Scope

To find out the most-likely lead poisoned area for children (under 6) in the New York State (excluding NYC) from publicly reported Elevated Blood Lead Level (EBLL) data. In this study, the EBLL is defined to have BLL >= 10 ug/dL in patient's blood test.

Data Approach

Each raw data sets is downloaded from free, public online resources. The data are later stored and processed with PANDAS using Python2.7, and are charted/listed to be presented in this report using Google Doc.

Data Source 1: NY State Dept. of Health

URL:https://health.data.ny.gov/Health/Childhood-Blood-Lead-Testing-and-Incidence-of-Bloo/d54z-enu8

Quick Table Facts:

Field Name	Values Description	non-missing data (%)	Used for Study?
CountyCode	Numeric County Code (1-123)	100%	Yes
County	Descriptive County Name	100%	Yes
Zip	5-digit NY State Zip Code	100%	Yes
Year	4-digit Year (2003-2012)	100%	Yes
Tests	Total # of tested children in the zip code in the year	33.10%	No
10-15 mcg/dL	Total # of tested children with BLL 10-15 ug/dL (1 ug = 1 mcg/dL)	3.14%	No
15+ mcg/dL	Total # of tested children with BLL >15 ug/dL (1 ug = 1 mcg/dL)	3.14%	No
Total Elevated Blood Levels	Total # of tested children with BLL >=10 ug/dL	5.05%	Yes
Percent	% of elevated blood lead levels children	5.05%	Yes
County Location	(longitude,latitude) of county's location	2.18%	No

Further data investigation:

1. There are 2,692 distinct zip codes in this table, however, from other online resource (Zip-codes.com) it shows 2,156 zip codes in the NY state. Further investigation can be done to check zip codes validity with USPS zipcode database.
2. Only 5.05% of total records having 'Percent' values, in fact, only 5.31% of zip codes (143 out of 2692) having the EBLL percent results. Further investigation can be done to learn how this 5.31% was chosen and to estimate the cost and benefits for studying the rest of 95% uncovered area.

Results:

The approach is as follows: (1) Identify the highest EBLL rate area in 2012 at the zip code level. Plot their EBLL rate from 2003 to 2012 to observe the trend. (2) Aggregate the results at the county level. This is for cross validation with the second data resource.

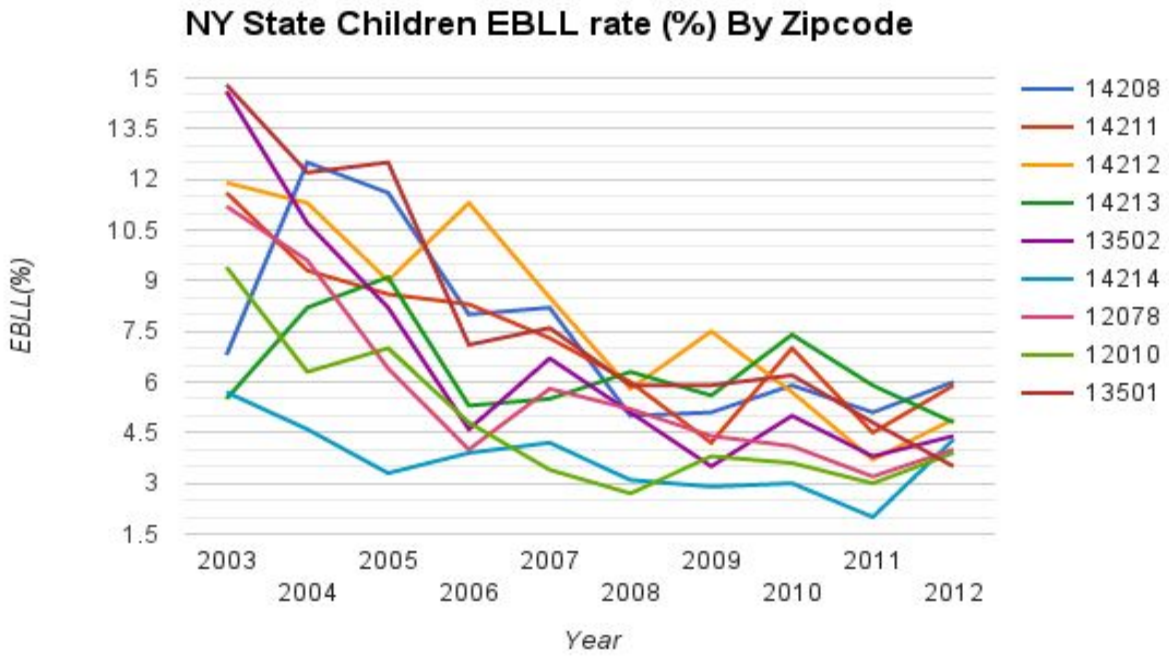

Figure 1: The top 10 Zip codes in NY State with highest children EBLL rate from 2003 to 2012. Each of records shows a trend that EBLL rate is decreasing from year 2003 to 2012. However, Zip codes 14208, 14211, 14212, and 14124 (all from Erie County, see Table 1) show a noticeable increment (1-2.5%) from year 2011 to 2012.

Table 1: The top 10 Zip codes with County Code and Name

CountyCode	CountyName	Zip
		14208
		14211
		14212
29	Erie County	**14213**
65	Oneida County	**13502**
29	Erie County	**14214**
117	Wayne County	**14489**
35	Fulton County	**12078**
57	Montgomery County	**12010**
65	Oneida County	**13501**

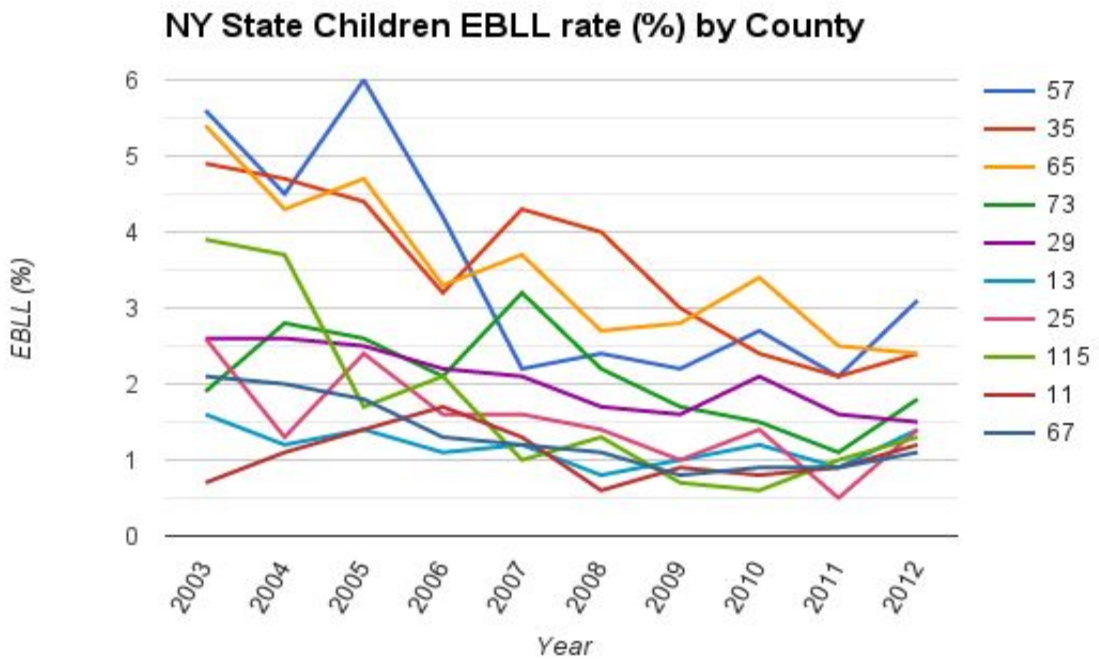

Figure 2: The top 10 County Codes in NY State with highest children EBLL rate from 2003 to 2012. Although in general the EBLL rate is decreasing in recent years, county code 57, 73, and 25 show a noticeable increment from year 2011 to 2012.

Data Source 2: CDC Lead State Surveillance Data - NY (2014)

URL: http://www.cdc.gov/nceh/lead/data/state/nydata.htm

Quick Table Facts:

Field Name	Values Description	non-missing data (%)	Used for Study?
County FIPS	3-digit County Code with leading zeros (001-123)	100%	Yes
County Name	Descriptive County Name	100%	Yes
# of Children Tested	Total # of tested children in the county	100%	No
Total # Children Tested 5-9 ug/dL	Total # of children with BBL between 5-9 ug/dL (potential EBLL case)	100%	No
Total Confirmed BLL >=10ug/dL (%)	Total # of children with BLL >=10ug/dL divided by # of tested children (%)	100%	Yes
Total Housing Units	Total # of housing units from the county from census 2000 data	100%	Yes
Pre-1950 Housing Units	Total # of housing units built before 1950 from the county from census 2000 data	100%	Yes

Further data investigation:

1. Although all 62 counties in NY state has tested data, the rates between tested children and general population are not the same. Further investigation can be done to adjust the interpretation by comparing the tested children and general population in each county.
2. Columns like 'Total # Children Tested 5-9 ug/dL' can be useful for the further analysis to re-evaluate the rate (with a new EBLL definition) to find out potential lead poisoning area earlier.

Results:

The CDC 2014 NY State surveillance data for lead poisoning is used to cross validate the findings from NY Dept. of Health data (see Figure 2). The CDC table is left joined to the original top 10 county table and to compare the EBLL results side-by-side.

Table 2: The top 10 county (from NY State Dept of Health) EBLL results compared with CDC BLL surveillance data in 2014. For NY State Dept data, column ‘avg EBLL’ is the average of EBLL % from 2003 to 2012; and column ‘pred EBLL’ is calculated as EBLL in 2012 plus the average of EBLL differences from 2003 to 2012. Data from CDC have higher EBLL rate (2-3 times) to the NY State Dept data, except for Delaware County in this table. ‘Pre-1950 Housing Unit’ are mostly greater or close to 50% among these counties.

		NY State Dept of Health			**CDC BLL Data 2014**	
County Code	CountyName	EBLL (%) in 2012	*avg EBLL (%)*	*Pred EBLL (%)*	EBLL (%) in 2014	Pre1950 House Unit (%)
57	Montgomery County	3.1	*3.5*	*2.82*	5.1	61.74
35	Fulton County	2.4	*3.54*	*2.12*	4.9	51.67
65	Oneida County	2.4	*3.52*	*2.07*	6.7	46.77
73	Orleans County	1.8	*2.09*	*1.79*	4.5	51.73
29	Erie County	1.5	*2.05*	*1.38*	4.2	44.61
13	Chautauqua County	1.4	*1.18*	*1.38*	3	55.88
25	Delaware County	1.4	*1.52*	*1.27*	0	39.14
115	Washington County	1.3	*1.73*	*1.01*	4.4	47.39
11	Cayuga County	1.2	*1.06*	*1.26*	3.3	50.62
67	Onondaga County	1.1	*1.32*	*0.99*	2.2	35.86

Conclusions

Erie County has 5 out of 10 zip code areas with the highest EBLL rates in 2012 from the NY State Dept of Health data. Among these areas, zip code 14214 (Buffalo) would be a good place to start with given its highest rate change (from 2.0% to 4.5%) from 2011 to 2012. The NY State Dept of Health data cover 5.31% of zip codes in the NY state and its county level EBLL rates are in general lower than the rates from CDC surveillance data.

Future Works

Data Platform

1. PANDAS is great to deal with aggregated tables and regression analysis. For data at granular levels (e.g. 4-million patients' blood test results in 10 years,) a relational database management system (RDMS) will be a better choice to store the data and to create aggregated results for PANDAS. Among RDMS, Amazon Redshift is a good candidate for its petabyte-scale performance (better than open source platforms) and its 'fee-for-services' feature (more economical than Teradata.) For data that is unstructured (e.g. MRI images, ECG time series,) the data is better to store in NoSQL systems.

2. Python is great for creating customized, good quality graphs for data visualization. For spatial visualization in this study, we can download NY Zip Code Tabulation Area (ZCTA) file and convert/compress it into a topojson file format to create the zip code area plot. Commercial softwares like Origin, or SAS EG can also create prompt interfaces for users to visualize the selected data without entering programming lines.

Data Credibility

1. After looking into the data, we found only 5.31% of zip codes in NY state have tested results. It is important to investigate how this 5.31% sample was selected, and how good are they representing the entire state.

2. Using second or more data sources to validate the original data. The data may need to be sampled or aggregated to be compared at the same level. Temporal data plots help to compare the trends and to identify abnormal results during the time interval.

Data Security/Privacy

1. The protected health information (PHI) data need to be de-identified for research purposes. HIPAA defines 2 de-identified methods : (1) Safe Harbor: to remove certain types of identifiers of an individual, and (2) Expert Determination: to apply statistical principles to 'mask' patient's identity. Some states may require more data privacy beyond HIPAA rules and we need to pay double attentions for these state-specific rules.

2. To transfer data we can use HIPAA-compliant data sharing services like Box, or CareCloud. Users need to double check how the data servers meet HIPAA and other security rules before storing/sharing the sensitive data.