery, and outpatient services) visit. More examples include a data set from CMS (Centers for Medicare & Medicaid Services) that contains information about providers who participate in Medicare[20] and New York State Cancer Mapping dataset[2,3] that consists of the number of people diagnosed with cancer (cancer counts, 2005-2009) in small geographic areas.

All these datasets provide street level location information for each patient, healthcare provider or facility site. These datasets cover a history of patient records, which also makes it possible for longitudinal analysis of historical disease patterns. The improved availability of health data combined with improved geospatial analysis and spatial statistics techniques has significant potential to uncover the spatial, and spatial-temporal patterns of diseases in a population at community level and provide new insights as to their causes and controls.[17-18]

Spatio-temporal data analysis for public health has a strong focus on locating patients and the agents of disease, studying the region level patterns and variations, and assessing the spatio-temporal trends on diseases and human health[8]. In the past, due to limited accessibility of heath data, public health studies were often limited at global level, and may not allow public health researchers and officials to adequately identify most at-risk populations, analyze, and monitor health events with fine-grained spatial resolutions, such as at the community or neighborhood level[12-16].

Our goal is to integrate open health data with a comprehensive set of spatio-temporal exposure data, which ranges from levels of various environmental pollutants to the socioeconomic status of persons at risk[3]. We focus on the spatio-temporal public health research with a fine-grained spatial resolution and consolidate a variety of spatial datasets into a data warehouse for scalable integrative spatial and spatial-temporal analytics[6].

In this paper, we introduced our preliminary study of spatial, temporal, and spatio-temporal analysis of disease patterns for New York State SPARCS data at the ZIP code level. The approach is generic and can be applied to finer spatial resolution at address level through geocoding patients' addresses and approximating them as census block group identifiers, which is an ongoing project.

The paper is organized as follows. We first present an overview of New York State SPARCS data with basic statistics for inpatient stay, emergency department visit, ambulatory surgery and outpatient visit, for the year 2014 (Table 1, Figure 1). We then studied spatial distributions for the top ranking diseases by discharge count, for the year 2014 (Table 2 and Figure 2-4). We then performed spatial clustering of asthma for the year 2014 (Figure 5) and analyzed the temporal trends for a selected group of diseases for the year 2003-2014 (Figure 6-8). At the end, we demonstrated our results as animation videos to visualize how the spatial clusters of asthma varies over time for the year 2005-2014, followed by discussions and conclusion.

Methods

Overview of New York State SPARCS Data

Population. In this work, we used four types of New York State SPARCS data according to the discharge claim type, namely inpatient stay (IP), emergency department visits (ED), ambulatory surgery (AS), and outpatient visits (OP) with basic demographics for the year 2014 given in Table 1. Any New York State healthcare facility certified to provide inpatient services, ambulatory surgery services, emergency department services or outpatient services is required to submit data to SPARCS. The purpose of SPARCS was to create a statewide data set to contribute to the goal of providing high quality medical care by serving as an information source[4].

Table 1. Demographics of New York State SPARCS data, 2014

Population/Discharge occurrence		Census*	Inpatient Stay	Emergency Department	Ambulatory Surgery	Outpatient Visit
		19,795,791	2,298,756	7,356,608	2,443,416	11,033,814
Age and Sex	Persons under 5 years	6.0%	12.5%	8.6%	1.8%	5.1%
	Persons under 18 years	21.3%	15.4%	20.2%	5.3%	13.3%
	Persons 65 years and over	15.0%	34.2%	13.9%	32.5%	24.5%
	Female persons	51.4%	56.2%	55.1%	55.9%	59.2%
Race	White alone	70.1%	57.4%	47.1%	65.8%	37.0%
	African American alone	17.6%	18.5%	25.5%	10.6%	23.8%
	American Indian and Alaska Native alone	1.0%	0.3%	0.2%	0.4%	0.4%
	Asian alone	8.8%	3.8%	2.5%	3.0%	3.4%
	Native Hawaiian and Other Pacific Islander	0.1%	0.0%	0.0%	0.0%	0.0%
	Two or More Races	2.4%	0.4%	0.3%	0.5%	0.1%
	Other Race or Unknown	-	19.6%	24.4%	19.7%	35.3%

* Estimates, July 1, 2015

Disease Categories. While the SPARCS data contains a comprehensive list of diagnoses and treatment procedure code for each discharge record. This paper focused on analyzing the spatio-temporal trends on disease categories based on the principal diagnostic code. The 'Principal/Primary Diagnosis' is the condition established after study to have been chiefly responsible for occasioning the admission of the patient to the hospital for care.[4] We used the Clinical Classifications Software (CCS) for grouping patient diagnoses into a manageable number of clinically meaningful categories. Figure 1 included the hospital discharge counts for the year 2014 for major disease groupings while the rest of this paper used more detailed disease categories from the single-level CCS.[5]

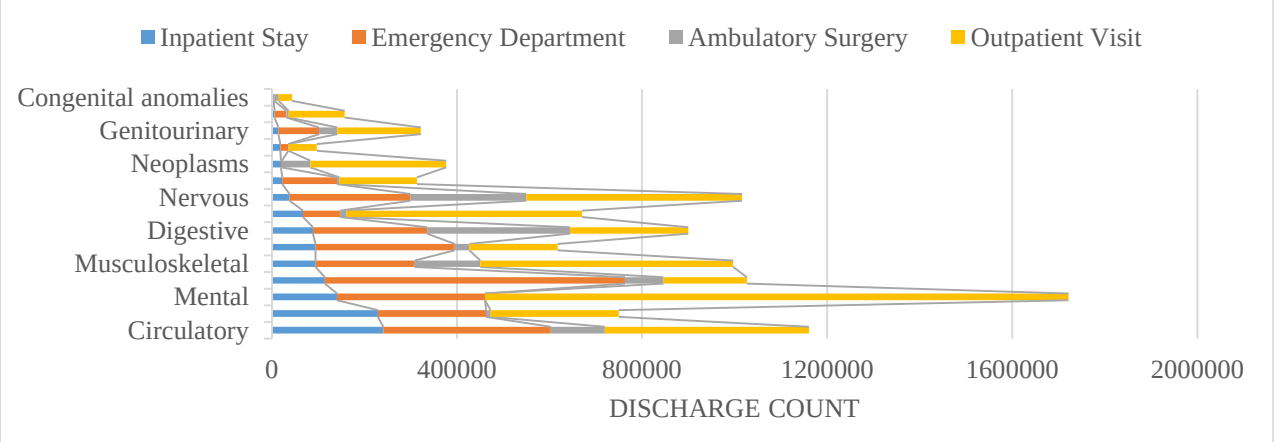

Figure 1. Occurrence Counts for Major Disease Groupings by Discharge Claim Type, New York State, 2014

Patients' Home Address and Hospital Admission Year. We approximate patients' home location by combining the 5-digit ZIP code number from the patients' home address and geographic data from TIGER/LINE data.[7] We then aggregated disease occurrences at ZIP code level and generated disease regional counts for the following spatio-temporal analyses. To evaluate the temporal trends, this work used the hospital admission year for the time range 2005-2014 for IP discharge and 2003-2014 for ED, AS, and OP discharge.

Spatio-temporal Analysis

We first used global spatial clustering analysis to determine whether a disease is dependent to patients' home locations. As a starting point to access all disease mappings of New York State, we examined the top ten diseases of inpatient stays and emergency department visits according to the discharge counts. We then checked the spatial variation and local spatial clustering for exemplary diseases according to their degree of spatial dependence. At last, we analyzed detailed temporal and spatio-temporal trends for one exemplary disease (asthma inpatient stay).

Hospital Discharge Rate. We used the hospital discharge rates to measure the probability of occurrence of a given disease in a population within a specified period of time. We calculated hospital discharge rates for inpatient stays, emergency department visits, ambulatory surgery, and outpatient visits respectively. The hospital discharge rate was calculated through dividing discharge counts by population counts from Census data[7]. In this paper, we evaluated both statewide rate and rates at ZIP code level. The discharge rates provide useful information about how common a disease is when compared to other diseases, or how common a disease in a specific location is as compared to the global baseline.

Table 2. Spatial Autocorrelation of Top Ten Diseases by Discharge Count for Inpatient Stay and Emergency Department Visit, New York State, 2014

Disease Name		Total Discharge Count	Hospital Discharge Rate Mean (St. Dev.)	Moran's I Index
Inpatient Stay (IP)	**Liveborn**	222,803	100 (165)	0.01
	Osteoarthritis	54,367	44 (93)	0.46*
	Congestive heart failure (non-hypertensive)	45,722	25 (34)	0.64*
	Mood disorders	43,209	122 (230)	0.54*
	Other complications of birth; puerperium affecting management of mother	36,480	15 (30)	0.70*
	Cardiac dysrhythmias	35,297	22 (40)	0.61*
	Complication of device; implant or graft	33,305	20 (34)	0.62*
	Diabetes mellitus with complications	33,040	15 (31)	0.70*
	Asthma	32,505	10 (23)	0.78*
	Acute myocardial infarction	31,249	22 (36)	0.53*
Emergency Department (ED)	Abdominal pain	342,294	189 (210)	0.05*
	Nonspecific chest pain	300,623	197 (449)	0.02
	Asthma	158,175	52 (191)	0.04*
	Other non-traumatic joint disorders	137,937	58 (73)	0.25*
	Other complications of pregnancy	134,195	49 (247)	0.02*
	Other injuries and conditions due to external causes	111,137	63 (114)	0.04*
	Other viral infections* **	109,746	35 (48)	0.38*
	Sprains and strains	109,429	81 (211)	0.02
	Superficial injury; contusion	102,582	79 (222)	0.03**
	Other gastrointestinal disorders	93,094	49 (140)	0.01

***** Significant at 1% confidence interval

****** Significant at 5% confidence interval

*** other viral infections include herpes zoster infection, herpes simplex infection, and other and unspecified viral infection.

Global Spatial Clustering. In spatial statistics, spatial autocorrelation measures how much close objects are in comparison with other close objects. To test whether there is spatial autocorrelation, this work used Moran's I (Table 2). Moran's I is a widely used global cluster test, which determines the degree of clustering or dispersion within a data set. The resulting values may range from 1 (perfect correlation), 0 (complete spatial randomness) to -1 (perfect dispersed).[8] For the hospital discharge rates, a positive spatial autocorrelation means that the areas with high discharge rates are close to other areas with high discharge rates.

Local Spatial Clustering. In addition to the global cluster test (with Moran's I index in Table 2) and visual analysis for mapping hospital discharge rates (in Figure 2-4), we then took cluster and outlier analysis with local Moran's I statistics[8] to quantitatively detect local clusters for the asthma IP rates (Figure 5). The local Moran's I statistics is a local cluster test that, given a set of weighted features, identifies statistically significant hot spots, cold spots, and spatial outliers.

Temporal Analysis and Spatio-temporal Animation. As a starting point for spatio-temporal analysis for SPARCS data, we used the admission dates to examine the temporal trends of the top ten IP diseases (excluding liveborn) for the year 2003-2014 in Figure 6-8. To observe patterns that emerge in the mapping of discharge rates as time passes, we made spatio-temporal animations to visualized how the spatial clusters varies over time for the year 2005-2014.

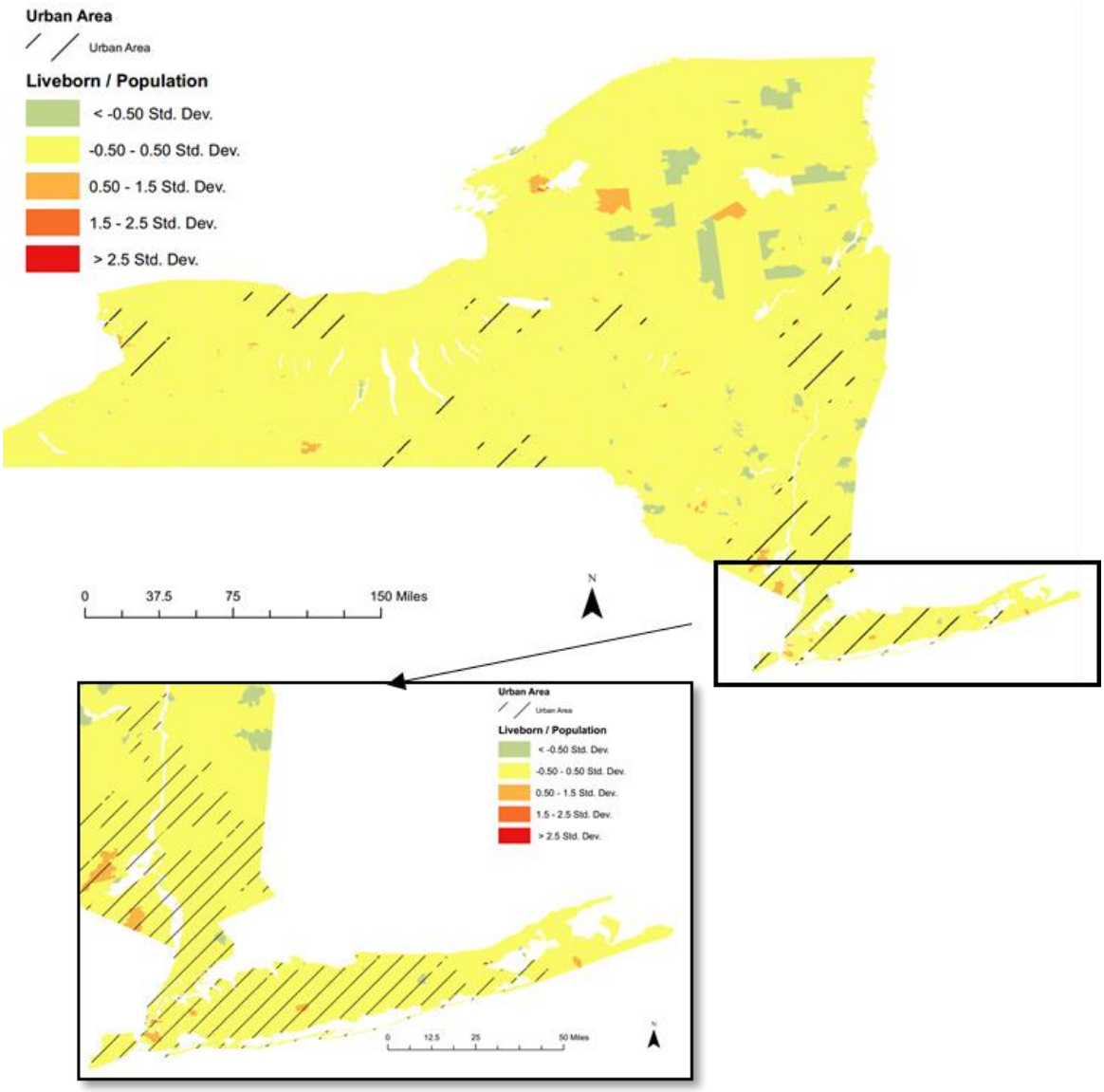

Figure 2. Liveborn Inpatient Discharge Rate per 10,000 Residents by ZIP Code, New York State, 2014

Results

Spatial Trends

According to the Moran's I indexes and the significant levels (Table 2), four types of disease discharge rates were spatially random, including IP liveborn, ED nonspecific chest pain, ED sprains and strains, and ED other gastrointestinal disorders (constipation, dysphagia, and other and unspecified gastrointestinal disorders). For the top ten IP diseases, asthma had the highest Moran's I index that indicates the highest spatial clustering tendency. For the top ten ED diseases, other viral infections (herpes zoster infection, herpes simplex infection, and other and unspecified viral infection) showed the highest spatial clustering tendency.

As shown in the choropleth map Figure 2-4, all ZIP code areas were classified into five classes according to their discharge rates. Class breaks were created with equal value ranges at intervals of one standard deviations using mean values and the standard deviations from the mean.

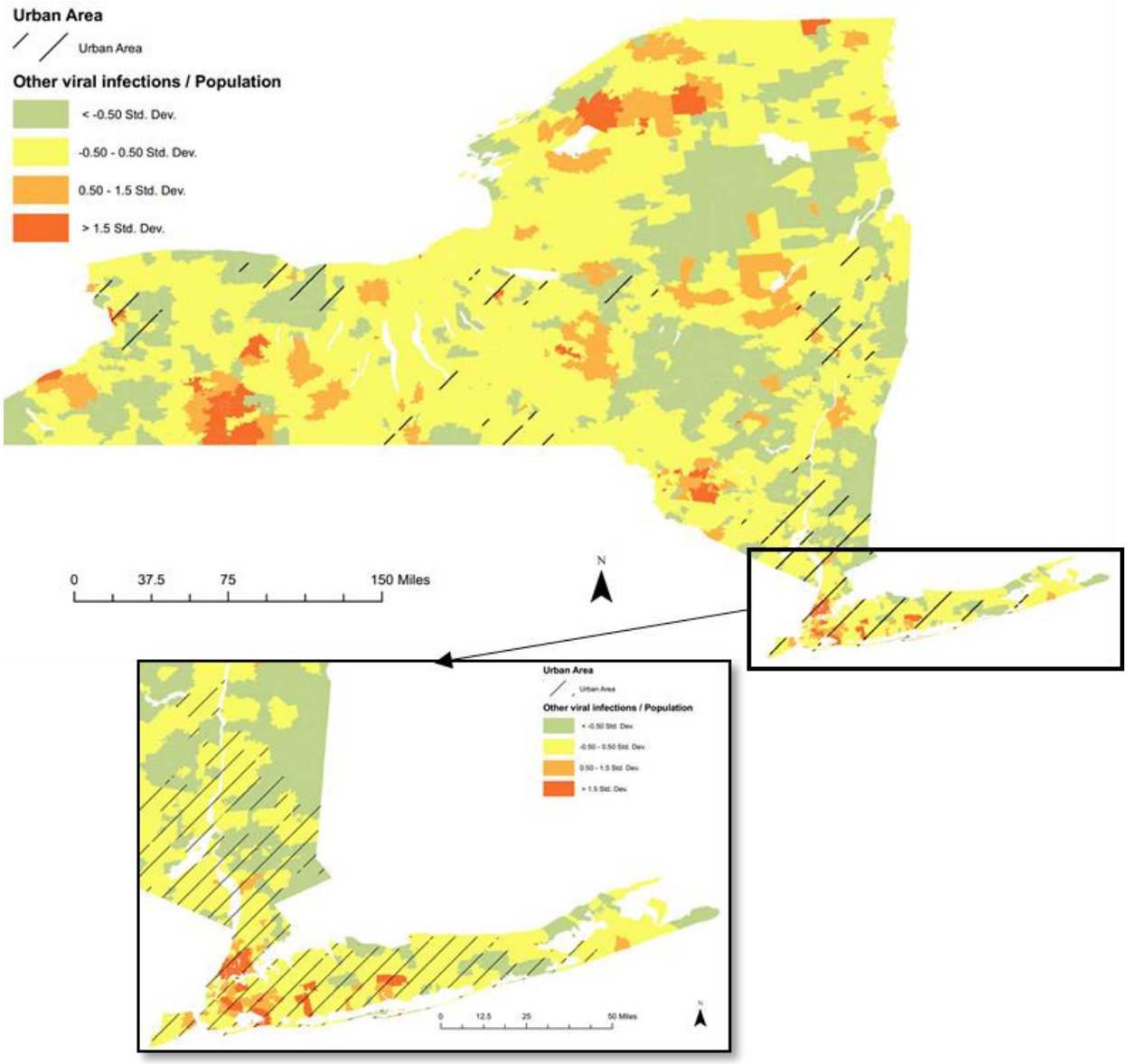

Figure 3. Viral Infections Emergency Department Visit Rate per 10,000 Residents by ZIP Code, New York State, 2014

Figure 2 visualized the spatial distribution of liveborn inpatient discharge rate by ZIP code, New York State (NYS), 2014. The liveborn rates for the majority part of NYS were within one standard deviation from the mean. Only a very small number of ZIP code areas were hotspots (areas with liveborn rate higher than 1.5 std. dev. in orange and red color in the map). This implied that the occurrence of liveborn was independent from the patients' home locations.

While Figure 2 showed an example for spatially independent disease, we used viral infection ED visit rate (Figure 3) and asthma inpatient discharge rate (Figure 4) as two examples for spatially dependent diseases. Both the viral infection (Figure 3) and asthma (Figure 4) exhibited a tendency of spatial clustering. Compared to the spatial distribution of liveborn rate, the maps of viral infection and asthma contained more hotspots (areas with rates higher than 1.5 std. dev. in orange and red color in the map), which indicates that the cases of viral infection or asthma more likely occurred near each other.

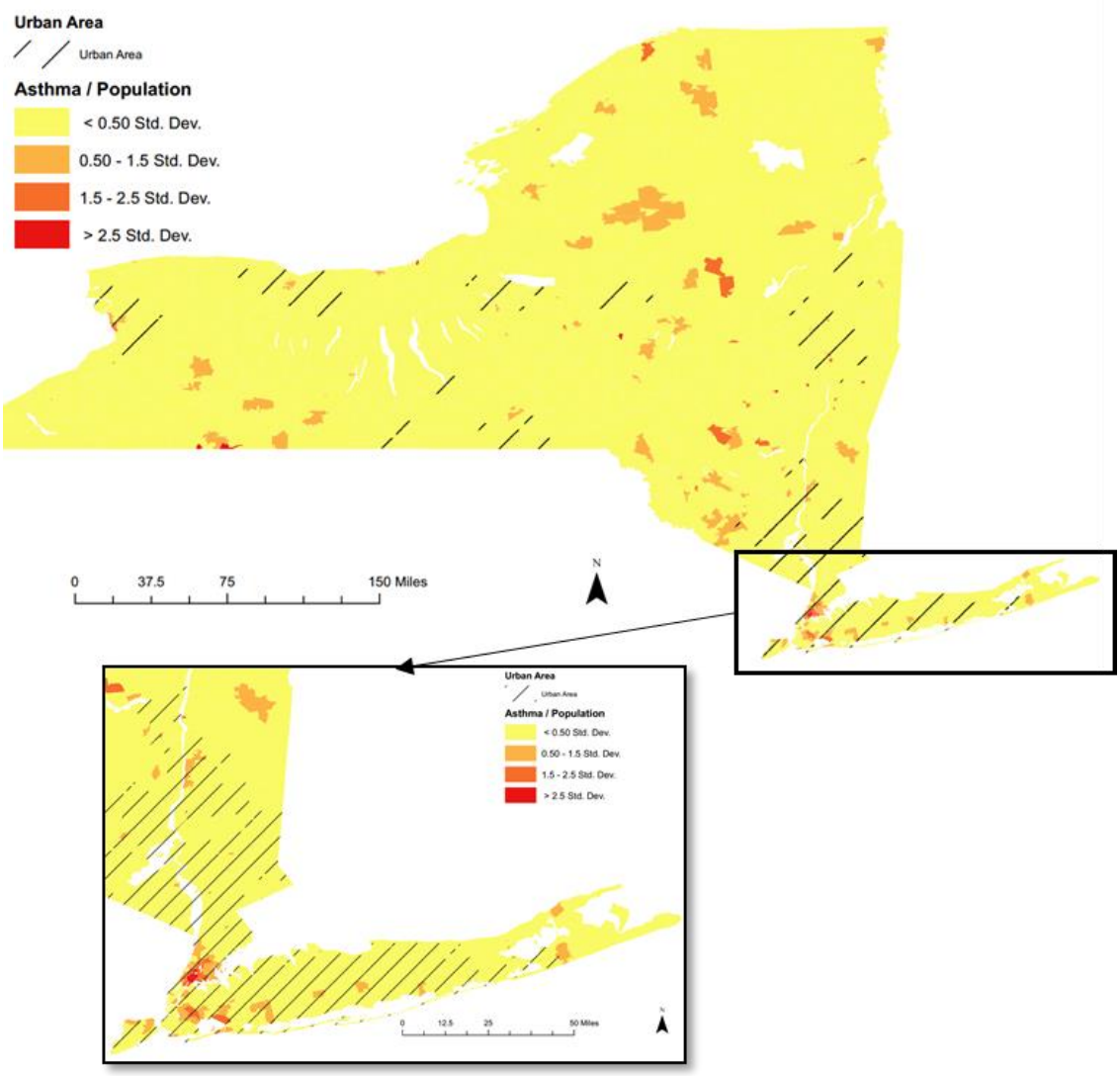

Figure 4. Asthma Inpatient Discharge Rate per 10,000 Residents by ZIP Code, New York State, 2014

We can see that the mapping of viral infections ED rates has no extreme hotspots (areas with rate higher than 2.5 std. dev. in red color in the map), which indicates a more equal distribution of the rates. The asthma IP rates, on the other hand, have more extreme hotspots, which indicates a tendency of more uneven spatially distribution. From Table 2, we can also tell that the Moran's I index values of ED visit rates are generally lower than those of IP discharge rates. It means that the ED visits rates generally have a lower degree of clustering tendency than IP discharge rates.

For the rural-urban difference, we cannot find a rural-urban disparity for viral infections or asthma. Most parts of rural areas and urban areas (areas in the map with line fill symbol) have no hotspots. The few cases of hotspots and higher

rates (with values higher than 1.5 std. dev.) areas seem more likely appear at New York City. With manual checking, we found the two hotspots areas for asthma are an area near JFK airport and upper Manhattan.

The cluster/outlier type field in Figure 5 distinguishes between a statistically significant cluster of high values (High-High cluster), cluster of low values (Low-Low cluster), outlier in which a high value is surrounded by low values (High-Low outlier), and outlier in which a low value is surrounded by high values (Low-High outlier). Statistical significance is set at the 95 percent confidence level. We applied the False Discovery Rate (FDR) correction to reduce this p-value threshold from 0.05 to a value that better reflects the 95 percent confidence level given multiple testing. The FDR procedure will potentially reduce the critical p-value in order to account for multiple testing and spatial dependency.[8]

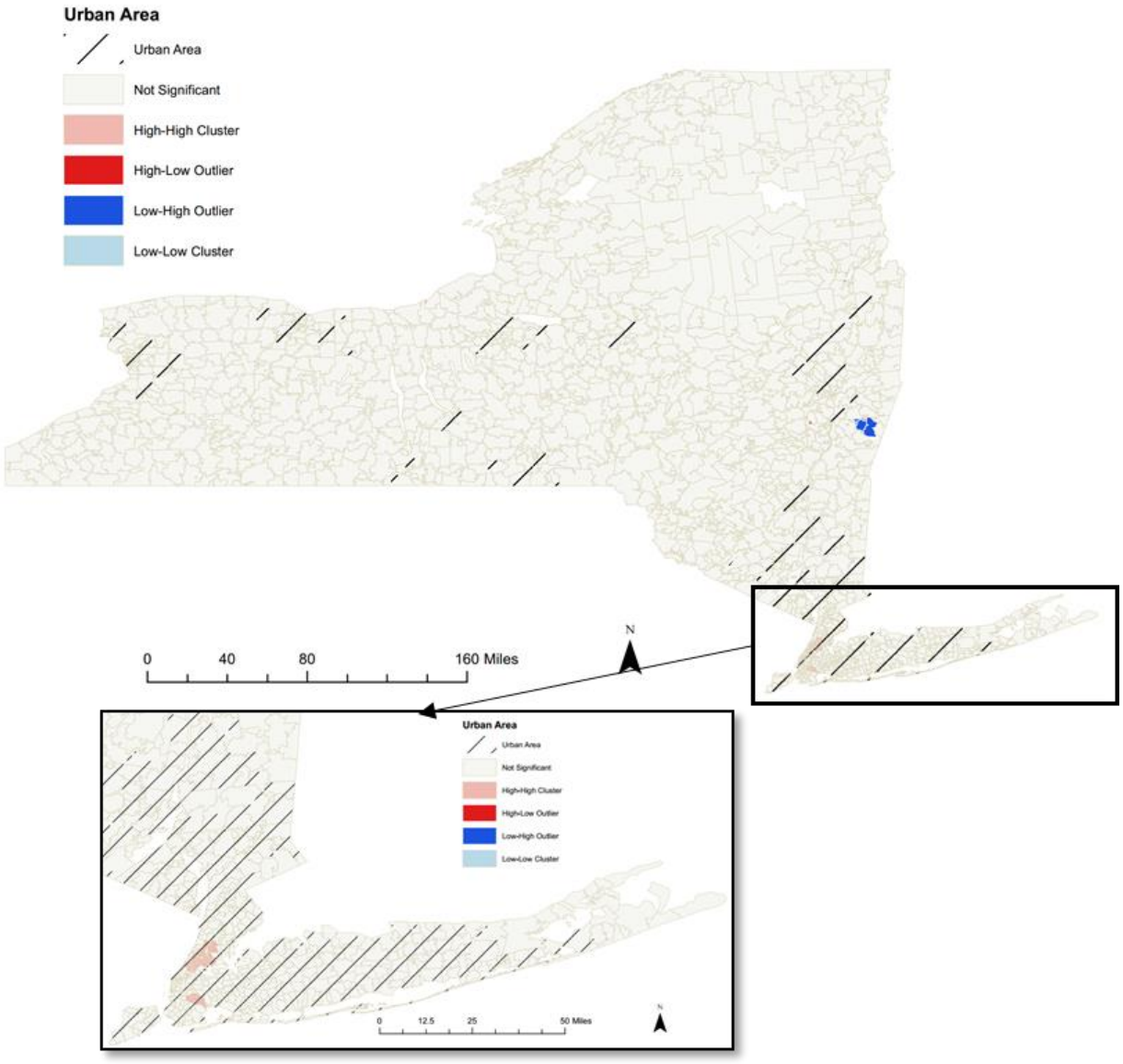

Figure 5. Spatial Clusters and Outliers of Asthma Inpatient Discharge Rate by ZIP Code, New York State, 2014

As shown in the mapping of spatial clustering for asthma IP rates, two areas located at near JFK airport and upper Manhattan were identified as the High-High cluster (in pink color in Figure 5), which represent statistically significant clusters of higher asthma IP rates. Such results confirmed the visual analysis for spatial distribution map in Figure 4 and prior research about the potential health impact of residential proximity to large NYS airports[9-10]. Many High-

Low outliers and Low-High outliers were also identified near the Albany city in northeastern New York State. Such finding, however, may require further research for the potential driving factors.

Temporal Trends

In Figure 6, most diseases have a down trend over the past decade (2005-2014). The osteoarthritis and congestive heart failure (non-hypertensive), however, have a rising trend which may require further research for the potential driving factors. We then took a close look at the temporal trends for asthma with different claim types (Figure 7). While inpatient stay, ambulatory surgery, and outpatient visits shared a similar down trends, the emergency department visit rates were rising over the past years. For inpatient discharge rates, we then broke down the asthma into its six subcategories (Figure 8) and found that the asthma subcategory (asthma other than chronical obstructive asthma with acute exacerbation) also had a rising trend. All these findings require further research for the potential driving factors.

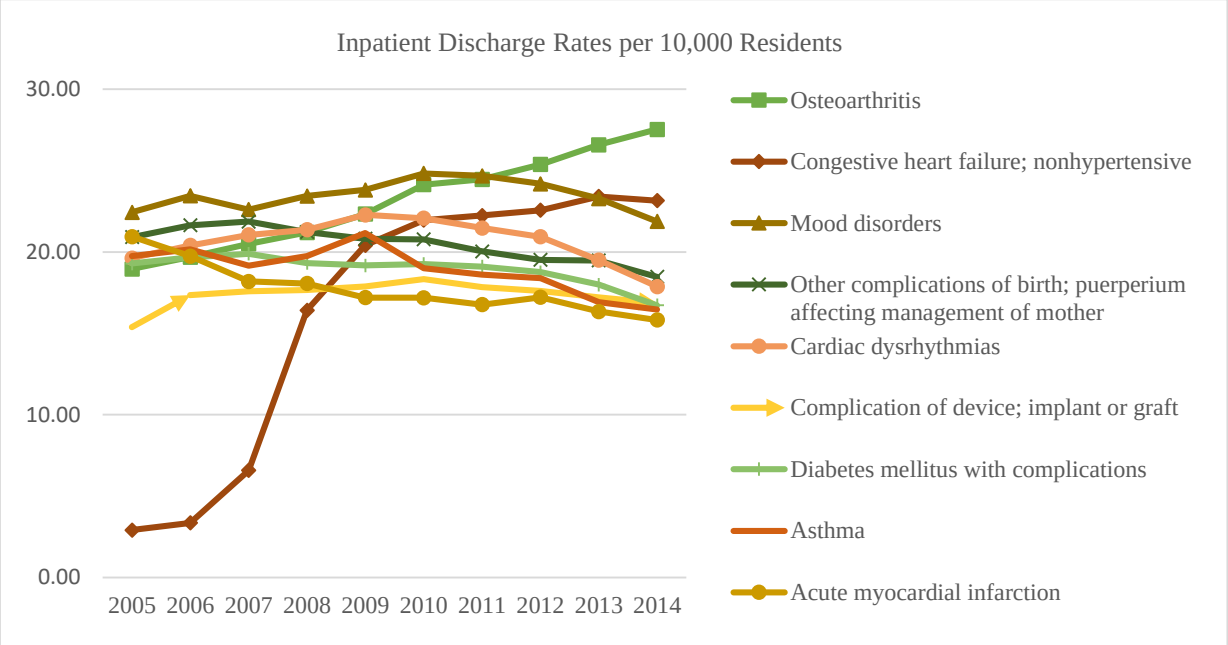

Figure 6. Temporal Trends of Top Ten Diseases by Discharge Count for Inpatient Stay, New York State, 2005-2014

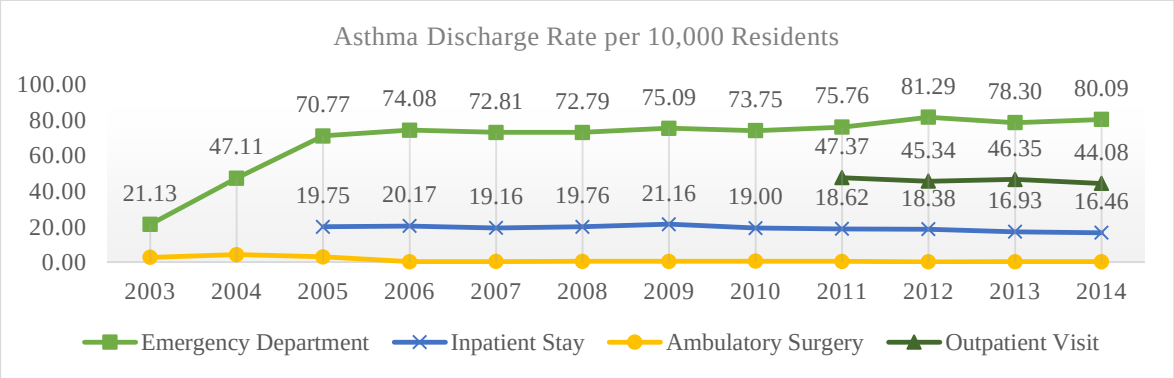

Figure 7. Temporal Trends of Asthma Discharge Rate for Four Discharge Claim Types, New York State, 2003-2014

Spatio-temporal Animation

We made two exemplary animations of spatial clusters for the year 2005-2014 for New York State https://vimeo.com/183126416 and New York City and Long Island area https://vimeo.com/183126718 separately. At the statewide scale, the distribution of spatial clusters over New York State changed more dramatically than that over

New York City and long island areas. For the areas near JFK airport, the spatial clusters with high rates emerged in 2006, disappeared during 2008-2012, and reappear after 2012. For the Low-High outliers and High-Low outliers appearing near the Albany city in northeastern New York State, we also observed a similar fluctuation trend over the years. To evaluate the potential factors that result in such fluctuation patterns, we will integrate SPARCS data with a comprehensive set of spatial impact factors in our future work, which ranges from levels of various environmental pollutants to demographical and socioeconomic status of persons at risk[3, 20].

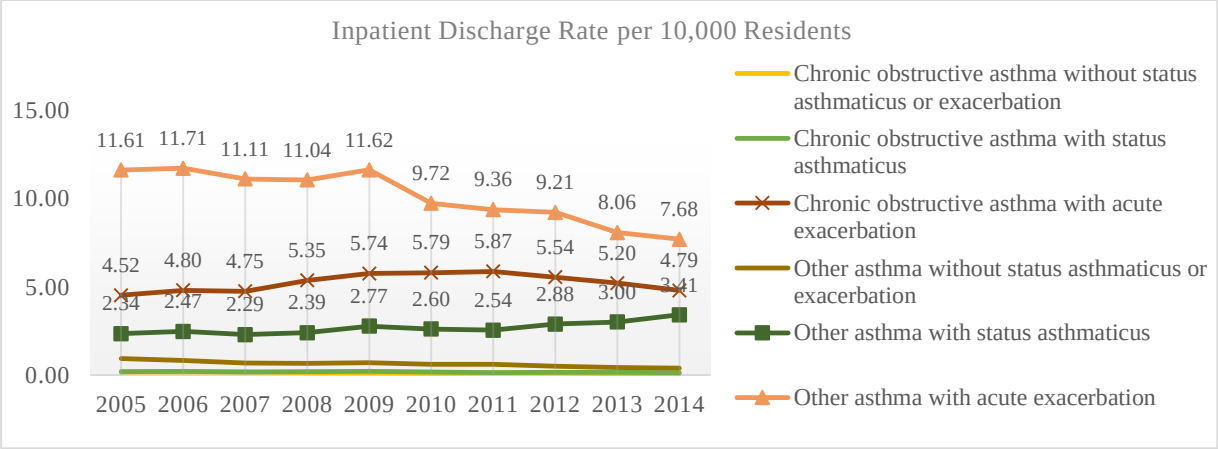

Figure 8. Temporal Trends of Inpatient Discharge Rate for Asthma's Six Subcategories, New York State, 2005-2014

Discussion and Future Work

This study provided our preliminary results of the spatio-temporal analyses on the New York State SPARCS data, which provides the analysis framework and acts as a baseline for more refined studies of spatio-temporal trends on diseases and human health in our future work. We examined the spatial variations for the top ranking diseases by different claim types and performed case studies on the spatial clustering patterns of asthma that draw consistent results with previous work.

While our long-term goal is to provide integrative spatial analytics at fine grained spatial resolution[3], this paper focused on the spatio-temporal analysis on the disease discharge rates at ZIP code level. Our future work will break down patients based on demographics, such as gender, age or age ranges, race and ethnicity groups. For discharge rates based patterns of the diseases, we could elaborate on age or race based analysis. More fined-grained temporal information such as patient admission or discharge dates could also be used for discovery of refined temporal patterns in our future work.

There has long been a demand for spatio-temporal data analysis at a fine geographic resolution for use in etiologic hypothesis generation, methodological evaluation and teaching. In our ongoing work, we are examining the comprehensive list of diagnoses and treatments, services, and charges for each hospital inpatient stay and outpatient visit. We will also take advantage of the street level location information and specific date information (admission date, procedure date, etc.) for each patient, healthcare provider and facility site.

After geocoding and approximating addresses into census block group identifiers, our framework for integrative spatial data analytics will provide spatial queries based on coordinates or boundaries, and enable integrating and correlating the health records with spatial exposure data at multiple resolutions. We will provide multi-dimensional analysis by grouping patients according to their demographic or socio-economic attributes. We will also study potential spatial clusters of disease distributions and correlations between disease risk and spatial impact factors. For example, we are interested in exploring potential hotspots of Hepatitis C or potential environment and weather factors that may have correlations with asthma.

Conclusion

Vast amounts of spatio-temporal big data are being increasingly generated and made available in the public health domain. Spatio-temporal analyses could provide new insights and create new forms of value to support community or neighborhood level public health studies. In this paper, we present our preliminary results of spatio-temporal

analysis for New York State SPARCS data. We focus on representative case studies for the top ranking diseases by discharge count. Our results provide much refined results on spatio-temporal trends of diseases, and demonstrate consistent spatial clustering patterns. The analysis framework we developed is generic and provides a foundation for advanced SPARCS data analysis at the community and neighborhood level in the future.

Acknowledgments

This work is supported in part by NSF ACI 1443054, by NSF IIS 1350885 and by NSF IIP1069147.

References

1. Statewide Planning and Research Cooperative System [Internet]. Health.ny.gov. 2016 [cited 9 March 2016]. Available from: https://www.health.ny.gov/statistics/sparcs/
2. Environmental Facilities and Cancer Mapping [Internet]. Health.ny.gov. 2016 [cited 9 March 2016]. Available from: https://www.health.ny.gov/statistics/cancer/environmental_facilities/mapping/
3. Chen X, Wang F. Integrative Spatial Data Analytics for Public Health Studies of New York State. To Appear in AMIA Annual Symposium Proceedings 2016. American Medical Informatics Association.
4. SPARCS Outpatient Data Dictionation [Internet]. 2016 [cited 9 September 2016]. Available from: https://www.health.ny.gov/statistics/sparcs/sysdoc/outpatientoutputdd.pdf
5. HCUP-US Tools & Software Page [Internet]. Hcup-us.ahrq.gov. 2016 [cited 9 September 2016]. Available from: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
6. Chen X, Vo H, Aji A, Wang F. High performance integrative spatial big data analytics. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data 2014 Nov 4 (pp. 11-14). ACM.
7. Branch G. TIGER/Line® - Geography - U.S. Census Bureau [Internet]. Census.gov. 2016 [cited 9 September 2016]. Available from: https://www.census.gov/geo/maps-data/data/tiger-line.html
8. Waller LA, Gotway CA. Applied spatial statistics for public health data. John Wiley & Sons; 2004 Aug 12.
9. [Internet]. 2016 [cited 9 September 2016]. Available from: http://www.health.ny.gov/statistics/ny_asthma/pdf/2013_asthma_surveillance_summary_report.pdf
10. Information on Asthma in New York State [Internet]. Health.ny.gov. 2016 [cited 9 September 2016]. Available from: http://www.health.ny.gov/statistics/ny_asthma/
11. Publications [Internet]. Health.ny.gov. 2016 [cited 9 September 2016]. Available from: http://www.health.ny.gov/environmental/public_health_tracking/program/publications
12. Jiang X, Cooper GF. A Bayesian spatio-temporal method for disease outbreak detection. Journal of the American Medical Informatics Association. 2010 Jul 1;17(4):462-71.
13. Flamand C, Fabregue M, Bringay S, Ardillon V, Quénel P, Desenclos JC, Teisseire M. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. Journal of the American Medical Informatics Association. 2014 Oct 1;21(e2):e232-40.
14. Mandal R, St-Hilaire S, Kie JG, Derryberry D. Spatial trends of breast and prostate cancers in the United States between 2000 and 2005. International Journal of Health Geographics. 2009 Sep 29;8(1):1.
15. Dijkstra A, Janssen F, De Bakker M, Bos J, Lub R, Van Wissen LJ, Hak E. Using spatial analysis to predict health care use at the local level: a case study of type 2 diabetes medication use and its association with demograpHic change and socioeconomic status. PloS one. 2013 Aug 30;8(8):e72730.
16. Kauhl B, Heil J, Hoebe CJ, Schweikart J, Krafft T, Dukers-Muijrers NH. The Spatial Distribution of Hepatitis C Virus Infections and Associated Indicators—An Application of a Geographically Weighted Poisson Regression for Evidence-Based Screening Interventions in Hotspots. PloS one. 2015 Sep 9;10(9):e0135656.
17. Richardson DB, Volkow ND, Kwan MP, Kaplan RM, Goodchild MF, Croyle RT. Spatial turn in health research. Science. 2013 Mar 22;339(6126):1390-2.
18. Kwan MP, editor. Geographies of health, disease and well-being: recent advances in theory and method. Routledge; 2016 Mar 23.
19. Medicare Provider Utilization and Payment Data - Centers for Medicare & Medicaid Services [Internet]. Cms.gov. 2016 [cited 21 September 2016]. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html
20. Oyana TJ, Rogerson P, Lwebuga-Mukasa JS. Geographic clustering of adult asthma hospitalization and residential exposure to pollution at a United States-Canada border crossing. American journal of public health. 2004 Jul;94(7):1250-7.